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

A.I. Memo No. 1606 May, 1997
C.B.C.L Paper No. 147

An Equivalence Between Sparse Approximation and Support Vector Machines

Federico Girosi

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1606.ps.Z

Abstract

This paper shows a relationship between two different approximation techniques: the Support Vector Machines (SVM), proposed by V. Vapnik (1995), and a sparse approximation scheme that resembles the Basis Pursuit De-Noising algorithm (Chen, 1995; Chen, Donoho and Saunders, 1995). SVM is a technique which can be derived from the Structural Risk Minimization Principle (Vapnik, 1982) and can be used to estimate the parameters of several different approximation schemes, including Radial Basis Functions, algebraic/trigonometric polynomials, B-splines, and some forms of Multilayer Perceptrons. Basis Pursuit De-Noising is a sparse approximation technique, in which a function is reconstructed by using a small number of basis functions chosen from a large set (the *dictionary*). We show that, if the data are noiseless, the modified version of Basis Pursuit De-Noising proposed in this paper is equivalent to SVM in the following sense: if applied to the same data set the two techniques give the same solution, which is obtained by solving the same quadratic programming problem. In the appendix we also present a derivation of the SVM technique in the framework of regularization theory, rather than statistical learning theory, establishing a connection between SVM, sparse approximation and regularization theory.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a ONR/ARPA grant under contract N00014-92-J-1879 and by MURI grant N00014-95-1-0600. Additional support is provided by Eastman Kodak Company, Daimler-Benz, Siemens Corporate Research, Inc. and AT&T.

1 Introduction

In recent years there has been an increasing interest in approximation techniques that use the concept of *sparsity* to perform some form of model selection. By sparsity we mean, in very general terms, a constraint that enforces the number of building blocks of the model to be small. Sparse approximation often appears in conjunction with the use of overcomplete or redundant representations, in which a signal is approximated as a linear superposition of basis functions taken from a large *dictionary* (Chen, 1995; Chen, Donoho and Saunders, 1995; Olshausen and Field, 1996; Daubechies, 1992; Mallat and Zhang, 1993; Coifman and Wickerhauser, 1992). In this case sparsity is used as a criterion to choose between different approximating functions with the same reconstruction error, favoring the one with the least number of coefficients. The concept of sparsity has also been used in linear regression, as an alternative to subset selection, in order to produce linear models that use a small number of variables and therefore have greater interpretability (Tibshirani, 1994; Breiman, 1993).

In this paper we discuss the relationship between an approximation technique based on the principle of sparsity and the Support Vector Machines (SVM) technique recently proposed by Vapnik (Vapnik, 1995; Vapnik, Golowich and Smola, 1996). SVM is a classification/approximation technique derived by V. Vapnik in the framework of Structural Risk Minimization, which aims at building "parsimonious" models, in the sense of VC-dimension. Sparse approximation technique are also "parsimonious", in the sense that they try to minimize the number of parameters of the model, so it is not surprising that some connections between SVM and sparse approximation exist. What is more surprising and less obvious is that SVM and a specific model of sparse approximation, which is a modified version of the Basis Pursuit De-Noising algorithm (Chen, 1995; Chen, Donoho and Saunders, 1995), are actually equivalent, in the case of noiseless data. By equivalent we mean the following: if applied to the same data set they give the same solution, which is obtained by solving the same quadratic programming problem. While the equivalence between sparse approximation and SVM for noiseless data is the main point of the paper, we also include a derivation of the SVM which is different from the one given by V. Vapnik, and that fits very well in the framework of regularization theory, the same one which is used to derive techniques like splines or Radial Basis Functions.

The plan of the paper is as follows: in section 2 we introduce the technique of SVM in the framework of regularization theory (the mathematical details can be found in appendix B). Section 3 introduces the notion of sparsity and presents an exact and approximate formulation of the problem. In section 4 we present a sparse approximation model, which is similar in spirit to the Basis Pursuit De-Noising technique of Chen, Donoho and Saunders (1995), and show that, in the case of noiseless data, it is equivalent to SVM. Section 5 concludes the paper and contains a series of remarks and observations. Appendix A contains some background material on Reproducing Kernel Hilbert Spaces, which are heavily used in this paper. Appendix B contains an explicit derivation of the SVM technique in the framework of regularization theory, and appendix C addresses the case in which data are noisy.

2 From Regularization Theory to Support Vector Machines

In this section we briefly sketch the ideas behind the Support Vector Machines (SVM) for regression, and refer the reader to (Vapnik, 1995) and (Vapnik, Golowich and Smola, 1996) for a full description of the technique. The reader should be warned that the way the theory is presented here is slightly different from the way it is derived in Vapnik's work. In this paper we will take a viewpoint which is closer to classical regularization theory (Tikhonov and Arsenin, 1977; Morozov, 1984; Bertero, 1986; Wahba, 1975, 1979, 1990), which might be more familiar to the reader, rather than the theory of uniform convergence in probability developed by Vapnik (Vapnik, 1982; Vapnik, 1995). A similar approach is described in (Smola and Schölkopf, 1998), although with a different formalism. In this section and in the following ones we will need some basic notions about Reproducing Kernel Hilbert Spaces (RKHS). For simplicity of exposition we put all the technical material about RKHS in appendix (A). Since the RKHS theory is very well developed we do not include many important mathematical technicalities (like the convergence of certain series, or the issue of semi-RKHS), because the goal here is just to provide the reader with a basic understanding of an already existing technique. The rigorous mathematical apparatus that we use can be mostly found in chapter 1 of the book of G. Wahba (1990) .

2.1 Support Vector Machines

The problem we want to solve is the following: we are given a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, obtained by sampling, with noise, some unknown function $f(\mathbf{x})$ and we are asked to recover the function f, or an approximation of it, from the data D. We assume that the function f underlying the data can be represented as:

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

where $\{\phi_n(\mathbf{x})\}_{n=1}^{\infty}$ is a set of given, linearly independent basis functions, and c_n and b are parameters to be estimated from the data. Notice that if one of the basis functions ϕ_n is constant then the term b is not necessary. The problem of recovering the coefficients c_n and b from the data set D is clearly ill-posed, since it has an infinite number of solutions. In order to make this problem well-posed we follow the approach of regularization theory (Tikhonov and Arsenin, 1977; Morozov, 1984; Bertero, 1986; Wahba, 1975, 1990) and impose an additional smoothness constraint on the solution of the approximation problem. Therefore we choose as a solution the function that solves the following variational problem:

$$\min_{f \in \mathcal{H}} H[f] = C \sum_{i=1}^{l} V(y_i - f(\mathbf{x}_i)) + \frac{1}{2}\Phi[f] \tag{2}$$

where $V(x)$ is some error cost function that is used to measure the interpolation error (for example $V(x) = x^2$), C is a positive number, $\Phi[f]$ is a smoothness functional and $\mathcal{H}$ is the set of functions over which the smoothness functional $\Phi[f]$ is well defined. The first term is enforcing closeness to the data, and the second smoothness, while C controls the tradeoff between these two terms. A large class of smoothness functionals, defined over elements of the form (1), can be defined as follows:

$$\Phi[f] = \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n} \tag{3}$$

where $\{\lambda_n\}_{n=1}^{\infty}$ is a decreasing, positive sequence.
That eq. (3) actually defines a smoothness functional can be seen in the following:

Example: Let us consider a one-dimensional case in which $x \in [0, 2\pi]$, and let us choose $\phi_n(x) = e^{inx}$, so that the c_n are the Fourier coefficients of the function f. Since the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is decreasing, the constraint that $\Phi[f] < \infty$ is a constraint on the rate of convergence to zero of the Fourier coefficients c_n, which is well known to control the differentiability properties of f. Functions for which $\Phi[f]$ is small have limited high frequency content, and therefore do not oscillate much, so that $\Phi[f]$ is a measure of smoothness. More examples can be found in appendix A.

