Functional Learning in Signal Processing via Least Squares *

J.E. Perkins [†] I.M.Y. Mareels [†] J.B. Moore [†] R. Horowitz [‡]

March 1992

Abstract

This paper addresses certain functional learning tasks in signal processing using familiar algorithms and analytical tools of least squares for autoregressive moving average exogonous input (ARMAX) models. The models can be viewed as conventional ARMAX models but with parameters dependent on variables such as inputs or states, termed function input variables. The functional dependence of the parameters on these variables is represented in terms of basis function expansions, or more generally interpolation function representations. The interpolation functions in a least squares identification of coefficients also turn out to be in essence spread functions that spread learning throughout the space of function input variables. Thus for a set of training sequences, or trajectories in function input space, system parameters and thereby system functionals can be updated. The idea is that these will have relevance for similar sequences or neighbouring trajectories.

The concept of persistence of excitation to achieve complete function learning, or equivalently, signal model learning is studied using least squares convergence results. Application of the proposed algorithms and theory within the signal processing context is addressed by means of simple illustrative examples.

1 Introduction

The current neural network literature has highlighted the task of functional learning for application within the fields of control systems, and signal processing. The idea is that some input-output function $f(\cdot)$ is learned by means of a training sequence of function inputs x_k and outputs y_k for $k = 1, 2, \ldots, r$ as $\hat{f}(\cdot)$. The function estimate $\hat{f}(\cdot)$ can then be used to achieve outputs y from inputs x as $y = \hat{f}(x)$.

Of course, neural networks are usually restricted to the set of parametrizations for $\hat{f}(\cdot)$ in terms of suitably parametrized sigmoid functions and weights in a multilayer network. The parameters and weights are learnt by various methods including backward propagation, and extended Kalman filters [1]. The representations are such that the functions are not linear in the parameters/weights

*Work partially supported by Boeing (BCAC).

[†]Department of Systems Engineering, Research School of Physical Sciences and Engineering, Australian National University, GPO Box 4, Canberra ACT 2601.

[‡]Department of Mechanical Engineering, University of California, Berkley, California, U.S.A.

so that standard least squares or weighted least squares parameter estimations techniques do not apply.

For a number of reasons it would be of interest to pursue the role of least squares techniques for functional learning where the functions are linear in the parameters (weights). Least squares methods can be truly recursive in that estimates can be updated as each new measurement arrives. They are readily implemented. Also their convergence properties are relatively well understood within the adaptive control and signal processing context where they are ubiquitous. It has been a natural development for such adaptive methods to evolve towards learning systems where the underlying task is functional learning rather than parameter estimation. Thus in the trivial functional learning environment when the function is linear and constant, it is appealing for the learning algorithms to specialize to the well understood least squares based parameter estimation schemes.

A key property of least squares algorithms is that their convergence depends on certain excitation conditions of the regression vectors, which in turn depend on external excitations. This property in the adaptive estimation context should carry over to the functional learning context. In earlier studies [2][3][4], the concept of functional persistence of excitation is developed for continuous-time deterministic systems in an infinite dimensional setting working with integral operators. The kernel functions allow information to spread in the function input space. Application studies for the control of robots are performed using discrete-time and gradient or least squares ideas. From this work, the question that naturally emerges is: What are fundamental results concerning functional learning and persistence of excitation in a least squares stochastic identification context?

In this paper we interpret a class of functional learning tasks as least squares parameter estimation tasks, or a system of lower order least squares parameter estimation tasks performed in parallel. One of the main ideas used in the paper is that in learning a function $f(\cdot)$ at a point γ from input-output measurements x_k, y_k, the closer x_k is to γ, the greater the influence of the pair x_k, y_k should be in learning $f(\gamma)$. Thus for x_k in the neighborhood of γ, the associated weightings are high relative to weightings for x_k outside the neighbourhood of γ. The weightings then control what can be termed the spread of learning.

The algorithm we propose, in its most general form, seeks function estimates, or rather function parametrization estimates at a set of points $\Gamma_I = [\gamma_1, \gamma_2, \ldots, \gamma_n]$ in the function input variable space Γ_x. As each new measurement pair x_k, y_k arrives, estimates on Γ_I are updated with the learning being strongest in the neighbourhood of x_k and diminished or even zero outside this neighbourhood. With estimates at Γ_I, then an interpolation function can be used to give estimates on Γ_x. In fact, in our algorithm, the interpolation function is also used to control the spreading out of learning. Because of this dual role for the interpolation function, we must select bisigmoidal functions $K_i(x)$ which decay to zero outside the neighbourhood of γ_i. Thus polynomial, spline, and Fourier basis or interpolation functions are not an appropriate practical choice, although some of our formulations allow such function representations.

A second main idea of the paper concerns the convergence of least squares algorithms in the functional learning context. Known convergence theory for the least squares algorithms can be applied. Thus in any calculation, convergence behaviour can be estimated on line in terms of persistence of excitation measures on variables used in the calculations, under appropriate assumptions. It is of course desirable to translate such excitation conditions onto external variables. We claim that the (functional) persistence of excitation conditions for consistent estimation of the function, under function reconstructability conditions are a natural generalization of the available theory for the parameter estimation context, making connections to related work [4].

So as to generalize least squares based adaptive schemes in signal processing and control, we will work with signal models which are natural generalizations of familiar input-output models in these fields.

We desire to learn the functional representation of the coefficients of the discrete-time "ARMAX" equation, specialized to the white noise case, namely

$$A(x_k)y_k = B(x_k)u_k + w_k \tag{1.1}$$

where w_k is zero mean white Gaussian noise, $u_k \in \Gamma_u$ the set of allowable inputs, and $x_k \in \Gamma_x$, the set of allowable function input variables. The vectors x_k, y_k, u_k are measurable at time k. Here $A(x_k) = 1 + a_1(x_k)q^{-1} + \ldots + a_n(x_k)q^{-n}$ and $B(x_k) = 1 + b_1(x_k)q^{-1} + \ldots + b_m(x_k)q^{-m}$ where q^{-1} is the unit delay operator. Given a set of noisy measurements $\{x_k, y_k, u_k\}$ we propose two different types of least squares algorithms to estimate (possibly matrix or vector) function representations $f(x)$ of the coefficients $a_i(\cdot)$, $b_i(\cdot)$.

Of course, (1.1) is a special case of the more general form

$$y_k = \Phi_k'\Theta(x_k) + w_k \tag{1.2}$$

When specialized to (1.1), then $\Theta'(\cdot) = [a_1(\cdot)\ldots a_n(\cdot)b_1(\cdot)\ldots b_m(\cdot)]$ and $\Phi_k' = [y_{k-1}\cdots y_{k-n}u_{k-1}\cdots u_{k-m}]$. Our objective is to estimate the (vector or matrix) function $\Theta(\cdot)$ from knowledge of the sequences x_k, y_k, Φ_k.

