MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1681 November 1999
C.B.C.L Paper No. 184

A note on the generalization performance of kernel classifiers with margin.

Theodoros Evgeniou and Massimiliano Pontil

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1681.ps

Abstract

We present distribution independent bounds on the generalization misclassification performance of a family of kernel classifiers with margin. Support Vector Machine classifiers (SVM) stem out of this class of machines. The bounds are derived through computations of the V_γ dimension of a family of loss functions where the SVM one belongs to. Bounds that use functions of margin distributions (i.e. functions of the slack variables of SVM) are derived.

This report describes research done at the Center for Biological & Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. This research was sponsored by the Office of Naval Research under contract No. N00014-93-1-0385 and contract No. N00014-95-1-0600. Partial support was also provided by Daimler-Benz AG, Eastman Kodak, Siemens Corporate Research, Inc., AT&T, Digital Equipment Corporation, Central Research Institute of Electrical Power Industry, and Honda.

1 Introduction

Deriving bounds on the generalization performance of kernel classifiers has been an important theoretical topic of research in recent years [4, 8, 9, 10, 12]. We present new bounds on the generalization performance of a family of kernel classifiers with margin, from which Support Vector Machines (SVM) can be derived. The bounds use the V_γ dimension of a class of loss functions, where the SVM one belongs to, and functions of the margin distribution of the machines (i.e. functions of the slack variables of SVM - see below).
We consider classification machines of the form:

$$\begin{array}{rc} \min & \sum_{i=1}^{m} V(y_i, f(\mathbf{x}_i)) \\ \text{subject to} & \|f\|_K^2 \leq A^2 \end{array} \tag{1}$$

where we use the following notation:

- $D_m = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, with $(\mathbf{x}_i, y_i) \in R^n \times \{-1, 1\}$ sampled according to an unknown probability distribution $P(\mathbf{x}, y)$, is the training set.
- $V(y, f(\mathbf{x}))$ is the loss function measuring the distance (error) between $f(\mathbf{x})$ and y.
- f is a function in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ defined by kernel K, with $\|f\|_K^2$ being the norm of f in $\mathcal{H}$ [11, 2]. We also call f a hyperplane, since it is such in the feature space induced by the kernel K [11, 10].
- A is a constant.

Classification of a new test point $\mathbf{x}$ is always done by simply considering *the sign* of $f(\mathbf{x})$.
Machines of this form have been motivated in the framework of statistical learning theory. We refer the reader to [10, 6, 3] for more details. In this paper we study the generalization performance of these machines for choices of the loss function V that are relevant for classification. In particular we consider the following loss functions:

- Misclassification loss function:

$$V(y, f(\mathbf{x})) = V^{msc}(yf(\mathbf{x})) = \theta(-yf(\mathbf{x})) \tag{2}$$

- Hard margin loss function:

$$V(y, f(\mathbf{x})) = V^{hm}(yf(\mathbf{x})) = \theta(1 - yf(\mathbf{x})) \tag{3}$$

- Soft margin loss function:

$$V(y, f(\mathbf{x})) = V^{sm}(yf(\mathbf{x})) = \theta(1 - yf(\mathbf{x}))(1 - yf(\mathbf{x})), \tag{4}$$

where θ is the Heavyside function. Loss functions (3) and (4) are "margin" ones because the only case they do not penalize a point $(\mathbf{x}, y)$ is if $yf(\mathbf{x}) \geq 1$. For a given f, these are the points that are correctly classified *and* have distance $\frac{|f(\mathbf{x})|}{\|f\|^2} \geq \frac{1}{\|f\|^2}$ from the surface $f(\mathbf{x}) = 0$ (hyperplane in the feature space induced by the kernel K [10]). For a point $(\mathbf{x}, y)$, quantity $\frac{yf(\mathbf{x})}{\|f\|}$ is its margin,

