Monograph:
Dodd, T.J. and Harrison, R.F. (2001) Iterative sparse interpolation in reproducing kernel Hilbert spaces. Research Report. ACSE Research Report no. 814 . Automatic Control and Systems Engineering, University of Sheffield

Iterative Sparse Interpolation in Reproducing Kernel Hilbert Spaces

Tony J. Dodd and Robert F. Harrison
Department of Automatic Control and Systems Engineering
University of Sheffield, Sheffield S1 3JD, UK
e-mail: {t.j.dodd, r.f.harrison}@shef.ac.uk

Research Report No. 814
December 2001

Abstract

The problem of interpolating data in reproducing kernel Hilbert spaces is well known to be ill-conditioned. In the presence of noise, regularisation can be applied to find a good solution. In the noise-free case, regularisation has the effect of over-smoothing the function and few data points are interpolated. In this paper an alternative framework, based on sparsity, is proposed for interpolation of noise-free data. Iterative construction of a sparse sequence of interpolants is shown to be well defined and produces good results.

1 Introduction

The method of regularisation is usually applied to function approximation where the available observations are noisy. Reproducing kernel Hilbert spaces (RKHS) provide a framework for such problems. However, in certain cases we may be interested in the interpolation problem whereby the observed data are assumed to be virtually noise-free. This is the case in numerical quadrature or where we seek to interpolate between solutions of simulations, for example in computational fluid dynamics, to reduce the high computational burden. Despite the lack of noise, such problems may still be ill-conditioned owing to a lack of effective linear independence of the basis functions. Round-off errors in the computer can then prevent the computation of interpolating solutions. An approach to overcoming this is to consider a sparse subset of the available data.

In this paper a novel theoretical framework for sparsity is described which is valid for both batch and iterative applications. In fact key results are given which ensure that a sequence of iterative sparse solutions will be well defined and convergent. Again, these results are novel in the context of RKHS. The motivation for sparsity is, unlike previous approaches, as an alternative to regularisation. In fact we wish to avoid regularisation in interpolation applications as the solution will be too smooth to interpolate the data. A simple iterative scheme for constructing a sparse set of data points is proposed based on minimising the condition number of the kernel (Gram) matrix.

Basic definitions and results on RKHS can be found in the papers (Aronszajin 1950; Wahba 1990). Additional useful references on RKHS which focus on linear time series analysis include . For function approximation RKHS are equivalent to the method of potential functions (Aizerman, Braverman, and Rozonoer 1964) for which iterative solutions based on stochastic approximation are well known (Fu 1968). More recently support vector machines and Gaussian processes have been introduced (Vapnik 1998; Williams 1999) which can be considered as particular examples of approximation in RKHS.

More generally, the approximation of functions in Hilbert spaces is described in (Kreyszig 1978) together with properties of projection operators. More particularly, the theory of generalised inverses for linear operators widely used in approximation theory can be found in (Groetsch 1978). For details on ill-conditioning in discrete problems, learning with discrete data and singular systems see (Bertero, De Mol, and Pike 1985; Neumaier 1998; Hansen 1998). Finally, for a similar approach to iterative sparsity as described in this paper see (Partington 1997) on matching pursuit in RKHS.

In the next section approximation from finite data in RKHS is described. The solution to the approximation problem is then presented in Section 3. In Section 4 the numerical stability of this solution is addressed. Sparse approximation solutions are analysed in detail in Section 5 with particular reference to iterative schemes, and finally, an illustrative example is given.

2 Point Observations from Hilbert Spaces

We assume that we have some unknown function f of interest but that we are able to observe its behaviour. The function belongs to some Hilbert space $\mathcal{F}$ defined on some parameter set $\mathcal{X}$. This set can be considered as an input set in the sense that for $x \in \mathcal{X}$, $f(x)$ represents the evaluation of f at x.

A finite set of observations $\{z_i\}_{i=1}^N$ of the function is made corresponding to inputs $\{x_i\}_{i=1}^N$. It is assumed that the space of all possible observations is a metric space $\mathcal{Z}$ (a metric space is required later in assessing the ill-conditioning of the problem). Neglecting the effects of errors, the observations arise as follows

$$z_i = L_i f \tag{1}$$

where $\{L_i\}_{i=1}^N$ is a set of linear evaluation functionals, defined on $\mathcal{F}$, which associate real numbers to the function f. We can represent the complete set of observations $[z_1, \ldots, z_N]^T$ in vector form as follows

$$z^N = Lf = \sum_{i=1}^{N} (L_i f) e_i \tag{2}$$

where $e_i \in \mathbb{R}^N$ is the ith standard basis vector.

