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

A.I. Memo No. 1656 May 1999
C.B.C.L Paper No. 172

On the V_γ Dimension for Regression in Reproducing Kernel Hilbert Spaces

Theodoros Evgeniou, 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-1656.ps

Abstract

This paper presents a computation of the V_γ dimension for regression in bounded subspaces of Reproducing Kernel Hilbert Spaces (RKHS) for the Support Vector Machine (SVM) regression ϵ-insensitive loss function, and general L_p loss functions. Finiteness of the V_γ dimension is shown, which also proves uniform convergence in probability for regression machines in RKHS subspaces that use the L_ϵ or general L_p loss functions. This paper presents a novel proof of this result also for the case that a bias is added to the functions in the RKHS.

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 National Science Foundation under contract No. IIS-9800032, 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

In this paper we present a computation of the V_γ dimension of real-valued functions $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|^p$ and (Vapnik's ϵ-insensitive loss function [10]) $L(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\epsilon$ with f in a bounded sphere in a Reproducing Kernel Hilbert Space (RKHS). We show that the V_γ dimension is finite for these loss functions, and compute an upper bound on it. We present two solutions to the problem. First we discuss a simple argument which leads to a loose upper bound on the V_γ dimension. Then we refine the result in the case of infinite dimensional RKHS, which is often the type of hypothesis spaces considered in the literature (i.e. Radial Basis Functions [8, 5]). Our result applies to standard regression learning machines such as Regularization Networks (RN) and Support Vector Machines (SVM). We also present a novel computation of the V_γ dimension when a bias is added to the functions, that is with f being of the form $f = f_0 + b$, where $b \in R$ and f_0 is in a sphere in an infinite dimensional RKHS.

For a regression learning problem using L as a loss function it is known [1] that finiteness of the V_γ dimension for all $\gamma > 0$ is a necessary and sufficient condition for uniform convergence in probability [10]. So the results of this paper have implications for uniform convergence both for RN and for SVM regression [4].

Previous related work addressed the problem of pattern recognition where L is an indicator function [**?**, 6]. The fat-shattering dimension [1] was considered instead of the V_γ one. A different approach to proving uniform convergence for RN and SVM is given in [12] where covering number arguments using entropy numbers of operators are presented. In both cases, regression as well as the case of non-zero bias b were marginally considered.

The paper is organized as follows. Section 2 outlines the background and motivation of this work. The reader familiar with statistical learning theory and RKHS can skip this section. Section 3 presents a simple proof of the results as well as some upper bounds to the V_γ dimension. Section 4 presents a refined computation in the case of infinite dimensional RKHS, as well as the case where the hypothesis space consists of functions of the form $f = f_0 + b$ where $b \in R$ and f_0 in a sphere in a RKHS. Finally, section 5 discusses possible extensions of this work.

2 Background and Motivation

We consider the problem of learning from examples as it is viewed in the framework of statistical learning theory [10]. We are given a set of l examples $\{(\mathbf{x}_1, y_1), .., (\mathbf{x}_l, y_l)\}$ generated by randomly sampling from a space $X \times Y$ with $X \subset R^d$, $Y \subset R$ according to an unknown probability distribution $P(\mathbf{x}, y)$. Throughout the paper we assume that X and Y are bounded. Using this set of examples the problem of learning consists of finding a function $f : X \to Y$ that can be used given any new point $\mathbf{x} \in X$ to predict the corresponding value y.

The problem of learning from examples is known to be ill-posed [10, 9]. A classical way to solve it is to perform Empirical Risk Minimization (ERM) with respect to a certain loss function, while restricting the solution to the problem to be in a "small" hypothesis space [10]. Formally this means minimizing the empirical risk $I_{\text{emp}}[f] = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(\mathbf{x}_i))$ with $f \in \mathcal{H}$, where L is the loss function measuring the error when we predict $f(\mathbf{x})$ while the actual value is y, and $\mathcal{H}$ is a given hypothesis space.

