15th Triennial World Congress, Barcelona, Spain

# ITERATIVE SOLUTION TO APPROXIMATION 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*

Abstract: A general framework for function approximation from finite data is presented based on reproducing kernel Hilbert spaces. Key results are summarised and the normal and regularised solutions are described. A potential limitation to these solutions for large data sets is the computational burden. An iterative approach to the least-squares normal solution is proposed to overcome this. Detailed proofs of convergence are given.

Keywords: Hilbert spaces, system identification, function approximation, Gaussian processes, iterative methods, least-squares approximation, regularisation

## 1. INTRODUCTION

Approximating functions given only finite data on the function is a problem common to system identification, nonlinear time series prediction and nonlinear predictive control. Neural networks and the NARMAX methodology are commonly used in these areas (Chen and Billings, 1992). Motivated by recent activity in kernel methods (Vapnik, 1998) we propose an alternative approach based on the idea of reproducing kernel Hilbert spaces (RKHS).

Basic definitions and results on RKHS can be found in the papers by Aronszajin (1950) and Wahba (1990) . Additional useful references on RKHS include the papers of Parzen (1961) and Kailath (1971) who focus on linear time series analysis. For function approximation, RKHS are equivalent to the method of potential functions (Aizerman *et al.*, 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.

Our main contribution is an iterative solution to approximation in RKHS with finite data including detailed proofs of convergence. The solution and assumptions for convergence are well known (Freiß and Harrison, 1999) but this is the first time, to the authors' knowledge, they have been presented for RKHS with finite data. The proofs are presented in detail unlike the basic assumption in Bertero (1988) which takes results from the continuous operator case and applies them to the finite data case without detailed proof. The iterative approach presented in Section 5 uses the basic formulation for general Hilbert spaces (Bertero *et al.*, 1985; Bertero *et al.*, 1988) and adapts the solution and proof for continuous operators in RKHS (Weiner, 1965). The latter only considers the time series case and not the more general function approximation problem addressed here.

In Section 2 the general problem of approximation in Hilbert spaces with finite data is described and specialised to RKHS in Section 3. The normal and regularised least-squares solutions to approximation in RKHS are then given in Section 4. The iterative approach to the normal solution is described in Section 5 together with detailed proofs of convergence. Finally, an illustrative system identification example is given in Section 6.

## 2. APPROXIMATION IN 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 $\mathcal{Z}$ (to permit the quantification of the effects of errors). 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 $i$th 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 *et al.*, 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. This is an inverse problem, the solution of which is given in Section 4. We now address the case where $\mathcal{F}$ is a RKHS.

## 3. REPRODUCING KERNEL HILBERT SPACES

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 $M$ such that

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

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 $k(x_i, \cdot)$ to refer to $\psi_i$ (i.e. the evaluation of the function $k(x_i, \cdot) = \psi_i$ at $x_j$ is $k(x_i, x_j)$). 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 $x_1, \ldots, x_n \in \mathcal{X}$, $a_i, \ldots, a_n \in \mathbb{R}$,

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

where $\|\cdot\|_{\mathcal{F}}$ is the corresponding norm in the RKHS. The following is then a standard theorem on RKHS.

*Theorem 3.1.* (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 then have a more common definition for RKHS.

*Definition 3.1.* (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

$$f(\cdot) = \sum_{i=1}^{L} c_i k(x_i, \cdot) \tag{5}$$

for $L \in \mathbb{Z}^+$ and $c_i \in \mathbb{R}$. 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.