In general L_i permits indirect observation (e.g. via derivatives of f), but we are concerned with the case

$$z_i = f(x_i) \tag{3}$$

leading to the exact interpolation problem.

The approximation problem can then be formulated as follows (Bertero, De Mol, and Pike 1985): given a class $\mathcal{F}$ of functions, and a set $\{z_i\}_{i=1}^N$ of values of linear functionals $\{L_i\}_{i=1}^N$ defined on $\mathcal{F}$, find in $\mathcal{F}$ a function f which satisfies Eq. 1.

By assuming that $\mathcal{F}$ is a Hilbert space, and further, the $\{L_i\}_{i=1}^N$ are continuous (hence bounded), it follows from the Riesz representation theorem that we can express the observations as (Akhiezer and Glazman 1981)

$$L_i f = \langle f, \psi_i \rangle_{\mathcal{F}}, \quad i = 1, \ldots, N \tag{4}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denotes the inner product in $\mathcal{F}$. The $\{\psi_i\}_{i=1}^N$ are a set of functions each belonging to $\mathcal{F}$ and uniquely determined by the functionals $\{L_i\}_{i=1}^N$.

The approximation problem can now be stated as follows: given the Hilbert space of functions $\mathcal{F}$, the set of functions $\{\psi_i\}_{i=1}^N \subset \mathcal{F}$ and the observations $\{z_i\}_{i=1}^N$, find a function $f \in \mathcal{F}$ such that Eq. 4 is satisfied. We now address the case where $\mathcal{F}$ is a RKHS.

Formally a RKHS is a Hilbert space of functions on some parameter set $\mathcal{X}$ with the property that, for each $x \in \mathcal{X}$, the evaluation functional L_i, which associates f with $f(x_i)$, $L_i f \to f(x_i)$, is a bounded linear functional (Wahba 1990). The boundedness means that there exists a scalar $M \geq 0$ such that

$$|L_i f| = |f(x_i)| \leq M \|f\|_{\mathcal{F}} \quad \text{for all } f \text{ in the RKHS}$$

where $\|\cdot\|_{\mathcal{F}}$ is the norm in the Hilbert space. But to satisfy the Riesz representation theorem the L_i must be bounded, hence any Hilbert space satisfying the Riesz theorem will be a RKHS.

We use $k_i = k(x_i, \cdot)$ to refer to ψ_i (i.e. the evaluation of the function $k(x_i, \cdot) = \psi_i$ at x_j is $k_{ij} = k(x_i, x_j)$). The inner product $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{F}}$ must equal $k(x_i, x_j)$ by the Riesz representation theorem. This leads to the following important result: $k(x_i, x_j)$ is positive definite since, for any $x_1, \ldots, x_n \in \mathcal{X}$, $a_i, \ldots, a_n \in \mathbb{R}$,

$$\begin{aligned}\sum_{i,j} a_i a_j k(x_i.x_j) &= \sum_{i,j} a_i a_j \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{F}} \\ &= \left\| \sum a_i k(x_i, \cdot) \right\|_{\mathcal{F}}^2 \geq 0.\end{aligned}$$

The following is then a standard theorem on RKHS.

Theorem 1 *(Aronszajin 1950) To every RKHS there corresponds a unique positive-definite function (the reproducing kernel) and, conversely, given a positive-definite function k on $\mathcal{X} \times \mathcal{X}$ we can construct a unique RKHS of real-valued functions on $\mathcal{X}$ with k as its reproducing kernel.*

We also have.