One example where functional learning in dynamical systems can arise is in gain scheduling for an aircraft controller where the function input variables x_k are the speed and altitude of the aircraft and $f(x_k)$ is the gain schedule. Another possible application area is in robotics, [4], where x_k could be the position, and orientation, of the robot hand in space. In these two cases the parameters of the linear system are functionally dependent on the position. The optimal control is then also a function of position. The aim is to learn the control function given calculations at discrete points.

In Section 2, some theorems are reviewed about functional representation, and least squares convergence. In Section 3, the standard type of least squares method is applied to functional learning, and in Section 4 the idea of interpolating functions is exploited for this context. Section 5 has some simulation results and observations on practical implementation. In Section 6. areas that require further investigation are discussed and concluding remarks made.

2 Preliminary Definitions and Theorems

This section on functional representation. and least squares convergence can be used as reference material for some of the results in the paper.

2.1 Function Representation

For some of the results to follow we focus on representing a function as a sum of simply parametrized functions. termed here representation functions. Examples of such representation functions are sigmoids, and bisigmoids. The definition of these functions are now recalled.

Definition 2.1 *A scalar sigmoid function of a scalar variable t is one of the form*

$$\sigma(t) = \begin{cases} 1 & t \rightarrow \infty \\ 0 & t \rightarrow -\infty \end{cases}$$

This general definition does not require continuity, however the sigmoids we are interested in are continuous. An example of such a scalar sigmoid function is $\sigma(t) = (1 + e^{-t})^{-1}$.

Definition 2.2 *A scalar bisigmoid is the difference of two offset sigmoid functions with the property*

$$\sigma_b(t) = \sigma(t) - \sigma(t-1) = \begin{cases} 0 & t \rightarrow \infty \\ 0 & t \rightarrow -\infty \end{cases}$$

We are interested in integrable bisigmoids generated by a monotonic sigmoid. An example of such a scalar bisigmoid is $\sigma_b(t) = (1 + e^{-t})^{-1} - (1 + e^{-t+1})^{-1}$.

Another function that is of interest is the familiar Gaussian function with covariance Σ_i, assuming $|\Sigma_i| \neq 0$, is $\sigma(t) = (\sqrt{2\pi|\Sigma_i|})^{-1} \exp(-t'\Sigma_i^{-1}t/2)$.

A theorem about functional representations on a compact interval is now reviewed. This theorem gives conditions for approximating an arbitrary integrable function to an arbitrary accuracy using a given error measure. These conditions justify the use of continuous sigmoids and integrable bisigmoids as representation functions.

We use the notation that $\mathbb{R}$ is the set of real numbers, $\boldsymbol{N}$ the set of natural numbers. Consider

$$G : \mathbb{R} \rightarrow \mathbb{R}$$

Let us define

$$\sum^{r}(G) \quad = \quad \{g : g(x) = \sum_{j=1}^{q} \beta_j G(y_j' x + z_j); x, y_j \in \mathbb{R}^n, q \in \boldsymbol{N}, z_j, \beta_j \in \mathbb{R}\}$$

Theorem 2.1 *Denote the unit cube in $I\!R^n$ by I_n. If $G(\cdot) \in L^1$, the space of absolutely integrable scalar functions, and $\int_{I_1} G(t)dt \neq 0$, then $\sum^r(G)$ is dense in $L^1(I_n)$*

Proof the proof of this can be found in [6] □

Similar theorems are given in [6],[5] that give conditions for dense function representations over the space of continuous functions. An immediate consequence of this theorem is that sums of continuous integrable bisigmoid functions are dense, in the L^1 sense, and can approximate integrable functions over finite domains.

2.2 Least Squares Convergence

The theory of least squares gives a method of finding the constant coefficient θ of the equation

$$y_k = \Phi_k'\theta + w_k \tag{2.1}$$

where y_k is an m vector, Φ_k is an $r \times m$ matrix, θ is an r vector, and w_k is an m vector of white Gaussian noise, independent of Φ_k and θ. Here the task is to select θ as to minimize a weighted square of the error. That is to minimize with respect to ζ

$$V_k(\zeta) = \frac{1}{k}\sum_{i=0}^{k}(y_i - \Phi_i'\zeta)'W_i(y_i - \Phi_i'\zeta) \tag{2.2}$$

Where $W_k = W_k' > 0$ are the weighting matrices. The optimal ζ at time k, denoted $\hat{\theta}_k$ is given from the recursion

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_k\Phi_k W_k(y_k - \Phi_k'\hat{\theta}_{k-1}) \;\; ; \quad \hat{\theta}_0 \tag{2.3}$$

$$P_k^{-1} = P_{k-1}^{-1} + \Phi_k W_k \Phi_k' \;\; ; \quad P_0 = P_0' > 0 \tag{2.4}$$

where P_k is an $m \times m$ matrix, and with appropriate initial conditions.

Theorem 2.2 *Consider the weighted least squares algorithm (2.3) (2.4) applied to the signal model (2.1). Then, as $k \to \infty$, $P_k \to P_\infty$, $\hat{\theta}_k \to \theta_\infty$ a.s.. Consider also that θ is a random variable with a normal probability density function $N[\theta_0, P_0]$, and that the noise w_k is independent with a probability density function $N[0, W_k^{-1}]$, then the conditional distribution of θ_k given $y_1 \ldots y_k$ has mean $\hat{\theta}_k$ given by (2.3) and covariance P_k given by (2.4). Moreover, if $P_\infty = 0$ a.s., then $\lim_{k\to\infty} \hat{\theta}_k = \theta$ a.s..*

Proof The proof of this can be found in [7]. □

Remarks:

1. Actually, if the regression vector Φ_k is not influenced by the estimates $\hat{\theta}_k$, then the initial condition restriction in the theorem can be relaxed. as indeed can the interpretation of W_k^{-1} as a noise covariance. See [7].

2. Convergence rates for $\hat{\theta}_k$ are according to the convergence rates for P_k. Precise results on this can be found in [8] for the case when θ is not required to be a random variable. Thus with w_k a martingale increment process with bounded second moments.

$$\|\hat{\theta}_k - \theta\|^2 = O(\frac{\log(tr(P_k^{-1}))}{\lambda_{\min}(P_k^{-1})}) \qquad \text{a.s.} \tag{2.5}$$

   where $\lambda_{\min}$ denotes the minimum eigenvalue. Of course, if for all j, and some N,

$$\beta I > \frac{1}{N}\sum_{i=j}^{j+N} \Phi_i W_i \Phi_i' > \alpha I \qquad \text{a.s.} \tag{2.6}$$

   then $\|\theta_k - \theta\|^2 = O(k^{-1}\log k)$.