Suppose that $k(x_1, x_2)$ is continuous and

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k^2(x_1, x_2)dx_1dx_2 < \infty \tag{6}$$

then there exists an orthonormal sequence of continuous eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ in $L_2(\mathcal{X})$ with associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ such that (Wahba, 1990)

$$\int_{\mathcal{X}} k(x_1, x_2)\phi_i(x_2)dx_2 = \lambda_i\phi_i(x_1), \quad i = 1, 2, \ldots \tag{7}$$

It can then be shown that if we let

$$f_i = \int_{\mathcal{X}} f(x)\phi_i(x)dx, \tag{8}$$

then $f \in \mathcal{F}$ if and only if (Wahba, 1990)

$$\sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i} < \infty \tag{9}$$

and

$$||f||_{\mathcal{F}}^2 = \sum_{i=1}^{\infty} \frac{f_i^2}{\lambda_i}. \tag{10}$$

Expanding $f$ in a Fourier series

$$f(x) = \sum_i f_i\phi_i(x). \tag{11}$$

For proofs of the foregoing results see Wahba (1990) .

## 4. NORMAL AND REGULARISED SOLUTIONS

Considering still the error free case, returning to the approximation problem of solving for $f \in \mathcal{F}$ in Eq. 4, we now assume that $\mathcal{F}$ is a RKHS and therefore the $\psi_i$ are given by $k(x_i, \cdot)$. The problem then is to find a function in the RKHS of the form, Eq. 5, which satisfies the data at the corresponding points. The solution will not be unique since we can only derive a finite number of values of $f$ from the observations. Assuming that the $k(x_i, \cdot)$ are linearly independent we can form a finite dimensional space $\mathcal{F}_N$, a subspace of $\mathcal{F}$. We can add to any solution in $\mathcal{F}_N$ any function orthogonal to this space to obtain a new solution.

We must then solve the following linear system

$$Kc = z^N \tag{12}$$

where $K$ is the kernel Gram matrix with elements $K_{ij} = \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{F}} = k(x_i, x_j)$.

This solution is the "normal" solution, $f^{\ddagger}$, and is guaranteed to exist and be unique as, within the set of solutions, there will always be one of minimal distance from the null element of $\mathcal{F}$.

It can be shown that the solution depends continuously on the data in the sense that, for a variation $\Delta z^N$ in $z^N$ and corresponding variation $\Delta f^{\ddagger}$ in $f^{\ddagger}$, $||\Delta f^{\ddagger}||_{\mathcal{F}} \to 0$ when $||\Delta z^N||_{\mathcal{Z}} \to 0$ (Bertero *et al.*, 1985). In the strict mathematical sense then, the problem of determining $f^{\ddagger}$ is well-posed. For large data sets, where the Gram matrix will have many small eigenvalues, much of the data does not effectively add any independent information about the function. In the presence of errors the problem will therefore be ill-conditioned.

If the functions $k(x_i, \cdot)$ are effectively dependent and the data $z_i$ are affected by errors then, in general, the normal solution no longer exists. Instead we must find a solution by minimising the norm of the errors in $\mathcal{Z}$, i.e. find an $f \in \mathcal{F}$ such that

$$\sum_{i=1}^{N} ||f(x_i) - z_i||_{\mathcal{Z}} = \text{minimum}. \tag{13}$$

However, this may still be ill-conditioned so we use instead, a solution corresponding to the minimiser of

$$J_{reg}[f] = \sum_{i=1}^{N} ||f(x_i) - z_i||_{\mathcal{Z}}^2 + \rho||f||_{\mathcal{F}}^2 \tag{14}$$

where $\rho \in \mathbb{R}^+$ is known as the regularisation parameter. We can rewrite Eq. 14 in terms of Eqs. 10 and 11

$$\sum_{i=1}^{N} \left\| \sum_j f_j\phi_j(x_i) - z_i \right\|_{\mathcal{Z}}^2 + \rho \sum_j \frac{f_j^2}{\lambda_j}. \tag{15}$$