Definition 1 *(Parzen 1961) A Hilbert space $\mathcal{F}$ is said to be a reproducing kernel Hilbert space, with reproducing kernel k, if the members of $\mathcal{F}$ are functions on some set $\mathcal{X}$, and if there is a kernel k on $\mathcal{X} \times \mathcal{X}$ having the following two properties; for every $x \in \mathcal{X}$ (where $k(\cdot, x_2)$ is the function defined on $\mathcal{X}$, with value at x_1 in $\mathcal{X}$ equal to $k(x_1, x_2)$):*

1. *$k(\cdot, x_2) \in \mathcal{F}$; and*
2. *$\langle f, k(\cdot, x_2)\rangle_{\mathcal{F}} = f(x_2)$*

for every f in $\mathcal{F}$.

We can then associate with $k(\cdot, \cdot)$ a unique collection of functions of the form

$$f(\cdot) = \sum_i a_i k(x_i, \cdot) \tag{5}$$

for $a_i \in \mathbb{R}$. A well defined inner product for this collection is (Wahba 1990)

$$\left\langle \sum_i a_i k(x_i, \cdot), \sum_j b_j k(x_j, \cdot) \right\rangle_{\mathcal{F}} = \sum_{i,j} a_i b_j \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{F}} = \sum_{i,j} a_i b_j k(x_i.x_j).$$

For this collection, norm convergence implies pointwise convergence and we can therefore adjoin all limits of Cauchy sequences of functions which are well defined as pointwise limits (Wahba 1990). The resulting Hilbert space is then a RKHS.

3 Approximation in RKHS

The functions k_i associated with the N observations $L_i f$ define a (linearly independent) basis for the N dimensional subspace $\mathcal{F}_N \subset \mathcal{F}$. Therefore we can express any $g \in \mathcal{F}_N$ as

$$g = \sum_{i=1}^{N} a_i k_i. \tag{6}$$

$\mathcal{F}_N$ is a closed subspace of $\mathcal{F}$ and therefore there exists a unique best approximation $\hat{f} \in \mathcal{F}_N$ to any $f \in \mathcal{F}$. In fact

$$\mathcal{F} = \mathcal{F}_N \oplus \mathcal{F}_N^{\perp} \tag{7}$$

where $F_N^{\perp}$ represents the orthogonal complement of the closed space F_N. We must therefore have $(f - \hat{f}) \perp \mathcal{F}_N$, i.e.

$$\langle f - \hat{f}, k_i \rangle_{\mathcal{F}} = 0, \quad \forall\ k_i. \tag{8}$$

Expressing this using Eq. 6

$$\left\langle f - \sum_{j=1}^{N} a_j k_j, k_i \right\rangle_{\mathcal{F}} = 0 \quad i = 1, \ldots, N \tag{9}$$

giving the N conditions

$$\langle f, k_i \rangle_{\mathcal{F}} - a_1 \langle k_1, k_i \rangle_{\mathcal{F}} - \cdots - a_N \langle k_N, k_i \rangle_{\mathcal{F}} = 0. \tag{10}$$

This is a nonhomogeneous system of N linear equations in N unknowns $a_1, \ldots, a_N$ having a unique normal solution given by the solution of the matrix equation

$$K a = z^N \tag{11}$$

where K is the Gram (or kernel) matrix defined by $[K]_{ij} = \langle k_i, k_j \rangle_{\mathcal{F}} = k(x_i, x_j) = k_{ij}$ (using the basic results of RKHS). Similarly z^N is the vector of observations with $\langle f, k_i \rangle_{\mathcal{F}} = L_i f = z_i$. A unique solution exists since the k_i are linearly independent and therefore the Gram determinant $\det(K) \neq 0$. Representing the Gram determinant by $G(k_1, \ldots, k_N)$ then the distance between f and $\hat{f}$ can be expressed as follows

$$\|f - \hat{f}\|_{\mathcal{F}}^2 = \frac{G(f, k_1, \ldots, k_N)}{G(k_1, \ldots, k_N)} \tag{12}$$

where, by definition,

$$G(f, k_1, \ldots, k_N) = \begin{vmatrix} \langle f, f \rangle & \langle f, k_1 \rangle & \cdots & \langle f, k_N \rangle \\ \langle k_1, f \rangle & \langle k_1, k_1 \rangle & \cdots & \langle k_1, k_N \rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle k_N, f \rangle & \langle k_N, k_1 \rangle & \cdots & \langle k_N, k_N \rangle \end{vmatrix} \tag{13}$$

where all inner products are in fact $\langle \cdot, \cdot \rangle_{\mathcal{F}}$.

