Pergamon

0005-1098(94)00131-6

Automatica, Vol. 31, No. 3, pp. 503–508, 1995

0005-1098/95 $9.50 + 0.00

Brief Paper

Worst-case Identification of Nonlinear Fading Memory Systems*

MUNTHER A. DAHLEH,† EDUARDO D. SONTAG,‡ DAVID N. C. TSE† and JOHN N. TSITSIKLIS†

Key Words—System identification; nonlinear systems; fading memory.

Abstract—In this paper, the problem of asymptotic identification for fading memory systems in the presence of bounded noise is studied. For any experiment, the worst-case error is characterized in terms of the diameter of the worst-case uncertainty set. Optimal inputs that minimize the radius of uncertainty are studied and characterized. Finally, a convergent algorithm that does not require knowledge of the noise upper bound is furnished. The algorithm is based on interpolating data with spline functions, which are shown to be well suited for identification in the presence of bounded noise—more so than other basis functions such as polynomials.

1. *Introduction*

Recently, there has been increasing interest among the control community in the problem of identifying plants for control purposes. This generally means that the identified model should approximate the plant in the operator topology, since this allows the immediate use of robust control tools for designing controllers (Dahleh and Khammash, 1990; Doyle, 1982). This problem is of special importance when the data are corrupted with bounded noise. The case where the objective is to minimize the prediction error for a fixed input has been analyzed by many researchers (see Ljung, 1987, and references therein). The problem is more interesting when the objective is to approximate the original system as an operator, a problem extensively discussed by Zames (1979), especially when the plant's order is not known a priori. For linear time-invariant plants, such an approximation can be achieved by uniformly approximating the frequency response (H_∞ norm) or the impulse response (ℓ_1 norm). In H_∞ identification, it was shown that robustly convergent algorithms can be furnished, when the available data is in the form of a corrupted frequency response, at a set of points dense on the unit circle (Helmicki *et al.*, 1991; Gu *et al.*, 1992; Gu and Khargonekar 1992). For the topology induced by the ℓ_1 norm, a complete study of asymptotic identification was furnished by Tse *et al.* (1993) for arbitrary inputs, and the question of optimal input design was addressed. Most work on input design has been done in stochastic settings (see e.g. Mehra, 1974; Zarrop, 1979), but recently there have also been some results in worst-case settings (Kacewicz and Milanese, 1992; Makila, 1991). Related work on the worst-case identification problem was also reported by Makila and Partington (1991), Poolla and Tikku (1994), Dahleh *et al.* (1993) and Jacobson and Nett (1991).

In this paper, the work of Tse *et al.* (1993) is extended to analyse the worst-case asymptotic identification of nonlinear fading memory systems. As in Tse *et al.* (1993), the study is done in two steps. The first is concerned with obtaining tight upper and lower bounds on the optimal achievable error by *any* identification algorithm. The bounds are functions of the input used for the experiments, and this can be arbitrary. The second step is then to study these bounds and characterize the inputs that will minimize them. In particular, simple topological conditions are furnished that guarantee the existence of an algorithm with a worst-case error within a factor of two from the lower bound. A near-optimal input is characterized so that the worst-case error is within a factor of two of the bound on the noise.

It is noted that for the results on arbitrary experiments, the suggested optimal algorithms are tuned to the knowledge of the bound on the noise. If, however, the near-optimal input is used, then an untuned algorithm can be provided that results in a worst-case error equal to the noise bound, δ. Such an algorithm is based on interpolating data by spline functions of several variables.

The rest of the paper is organized as follows: Section 2 gives a formal definition of nonlinear fading memory systems. Section 3 describes the identification set-up. Section 4 characterizes the asymptotically optimal algorithms and the associated optimal worst-case errors for a given input. The problem of optimal inputs is addressed in Section 5. An optimal untuned algorithm is developed in Section 6. Section 7 contains our conclusions.