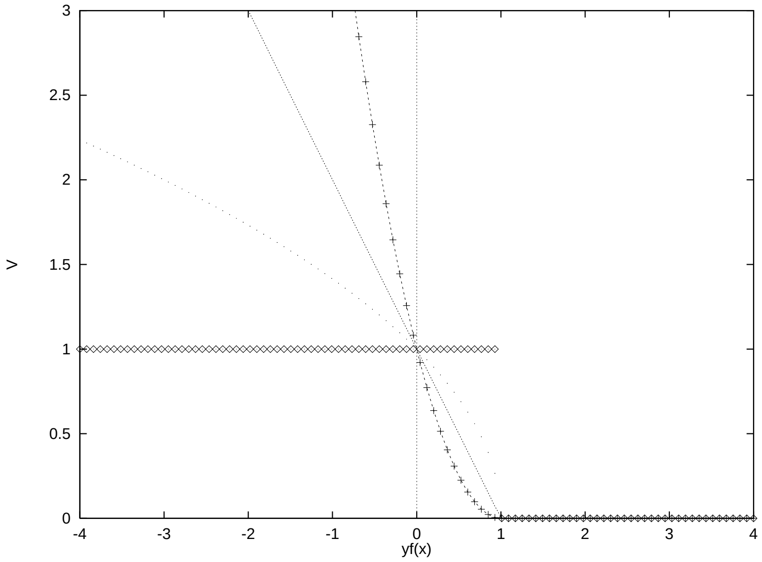

Figure 1: Hard margin loss (line with diamond-shaped points), soft margin loss (solid line), nonlinear soft margin with $\sigma = 2$ (line with crosses), and $\sigma = \frac{1}{2}$ (dotted line)

and the probability of having $\frac{yf(\mathbf{x})}{\|f\|} \geq \delta$ is called the *margin distribution* of hypothesis f. For SVM, quantity $\theta(1-y_i f(\mathbf{x}_i))(1-y_i f(\mathbf{x}_i))$ is known as the *slack variable* corresponding to training point $(\mathbf{x}_i, y_i)$ [10].

We will also consider the following family of margin loss functions (nonlinear soft margin loss functions):

$$V(y, f(\mathbf{x})) = V^{\sigma}(yf(\mathbf{x})) = \theta(1 - yf(\mathbf{x}))(1 - yf(\mathbf{x}))^{\sigma}. \tag{5}$$

Loss functions (3) and (4) correspond to the choice of $\sigma = 0, 1$ respectively. In figure 1 we plot some of the possible loss functions for different choices of the parameter σ.

To study the statistical properties of machines (1) we use some well known results that we now briefly present. First we define some more notation, and then state the results from the literature that we will use in the next section.

We use the following notation:

- $R^V_{emp}(f) = \sum_{i=1}^m V(y_i, f(\mathbf{x}_i))$ is the empirical error made by f on the training set D_m, using V as the loss function.

- $R^V(f) = \int_{R^n \times \{-1,1\}} V(y, f(\mathbf{x}))\ P(\mathbf{x}, y)\ d\mathbf{x}\ dy$ is the expected error of f using V as the loss function.

- Given a hypothesis space of functions $\mathcal{F}$ (i.e. $\mathcal{F} = \{f \in \mathcal{H} : \|f\|^2 \leq A^2\}$), we note by $h^{V\mathcal{F}}_{\gamma}$ the V_γ dimension of the loss function $V(y, f(\mathbf{x}))$ in $\mathcal{F}$, which is defined as follows [1]:

Definition 1.1 *Let $A \leq V(y, f(\mathbf{x})) \leq B$, $f \in \mathcal{F}$, with A and $B < \infty$. The V_γ-dimension of V in $\mathcal{F}$ (of the set of functions $\{V(y, f(\mathbf{x}))\ \mid\ f \in F\}$) is defined as the the maximum number h of vectors $(\mathbf{x}_1, y_1)\ldots,(\mathbf{x}_h, y_h)$ that can be separated into two classes in all 2^h possible ways using rules:*

$$\textit{class 1 if: } V(y_i, f(\mathbf{x}_i)) \geq s + \gamma$$
$$\textit{class -1 if: } V(y_i, f(\mathbf{x}_i)) \leq s - \gamma$$

for $f \in \mathcal{F}$ and some $s \geq 0$. If, for any number m, it is possible to find m points $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_m, y_m)$ that can be separated in all the 2^m possible ways, we will say that the V_γ-dimension of V in $\mathcal{F}$ is infinite.

If instead of a fixed s for all points we use a different s_i for each $(\mathbf{x}_i, y_i)$, we get what is called the *fat-shattering* dimension fat_γ [1]. Notice that definition (1.1) includes the special case in which we directly measure the V_γ dimension of the space of functions F, i.e. $V(y, f(\mathbf{x})) = f(\mathbf{x})$. We will need such a quantity in theorem 2.2 below.
Using the V_γ dimension we can study the statistical properties of machines of the form (1) based on a standard theorem that characterizes the generalization performance of these machines.