We can express the dual form to Eq. 6 using

$$k^i = \sum_{j=1}^{N} K^{ij} k_j, \tag{14}$$

where $K^{ij} = [K^{-1}]_{ij}$, as

$$\hat{f} = \sum_{i=1}^{N} z_i k^i. \tag{15}$$

Note that the k^i, k_j satisfy $\langle k^i, k_j \rangle_{\mathcal{F}} = \delta_{ij}$.

This dual representation is important as it demonstrates that the solution depends continuously on the data, i.e. $\|\delta \hat{f}\|_{\mathcal{F}} \to 0$ when $\|\delta z^N\|_{\mathcal{Z}} \to 0$ where δz^N represents a variation of z^N and $\delta \hat{f}$ the corresponding variation of $\hat{f}$.

4 Numerical Stability

In the previous section it was stated that the normal solution depends continuously on the data. We will now explore this in more detail. Although we are interested in the interpolation problem we assume that the observations are affected by noise. Even in the so-called noise-free case the observations will be affected by the numerical precision of the computer. If the problem is ill-conditioned this limit in numerical precision can mean that no solution can be found. However, we may desire a solution in any case. The analysis in this section will guide us in how to construct such solutions.

In order to facilitate the analysis we introduce the adjoint operator L^* defined by

$$\langle Lf, z^N \rangle_{\mathcal{Z}} = \langle f, L^* z^N \rangle_{\mathcal{F}} \tag{16}$$

where $\mathcal{Z} \subset \mathbb{R}^N$. The adjoint operator transforms the observation vector z^N into an element of $\mathcal{F}$, or more precisely the finite dimensional subspace $\mathcal{F}_N$.

The inverse interpolation problem is then well known to be the solution of the following equation

$$L^* L \hat{f} = L^* z^N \tag{17}$$

which is equivalent to the orthogonal projection solution previously described. This can be shown using the identity $L^\dagger = (L^*L)^\dagger L^* = L^*(LL^*)^\dagger$ (where $\dagger$ denotes the matrix generalised inverse) and the results that (Dodd and Harrison 2001)

$$L^* = \sum_{i=1}^{N} k_i e_i \tag{18}$$

and

$$\hat{L} = LL^* = \sum_{i,j=1}^{N} k_{ij} e_j e_i^T \tag{19}$$

in a RKHS. The latter shows that LL^* is equivalent to the kernel matrix K. The solution for the interpolant is then given by

$$\hat{f} = L^\dagger z^N = L^*(LL^*)^\dagger z^N = k^T K^{-1} z^N \tag{20}$$

for K full rank and $k = [k_1, \ldots, k_N]$. This is equivalent to Eq. 15.

Given that the we assume the solution is an exact interpolant we have

$$z^N = L\hat{f} \tag{21}$$

and perturbing the observations with noise the solution is perturbed accordingly

$$z^N + \delta z^N = L(\hat{f} + \delta\hat{f}). \tag{22}$$

Subtracting $z^N = L\hat{f}$ from both sides the errors in the observations and interpolant are given by

$$\delta z^N = L\delta\hat{f}. \tag{23}$$

from which

$$\delta\hat{f} = L^\dagger \delta z^N. \tag{24}$$

Taking norms

$$||\delta\hat{f}||_{\mathcal{F}} \leq ||L^\dagger|| ||\delta z^N||_{\mathcal{Z}} \tag{25}$$

where $||\cdot||$ denotes the appropriate (operator) norm. But from Eq. 21 we also have

$$||z^N||_{\mathcal{Z}} \leq ||L|| ||\hat{f}||_{\mathcal{F}} \tag{26}$$

from which

$$\frac{1}{||\hat{f}||_{\mathcal{F}}} \leq \frac{||L||}{||z^N||_{\mathcal{Z}}}. \tag{27}$$

Substituting into Eq. 25

$$\frac{||\delta\hat{f}||_{\mathcal{F}}}{||\hat{f}||_{\mathcal{F}}} \leq ||L^\dagger|| ||L|| \frac{||\delta z^N||_{\mathcal{Z}}}{||z^N||_{\mathcal{Z}}} \tag{28}$$

where $||L^\dagger|| ||L|| = C(L)$ is known as the condition number.