In this paper, we consider hypothesis spaces of functions which are hyperplanes in some feature space:

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} w_n \phi_n(\mathbf{x}) \tag{1}$$

with:

$$\sum_{n=1}^{\infty} \frac{w_n^2}{\lambda_n} < \infty \tag{2}$$

where $\phi_n(\mathbf{x})$ is a set of given, linearly independent basis functions, λ_n are given non-negative constants such that $\sum_{n=1}^{\infty} \lambda_n^2 < \infty$. Spaces of functions of the form (1) can also be seen as Reproducing Kernel Hilbert Spaces (RKHS) [2, 11] with kernel K given by:

$$K(\mathbf{x}, \mathbf{y}) \equiv \sum_{n=1}^{\infty} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{y}). \tag{3}$$

For any function f as in (1), quantity (2) is called the RKHS norm of f, $\|f\|_K^2$, while the number D of features ϕ_n (which can be finite, in which case all sums above are finite) is the dimensionality of the RKHS.
If we restrict the hypothesis space to consist of functions in a RKHS with norm less than a constant A, the general setting of learning discussed above becomes:

$$\begin{array}{rc} \text{Minimize :} & \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(\mathbf{x}_i)) \\ \text{subject to :} & \|f\|_K^2 \leq A^2. \end{array} \tag{4}$$

An important question for any learning machine of the type (4) is whether it is consistent: as the number of examples $(\mathbf{x}_i, y_i)$ goes to infinity the expected error of the solution of the machine should converge in probability to the minimum expected error in the hypothesis space [10, 3]. In the case of learning machines performing ERM in a hypothesis space (4), consistency is shown to be related with uniform convergence in probability [10], and necessary and sufficient conditions for uniform convergence are given in terms of the V_γ dimension of the hypothesis space considered [1, 7], which is a measure of complexity of the space.
In statistical learning theory typically the measure of complexity used is the VC-dimension. However, as we show below, the VC-dimension in the above learning setting in the case of infinite dimensional RKHS is infinite both for L_p and L_ϵ, so it cannot be used to study learning machines of the form (4). Instead one needs to consider other measures of complexity, such as the V_γ dimension, in order to prove uniform convergence in infinite dimensional RKHS. We now present some background on the V_γ dimension [1].
In the case of indicator functions the definition of the V_γ dimension is:

Definition 2.1 *The V_γ-dimension of a set of indicator functions $\{\theta(f(\mathbf{x})),\ f \in \mathcal{H}\}$ (where θ is the heavyside function), is the maximum number h of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_h$ that can be separated into two classes in all 2^h possible ways using functions of the set and the rules:*

$$\begin{array}{c} \textit{class 1 if: } f(\mathbf{x}) \geq s + \gamma \\ \textit{class -1 if: } f(\mathbf{x}) \leq s - \gamma \end{array}$$

for some $s \geq 0$. If, for any number N, it is possible to find N points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ that can be separated in all the 2^N possible ways, we will say that the V_γ-dimension of the set is infinite.

For $\gamma = s = 0$ this becomes the definition of the VC dimension of a set of indicator functions $\theta(f(\mathbf{x}))$ [10]. Moreover, in the case of hyperplanes (1), the V_γ dimension has also been referred to in the literature [10] as the VC dimension of hyperplanes with margin. In order to avoid confusion with names, we call the VC dimension of hyperplanes with margin as the V_γ dimension of hyperplanes (for appropriate γ depending on the margin, as discussed below). The VC dimension has been used to bound the growth function $\mathcal{G}^{\mathcal{H}}(l)$. This function measures the maximum number of ways we can separate l points using functions from hypothesis space $\mathcal{H}$. If h is the VC dimension, then $\mathcal{G}^{\mathcal{H}}(l)$ is 2^l if $l \leq h$, and $\leq (\frac{el}{h})^h$ otherwise [10].
In the case of real valued function (which is the case we are interested in here) the definition of the V_γ dimension is:

Definition 2.2 *Let $C \leq L(y, f(\mathbf{x})) \leq B$, $f \in \mathcal{H}$, with C and $B < \infty$. The V_γ-dimension of L in $\mathcal{H}$ (of the set $\{L(y, f(\mathbf{x})),\ f \in \mathcal{H}\}$) is defined as 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:*

$$\begin{array}{c} \textit{class 1 if: } L(y_i, f(x_i)) \geq s + \gamma \\ \textit{class -1 if: } L(y_i, f(x_i)) \leq s - \gamma \end{array}$$

for $f \in \mathcal{H}$ and some $C + \gamma \leq s \leq B - \gamma$. If, for any number N, it is possible to find N points $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_N, y_N)$ that can be separated in all the 2^N possible ways, we will say that the V_γ-dimension of L in $\mathcal{H}$ is infinite.