2. *Fading memory systems*

Let $\mathcal{U}$ be the set of one-sided infinite sequences whose ℓ_∞ norm is bounded by 1. This can be viewed as the input set that contains the possible inputs that can be used for performing the identification experiments. We consider the set of models $\mathcal{K}$ as discrete-time, causal functions from $\mathcal{U}$ to $\mathbb{R}^\infty$; a plant $h \in \mathcal{K}$ takes as input a sequence $u = (u_0, u_1, \ldots)$ to give an output sequence $(h_0(u), h_1(u), \ldots)$. We assume that $h \in \mathcal{K}$ further satisfies the following properties:

(1) $h_n(u)$ depends continuously on $u_0, \ldots, u_{n-1}$;

(2) h has equilibrium-initial behavior:

$$h_{n+1}(0u) = h_n(u) \quad \forall n,$$

where $0u$ is the input $0, u_0, u_1, \ldots$. (In general, we shall use the notation vw for concatenation, i.e. first apply the finite sequence v, then w. Since we are dealing with causal systems, we shall slightly abuse the notation and write $h_n(w)$ to mean $h_n(u)$, where u is any infinite sequence whose first n elements are given by the finite sequence w.); and

(3) h has fading memory (FM): for each $\varepsilon > 0$, there is some $T = T(\varepsilon)$ such that for every k, every $t \geq T$ and every finite sequence $v \in [-1, 1]^k$, $w \in [-1, 1]^t$

$$|h_{t+k}(vw) - h_t(w)| < \varepsilon.$$

* Received 2 December 1992; revised 23 August 1993; received in final form 1 June 1994. This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Paul van den Hof under the direction of Editor T. Söderström. Corresponding author Professor M. A. Dahleh. E-mail dahleh@lids.mit.edu.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

‡ Department of Mathematics, SYCON, Rutgers University, New Brunswick, NJ 08903, U.S.A.

To measure the identification error, we shall use the operator gain

$$\|h\|_{\mathscr{X}} = \sup_{u \in \mathscr{U}} \|h(u)\|_\infty.$$

It can be seen from the defining properties that systems in $\mathscr{X}$ must necessarily have bounded operator gain. This is a good norm to consider for robust control applications. However, it should be noted that this norm is different from the standard definition of the gain of a nonlinear operator, which is readily suitable for robust control applications. For the above induced norm to be useful, an upper bound on the amplitudes of input signals has to be known a priori. In the above definition, this bound is normalized to one.

2.1. *Examples of FM systems*

2.1.1. *Stable LTI systems.* For each $h \in \ell_1$, consider the input/output map $u \mapsto u * h$. It is clear that these systems satisfy the above conditions. The operator-induced norm in this case is just the ℓ_1 norm.

2.1.2. *Hammerstein systems.* These are systems formed by composition of a stable LTI system followed by a memoryless nonlinear element:

$$y_n = g((u * h)_n)$$

for some $h \in \ell_1$ and some continuous functions $g: \mathbb{R} \to \mathbb{R}$. It is easy to verify that these systems satisfy the first two conditions above. If we assume further that g is uniformly continuous then it can be seen that the system also has fading memory.

For further details on fading memory operators, see Boyd and Chua (1985) and Shamma and Zhao (1990).

3. *Identification set-up*

The plant to be identified is known to be in a model set $\mathscr{M} \subset \mathscr{X}$. An input u is selected from the set $\mathscr{U}$. We assume that the observed output y is corrupted by some additive disturbance d that is unknown but magnitude-bounded, $\|d\|_\infty \le \delta$, i.e. if h is the system then the observed output is

$$y = h(u) + d.$$

An identification algorithm is a sequence of mappings $\phi = \{\phi_n\}$ generating at each time an estimate $\phi_n(P_n u, P_n y) \in \mathscr{X}$ of the unknown plant. Here P_n is the truncation operator defined by $P_n(u_0, u_1, \ldots, u_n, u_{n+1}, \ldots) = (u_0, \ldots, u_n, 0, 0, \ldots)$.

Given an identification algorithm and a chosen input, we should like to consider the limiting situation when longer and longer output sequences are observed. To this end, the worst-case asymptotic error, $e_\infty(\phi, u, \delta)$ of an algorithm ϕ is defined as the smallest number r such that for all plants $h \in \mathscr{M}$ and for all disturbances d with $\|d\|_\infty \le \delta$

$$\limsup_{n \to \infty} \|\phi(P_n u, P_n(u * h + d)) - h\|_{\mathscr{X}} \le r.$$

Equivalently

$$e_\infty(\phi, u, \delta) \equiv \sup_{h \in \mathscr{M}} \sup_{\|d\|_\infty \le \delta} \limsup_{n \to \infty} \|\phi(P_n u, P_n(u * h + d)) - h\|_{\mathscr{X}}.$$

The interpretation of this definition is that no matter what the true plant and the disturbances are, the plant can be eventually approximated to within $e_\infty(\phi, u, \delta)$, using the estimates generated by the identification algorithm. The convergence rate may depend on the plant and noise, i.e. for a given ε, there exists some $N(d, h, \varepsilon)$ such that

$$\|\phi_n(y) - h\|_{\mathscr{X}} < e_\infty(\phi, u, \delta) + \varepsilon$$

whenever $n \ge N$. We say that the convergence is uniform if $N(y, h, \varepsilon)$ depends only on ε. For more motivation and discussion on these definitions, see Tse *et al.* (1993).