The operators $\hat{L} = LL^*$ and $\tilde{L} = L^*L$ are self-adjoint, non-negative definite operators in $\mathcal{Z}$ and $\mathcal{F}$ respectively (Edmunds and Evans 1987). We can therefore form the eigenvalue-eigenvector decompositions of these operators. The positive eigenvalues are the same and we assume they are arranged to form the non-decreasing sequence

$$\alpha_1^2 \geq \alpha_2^2 \geq \cdots \geq \alpha_N^2. \tag{29}$$

We denote by v_i the eigenvector of $\hat{L}$ associated with α_i^2. Similarly the ith eigenfunction of $\tilde{L}$ is denoted by u_i. The v_i and u_i form orthonormal bases in $\mathcal{Z}$ and $\mathcal{F}_N$ respectively.

The set $\{\alpha_i, u_i, v_i\}_{i=1}^N$ is called the singular system of L. As the u_i form an orthonormal basis for $\mathcal{F}_N$ we can express $\hat{f}$ in terms of this basis as follows (Bertero, De Mol, and Pike 1985)

$$\hat{f} = \sum_{i=1}^{N} \frac{1}{\alpha_i} \langle z^N, v_i \rangle_{\mathcal{Z}} u_i. \tag{30}$$

It can also be shown that $\|L\| = \alpha_1$ and $\|L^\dagger\| = \alpha_N$ so that the condition number (Bertero, De Mol, and Pike 1985)

$$C(L) = \frac{\alpha_1}{\alpha_N}. \tag{31}$$

Returning to Eq. 28 then for a large condition number, $C(L)$, small errors in z^N will result in large (but bounded) errors in the solution. So whilst the solution strictly depends continuously on the data large errors are possible. In the case where $C(L) \gg 1$, i.e. $\alpha_1 \gg \alpha_N$, the problem of computing $\hat{f}$ is ill-conditioned. In the case where $\mathcal{Z}$ is the usual Euclidean space then the condition number is the square root of the ratio between the largest and smallest eigenvalues of the kernel matrix (Bertero, De Mol, and Pike 1985).

The ill-conditioning arises as the kernel basis functions are effectively linearly dependent. This manifests itself in many small singular values α_i. Because the terms $\langle z^N, v_i \rangle_{\mathcal{Z}}$ typically do not decay as fast as the singular values the solution is dominated by those terms corresponding to the smallest α_i. These are usually found to be the most oscillatory in practice (Hansen 1998). Hence the solution is dominated by highly oscillatory terms.

Various forms of regularisation have been proposed to overcome this problem. They have the basic characteristic of altering the singular values such that those corresponding to the highly oscillatory components are damped. In the case of singular value decomposition the factors $1/\alpha_i$ corresponding to those α_i below a threshold are simply set to zero. However we are interested in the case of interpolation in the presence of, at most, very small amounts of noise. The problem can still be ill-conditioned due to the effective linear dependence of the kernel basis functions in the presence even of the round-off errors of the computer. Any regularisation will have the effect of smoothing the approximation so that it no longer interpolates the data.

5 Sparse Solutions

The (potential) numerical instability is caused by the effective linear dependence of the kernel basis functions. Therefore by carefully selecting a subset of the kernel functions these can be chosen to be more linearly independent resulting once again in a stable problem. In this section various results will be shown relating to such a sparse set.

Consider now a subset of the $\{k_i\}$, $\{k_l, \ldots, k_p\}$, consisting of m elements where $1 \leq l \leq p \leq N$. This set now spans a new subspace, denoted $\mathcal{F}_m$, such that $\mathcal{F}_m \subset \mathcal{F}_N \subset \mathcal{F}$. The direct projection of f onto $\mathcal{F}_m$ follows exactly the analysis in Section 3. The solution can then be expressed

$$\hat{f}_m = \sum_{i \in I} b_i k_i \tag{32}$$

where $I = \{l, \ldots, p\}$ is the subset index set and the b_i are found as previously described using the reduced Gram matrix consisting of only terms $k_{ij}, i, j \in I$.

