Margin Error and Generalization Capabilities of Multi-Class Discriminant Systems

André Elisseeff, ERIC[1] Yann Guermeur, LORIA[2]
Hélène Paugam-Moisy, ERIC[3]

NeuroCOLT2 Technical Report Series

NC2-TR-1999-051-R

December 10, 1999[4]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

[1]email: aelissee@univ-lyon2.fr ERIC, Université Lumière Lyon 2, 5, avenue Pierre Mendès-France, 69676 Bron Cedex, France

[2]email: Yann.Guermeur@loria.fr LORIA INRIA-Lorraine, Campus Scientifique, B.P. 239, 54506 Vandœuvre-lès-Nancy cedex, France

[3]email: hpaugam@univ-lyon2.fr ERIC, Université Lumière Lyon 2, 5, avenue Pierre Mendès-France, 69676 Bron Cedex, France

[4]Received December 10, 1999, revised June 15, 2001

Abstract

The theory and practice of discriminant analysis have been mainly developed for two-class problems (computation of dichotomies). This phenomenon can easily be explained, since there is an obvious way to perform multi-category discrimination tasks using solely models computing dichotomies. It consists in dividing the problem at hand into several *one-against-all* ones and applying a simple rule to construct the global discriminant function from the partial ones. Adopting a direct approach, however, should make it possible to improve the results, let them be theoretical (bounds on the expected risk) or practical (values of the empirical risk and the expected risk). Although multi-category extensions of the main models computing dichotomies can often be conceived simply, as in the case of multi-layer perceptrons, in other cases, this cannot be done readily but at the expense of the loss of part of the theoretical foundations. This is for instance the main shortcoming of the multi-category support vector machines developed so far. One of the major difficulties of multi-category discriminant analysis rests in the fact that it requires specific uniform convergence results. Indeed, the uniform strong laws of large numbers established for dichotomies do not extend nicely to multi-category problems. They become significantly looser. This is problematical indeed, since the question of the quality of bounds is of central importance if one wants to implement with confidence the structural risk minimization inductive principle, which is precisely grounding the support vector method.

In this paper, building upon the notions of margin used in the context of statistical learning theory and boosting theory, and the corresponding generalization error bounds, we derive sharper bounds on the expected risk (generalization error) of multi-class vector-valued discriminant models. The main result is an extension of a lemma by Bartlett. After a discussion about the notion of margin and its use for two-class discriminant analysis, we derive the main theorem and its corollaries. We then show how to bound the capacity measure for sets of functions of interest and study specifically the case of the multivariate linear regression model, which is of particular importance, since it is directly related to multi-category support vector machines. Finally, the bound is assessed on a real-world biocomputing problem.

This work, which aims at establishing foundations for the statistical analysis of multi-class models, based on a new notion of margin, paves the way for the theoretical study of the existing multi-class support vector machines and the design of new ones.

1 Introduction

The study of the generalization capabilities of learning systems falls within the province of one of the four main parts of Vapnik's statistical learning theory [40]: the theory of bounds. Results about the uniform convergence of the *empirical risk* (training error) to the *expected risk* (generalization error) have been first stated for binary-valued functions [41], then for real-valued functions [42]. Many studies have been devoted to the computation of upper bounds on the sample complexity of multi-class discriminant systems. They can be divided into two groups. In the first one (see for instance [2]), the discrimination task in divided into multiple two-class ones. In the second one, the bounds are based on the definition of specific combinatorial quantities such as the *Natarajan dimension* [28] or the *Graph dimension* [28, 34], which characterize the complexity or *capacity* of the family of functions implemented. These quantities are extensions of the VC dimension [40] to the multi-class case. A synthetic study of the properties of the extensions of the VC dimension to $\{0, \dots, n\}$-valued functions can be found in [7]. In [6], Bartlett states original bounds on the generalization error of binary classifiers. These results lead to many improvements of previous works about discriminant systems by exhibiting a controllable feature of learning systems: the margin. For multi-class discrimination, interesting developments have been done by Schapire *et al.* in [33], where a notion of multi-class margin is introduced. They establish error bounds for boosting methods but their calculations cannot readily be extended to all multi-class systems. By deriving an extension of a lemma by Bartlett, this paper aims at giving tighter error bounds for general multi-class systems built from vector-valued functions, and establishes subsequently a theoretical framework appropriate for the study and development of multi-class Support Vector Machines (M-SVMs).

Let us first set the statistical framework of this study. Let $\mathcal{X}$ be the input or instance space and $\mathcal{C}$ a set of Q categories. $\mathcal{X}$ and $\mathcal{C}$ are supposed to be associated through a fixed but unknown joint probability distribution P (throughout, we ignore issues of measurability, and assume that all sets considered are measurable). Solving the discrimination problem then consists in finding in a given family of classifiers $\mathcal{F}$, *i.e.* a set of functions f from $\mathcal{X}$ into $\mathcal{C}$, the one with minimal *expected risk*. For all f in $\mathcal{F}$, the risk is classically defined as:

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

where $\mathbb{1}_{\{f(x)\neq C\}} = 1$ if $f(x) \neq C$, $\mathbb{1}_{\{f(x)\neq C\}} = 0$ otherwise. Similarly, the *empirical risk* of f, on a subset $s_m = \{(x_i, C(x_i))\}$, $(1 \leq i \leq m)$, of $\mathcal{X} \times \mathcal{C}$ is defined as:

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

This framework is known in the learning theory as the *agnostic learning model* [25]. In this paper, we address specifically the case where the family of discriminant functions considered is generated by a family of real-valued multivariate functions $\mathcal{H}$. Precisely, each function h in $\mathcal{H}$ is associated with the function f satisfying :

$$\forall x \in \mathcal{X},\; f(x) = C_k \Longleftrightarrow h_k(x) = \max_{1\leq l\leq Q} h_l(x)$$

where C_k is the k^{th} category of $\mathcal{C}$. With this definition at hand, an ambiguity can occur if x belongs to a decision boundary. We will bypass this difficulty by considering that in this case, the classifier systematically misclassifies the example. From a formal point of view, this can simply be achieved by introducing a an additional "dummy" category.

Many discriminant models belong to this framework. It is in particular the case when the h_k provide estimates of the class posterior probabilities, which occurs for instance when the model is a neural network and the training criterion is adequately selected [30]. Then, assigning a pattern to the class corresponding to the highest output amounts to implementing Bayes' estimated decision rule. This choice is justified by the fact that this rule is the one corresponding to the lowest error rate. We will make this assumption in what follows, and, when no confusion is possible, simply identify the sets $\mathcal{F}$ and $\mathcal{H}$, writing for instance $R(h)$ in place of $R(f)$. Note that in this context, two-class problems are degenerate ones, since determining the decision boundary in that case only requires the computation of one single real-valued function, which can be seen as the difference between h_1 and h_2. As a consequence, this case will be considered independently here. To distinguish two-class problems from problems with 3 classes and over, a classifier will be called in the first case a *binary classifier* and in the second case a *multiclassifier*. Alternatively, when the context is clear enough, the generic term classifier will also be used.

In order to choose the optimal function $h \in \mathcal{H}$, we make the assumption that a training set $s_m = \{(x_i, y_i)\}$, $(1 \leq i \leq m)$, made up of labelled examples, iid according to the joint distribution on $S = \mathcal{X} \times \mathcal{C}$ to be infered, is available. In what follows, $C(x_i)$ will denote indifferently the category of the pattern x_i, or the index of this category, while y_i will be the corresponding canonical coding, i.e. $C(x_i) = C_l \Longleftrightarrow y_i = [y_{ik}] \in \mathbb{R}^Q$, where $y_{ik} = -1^{1-\delta_{kl}}$, and δ is Kronecker's symbol.

The organization of this paper is as follows. In the following section, a survey of the different definitions of margins introduced in learning theory to study two-class problems highlights the central role played by this notion for generalization capabilities assessment. Section 3 deals with the notion of multi-class margin on which the results of this work are grounded. Section 4 is devoted to the main theorem, a bound on the expected risk of general multi-class discriminant models, where the capacity measure is a covering number. From this theorem, several corollaries can be derived, which are discussed in Section 5. In order for the formula to be of practical use, efficient tools must be available to bound the covering number of interest. This question is tackled in Section 6, where tight bounds on this quantity are derived from functional analysis results. At last, we present in Section 7 an application in predictive structural biology, and show how to use the results to obtain confidence bounds on the empirical error.

2 Margin for two-class discrimination

In this section, we discuss the notion of margin for binary classifiers. Since the early sixties, the notion of margin has been widely used in the learning community to study two-class discrimination. It is for instance involved in the analyse of the convergence of the perceptron algorithm (Novikoff's theorem, see [40] for a proof and a discussion). A first quantity connecting the notion of margin and the rate of convergence of the empirical risk is the fat-shattering dimension [23, 24]. In order to introduce this capacity measure, we must first give preliminary definitions.

Definition 1 (Growth function) *Let $\mathcal{G}$ be a set of indicator functions (functions taking two values) on $\mathcal{X}$. Let $s_m = \{x_1, \ldots, x_m\} \subset \mathcal{X}^m$ and let $\Pi_{\mathcal{G}}(s_m)$ be the number of different* dichotomies *on s_m computed by the functions of $\mathcal{G}$. For all positive integer m, the growth function $\Pi_{\mathcal{G}}$ is defined by:*

$$\Pi_{\mathcal{G}}(m) = \max\left\{\Pi_{\mathcal{G}}(s_m) : s_m \in \mathcal{X}^m\right\}.$$

Note that this definition is the one used in the PAC community (see for instance [4]). For Vapnik, the growth function is the natural logarithm of this expression [40]. This difference has no incidence here.

Definition 2 (VC dimension) *The VC dimension of a set $\mathcal{G}$ of indicator functions is the maximum number d_{VC} of vectors that can be* shattered, *i.e. separated into two classes in all $2^{d_{VC}}$ possible ways, using functions of $\mathcal{G}$. Thus:*

$$d_{VC} = max\left\{m : \Pi_{\mathcal{G}}(m) = 2^m\right\}$$

If this maximum does not exist, the VC dimension is equal to infinity.