The optimal worst-case asymptotic error $E_\infty(u, \delta)$ is defined as the smallest error achievable by any algorithm:

$$E_\infty(u, \delta) \equiv \inf_{\phi} e_\infty(\phi, u, \delta).$$

Any algorithm for which the infimum is attained is said to be asymptotically optimal. We shall obtain a general characterization of the asymptotically optimal algorithms and the resulting optimal error, for any given input u. We shall then find conditions on the input u to make this optimal worst-case asymptotic error small.

4. *Asymptotically optimal identification*

The characterization of asymptotically optimal algorithms and optimal asymptotic errors is in terms of the uncertainty set, an important notion in information-based complexity theory. The uncertainty set $S_n(u, y, \delta)$ at time n is the set of all systems in the given model set $\mathscr{M}$ that are consistent with the observed data up until time n:

$$S_n(u, y, \delta) = \{h \in \mathscr{M}: \|P_n(y - h(u))\|_\infty \le \delta\}.$$

These are the plants that can give rise to the observed output for some valid disturbance sequence. The infinite-horizon uncertainty set is

$$S_\infty(u, y, \delta) = \{h \in \mathscr{M}: \|y - h(u)\|_\infty \le \delta\}.$$

For a given set $A \subset \mathscr{X}$, define the diameter of the set as

$$\operatorname{diam}(A) = \sup_{g, h \in A} \|g - h\|_{\mathscr{X}},$$

and let $D(u, \delta)$ be the diameter of the worst-case infinite-horizon uncertainty set:

$$D(u, \delta) = \sup_{h \in \mathscr{M}} \sup_{\|d\|_\infty \le \delta} \operatorname{diam}[S_\infty(u, u * h + d, \delta)].$$

Under appropriate topological conditions on the model set, this quantity characterizes the optimal asymptotic worst-case error so that the infinite-horizon experiment can essentially be viewed as the limit of the finite-horizon experiments. We call this notion consistency. The following consistency result is a generalization of Proposition 3.3, Theorem 3.4 and Proposition 3.9 in the LTI case (Tse *et al.*, 1993) to our present setting. The proof is essentially the same, and is omitted.

Theorem 4.1. If the model set $\mathscr{M} \subset \mathscr{X}$ is σ-compact (i.e. $\mathscr{M}$ is a countable union of compact sets) then

$$\tfrac{1}{2} D(u, \delta) \le E_\infty(u, \delta) \le D(u, \delta).$$

Furthermore, if $\mathscr{M}$ is compact then the convergence can be made uniform.

In the σ-compact case, an algorithm achieving an asymptotic error within the above bounds can be realized using the principle of Occam's razor. Let $\mathscr{M} = \bigcup_i \mathscr{M}_i$, where the $\mathscr{M}_i$ are compact and increasing. This decomposition gives a complexity index to each plant h in $\mathscr{M}$, defined as the smallest i for which $h \in \mathscr{M}_i$. At each time n, the algorithm simply returns, as an estimate, any plant in the uncertainty set S_n with the smallest complexity index. Note that since the disturbance bound δ is required to compute the uncertainty set, this algorithm is tuned to this information. On the other hand, if $\mathscr{M}$ is compact, one can use an algorithm that simply returns the plant in $\mathscr{M}$ that fits best the input/output data observed so far. This algorithm attains an asymptotic error within the above bounds with a uniform rate of convergence. It is also untuned to the disturbance bound δ.

A slight extension of the above result yields essentially the same bounds for the case when $\mathscr{M}$ is separable. The proof is along the same lines as the proofs of Lemma 4.5 and Proposition 4.6 in Tse *et al.* (1993). The optimal algorithm has roughly the same structure as that for the σ-compact case.