When the smoothness functional has the form (3) it is easy to prove (appendix B) that, independently on the form of the error function V, the solution of the variational problem (2) has always the form:

$$f(\mathbf{x}) = \sum_{i=1}^{l} a_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{4}$$

where we have defined the (symmetric) *kernel function* K as:

$$K(\mathbf{x}; \mathbf{y}) = \sum_{n=1}^{\infty} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{y}) \tag{5}$$

The kernel K can be seen as the kernel of a Reproducing Kernel Hilbert Space (RKHS), a concept that will be used in section (4). Details about RKHS and examples of kernels can be found in appendix A and in (Girosi, 1997).
If the cost function V is quadratic the unknown coefficients in (4) can be found by solving a linear system. When the kernel K is a radially symmetric function eq. (4) describe a Radial Basis Functions approximation scheme, which is closely related to smoothing splines, and when K is of the form $K(\mathbf{x} - \mathbf{y})$ eq. (4) is a Regularization Network (Girosi, Jones and Poggio, 1995). When the cost function V is not quadratic anymore the solution of the variational problem (2) has still the form (4) (Smola and Schölkopf, 1998; Girosi, Poggio and Caprile, 1991), but the coefficients a_i cannot be found anymore by solving a linear system. V. Vapnik (1995) proposed to use a particularly interesting form for the function V, which he calls the *ϵ-insensitive cost function*, which we plot in figure (1):

$$V(x) = |x|_\epsilon \equiv \begin{cases} 0 & \text{if } |x| < \epsilon \\ |x| - \epsilon & \text{otherwise.} \end{cases} \tag{6}$$

The ϵ-insensitive cost function is similar to some of the functions used in robust statistics (Huber, 1981), which are known to provide robustness against outliers. However the function (6) is not only a robust cost function, but also assigns zero cost to errors which are smaller then ϵ. In other words, according to the cost function $|x|_\epsilon$ any function that comes closer than ϵ to the data points is a perfect interpolant. In a sense, the parameter ϵ represents, therefore, the resolution at which we want to look at the data. When the ϵ-insensitive cost function is used in conjunction with the variational approach of (2), one obtains the approximation scheme known as SVM, which has the form

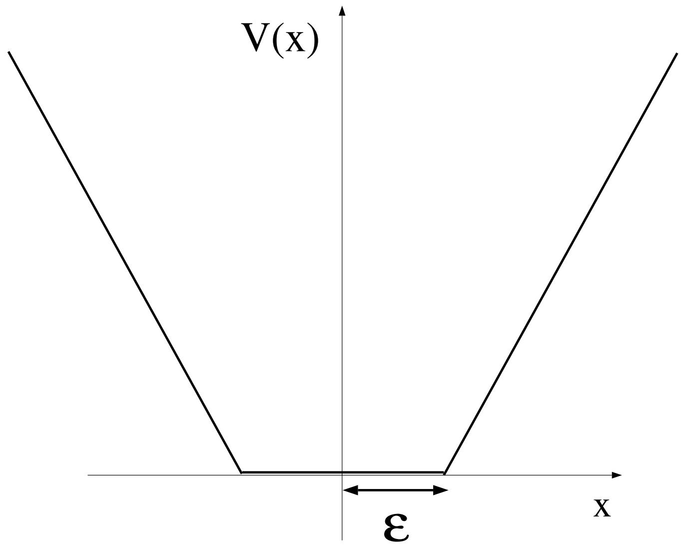

Figure 1: Vapnik's ϵ-insensitive cost function $V(x) = |x|_\epsilon$.

$$f(\mathbf{x}, \alpha, \alpha^*) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) K(\mathbf{x}; \mathbf{x}_i) + b, \tag{7}$$

where α_i^* and α_i are some positive coefficients which solve the following Quadratic Programming (QP) problem:

$$\min_{\alpha, \alpha^*} R(\alpha^*, \alpha) = \epsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i (\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\mathbf{x}_i; \mathbf{x}_j), \tag{8}$$

subject to the constraints

$$\begin{array}{ll} 0 \leq \boldsymbol{\alpha}^*, \boldsymbol{\alpha} \leq C & \\ \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 & \\ \alpha_i \alpha_i^* = 0 & \forall i = 1, \ldots, l \end{array} \tag{9}$$

Notice that the parameter b does not appear in the QP problem, and we show in appendix (B) that it is determined from the knowledge of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$. It is important to notice that it is possible to prove that the last of the constraints above ($\alpha_i \alpha_i^* = 0$) is *automatically* satisfied by the solution and it could be dropped from the formulation. We include this constraint just because it will be useful in section 4.
Due to the nature of this quadratic programming problem, only a number of coefficients $\alpha_i^* - \alpha_i$ will be different from zero, and the input data points $\mathbf{x}_i$ associated to them are called *support vectors*. The number of support vectors depends on both C and ϵ. The parameter C weighs the data term in functional (2) with respect to the smoothness term, and in regularization theory is known to be related to the amount of the noise in the data. If there is no noise in the data the optimal value for C is infinity, which forces the data term to be zero. In this case SVM will find, among all the functions which have interpolation errors smaller than ϵ, the one that minimizes the smoothness functional $\Phi[f]$. The parameters C and ϵ are two free parameters of the theory, and their choice is left to the user, as well as the choice of the kernel K, which determines the smoothness properties of the solution and should reflect prior knowledge on the data. For certain choices of K some well known approximation schemes are recovered, as shown in table (1). We refer the reader to the book of Vapnik (1995) for more details about SVM, and for the original derivation of the technique.

Kernel Function	Approximation Scheme
$K(\mathbf{x}; \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$	Gaussian RBF
$K(\mathbf{x}; \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d$	Polynomial of degree d
$K(\mathbf{x}; \mathbf{y}) = \tanh(\mathbf{x} \cdot \mathbf{y} - \theta)$	(only for some values of θ) Multi Layer Perceptron
$K(x; y) = B_{2n}(x - y)$	B-splines
$K(x; y) = \frac{\sin(d+1/2)(x-y)}{\sin\frac{(x-y)}{2}}$	Trigonometric polynomial of degree d

Table 1: Some possible kernel functions and the type of decision surface they define. The last two kernels are one-dimensional: multidimensional kernels can be built by tensor products of one-dimensional ones. The functions B_n are piecewise polynomials of degree n, whose exact definition can be found in (Schumaker, 1981)

3 Sparse Approximation

In recent years there has been a growing interest in approximating functions using linear superpositions of basis functions selected from a large, redundant set of basis functions, called *dictionary*. It is not the purpose of this paper to discuss the motivations that lead to this approach, and refer the reader to (Chen, 1995; Chen, Donoho and Saunders, 1995; Olshausen and Field, 1996; Harpur and Prager, 1996; Daubechies, 1992; Mallat and Zhang, 1993; Coifman and Wickerhauser, 1992) for further details. A common aspects of these technique is that one seeks an approximating function of the form:

$$f(\mathbf{x};\mathbf{a}) = \sum_{i=1}^{n} a_i \varphi_i(\mathbf{x}) \tag{10}$$

where $\varphi \equiv \{\varphi_i(\mathbf{x})\}_{i=1}^n$ is a fixed set of basis functions that we will call *dictionary*. If n is very large (possibly infinite) and φ is not an orthonormal basis (for example it could be a frame, or just a redundant, finite set of basis functions) it is possible that many different sets of coefficients will achieve the same error on a given data set. A *sparse* approximation scheme looks, among all the approximating functions that achieve the same error, for the one with the smallest number of non-zero coefficients. The sparsity of an approximation scheme can also be invoked whenever the number of basis functions initially available is considered, for whatever reasons, too large (this situation arises often in Radial Basis Functions applied to a very large data set).
More formally we say that an approximating function of the form (10) is sparse if the coefficients have been chosen so that they minimize the following cost function:

$$E[\mathbf{a},\boldsymbol{\xi}] = \|f(\mathbf{x}) - \sum_{i=1}^{n} \xi_i a_i \varphi_i(\mathbf{x})\|_{L_2}^2 + \lambda(\sum_{i=1}^{n} \xi_i)^p \tag{11}$$

where $\{\xi_i\}_{i=1}^n$ is a set of binary variables, with values in $\{0,1\}$,$\|\cdot\|_{L_2}$ is the usual L_2 norm, and p is a positive number that we set to one unless otherwise stated. It is clear that, since the L_0 norm of a vector counts the number of elements of that vector which are different from zero, the cost function above can be replaced by the cost function:

$$E[\mathbf{a}] = \|f(\mathbf{x}) - \sum_{i=1}^{n} a_i \varphi_i(\mathbf{x})\|_{L_2}^2 + \lambda\|\mathbf{a}\|_{L_0}^p \tag{12}$$

The problem of minimizing such a cost function, however, is extremely difficult because it involves a combinatorial aspect, and it will be impossible to solve in practical cases. In order to circumvent this problem, approximated versions of the cost function above have been proposed. For example, in (Chen, 1995; Chen, Donoho and Saunders, 1995) the authors use the L_1 norm as an approximation of the L_0, obtaining an approximation scheme that they call *Basis Pursuit De-Noising*. In related work, Olshausen and Field (1996) enforce sparsity by considering the following cost function:

$$E[\mathbf{a}] = \|f(\mathbf{x}) - \sum_{i=1}^{n} a_i \varphi_i(\mathbf{x})\|_{L_2}^2 + \epsilon \sum_{j=1}^{n} S(a_j) \tag{13}$$

where the function S was chosen in such a way to approximately penalize the number of non-zero coefficients. Examples of some the choices considered by Olshausen and Field (1996) are reported in table (2).