Considering the case where $\mathcal{Z} = L_2$, i.e. $||\cdot||_{\mathcal{Z}} = ||\cdot||_2$ then to minimise Eq. 15 we minimise w.r.t. the $f_i$. The solution for $c$ is then given by

$$(K + \rho I)c = z^N \tag{16}$$

and

$$f(x) = \sum_{i=1}^{N} c_i k(x, x_i). \tag{17}$$

## 5. ITERATIVE SOLUTION

Consider now the case where we wish to compute the solution iteratively. The adjoint operator of $L$, $L^*$, is defined through

$$\langle Lf, z^N\rangle_{\mathcal{Z}} = \langle f, L^*z^N\rangle_{\mathcal{F}} \tag{18}$$

and transforms the observation vector $z^N$ into an element of $\mathcal{F}$, or more precisely the finite dimensional subspace $\mathcal{F}_N$. The adjoint operator in a RKHS is determined by (Appendix A)

$$L^*z^N = \sum_{i=1}^{N} k(x_i, \cdot)z_i \tag{19}$$

and also we show that

$$\hat{L} = LL^* = \sum_{j=1}^{N}\sum_{i=1}^{N} k(x_i, x_j)e_je_i^T \tag{20}$$

which is equivalent to the kernel matrix K. We can re-express Eqs. 16 and 17 as

$$f(x) = L^*(LL^* + \rho I)^{-1}z^N. \tag{21}$$

In the case where $\rho = 0$ the solution is the minimum of $||Lf - z^N||_{\mathcal{Z}}$ which is given by the solution of $L^*Lf = L^*z^N$. We denote this solution by $f^{\dagger}$.

Consider now an iterative solution for $f^\dagger$, then, defining a sequence of estimates as $\{f^n\}_{n=1}^\infty$, the method of successive approximations estimates $f^{n+1}$ in terms of $f^n$ as

$$f^{n+1} = f^n - \gamma_n \tilde{f}^n \tag{22}$$

where $f^0 \in \mathcal{F}$, $\gamma_n \in \mathbb{R}^+$ and $\tilde{f}^n$ is the residual

$$\tilde{f}^n = L^* L f^n - L^* z^N. \tag{23}$$

In practice the iterations must be made on finite dimensional objects. Returning to the basic solution in RKHS, Eq. 17, $f^n$ can be expressed, using the adjoint operator, as a linear combination of the $c_i$

$$f^n = L^* c^n \tag{24}$$

where $c^n = [c_1^n, \ldots, c_N^n]^T$. Also

$$\tilde{f}^n = L^* \tilde{c}^n, \quad \tilde{c}^n = LL^* c^n - z^N. \tag{25}$$

The method of successive approximations estimates the coefficients as

$$c^0 \in \mathbb{R}^N, \quad c^{n+1} = c^n - \gamma_n \tilde{c}^n \tag{26}$$

where the $\gamma_n$ are chosen as below. The function at each iteration is determined by $f^n = L^* c^n = \sum_{j=1}^N c_j^n k(x_j, \cdot)$.

To complete the iterative scheme we need to define a schedule for the parameters $\gamma_n$ and together with this prove convergence in the sense that $\|\tilde{f}^n\|^2 \to 0$ when $n \to \infty$.

*Theorem 5.1.* Let $\{\gamma_n\}_{n=1}^\infty$ satisfy:

(1) $0 < \gamma_n < 2/\lambda_{max}, \forall n$, where $\lambda_{max}$ is the largest eigenvalue of $LL^* = K$; and
(2) $\sum_{n=1}^\infty \gamma_n = \infty$.

Define the iteration $f^n = L^* c^n = \sum_{i=1}^N c_i^n k(x_i, \cdot)$ together with $f^0 \in \mathcal{F}$ (i.e. $c^0 \in \mathbb{R}^N$) arbitrary, $c^{n+1} = c^n - \gamma_n \tilde{c}^n$, $\tilde{c}^n = LL^* c^n - z^N$, then

$$\|\tilde{f}^n\|_{\mathcal{F}}^2 = \|L^* \tilde{c}^n\|_{\mathcal{F}}^2 \to 0.$$

as $n \to \infty$.

Proof.