Theorem 1.1 (Alon et al., 1993) *Let $A \leq V(y, f(\mathbf{x})) \leq B$, $f \in \mathcal{F}$, $\mathcal{F}$ be a set of bounded functions. For any $\epsilon \geq 0$, for all $m \geq \frac{2}{\epsilon^2}$ we have that if $h_\gamma^{V_\mathcal{F}}$ is the V_γ dimension of V in $\mathcal{F}$ for $\gamma = \alpha\epsilon$ ($\alpha \geq \frac{1}{48}$), $h_\gamma^{V_\mathcal{F}}$ finite, then:*

$$Pr\left\{\sup_{f\in\mathcal{F}}\left|R_{\text{emp}}^V(f) - R^V(f)\right| > \epsilon\right\} \leq \mathcal{G}(\epsilon, m, h_\gamma^{V_\mathcal{F}}), \tag{6}$$

where $\mathcal{G}$ is an increasing function of $h_\gamma^{V_\mathcal{F}}$ and a decreasing function of ϵ and m, with $\mathcal{G} \to 0$ as $m \to \infty$.

In [1] the fat-shattering dimension was used, but a close relation between that and the V_γ dimension [1] make the two equivalent for our purpose[1]. Closed forms of $\mathcal{G}$ can be derived (see for example [1]) but we do not present them here for simplicity of notation. Notice that since we are interested in classification, we only consider $\epsilon < 1$, so we will only discuss the case $\gamma < 1$ (since γ is about $\frac{1}{48}\epsilon$).
In "standard" statistical learning theory the VC dimension is used instead of the V_γ one [10]. However, for the type of machines we are interested in the VC dimension turns out not to be appropriate: it is not influenced by the choice of the hypothesis space $\mathcal{F}$ through the choice of A, and in the case that $\mathcal{F}$ is an infinite dimensional RKHS, the VC-dimension of the loss functions we consider turns out to be infinite (see for example [5]). Instead, scale-sensitive dimensions (such as the V_γ or fat-shattering one [1]) have been used in the literature, as we will discuss in the last section.

2 Main results

We study the loss functions (2 - 5). For classification machines the quantity we are interested in is the expected misclassification error of the solution f of problem 1. With some abuse of notation we note this with R^{msc}. Similarly we will note with R^{hm}, R^{sm}, and R^σ the expected risks using loss functions (3), (4) and (5), respectively, and with R^{hm}_{emp}, R^{sm}_{emp}, and R^σ_{emp}, the

[1]In [1] it is shown that $V_\gamma \leq \text{fat}_\gamma \leq \frac{1}{\gamma}V_{\frac{\gamma}{2}}$.

corresponding empirical errors. We will not consider machines of type (1) with V^{msc} as the loss function, for a clear reason: the solution of the optimization problem:

$$\begin{array}{rc} \min & \sum_{i=1}^{m} \theta(-y_i f(\mathbf{x}_i)) \\ \text{subject to} & \|f\|_K^2 \leq A^2 \end{array}$$

is independent of A, since for any solution f we can always rescale f and have the same cost $\sum_{i=1}^{m} \theta(-y_i f(\mathbf{x}_i))$.
For machines of type (1) that use V^{sm} or V^{σ} as the loss function, we prove the following:

Theorem 2.1 *The V_γ dimension h for $\theta(1-yf(\mathbf{x}))(1-yf(\mathbf{x}))^{\sigma}$ in hypothesis spaces $\mathcal{F}_A = \{f \in \mathcal{H} | \|f\|_K^2 \leq A^2\}$ (of the set of function $\{\theta(1-yf(\mathbf{x}))(1-yf(\mathbf{x}))^{\sigma} \mid f \in \mathcal{F}_A\}$) and $y \in \{-1, 1\}$, is finite for $\forall\ 0 < \gamma$. If D is the dimensionality of the RKHS $\mathcal{H}$, R^2 is the radius of the smallest sphere centered at the origin containing the data $\mathbf{x}$ in the RKHS, and $B > 1$ is an upper bound on the values of the loss function, then h is upper bounded by:*

- $O(min(D, \frac{R^2A^2}{\gamma^{\frac{2}{\sigma}}}))$ *for* $\sigma < 1$
- $O(min(D, \frac{(\sigma B^{\frac{\sigma-1}{\sigma}})^2 R^2 A^2}{\gamma^2}))$ *for* $\sigma \geq 1$

Proof
The proof is based on the following theorem [7] (proved for the fat-shattering dimension, but as mentioned above, we use it for the "equivalent" V_γ one).