3. In the noise free case it can be shown that the convergence of $\hat{\theta}_k$ to θ is at least exponential when Φ_k satisfies (2.6).

3 Least Squares via Basis Functions (one dimensional problem)

3.1 The Signal Model

Here we examine a standard problem in (deterministic) approximation theory, in order to gain insights for the (stochastic) learning problem which is the focus of this paper. In particular, we work with basis function expansions and employ least squares parameter estimation for estimating the coefficients in a basis function expansion.

Consider for simplicity the square integrable functions

$$f : \Gamma_x \to \mathbb{R}, x \mapsto y = f(x) \tag{3.1a}$$

$$K_i : \Gamma_x \to \mathbb{R}, x \mapsto K_i(x) \tag{3.1b}$$

where $\Gamma_x \subset \mathbb{R}$. Let us investigate finite representations estimating $f(x)$ of the form

$$\hat{f}(x;\hat{Q}) = \sum_{i=1}^{n} K_i(x)\hat{q}_i = K_B'(x)\hat{Q} \tag{3.2}$$

where

$$\hat{Q} = [\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_n] \quad , \quad K_B'(x) = [K_1(x), K_2(x), \cdots, K_n(x)] \quad .$$

Here $K_i(\cdot)$ are known square integrable basis functions and $\hat{Q}$ is a parameter vector estimate.

We observe data points (x_k, y_k) generated as:

$$y_k = f(x_k) + w_k$$

where w_k is a sequence of white Gaussian noise independent of position x_k.

3.2 Measures of Error and Minimization Task

We consider now in what sense we wish the function representation to approximate the function. Let us work with a global measure of the error $f(x) - \hat{f}(x; \hat{Q})$, for all $x \in \Gamma_x$ under (3.1a), (3.2). An example of such a measure is

$$d_2(\hat{Q}) = [\int_{\Gamma_x} \|f(x) - \hat{f}(x; \hat{Q})\|^2 dx]^{\frac{1}{2}} \tag{3.3}$$

which is the mean square error measure. With $f(x_k)$ available only at a discrete set of points $x_k \in \Gamma_x$, it makes sense to consider a restricted measure of the mean square error as

$$d_2^{(r)}(\hat{Q}) = \frac{1}{r}[\sum_{k=1}^{r} \|f(x_k) - \hat{f}(x_k; \hat{Q})\|^2]^{\frac{1}{2}} \tag{3.4}$$

In approximating functions (3.1a) by function representations (3.2), the minimization task we focus on is as follows

$$\min_{\hat{Q}} d_2(\hat{Q}) \tag{3.5}$$

or the closely related index

$$\min_{\hat{Q}} d_2^{(r)}(\hat{Q}) \tag{3.6}$$

Remarks

1. It is really the error measure $d_2(Q)$ that is of interest, because this gives a measure of the error at both the points that have been visited and those for which a function estimate is given. In any application only measurements at a finite set of points are available so $d_2^{(r)}(Q)$ is the only realistic error measure to work with. In the situation that $f(x)$ is smooth and the points x_k are chosen in a uniformly dense way, then standard calculus theory tells us that the $d_2^{(r)}$ error measure approaches the d_2 error measure.

2. Another example of an error measure which is appropriate in some situations is

$$d_\infty(\hat{Q}) = \max_{\Gamma_x} \|f(x) - \hat{f}(x; \hat{Q})\|$$

There is in fact a whole family of possible error measures of the form

$$d_p(\hat{Q}) = [\int_{\Gamma_x} ||f(x) - \hat{f}(x;\hat{Q})||_p dx]^{\frac{1}{p}}$$

which may have merit for particular applications. In the sequal however we are concerned only with the d_2 error measure.

3. The error measure only considers the functional representation on the region Γ_x. It will be dependent on the aplication as to whether values should be truncated outside this region or not.

3.3 Allowable Basis Functions and Reconstructability

If one function $f(x)$ is to be represented as a sum of other functions, it is necessary that the possible function summations $\hat{f}(x, Q)$ be sufficiently rich to allow a reasonable approximation. Representation theorems like 2.2 are important in giving conditions as to what functions can be used in such representations. There are obvious disadvantages if there exist $Q_1 \neq Q_2$ such that $\hat{f}(x, Q_1) = \hat{f}(x, Q_2)$ for all $x \in \Gamma_x$. It is also necessary that the measurements that are used to choose the function representation are sufficiently rich to characterize the behaviour of the function being approximated. There is a need in some of the theory to follow for restrictions on the function representations as well as on the class of function that is estimated. Of particular interest are allowable basis function representations and the class of reconstructable functions.

Definition 3.1 *The set of square integrable basis functions $K_B(x)$ is termed* allowable *if and only if*

$$\infty > [\int_{\Gamma_x} K_B(x)K'_B(x)dx] > 0 \tag{3.7}$$

Definition 3.2 *The function $f(x)$ is said to be reconstructable if it is in the model set of functions $\hat{f}(x;\hat{Q})$ of (3.2) that is*

$$f(x) = K'_B(x)Q \quad \textit{for some vector } Q. \tag{3.8}$$

Theorem 3.1 *The minimization task (3.5) under (3.1a) (3.2) has a unique critical point if and only if the elements of $K_B(x)$ are allowable. This optimal $\hat{Q}$, denoted $\hat{Q}^*$, is given by*

$$\hat{Q}^* = (\int_{\Gamma_x} K_B(x)K'_B(x)dx)^{-1} \int_{\Gamma_x} f(x)K'_B(x)dx \tag{3.9}$$

Moreover, when $f(x)$ is reconstructable with respect to the class of functions $\hat{f}(x;\hat{Q})$ of (3.2), then $f(x)$ is uniquely parametrized as in (3.8) with $Q = \hat{Q}^*$ given in (3.9).