The fat-shattering dimension is defined as follows (cf. [23, 35, 3, 6, 36]):

Definition 3 (Fat-shattering dimension) *Let $\mathcal{G}$ be a set of real-valued functions, and γ a positive real. A set of points $s_m = \{x_i\}$, $(1 \leq i \leq m)$ is said to be γ-shattered by $\mathcal{G}$ if there is a vector $b = [b_i] \in \mathbb{R}^m$ such that, for all binary vector $y = [y_i] \in \{-1, 1\}^m$, there is a function $g_y \in \mathcal{G}$ satisfying*

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

The fat-shattering *dimension* $fat_{\mathcal{G}}$ *of the set* $\mathcal{G}$ *is a function from the positive real numbers to the integers which maps a value* γ *to the size of the largest* γ*-shattered set, if it is finite, or to infinity otherwise.*

To some extent, one can consider that the fat-shattering dimension plays in the context of function approximation the role that the VC dimension plays in the context of two-class discriminant analysis. Indeed, in the case of dichotomy computation, the necessary and sufficient condition for non trivial consistency of the Empirical Risk Minimization (ERM) inductive principle is the finiteness of the VC dimension [40]. Similarly, for real-valued functions, this consistency rests on the finiteness of the fat-shattering dimension [3][1]. In the usual case where a two-class discriminant function is obtained by applying a threshold function on the output of a real-valued one, bounding the fat-shattering dimension of the latter makes it possible to study the generalization properties of the former. In [35] and in [6], error bounds for discriminant models are derived in that way. Below is one of them.

Theorem 1 (From Bartlett [6], Corollary 9) *Suppose* $s = ((x_1, y_1), \ldots, (x_m, y_m))$ *is chosen by* m *independent draws from* P*. Then with probability at least* $1 - \delta$*, for all* $\gamma \in]0, 1]$*, for every classification system* $\mathcal{F}$ *of functions* f *taking* x *as input and computing their outputs by signing a function* $h \in \mathcal{H}$*:* $f(x) = sign(h(x))$*, we have*

$$R(h) < R^{\gamma}_{s_m}(h) + \tilde{O}\left(\sqrt{\frac{d_\gamma - \ln(\gamma\delta)}{m}}\right)$$

where $d_\gamma = fat_{\mathcal{H}}(\gamma/32)$ *and*

$$R^{\gamma}_{s_m}(h) = \frac{1}{m}\left|\{(x_i, y_i) \ / \ f(x_i)y_i < \gamma\}\right|$$

(The notation $\tilde{O}$ *means that the logarithmic terms are neglected).*

In his original work, Bartlett calls the quantity $R^{\gamma}_{s_m}(h)$ a *margin error.* The term $\tilde{O}\left(\sqrt{\frac{d_\gamma - \ln(\gamma\delta)}{m}}\right)$ will be referred to as the *control term.* Since the margin error is an increasing function of γ, whereas d_γ, and by way of consequence the control term, are decreasing functions of it, making the best use of Theorem 1 amounts to finding an adequate value for this degree of freedom. In Figure 1, hypothetic values of the margin error and the control term are represented as a function of γ. In practice, the increase of the margin error with γ should be relatively large compared to the decrease of the control term. However, this assertion depends greatly on the geometrical configuration of the points in the learning set.

The term γ is referred to as a margin in [6]. In order to compare it with the other notion of margin, the following definition must be introduced:

Definition 4 (Two-class margin) *A discriminant function* f *associated with a real-valued function* h *is said to have a* two-class margin γ_0 *over a set* s *if*

$$\gamma_0 = \min_{(x_i, y_i) \in s} \{y_i h(x_i) \ / \ y_i h(x_i) > 0\}$$

This expression is related to the notions of margin used in the context of boosting (see for instance [33]), with the difference that it is always non negative. With this definition at hand, it results from Theorem 1 that in order to get a low generalization error, a learning system has to correctly classify a large part of the learning set

[1]Note further that the necessary and sufficient conditions corresponding to the case of families of functions taking their values in a finite discrete set, case in which multi-category discriminant models fall, are exhibited in [7].

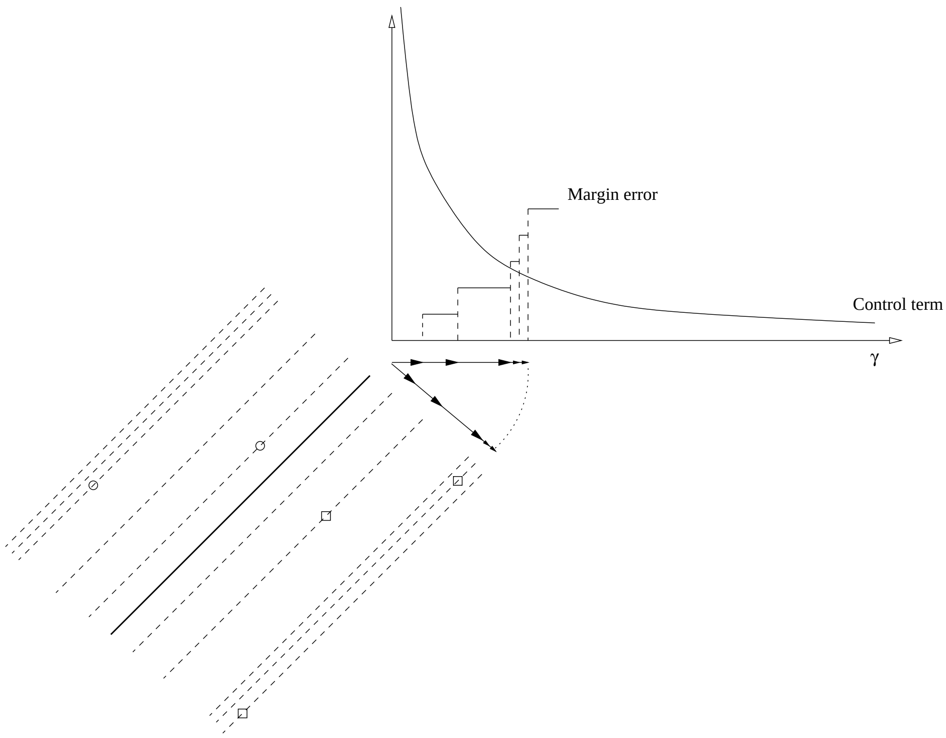

Figure 1 Compromise between a large margin error and a large confidence interval, both of which are controlled by the value of γ. The bottom left part of the figure represents the separation performed by a classifier $f(x) = sign\,(\gamma(\langle w, x\rangle + b))$, where (w, b) defines a canonical hyperplane (see Definition 5). The dashed lines correspond to the values of the margin error when γ varies. The best choice for γ depends on a compromise between the increase of the margin error and the decrease of the control term.

with a large margin. However, there is no point in achieving this goal by scaling the function h, due to the concomitant increase of the fat-shattering dimension. Before Bartlett's work and the introduction of the fat-shattering dimension, Vapnik presented in [39] the maximal margin hyperplane. The idea of maximal margin has then been reconsidered ten years later to give birth to SVMs and all the related algorithms for their non-linear extension [13]. The principle of SVMs rests on the definition of a canonical hyperplane.

Definition 5 (Canonical hyperplane) *Let a hyperplane in a Euclidean space be characterized by the couple* (w, b). *Let* s *be a set of points in this space satisfying:*

$$\min_{x \in s} |\langle w, x\rangle + b| > 0$$

Then the canonical form *of the hyperplane with respect to* s *corresponds to the couple* (w_0, b_0) *(equal to* (w, b) *up to a multiplicative constant), such that*

$$\min_{x \in s} |\langle w_0, x_0\rangle + b| = 1 \tag{3}$$

A hyperplane in canonical form has the property that its distance from s is equal to $\frac{1}{\|w\|}$. The complexity of these hyperplanes can be expressed in terms of their VC dimension as in the following theorem:

Theorem 2 (Vapnik [40]) *Suppose s is a subset of the input space contained in a ball of radius R about some point. Consider the set of hyperplanes in canonical form with respect to s that satisfy $\|w\| \leq A$, and let $\mathcal{F}$ be the class of corresponding linear threshold functions,*

$$f(x) = sign(\langle w, x \rangle + b)$$

Then the restriction of $\mathcal{F}$ to the points of s has VC dimension bounded above by

$$\min\{R^2 A^2, n\} + 1$$

The drawback of this theorem is that it cannot be directly used to compute the well-known *guaranteed risks* involving the VC dimension (e.g. bounds given in [40]). Indeed, this requires to bound the VC dimension associated with the whole input space and not solely the "training" set. This remark can be found in [11] and [35], where the problem has been solved by modifying Vapnik's result. The following theorem, which directly springs from [35] Corollary 5.4, illustrates the way γ can be used to control the generalization performance.

Theorem 3 *Let $\mathcal{H}$ be a set of hyperplanes in a n-dimensional space, with $\|w\| \leq A$. Assume that all the inputs are in a ball of radius R. The fat-shattering dimension of $\mathcal{H}$ can be bounded above by*

$$fat_{\mathcal{H}}(\gamma) \leq \min\{9\frac{R^2 A^2}{\gamma^2}, n+1\} + 1$$

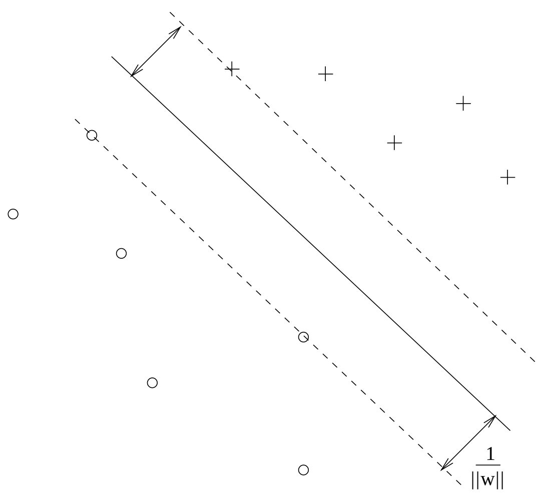

Figure 2 Biclass margin and canonical hyperplane - the geometrical margin is materialized with dashed lines

Since Vapnik's notion of margin characterizes the distance from the training set to the hyperplane (see Figure 2), we will call it in what follows the *geometrical margin*, to distinguish it from the two-class margin (see Definition 4), which depends on the real-valued function from which the binary classifier has been built. For canonical hyperplanes, provided at least one point is correctly classified, the two-class margin γ_0 is always equal to 1 as a consequence of Definition 5, contrary to the geometrical margin. Consequently, the margin error $R_s^\gamma(h)$ is equal to the standard empirical risk as long as the value of γ is inferior to 1. Vapnik considers solely the geometrical margin of hyperplanes in the case of linearly separable learning sets (cf. [40], Chap. 10) and null empirical risks. Let us express these hypotheses in the framework developed so far. By definition, $0 < \gamma \leq \gamma_0 = 1$. Consequently, due to Theorem 3, finding an optimal hyperplane with respect to the bound of Theorem 1 simply amounts to finding a hyperplane of minimal norm subject to the constraints induced by the nullity of the (margin) error, which rises no difficulty, since these constraints no longer involve the parameter γ. Note that once the corresponding value of the norm is known, the minimun of the bound is obtained by setting $\gamma = \gamma_0 = 1$. This training procedure is exactly the support vector method [40], which thus finds here another theoretical basis.