For $\gamma = 0$ and for s being free to change values for each separation of the data, this becomes the VC dimension of the set of functions [10].
Using the V_γ dimension Alon et al. [1] gave necessary and sufficient conditions for uniform convergence in probability to take place in a hypothesis space $\mathcal{H}$. In particular they proved the following important theorem:

Theorem 2.1 *(Alon et al. , 1997) Let $C \leq L(y, f(\mathbf{x}))) \leq B$, $f \in \mathcal{H}$, $\mathcal{H}$ be a set of bounded functions. The ERM method uniformly converges (in probability) if and only if the V_γ dimension of L in $\mathcal{H}$ is finite for every $\gamma > 0$.*

It is clear that if for learning machines of the form (4) the V_γ dimension of the loss function L in the hypothesis space defined is finite for $\forall \gamma > 0$, then for these machines uniform convergence takes place. In the next section we present a simple proof of the finiteness of the V_γ dimension, as well as an upper bound on it.

2.1 Why not use the VC-dimension

Consider first the case of L_p loss functions. Consider an infinite dimensional RKHS, and the set of functions with norm $\|f\|_K^2 \leq A^2$. If for any N we can find N points that we can shatter using functions of our set according to the rule:

$$\begin{array}{rl} \text{class 1 if :} & |y - f(\mathbf{x})|^p \geq s \\ \text{class } -1 \text{ if :} & |y - f(\mathbf{x})|^p \leq s \end{array}$$

then clearly the VC dimension is infinite. Consider N points $(\mathbf{x}_i, y_i)$ with $y_i = 0$ for all i, and $\mathbf{x}_i$ be such that the smallest eigenvalue of matrix G with $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is $\geq \lambda$. Since we are

in infinite dimensional RKHS, matrix G is always invertible [11], so $\lambda > 0$ since G is positive definite.
For any separation of the points, we consider a function f of the form $f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$, which is a function of the form (1). We need to show that we can find coefficients α_i such that the RKHS norm of the function is $\leq A^2$. Notice that the norm of a function of this form is $\boldsymbol{\alpha}^T G \boldsymbol{\alpha}$ where $(\boldsymbol{\alpha})_i = \alpha_i$ (throughout the paper bold letters are used for noting vectors). Consider the set of linear equations

$$\begin{array}{rl} \mathbf{x}_j \in \text{class } 1: & \sum_{i=1}^{N} \alpha_i G_{ij} = s^{\frac{1}{p}} + \eta \quad \eta > 0 \\ \mathbf{x}_j \in \text{class } -1: & \sum_{i=1}^{N} \alpha_i G_{ij} = s^{\frac{1}{p}} - \eta \quad \eta > 0 \end{array}$$

Let $s = 0$. If we can find a solution $\boldsymbol{\alpha}$ to this system of equations such that $\boldsymbol{\alpha}^T G \boldsymbol{\alpha} \leq A^2$ we can perform this separation, and since this is any separation we can shatter the N points. Notice that the solution to the system of equations is $G^{-1}\boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is the vector whose components are $(\boldsymbol{\eta})_i = \eta$ when $\mathbf{x}_i$ is in class 1, and $-\eta$ otherwise. So we need $(G^{-1}\boldsymbol{\eta})^T G (G^{-1}\boldsymbol{\eta}) \leq A^2 \Rightarrow \boldsymbol{\eta}^T G^{-1} \boldsymbol{\eta} \leq A^2$. Since the smallest eigenvalue of G is $\geq \lambda > 0$, $\boldsymbol{\eta}^T G^{-1} \boldsymbol{\eta} \leq \frac{\boldsymbol{\eta}^T \boldsymbol{\eta}}{\lambda}$. Moreover $\boldsymbol{\eta}^T \boldsymbol{\eta} = N\eta^2$. So if we choose η small enough such that $\frac{N\eta^2}{\lambda} \leq A^2 \Rightarrow \eta^2 \leq \frac{A^2\lambda}{N}$, the norm of the solution is less than A^2, which completes the proof.
For the case of the L_ϵ loss function the argument above can be repeated with $y_i = \epsilon$ to prove again that the VC dimension is infinite in an infinite dimensional RKHS.
Finally, notice that the same proof can be repeated for finite dimensional RKHS to show that the VC dimension is never less than the dimensionality D of the RKHS, since it is possible to find D points for which matrix G is invertible and repeat the proof above. As a consequence the VC dimension cannot be controlled by A^2.