Proof Consider the minimization of d_2 under (3.1a) (3.2) as in (3.8). Upon differentiation, it is evident that any critical point must satisfy

$$-2\int_{\Gamma_x} K_B(x)[f(x) - K_B'(x)\hat{Q}]dx = 0\,. \tag{3.10}$$

The critical point is unique if and only if (3.7) holds and is given by (3.9). Under (3.8), $Q = \hat{Q}^*$. □

Remarks:

1. If $K_B(x)$ is not allowable, then there will be an infinite number of critical points of the minimization.
2. As n increases, the class of reconstructable $f(x)$ becomes larger. In order to represent an arbitrary function with arbitrarily small error, it is necessary that n approach infinity.
3. For $f(\cdot)$ known to be frequency band limited in a spatial sense, suitable choices of K_i are

   $$K_i(x) = \begin{cases} \sin(\frac{i}{2}x) & i \text{ even} \\ \cos(\frac{i-1}{2}x) & i \text{ odd} \end{cases}$$

   For $f(\cdot)$ known to be a polynomial of degree less than or equal to some fixed value, an appropriate choice would be

   $$K_i(x) = x^i$$

Definition 3.3 *The set of ppoints x_k is sufficiently rich on $K_B(\cdot)$ if for all k, $\hat{f}(x_k, Q_1) = \hat{f}(x_k, Q_2)$ then $Q_1 = Q_2$.*

This is an obvious discretization of the condition that Q is uniquely determined. A necessary and sufficient condition to guarantee that x_k is sufficiently rich is that

$$\sum_{k=j}^{\infty} K_B(x_k)K_B'(x_k) > 0 \tag{3.11}$$

A stronger condition is that there exists an N such that for all $j > 0$

$$\delta I > \frac{1}{N}\sum_{k=j}^{j+N} K_B(x_k)K_B'(x_k) > \eta I > 0 \tag{3.12}$$

for some $\delta, \eta > 0$. This condition is termed persistence of excitation, and means that in every set of N measurements there is sufficient information to choose a unique Q, thus giving fast learning. Observe that $K_B(x)$ being allowable is a sufficient condition for the existence of such persistently exciting sequences.

3.4 Recursive Least Squares Algorithm

In order to minimize $d_2^{(r)}$ of (3.4) for $r = 1, 2, \ldots$, given a sequence $\{x_k, y_k\}$, standard least squares derivations leads to a recursive estimate of Q, denoted $\hat{Q}_k$, as

$$\hat{Q}_k = \hat{Q}_{k-1} + P_k K_B(x_k)[y_k - K_B'(x_k)\hat{Q}_{k-1}] \tag{3.13}$$

$$P_k^{-1} = P_{k-1}^{-1} + K_B(x_k)K_B'(x_k) \tag{3.14}$$

with suitable initial conditions $\hat{Q}_0$, $P_0 = P_0' > 0$.

Theorem 3.2 *Consider that K_B is allowable as defined in (3.7), and $f(\cdot)$ is reconstructable with regard to $\hat{f}(\cdot;\cdot)$ of (3.2). Then provided the P_k as defined in (3.14) approach zero as $k \to \infty$, the parameter estimates $\hat{Q}_k$ of (3.13) converge as*

$$\lim_{k\to\infty} \hat{Q}_k = \hat{Q}^* \text{ a.s.} \tag{3.15}$$

If the persistence of excitation condition (3.12) is satisfied then

$$tr(P_k), \lambda_{\min}(P_k) = O(k) \qquad \text{a.s.} \tag{3.16}$$

and

$$||\hat{Q}_k - \hat{Q}^*||^2 = O(k^{-1}\log k) \qquad \text{a.s.} \tag{3.17}$$

Proof Standard least squares theory of Theorem 2.2 applies □

Remarks:

1. The condition (3.12) can be seen to correspond to the continuous time persistence of excitation condition (3.3) in [4].

2. What happens if $f(x)$ is not reconstructable but K_B is allowable? There is a reconstructable $f^*(x)$ that is closest in mean square to $f(x)$. The difference between $f(x)$ and $f^*(x)$ is orthogonal to $K_B(x)$ and hence the learning of $f^*(x)$ from y_k is covered by Theorem 3.2.

3. In the non persistence of excitation case, where

$$P_\infty = (\sum_{k=0}^{\infty} K_B(x_k)K_B'(x_k))^{-1} = 0$$

the algorithm still converges with a rate given "loosely" by the rate of convergence of P_k to 0.

4. Of course, by monitoring P_k it would become clear if $P_k \not\to 0$. To achieve convergence more excitation of x_k is required. In any practical applications, persistence of excitation could be a difficult property to ensure a priori. (See Remark 2 following Theorem 2.2)

5. In applying the basis function approach, as above, to the signal model (1.2), there are two possible approaches. The first is, in the case when $\Phi_k = \Phi(x_k)$, to estimate the product $\Phi'(\cdot)\Theta(\cdot)$ as the unknown function, ignoring the fact that Φ_k is known. This would be particularly unattractive if $\Theta(\cdot)$ is a simple function and Φ_k is not. The second approach is to introduce Φ_k into the analysis replacing $K_B(x)$ by $K_B(x) \otimes \Phi_k$, where $\otimes$ denotes the Kronecker product. Of course, then $\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} (K_B(x_k) \otimes \Phi_k)(\Phi_k \otimes K_B(x_k))'$ will not generally be diagonal. Consequently, there is no particular advantage to work with orthogonal $K_B(\cdot)$. This second approach is developed further in the next section.

4 Interpolation Functions in Least Squares

4.1 Signal Model

Consider now a method for identifying signal models (1.2) using interpolation function representations for $\Theta(x)$. Thus $\Theta(x)$ is approximated as

$$\hat{\Theta}(x; \hat{Q}) = \sum_{i=1}^{n} K_i(x)\hat{q}_i \tag{4.1}$$

with $\hat{Q}' = [\hat{q}_1' \cdots \hat{q}_n']$. Here $\hat{q}_i$ is an m vector, $\hat{Q}$ is an mn vector and $K_i(x)$ is a scalar function of x. Here $\Gamma_I = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ is a preselected set of points in Γ_x, and we work with $K_i(x)$ as a scalar interpolating function between the points in Γ_I and those in Γ_x. Of course, one could specialize $K_i(x)$ to be orthogonal basis functions so that (4.1) is a basis function expansion, and build on the methods of the previous section. Here we prefer to think of $\hat{q}_i$ as close to $\Theta(\gamma_i)$ so that (4.1) allows an interpolation for $x \notin \Gamma_I$. Given $K_i(\cdot)$, and estimates of $\hat{q}_i$, then $\hat{\Theta}(x; \hat{Q})$ can be evaluated at any x using (4.1).

4.2 Reconstructability

Under reconstructability of $\Theta(x)$ as a function $\hat{\Theta}(x; \hat{Q})$ of the form given in (4.1), then for some parameter vector $\hat{Q}$, denoted Q^*, $\hat{\Theta}(x; \hat{Q})$ satisfies $\hat{\Theta}(x; Q^*) = \Theta(x)$. Thus in the case that $K_i(x)$ are integrable bisigmoids suitably shifted by affine maps, Theorem 2.1 tells us that as n becomes infinite, the class of functions (4.1) are dense in the space of continuous functions. Now under reconstructability, (1.2) can be written as

$$y_k = \Phi_k \sum_{i=1}^{n} K_i(x)q_i^* + w_k = \Phi_I'(x_k)Q^* + w_k \tag{4.2}$$

where [with scalar $K_i(\cdot)$]

$$\Phi_I'(x_k) = [K_1(x_k)\Phi_k' \;\; K_2(x_k)\Phi_k' \;\cdots\; K_n(x_k)\Phi_k'] = K_I(x_k) \otimes \Phi_k \tag{4.3}$$

$$K_I(x_k) = [K_1(x_k) \;\; K_2(x_k) \;\cdots\; K_n(x_k)]$$

Here $\Phi_I(\cdot)$ is known. and Q^* is to be estimated.