For non linearly separable learning sets, both notions of margin are still defined and the previous considerations hold. However, a compromise is then to be found between training accuracy (low margin error) and confidence (low control term). Simply maximizing the margin, as in the basic support vector method, is no longer appropriate. The introduction of the *soft margin* in the SVM algorithm [40] can be interpreted as a way to implement this trade-off[2]. This notion of margin appears to be intimately connected to the support vector method and is beyond the scope of this report.

To sum up, three definitions of margin can be found in learning theory literature. The first one, referred here as the two-class margin, is related to fat-shattering dimensions and exhibits the drawback to depend on the real-valued function which defines the binary classifier. Hence, a binary classifier has many margins since it can be built from many different real-valued functions. Theorem 1 tells us what classifier to choose to have a low generalization error bound. The second one is the geometrical margin: it corresponds to a geometric feature and is unique for a given classifier. Maximizing it requires the carrying of a convention, which is implemented in the context of SVM by working with hyperplanes in canonical form. This maximization is not always the best objective, as pointed out in the discussion. The third one, the soft margin, has been defined in the support vector method to deal with the aforementioned compromise between margin error and control term.

3 Margin for multi-class discrimination

The notion of multi-class margin we will introduce is an extension of the two-class one. Before putting it more formally, let us look at the way the problem is dealt with by other authors. In [40], Vapnik describes an experiment on digit recognition by means of SVMs. To perform this task, ten linear SVMs are used, each one trained to distinguish one particular digit from all the others. As a consequence, the training procedure amounts to maximizing a set of ten (two-class) geometrical margins. Figure 3 depicts the result of such an approach for a three-class discrimination problem. Three linear SVMs have be trained to differentiate one class from the other two. The figure on the left presents a possible result for the learning set pictured by squares (class 1), circles (class 2) and crosses (class 3). The SVMs maximize their respective geometrical margins.

Figure 3 (Right) shows how the space is labelled accordingly, by applying Bayes' estimated decision rule, that is, by taking the index of the largest output: $f(x) = argmax_i\{\langle w_i, x\rangle + b_i\}$, where (w_i, b_i) are the parameters of the classifier associated with class i. The question now arises to determine to what extent these two-class (geometrical) margins can be used to analyse multi-class systems. Looking at Figure 5, the distance from a point to the decision boundary appears not to be directly related to the notion of two-class geometrical margin. This is especially

[2]The soft margin also presents an interest for linearly separable cases, especially when the input space has an infinite dimension.

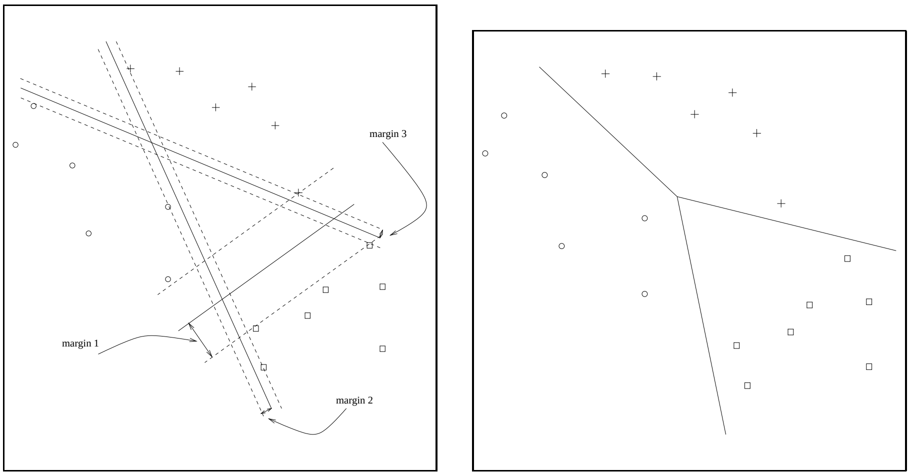

Figure 3 Margin for multi-class discriminant systems. **Left:** three-class problem with the linear separations between each class and the two others. **Right:** corresponding separation for the three-class problem. Each hyperplan is associated with a class (circle, square or cross). The separation is computed by taking the class of the hyperplan the "output" of which is the largest.

clear for the circles. Actually, the smallest distance between any of the learning elements and the decision boundary is larger than all the margins shown in Figure 3. Thus, the two-class geometrical margins do not represent here a clear geometrical feature of the overall separation between the three classes. Other techniques in pattern recognition which aim at designing large margin classifiers are boosting methods [10, 14]. Although theory in the field has mainly been developed for dichotomy computation, results have also been derived for multi-class problems. In [33], a definition of a multi-class margin is given, which applies to couples (x, y). For this definition, an error bound is derived (Theorem 6). Although elegant, the result cannot be extended simply to to other learning systems. In what follows, we introduce a multi-class margin which appears as an extension of the one of Schapire *et al.*. This requires additional definitions to be preliminary given.

Definition 6 (Canonical function) *For a given function h in $\mathcal{H}$, 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 $h_l(x) = \max_{k \neq M_1(x)} h_k(x)$. Define $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, as the function from $\mathcal{X}$ into $\mathbb{R}^Q$, satisfying*

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

Let θ denotes the step-function

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

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

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

With these definitions at hand, x is correctly classified (without ambiguity) if and only if the vector $\left[2\hat{\theta}\left(\Delta h_k(x)\right) - 1\right]$ is equal to the canonical coding of its category.

By denoting $z = (x, y) \in S$, we have

$$R(h) = \int_S \hat{\theta} \left[Q - \sum_{k=1}^{Q} \hat{\theta} \left(\Delta h_k(x) y_k \right) \right] dP(z)$$

which simplifies in

$$R(h) = \int_S \left[1 - \hat{\theta} \left(\Delta h_{M_1(h,x)}(x) y_{M_1(h,x)} \right) \right] dP(z)$$

Remark that in this expression, an ambiguity, corresponding to $\Delta h_{M_1(h,x)}(x) = 0$, leads to an error irrespective of the value of $y_{M_1(h,x)}$. The notion of multi-class margin can then be defined from function Δh:

Definition 7 (Multi-class margin) *A mutli-class discriminant function f, built from a vector-valued function $h = [h_k]$, $(1 \leq k \leq Q)$, is said to have a multi-class margin γ_0 over a set s, if:*

$$\gamma_0 = \min_{(x,y) \in s} \left\{ \Delta h_{M_1(h,x)}(x) \ / \ \Delta h_{M_1(h,x)}(x) y_{M_1(h,x)} > 0 \right\}$$

Figure 4 illustrates the way the multi-class margin γ_0 is computed for a three-class problem.

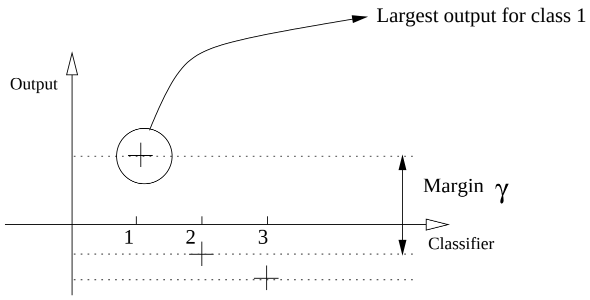

Figure 4 The value of the multi-class margin γ is obtained by taking the largest output and by considering the difference between this output and the others.

This definition generalizes the two-class one. In order to prove this, suffice it to notice that the real-valued function h_0 associated with a two-class discriminant function can always be written has $h_0 = \frac{1}{2}(h_1 - h_2)$, with h_1 and h_2 being associated with the two categories. Similarly, the *geometrical* multi-class margin is defined as the smallest distance between an element of the learning set and the separation between classes. Scaling all the h_k by the same value K will change neither the decision boundary nor the geometrical margin, whereas the multi-class margin γ_0 will be scaled by K. These definitions are illustrated in Figures 3 and 5. Let us consider as previously a three-class problem. These classes are still represented by circles, crosses and squares. The set $\mathcal{H}$ is assumed to be linear applications from $\mathcal{X}$ into $\mathbb{R}$. The left picture in 3 shows three separations (solid line) with their corresponding two-class geometrical margin (interspace between dashed lines). These margins are calculated for each function h_1, h_2 and h_3 separately. The right Figure 3 shows the separation between the different classes. Figure 5 represents the same separation with the margins 1, 2 and 3. The smallest of these margins is the multi-class geometrical margin. We see in this figure that multi-class margin and two-class margins are different notions. The latter does not take into account the