Theorem 2.2 [Gurvits, 1997] The V_γ dimension h of the set of functions[2] $\mathcal{F}_A = \{f \in \mathcal{H} | \|f\|_K^2 \leq A^2\}$ is finite for $\forall\ \gamma > 0$. If D is the dimensionality of the RKHS, then $h \leq O(min(D, \frac{R^2A^2}{\gamma^2}))$, where R^2 is the radius of the smallest sphere in the RKHS centered at the origin here the data belong to.

Let $2N$ be the largest number of points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{2N}, y_{2N})\}$ that can be shattered using the rules:

$$\begin{array}{rl} \text{class 1 if} & \theta(1-y_i f(\mathbf{x}_i))(1-y_i f(\mathbf{x}_i))^{\sigma} \geq s + \gamma \\ \text{class } -1 \text{ if} & \theta(1-y_i f(\mathbf{x}_i))(1-y_i f(\mathbf{x}_i))^{\sigma} \leq s - \gamma \end{array} \quad (7)$$

for some s with $0 < \gamma \leq s$. After some simple algebra these rules can be decomposed as:

$$\begin{array}{rl} \text{class 1 if} & f(\mathbf{x}_i) - 1 \leq -(s+\gamma)^{\frac{1}{\sigma}} \text{ (for } y_i = 1 \text{)} \\ \text{or} & f(\mathbf{x}_i) + 1 \geq (s+\gamma)^{\frac{1}{\sigma}} \text{ (for } y_i = -1 \text{)} \\ \text{class } -1 \text{ if} & f(\mathbf{x}_i) - 1 \geq -(s-\gamma)^{\frac{1}{\sigma}} \text{ (for } y_i = 1 \text{)} \\ \text{or} & f(\mathbf{x}_i) + 1 \leq (s-\gamma)^{\frac{1}{\sigma}} \text{ (for } y_i = -1 \text{)} \end{array} \quad (8)$$

From the $2N$ points at least N are either all class -1, or all class 1. Consider the first case (the other case is exactly the same), and for simplicity of notation let's assume the first N points are

[2]As mentioned above, in this case we can consider $V(y, f(\mathbf{x})) = f(\mathbf{x})$.

class -1. Since we can shatter the $2N$ points, we can also shatter the first N points. Substituting y_i with 1, we get that we can shatter the N points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ using rules:

$$\begin{array}{rl} \text{class } 1 \text{ if } & f(\mathbf{x}_i) + 1 \geq (s+\gamma)^{\frac{1}{\sigma}} \\ \text{class } -1 \text{ if } & f(\mathbf{x}_i) + 1 \leq (s-\gamma)^{\frac{1}{\sigma}} \end{array} \tag{9}$$

Notice that the function $f(\mathbf{x}_i) + 1$ has RKHS norm bounded by A^2 plus a constant C (equal to the inverse of the eigenvalue corresponding to the constant basis function in the RKHS - if the RKHS does not include the constant functions, we can define a new RKHS with the constant and use the new RKHS norm). Furthermore there is a "margin" between $(s+\gamma)^{\frac{1}{\sigma}}$ and $(s-\gamma)^{\frac{1}{\sigma}}$ which we can lower bound as follows.
For $\sigma < 1$, assuming $\frac{1}{\sigma}$ is an integer (if not, we can take the closest lower integer),

$$\frac{1}{2}\left((s+\gamma)^{\frac{1}{\sigma}} - (s-\gamma)^{\frac{1}{\sigma}}\right) = \frac{1}{2}((s+\gamma)-(s-\gamma))\left(\sum_{k=0}^{\frac{1}{\sigma}-1}(s+\gamma)^{\frac{1}{\sigma}-1-k}(s-\gamma)^k\right) \geq \gamma\gamma^{\frac{1}{\sigma}-1} = \gamma^{\frac{1}{\sigma}}. \tag{10}$$