Theorem 4.2. If $\mathscr{M}$ is separable then

$$\tfrac{1}{2} D(u, \delta) \le E_\infty(u, \delta) \le \lim_{x \downarrow \delta} D(u, x).$$

To apply the above results, we now look at the topological structure of some classes of fading memory systems under the operator-induced norm.

Consider first the class of stable LTI systems. Since this corresponds to the space ℓ_1, which is separable, Theorem 4.2 is applicable in this case. More generally, we can in fact prove the following.

Theorem 4.3. The class of all fading memory systems is separable.

Proof. Define the class of pth-order memory systems, $\mathcal{M}_p$, to be the set of all f such that for every k and for every $t > p$ and all finite sequences $v \in [-1, 1]^k$ and $w \in [-1, 1]^t$, $f_{t+k}(vw) = f_t(w)$. It is clear that any fading memory system can be approximated (in the operator-induced norm) arbitrarily closely by a pth-order memory system for sufficiently large p. Hence it suffices to prove that $\mathcal{M}_p$ is separable for all p.

Now, given any $f \in \mathcal{M}_p$, we can find some continuous function $g: [-1, 1]^p \to \mathbb{R}$ such that for all time n and all input u

$$f_n(u) = g(u_{n-p}, \ldots, u_{n-1}).$$

We call g the memory function for f. Hence we have $\|f\| = \|g\|_\infty$, where the infinity norm is taken over $[-1, 1]^p$. But the space of continuous functions with the uniform topology induced by the ℓ_∞ norm, denoted by $C([-1, 1]^p)$, is separable, and hence so is $\mathcal{M}_p$. ■

This means that when we look at fading memory systems, we can apply Theorem 4.2, and reduce the analysis of the asymptotic optimal error to the analysis of the worst-case infinite-horizon diameter.

5. *Optimal inputs*

We now turn to the question of optimal inputs, i.e. inputs u that minimize the worst-case infinite-horizon diameter $D(u, \delta)$. First we state a simple lower bound.

Lemma 5.1. Assume $\mathcal{M}$ is path-connected and diam $(\mathcal{M}) \geq 2\delta$. Then $D(u, \delta) \geq 2\delta$ for all $u \in \mathcal{U}$.

Proof. See Tse *et al.* (1993). ■

Since diam $(\mathcal{M}) > 2\delta$ for most reasonable model sets, the above results gives a general lower bound. We now investigate how to choose an input that achieves this bound.

Recall that $\mathcal{M}$ is balanced if $h \in \mathcal{M}$ implies $-h \in \mathcal{M}$. For balanced and convex model sets, it is well-known from information-based complexity theory (Traub and Wozniakowski, 1980) that the worst-case diameter is equal to the diameter of the uncertainty set when the output is identically zero. The following lemma summarizes this.

Lemma 5.2. Assume that $\mathcal{M}$ is balanced and convex. Then, for all $u \in \mathcal{U}$, $\delta > 0$

$$D(u, \delta) = \text{diam}\,[S_\infty(u, 0, \delta)].$$

Call an input $u \in \mathcal{U}$ persistently exciting for $\mathcal{M}$ if the following property holds:

$$\|h(u)\|_\infty = \|h\|_\mathcal{M} \quad \forall h \in \mathcal{M}.$$

The following result says that persistently exciting inputs are optimal.

Theorem 5.1. Assume that $\mathcal{M}$ is balanced and convex.

(1) If the input u is persistently exciting, then $D(u, \delta) \leq 2\delta$ for all $\delta > 0$.

(2) If u is persistently exciting then $D(u, \delta) = 2\delta$ for each $0 < \delta \leq \frac{1}{2}$ diam $(\mathcal{M})$.

Proof. (1) By Lemma 5.2, for all $\delta > 0$

$$D(u, \delta) = 2 \sup\{\|h\| : h \in \mathcal{M}, \|h(u)\|_\infty \leq \delta\}.$$

Pick any $h \in \mathcal{M}$ such that $\|h(u)\|_\infty \leq \delta$. If u is persistently exciting, this means that also $\|h\|_\mathcal{M} \leq \delta$, so $D(u, \delta) \leq 2\delta$.

(2) From Lemma 5.1, $D(u, \delta) \geq 2\delta$ for such δ. The result follows from (1). ■

It follows from Theorems 4.2 and 4.3, Lemma 5.1 and the above theorem that one can achieve nearly optimal asymptotic identification for the entire class of fading memory systems if one uses a persistently exciting input.