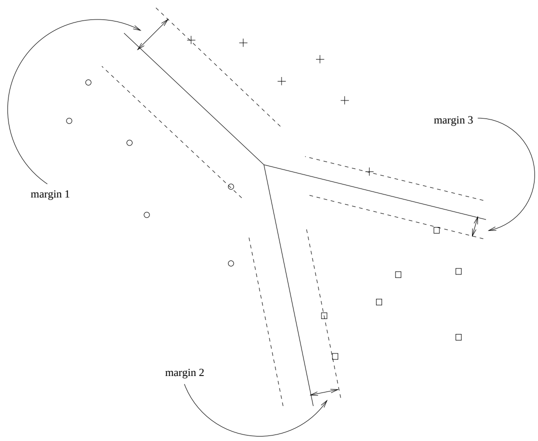

Figure 5 Geometrical margins of a multi-class linear discriminant model. The multi-class margin is the minimum of the three interspaces 1, 2 and 3.

way the multiclassifier computes its output. Furthermore, it is possible to have a small two-class margin and a (relatively) large multi-class margin. The two-class margin has been shown to be of importance to analyse the generalization error of a binary classifier. The fraction of elements in the learning set which are closer to the separation between class 1 and class -1 than a distance γ is indeed an approximation of the generalization error. The larger γ is, the more accurate the approximation is. For multi-class discrimination, the fraction of elements in the learning set which are closer to the separation between the different classes than a distance γ also plays an important role. This fraction will be called the *empirical margin risk* and is formally defined as

Definition 8 (Empirical margin risks) *The empirical risk with margin* $\gamma \in (0,1]$ *on a training set* s_m *of size* m *is*

$$
\begin{aligned}
R^{\gamma}_{s_m}(h) &= \frac{1}{m}\left|\left\{(x_i,y_i) \in s_m / \sum_{k=1}^{Q} \theta\left(\Delta h_k(x_i)y_{ik} - \gamma\right) - Q < 0\right\}\right| \\
&= \frac{1}{m}\left|\left\{(x_i,y_i) \in s_m \;/\; \Delta h_{C(x_i)}(x_i) < \gamma\right\}\right|
\end{aligned}
$$

$R^{\gamma}_{s_m}(h)$ can be used for two-class discrimination as well as multi-class discrimination. For $0 < \gamma \leq \gamma_1$ where γ_1 is the geometrical margin, it corresponds to the empirical classification error. In the case of a two-class problem, this risk is the same as the margin error defined in [6] which has been used to bound the generalization error.

4 Bounds for the generalization error

In this section, building upon this definition of the empirical risk, we extend a lemma of Bartlett to derive an upper bound of the generalization error of multi-class discriminant models. For binary classification (dichotomies), this error is usually bounded as follows:

> Let $\mathcal{F}$ be a set of $\{0,1\}$-valued functions f defined on an input space $\mathcal{X}$. Let P be any probability measure on $S = \mathcal{X} \times \{0,1\}$. Then, with

probability $1-\delta$, for all $s_m \in S^m$ and for all $f \in \mathcal{F}$

$$R(f) \leq R_{s_m}(f) + O\left(\sqrt{\frac{\ln(\Pi_{\mathcal{F}}(2m)) - \ln \delta}{m}}\right)$$

where $\Pi_{\mathcal{F}}$ is the growth function defined in Section 2. In practice, an upper bound on this function is often obtained from an upper bound on the VC dimension of $\mathcal{F}$ and the Sauer-Shelah lemma [32]. In regression, the same kind of bound holds but with covering numbers instead of the growth function. To get a better understanding of the similarity between both theories (discrimination and regression), it is worthwhile highlighting the link between the growth function and covering numbers.

Definition 9 (Pseudo-metric space) *A pseudo-metric on a set E is a function ρ from $E \times E$ into $\mathbb{R}_+$ such that*

$$\forall (x_1, x_2, x_3) \in E^3 \begin{cases} x_1 = x_2 \Longrightarrow \rho(x_1, x_2) = 0 \\ \rho(x_1, x_2) = \rho(x_2, x_1) \\ \rho(x_1, x_3) \leq \rho(x_1, x_2) + \rho(x_2, x_3) \end{cases}$$

A space endowed with a pseudo-metric is called a *pseudo-metric space.* Note that a metric is simply a pseudo-metric with the additional *separability* condition: $\rho(x_1, x_2) = 0 \Longrightarrow x_1 = x_2$.

Definition 10 (Covering number) *Let (E, ρ) be a pseudo-metric space, and $B(c, r)$ the closed ball of radius r and centre c in E. The* covering number $\mathcal{N}(\epsilon, H, \rho)$ *of a set $H \subset E$ is the smallest cardinality of the sets $\bar{H} \subset E$ such that*

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

The sets $\bar{H}$ satisfying this property are called ϵ-covers of H: each element in H is at a distance less than ϵ of an element in $\bar{H}$.

See [26, 12] for the fundamental results regarding covering numbers. In general, for a given ϵ, a cover of minimal cardinality may not be included in the set H. However, from any ϵ-cover of H, one can readily construct a 2ϵ-cover of H of same cardinality and included in H. If the set H is discrete and the distance between two elements in H is larger than 2ϵ, then, $\mathcal{N}(\epsilon, H, \rho) = |H|$ (keeping in mind that H is its own ϵ-cover for any positive value of ϵ). Turning to the case of spaces of binary functions (i.e. with our notations $H = \mathcal{F}$), by considering the appropriate pseudo-metric space[3], the value of $\Pi_{\mathcal{G}}(s_m)$ which appears in the definition of a growth function can be seen as a covering number. Indeed, consider the following pseudo-metric:

Definition 11 *Let $\mathcal{F}$ be a set of $\{-1, 1\}$-valued functions over $\mathcal{X}$ and $(f, \bar{f}) \in \mathcal{F}^2$. For $s \subset \mathcal{X}$, define the pseudo-metric $d_{l_\infty(s)}$ by*

$$d_{l_\infty(s)}(f, \bar{f}) = \max_{x \in s} \left|f(x) - \bar{f}(x)\right|$$

If the set $\mathcal{F}$ is endowed with the pseudo-metric $d_{l_\infty(s)}$, the pseudo-distance between two functions is either zero (the functions coincide on the learning set s) or equal to 2 (there is at least one difference). Consequently

$$\Pi_{\mathcal{F}}(s) = \mathcal{N}(1, \mathcal{F}, d_{l_\infty(s)})$$

[3]For now on, for the sake of simplicity, we will not specifiy which of the pseudo-metrics we use are actually metrics.

and the growth function can be expressed as

$$\Pi_{\mathcal{F}}(2m) = \mathcal{N}_\infty(1, \mathcal{F}, 2m)$$

where $\mathcal{N}_\infty(1, \mathcal{F}, 2m) = \sup_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{N}(1, \mathcal{F}, d_{l_\infty(s_{2m})})$. Actually, covering numbers play an important role in classification, which lies beyond their connection with the growth function. From Bartlett's results [6], the following lemma can be direclty derived:

Lemma 1 (From Bartlett [6], Corollary 9) *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}$. Let $0 < \gamma \leq 1$ and $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ be the piecewise-linear squashing function defined as*

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

Let $\pi_\gamma(\mathcal{H}) = \{\pi_\gamma \circ h \ / \ h \in \mathcal{H}\}$. Let s_m be a learning set chosen by m independent draws from a probability P. With probability $1 - \delta$, $\forall \gamma \in]0, 1]$, for all h in $\mathcal{H}$

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

This lemma is the key point in the proof of Theorem 1. Remark, as in this theorem, that scaling the function h by an appropriate real can decrease the upper bound on the generalization error (guaranteed risk). For multi-class problems, we have discussed the fact that using the two-class margin directly was not utterly satisfactory. For the same reason, Lemma 1 cannot be used really efficiently to bound the generalization error of multi-class models. This has been highlighted in [17], on a real-world application. One would expect that a multi-class version of this lemma would give better results for these models. This extension is the subject of the following section.

4.1 Extension to multi-class discrimination

This section establishes a way to extend Lemma 1 for multi-class discrimination. To that end, a new pseudo-metric is required.

Definition 12 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, and $(h, \bar{h}) \in \mathcal{H}^2$, define*

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

Let $\mathcal{N}_{\infty,1}(\gamma/2, \mathcal{H}, 2m) = \sup_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{N}(\gamma/2, \mathcal{H}, d_{l_\infty, l_1(s_{2m})})$. With these definitions at hand, the following result holds.

Theorem 4 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}$. Let $\Delta\mathcal{H}$ be the set of functions derived from $\mathcal{H}$ according to Definition 6. Let s_m be a learning set obtained by m independent draws from a probability P. With probability $1 - \delta$, $\forall h \in \mathcal{H}$*

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

where $\Delta\mathcal{H}^\gamma = \{\pi_\gamma(\Delta h) \ / \ h \in \mathcal{H}\}$.

Proof

The sketch of the proof follows classical proofs in statistical learning theory (see [29] for an example). We will split it into three steps : first symmetrization, second symmetrization and exponential bound. Before beginning the calculations, let us present some quantities which will be of interest in our reasoning. The empirical margin risk has been introduced in the previous section. The *margin risk* is the corresponding extension to the whole product space $S = \mathcal{X} \times \mathcal{C}$.

Definition 13

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

The following risks will also prove useful:

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

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

For the sake of simplicity, we will write from now on Δh_k^{γ} in place of $\pi_{\gamma}\left(\Delta h_k\right)$ and define Δh^{γ} accordingly. 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

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

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 [40] (Section 4.5), we prove the following result:

Lemma 2

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

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 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

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

$$\int_{\mathcal{Q}} dP(s_m) \underbrace{\int_{S^m} \theta \left[\sup_{h \in \mathcal{H}} \left(\mathcal{R}^{\gamma}_{\tilde{s}_m}(h) - R^{\gamma}_{s_m}(h) \right) - \epsilon + \frac{1}{m} \right] dP(\tilde{s}_m)}_{I} \tag{4}$$

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

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

From

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

a sufficient condition for

$$\mathcal{R}^{\gamma}_{\tilde{s}_m}(h^*) - R^{\gamma}_{s_m}(h^*) \geq \epsilon - \frac{1}{m}$$

is deduced:

$$\mathcal{R}^{\gamma}_{\tilde{s}_m}(h^*) \geq \mathcal{R}^{\gamma}(h^*) - \frac{1}{m}$$

Since $\mathcal{R}^{\gamma}_{\tilde{s}_m}(h^*) \geq \mathcal{R}^{\gamma}(h^*) - \frac{1}{m} \Longrightarrow \mathcal{R}^{\gamma}_{\tilde{s}_m}(h^*) - R^{\gamma}_{s_m}(h^*) \geq \epsilon - \frac{1}{m}$, it comes

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

Furthermore

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

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

The right-hand side is always greater than $1-\frac{1}{e}$, and thus also greater than $\frac{1}{2}$. Substituting this lower bound of I into (4) yields

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

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

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

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

$$P_{s_m}\left(\sup_{h\in\mathcal{H}}\left(R(h) - R^\gamma_{s_m}(h)\right) \geq \epsilon\right) \leq 2P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) \tag{5}$$

springs then directly from the fact that $R(h) \leq \mathcal{R}^\gamma(h)$.