3 A simple upper bound on the V_γ dimension

Below we always assume that data X are within a sphere of radius R in the feature space defined by the kernel K of the RKHS. Without loss of generality, we also assume that y is bounded between -1 and 1. Under these assumptions the following theorem holds:

Theorem 3.1 *The V_γ dimension h for regression using L_p or L_ϵ loss functions for hypothesis spaces $\mathcal{H}_A = \{f(\mathbf{x}) = \sum_{n=1}^{\infty} w_n \phi_n(\mathbf{x}) \mid \sum_{n=1}^{\infty} \frac{w_n^2}{\lambda_n} \leq A^2\}$ and y bounded, is finite for $\forall \gamma > 0$. If D is the dimensionality of the RKHS, then $h \leq O(min(D, \frac{(R^2+1)(A^2+1)}{\gamma^2}))$.*

Proof. Let's consider first the case of the L_1 loss function. Let B be the upper bound on the loss function. From definition 2.2 we can decompose the rules for separating points as follows:

$$\begin{array}{rl} \text{class 1 if} & y_i - f(\mathbf{x}_i) \geq s + \gamma \\ \text{or} & f(\mathbf{x}_i) - y_i \geq s + \gamma \\ \text{class } -1 \text{ if} & y_i - f(\mathbf{x}_i) \leq s - \gamma \\ \text{or} & f(\mathbf{x}_i) - y_i \leq s - \gamma \end{array} \tag{5}$$

for some $\gamma \leq s \leq B - \gamma$. For any N points, the number of separations we can get using rules (5) is not more than the number of separations we can get using the product of two indicator functions (of hyperplanes with margin):

$$\begin{array}{rrl} \text{function (a)}: & \text{class } 1 \text{ if} & y_i - f_1(\mathbf{x}_i) \geq s_1 + \gamma \\ & \text{class } -1 \text{ if} & y_i - f_1(\mathbf{x}_i) \leq s_1 - \gamma \\ \text{function (b)}: & \text{class } 1 \text{ if} & f_2(\mathbf{x}_i) - y_i \geq s_2 + \gamma \\ & \text{class } -1 \text{ if} & f_2(\mathbf{x}_i) - y_i \leq s_2 - \gamma \end{array} \tag{6}$$

where f_1 and f_2 are in $\mathcal{H}_A$, $\gamma \leq s_1, s_2 \leq B - \gamma$. For $s_1 = s_2 = s$ and for $f_1 = f_2 = f$ we recover (5): for example, if $y - f(\mathbf{x}) \geq s + \gamma$ then indicator function (a) will give -1, indicator function (b) will give also -1, so their product will give $+1$ which is what we get if we follow (5). So since we give more freedom to f_1, f_2, s_1, s_2 clearly we can get at least as many separations for any set of points than the number of separations we would get using (5).

As mentioned in the previous section, for any N points the number of separations is bounded by the growth function. Moreover, for products of indicator functions it is known [10] that the growth function is bounded by the product of the growth functions of the indicator functions. Furthermore, the indicator functions in (6) are hyperplanes with margin in the $D+1$ dimensional space of vectors $\{\phi_n(\mathbf{x}), y\}$ where the radius of the data is $R^2 + 1$, the norm of the hyperplane is bounded by $A^2 + 1$, (where in both cases we add 1 because of y), and the margin is at least $\frac{\gamma^2}{A^2+1}$. The V_γ dimension h_γ of these hyperplanes is known [10, ?] to be bounded by $h_\gamma \leq \min((D+1)+1, \frac{(R^2+1)(A^2+1)}{\gamma^2})$. So the growth function of the separating rules (5) is bounded by $\mathcal{G}(l) \leq (\frac{el}{h_\gamma})^{h_\gamma} (\frac{el}{h_\gamma})^{h_\gamma}$ whenever $l \geq h_\gamma$. If h_γ^{reg} is the V_γ dimension, then h_γ^{reg} cannot be larger than the larger number l for which inequality $2^l \leq (\frac{el}{h_\gamma})^{h_\gamma} (\frac{el}{h_\gamma})^{h_\gamma}$ holds. From this we get that $l \leq 5 h_\gamma$, therefore $h_\gamma^{reg} \leq 5 \min\,(D+2, \frac{(R^2+1)(A^2+1)}{\gamma^2})$ which proves the theorem for the case of L_1 loss functions.