Corollary 5.1. Let $\mathcal{M} = \mathcal{X}$, the class of all fading memory systems. Then for any identification algorithm ϕ and any input u, the worst-case asymptotic error $e_\infty(\phi, u, \delta)$ is lower bounded by δ. If u is persistently exciting then there is an algorithm that can achieve an asymptotic error of less than 2δ.

A natural question that arises at this point is whether persistently exciting inputs exist. In the stable LTI case, this was shown to be the case (Tse *et al.*, 1993). The next theorem shows that they also exist when the model set consists of nonlinear fading memory systems.

Theorem 5.2. Let the model set $\mathcal{M}$ be some subset of the set of fading memory systems. Let W be any countable dense subset of $[-1, 1]$ and consider any input $\omega_0 \in [-1, 1]^\infty$ that contains all possible finite sequences of elements of W. Then ω_0 is persistently exciting.

Proof. Assume that $h \in \mathcal{M}$, $\|h\| = K < \infty$. Pick any $\varepsilon > 0$. Let $T = T(\varepsilon)$ as in the definition of FM. By definition of the sup norm, there are some ω and some T_1 so that

$$\sup_{0 \leq t < T_1} |h_t(\omega)| > K - \varepsilon.$$

Using the equilibrium-initial assumption and replacing ω by $0^T\omega$ and T_1 by $T + T_1$, we may assume that

$$\sup_{T \leq t < T_1} |h_t(\omega)| > K - \varepsilon.$$

By density of W and continuity of $h_t(\omega)$ on past values of ω, we may further assume that $\omega(0), \ldots, \omega(T_1 - 1)$ take values in W. From the construction of ω_0, there is some k such that

$$\omega_0(k) = \omega(0), \ \omega_0(k+1) = \omega(1), \ \ldots, \ \omega_0(k + T_1 - 1) = \omega(T_1 - 1).$$

Let v be the finite sequence $\omega_0(0), \omega_0(1), \ldots, \omega_0(k-1)$ and let w be the finite sequence $\omega(0), \omega(1), \ldots, \omega(T_1 - 1)$, which is equal to $\omega_0(k), \omega_0(k+1), \ldots, \omega_0(k + T_1 - 1)$. So vw is the same as the first $T_1 + k - 1$ elements of ω_0.

By the FM property applied to these inputs, we have

$$|h_{t+k}(vw) - h_t(w)| < \varepsilon \quad \text{for each } t \geq T$$

(using the notational convention mentioned above for $h_s(w)$ if the length of w is larger than s). Then for such t

$$|h_{t+k}(\omega_0)| = |h_{t+k}(vw)| \geq |h_t(w)| - \varepsilon,$$

so

$$\|h(\omega_0)\| \geq \sup_{T+k \leq r < T_1 + k} |h_r(\omega_0)| \geq K - 2\varepsilon.$$

Thus we conclude that $K = \|h\| \geq \|h(\omega_0)\| > K - 2\varepsilon$ for all $\varepsilon > 0$, so $\|h(\omega_0)\| = K$ as wanted. ■

6. *An untuned algorithm*

As remarked earlier, the asymptotically optimal algorithms for σ-compact and separable model sets are tuned to the knowledge of δ, the bound on the disturbance. It will be shown that for fading memory systems, we can achieve asymptotically optimal identification without knowing δ, provided that we use a persistently exciting input. This is in fact a generalization of a result by Makila (1991), which was proved in the context of stable LTI systems.

We shall make use of multivariate piecewise-linear spline functions to interpolate between the measured data to form an approximation to the unknown plant. This is a generalization of the univariate linear spline, but, because in higher dimension there is no natural ordering of the data points, the description of the interpolant is more complicated.

Consider the cube $I = [-1, 1]^p \subset \mathbb{R}^p$. Let $x_1, x_2, \ldots, x_m$, $m > p$, be m points in the interior of the cube. We wish to construct a continuous, piecewise-linear function $f: I \to \mathbb{R}$ such that $f(x_i) = y_i$, $i = 1, 2, \ldots, m$, where the y_i are given data values to interpolate.

To facilitate the discussion, we need to first define several basic geometric concepts. A p-dimensional simplex S in $\mathbb{R}^p$ is the convex hull of $p + 1$ affinely independent points. Each of

these points is a vertex of S. The convex hull F of any subset of these $p+1$ points form a face of S if there exists a hyperplane H such that S lies entirely on one side of H and $S \cap H = F$. If F is the convex hull of p points, it is called a facet. A point v outside S is said to be separated from S by a face F if v and S lie on opposite sides of the $p-1$ dimensional hyperplane containing F. A more elaborate discussion of these concepts can be found in Grotschel *et al.* (1987).