$S(x)$
$\|x\|$
$-\exp(-x^2)$
$\log(1+x^2)$

Table 2: Some choices for the penalty function S in eq. (13) considered by Olshausen and Field (1996).

In the case in which $S(x) = |x|$, that is the Basis Pursuit De-Noising case, it is simple to see how the cost function (13) is an approximated version of the one in (11). In order to see this, let us allow the variables ξ_i to assume values in $\{-1, 0, 1\}$, so that the cost function (11) can be rewritten as

$$E[\mathbf{a}, \boldsymbol{\xi}] = \|f(\mathbf{x}) - \sum_{i=1}^{n} \xi_i a_i \varphi_i(\mathbf{x})\|_{L_2}^2 + \lambda \sum_{i=1}^{n} |\xi_i| \ . \tag{14}$$

If we now let the variables ξ_i assume values over the all real line, and *assuming that the coefficients a_i are bounded*, it is clear that the coefficients a_i are redundant, and can be dropped from the cost function. Renaming the variables ξ_i as a_i, we then have that the approximated cost function is

$$E[\mathbf{a}] = \|f(\mathbf{x}) - \sum_{i=1}^{n} a_i \varphi_i(\mathbf{x})\|_{L_2}^2 + \lambda \|\mathbf{a}\|_{L_1} \ , \tag{15}$$

which is the one proposed in the Basis Pursuit De-Noising method of Chen, Donoho and Saunders (1995).

4 An Equivalence Between Support Vector Machines and Sparse Coding

The approximation scheme proposed by Chen, Donoho and Saunders, (1995) has the form described by eq. (10), where the coefficients are found by minimizing the cost function (15). We now make the following choice for the basis functions φ_i:

$$\varphi_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i) \quad \forall i = 1, \ldots, l$$

where $K(\mathbf{x}; \mathbf{y})$ is the reproducing kernel of a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ (see appendix A) and $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ is a data set which has been obtained by sampling, *in absence of noise*, the target function f. We make the explicit assumption that the target function f belongs to the RKHS $\mathcal{H}$. The reader unfamiliar with RKHS can think of $\mathcal{H}$ as a space of smooth functions, for example functions which are square integrable and whose derivatives up to a certain order are also square integrable. The norm $\|f\|_{\mathcal{H}}^2$ in this Hilbert space can be thought as a linear combination of the L_2 norm of the function and the L_2 norm of its derivatives (the specific degree of smoothness and the linear combination depends on the specific kernel K). It follows from eq. (10) that our approximating function is:

$$f^*(\mathbf{x}) \equiv f(\mathbf{x};\mathbf{a}) = \sum_{i=1}^{l} a_i K(\mathbf{x};\mathbf{x}_i) \qquad (16)$$

This model is similar to the one of SVM (eq. 7), except for the constant b, and if $K(\mathbf{x};\mathbf{y}) = G(\|\mathbf{x}-\mathbf{y}\|)$, where G is a positive definite function, it corresponds to a classical Radial Basis Functions approximation scheme (Micchelli, 1986; Moody and Darken, 1989; Powell, 1992).
While Chen et al., in their Basis Pursuit De-Noising method, measure the reconstruction error with an L_2 criterion, we measure it by the true distance, in the $\mathcal{H}$ norm, between the target function f and the approximating function f^*. This measure of distance, which is common in approximation theory, is better motivated than the L_2 norm because it not only enforces closeness between the target and the model, but also between their derivatives, since $\|\cdot\|_{\mathcal{H}}$ is a measure of smoothness. We therefore look for the set of coefficients $\mathbf{a}$ that minimize the following cost function:

$$E[\mathbf{a}] = \frac{1}{2}\|f(\mathbf{x}) - \sum_{i=1}^{l} a_i K(\mathbf{x};\mathbf{x}_i)\|^2_{\mathcal{H}} + \epsilon\|\mathbf{a}\|_{L_1} \qquad (17)$$

where $\|\cdot\|_{\mathcal{H}}$ is the standard norm in $\mathcal{H}$. We consider this to be a modified version of the Basis Pursuit De-Noising technique of Chen (1995) and Chen, Donoho and Saunders (1995).
Notice that it looks from eq. (17) that the cost function E cannot be computed because it requires the knowledge of f (in the first term). This would be true if we had $\|\cdot\|_{L_2}$ instead of $\|\cdot\|_{\mathcal{H}}$ in eq. (17), and it would force us to consider the approximation:

$$\|f(\mathbf{x}) - f^*(\mathbf{x})\|^2_{L_2} \approx \frac{1}{l}\sum_{i=1}^{l}(y_i - f^*(\mathbf{x}_i))^2 \qquad (18)$$

However, because we used the norm $\|\cdot\|_{\mathcal{H}}$, we will see in the following that (surprisingly) no approximation is required, and the expression (17) can be computed exactly, up to a constant (which is obviously irrelevant for the minimization process).
For simplicity we assume that the target function f has zero mean in $\mathcal{H}$, which means that its projection on the constant function $g(\mathbf{x}) = 1$ is zero:

$$< f, 1 >_{\mathcal{H}} = 0$$

Notice that we are not assuming that the function $g(\mathbf{x}) = 1$ belongs to $\mathcal{H}$, but simply that the functions that we consider, including the reproducing kernel K, have a finite projection on it. In particular we normalize K in such a way that $< 1, K(\mathbf{x};\mathbf{y}) >_{\mathcal{H}} = 1$. We impose one additional constraints on this problem:

- We want to guarantee that the approximating function f^* has also zero mean in $\mathcal{H}$:

$$< f^*, 1 >_{\mathcal{H}} = 0 \qquad (19)$$

Substituting eq. (16) in eq. (19), and using the fact that K has mean equal to 1, we see that this constraint implies that:

$$\sum_{i=1}^{l} a_i = 0 \quad . \qquad (20)$$

We can now expand the cost function E of equation (17) as

$$E[\mathbf{a}] = \frac{1}{2}\|f\|^2_{\mathcal{H}} - \sum_{i=1}^{l} a_i < f(\mathbf{x}), K(\mathbf{x};\mathbf{x}_i) >_{\mathcal{H}} + \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j < K(\mathbf{x};\mathbf{x}_i), K(\mathbf{x};\mathbf{x}_j) >_{\mathcal{H}} + \epsilon \sum_{i=1}^{l} |a_i|$$

Using the reproducing property of the kernel K we have:

$$< f(\mathbf{x}), K(\mathbf{x};\mathbf{x}_i) >_{\mathcal{H}} = f(\mathbf{x}_i) \equiv y_i \tag{21}$$

$$< K(\mathbf{x};\mathbf{x}_i), K(\mathbf{x};\mathbf{x}_j) >_{\mathcal{H}} = K(\mathbf{x}_i;\mathbf{x}_j) \tag{22}$$

Notice that in eq. (21) we explicitly used the assumption that the data are noiseless, so that we know the value y_i of the target function f at the data points $\mathbf{x}_i$. We can now rewrite the cost functions as:

$$E[\mathbf{a}] = \frac{1}{2}\|f\|^2_{\mathcal{H}} - \sum_{i=1}^{l} a_i y_i + \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j K(\mathbf{x}_i;\mathbf{x}_j) + \epsilon \sum_{i=1}^{l} |a_i| \tag{23}$$

We now notice that the L_1 norm of $\mathbf{a}$ (the term with the absolute value in the previous equation), can be rewritten more easily by decomposing the vector $\mathbf{a}$ in its "positive" and "negative" parts as follows:

$$\mathbf{a} = \mathbf{a}^+ - \mathbf{a}^- \quad \mathbf{a}^+, \mathbf{a}^- \geq 0, \quad a_i^+ a_i^- = 0 \ \ \forall i = 1, \ldots, l.$$