Second symmetrization

Let us now consider the set of all permutations σ over S^{2m} such that for all $s_{2m} = (s_m, \tilde{s}_m)$ in S^{2m}, σ realizes a transposition between two elements of same ranking in s_m and $\tilde{s}_m$. Let $\mathcal{U}$ be the uniform distribution over the set of all such permutations σ. For every sample $t \in S^m$, t^σ 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 (5) is not affected by a transposition σ. Consequently

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

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

The set $(s_m, \tilde{s}_m)$ being fixed, $\Delta\bar{\mathcal{H}}^\gamma$ will refer to a $\frac{\gamma}{2}$-cover of the set $\Delta\mathcal{H}^\gamma = \{\pi_\gamma(\Delta h) \,/\, h \in \mathcal{H}\}$ with respect to the pseudo-metric $d_{l_\infty,l_1(s_{2m})}$. By definition

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

$$\max\left(\max_{(x,y)\in s_m}\sum_{k=1}^{Q}|\Delta h^\gamma_k(x) - \Delta\bar{h}^\gamma_k(x)|,\ \max_{(\tilde{x},\tilde{y})\in\tilde{s}_m}\sum_{k=1}^{Q}|\Delta h^\gamma_k(\tilde{x}) - \Delta\bar{h}^\gamma_k(\tilde{x})|\right) \leq \frac{\gamma}{2}$$

For a given $\gamma > 0$, define

$$d_{l_1(x,y)}(\Delta h, \gamma y) = \sum_{k=1}^{Q}|\Delta h_k(x) - \gamma y_k|$$

For all $\left(\Delta h^\gamma, \Delta\bar{h}^\gamma\right)$, the triangular inequality gives

$$\left\{\begin{array}{l} d_{l_1(\tilde{x}^\sigma,\tilde{y}^\sigma)}(\Delta h^\gamma, \gamma\tilde{y}^\sigma) \geq \gamma \\ d_{l_\infty,l_1(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2} \end{array}\right. \Longrightarrow d_{l_1(\tilde{x}^\sigma,\tilde{y}^\sigma)}(\Delta\bar{h}^\gamma, \gamma\tilde{y}^\sigma) \geq \frac{\gamma}{2}$$

which is equivalent to

$$\left\{\begin{array}{l} \sum_{k=1}^{Q}\left|\Delta h_k^{\gamma}(\tilde{x}^{\sigma})-\gamma\tilde{y}_k^{\sigma}\right|\geq\gamma \\ \\ d_{l_\infty,l_1(s_{2m})}(\Delta h^{\gamma},\Delta\bar{h}^{\gamma})\leq\frac{\gamma}{2}\end{array}\right. \Longrightarrow \sum_{k=1}^{k=Q}\left|\Delta\bar{h}_k^{\gamma}(\tilde{x}^{\sigma})-\gamma\tilde{y}_k^{\sigma}\right|\geq\frac{\gamma}{2}$$

Similarly

$$\left\{\begin{array}{l} d_{l_1(x^{\sigma},y^{\sigma})}(\Delta h^{\gamma},\gamma y^{\sigma})=0 \\ \\ d_{l_\infty,l_1(s_{2m})}(\Delta h^{\gamma},\Delta\bar{h}^{\gamma})\leq\frac{\gamma}{2}\end{array}\right. \Longrightarrow d_{l_1(x^{\sigma},y^{\sigma})}(\Delta\bar{h}^{\gamma},\gamma y^{\sigma})\leq\frac{\gamma}{2}$$

From this, by considering the converse, we deduce that

$$\left\{\begin{array}{l} d_{l_1(x^{\sigma},y^{\sigma})}(\Delta\bar{h}^{\gamma},\gamma y^{\sigma})\geq\frac{\gamma}{2} \\ \\ d_{l_\infty,l_1(s_{2m})}(\Delta h^{\gamma},\Delta\bar{h}^{\gamma})\leq\frac{\gamma}{2}\end{array}\right. \Longrightarrow d_{l_1(x^{\sigma},y^{\sigma})}(\Delta h^{\gamma},\gamma y^{\sigma})\neq 0$$

which is the same as

$$\left\{\begin{array}{l} \sum_{k=1}^{Q}\left|\Delta\bar{h}_k^{\gamma}(x^{\sigma})-\gamma y_k^{\sigma}\right|\geq\frac{\gamma}{2} \\ \\ d_{l_\infty,l_1(s_{2m})}(\Delta h^{\gamma},\Delta\bar{h}^{\gamma})\leq\frac{\gamma}{2}\end{array}\right. \Longrightarrow \exists l\in\{1,..,Q\}\ /\ \Delta h_l^{\gamma}(x^{\sigma})\neq\gamma y_l^{\sigma}$$

Consequently, defining $A\left(\Delta\bar{h}^{\gamma},s_m\right) := \frac{1}{m}\left|\left\{i:\sum_{k=1}^{Q}\left|\Delta\bar{h}_k^{\gamma}(\tilde{x}_i^{\sigma})-\gamma\tilde{y}_{ik}^{\sigma}\right|\geq\frac{\gamma}{2}\right\}\right|$, we have:

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

$$\sup_{(s_m,\tilde{s}_m)\in S^{2m}}\mathcal{U}\left\{\sigma:\sup_{\Delta\bar{h}^{\gamma}\in\Delta\bar{H}^{\gamma}}A\left(\Delta\bar{h}^{\gamma},s_m\right)-A\left(\Delta\bar{h}^{\gamma},\tilde{s}_m\right)\geq\epsilon-\frac{1}{m}\right\}$$

$$\leq\left|\Delta\bar{\mathcal{H}}^{\gamma}\right|\sup_{\Delta\bar{h}^{\gamma}\in\Delta\bar{H}^{\gamma}}\mathcal{U}\left\{\sigma:A\left(\Delta\bar{h}^{\gamma},s_m^{\sigma}\right)-A\left(\Delta\bar{h}^{\gamma},\tilde{s}_m^{\sigma}\right)\geq\epsilon-\frac{1}{m}\right\}$$

$$\leq\mathcal{N}_{\infty,1}(\gamma/2,\Delta\mathcal{H}^{\gamma},2m)\sup_{(a_i,b_i)}P\left(\frac{1}{m}\sum_i(a_i-b_i)\beta_i\geq\epsilon-\frac{1}{m}\right) \tag{6}$$

where a_i is equal to 1 if $\sum_{k=1}^{Q}\left|\Delta\bar{h}_k^{\gamma}(\tilde{x}_i^{\sigma})-\gamma\tilde{y}_{ik}^{\sigma}\right|\geq\frac{\gamma}{2}$ and is equal to 0 otherwise. Similarly, b_i is equal to 1 f $\sum_{k=1}^{Q}\left|\Delta\bar{h}_k^{\gamma}(x_i^{\sigma})-\gamma y_{ik}^{\sigma}\right|\geq\frac{\gamma}{2}$ and is equal to 0 otherwise. The probability P is a uniform distribution over the $\beta_i\in\{-1,1\}$ which are iid.

Exponential bounds

The right-hand side of (6) is bounded by Hoeffding's inequality (see for example [22, 29]).

Theorem 5 (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 (6) leads to

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

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

and finally,

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

∎

A related result has been derived by Schapire *et al.* [33] in the context of boosting. They have established error bounds for multi-class systems built from convex combinations of functions. Their result, despite its originality and its elegancy, cannot be applied in a general framework. For instance, the study of multi-class linear systems, such as multi-class linear support vector machines, cannot be understood with their results. The error bounds exposed here can be interpreted as a more general result. The final generalization error bound depends on the value of the covering number appearing in the lemma. Thus, to compare the latter to Schapire *et al.*'s results, we have to bound the covering number $\mathcal{N}_{\infty,1}(\gamma/2,\Delta\mathcal{H}^{\gamma},2m)$. In Section 6, we show how to derive bounds for these quantities but do not analyse the case of boosting algorithms. This study is currently in progress.

5 Derived Results

In analogy with classical results (see for instance [40], Chapter 4, or [4]), we can also obtain bounds for the relative deviation between the empirical margin risk and the generalization error,

Lemma 3 *Suppose* $\gamma>0$,

$$P_{s_m}\left(\exists h\in\mathcal{H}\ \textit{satisfying}\ \frac{R(h)-R^{\gamma}_{s}(h)}{\sqrt{R(h)}}>\epsilon\right)\leq 4\mathcal{N}_{\infty,1}(\gamma/2,\Delta\mathcal{H}^{\gamma},2m)\exp(-m\epsilon^2/2) \tag{7}$$

where P_{s_m} *is a probability over* $s_m\in S^m$.

Proof

Since the proof follows the same sketch as the proof of Theorem 4.2 in [40], we will only give its outlines here.