For general L_p loss functions we can follow the same proof where (5) now needs to be rewritten as:

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

Moreover, for $p > 1$, $(s+\gamma)^{\frac{1}{p}} \geq s^{\frac{1}{p}} + \frac{\gamma}{pB}$ (since $\gamma = \left((s+\gamma)^{\frac{1}{p}}\right)^p - \left(s^{\frac{1}{p}}\right)^p \leq ((s+\gamma)^{\frac{1}{p}} - s^{\frac{1}{p}})(pB)$) and $(s-\gamma)^{\frac{1}{p}} \leq s^{\frac{1}{p}} - \frac{\gamma}{pB}$ (similarly). Repeating the same argument as above, we get that the V_γ dimension is bounded by $5 \min\,(D+2, \frac{(pB)^2(R^2+1)(A^2+1)}{\gamma^2})$. Finally, for the L_ϵ loss function (5) can be rewritten as:

$$\begin{array}{rl} \text{class } 1 \text{ if} & y_i - f(\mathbf{x}_i) \geq s + \gamma + \epsilon \\ \text{or} & f(\mathbf{x}_i) - y_i \geq s + \gamma + \epsilon \\ \text{class } -1 \text{ if} & y_i - f(\mathbf{x}_i) \leq s - \gamma + \epsilon \\ \text{or} & f(\mathbf{x}_i) - y_i \leq s - \gamma + \epsilon \end{array} \tag{8}$$

where calling $s' = s + \epsilon$ we can simply repeat the proof above and get the same upper bound on the V_γ dimension as in the case of the L_1 loss function. (Notice that the constraint $\gamma \leq s \leq B - \gamma$ is not taken into account. We believe that taking this into account may slightly change the V_γ dimension for L_ϵ). $\Box$

Notice that these results imply that in the case of infinite dimensional RKHS the V_γ dimension is still finite and is influenced only by $5\frac{(R^2+1)(A^2+1)}{\gamma^2}$. In the next section we present a more refined upped bound of the V_γ dimension in the infinite dimensional case.

4 A more refined computation of the V_γ dimension

Below we assume that the data $\mathbf{x}$ are bounded so that for any $N \times N$ matrix G with entries $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ (where K is, as mentioned in the previous section, the kernel of the RKHS considered) the largest eigenvalue of G is always $\leq R^2$. Notice that R^2 is related to the radius of the data $\mathbf{x}$ in the "feature space" $\{\phi_n(\mathbf{x})\}$ typically considered in the literature [10, **?**, 6]. We also note with B the upper bound of $L(y, f(\mathbf{x}))$.

Theorem 4.1 *The V_γ dimension for regression using L_1 loss function and for hypothesis space $\mathcal{H}_A = \{f(\mathbf{x}) = \sum_{n=1}^{\infty} w_n \phi_n(\mathbf{x}) + b \mid \sum_{n=1}^{\infty} \frac{w_n^2}{\lambda_n} \leq A^2\}$ is finite for $\forall \gamma > 0$. In particular:*

1. *If b is constrained to be zero, then $V_\gamma \leq \left[\frac{R^2 A^2}{\gamma^2}\right]$*
2. *If b is a free parameter, $V_\gamma \leq 4\left[\frac{R^2 A^2}{\gamma^2}\right]$*

Proof of part 1.
Suppose we can find $N > \left[\frac{R^2 A^2}{\gamma^2}\right]$ points $\{(x_1, y_1), ..., (x_N, y_N)\}$ that we can shatter. Let $s \in [\gamma, B - \gamma]$ be the value of the parameter used to shatter the points.
Consider the following "separation"[1]: if $|y_i| < s$, then (x_i, y_i) belongs in class 1. All other points belong in class -1. For this separation we need:

$$\begin{array}{ll} |y_i - f(x_i)| \geq s + \gamma, & \text{if } |y_i| < s \\ |y_i - f(x_i)| \leq s - \gamma, & \text{if } |y_i| \geq s \end{array} \tag{9}$$