For $\sigma \geq 1$, σ integer (if not, we can take the closest upper integer) we have that:

$$\begin{aligned} 2\gamma &= \left((s+\gamma)^{\frac{1}{\sigma}}\right)^{\sigma} - \left((s-\gamma)^{\frac{1}{\sigma}}\right)^{\sigma} = ((s+\gamma)^{\frac{1}{\sigma}} - (s-\gamma)^{\frac{1}{\sigma}})\left(\sum_{k=0}^{\sigma-1}((s+\gamma)^{\frac{1}{\sigma}})^{\sigma-1-k}((s-\gamma)^{\frac{1}{\sigma}})^k\right) \leq \\ &\leq ((s+\gamma)^{\frac{1}{\sigma}} - (s-\gamma)^{\frac{1}{\sigma}})\sigma B^{\frac{\sigma-1}{\sigma}} \end{aligned}$$

from which we obtain:

$$\frac{1}{2}\left((s+\gamma)^{\frac{1}{\sigma}} - (s-\gamma)^{\frac{1}{\sigma}}\right) \geq \frac{\gamma}{\sigma B^{\frac{\sigma-1}{\sigma}}} \tag{11}$$

Therefore N cannot be larger than the V_γ dimension of the set of functions with RKHS norm $\leq A^2 + C$ and margin at least $\gamma^{\frac{1}{\sigma}}$ for $\sigma < 1$ (from eq. (10)) and $\frac{\gamma}{\sigma B^{\frac{\sigma-1}{\sigma}}}$ for $\sigma \geq 1$ (from eq. (11)). Using theorem 2.2, and ignoring constant factors (also ones because of C), the theorem is proved.
$\square$

In figure 2 we plot the V_γ dimension for $R^2A^2 = 1$, $B = 1$, $\gamma = 0.9$, and D infinite. Notice that as $\sigma \to 0$, the dimension goes to infinity. For $\sigma = 0$ the V_γ dimension becomes the same as the VC dimension of hyperplanes, which is infinite in this case. For σ increasing above 1, the dimension also increases: intuitively the margin γ becomes smaller relatively to the values of the loss function.
Using theorems 2.1 and 1.1 we can bound the expected error of the solution f of machines (1):

$$Pr\left\{\left|R^V_{\text{emp}}(f) - R^V(f)\right| > \epsilon\right\} \leq \mathcal{G}(\epsilon, m, h_\gamma), \tag{12}$$

where V is V^{sm} or V^σ. To get a bound on the expected misclassification error $R^{msc}(f)$ we use the following simple observation:

$$V^{msc}(y, f(\mathbf{x})) \leq V^\sigma(y, f(\mathbf{x})) \text{ for } \forall\, \sigma, \tag{13}$$

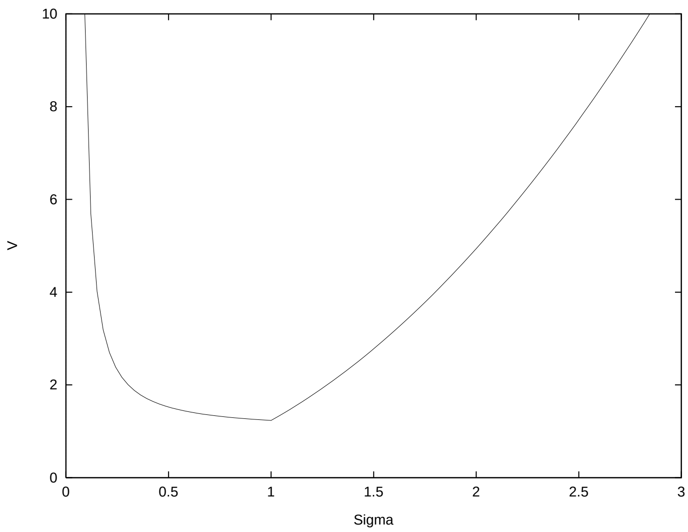

Figure 2: Plot of the V_γ dimension as a function of σ for $\gamma = .9$

So we can bound the expected misclassification error of the solution of machine (1) under V^{sm} and V^σ using the V_γ dimension of these loss functions and the empirical error of f measured using again these loss functions. In particular we get that for $\forall\sigma$, with probability $1 - \mathcal{G}(\epsilon, m, h_\gamma^{V^\sigma_\mathcal{F}})$:

$$R^{msc}(f) \leq R^\sigma_{emp}(f) + \epsilon \tag{14}$$

where ϵ and γ are related as stated in theorem 1.1.