First, remark that the function $x\mapsto\frac{x-y}{\sqrt{x}}$ is increasing for fixed y. Since $R(h)\leq\mathcal{R}^{\gamma}(h)$, we have,

$$\frac{R(h)-R^{\gamma}(h)}{\sqrt{R(h)}}\leq\frac{\mathcal{R}^{\gamma}(h)-R^{\gamma}(h)}{\sqrt{\mathcal{R}^{\gamma}(h)}}$$

and

$$P_{s_m}\left(\exists h\in\mathcal{H}\ \text{satisfying}\ \frac{R(h)-R^{\gamma}(h)}{\sqrt{R(h)}}\geq\epsilon\right)\leq$$

$$P_{s_m}\left(\exists h\in\mathcal{H}\ \text{satisfying}\ \frac{\mathcal{R}^{\gamma}(h)-R^{\gamma}(h)}{\sqrt{\mathcal{R}^{\gamma}(h)}}\geq\epsilon\right)$$

We will now focus on the right-hand side of this inequality. Following the same reasoning as in the proof of Lemma 4.1 Section 4.7 in [40], we have:

$$P_{s_m}\left(\exists h \in \mathcal{H} \text{ satisfying } \frac{\mathcal{R}^\gamma(h) - R^\gamma(h)}{\sqrt{\mathcal{R}^\gamma(h)}} \geq \epsilon\right) \leq$$

$$4P_{s_m,\tilde{s}_m}\left(\exists h \in \mathcal{H} \text{ satisfying } \frac{R^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h)}{\sqrt{R^\gamma_{s_m,\tilde{s}_m}(h) + \frac{1}{2m}}} \geq \epsilon\right)$$

where $R^\gamma_{s_m,\tilde{s}_m}(h) = \frac{1}{2}(R^\gamma_{s_m}(h) + R^\gamma_{\tilde{s}_m}(h))$. By introducing the permutation σ as in the proof of Theorem 4 and by fixing the learning sets s_m and $\tilde{s}_m$ and defining the $\gamma/2$-cover, we have,

$$\begin{aligned} & P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\frac{\mathcal{R}^\gamma(h)-R^\gamma(h)}{\sqrt{\mathcal{R}^\gamma(h)}} \geq \epsilon\right)\right) \leq \\ & 4\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \sup_{(a_i,b_i)} P\left(\frac{1}{m}\frac{\sum_{i=1}^m (a_i - b_i)\beta_i}{\sqrt{\frac{1}{2m}\sum_{i=1}^m (a_i+b_i) + \frac{1}{2m}}} \geq \epsilon - \frac{1}{m}\right) \end{aligned} \quad (8)$$

where the terms a_i, b_i, β_i, as well as P are defined as in (6). To bound (8), we use the Bernstein's inequality (see for instance [29]), as it is done in [20], Lemma 9 Section 9.3.

Theorem 6 (Berstein's inequality) *Let $\{X_1, X_2, \ldots, X_n\}$ be n independent random variables with zero means and bounded ranges: $|X_i| \leq M$. Write σ_i^2 for the variance of X_i. Suppose $V_n \geq \sum_{i=1}^n \sigma_i^2$. Then, for all $\eta > 0$,*

$$P\left(|\sum_{i=1}^n X_i| \geq \eta\right) \leq 2\exp\left(\frac{-\eta^2}{2\left(V_n + \frac{1}{3}M\eta\right)}\right)$$

The desired result follows then directly. ∎

From this lemma, we can deduce the following corollary,

Corollary 1 *For all $\alpha > 0$,*

(*i*) $P\left[(R(h) > \epsilon) \wedge (R^\gamma_s(h) = 0)\right] \leq 4\mathcal{N}_{\infty,Q}(\gamma/2, \Delta\mathcal{H}^\gamma)\exp(-m\epsilon/2)$

(*ii*) $P\left[R(h) > (1+\alpha)R^\gamma_s(h) + \epsilon\right] \leq 4\mathcal{N}_{\infty,Q}(\gamma/2, \Delta\mathcal{H}^\gamma)\exp(-m\epsilon\frac{\alpha^2}{2(1+\alpha)^2})$

The proof of this corollary is omitted since it mimics the proof of Corollary 7 in [6]. Lemma 4 and 3 are derived for a fixed value of γ. To apply our result, it is thus necessary to fix the margin and then to look at the empirical margin risk. As Bartlett says in [6], it seems unnatural since γ will be observed after the examples are seen. The solution to this problem springs from the following proposition proved in [6] which allows us to give results for all values of the margin γ between 0 and 1.

Proposition 1 (Bartlett, prop. 8 in [6]) *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 $0 < a \leq 1$ and $0 < \delta \leq 1$, $\bigcup_{\alpha\in(0,1]} E(a\alpha, \alpha, \delta\alpha(1-a))$ is measurable; and*

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

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

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

From this, we deduce

Corollary 2 *With probability at least* $1-\delta$ *over* s_m, *for every* $h \in \mathcal{H}$ *such that* $R^{\gamma}_{s_m}(h) = 0$ *and every* $\gamma \in (0,1]$, *we have,*

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

With probability at least $1-\delta$ *over* s_m, *for every* $h \in \mathcal{H}$ *and every* $\gamma \in (0,1]$, *we have,*

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

Proof

The proof of this corollary is exactly the same as in [6]. Just consider the sets:

- ⋄ $E(\gamma, \gamma, \delta)$ which will be the sets of s_m such that, $\exists h \in \mathcal{H}:\ R^{\gamma}_{s_m}(h) = 0$ and

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

  for the first part of the corollary, or consider the sets:

- ⋄ $E(\gamma, \gamma, \delta)$ which will be the sets of s_m such that, $\exists h \in \mathcal{H}$ and

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

  for the second part.

Apply then Proposition 1 with $a = 1/2$ and $\alpha = \gamma$. ∎

To make our result practical, it is necessary to bound covering numbers.

6 How to bound covering numbers

Bounds on covering numbers can be obtained by different methods, for instance combinatorial methods using the fat-shattering dimension [3]. Furthermore, technical tricks can be used for specific learning systems. A great many of them can be found in [20] or in [38]. In this section, we adopt a functional analysis approach.

6.1 Functional analysis approach

The functional analysis approach implemented here is largely inspired by the work of Williamson *et al.* [44, 37]. It can be decomposed into four steps.

1 - Evaluation function

Let $\mathcal{G}$ be a subset of a Banach space[4] E containing all the functions $h_k - h_l$, for $k, l \in \{1, .., Q\}^2$, $h_k \in \mathcal{F}$. For $s_m = \{x_1, \dots, x_{2m}\} \in \mathcal{X}^{2m}$, we define the linear operator:

$$\begin{array}{rccc} T: & \mathcal{G}^{Q(Q-1)/2} \subset E^{Q(Q-1)/2} & \longrightarrow & M_{2m,Q(Q-1)/2}(\mathbb{R}) \\ & (h_k - h_l)_{1 \le k < l \le Q} & \mapsto & T(h_1, \dots, h_Q) \end{array}$$

where,

$$T(h_1, \dots, h_Q) = \overset{\xleftrightarrow{\quad Q(Q-1)/2 \quad}}{\begin{bmatrix} (h_k(x_1) - h_l(x_1))_{1 \le k < l \le Q} \\ \vdots \\ (h_k(x_{2m}) - h_l(x_{2m}))_{1 \le k < l \le Q} \end{bmatrix}}$$

The application T is an evaluation operator which captures information about the function Δh. Given the definition of Δh, except for the sign, and a coefficient $\frac{1}{2}$, for each index i, the Q components of $\Delta h(x_i)$ can be found among the $Q(Q-1)/2$ components of the i^{th} row of $T(h_1, \dots, h_Q)$. Note that, in order to obtain this property, it is indeed necessary to consider the $Q(Q-1)/2$ possible differences of the form $h_k - h_l$, for $k < l$, since the values $M_1(h, x)$ and $M_2(h, x)$ are unknown *a priori*: they depend on the functions h_k and x. Note further that the value of $M_2(h, x)$ plays a minor role here, since $\Delta h_{M_1(h,x)}(x) = -\Delta h_{M_2(h,x)}(x)$.

2 - Cover of the image of T

Let us assume that $\mathcal{G}_Q = \mathcal{G}^{Q(Q-1)/2}$ is included in a ball B_Λ of radius Λ in $E^{Q(Q-1)/2}$. We postpone the discussion of this hypothesis later. Let $\|.\|_{\ell_\infty,1}$ be the norm on $M_{n,p}(\mathbb{R})$ defined by:

$$\forall A = [a_{ij}] \in M_{n,p}(\mathbb{R}), \quad \|A\|_{\ell_\infty,1} = \max_{1 \le i \le n} \sum_{j=1}^{p} |a_{ij}|$$

This norm has the following property:

Proposition 2 *For Δg^γ and Δh^γ in $\Delta \mathcal{H}^\gamma$,*

$$d_{l_\infty, l_1(s_{2m})}(\Delta g^\gamma, \Delta h^\gamma) \le \|T(g_1, \dots, g_Q) - T(h_1, \dots, h_Q)\|_{\ell_\infty,1} \tag{9}$$

Proof

By Definition 12, $d_{l_\infty, l_1(s_{2m})}(\Delta g^\gamma, \Delta h^\gamma) = \max_{x_i \in s_{2m}} \sum_{k=1}^{Q} |\Delta g_k^\gamma(x_i) - \Delta h_k^\gamma(x_i)|$. From an exhaustive study of the different cases:

- ⋄ $(k = M_1(g, x_i) \wedge k = M_1(h, x_i))$;
- ⋄ $(k = M_1(g, x_i) \wedge k \neq M_1(h, x_i))$;
- ⋄ $(k \neq M_1(g, x_i) \wedge k \neq M_1(h, x_i))$.