This means that: for points in class 1 f takes values either $y_i + s + \gamma + \delta_i$ or $y_i - s - \gamma - \delta_i$, for $\delta_i \geq 0$. For points in the second class f takes values either $y_i + s - \gamma - \delta_i$ or $y_i - s + \gamma + \delta_i$, for $\delta_i \in [0, (s - \gamma)]$. So (9) can be seen as a system of linear equations:

$$\sum_{n=1}^{\infty} w_n \phi_n(\mathbf{x}_i) = t_i. \tag{10}$$

with t_i being $y_i + s + \gamma + \delta_i$, or $y_i - s - \gamma - \delta_i$, or $y_i + s - \gamma - \delta_i$, or $y_i - s + \gamma + \delta_i$, depending on i. We first use lemma 1 to show that for any solution (so t_i are fixed now) there is another solution with not larger norm that is of the form $\sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$.

Lemma 4.1 *Among all the solutions of a system of equations (10) the solution with the minimum RKHS norm is of the form: $\sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$ with $\boldsymbol{\alpha} = G^{-1}\boldsymbol{t}$.*

For a proof see the Appendix. Given this lemma, we consider only functions of the form $\sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$. We show that the function of this form that solves the system of equations (10) has norm larger than A^2. Therefore any other solution has norm larger than A^2 which implies we cannot shatter N points using functions of our hypothesis space.
The solution $\boldsymbol{\alpha} = G^{-1}\boldsymbol{t}$ needs to satisfy the constraint:

$$\boldsymbol{\alpha}^T G \boldsymbol{\alpha} = \boldsymbol{t}^T G^{-1} \boldsymbol{t} \leq A^2$$

[1] Notice that this separation might be a "trivial" one in the sense that we may want all the points to be +1 or all to be -1 i.e. when all $|y_i| < s$ or when all $|y_i| \geq s$ respectively.

Let λ_{max} be the largest eigenvalue of matrix G. Then $\boldsymbol{t}^T G^{-1}\boldsymbol{t} \geq \frac{\boldsymbol{t}^T\boldsymbol{t}}{\lambda_{max}}$. Since $\lambda_{max} \leq R^2$, $\boldsymbol{t}^T G^{-1}\boldsymbol{t} \geq \frac{\boldsymbol{t}^T\boldsymbol{t}}{R^2}$. Moreover, because of the choice of the separation, $\boldsymbol{t}^T\boldsymbol{t} \geq N\gamma^2$ (for example, for the points in class 1 which contribute to $\boldsymbol{t}^T\boldsymbol{t}$ an amount equal to $(y_i+s+\gamma+\delta_i)^2$: $|y_i| < s \Rightarrow y_i+s > 0$, and since $\gamma+\delta_i \geq \gamma > 0$, then $(y_i+s+\gamma+\delta_i)^2 \geq \gamma^2$. Similarly each of the other points "contribute" to $\boldsymbol{t}^T\boldsymbol{t}$ at least γ^2, so $\boldsymbol{t}^T\boldsymbol{t} \geq N\gamma^2$). So:

$$\boldsymbol{t}^T G^{-1}\boldsymbol{t} \geq \frac{N\gamma^2}{R^2} > A^2$$

since we assumed that $N > \frac{R^2A^2}{\gamma^2}$. This is a contradiction, so we conclude that we cannot get this particular separation. $\square$

Proof of part 2.
Consider N points that can be shattered. This means that for any separation, for points in the first class there are $\delta_i \geq 0$ such that $|f(x_i) + b - y_i| = s + \gamma + \delta_i$. For points in the second class there are $\delta_i \in [0, s-\gamma]$ such that $|f(x_i)+b-y_i| = s-\gamma-\delta_i$. As in the case $b = 0$ we can remove the absolute values by considering for each class two types of points (we call them type 1 and type 2). For class 1, type 1 are points for which $f(x_i) = y_i + s + \gamma + \delta_i - b = t_i - b$. Type 2 are points for which $f(x_i) = y_i - s - \gamma - \delta_i - b = t_i - b$. For class 2, type 1 are points for which $f(x_i) = y_i + s - \gamma - \delta_i - b = t_i - b$. Type 2 are points for which $f(x_i) = y_i - s + \gamma + \delta_i - b = t_i - b$. Variables t_i are as in the case $b = 0$. Let $S_{11}, S_{12}, S_{-11}, S_{-12}$ denote the four sets of points (S_{ij} are points of class i type j). Using lemma 1, we only need to consider functions of the form $f(x) = \sum_{i=1}^N \alpha_i K(x_i, x)$. The coefficients α_i are given by $\boldsymbol{\alpha} = G^{-1}(\boldsymbol{t} - \boldsymbol{b})$ there $\boldsymbol{b}$ is a vector of b's. As in the case $b = 0$, the RKHS norm of this function is at least

$$\frac{1}{R^2}(\boldsymbol{t} - \boldsymbol{b})^T(\boldsymbol{t} - \boldsymbol{b}). \tag{11}$$