The direct projection of the approximation $\hat{f}_N$ using the whole basis set onto $\mathcal{F}_m$ can be found as follows. We require

$$\langle \hat{f}_N - \hat{f}_m, k_i \rangle_{\mathcal{F}} = 0, \quad i \in I \tag{33}$$

which can be expressed as

$$\left\langle \hat{f}_N - \sum_{j \in I} b_j k_j, k_i \right\rangle_{\mathcal{F}} = 0, \quad i \in I. \tag{34}$$

But $\hat{f}_N = \sum_{j=1}^{N} a_j k_j$ and hence

$$\left\langle \sum_{j=1}^{N} a_j k_j - \sum_{j \in I} b_j k_j, k_i \right\rangle_{\mathcal{F}} = 0, \quad i \in I \tag{35}$$

which is a set of m homogeneous linear equations in m unknowns $b_l, \ldots, b_p$. Expanding and rearranging

$$\begin{aligned} b_l \langle k_l, k_i \rangle_{\mathcal{F}} + \cdots + b_p \langle k_p, k_i \rangle_{\mathcal{F}} = \\ a_1 \langle k_1, k_i \rangle_{\mathcal{F}} + \cdots + a_N \langle k_N, k_i \rangle_{\mathcal{F}}. \end{aligned} \tag{36}$$

But the r.h.s. is, using Eq. 10, nothing more than $\langle f, k_i \rangle_{\mathcal{F}} = z_i$. We therefore arrive at the expected equation for the parameters b_i

$$K_m b = z^m \tag{37}$$

where the subscript m indicates that the Gram matrix and observation vector are now over the subset of m points comprising I only.

The equivalence of the projection from f to $\hat{f}^m$ directly and that via $\hat{f}^N$ is encapsulated in the following (general) partial ordering theorem.

Theorem 2 *Let P_1 and P_2 be projections defined on a Hilbert space $\mathcal{H}$. Denote by $\mathcal{H}_1 = P_1(\mathcal{H})$ and $\mathcal{H}_2 = P_2(\mathcal{H})$ the subspaces onto which $\mathcal{H}$ is projected by P_1 and P_2. Then the following conditions are equivalent*

1. $P_2 P_1 = P_1 P_2 = P_1$;
2. $\mathcal{H}_1 \subset \mathcal{H}_2$;
3. $\|P_1 h\| \leq \|P_2 h\|$ *for all* $h \in \mathcal{H}$; *and*
4. $P_1 \leq P_2$.

Proof. See (Kreyszig 1978). □

Specialising to our case we have $P_1 = P_m : \mathcal{F}_m = P_m(\mathcal{F})$ and $P_2 = P_N : \mathcal{F}_N = P_N(\mathcal{F})$ where by construction $\mathcal{F}_m \subset \mathcal{F}_N$. We then have $P_N P_m = P_m P_N = P_m$, $\|P_m f\| \leq \|P_N f\|$ and $P_m \leq P_N$. The latter is known as partial ordering defined by $P_m \leq P_N$ if and only if $\langle P_m f, f\rangle_{\mathcal{F}} \leq \langle P_N f, f\rangle_{\mathcal{F}}$ for all $f \in \mathcal{F}$.

The following lemmas on positivity and difference of projections are required in the subsequent theorem.

Lemma 1 *For any projection P on a Hilbert space,*

1. $\langle Ph, h\rangle = \|Ph\|^2$;
2. $P \geq 0$; *and*
3. $\|P\| \leq 1$

for all $h \in \mathcal{H}$.

Proof. See (Kreyszig 1978). □

Lemma 2 *Let P_1 and P_2 be projections on a Hilbert space $\mathcal{H}$. Then:*

1. *the difference $P = P_2 - P_1$ is a projection on $\mathcal{H}$ if and only if $\mathcal{H}_1 \subset \mathcal{H}_2$, where $\mathcal{H}_i = P_i(\mathcal{H})$; and*
2. *if $P = P_2 - P_1$ is a projection, P projects $\mathcal{H}$ onto $\mathcal{H}_0$, where $\mathcal{H}_0$ is the orthogonal complement of $\mathcal{H}_1$ in $\mathcal{H}_2$.*

Proof. See (Kreyszig 1978). □

Therefore $P_m - P_N$ is a projection onto the orthogonal projection of $\mathcal{F}_m$ in $\mathcal{F}_N$.

We now give a theorem (as presented in (Kreyszig 1978)) on the convergence of iterative sparse solutions. The proof is also given as it contains a further useful result on the norm of the differenc between the sparse approximation $\hat{f}_m$ and the full approximation $\hat{f}_N$.