The first step is to find a set of simplices $\{S_j\}$ such that (1) their combined vertex set is $\{x_1, \ldots, x_m\}$, (2) the simplices only intersect at common faces and (3) their union gives the convex hull of the vertex set. This can be done inductively as follows: for $m = p+1$, the set simply consists of one simplex, which is the convex hull of the $p+1$ points. Suppose now we obtain a set of simplices $S_1, S_2, \ldots, S_d$ to cover $m > p$ points, and consider one additional point x_{m+1}. If $x_{m+1} \in S_k$ for some k then we can simply replace S_k with the $p+1$ simplices formed by x_{m+1} with each of the faces of S_k. It is easy to see that these $p+1$ simplices only intersect at common faces and their union is S_k, so that the updated set of simplices now covers the $m+1$ points. On the other hand, if x_{m+1} lies outside $P = \bigcup_{i=1}^{d} S_i$ then for each facet F of some S_k that separates x_{m+1} from P, we add a simplex formed by x_{m+1} with F to the set. It can also be proved that these added simplices together with the original ones satisfy the three conditions.

Given these simplices, we can now define our interpolating linear spline f as follows: first define $f(x_i) = y_i$ at the given data points. For other $x \in [-1, 1]^p$, if $x \in S_j$ for some j, let $f(x) \equiv \Sigma_i \alpha_i f(u_i)$, where the u_i are the vertices of S_j and $x = \Sigma_i \alpha_i u_i$. It is easy to check that because of the above three conditions on the simplices, f is well defined and continuous. To extend f continuously outside $P = \bigcup_j S_j$, define $f(x)$ to be equal to the value of f at the nearest point in P to x. Since P is convex, this nearest point is unique, and this guarantees the continuity of this extension.

If we view this interpolating process as an operator T_m mapping the data vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ to the piecewise-linear interpolating function $(T_m(\mathbf{y}))(x)$ then we can see that this operator is linear, and its gain, defined as

$$\|T_m\| \equiv \sup_{\|\mathbf{y}\|_\infty = 1} \|T_m(\mathbf{y})\|_\infty, \tag{1}$$

is equal to one. This simple fact ensures that, no matter how many data are obtained, noise in the data will not be amplified in the interpolating process. This property of linear splines, which is not shared by methods such as global polynomial interpolation, turns out to be the key to guaranteeing the consistency of the estimates. A similar situation is encountered in linear system identification from frequency-response data (Helmicki *et al.*, 1991), where one-dimensional splines are used instead of polynomials to interpolate the noisy data to guarantee robustness of the identification procedure.

With the above basic discussions on multivariate linear splines, we may now state the main result of this section.

Theorem 6.1. Let the model set $\mathcal{M} = \mathcal{X}$, the set of all fading memory systems. If the input u is persistently exciting then there is an algorithm that can achieve an optimal worst-case asymptotic error $e_\infty(\phi, u, \delta) = \delta$. This algorithm does not require the knowledge of δ in computing the estimates.

Proof. The structure of the algorithm is as follows. We view the model set $\mathcal{M}$ as the closure of the finite-memory systems $\mathcal{M}_p$, $p = 0, 1, 2, \ldots$. We start by assuming that the true system is in $\mathcal{M}_0$. Data is observed until time $n(0)$, after which the algorithm comes up with an estimate $\hat{h}^{(0)} \in \mathcal{M}_0$. Then it moves onto the next model set $\mathcal{M}_1$, and waits until time $n(1)$ before coming up with an estimate $\hat{h}^{(1)} \in \mathcal{M}_1$. The algorithm continues to move onto the model set of one higher order, to produce a new estimate. It will be shown below how the time $n(p)$ is specified and the estimate $\hat{h}^{(p)}$ is computed for each p.

Let h be the true system. Let $\{\delta_p\}$ be any sequence that goes to zero monotonically.