Unfortunately we cannot use theorems 2.1 and 1.1 for the V^{hm} loss function. For this loss function, since it is a binary-valued function, the V_γ dimension is the same as the VC-dimension, which, as mentioned above, is not appropriate to use in our case. Notice, however, that for $\sigma \to 0$, V^σ approaches V^{hm} pointwise (from theorem 2.1 the V_γ dimension also increases towards infinity). Regarding the empirical error, this implies that $R^\sigma \to R^{hm}$, so, theoretically, we can still bound the misclassification error of the solution of machines with V^{hm} using:

$$R^{msc}(f) \leq R^{hm}_{emp}(f) + \epsilon + \max(R^\sigma_{emp}(f) - R^{hm}_{emp}(f), 0), \tag{15}$$

where $R^\sigma_{emp}(f)$ is measured using V^σ for some σ. Notice that changing σ we get a family of bounds on the expected misclassification error. Finally, we remark that it could be interesting to extend theorem 2.1 to loss functions of the form $\theta(1-yf(\mathbf{x}))h(1-yf(\mathbf{x}))$, with h any continuous monotone function.

3 Discussion

In recent years there has been significant work on bounding the generalization performance of classifiers using scale-sensitive dimensions of real-valued functions out of which indicator functions can be generated through thresholding (see [4, 9, 8],[3] and references therein). This is unlike the "standard" statistical learning theory approach where classification is typically studied using the theory of indicator functions (binary valued functions) and their VC-dimension [10].

The work presented in this paper is similar in spirit with that of [3], but significantly different as we now briefly discuss.
In [3] a theory was developed to justify machines with "margin". The idea was that a "better" bound on the generalization error of a classifier can be derived by excluding training examples on which the hypothesis found takes a value close to zero (as mentioned above, classification is performed after thresholding a real valued function). Instead of measuring the empirical misclassification error, as suggested by the standard statistical learning theory, what was used was the number of misclassified training points *plus* the number of training points on which the hypothesis takes a value close to zero. Only points classified correctly with some "margin" are considered correct. In [3] a different notation was used: the parameter A in equation (1) was fixed to 1, while a margin ψ was introduced inside the hard margin loss, i.e $\theta(\psi - yf(x))$. Notice that the two notations are equivalent: given a value A in our notation we have $\psi = A^{-1}$ in the notation of [3]. Below we adapt the results in [3] to the setup of this paper, that is, we set $\psi = 1$ and let A vary. Two main theorems were proven in [3].

Theorem 3.1 (Bartlett, 1998) *For a given A, with probability $1-\delta$, every function f with $\|f\|_K^2 \leq A^2$ has expected misclassification error $R^{msc}(f)$ bounded as:*

$$R^{msc}(f) < R^{hm}_{emp}(f) + \sqrt{\frac{2}{m}(dln(34em/d)\log_2(578m) + ln(4/\delta)}, \tag{16}$$

where d is the fat-shattering dimension fat_γ of the hypothesis space $\{f : \|f\|_K^2 \leq A^2\}$ for $\gamma = \frac{1}{16A}$.

Unlike in this paper, in [3] this theorem was proved without using theorem 1.1. Although practically both bound (16) and the bounds derived above are not tight and therefore not practical, bound (16) seems easier to use than the ones presented in this paper.
It is important to notice that, like bounds (12), (14), and (15), theorem 3.1 holds for a fixed A [3]. In [3] theorem 3.1 was extended to the case where the parameter A (or ψ in the notations of [3]) is not fixed, which means that the bound holds for all functions in the RKHS. In particular the following theorem gives a bound on the expected misclassification error of a machine that holds *uniformly* over all functions:

Theorem 3.2 (Bartlett, 1998) *For any f with $\|f\|_K < \infty$, with probability $1-\delta$, the misclassification error $R^{mcs}(f)$ of f is bounded as:*

$$R^{msc}(f) < R^{hm}_{emp}(f) + \sqrt{\frac{2}{m}(dln(34em/d)\log_2(578m) + ln(8\|f\|/\delta)}, \tag{17}$$

where d is the fat-shattering dimension fat_γ of the hypothesis space consisting of all functions in the RKHS with norm $\leq \|f\|_K^2$, and with $\gamma = \frac{1}{32\|f\|}$.