Theorem 3 *Let $\{P_j\}$ be a monotone increasing sequence of projections P_j defined on a Hilbert space $\mathcal{H}$. Then $\{P_j\}$ is strongly operator convergent, say, $P_j h \to Ph$ for every $h \in \mathcal{H}$, and the limit operator P is a projection defined on $\mathcal{H}$.*

Proof. Let $i < j$. By assumption, $P_i \leq P_j$, so that $P_i(\mathcal{H}) \subset P_j(\mathcal{H})$ by Theorem 2 and $P_j - P_i$ is a projection by Lemma 2. Hence for every fixed $h \in \mathcal{H}$ we obtain by Lemma 1

$$\begin{aligned} \|P_j h - P_i h\|^2 &= \|(P_j - P_i)h\|^2 \\ &= \langle (P_j - P_i)h, h\rangle \\ &= \langle P_j h, h\rangle - \langle P_i h, h\rangle \qquad (38) \\ &= \|P_j h\|^2 - \|P_i h\|^2. \qquad (39) \end{aligned}$$

Now $\|P_j\| \leq 1$ by Lemma 1, so that $\|P_i h\| \leq \|h\|$ for every j. Hence $\{\|P_j\|\}$ is a bounded sequence of numbers which is also monotone by Theorem 2 since $\{P_j\}$ is monotone. Hence $\{\|P_j\|\}$ converges. From this and Eq. 38 $\{\|P_j\|\}$ is a Cauchy sequence. Since $\mathcal{H}$ is complete (by definition), $\{P_j\}$ converges. The limit depends on h, say, $P_j h \to Ph$. This defines an operator P on $\mathcal{H}$ which is a projection. □

Consider now the set of iterated sparse solutions $\hat{f}_1, \ldots, \hat{f}_m$ where each solution is derived from the previous one with the addition of a further basis function, i.e. $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_m$. The associated sequence of projection operators $\{P_i\}$ will be a monotone increasing sequence which is therefore strongly operator convergent. In the limit then we have the sequence converging to P_N corresponding to all the available basis functions.

We can also quantify the error between $\hat{f}_m$ and $\hat{f}_N$ as follows. Using Eq. 39

$$\begin{aligned} \|P_N f - P_m f\|^2 &= \|P_N f\|^2 - \|P_m f\|^2 \\ &= \|\hat{f}_N\|^2 - \|\hat{f}_m\|^2. \end{aligned}$$

This can then be expressed in terms of the kernel functions as:

$$\begin{aligned} &\|\hat{f}_N\|^2 - \|\hat{f}_m\|^2 \\ &= \left\|\sum_{i=1}^{N} a_i k_i\right\|^2 - \left\|\sum_{i\in I} b_i k_i\right\|^2 \\ &= \sum_{i,j=1}^{N} a_i a_j \langle k_i, k_j\rangle_{\mathcal{F}} - \sum_{i,j\in I} b_i b_j \langle k_i, k_j\rangle_{\mathcal{F}} \\ &= \sum_{i,j=1}^{N} a_i a_j k_{ij} - \sum_{i,j\in I} b_i b_j k_{ij}. \end{aligned}$$

6 An Illustrative Example

The question is then how to select such a sparse set. Recall that the ill-conditioning of the problem is determined by $C(L)$ which is equal to the ratio

of the smallest and largest singular values of the Gram matrix (or square root of the ratio of the corresponding eigenvalues).

The approach we therefore propose to construct the iterative sparse set is as follows. The first data point is selected as that which minimises the error over the training data. Successive data points are then added as those which minimise the corresponding condition number of the kernel matrix. Data points are then added until a threshold condition number C_T is achieved. The interpolant is then estimated using this sparse set.