Using this decomposition we have

$$\|\mathbf{a}\|_{L_1} = \sum_{i=1}^{l}(a_i^+ + a_i^-) \ . \tag{24}$$

Disregarding the constant term in $\|f\|^2_{\mathcal{H}}$ and taking in account the constraint (20), we conclude that the minimization problem we are trying to solve is equivalent to the following quadratic programming (QP) minimization problem:

Problem 4.1 *Solve:*

$$\min_{a_i^+, a_i^-} \left[-\sum_{i=1}^{l}(a_i^+ - a_i^-) y_i + \frac{1}{2}\sum_{i,j=1}^{l}(a_i^+ - a_i^-)(a_j^+ - a_j^-) K(\mathbf{x}_i;\mathbf{x}_j) + \epsilon \sum_{i=1}^{l}(a_i^+ + a_i^-) \right] \tag{25}$$

subject to the constraints:

$$\begin{array}{ll} \mathbf{a}^+, \mathbf{a}^- & \geq 0 \\ \sum_{i=1}^{l}(a_i^+ - a_i^-) & = 0 \\ a_i^+ a_i^- & = 0 \quad \forall i = 1, \ldots, l \end{array} \tag{26}$$

If we now rename the coefficients as follows:

$$a_i^+ \Rightarrow \alpha_i^*$$
$$a_i^- \Rightarrow \alpha_i$$

we notice that the QP problem defined by equations (25) and (26) is the same QP problem that we need to solve for training a SVM with kernel K (see eq. 8 and 9) in the case in which the data are noiseless. In fact, as we argued in section 2.1, the parameter C of a SVM should be set to infinity when the data are noiseless. Since the QP problem above is the same QP problem of SVM, we can use the fact that the constraint $\alpha_i \alpha_i^* = 0$ is automatically satisfied by the SVM solution (see appendix B) to infer that the constraint $a_i^+ a_i^- = 0$ is also automatically satisfied in the problem above, so that it does not have to be included in the QP problem. Notice also that the constant term b which appears in (7) does not appear in our solution. We argue in appendix B that for most commonly used kernels K this term is not needed, because it is already implicitly included in the model. We can now make the following:

Statement 4.1 *When the data are noiseless, the modified version of Basis Pursuit De-Noising of eq. (17), with the additional constraint (19), gives the same solution of SVM, and the solution is obtained by solving the same QP problem of SVM.*

As expected, the solution of the Basis Pursuit De-Noising is such that only a subset of the data points in eq. (16) has non-zero coefficients, the so-called support vectors. The number of support vectors, that is the degree of sparsity, is controlled by the parameter ϵ, which is the only free parameter of this theory.

5 Conclusions and remarks

In this paper we showed that, in the case of noiseless data, SVM can be derived without using any result from VC theory, but simply enforcing a sparsity constraint in an approximation scheme of the form

$$f(\mathbf{x}; \mathbf{a}) = \sum_{i=1}^{l} a_i K(\mathbf{x}; \mathbf{x}_i)$$

together with the constraint that, assuming that the target function has zero mean, the approximating function should also have zero mean. This makes a connection between a technique such as SVM, which is derived in the framework of Structural Risk Minimization, and Basis Pursuit De-Noising, a technique which has been proposed starting from the principle of sparsity. Some observations are in order:

- This result shows that SVM provide an interesting solution to an old-standing problem: the choice of the centers for Radial Basis Functions. If the number of data points is very large we do not want to place one basis function at every data point, but rather at a (small) number of other locations, called "centers". The choice of the centers is often done by randomly choosing a subset of the data points. SVM provides a subset of the data points (the support vectors) which is "optimal" in the sense of the trade-off between interpolation error and number of basis functions (measured in the L_1 norm). SVM can be therefore seen as a "sparse" Radial Basis Functions in the case in which the kernel is radially symmetric.

- One can regard this result as an additional motivation to consider sparsity as an "interesting" constraint. In fact, we have shown here that, under certain conditions, sparsity leads to SVM, which is related to the Structural Risk Minimization principle, and is extremely well motivated in the theory of uniform convergence in probability.

- The result holds because in both this and Vapnik's formulation the cost function contains both an "L_2-type" and an "L_1-type" norm. However, the Support Vector method has an "L_1-type" norm in the error term, and an L_2 norm in the "regularization" term, while the cost function (17) we consider has an "L_2-type" norm in the error term and an L_1 norm in the "regularization" term.

- This results holds due to the existence of the reproducing property of the RKHS. If the norm $\|\cdot\|_{\mathcal{H}}$ were replaced by the standard L_2 norm or any other Sobolev norm the cost function would contain the scalar product in L_2 between the unknown function f and the kernel $K(\mathbf{x};\mathbf{x}_i)$, and the cost function could not be computed. If we replace the RKHS norm with the training error on a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ (as in Basis Pursuit De-Noising) the cost function could be computed, but it would lead to a different QP problem. Notice that the cost function contains the actual distance between the approximating and the unknown function, which is exactly the quantity that we want to minimize.

- As a side effect, this paper provides a derivation of the SVM algorithm in the framework of regularization theory (see appendix B). The advantage of this formulation is that it is particularly simple to state, and it is easily related to other well known techniques, such as smoothing splines and Radial Basis Functions. The disadvantage is that it hides the connection between SVM and the theory of VC bounds, and does not make clear what induction principle is being used. When the output of the target function is restricted to be 1 or -1, that is we consider a classification problem, Vapnik shows that SVM minimize an upper bound on the *generalization error*, rather than minimizing the training error within a fixed architecture. Although this is rigorously proved only in the classification case, this is a very important property, that makes SVM extremely well founded from the mathematical point of view. This motivation, however, is missing when the regularization theory approach is used to derive SVM.

- The equivalence between SVM and sparsity has only been shown in the case of noiseless data. In order to maintain the equivalence in the case of noisy data, one should prove that the presence of noise in the problem (17) leads to the additional constraint $\boldsymbol{\alpha}^*, \boldsymbol{\alpha} \leq C$ as in SVM, where C is some parameter inversely related to the amount of noise. In appendix C we sketch a tentative solution to this problem. This solution, however, is not very satisfactory because is purely formal, and it does not explain what assumptions are made on the noise in order to maintain the equivalence.

Acknowledgments I would like to thank T. Poggio and A. Verri for their useful comments and B. Olshausen for the long discussions on sparse approximation.

A Reproducing Kernel Hilbert Spaces

In this paper, a Reproducing Kernel Hilbert Space (RKHS) (Aronszajn, 1950) is defined a Hilbert space of functions defined over some domain $\Omega \subset R^d$ with the property that, for each $\mathbf{x} \in \Omega$, the evaluation functionals $\mathcal{F}_{\mathbf{x}}$ defined as

$$\mathcal{F}_{\mathbf{x}}[f] = f(\mathbf{x}) \quad \forall f \in \mathcal{H}$$

are linear, bounded functionals. It can be proved that to each RKHS $\mathcal{H}$ it corresponds a positive definite function $K(\mathbf{x}, \mathbf{y})$, which is called the *reproducing kernel* of $\mathcal{H}$. The kernel of $\mathcal{H}$ has the following *reproducing property*:

$$f(\mathbf{x}) = < f(\mathbf{y}), K(\mathbf{y}; \mathbf{x}) >_{\mathcal{H}} \quad \forall f \in \mathcal{H} \tag{27}$$

where $< \cdot, \cdot >_{\mathcal{H}}$ denotes the scalar product in $\mathcal{H}$. The function K acts in a similar way to the delta function in L_2, although L_2 is not a RKHS (its elements are not necessarily defined pointwise). Here we sketch a way to construct a RKHS, which is relevant to our paper. The mathematical details (such the convergence or not of certain series) can be found in the theory of integral equations (Hochstadt, 1973; Cochran, 1972; Courant and Hilbert, 1962), which is very well established, so we do not discuss them here. In the following we assume that $\Omega = [0, 1]^d$ for simplicity. The main ideas will carry over to the case $\Omega = R^d$, although with some modifications, as we will see in section (A.2).
Let us assume that we find a sequence of positive numbers λ_n and linearly independent functions $\phi_n(\mathbf{x})$ such that they define a function $K(\mathbf{x}; \mathbf{y})$ in the following way[1]:

$$K(\mathbf{x}; \mathbf{y}) \equiv \sum_{n=1}^{\infty} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{y}) \tag{28}$$

where the series is well defined (for example it converges uniformly). A simple calculation shows that the function K defined in eq. (28) is positive semi-definite. Let us now take as Hilbert space the set of functions of the form