The b that minimizes (11) is $\frac{1}{N}(\sum_{i=1}^N t_i)$. So (11) is at least as large as (after replacing b and doing some simple calculations) $\frac{1}{2NR^2}\sum_{i,j=1}^N (t_i - t_j)^2$.
We now consider a particular separation. Without loss of generality assume that $y_1 \leq y_2 \leq \ldots \leq y_N$ and that N is even (if odd, consider $N-1$ points). Consider the separation where class 1 consists only of the "even" points $\{N, N-2, \ldots, 2\}$. The following lemma is shown in the appendix:

Lemma 4.2 *For the separation considered,* $\sum_{i,j=1}^N (t_i - t_j)^2$ *is at least as large as* $\frac{\gamma^2(N^2-4)}{2}$.

Using Lemma 4.2 we get that the norm of the solution for the considered separation is at least as large as $\frac{\gamma^2(N^2-4)}{4NR^2}$. Since this has to be $\leq A^2$ we get that $N - \frac{4}{N} \leq 4\left[\frac{R^2A^2}{\gamma^2}\right]$, which completes the proof (assume $N > 4$ and ignore additive constants less than 1 for simplicity of notation). $\square$

In the case of L_p loss functions, using the same argument as in the previous section we get that the V_γ dimension in infinite dimensional RKHS is bounded by $\frac{(pB)^2R^2A^2}{\gamma^2}$ in the first case of theorem 4.1, and by $4\frac{(pB)^2R^2A^2}{\gamma^2}$ in the second case of theorem 4.1. Finally for L_ϵ loss functions the bound on the V_γ dimension is the same as that for L_1 loss function, again using the argument of the previous section.

5 Conclusion

We presented a novel approach for computing the V_γ dimension of RKHS for L_p and L_ϵ loss functions. We conclude with a few remarks. First notice that in the computations we did not take into account ϵ in the case of L_ϵ loss function. Taking ϵ into account may lead to better bounds. For example, considering $|f(x)-y|_\epsilon^p, p>1$ as the loss function, it is clear from the proofs presented that the V_γ dimension is bounded by $\frac{p^2(B-\epsilon)^2R^2A^2}{\gamma^2}$. However the influence of ϵ seems to be minor (given that $\epsilon << B$). Furthermore, it may be possible to extend the computations for more general loss functions.
An interesting observations is that the eigenvalues of the matrix G appear in the computation of the V_γ dimension. In the proofs we took into account only the largest and smallest eigenvalues. If similar computations are made to compute the number of separations for a given set of points, then it is possible that all the eigenvalues of G are taken into account. This leads to interesting relations with the work in [12]. Finally, the bounds on the V_γ dimension can be used to get bounds on the generalization performance of regression machines of the form (4) [1, 4].

Acknowledgments
We would like to thank S. Mukherjee, T. Poggio, R. Rifkin, and A. Verri for useful discussions and comments.

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] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Number 31 in Applications of mathematics. Springer, New York, 1996.

[4] 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.

[5] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

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

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

[8] M.J.D. Powell. The theory of radial basis functions approximation in 1990. In W.A. Light, editor, *Advances in Numerical Analysis Volume II: Wavelets, Subdivision Algorithms and Radial Basis Functions*, pages 105–210. Oxford University Press, 1992.

[9] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems.* W. H. Winston, Washington, D.C., 1977.

[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.