We consider the simple example of approximating a sinc function as used in (Vapnik 1998) to demonstrate the sparsity of the representations given by support vector machines. One hundred random uniformly distributed samples were generated in the interval $[0, 1]$. The sinc function was then evaluated using

$$z_i = \frac{\sin(20x_i - 10)}{20x_i - 10} \tag{40}$$

with $z_i = 1$ for $x_i = 0.5$. The RKHS chosen was that corresponding to the kernel $k(x, x') = \exp(-\sigma\|x - x'\|^2)$ with $\sigma = 100$. The condition number of the associated kernel matrix was approximately 2×10^{18} and no solution could be computed using all the data. Various threshold condition numbers were then investigated. With $C_T = 100$ the interpolant in Figure 1 was found which utilises only 15 data points. We note that the data points are approximately equally spaced. This spacing is determined by a sparsity trade-off between C_T and the kernel width. In the case of evenly spaced data the separation of the data points would in fact be equal arising due to equal splitting of the intervals between already selected data points.

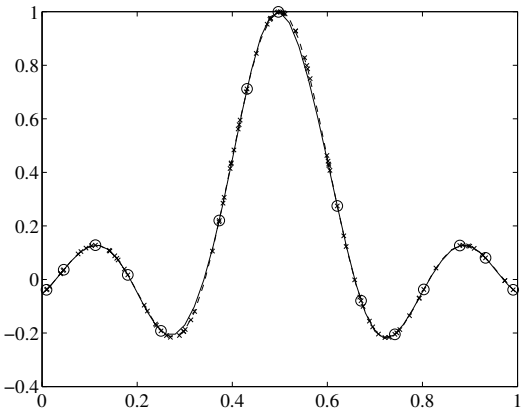

Figure 1: Typical predicted output ('- -') for the sparse solution and actual true output ('–'). The original data points ('×') and iterated sparse set are also shown ('o').

7 Concluding Remarks

A framework for function interpolation in the presence of finite data has been presented based on the idea of RKHS. The function of interest is treated as belonging to a RKHS, which is uniquely determined by a positive definite function called the reproducing kernel. In the case when the basis kernel functions are linearly independent the solution is well-conditioned and can be calculated using a simple matrix inverse. However, in the presence of even very small errors in the data (such as resulting from the numerical precision of the computer) the problem will be ill-conditioned if the kernels are effectively linearly dependent. This is valid even in the case where we are interested in interpolation of noise-free data. In order to overcome this problem and still interpolate the data it is necessary to consider a sparse version of the data set. The set of such sparse solutions can be constructed iteratively by addition of data points. Such a sequence of solutions is strongly convergent. Results were shown to demonstrate the interpolation of noise-free data using an example iterative sparse scheme.

Acknowledgements

The authors would like to thank the UK EPSRC for their financial support under Grant No. GR/R15726/01.

References

Aizerman, M., E. Braverman, and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automatika i Telemekhanika 25*(6), 821–837.

Akhiezer, N. and I. Glazman (1981). *Theory of Linear Operators in Hilbert Space*, Volume I. Pitman.

Aronszajin, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337–404.

Bertero, M., C. De Mol, and E. Pike (1985). Linear inverse problems with discrete data. I: General formulation and singular system analysis. *Inverse Problems 1*, 301–330.

Dodd, T. and R. Harrison (2001). The method of successive approximations for reproducing kernel Hilbert spaces. Technical Report 805, Department of Automatic Control and Systems Engineering, University of Sheffield, UK.

Edmunds, D. and W. Evans (1987). *Spectral Theory and Differential Operators*. Oxford Mathematical Monographs. Clarendon Press.

Fu, K. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*, Volume 52 of *Mathematics in Science and Engineering*. Academic Press.

Groetsch, C. (1978). *Generalized Inverses of Linear Operators*. Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker.

Hansen, P. (1998). Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. Technical report, Department of Mathematical Modelling, Technical University of Denmark.

Kailath, T. (1971). RKHS approach to detection and estimation problems - Part I: Deterministic signals in Gaussian noise. *IEEE Transactions on Information Theory IT-17*(5), 530–549.

Kreyszig, E. (1978). *Introductory Functional Analysis with Applications*. John Wiley & Sons.

Neumaier, A. (1998, September). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review 40*(3), 636–666.

Partington, J. (1997). *Interpolation, Identification, and Sampling*, Volume 17 of *London Mathematical Society Monographs New Series*. Clarendon Press.

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics 32*, 951–989.

Vapnik, V. (1998). *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 50 of *Series in Applied Mathematics*. Philadelphia: SIAM.

Williams, C. (1999). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning in Graphical Models*, pp. 599–621. The MIT Press.