4.3 Allowable Interpolation Functions

Using the methodology of Section 3.3 we find conditions that allow unique identification of Q. This requires conditions on both the basis function K_i and the data sequence, as in Section 3.3.

The class of allowable $K_I(\cdot)$ is equivalent to the class of allowable $K_B(\cdot)$.

The condition for unique identifiability using discrete measurements requires now that

$$\sum_{k=0}^{\infty} \Phi_I(x_k)\Phi_I'(x_k) > 0$$

which is dependent on both the state domain x_k trajectory in Γ_x and the time domain regression vector Φ_k. It is not immediately clear how to interpret this excitation condition when excitation in both the time domain and the state domain are involved. One way to indicate the difference between Φ_k and x_k is to use time scale separation.

Definition 4.1 *Suppose there is given a continuous function $K_I(x)$ with a Lipshitz constant c, such that $0 \leq K_i(x) \leq 1$, $\int_{\Gamma_x} K_I(x)K_I'(x)dx > aI$, and a sequence Φ_k. Then the transformation $T(x_k)$ is said to be* slowly varying *with respect to Φ_k, and K_I, if $\exists$ an $\epsilon < \frac{1}{5}\eta\alpha\beta^{-1}c^{-1}$, N, α, β such that for all k*

$$\|T^l x_k - x_k\| \leq \epsilon \;\; \forall l \in \{0, 1, \cdots, N\}$$

$$\text{and} \qquad \beta I > \frac{1}{N}\sum_{k=l}^{l+N-1} \Phi_k\Phi_k' > \alpha I > 0 \qquad \text{hold.} \tag{4.4}$$

Theorem 4.1 *Assume that $bI > \int_{\Gamma_x} K_I(x)K_I'(x)dx > aI > 0$. and Φ_k satisfies (4.4). Assume also that $\{x_k\}_0^\infty$ is given by $x_k = T(x_{k-1})$ where T is a mapping from Γ_x to Γ_x such that x_k satisfies (3.12), and is slowly varying with respect to Φ_k, K_I. Then $\Phi_I(\cdot)$ satisfies*

$$\beta I > \frac{1}{S}\sum_{k=l}^{l+S-1} \Phi_I(x_k)\Phi_I'(x_k) > \delta I > 0 \tag{4.5}$$

for some finite S, β, δ and all l.

Proof By the definition of Φ_I, then simple manipulations give,

$$\frac{1}{S}\sum_{k=0}^{S-1} \Phi_I(x_k)\Phi_I'(x_k) = \frac{1}{S}\sum_{k=0}^{S-1} K_I(x_k)K_I'(x_k) \otimes \Phi_k\Phi_k'$$

$$= \frac{N}{S}\sum_{i=0}^{S/N} K_I(x_{iN})K_I'(x_{iN}) \otimes \frac{1}{N}\sum_{k=iN}^{(i+1)N-1} \Phi_k\Phi_k' + R$$

where the remainder can be overbounded by $|R| \leq 2\epsilon c\beta$. Because x_k are persistently exciting there exists a finite S such that $\sum_{i=0}^{S} K_I(x_i)K_I'(x_i) > \eta I > 0$. Thus

$$\frac{N}{S}\sum_{i=0}^{\frac{S}{N}} K_I(x_{iN})K_I'(x_{iN}) > \eta I - 2\epsilon c$$

Hence

$$\frac{1}{S}\sum_{k=0}^{S-1} \Phi_I(x_k)\Phi_I'(x_k) \geq aI \otimes \alpha I - 4\epsilon c\beta I \otimes I > \frac{1}{5}\eta\alpha I \otimes I$$

Hence there exists a finite S such that for all j

$$\sum_{k=j}^{j+S} \Phi_I(x_k)\Phi_I'(x_k) > \delta I$$

The proof for the upper bound is similar. □

Remarks:

1. The condition (4.5) can be seen to parallel the continuous time persistence of excitation condition (3.3) in [4].
2. One method of ensuring that this condition is satisfied is to fix x_k for N iterations while Φ_k spans the space. Then the $\hat{Q}$ need only be updated every Nth iteration.
3. It is possible to relax the condition that T be slowly varying. This may be seen by rearranging the ordering of finite groups of samples so that the reordered samples are slowly varying. That this is allowable follows from the uniform convergence of the sample means.
4. As the number of interpolating functions, n, tends to infinity the size of the vector $K_I(\cdot)$ will tend to infinity, but it is always rank 1. Noting that $S \geq n$ then S must tend to infinity in order to satisfy condition (4.5). Thus persisance of excitation is unrealistic.

4.4 Least Squares Algorithm

The standard least squares recursions associated with (4.2) are

$$\hat{Q}_k = \hat{Q}_{k-1} + P_k(\Gamma_I)\Phi_I(x_k)[y_k - \Phi_I'(x_k)\hat{Q}_{k-1}] \tag{4.6}$$

$$P_k^{-1}(\Gamma_I) = P_{k-1}^{-1}(\Gamma_I) + \Phi_I(x_k)\Phi_I'(x_k) \tag{4.7}$$

At any time k, the signal model parameter $\Theta(x_k)$ can be estimated using (4.1).

Theorem 4.2 *Consider that $K_I(\cdot)$ is allowable, $\Theta(\cdot)$ is reconstructable as a function $\hat{\Theta}(x; Q)$ of the form given in (4.1). Consider also that in (4.7), P_k approaches zero as $k \to \infty$. Then in (4.6)*

$$\lim_{k\to\infty} \hat{Q}_k = Q \text{ a.s..} \tag{4.8}$$

Furthermore, if (4.5) holds then

$$tr(P_k), \lambda_{\min}(P_k) = O(k) \qquad \text{a.s.} \tag{4.9}$$

and

$$||\hat{Q}_k - Q||^2 = O(k^{-1} \log k) \qquad \text{a.s..} \tag{4.10}$$

Proof This follows the proof of Theorem 3.2 □

Remarks

1. If (4.5) does not hold, this algorithm can be implemented with a check on P_k to watch for convergence. If P_k is not going to zero, x_k must be further excited. It may be that there is little learning of the function $\Theta(x)$ in the vicinity of a subset of Γ_I. Then it makes sense to select x_k trajectories in the vicinity of that subset.