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}) \tag{29}$$

in which the scalar product is defined as:

$$< \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}), \sum_{n=1}^{\infty} d_n \phi_n(\mathbf{x}) >_{\mathcal{H}} \equiv \sum_{n=1}^{\infty} \frac{c_n d_n}{\lambda_n} \tag{30}$$

Assuming that all the evaluation functionals are bounded, it is now easy to check that such an Hilbert space is a RKHS with reproducing kernel given by $K(\mathbf{x}; \mathbf{y})$. In fact we have

$$< f(\mathbf{x}), K(\mathbf{x}; \mathbf{y}) >_{\mathcal{H}} = \sum_{n=1}^{\infty} \frac{c_n \lambda_n \phi_n(\mathbf{y})}{\lambda_n} = \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{y}) = f(\mathbf{y}).$$

We conclude that it is possible to construct a RKHS whenever a function K of the form (28) is available. The norm in this RKHS has the form:

$$\|f\|_{\mathcal{H}}^2 = \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n} \tag{31}$$

It is well known that expressions of the form (28) actually abound. In fact, it follows from Mercer's theorem (Hochstadt, 1972) then any function $K(\mathbf{x}; \mathbf{y})$ which is the kernel of a positive operator[2] in $L_2(\Omega)$ has an expansion of the form (28), in which the ϕ_i and the λ_i are respectively,

[1] When working with complex functions $\phi_n(\mathbf{x})$ this formula should be replaced with $K(\mathbf{x}; \mathbf{y}) \equiv \sum_{n=1}^{\infty} \lambda_n \phi_n(\mathbf{x}) \phi_n^*(\mathbf{y})$

[2] We remind the reader that positive operators in L_2 are self-adjoint operators such that $< Kf, f > \geq 0$ for all $f \in L_2$.

the orthogonal eigenfunctions and the positive eigenvalues of the operator corresponding to K. In (Stewart, 1976) it is reported that the positivity of the operator associated to K is equivalent to the statement that the kernel K is positive definite, that is the matrix $K_{ij} = K(\mathbf{x}_i; \mathbf{x}_j)$ is positive definite for all choices of distinct points $\mathbf{x}_i$. Notice that a kernel K could have an expansion of the form (28) in which the ϕ_n are not necessarily its eigenfunctions.
The case in which $\Omega = R^d$ is similar, with the difference that the eigenvalues may assume *any* positive value, so that there will be a non-countable set of orthogonal eigenfunctions. In the following section we provide a number of examples of these different situations, that also show why the norm $|f\|^2_{\mathcal{H}}$ can be seen as a smoothness functional.

A.1 Examples: RKHS over $[0, 2\pi]$

Here we present a simple way to construct meaningful RKHS of functions of one variable over $[0, 2\pi]$. In the following all the normalization factors will be set to 1 for simplicity.
Let us consider any function $K(x)$ which is continuous, symmetric, periodic, and whose Fourier coefficients λ_n are positive. Such a function can be expanded in a uniformly convergent Fourier series:

$$K(x) = \sum_{n=0}^{\infty} \lambda_n \cos(nx) \; . \tag{32}$$

An example of such a function is

$$K(x) = 1 + \sum_{n=1}^{\infty} h^n \cos(nx) == \frac{1}{2\pi} \frac{1 - h^2}{1 - 2h\cos(x) + h^2}$$

where $h \in (0, 1)$.
It is easy to check that, if (32) holds, then we have:

$$K(x - y) = 1 + \sum_{n=1}^{\infty} \lambda_n \sin(nx)\sin(ny) + \sum_{n=1}^{\infty} \lambda_n \cos(nx)\cos(ny) \tag{33}$$

which is of the form (28) in which the set of *orthogonal* functions ϕ_n has the form:

$$\{\phi_i(x)\}_{i=0}^{\infty} \equiv (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(nx), \cos(nx), \ldots) \; .$$

Therefore, given any function K which is continuous, periodic and symmetric we can then define a RKHS $\mathcal{H}$ over $[0, 2\pi]$ by defining a scalar product of the form:

$$< f, g >_{\mathcal{H}} \equiv \sum_{n=0}^{\infty} \frac{f_n^c g_n^c + f_n^s g_n^s}{\lambda_n}$$

where we use the following symbols for the Fourier coefficients of a function f:

$$f_n^c \equiv < f, \cos(nx) > \; , \quad f_n^s \equiv < f, \sin(nx) >$$

The functions in $\mathcal{H}$ are therefore functions in $L_2([0, 2\pi])$ whose Fourier coefficients satisfy the following constraint:

$$\|f\|^2_{\mathcal{H}} = \sum_{n=0}^{\infty} \frac{(f_n^c)^2 + (f_n^s)^2}{\lambda_n} < +\infty \tag{34}$$

Since the sequence λ_n is decreasing, the constraint that the norm (34) has to be finite can be seen as a constraint on the rate of decrease to zero of the Fourier coefficients of the function f, which is known to be related to the smoothness properties of f. Therefore, choosing different kernels K is equivalent to choose RKHS of functions with different smoothness properties, and the norm (34) can be used as the smoothness functional $\Phi[f]$ in the regularization approach sketched in section 2. The relationship between the kernel K and the smoothness properties of the functions in the corresponding RKHS will become more clear in the next section, where we discuss the extension of this approach to the infinite domain $\Omega = R^d$.

A.2 Examples: RKHS over R^d

When the domain Ω over which we wish to define a RKHS becomes the whole space R^d most of the results of the previous section still apply, with the difference that the spectrum of K becomes (usually) the whole positive axis, and it is not countable anymore.
For translation invariant kernels, that is positive definite functions of the form $K(\mathbf{x}-\mathbf{y})$, the following decomposition holds:

$$K(\mathbf{x}-\mathbf{y}) = \int_{R^d} d\mathbf{s} \tilde{K}(\mathbf{s}) e^{i\mathbf{s}\cdot\mathbf{x}} e^{-i\mathbf{s}\cdot\mathbf{y}} \tag{35}$$

Equation (35)is the analog of (28) over an infinite domain, and one can go from the case of bounded Ω to the case of $\Omega = R^d$ by the following substitutions:

$$\begin{array}{lll} n & \Rightarrow & \mathbf{s} \\ \lambda_n & \Rightarrow & \tilde{K}(\mathbf{s}) \\ \phi_n(\mathbf{x}) & \Rightarrow & e^{i\mathbf{s}\cdot\mathbf{x}} \\ \sum_{n=1}^{\infty} & \Rightarrow & \int_{R^d} d\mathbf{s} \end{array}$$

We conclude then that any positive definite function of the form $K(\mathbf{x}-\mathbf{y})$ defines a RKHS over R^d by defining a scalar product of the form

$$< f, g >_{\mathcal{H}} \equiv \int d\mathbf{s} \, \frac{\tilde{f}(\mathbf{s}) \tilde{g}^*(\mathbf{s})}{\tilde{K}(\mathbf{s})} \tag{36}$$

The reproducing property of K is easily verified:

$$< f(\mathbf{x}), K(\mathbf{x}-\mathbf{y}) >= \int d\mathbf{s} \, \frac{\tilde{f}(\mathbf{s}) \tilde{K}(\mathbf{s}) e^{-i\mathbf{y}\cdot\mathbf{s}}}{\tilde{K}(\mathbf{s})} = f(\mathbf{y})$$

and the RKHS becomes simply the subspace of $L_2(R^d)$ of the functions such that

$$\|f\|^2_{\mathcal{H}} = \int d\mathbf{s} \, \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{K}(\mathbf{s})} < +\infty \tag{37}$$

Functionals of the form (37) are known to be *smoothness* functionals. In fact, the rate of decrease to zero of the Fourier transform of the kernel will control the smoothness property of the function in the RKHS. Consider for example, in one dimension, the kernel $K(x) = e^{-|x|}$, whose Fourier Transform is $\tilde{K}(s) = (1+s^2)^{-1}$. The RKHS associated to this kernel contain functions such

$$\|f\|^2_{\mathcal{H}} = \int ds \, \frac{|\tilde{f}(s)|^2}{(1+s^2)^{-1}} = \|f\|^2_{L_2} + \|f'\|^2_{L_2} < \infty$$

This is the well known Sobolev space W_2^1, where we denote by W_2^m the set of functions whose derivatives up to order m are in L_2 (Yosida, 1974). Notice that the norm induced by the scalar product (36) is the smoothness functional considered by Girosi, Jones and Poggio (1995) in their approach to regularization theory for function approximation. This is not surprising, since RKHS have been known to play a central role in spline theory (Wahba, 1990). Notice also that in spline theory one actually deals with semi-RKHS, in which the norm $\|\cdot\|_{\mathcal{H}}$ has been substituted with a semi-norm. Semi-RKHS share most of the properties of RKHS, but their theory becomes a little more complicated because of the null space of the semi-norm, which has to be taken in account. Details about semi-RKHS can be found in (Wahba, 1990).