Appendix

Proof of Lemma 4.1

We introduce the $N \times \infty$ matrix $A_{in} = \sqrt{\lambda_n}\phi_n(\mathbf{x}_i)$ and the new variable $z_n = \frac{w_n}{\sqrt{\lambda_n}}$. We can write system (10) as follows:

$$A\mathbf{z} = \mathbf{t}. \tag{12}$$

Notice that the solution of the system of equation 10 with minimum RKHS norm, is equivalent to the Least Square (LS) solution of equation 12. Let us denote with $\mathbf{z}^0$ the LS solution of system 12. We have:

$$\mathbf{z}^0 = (A^\top A)^+ A^\top \mathbf{t} \tag{13}$$

where $+$ denotes pseudoinverse. To see how this solution looks like we use Singular Value Decomposition techniques:

$$\begin{aligned} A &= U\Sigma V^\top, \\ A^\top &= V\Sigma U^\top, \end{aligned}$$

from which $A^\top A = V\Sigma^2 V^\top$ and $(A^\top A)^+ = V_N \Sigma_N^{-2} V_N^\top$, where Σ_N^{-1} denotes the $N \times N$ matrix whose elements are the inverse of the nonzero eigenvalues. After some computations equation (13) can be written as:

$$\mathbf{z}^0 = V\Sigma_N^{-1} U_N^\top \mathbf{t} = (V\Sigma_N U_N^\top)(U_N \Sigma_N^{-2} U_N^\top)\mathbf{t} = AG^{-1}\mathbf{t}. \tag{14}$$

Using the definition of $\mathbf{z}^0$ we have that

$$\sum_{n=1}^{\infty} w_n^0 \phi_n(\mathbf{x}) = \sum_{n=1}^{\infty}\sum_{i=1}^{N} \sqrt{\lambda_n}\phi_n(\mathbf{x}) A_{ni}\alpha_i. \tag{15}$$

Finally, using the definition of A_{in} we get:

$$\sum_{n=1}^{\infty} w_n^0 \phi_n(\mathbf{x}) = \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x}_i)\alpha_i$$

which completes the proof. $\square$

Proof of Lemma 4.2

Consider a point (x_i, y_i) in S_{11} and a point (x_j, y_j) in S_{-11} such that $y_i \geq y_j$ (if such a pair does not exist we can consider another pair from the cases listed below). For these points $(t_i - t_j)^2 = (y_i + s + \gamma + \delta_i - y_j - s + \gamma + \delta_j)^2 = ((y_i - y_j) + 2\gamma + \delta_i + \delta_j)^2 \geq 4\gamma^2$. In a similar way (taking into account the constraints on the δ_i's and on s) the inequality $(t_i - t_j)^2 \geq 4\gamma^2$ can be shown to hold in the following two cases:

$$\begin{array}{lll} (x_i, y_i) \in S_{11}, & (x_j, y_j) \in S_{-11} \bigcup S_{-12}, & y_i \geq y_j \\ (x_i, y_i) \in S_{12}, & (x_j, y_j) \in S_{-11} \bigcup S_{-12}, & y_i \leq y_j \end{array} \tag{16}$$

Moreover, clearly

$$\sum_{i,j=1}^{N} (t_i - t_j)^2 \geq 2 \left[\sum_{i \in S_{11}} \left(\sum_{j \in S_{-11} \bigcup S_{-12}, y_i \geq y_j} (t_i - t_j)^2 \right) + \sum_{i \in S_{12}} \left(\sum_{j \in S_{-11} \bigcup S_{-12}, y_i \leq y_j} (t_i - t_j)^2 \right) \right].$$

Using the fact that for the cases considered $(t_i - t_j)^2 \geq 4\gamma^2$, the right side is at least

$$\begin{array}{l} 8\gamma^2 \sum_{i \in S_{11}} (\text{number of points j in class } -1 \text{ with } y_i \geq y_j) + \\ +8\gamma^2 \sum_{i \in S_{12}} (\text{number of points j in class } -1 \text{ with } y_i \leq y_j) \end{array} \tag{17}$$

Let I_1 and I_2 be the cardinalities of S_{11} and S_{12} respectively. Because of the choice of the separation it is clear that (17) is at least

$$8\gamma^2 \left((1 + 2 + \ldots + I_1) \right) + (1 + 2 + \ldots + (I_2 - 1)))$$

(for example if $I_1 = 2$ in the worst case points 2 and 4 are in S_{11} in which case the first part of (17) is exactly 1+2). Finally, since $I_1 + I_2 = \frac{N}{2}$, (17) is at least $8\gamma^2 \frac{N^2-4}{16} = \frac{\gamma^2(N^2-4)}{2}$, which proves the lemma. □