2. If the $K_i(x)$ are chosen to be bisigmoids, generated by monotonic sigmoids, centered on γ_i then straightforward reasoning shows that $P_m^{-1} = \sum_{k=1}^{m} \Phi_I(x_k)\Phi_I'(x_k)$ is diagonally dominant. (Each Φ_k has one element that is greater than the others, and decreases symmetrically away from this element, hence $\Phi_I(x_k)\Phi_I'(x_k)$ is diagonally dominant.) Using this approach q_i is a first approximation of $\Theta(\gamma_i)$. Also a new measurement pair (x_k, y_k) primarily updates the q_i for which x_k is near γ_i, and has a diminishing effect as $|x_k - \gamma_i|$ increases.

3. Following on from Remark 2, with an appropriately truncated K_I, then P_m^{-1} is diagonal, and $q_i = \Theta(\gamma_i)$ for all γ_i. Certain γ_i selection and appropriate truncation could lead to P_m^{-1} being (say) tridiagonal. Diagonal, tridiagonal, or such truncation of P_m would then lead to computational savings at the expense of introducing limits to spreading the learning and the interpolation.

4. Remark 2 suggests that for bisigmoid representations even in the absence of any K_I truncation, by using only the diagonal part of P_k, or tridiagonal part (say), the computational effort will be reduced with some loss in spread of learning, but not in interpolation spread. The accuracy of such an approach is dependent on the "width" of the function K_I. We do not present here any theory for this case when the K_I are not truncated, but P_k is diagonalized. Simulation results in Section 5 support the proposed method for computational effort reduction.

5. In neural networks, nonlinear functions are represented as sums of sigmoid functions, suitably biased, which are dense in function space. One might think that it is reasonable for K_I to be chosen to be offset sigmoids. Remarks 2 and 3 above do not apply with this choice of interpolation function, nor is there physical meaning to the parameter q_i. We do not explore such selections further.

6. It can be seen that when there is only one γ_i and $K_i(x) = 1$, that is, $Q = q_1, \Theta(x) = Q$, then the algorithm collapses to the standard least squares parameter estimation algorithm.

7. With the choice of $K_i(\cdot)$ as

$$K_i(x_k) = \left\{ \begin{array}{ll} 1 & \text{if} \quad \|x_k - \gamma_i\| < \|\gamma_i - \gamma_j\|/2 \ \forall \, \gamma_j \neq \gamma_i \\ 0 & \qquad\qquad\qquad \text{otherwise} \end{array} \right\} \tag{4.11}$$

then only one of the Φ_I are nonzero and the basis function algebra is recovered. (The basis function is a rectangular pulse of height 1). In this case P_k is block diagonal and the computational effort is minimal as only one of the q_i are updated at each iteration. Such a truncated interpolation function as (4.11) effectively decides which γ_i neighbourhood a measurement is in, and then upgrades the associated q_i estimate with a stepsize which is independent of the "distance" from x_k to γ_i within the neighbourhood of γ_i.

8. When there is only partial excitation of the region Γ_x there can still be some useful results. If the region $\Gamma' \subset \Gamma_x$ is persistently excited while the whole of the region Γ_x is not excited then there is no unique estimate of the function over the region Γ_x but there is a unique function value representation on Γ'.

5 Numerical Simulation

Consider the reconstructable system (1.2) where

$$\Theta(x) = 2K_1(x) + 3K_2(x) + 2K_3(x) + K_4(x)$$

$K_i(x) = e^{-64(x-\gamma_i)^2}$, $\Gamma_x = [0,1]$, and $\gamma_i = \frac{i-1}{3}$. Figure 5.1 shows the time evolution of the parameter estimates $\hat{Q}$ when the least squares recursion (4.6) is used. It can be observed that as the theory predicted the parameter estimates converge to the true value. We suggested earlier that calculations could be simplified in the case of K_i being bisigmoid by considering only the diagonal elements of P. Figure 5.2 shows the evolution of the parameter $\hat{Q}$ when the suboptimal version of (4.6), taking only the diagonal part of P_k, is used. This example demonstrates the marginally slower response expected using the diagonalized algorithm (performance can be expected to be sacrificed since the calculations are simplified) compared to the full algorithm.

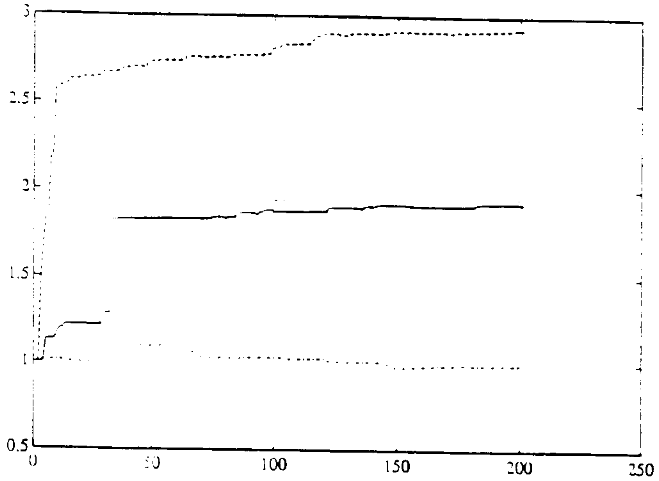

Figure 5.1: Parameter estimation for a reconstructable system using (4.6)

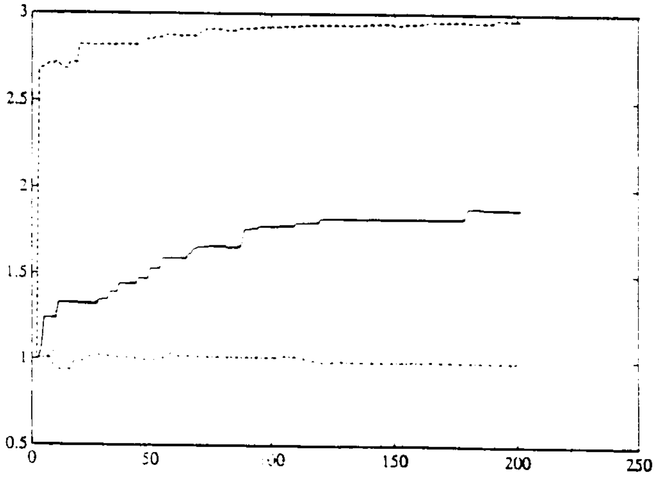

Figure 5.2: Parameter estimation for a reconstructable system using (4.6) with diagonalized P_k