A.3 Finite Dimensional RKHS

When the set of basis functions ϕ_i has finite cardinality N, the construction of a RKHS sketched in the previous section (eq. 1 and 30) is always well defined, as long as the basis functions are linearly independent. Notice that the functions ϕ_i do not have to be orthogonal, and that they will *not* be the eigenfunctions of K. It is interesting to notice that in this case we can define a different set of basis functions $\tilde{\phi}_i$, which we call the *dual* basis functions, with some interesting properties. The dual basis functions are defined as:

$$\tilde{\phi}_i(\mathbf{x}) = \sum_{i,j=1}^{N} M_{ij}^{-1} \phi_j(\mathbf{x}) \tag{38}$$

where M^{-1} is the inverse of the matrix M:

$$M_{ij} \equiv < \phi_i, \phi_j > \tag{39}$$

and the scalar product is taken in L_2. It is easy to verify that, for any function of the space, the following identity holds:

$$f(\mathbf{x}) = \sum_{i=1}^{N} < f, \tilde{\phi}_i > \phi_i(\mathbf{x}) = \sum_{i=1}^{N} < f, \phi_i > \tilde{\phi}_i(\mathbf{x}) \tag{40}$$

where the second part of the identity comes from the fact that the dual basis of the dual basis is the original basis. From here we conclude that, for any choices of positive λ_i, the set of functions spanned by the functions ϕ_i form a RKHS whose norm is

$$\|f\|_{\mathcal{H}}^2 \equiv \sum_{i=1}^{N} \frac{< f, \tilde{\phi}_i >^2}{\lambda_i} \tag{41}$$

Notice that while the elements of the (dual) basis are not orthogonal to each other, orthogonality relationships hold between elements of the basis and the elements of the dual basis:

$$< \tilde{\phi}_i, \phi_j >= \delta_{ij} \tag{42}$$

As a consequence, it is also possible to show that, defining the dual kernel as:

$$\tilde{K}(\mathbf{x}; \mathbf{y}) \equiv \sum_{i=1}^{N} \lambda_i \tilde{\phi}_i(\mathbf{x}) \tilde{\phi}_i(\mathbf{y}) \tag{43}$$

the following relationships hold:

$$\int d\mathbf{y}\ K(\mathbf{x};\mathbf{y})\tilde{\phi}_i(\mathbf{y}) = \lambda_i\phi_i(\mathbf{x})$$

$$\int d\mathbf{y}\ \tilde{K}(\mathbf{x};\mathbf{y})\phi_i(\mathbf{y}) = \lambda_i\tilde{\phi}_i(\mathbf{x})$$

A.3.1 A RKHS of polynomials

In this section we present a particular RKHS, which allows us to partially answer to an old standing question: is it possible to derive, in the framework of regularization theory, an approximation scheme of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i\sigma(\mathbf{x}\cdot\mathbf{w}_i + \theta) \tag{44}$$

where σ is some continuous, one dimensional function? Connections between approximation schemes of the form (44) and regularization theory have been presented before (Girosi, Jones, and Poggio, 1995), but always involving approximation or extension of the original regularization theory framework. A positive answer to the previous question is given by noticing that it is possible to define a RKHS whose kernel is

$$K(\mathbf{x};\mathbf{y}) = (1 + \mathbf{x}\cdot\mathbf{y})^d$$

where d is any integer. In fact, the kernel K above has an expansion of the form:

$$K(\mathbf{x};\mathbf{y}) \equiv \sum_{n=1}^{N} \lambda_n\phi_n(\mathbf{x})\phi_n(\mathbf{y}) \tag{45}$$

where the ϕ_n are the monomials of degree up to d, which constitutes a basis in the set of polynomials of degree d, and the λ_i are some positive numbers. The kernel above therefore can be used to give the structure of RKHS to the set of polynomials of degree d (in arbitrary number of variables), and the norm defined by eq. (41) can be used a smoothness functional in the regularization theory approach (see appendix B) to derive an approximating scheme of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{l} c_i(1 + \mathbf{x}\cdot\mathbf{x}_i)^d$$

which is a special case of eq. (44). The fact that the kernel $K(\mathbf{x};\mathbf{y}) = (1 + \mathbf{x}\cdot\mathbf{y})^d$ has an expansion of the form (45) had been reported by Vapnik (1995). Here we give some examples, from which the reader can infer the general case.
Let us start from the one-dimensional case, where one chooses:

$$\lambda_n = \begin{pmatrix} d \\ n \end{pmatrix} \qquad \phi_n(x) = x^n \quad n = 0,\ldots,d \quad .$$

It is now easy to see that

$$K(x;y) = \sum_{n=0}^{d} \lambda_n\phi_n(x)\phi_n(y) = \sum_{n=0}^{d} \begin{pmatrix} d \\ n \end{pmatrix} (xy)^n = (1 + xy)^d$$

A similar result, although with a more complex structure of the coefficients λ_n, is true in the multivariate case. For example in two variables we can define:

$$\{\phi_i(\mathbf{x})\}_{i=1}^{6} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

$$\{\lambda_i\}_{i=1}^{6} = (1, \sqrt{2}\ , \sqrt{2}\ , \sqrt{2}\ 1, 1)$$

In this case it is easy to verify that:

$$K(\mathbf{x};\mathbf{y}) = \sum_{n=1}^{6} \lambda_n \phi_n(\mathbf{x})\phi_n(\mathbf{y}) = (1 + \mathbf{x}\cdot\mathbf{y})^2 \tag{46}$$

The same kernel can be obtained in 3 variables by choosing:

$$\{\phi_i(\mathbf{x})\}_{i=1}^{10} = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3)$$

$$\{\lambda_i\}_{i=1}^{10} = (1, \sqrt{2}, \sqrt{2}\ , \sqrt{2}\ , 1, 1, 1, \sqrt{2}\ , \sqrt{2}\ , \sqrt{2}\ )$$

In 2 variables, we can also make the following choice:

$$\{\phi_i(\mathbf{x})\}_{i=1}^{9} = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2, x_1^2x_2, x_1x_2^2, x_1^3\ , x_2^3\ )$$

$$\{\lambda_i\}_{i=1}^{9} = (1, \sqrt{3}\ , \sqrt{3}\ , \sqrt{3}\ , \sqrt{3}\ , \sqrt{6}, \sqrt{3}\ , \sqrt{3}\ , 1\ , 1)$$

and it is easy to see that:

$$K(\mathbf{x};\mathbf{y}) = \sum_{n=1}^{9} \lambda_n \phi_n(\mathbf{x})\phi_n(\mathbf{y}) = (1 + \mathbf{x}\cdot\mathbf{y})^3$$

While it is difficult to write a closed formula for the coefficients λ_i in the general case, it is obvious that such coefficients can always be found. This, therefore, leaves unclear what form of smoothness is imposed by using the norm in this particular RKHS, and shows that this example is quite an academic one. A much more interesting case is the one in which the function σ in eq. (44) is a sigmoid or some other activation function. If it were possible to find an expansion of the form:

$$\sigma(\mathbf{x}\cdot\mathbf{y} + \theta) = \sum_{n=1}^{\infty} \lambda_n \phi_n(\mathbf{x})\phi_n(\mathbf{y})$$

for some fixed value of θ, this would mean that a scheme of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \sigma(\mathbf{x}\cdot\mathbf{x}_i + \theta)$$

could be derived in a regularization theory framework. Vapnik (1995) reports that if $\sigma(x) = \tanh(x)$ the corresponding kernel is positive definite for some values of θ, but the observation is of experimental nature, so that we do not know what the λ_i or the ϕ_n are, and therefore we do not what kind of functions the corresponding RKHS contains.