Notice that the only differences between (16) and (17) are the $ln(8\|f\|/\delta)$ instead of $ln(4/\delta)$, and that $\gamma = \frac{1}{32\|f\|}$ instead of $\gamma = \frac{1}{16A}$.
So far we studied machines of the form (1), where A is fixed *a priori*. In practice learning machines used, like SVM, do not have A fixed a priori. For example in the case of SVM the problem is formulated [10] as minimizing:

$$\min \quad \sum_{i=1}^m \theta(1 - y_i f(\mathbf{x}_i))(1 - y_i f(\mathbf{x}_i)) + \lambda\|f\|_K^2 \tag{18}$$

where λ is known as the *regularization parameter*. In the case of machines (18) we do not know the norm of the solution $\|f\|_K^2$ before actually solving the optimization problem, so it is not clear what the "effective" A is. Since we do not have a fixed upper bound on the norm $\|f\|_K^2$ *a priori*, we **cannot** use the bounds of section 2 or theorem 3.1 for machines of the form (18). Instead, we need to use bounds that hold uniformly for *all* A (or ψ if we follow the setup of [3]), for example the bound of theorem 3.2, so that the bound also holds for the solution of (18) we find. In fact theorem 3.2 has been used directly to get bounds on the performance of SVM [4]. A straightforward applications of the methods used to extend theorem 3.1 to 3.2 can also be used to extend the bounds of section 2 to the case where A is not fixed (and therefore hold for all f with $\|f\| < \infty$), and we leave this as an exercise.
There is another way to see the similarity between machines (1) and (18). Notice that the formulation (1) the regularization parameter λ of (18) can be seen as the *Lagrange multiplier* used to solve the constrained optimization problem (1). That is, problem (1) is equivalent to:

$$\max_\lambda \min_f \sum_{i=1}^{m} V(y_i, f(\mathbf{x}_i)) + \lambda(\|f\|_K^2 - A^2) \tag{19}$$

for $\lambda \geq 0$, which is similar to problem (18) that is solved in practice. However in the case of (19) the Lagrange multiplier λ is not known before having the training data, unlike in the case of (18).
So, to summarize, for the machines (1) studied in this paper, A is fixed a priori and the "regularization parameter" λ is not known a priori, while for machines (18) the parameter λ is known a priori, but the norm of the solution (or the effective A) is not known a priori. As a consequence we can use the theorems of this paper for machines (1) but not for (18). To do the second we need a technical extension of the results of section 2 similar to the extension of theorem 3.1 to 3.2 done in [3]. On the practical side, the important issue for both machines (1) and (18) is how to choose A or λ. We believe that the theorems and bounds discussed in sections 2 and 3 cannot be practically used for this purpose. Criteria for the choice of the regularization parameter exist in the literature - such as cross validation and generalized cross validation - (for example see [10, 11],[6] and references therein), and is the topic of ongoing research. Finally, as our results indicate, the generalization performance of the learning machines can be bounded using any function of the slack variables and therefore of the margin distribution. Is it, however, the case that the slack variables (margin distributions or any functions of these) are *the* quantities that control the generalization performance of the machines, or there are other important geometric quantities involved? Our results suggest that there are many quantities related to the generalization performance of the machines, but it is not clear that these are the most important ones.

Acknowledgments We wish to thank Peter Bartlett for useful comments. *Acknowledgments*

References

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

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.

[3] P. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important that the size of the network. *IEEE Transactions on Information Theory*, 1998.

[4] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machine and other patern classifiers. In C. Burges B. Scholkopf, editor, *Advances in Kernel Methods–Support Vector Learning*. MIT press, 1998.

[5] T. Evgeniou and M. Pontil. On the v-gamma dimension for regression in reproducing kernel hilbert spaces. A.i. memo, MIT Artificial Intelligence Lab., 1999.

[6] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. A.I. Memo No. 1654, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1999.

[7] L. Gurvits. A note on scale-sensitive dimension of linear bounded functionals in banach spaces. In *Proceedings of Algorithm Learning Theory*, 1997.

[8] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998. To appear. Also: NeuroCOLT Technical Report NC-TR-96-053, 1996, ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports.

[9] J. Shawe-Taylor and N. Cristianini. Robust bounds on generalization from the margin distribution. Technical Report NeuroCOLT2 Technical Report NC2-TR-1998-029, NeuroCOLT2, 1998.

[10] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[11] G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[12] R. Williamson, A. Smola, and B. Scholkopf. Generalization performance of regularization networks and support vector machines via entropy numbers. Technical Report NC-TR-98-019, Royal Holloway College University of London, 1998.