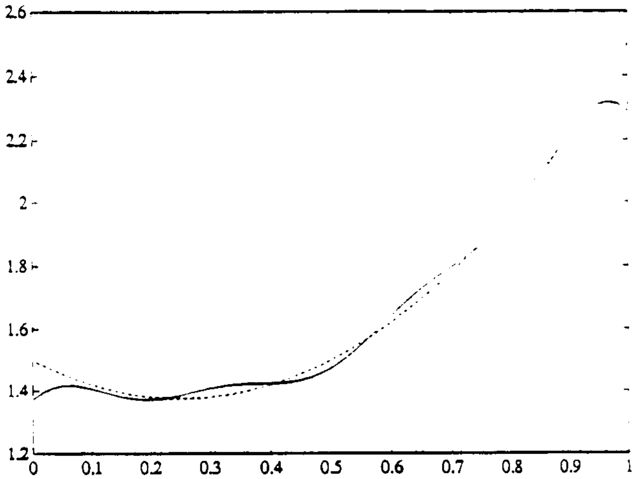

Figure 5.3: Parameter estimation in the case when the parameter is not reconstructable

We now consider an example where the function to be learnt is not reconstructable. Figure 5.3 shows a typical result of estimating the parameter function $\Theta(x) = 1.5 + 2x^2 - x$ of an ARMAX model when the parameter function is not reconstructable. In this case Φ_k is taken to be a uniformly distributed random number between 0 and 1. The noise term is neglected. There are 4 equally spaced Gaussian interpolating functions, located on the boundary and interior of $\Gamma_x = [0,1]$ at $\gamma_i = \frac{i-1}{3}$, each one of the form $K_i(x) = e^{-20(x-\gamma_i)^2}$. The recursion proceeded for 100 iterations. Notice that the final estimates are reasonably accurate, that is we converge to the best least squares estimate. Figure 5.4 shows the time evolution of the parameter estimate for this set of data. Notice the bursts in learning according to the excitation. It can be seen that the algorithm learns well despite the lack of reconstructability.

Computer simulations have shown the importance (when functions are not reconstructable) of choosing appropriate interpolation functions. Too wide an interpolation leads to a blurring of detail, while too narrow an interpolation leads to "egg-carton" estimates. Figures 5.6 to 5.8 demonstrate this when estimating $\Theta(x) = xx'$ as the sum of sixteen bisigmoid, and can be compared to Figure 5.5 which shows the actual value of $\Theta(x)$. In these simulations we have selected $K_i(x) = e^{-a16(x-\gamma_i)^2}$ where a is set to 1, 3, 0.05 respectively. An estimate of the d^2 error is 10.31, 0.4278, 26.44 respectivly. For simplicity the noise sequence in these simulations has been set to zero. Thus although non-reconstructable functions can be considered the nature of the interpolating needs to be considered in order to obtain a reasonable approximation.

If finer structure is required it is suggested that extra γ_i can be introduced while reducing the spread of K_I. A sensible initial value for the associated q_i would be the previous predicted value of $\Theta(\gamma_i)$. This can be seen in Figure 5.9 where an estimate of $x^2 - (x-2)^{-1}$ is made using 4 and 8 γ_i. The inverse

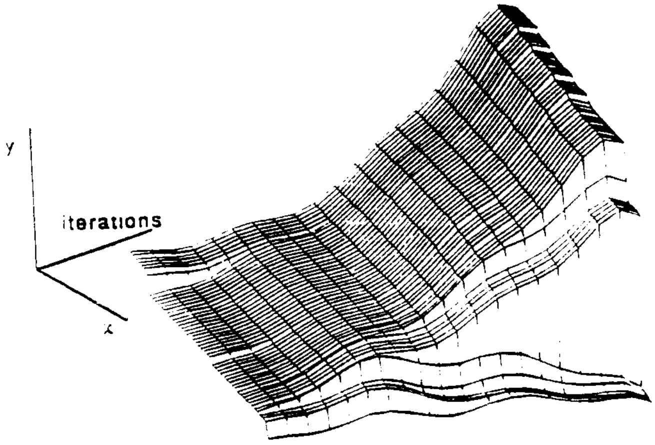

Figure 5.4: Parameter function evolution

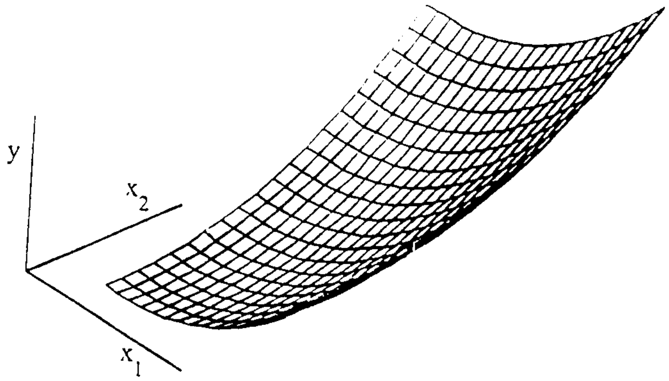

Figure 5.5: The parameter function to be estimated on the region Γ_x.

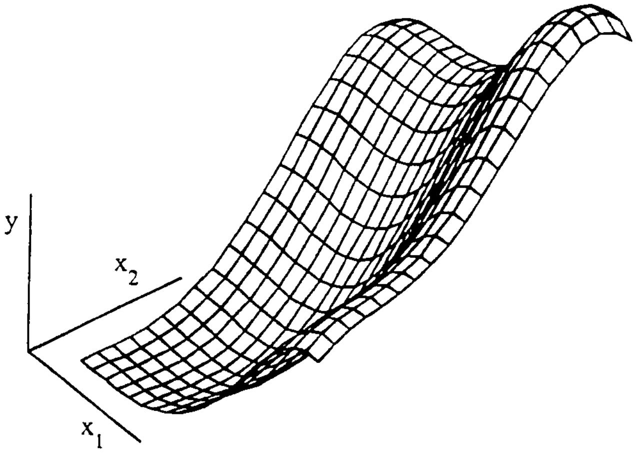

Figure 5.6: Parameter function estimation in the case when the K_I are chosen to give an even coverage of the region Γ_x