Fix p, and let the time $n \in [n(p-1), n(p)]$. (This is when the algorithm is collecting data to compute an estimate in $\mathcal{M}_p$.) Consider all the blocks

$$(u_{n-p+1}, \ldots, u_{n-1}, u_n) \quad \forall n = n(p-1), \ldots, n(p)$$

in the input as data points in the cube $[-1, 1]^p$. We maintain a simplicial structure in $[-1, 1]^p$ with these data points as a vertex set, and the structure is incrementally modified more or less according to the procedure discussed earlier, with a slight twist. Let $C_n = \bigcup_j S_j$ be the union of the simplices at time n, and let d_n be the distance between C_n and the corner of $[-1, 1]^p$ farthest away from C_n. At time $n+1$, one more data point is obtained. If $d_n < \delta_p$ and the new data point lies outside C_n then discard the new point. Otherwise update the simplicial structure as described earlier.

Let $n(p)$ be the earliest time such that $d_{n(p)} < \delta_p$ and the diameter of the largest simplex in C_n is less than δ_p. At this time, the algorithm returns an estimate $\hat{h}^{(p)} \equiv \phi_{n(p)}(h(u) + d)$ to be the pth-order system with memory function as the piecewise-linear spline interpolant of the current simplex structure.

We now claim that $n(p) < \infty$ for every p. First we see that, because the input is persistently exciting, the p-blocks in u are dense in $[-1, 1]^p$. (Otherwise, there is a ball in $[-1, 1]^p$ that does not contain any blocks in u, and we can construct a p-step finite memory system with a continuous function $f: \mathbb{R}^p \to \mathbb{R}$ to be positive at the centre of the ball and zero outside. Then applying the input u to the system will give a zero output while the gain of the system is non-zero, thus contradicting the persistent excitedness of u.) Hence there exists a time $m(p)$ such that $d_{m(p)} < \delta_p$. After this time, the convex hull C_n no longer expands. All the changes consist of further partitioning of the simplices inside C_n due to the new data points. Because the data points are dense, it can be seen that the diameter of the largest simplex must go to zero. Hence, $n(p)$ is finite.

We now claim that

$$\limsup_{p \to \infty} \|\phi_{n(p)}(h(u) + d) - h\| \leq \delta$$

for all d, $\|d\|_\infty \leq \delta$. This, combined with our lower-bound result (Corollary 5.1), shows that the asymptotic error is exactly δ. Note also that the algorithm defined above does not use the value of δ in computing the estimate.

Take any $\epsilon > 0$. There exists some q such that $\mathcal{M}_q$ contains a system h^ϵ with

$$\|h - h^\epsilon\| \leq \epsilon. \tag{2}$$

Let g^ϵ be the (q-step) memory function for h^ϵ. For $p \geq q$, $\phi_{n(p)}(h(u) + d)$ is the spline interpolant that approximates the unknown memory function, and $y = h(u) + d$ is the output. We can also extend g^ϵ to a function on $[-1, 1]^p$ that depends only on the last q coordinates. Now

$$\begin{aligned}
&\|\phi_{n(p)}(h(u) + d) - g^\epsilon\|_\infty \\
&\quad = \|\phi_{n(p)}(h(u)) + \phi_{n(p)}(d) - g^\epsilon\|_\infty \\
&\qquad \text{(by linearity of the interpolation operator)} \\
&\quad \leq \|\phi_{n(p)}(h(u)) - g^\epsilon\|_\infty + \|\phi_{n(p)}(d)\|_\infty \\
&\quad \leq \|\phi_{n(p)}(h(u)) - g^\epsilon\|_\infty + \delta \quad \text{(by (1))}.
\end{aligned}$$

The first term is the interpolation error when the data is noiseless, whereas the second term is the error due to the noise. We now show that the first term can be made arbitrarily small for large p.

Since g^ϵ is continuous, g^ϵ is a uniformly continuous function on $[-1, 1]^q$. Choose ϵ_1 such that

$$\|x_1 - x_2\|_2 \leq \epsilon_1 \Rightarrow \|g^\epsilon(x_1) - g^\epsilon(x_2)\|_2 \leq \epsilon. \tag{3}$$

Now pick p sufficiently large that $\delta_p < \epsilon_1$ and $p > q$. Let $g^p(x) \equiv \phi_{n(p)}(h(u))$.