(a) *Monotonicity.*

$$\tilde{f}^{n+1} = L^* L f^{n+1} - L^* z^N$$

but $f^{n+1} = L^* c^{n+1}$ and $c^{n+1} = c^n - \gamma_n \tilde{c}^n$, therefore $f^{n+1} = L^*(c^n - \gamma_n \tilde{c}^n)$ from which

$$\begin{aligned}\tilde{f}^{n+1} &= L^* L L^* (c^n - \gamma_n \tilde{c}^n) - L^* z^N \\ &= L^* L f^n - L^* z^N - \gamma_n L^* L \tilde{f}^n.\end{aligned}$$

Define

$$\begin{aligned}\Delta\|\tilde{f}^n\|_{\mathcal{F}}^2 &= \|\tilde{f}^n\|_{\mathcal{F}}^2 - \|\tilde{f}^{n+1}\|_{\mathcal{F}}^2 \\ &= \|\tilde{f}^n\|_{\mathcal{F}}^2 - \|\tilde{f}^n - \gamma_n L^* L \tilde{f}^n\|_{\mathcal{F}}^2\end{aligned}$$

and thus

$$\begin{aligned}\Delta\|\tilde{f}^n\|_{\mathcal{F}}^2 &= \|\tilde{f}^n\|_{\mathcal{F}}^2 - \|\tilde{f}^n\|_{\mathcal{F}}^2 - \gamma_n^2 \langle L^* L \tilde{f}^n, L^* L \tilde{f}^n \rangle_{\mathcal{F}} \\ &\quad + 2\gamma_n \langle \tilde{f}^n, L^* L \tilde{f}^n \rangle_{\mathcal{F}} \\ &= 2\gamma_n \langle \tilde{f}^n, L^* L \tilde{f}^n \rangle_{\mathcal{F}} - \gamma_n^2 \langle L^* L \tilde{f}^n, L^* L \tilde{f}^n \rangle_{\mathcal{F}} \\ &= 2\gamma_n \langle L \tilde{f}^n, L \tilde{f}^n \rangle_{\mathcal{Z}} - \gamma_n^2 \langle LL^* L \tilde{f}^n, L \tilde{f}^n \rangle_{\mathcal{Z}}\end{aligned}$$

using Eq. 18. Now

$$\frac{\langle L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}}}{\langle LL^* L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}}} = \frac{\sum_{j=1}^N (\tilde{f}_j^n)^2}{\sum_{j=1}^N (\tilde{f}_j^n)^2 \lambda_j} \tag{27}$$

where $\lambda_j$ is the $j$th eigenvalue of $LL^* = K$. Eq. 27 therefore satisfies

$$\frac{2\langle L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}}}{\langle LL^* L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}}} \geq \frac{2\sum_{j=1}^N (\tilde{f}_j^n)^2}{\lambda_{max} \sum_{j=1}^N (\tilde{f}_j^n)^2} = \frac{2}{\lambda_{max}}.$$

But by assumption $\gamma_n < 2/\lambda_{max}$ therefore

$$\Delta\|\tilde{f}^n\|_{\mathcal{F}}^2 = 2\gamma_n \langle L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}} - \gamma_n^2 \langle LL^* L\tilde{f}^n, L\tilde{f}^n \rangle_{\mathcal{Z}} \geq 0$$

□

(b) *Convergence.*

It was shown above that the residuals satisfy

$$\tilde{f}^{n+1} = L^* L f^n - L^* L \gamma_n \tilde{f}^n - L^* z^N$$

hence

$$\begin{aligned}\tilde{f}^{n+1} &= \tilde{f}^n - \gamma_n L^* L \tilde{f}^n = (I - \gamma_n L^* L)\tilde{f}^n \\ &= \prod_{k=1}^n (I - \gamma_k L^* L)\tilde{f}^0 = \prod_{k=1}^n (I - \gamma_k L^* L) g\end{aligned}$$

where $g = \tilde{f}^0 \in \mathcal{F}$ which can be expanded as (c.f. Eq. 11)

$$g = \sum_i g_i \phi_i(x) \tag{28}$$

and

$$L^* L g = \sum_i g_i \lambda_i \phi_i(x) \tag{29}$$

where $\lambda_i$ now refer to the eigenvalues of $L^* L$ [1].