[4] Complete normed linear space.

using the fact that π_γ is 1-lipschitzian, it appears that we have:

$$\forall (i,k) \in \{1,\dots,2m\} \times \{1,\dots,Q\}, \ \ \exists l(i,k) \in \{1,\dots,Q\} \setminus \{k\} \ \ /$$

$$|\Delta g_k^\gamma(x_i) - \Delta h_k^\gamma(x_i)| \leq \frac{1}{2} \left| g_{l(i,k)}(x_i) - g_k(x_i) - h_{l(i,k)}(x_i) + h_k(x_i) \right|$$

For each couple (i,k), let $l(i,k)$ be one of the indexes satisfying the aforementioned property. By summation over k, we get:

$$\forall i \in \{1,\dots,2m\},$$

$$\sum_{k=1}^{Q} |\Delta g_k^\gamma(x_i) - \Delta h_k^\gamma(x_i)| \leq \frac{1}{2} \sum_{k=1}^{Q} \left| g_{l(i,k)}(x_i) - g_k(x_i) - h_{l(i,k)}(x_i) + h_k(x_i) \right| \quad (10)$$

Considering now the right-hand side of (10), it appears that, except for the sign, all the terms in the sum belong to the i^{th} row of $T(g_1,\dots,g_Q) - T(h_1,\dots,h_Q)$. Furthermore, each term of this row appears at most 2 times in the sum. Indeed, there is only one configuration where two terms in the right-hand side of (10) can be identical with respect to the indexes involved. It corresponds to the existence of a couple (k_1,k_2) such that $k_2 = l(i,k_1)$ and $k_1 = l(i,k_2)$. One example of this situation one can think of, if $M_1(g,x_i) = M_1(h,x_i)$ and $M_2(g,x_i) = M_2(h,x_i)$, is the following : $k_1 = M_1(g,x_i)$, $k_2 = M_2(g,x_i)$. Let $T_{i,j}(g,h)$ be the general term of $T(g_1,\dots,g_Q) - T(h_1,\dots,h_Q)$. We have thus:

$$\forall i \in \{1,\dots,2m\},$$

$$\sum_{k=1}^{Q} |\Delta g_k^\gamma(x_i) - \Delta h_k^\gamma(x_i)| \leq \frac{1}{2}.2 \sum_{j=1}^{Q(Q-1)/2} |T_{ij}(g,h)| \quad (11)$$

Taking the maximum over the x_i, we get finally

$$\max_{x_i \in s_{2m}} \sum_{k=1}^{Q} |\Delta g_k^\gamma(x_i) - \Delta h_k^\gamma(x_i)| \leq \max_{x_i \in s_{2m}} \sum_{j=1}^{Q(Q-1)/2} |T_{ij}(g,h)| \quad (12)$$

By definition of $d_{l_\infty, l_1(s_{2m})}$ and $\|.\|_{\ell_{\infty,1}}$, this expression is equivalent to (9), which concludes the proof. ∎

We consider N_γ, a γ cover of $T(\mathcal{G}_Q)$ for the norm $\|.\|_{\ell_{\infty,1}}$. We suppose for the moment that all points in N_γ are in $T(\mathcal{G}_Q)$.

3 - Cover of $\Delta\mathcal{H}^\gamma$

$\forall \Delta h \in \Delta\mathcal{H}$ built from $h = [h_1,\dots,h_Q] \in \mathcal{H}$, we have:

$$T(h_1,\dots,h_Q) \in T(\mathcal{G}_Q)$$

Let t be an element in the cover $N_\gamma \subset T(\mathcal{G}_Q)$ such that:

$$\|T(h_1,\dots,h_Q) - t\|_{\ell_{\infty,1}} \leq \gamma$$

we consider an inverse $(g_1 - g_2, \dots, g_{Q-1} - g_Q)$ of t by T in $\mathcal{G}_Q$. This inverse is related to a multiclassifier g built from $(g_1,\dots,g_Q)$. According to (9), we have then:

$$\|\pi_\gamma(\Delta h) - \pi_\gamma(\Delta g)\|_{\ell_{\infty,1}} \leq \|T(h_1,\dots,h_Q) - t\|_{\ell_{\infty,1}} \leq \gamma \quad (13)$$

We have thus built a γ-cover of $\Delta\mathcal{H}^\gamma$ from a γ-cover of $T(\mathcal{G}_Q)$. We must now cope with the additional requirement that the latter must have all its elements in

$T(\mathcal{G}_Q)$. As pointed out before, this is not a problem since, from any γ-cover of $T(\mathcal{G}_Q)$, it is possible to derive a 2γ-cover of $T(\mathcal{G}_Q)$ of same cardinality and included in $T(\mathcal{G}_Q)$.

4 - Bound on the covering number of $T(\mathcal{G}_Q)$

To obtain the final bound, it remains to bound the covering number of $T(\mathcal{G}_Q) \subset T(B_\Lambda)$. To that end, we can make use of results that can be found in functional analysis books such as [12]. They are also exposed in [44] or in [38] in the context of learning theory. We have:

Proposition 3 ([12], prop.1.3.1) *If T, a bounded linear operator acting between arbitrary real Banach spaces (E, ρ_E) and (F, ρ_F), has a rank equal to r, then,*

$$\mathcal{N}(\gamma, T(U_{\rho_E}), \rho_F) \leq \left(\frac{4\|T\|}{\gamma}\right)^r$$

where U_{ρ_E} is the unit ball of E and $\|T\| = \sup_{f \in E} \frac{\|T(f)\|_{\rho_F}}{\|f\|_{\rho_E}}$.

One can make use of this proposition without considering a ball centered on the origine, thanks to the following simple result, a proof of which can for instance be found in [44]:

Proposition 4 *For all balls B_Λ of radius Λ centered in a :*

$$T(B_\Lambda) = \Lambda T(U_{\rho_E}) + T(a)$$

Thus, the covering number of $T(B_\Lambda)$ can also be bounded thanks to this proposition. A bound on $T(\mathcal{G}_Q)$ is deduced because the covering number of a set is always smaller than the covering number of the set where it is included.

6.2 Particular case

It is possible to directly apply the approach we just described in a simple example. Let us assume that $\mathcal{F}$ is the set of linear applications from $\mathbb{R}^d$ onto $\mathbb{R}$:

$$\mathcal{F} = \left\{ \langle w, x \rangle \ / \ w \in \mathbb{R}^d \right\}$$

Suppose that inputs x are in the unit ball: $\|x\| \leq 1$. The set $\mathcal{F}^Q$ is identical to $\{w_i, \ i = 1, .., Q\} \subset \mathbb{R}^{Qd}$ (endowed with the Euclidian norm). Assume that $\sum_{k=1}^{Q} w_k = 0_d$, and that,

$$\sum_{1 \leq i < j \leq Q} \|w_i - w_j\|^2 \leq \frac{1}{2} \tag{14}$$

This last hypothesis does not restrict the study, since multiplying each w_i by the same scalar does not change the resulting classifier. Thus, in order to have (14), one can divide all w_i, $i \in \{1, .., Q\}$ by the square root of $2\sum_{i<j}^{Q} \|w_i - w_j\|^2$. In that case, $\mathcal{F}^{Q(Q-1)/2}$ is in the ball centered at the origin and of radius $1/2$ in $E^{Q(Q-1)/2}$ (recall that here $E = \mathbb{R}^d$ endowed with the usual Euclidian norm). Application T has a finite rank which is smaller than Qd. Let us compute the norm of this application. For $h_1 = w_1, \ldots, h_Q = w_Q$ with,

$$\sum_{i<j}^{Q} \|w_i - w_j\|^2 \leq 1$$

We have:

$$\begin{aligned} \|T(h_1,\ldots,h_Q)\|_{\ell_{\infty,1}} &= \max_{x\in\mathcal{X}} \sum_{i<j}^{Q} |h_i(x)-h_j(x)| &(15)\\ &= \max_{x\in\mathcal{X}} \sum_{i<j}^{Q} |\langle w_i - w_j, x\rangle| &(16)\end{aligned}$$

But, $\|x\| \leq 1$, hence:

$$\begin{aligned} \|T(h_1,\ldots,h_Q)\|_{\ell_{\infty,1}} &\leq \max_{x\in\mathcal{X}} \sum_{i<j}^{Q} \|w_i - w_j\| &(17)\\ &\leq \frac{Q}{\sqrt{2}} \underbrace{\sqrt{\sum_{i<j}^{Q} \|w_i - w_j\|^2}}_{\leq 1} &(18)\end{aligned}$$

and finally:

$$\|T(h_1,\ldots,h_Q)\|_{\ell_{\infty,1}} \leq \frac{Q}{\sqrt{2}}$$

Consequently, according to Proposition 3, since $\Lambda = \frac{1}{2}$ and $F = \mathbb{R}^{Q(Q-1)m}$ endowed with the norm $\|.\|_{\ell_{\infty,1}}$:

$$\mathcal{N}(\gamma, T(B_\Lambda), \rho_F) \leq \left(\frac{Q}{2\sqrt{2}\gamma}\right)^{Qd}$$

Since, from any $\frac{\gamma}{2}$-cover, it is possible to build a γ-cover whose elements are in $T(U_{\rho_E})$, following the same reasoning as in the previous section yields:

$$\mathcal{N}\left(\gamma, \Delta\mathcal{H}^\gamma, \|.\|_{\ell_{\infty,1}}\right) \leq \left(\frac{Q}{\sqrt{2}\gamma}\right)^{Qd} \tag{19}$$

This bound is independant of $\mathcal{X}$, thus:

$$\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq \left(\frac{\sqrt{2}Q}{\gamma}\right)^{Qd}$$

This bound, combined with Theorem 4, provides us with a bound on the generalization error of multiclassifiers built from linear systems. We have explicited a method for linear systems in finite dimensional spaces, but our approach can be readily applied to any systems, which are expansions of Mercer kernels [1]. The latter can indeed be interpreted as linear systems in infinite dimensional spaces. The dimension of the input space appears in our method only to bound the covering number of the application T thanks to Proposition 3. In an infinite dimensional space, similar propositions can be derived for T which has then an "infinite" rank. They are exposed in [12] and their application to learning theory is developed in [44] and [38]. The method we just presented can thus be generalized to infinite dimensional spaces provided Proposition 3 is replaced.