B Derivation of the SVM Algorithm

B.1 Generalities on Regularization Theory

Let us look more closely at the solution of the variational problem (2):

$$\min_{f\in\mathcal{H}} H[f] = C\sum_{i=1}^{l} V(y_i - f(\mathbf{x}_i)) + \frac{1}{2}\Phi[f]$$

We assume that $\mathcal{H}$ is a RKHS with kernel K and that the smoothness functional $\Phi[f]$ is:

$$\Phi[f] = \|f\|^2_{\mathcal{H}}$$

This is equivalent to assume that the functions in $\mathcal{H}$ have a unique expansion of the form:

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

and that their norm is:

$$\|f\|^2_{\mathcal{H}} = \sum_{n=1}^{\infty}\frac{c_n^2}{\lambda_n}$$

In this derivation we do not have the coefficient b which appears in (1), since we argued before that if one of the ϕ_i is constant, which is usually the case, this term is not necessary.
We can think of the functional $H[f]$ as a function of the coefficients c_n. In order to minimize $H[f]$ we take its derivative with respect to c_n and set it equal to zero, obtaining the following:

$$-C\sum_{i=1}^{l} V'(y_i - f(\mathbf{x}_i))\phi_n(\mathbf{x}_i) + \frac{c_n}{\lambda_n} = 0 \tag{47}$$

Let us now define the following set of unknowns:

$$a_i \equiv CV'(y_i - f(\mathbf{x}_i))$$

Using eq. (47) we can express the coefficients c_n as a function of the a_i:

$$c_n = \lambda_n\sum_{i=1}^{l} a_i\phi_n(\mathbf{x}_i)$$

The solution of the variational problem has therefore the form:

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} c_n\phi_n(\mathbf{x}) = \sum_{n=1}^{\infty}\sum_{i=1}^{l} a_i\lambda_n\phi_n(\mathbf{x}_i)\phi_n(\mathbf{x}) = \sum_{i=1}^{l} a_i K(\mathbf{x};\mathbf{x}_i) \tag{48}$$

where we have used the expansion (28). This shows that, independently of the form of V, the solution of the regularization functional (2) is always a linear superposition of kernel functions, one for each data point. The cost function V affects the computation of the coefficients a_i. In fact, plugging eq. (48) back in the definition of the a_i we obtain the following set of equations for the coefficients a_i:

$$a_i = CV' \left(y_i - \sum_{j=1}^{l} K_{ij} a_j \right) \quad , \quad i = 1, \ldots, l$$

where we have defined $K_{ij} = K(\mathbf{x}_i; \mathbf{x}_j)$. In the case in which $V(x) = x^2$ we obtain the standard regularization theory solution (see Girosi, Jones and Poggio, 1995 for an alternative derivation):

$$(K + \gamma I)\ \mathbf{a} = \mathbf{y}$$

where we have defined $\gamma \equiv \frac{1}{C}$.

B.2 The SVM algorithm in the Regularization Theory Framework

Following Vapnik (1995) we now consider the case of the ϵ-insensitive cost function $V(x) = |x|_\epsilon$. In this case the approach sketched above is problematic because V is not differentiable at $x = \epsilon$ (although it still makes sense everywhere else). In order to make our notation consistent with Vapnik's one, we have to modify slightly the model proposed in the previous section. Vapnik explicitly takes into account an offset in the model, so that equation (1) is replaced by

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}) + b \tag{49}$$

The smoothness functional remains unchanged (so that the smoothness does not depend on b):

$$\Phi[f] = \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n}$$

Also, we scale the functional in (2) of a factor $\frac{1}{2\lambda} \equiv C$, obtaining the following variational problem:

$$\min_{f \in \mathcal{H}} H[f] = C \sum_{i=1}^{l} |y_i - f(\mathbf{x}_i)|_\epsilon + \frac{1}{2}\Phi[f]$$

Since it is difficult to deal with the function $V(x) = |x|_\epsilon$, the problem above is replaced by the following equivalent[3] problem, in which an additional set of variables is introduced:

$$\min_{f \in \mathcal{H}} H[f] = C \sum_{i=1}^{l} (\xi_i + \xi_i^*) + \frac{1}{2}\Phi[f] \tag{50}$$

subject to

$$\begin{array}{lll} f(\mathbf{x}_i) - y_i & \leq & \epsilon + \xi_i \qquad i = 1, \ldots, l \\ y_i - f(\mathbf{x}_i) & \leq & \epsilon + \xi_i^* \qquad i = 1, \ldots, l \\ \xi_i & \geq & 0 \qquad\quad\ \ i = 1, \ldots, l \\ \xi_i^* & \geq & 0 \qquad\quad\ \ i = 1, \ldots, l \end{array} \tag{51}$$

The equivalence of the variational problem is established just noticing that in the problem above a (linear) penalty is paid only when the absolute value of the interpolation error exceeds ϵ, which correspond to Vapnik's ϵ-insensitive cost function. Notice that when of the two top constraints is

[3]By *equivalent* we mean that the function that minimizes the two functionals is the same

satisfied with some non-zero ξ_i (or ξ_i^*), the other is automatically satisfied with a zero value for ξ_i^* (or ξ_i). In order to solve the constrained minimization problem above we use the technique of Lagrange multipliers. The Lagrangian corresponding to the problem above is:

$$\begin{aligned} \mathcal{L}(f, \boldsymbol{\xi}, \boldsymbol{\xi}^*; \boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \mathbf{r}, \mathbf{r}^*) &= C\sum_{i=1}^{l}(\xi_i + \xi_i^*) + \frac{1}{2}\Phi[f] + \sum_{i=1}^{l}\alpha_i^*(y_i - f(\mathbf{x}_i) - \epsilon - \xi_i^*) + \\ &\quad + \sum_{i=1}^{l}\alpha_i(f(\mathbf{x}_i) - y_i - \epsilon - \xi_i) - \sum_{i=1}^{l}(r_i\xi_i + r_i^*\xi_i^*) \end{aligned} \tag{52}$$

where $\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \mathbf{r}, \mathbf{r}^*$ are positive Lagrange multipliers. The solution of the constrained variational problem above is now obtained by minimizing the Lagrangian (52) with respect to f (that is with respect to the c_n and to b), $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ and maximizing (in the positive quadrant) with respect to $\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \mathbf{r}, \mathbf{r}^*$. Since the minimization step is now unconstrained, we set to zero the derivatives with respect to c_n, b, $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$, obtaining:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial c_n} = 0 &\Rightarrow c_n = \lambda_n \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi_n(\mathbf{x}_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 &\Rightarrow \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_n} = 0 &\Rightarrow r_n = C - \alpha_n \\ \frac{\partial \mathcal{L}}{\partial \xi_n^*} = 0 &\Rightarrow r_n^* = C - \alpha_n^* \end{aligned}$$

Substituting the expression for the coefficients c_n in the model (49) we then conclude that the solution of the problem (50) is a function of the form

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(\mathbf{x}; \mathbf{x}_i) + b \tag{53}$$

Substituting eq. (53) in the Lagrangian, we obtain an expression that should now be maximized (in the positive quadrant) with respect to $\boldsymbol{\alpha}, \boldsymbol{\alpha}^*, \mathbf{r}, \mathbf{r}^*$, with the additional constraints listed above. Noticing that the relationship between r_n (r_n^*) and α_n (α_n^*) implies that $\boldsymbol{\alpha} \leq C$ and $\boldsymbol{\alpha}^* \leq C$, and minimizing $-\mathcal{L}$ rather than maximizing $\mathcal{L}$, we now obtain the following QP problem:

Problem B.1

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \epsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l}y_i(\alpha_i^* - \alpha_i) + \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\mathbf{x}_i; \mathbf{x}_j),$$

subject to the constraints

$$\begin{gathered} 0 \leq \boldsymbol{\alpha}^*, \boldsymbol{\alpha} \leq C \\ \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \end{gathered}$$

This is the QP problem that has to be solved in order to compute the SVM solution. It is useful to write and discuss the Kuhn-Tucker conditions:

$$
\begin{aligned}
\alpha_i(f(\mathbf{x}_i) - y_i - \epsilon - \xi_i) = 0 \quad & i = 1, \dots, l \\
\alpha_i^*(y_i - f(\mathbf{x}_i) - \epsilon - \xi_i^*) = 0 \quad & i = 1, \dots, l \\
(C - \alpha_i)\xi_i = 0 \quad & i = 1, \dots, l \\
(C - \alpha_i^*)\xi_i^* = 0 \quad & i = 1, \dots, l
\end{aligned}
$$