We can then write

$$\prod_{k=1}^n (I - \gamma_k L^* L) g = \sum_i g_i \prod_{k=1}^n (1 - \gamma_k \lambda_i) \phi_i$$

and hence

$$\|\tilde{f}^{n+1}\|_{\mathcal{F}}^2 = \left\| \prod_{k=1}^n (I - \gamma_k L^* L) g \right\|_{\mathcal{F}}^2 = \sum_i \frac{g_i^2}{\lambda_i} \prod_{k=1}^n (1 - \gamma_k \lambda_i)^2$$

(c.f. Eq. 10). Using the assumed inequality on $\gamma_k$

$$1 - \frac{2\lambda_i}{\lambda_{max}} < 1 - \gamma_k \lambda_i < 1 \Rightarrow (1 - \gamma_k \lambda_i)^2 < 1.$$

[1] Note that $L^* L$ has the same positive eigenvalues as $LL^*$ with the same multiplicity. We therefore use the same notation to refer to the eigenvalues of each, although strictly $L^* L$ possesses more eigenvalues. The main result applies only to $\lambda_{max}$ which is common to both (Bertero *et al.*, 1985).

Then for any $L \in \mathbb{Z}^+$

$$\|\tilde{f}^{n+1}\|_{\mathcal{F}}^2 \leq \sum_{i=1}^{L} \frac{g_i^2}{\lambda_i} \prod_{k=1}^{n} (1-\gamma_k\lambda_i)^2 + \sum_{i>L} \frac{g_i^2}{\lambda_i}.$$

For fixed $L$, let $n \to \infty$, and since $\sum_{i=1}^{\infty} \gamma_i = \infty$ (by assumption) and $(1-\gamma_k\lambda_i)^2 < 1$ the first term tends to 0. Now let $L \to \infty$, $g \in \mathcal{F} \Rightarrow \sum_{i=1}^{\infty} g_i^2/\lambda_i < \infty$, and therefore the second term is the tail of a convergent series and therefore tends to 0.□

Defining further $c^0 = 0$ (and therefore $f^0 = 0$) then for any $n$, $\|f^n\|_{\mathcal{F}} \leq \|f^{n+1}\|_{\mathcal{F}}$ and therefore $\|f^n\|_{\mathcal{F}} \leq \|f^{\dagger}\|_{\mathcal{F}}$ (Bertero *et al.*, 1988). It follows that the method of successive approximations defines a regularisation scheme where the inverse of the number of iterations plays the role of the regularisation parameter.

## 6. EXAMPLE

As an example of the application of the iterative RKHS approach consider the discrete-time nonlinear dynamical system (Billings and Voon, 1986)

$$\begin{aligned} y(t) = {} & 0.5y(t-1) + 0.3y(t-1)u(t-1) \\ & + 0.2u(t-1) + 0.05y^2(t-1) + 0.6u^2(t-1) \end{aligned}$$

with the observations generated as $z(t) = y(t) + \varepsilon(t)$ where $\varepsilon(t) \sim N(0, 0.1)$ (note that this is a very noisy signal with a signal-to-noise ratio of approximately 30%). In identifying the system the data were generated from an initial condition of $y(1) = 0.1$ and the control input was sampled as $u(t) \sim N(0.2, 0.1)$. The RKHS approach was then applied to estimate a model of the form $y(t) = f(y(t-1), u(t-1))$ (in practise the exact model structure would normally be determined from the data).

Throughout, the reproducing kernel used is the Gaussian function, $k(x_i, x_j) = \exp(-\beta\|x_i - x_j\|_2^2)$ where $\beta \in \mathbb{R}^+$. Ten different sets of training and testing data with 500 samples each were used. In order to estimate $\beta$ and $\lambda$ for the static case a further 500 independent validation data were used. An appropriate value of $\beta$ was decided on as 0.1 and the $\lambda$ corresponding to the minimum of the validation MSE was 0.02. A value of $\gamma_n$ of 0.002 was chosen which always ensured that the conditions in Theorem 5.1 were satisfied.