7 Applications

Our approach has been assessed on an open problem in predictive structural biology: the problem of globular protein secondary structure prediction. Seen as a pattern recognition problem, it is a 3-category discrimination task which consists in assigning to each residue (amino acid) in a sequence its conformational state: α-helix, β-strand or coil. This problem appears particularly appropriate to assess bounds on the expected risk. This springs from two reasons. First, the rapid strides made in molecular biology provide us which huge quantities of data, making it possible to study the models of interest in the whole spectrum of regimes, ranging from the small samples to the asymptotic behaviour. Second, studies based on biological and statistical evidence have given birth to reliable estimates of the optimal level of accuracy associated with the different strategies implemented to tackle the problem (see [46] for an upper bound on the accuracy of predictions based on local information and single sequences, or [5] for an overview on performance obtained by using multiple alignments). Apart from the fact that the problem is of central importance in biology, it thus appears interesting due to the possibility to characterize the significance of the results observed (see also [31]). We used the MLR combiner described in [18] to combine three of the current best prediction methods: SOPMA [16], GOR IV [15] and SIMPA96 [27]. The MLR ensemble method, designed to combine discriminant models estimating the class posterior probabilities, performs on the *predictors* a multivariate linear regression of the form:

$$h(x) = \begin{bmatrix} w_1^T \\ \vdots \\ w_k^T \\ \vdots \\ w_Q^T \end{bmatrix} x = \begin{bmatrix} w_{111} \dots w_{1jl} \dots w_{1PQ} \\ \vdots \\ w_{k11} \dots w_{kjl} \dots w_{kPQ} \\ \vdots \\ w_{Q11} \dots w_{Qjl} \dots w_{QPQ} \end{bmatrix} \begin{bmatrix} x_{11} \\ \vdots \\ x_{jl} \\ \vdots \\ x_{PQ} \end{bmatrix}$$

where P is the number of classifiers and Q, as before, is the number of categories. x_{jl} is the estimate of the posterior probability of the l^{th} category produced by the classifier of index j. Let $v = \left[w_1^T, \dots, w_k^T, \dots w_Q^T\right] \in \mathbb{R}^{PQ^2}$. Provided the objective function is adequately chosen (see for instance [30, 8]), a sufficient condition for the components $h_k(x)$ to be probability estimates is for v to belong to the convex set V given by:

$$V = \left\{ \; v \in \mathbb{R}_+^{PQ^2} / \left\{ \begin{array}{l} \forall (l,m), \; \sum_k (w_{k,l,m} - w_{k,l,Q}) = 0 \\ 1_{PQ^2}^T v = Q \end{array} \right. \right\}$$

In order for the predictors to satisfy the aforementioned hypotheses, the outputs of the prediction methods were preliminary post-processed by a neural network. These estimates, as required, are non-negative and sum to unity. Consequently, for each residue, the input of the combiner was a vector $x \in \mathbb{R}^9$ made up of three estimates of the posterior probability of each conformational state. The training set was a recent release of the PDBSELECT database [21]. This database contains 629 chains made up of 147518 residues. The combiner was trained with the least-squares method. Figure 6 summarizes the overall setting of the model. In brief, and to use previous notations, we have: $\mathcal{F} = \{w \in \mathbb{R}^9\}$. Solving a simple quadratic programming problems, it appears that the constraint $v \in V$ implies $\|w\| \leq 1$. $m = 147518$, $Q = 3$, $d = 9$ and $\|x\| \leq \sqrt{3} \;\; \forall x \in \mathbb{R}^9$. The latter property springs from the fact that each prediction method feeding the linear model outputs probability estimates.

In order to apply Theorem 4, a covering number is to be bounded. Implementing the method explained in Section 6 gives:

$$R(h) \leq \underbrace{0.271}_{e.m.r.} + \underbrace{0.025}_{c.t.} = 0.296$$

This bound has been calculated for $\gamma = 0.1$. It is made up of two terms: the empirical margin risk (e.m.r.) and the control term (c.t.). In order to compare

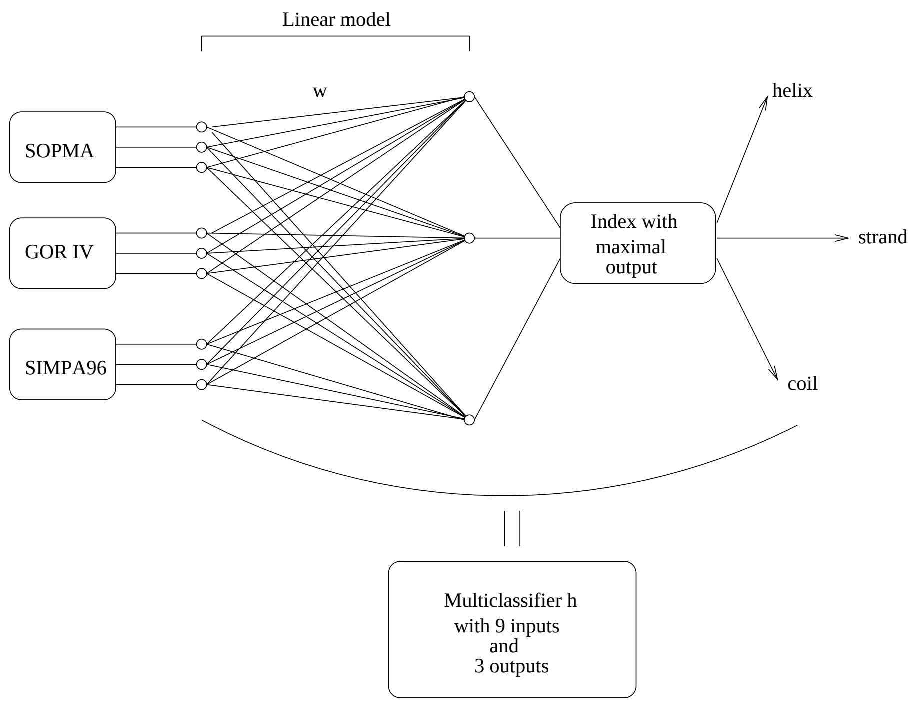

Figure 6 Combination of protein secondary structure prediction methods with a MLR combiner.

Generalization error estimate	Estimate / upper bound
Cross-validation	0.287
VC bound (Graph dimension)	0.311
This study	0.296

Table 1 Estimates of the expected risk of the MLR model.

the resulting upper bound on the generalization error of h with other estimates, a growth function of the multiclassifier was estimated. This multi-class growth function is the extension of the two-class one associated with the *graph dimension* [28], and can be used identically to bound the generalization error of a multiclassifier (see [34, 4] for details). We also implemented a cross-validation procedure introduced in [45], which is considered as providing a good experimental estimate of the generalization error but is computationnally demanding. The results of this experiment can be found in Table 7. Compared to the other theoretical approach based on a combinatorial quantity, our approach yields a more realistic bound and seems thus more appropriate to study the behaviour of the model. This experimental evidence is in accordance with the one detailed in [17].

One of the reasons why our approach peformed satisfactorily on that application is the fact that when bounding the covering number, the specificities of the implementation could be taken into accound, such as the fact that all the inputs are probability estimates (live in a specific convex set). The other theoretical method, based on the growth function, did not provide us with a way to plug this knowledge in the formula. It is worthwile noticing, for instance, that no extension of the VC dimension to the multi-class case is affected here by the constraint $v \in V$ [19].

8 Conclusion

In this paper, different notions of margin for two-class and multi-class discriminant systems built from real-valued functions have been detailed. The motivation to study these margins is twofold. First, experimental evidence strongly suggests that uniform convergence results based on margins are superior to the classical ones. Second, margins provide us with a way to define new inductive principles. The most famous example of a training criterion intended to minimize directly the generalization error is given by Vapnik's support vector machine method [40]. Up to now, however, this machine could only compute dichotomies or perform regression tasks. Theorem 4 can be readily applied to determine the learning algorithm of multi-class support vector machines (linear and non-linear). This would represent a significant improvement, since no bound on the guaranteed risk has been derived for the sole support vector machine currently performing multi-class discrimination task, and developed independently by Vapnik [40], Weston & Watkins [43] and Bredensteiner & Bennett [9], among others. This means that for this SVM, the significance of the implementation of the structural risk minimization inductive principle is very difficult to assess. On the contrary, the extension of Bartlett's result provides us with a means to minimize simultaneously the two terms of the bound on the expected risk. This should contribute to decrease the wide gap still parting theoretical results from empirical ones in the field of statistical learning.

Acknowledgement

The authors gratefully acknowledge the support of the ESPRIT funded Working Group N. 27150 "Neural Networks and Computational Learning Theory". We would also like to thank O. Teytaud and G. Gavin for interesting discussions. Special thanks to F. d'Alché for discussions and for pointing out to us interesting bibliography.

References

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

[2] E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research*, 2000. (to appear).

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

[4] M. Anthony. Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. *Neural Computing Surveys*, 1:1–47, 1997.

[5] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[6] P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[7] 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.

[8] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

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

[10] L. Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[11] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

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

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

[14] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[15] J. Garnier, J.-F. Gibrat, and B. Robson. GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. *Methods Enzymol.*, 266:540–553, 1996.

[16] C. Geourjon and G. Deléage. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS*, 11(6):681–684, 1995.

[17] 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.

[18] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5):413–421, 1999.

[19] Y. Guermeur, H. Paugam-Moisy, and P. Gallinari. Multivariate linear regression on classifier outputs: a capacity study. In *ICANN'98*, pages 693–698, 1998.

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

[21] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.*, 3:522–524, 1994.

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

[23] 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.

[24] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and Systems Sciences*, 48(3):464–497, 1994.

[25] M.J. Kearns, R.E. Schapire, and L.M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–142, 1994.

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

[27] J.-M. Levin. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.*, 10(7):771–776, 1997.

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

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

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

[31] B. Rost, C. Sander, and R. Schneider. Redefining the Goals of Protein Secondary Structure Prediction. *J. Mol. Biol.*, 235:13–26, 1994.

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

[33] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

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

[35] 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.

[36] J. Shawe-Taylor and N. Cristianini. Robust Bounds on Generalization from the Margin Distribution. Technical Report NC2-TR-1998-029, NeuroCOLT2, 1998.

[37] A. Smola, A. Elisseeff, B. Schölkopf, and R.C. Williamson. Entropy numbers for convex combinations and multilayer networks. In Smola, Bartlett, Schölkopf, and Schuurmans, editors, *Advances in Large Margin Classifiers.* MIT Press, 1999.

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

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

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

[41] 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.

[42] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of the means to their expectations. *Theory of Probability and its Applications*, 26:532–553, 1981.

[43] 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.

[44] 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*, 1999. (to appear).

[45] D.H. Wolpert. Stacked Generalization. *Neural Networks*, 5:241–259, 1992.

[46] X. Zhang, J.P. Mesirov, and D.L. Waltz. Hybrid System for Protein Secondary Structure Prediction. *J. Mol. Biol.*, 225:1049–1063, 1992.