The input data points $\mathbf{x}_i$ for which α_i or α_i^* are different from zero are called *support vectors*. Few observations are in order:

- The Lagrange multipliers α_i and α_i^* cannot be simultaneously different from zero, so that the constraint $\alpha_i \alpha_i^* = 0$ holds.
- The support vectors are those data points $\mathbf{x}_i$ at which the interpolation error is either greater or equal to ϵ. Points at which the interpolation error is smaller than ϵ are never support vectors, and do not enter in the determination of the solution. Once they have been found, they could be removed from the data set, and if the SVM were run again on the new data set the same solution would be found.
- Any of the support vectors for which $0 < \alpha_i < C$ (and therefore $\xi_i = 0$) can be used to compute the parameter b. In fact, in this case it follows from the Kuhn-Tucker conditions that:

  $$f(\mathbf{x}_i) = \sum_{j=1}^{l} \alpha_j K(\mathbf{x}_i; \mathbf{x}_j) + b = y_i + \epsilon$$

  (a similar argument holds for the α_i^*).
- If $\epsilon = 0$ then all the points become support vectors;
- Because of the constraint $\alpha_i \alpha_i^* = 0$, defining

  $$\mathbf{a} = \boldsymbol{\alpha}^* - \boldsymbol{\alpha}$$

  and using eq. (24) the QP problem B.1 can be written as follows:

  Problem B.2

  $$\min_{\mathbf{a}} E^*[\mathbf{a}] = \epsilon \|\mathbf{a}\|_{L_1} - \mathbf{a} \cdot \mathbf{y} + \frac{1}{2} \mathbf{a} \cdot K \mathbf{a}$$

  subject to the constraints

  $$
  \begin{aligned}
  -C \leq a_i \leq C \\
  \mathbf{a} \cdot \mathbf{1} = 0
  \end{aligned}
  $$

Important note: Notice that if one the basis functions ϕ_i is constant, then the parameter b in (49) could be omitted. *The RKHS described in appendix A all have this property.*

C Noisy case: an equivalence?

It is natural to ask whether the result of this paper can be extended to the case of noisy data. I will sketch here an argument to show that there is still a relationship between SVM and sparse approximation, when data are noisy, although the relationship is much less clear. In the presence of additive noise we have

$$f(\mathbf{x}_i) = y_i + \delta_i \ ,$$

where y_i are the measured value of f, and δ_i are random variables with unknown probability distribution. Substituting y_i with $y_i + \delta_i$ in eq. (23), disregarding the constant term in $\|f\|^2_{\mathcal{H}}$, and defining

$$E^*[\mathbf{a}] = -\sum_{i=1}^{l} a_i y_i + \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j K(\mathbf{x}_i;\mathbf{x}_j) + \epsilon \sum_{i=1}^{l} |a_i|$$

we conclude that we need to minimize the following QP problem:

Problem C.1

$$\min_{\mathbf{a}} [E^*[\mathbf{a}] - \mathbf{a}\cdot\boldsymbol{\delta}]$$

subject to the constraint:

$$\mathbf{a}\cdot\mathbf{1} \quad = 0$$

where the vector $\boldsymbol{\delta}$ *is unknown.*

In order to understand how to deal with the fact that we do not know $\boldsymbol{\delta}$, let us consider a different QP problem:

Problem C.2

$$\min_{\mathbf{a}} E^*[\mathbf{a}]$$

subject to the constraints:

$$\begin{aligned} \mathbf{a}\cdot\mathbf{1} \quad &= 0 \\ \mathbf{a} \quad &\geq \boldsymbol{\eta} \\ \mathbf{a} \quad &\leq \boldsymbol{\eta}^* \end{aligned}$$

where the box parameters $\boldsymbol{\eta}$ *and* $\boldsymbol{\eta}^*$ *are unknown.*

We solve problem C.2 using the Lagrange multipliers technique for the inequality constraints, obtaining the following dual version of problem C.2:

Problem C.3

$$\max_{\boldsymbol{\beta},\boldsymbol{\beta}^*}\min_{\mathbf{a}} [E^*[\mathbf{a}] - \mathbf{a}\cdot(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \boldsymbol{\beta}\cdot\boldsymbol{\eta} - \boldsymbol{\beta}^*\cdot\boldsymbol{\eta}^*]$$

subject to the constraint:

$$\begin{aligned} \mathbf{a}\cdot\mathbf{1} \quad &= 0 \\ \boldsymbol{\beta},\boldsymbol{\beta}^* \quad &\geq 0 \end{aligned}$$

where $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^*$ *are vectors of Lagrange multipliers.*

Notice now that the choice of the box parameters $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ uniquely determines $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, and that setting $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, problems C.1 and C.3 are identical for what concerns the $\mathbf{a}$ vector: they both require to solve a QP problem in which a linear term contains unknown coefficients. Therefore, solving problem C.1 with unknown $\boldsymbol{\delta}$ seems to be formally equivalent to solving problem C.3 with unknown box parameters. This suggests the following argument: 1) solving C.1 with unknown $\boldsymbol{\delta}$ is formally equivalent to solving problem C.3 with unknown box parameters; 2) in absence of any information on the noise, and therefore on the box parameters, we could set the box parameters to $\boldsymbol{\eta}^* = -\boldsymbol{\eta} = C\mathbf{1}$ for some unknown C; 3) for $\boldsymbol{\eta}^* = -\boldsymbol{\eta} = C\mathbf{1}$ problem C.3 becomes the usual QP problem of SVM (problem B.1); 4) therefore, in total absence of information on the noise, problem C.1 leads to the same QP problem of SVM, making the equivalence between sparse approximation and SVM complete. However this argument is not very rigorous, because it does not make clear how the assumptions on $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ are reflected on the noise vector $\boldsymbol{\delta}$. However, the formal similarity of the problems C.3 and C.1 seems to point in the right direction, and an analysis of the relationship between $\boldsymbol{\eta}$, $\boldsymbol{\eta}^*$ and $\boldsymbol{\delta}$ could lead to useful insights on the assumptions which are made on the noise in the SVM technique.

References

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

[2] M. Bertero. Regularization methods for linear inverse problems. In C. G. Talenti, editor, *Inverse Problems*. Springer-Verlag, Berlin, 1986.

[3] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifier. In *Proc. 5th ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992.

[4] L. Breiman. Better subset selection using the non-negative garotte. Technical report, Department of Statistics, University of California, Berkeley, 1993.

[5] S. Chen, , D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

[6] S. Chen. *Basis Pursuit*. PhD thesis, Department of Statistics, Stanford University, November 1995.

[7] J.A. Cochran. *The analysis of linear integral equations*. McGraw-Hill, New York, 1972.

[8] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

[9] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[10] I. Daubechies. *Ten lectures on wavelets*. CBMS-NSF Regional Conferences Series in Applied Mathematics. SIAM, Philadelphia, PA, 1992.

[11] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. A.I. Memo 1606, MIT Artificial Intelligence Laboratory, 1997. (available at the URL: http://www.ai.mit.edu/people/girosi/svm.html).

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

[13] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximation and learning: outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[14] G.F. Harpur and R.W. Prager. Development of low entropy coding in a recurrent network. *Network*, 7:277–284, 1996.

[15] H. Hochstadt. *Integral Equations*. Wiley Classics Library. John Wiley & Sons, 1973.

[16] P.J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[17] S. Mallat and Z. Zhang. Matching Pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[18] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[19] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[20] V.A. Morozov. *Methods for solving incorrectly posed problems*. Springer-Verlag, Berlin, 1984.

[21] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[22] 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.

[23] L.L. Schumaker. *Spline functions: basic theory*. John Wiley and Sons, New York, 1981.

[24] A. Smola and B. Schölkopf. From regularization operators to support vector kernels. In *Advances in Neural Information Processings Systems 10*. Morgan Kaufmann Publishers, 1998.

[25] J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain J. Math.*, 6:409–434, 1976.

[26] R. Tibshirani. Regression selection and shrinkage via the lasso. Technical report, Department of Statistics, University of Toronto, June 1994. ftp://utstat.toronto.edu/pub/tibs/lasso.ps.

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

[28] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[29] V. Vapnik, S.E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processings Systems 9*, pages 281–287, San Mateo, CA, 1997. Morgan Kaufmann Publishers.

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

[31] G. Wahba. Smoothing noisy data by spline functions. *Numer. Math*, 24:383–393, 1975.

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

[33] K. Yosida. *Functional Analysis*. Springer, Berlin, 1974.