Static and iterated models were then estimated for the ten data sets, for the iterative models 10,000 iterations were performed. An example prediction over the first 100 samples of one of the test sets for an iterative model is shown in Figure 1. In all cases the static and iterative models were very close as can be seen in Figure 2 which compares the estimated parameters. Note that the static parameters are scaled by 0.3257 which is necessary due to the effect of the regularisation.

The average MSE over the data sets for the static and iterative solutions are 0.0011 and 0.0012 respectively

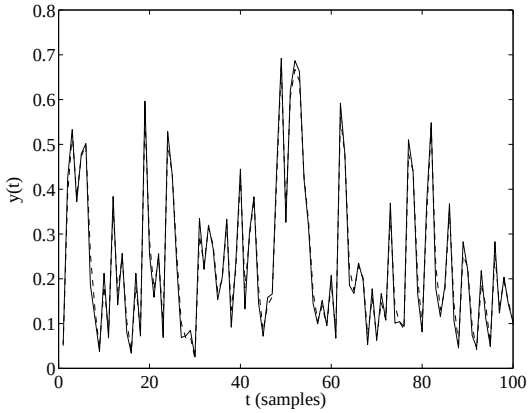


Fig. 1. Typical predicted output ('–') for the iterative solution and actual noise free true output ('- -').

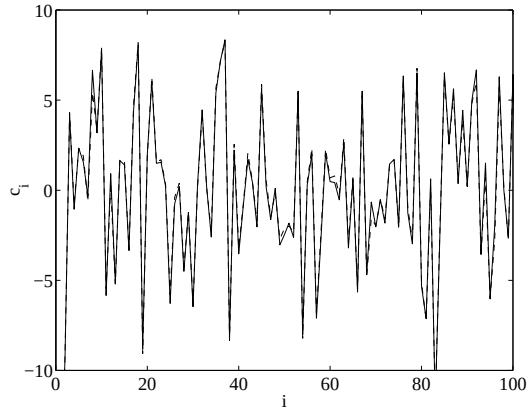


Fig. 2. Example comparison of first 100 parameters for iterative ('–') and static ('- -') solutions. Note the static parameters are scaled by 0.3257.

which compares favourably to the noise variance of 0.01. The average performance of the static solution is marginally better than the iterative solution. However, in four of the ten data sets the iterative solution was better. This is a feature of the particular test sets used.

## 7. CONCLUSIONS

A framework for normal and regularised function approximation 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 certain instances (e.g. large data sets) it may be necessary to solve iteratively for the function. An iterative approach to the least-squares, normal solution was presented, including detailed proofs of convergence, using the method of successive approximations. The approach was demonstrated using a system identification problem.

## ACKNOWLEDGEMENT

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

## 8. REFERENCES


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

Akhiezer, N.I. and I.M. Glazman (1981). *Theory of Linear Operators in Hilbert Space*. Vol. 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.R. Pike (1985). Linear inverse problems with discrete data. I: General formulation and singular system analysis. *Inverse Problems* **1**, 301–330.

Bertero, M., C. De Mol and E.R. Pike (1988). Linear inverse problems with discrete data. II: Stability and regularisation. *Inverse Problems* **4**, 573–594.

Billings, S.A. and W.S.F. Voon (1986). Correlation-based model validity tests for non-linear models. *International Journal of Control* **44**, 235–244.

Chen, S. and S.A. Billings (1992). Neural networks for nonlinear dynamic system modeling and identification. *International Journal of Control* **56**(2), 319–346.

Freiß, T.T. and R.F. Harrison (1999). A kernel-based adaline for function approximation. *Journal of Intelligent Data Analysis* **3**, 307–313.

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

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.

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

Vapnik, V.N. (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*. Vol. 50 of *Series in Applied Mathematics*. SIAM. Philadelphia.