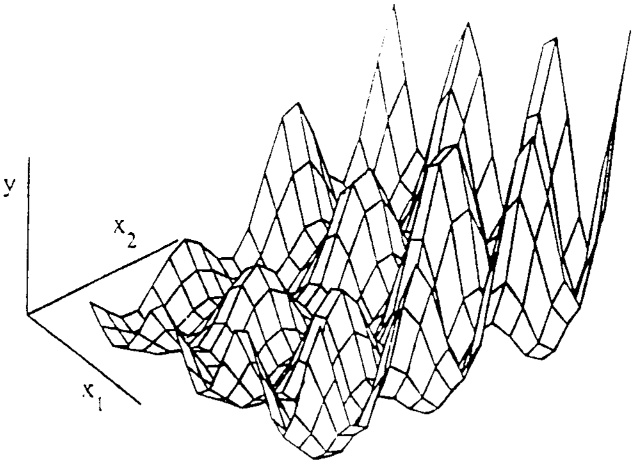

Figure 5.7: Parameter function estimation in the case when the K_I are chosen too narrow to cover the region Γ_x

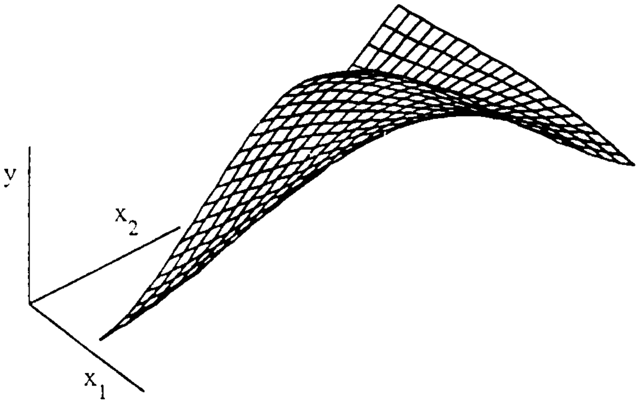

Figure 5.8: Parameter function estimation in the case when the K_I are chosen too broad to resolve information in the region Γ_x

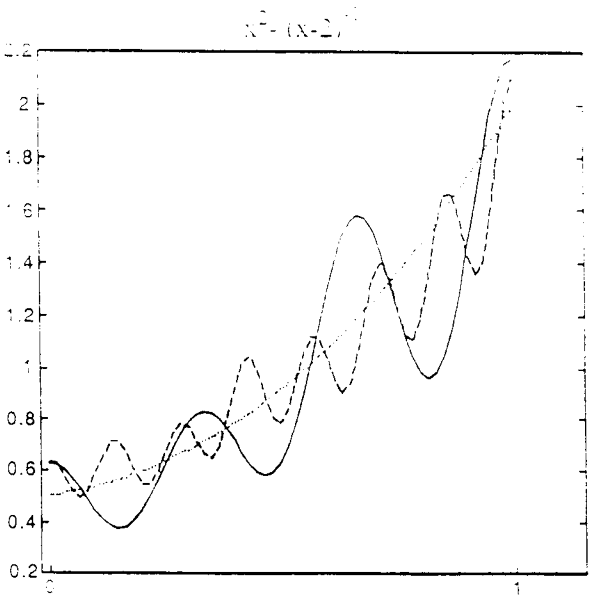

Figure 5.9: Comparison of parameter function estimation using 4 and 8 γ_i

variance of the interpolating Gaussian was chosen to be 3 times the square of the number of γ_i. This increase in number of interpolating functions corresponds to increasing the size of the class of reconstructable functions and thus decreasing the necessary error.

The positioning of the interpolating functions influences the precision of the function estimation in the case where the function is not reconstructable. If the γ_i are uniformly distributed in the domain and the K_I are fixed bisigmoids then edge effects are observed, as shown in Figure 5.10, which estimates the same surface as Figure 5.6 but with γ_I now uniformly distributed over the interior of the region. This can be prevented by placing γ_i on the edge of the domain as was done in in the previous figures, thus preventing the edge bisigmoids from covering a larger region than the interior bisigmoids.

6 Conclusion

We have shown how a least squares algorithm or a system of such can be applied in functional learning. Crucial to the success of the algorithms is the selection of interpolation functions, not only to interpolate between parameter estimates at a set of points in the function space, but also to spread learning from the data to achieve estimates at the set of points in question. Convergence properties of this algorithm for stochastic models are established using standard least squares results. The results here have been developed for ARMAX models with coefficients being functionals of some input variables. Simulation studies have shown various trade offs in the selection of the interpolation function expansions. There are still open questions concerning optimization of the choice of interpolation functions, and guaranteeing identifiability in any practical application.

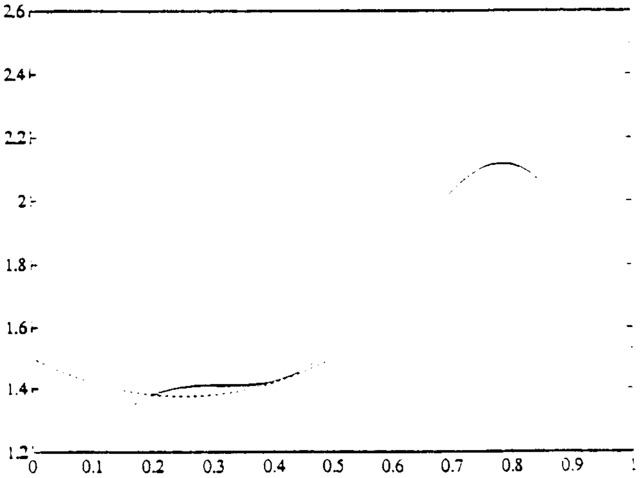

Figure 5.10: Parameter function estimation when there are no γ_i located on the boundary of Γ_x

References

[1] S. Singhal, L. Wu, "Training multilayer perceptrons with the extended Kalman algorithm", Advances in neural information processing system I

[2] W. Messner, R. Horowitz, W. Kao, M. Boals, "A new adaptive learning rule", IEEE Trans AC to appear.

[3] R. Horowitz, W. Messner, J.B. Moore, "Exponential convergence of a learning controller for robot manipulators", IEEE Trans AC to appear.

[4] J.B. Moore, R. Horowitz, W. Messner, "Functional persistence of excitation and observability", Proc ACC conference San Diege May 1990.

[5] M. Stinchcome, H. White, "Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions", Neural Networks 1989

[6] G. Cybenko, "Approximation by superpositions of a sigmoidal function", Mathematics of Control, Signals, and Systems 1989

[7] J. Sternby, "On consistency for the method of lest squares using Martingale theory", IEEE Transactions on Automatic Control, AC-22 1977, p346

[8] H. Chen, L. Gou, "Convergence rates of least square identification and adaptive control for signal processing", Int. J. Control, Vol 44. 1986, p1459-1476

(b) Let $\Phi : M \to \mathbb{R}$ be a Morse-Bott function on a Riemannian manifold M. Then the ω-limit set $L_\omega(x)$, $x \in M$, for the gradient flow (5) is a single critical point of Φ. Every solution of the gradient flow (5) converges as $t \to +\infty$ to an equilibrium point.