Weiner, H.J. (1965). The gradient iteration in time series analysis. *Journal of the Society for Industrial and Applied Mathematics* **13**(4), 1096–1101.

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


## Appendix A. PROOFS RELATING TO ADJOINT OPERATORS IN RKHS

Consider the operator $L : \mathcal{F} \to \mathcal{Z}$ where $\mathcal{Z}$ is the $N$ dimensional Euclidean space with inner product $\langle g, h\rangle_{\mathcal{Z}} = \sum_{i=1}^{N} g_i h_i$, for $g, h \in \mathcal{Z}$, then, for $z^N \in \mathcal{Z}$, $f \in \mathcal{F}$ the adjoint operator $L^*$ is defined by

$$\langle Lf, z^N\rangle_{\mathcal{Z}} = \langle f, L^* z^N\rangle_{\mathcal{F}} \tag{A.1}$$

and transforms the observation vector $z^N$ into an element of $\mathcal{F}$ or more precisely the finite dimensional subspace $\mathcal{F}_N$. In a RKHS the operator $L$ acting on $f$ has the form $Lf = \sum_{i=1}^{N} \langle f, k(x_i, \cdot)\rangle_{\mathcal{F}} \cdot e_i$, where $e_i \in \mathbb{R}^N$ is the $i$th standard basis vector. The following results apply to the operator $L$ and its adjoint $L^*$.

*Theorem A.1.* Given the operator $L$ and its adjoint $L^*$ defined by

$$\langle Lf, z^N\rangle_{\mathcal{Z}} = \langle f, L^* z^N\rangle_{\mathcal{F}} \tag{A.2}$$

then in a RKHS with $Lf = \sum_{i=1}^{N} \langle f, k(x_i, \cdot)\rangle_{\mathcal{F}} \cdot e_i$ the adjoint $L^*$ is given by

$$L^* z^N = \sum_{i=1}^{N} z_i k(x_i, \cdot). \tag{A.3}$$

Proof. Solving for the LHS of Eq. A.2

$$\langle Lf, z^N\rangle_{\mathcal{Z}} = \sum_{i=1}^{N} \langle f, k(x_i, \cdot)\rangle_{\mathcal{F}} \cdot z_i = \sum_{i=1}^{N} f(x_i) z_i. \tag{A.4}$$

By assumption we set $L^* z^N = \sum_{i=1}^{N} z_i k(x_i, \cdot)$ and solving for the RHS of Eq. A.2

$$\langle f, L^* z^N\rangle_{\mathcal{F}} = \left\langle f, \sum_{i=1}^{N} z_i k(x_i, \cdot)\right\rangle_{\mathcal{F}} = \sum_{i=1}^{N} z_i \langle f, k(x_i, \cdot)\rangle_{\mathcal{F}} \tag{A.5}$$

the latter due to the linearity property of the inner product. But this is simply equal to $\sum_{i=1}^{N} z_i f(x_i)$.□

*Theorem A.2.* The operator $LL^*$ is equal to the kernel (Gram) matrix $K$, i.e.

$$LL^* = \sum_{j=1}^{N} \sum_{i=1}^{N} k(x_i, x_j) e_j e_i^T. \tag{A.6}$$

Proof. The operator $LL^*$ acting on $z^N$ can be expressed, using the previous results, as follows:

$$\begin{aligned} LL^* z^N &= L\left(\sum_{i=1}^{N} z_i k(x_i, \cdot)\right) \\ &= \sum_{j=1}^{N} \left\langle \sum_{i=1}^{N} z_i k(x_i, \cdot), k(x_j, \cdot)\right\rangle_{\mathcal{F}} \cdot e_j \\ &= \sum_{j=1}^{N} \sum_{i=1}^{N} z_i \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{F}} \cdot e_j \\ &= \sum_{j=1}^{N} \sum_{i=1}^{N} z_i k(x_i, x_j) e_j \end{aligned}$$

and therefore the operator $LL^* = \sum_{j=1}^{N} \sum_{i=1}^{N} k(x_i, x_j) e_j e_i^T$.□