Now, for any $x \in C_{n(p)}$, the convex hull, let $x = \Sigma_i \alpha_i x_i$, where x_i are the vertices of the simplex containing x. Since g^p agrees with the noiseless output data at the vertices, by (2), for each i

$$|g^p(x_i) - g^\epsilon(x_i)| \leq \epsilon. \tag{4}$$

We have

$$|g^p(x) - g^\epsilon(x)| = \left|\sum_i \alpha_i g^p(x_i) - g^\epsilon(x)\right|$$
$$\leq \left|\sum_i \alpha_i g^\epsilon(x_i) - g^\epsilon(x)\right| + \epsilon \quad \text{(by (4))}$$
$$\leq \sum_i \alpha_i |g^\epsilon(x_i) - g^\epsilon(x)| + \epsilon$$
$$\leq 2\epsilon$$

by (2), since $\|x_i - x\|$ is less than the diameter of the simplex, which is smaller than ϵ_1.

Now for x outside $C_{n(p)}$, let x' be the point in $C_{n(p)}$ that is closest to x. By definition of $n(p)$, the distance of x' from x is at most $\delta_p < \epsilon_1$. Hence

$$|g^p(x) - g^\epsilon(x)| = |g^p(x') - g^\epsilon(x)|$$
(by definition of the interpolant)
$$\leq |g^p(x') - g^\epsilon(x')| + \epsilon \quad \text{(by (2))}$$
$$\leq 3\epsilon \quad \text{(from above).}$$

Therefore if h^p is the finite-memory system with memory function $\phi_{n(p)}(h(u) + d)$ then

$$\|h^p - h\| \leq \|g^p - g^\epsilon\|_\infty + \epsilon \leq \delta + 4\epsilon.$$

Since this is true for all ϵ, it follows that

$$\limsup_{p\to\infty} \|h^p - h\| \leq \delta,$$

as desired. ■

Compared with the corresponding result in the LTI case (Makila, 1991), it can be seen that, while in the fading memory case we can attain the lower bound of δ in the asymptotic error, in the LTI case one can only guarantee an asymptotic error of less than 2δ. While this seems at first paradoxical, since the space of LTI plants is a subset of the fading memory plants, it shows that in fact one can reduce the worst-case error by allowing the algorithm to return nonlinear plants as estimates even if it was known a priori that the true plant is linear. This is due to the geometry of the subset of LTI plants in relation to the bigger space of nonlinear fading memory plants.

Finally, we should like to make a comment about the algorithmic and time complexity of this identification problem. For a given experiment length n, the complexity of the proposed identification algorithm is roughly proportional to the number of simplices in the imposed simplex structure. This is in turn bounded by $O(n^p)$, where p is the order of the memory function, since each simplex has $p + 1$ vertices. In terms of time complexity, it can easily be seen that, in general, the time needed to identify a system to a prescribed accuracy grows exponentially with the order of the system, even when there is no noise. For example, if we assume a certain Lipschitz condition on the order p memory function g, such as $|g(x) - g(y)| < M\|x - y\|$, then, to identify the function up to accuracy ϵ (in the $\|\cdot\|_\infty$ norm) the number of data points needed is at least the minimum number of ϵ-balls to cover $[-1, 1]^p$. Since the volume of an ϵ-ball is proportional to ϵ^p, it is clear that this minimum number is at least proportional to ϵ^{-p}, and hence so is the experiment length. This means that if p is large, the experiment length will be very long if we make no further assumption on the unknown plant.

It is interesting to compare this situation with the problem of identifying linear finite-impulse response systems. For nonlinear systems, the time complexity is exponential in the order, whether or not there is noise. For the linear case, while it takes only linear time to identify an FIR system exactly when there is no noise, it has been shown (Dahleh *et al.* 1993; Poolla and Tikku, 1994) that the time complexity immediately becomes exponential once we introduce any worst-case noise. Moreover, it has been demonstrated that if we are willing to put a probability distribution on the noise, polynomial time complexity can often be obtained (Tse and Tsitsiklis, 1990). These facts show that while in the nonlinear case the plant uncertainty determines the time complexity of the identification, in the linear case the complexity is sensitive to how the noise is modeled.

7. *Conclusions*

A framework for the analysis of asymptotic worst-case idenfication of LTI systems has been extended to the setting of nonlinear fading memory systems. For model sets that are either σ-compact or separable, and for any experiment, the optimal worst-case error is always bounded by twice the lower bound, which is the diameter of a certain uncertainty set. Optimal inputs that minimize this diameter have been characterized. It has also been shown that accurate asymptotic identification can be achieved by an optimal input, using an untuned algorithm based on spline interpolation.

Acknowledgements—M. A. Dahleh's research was supported by AFOSR under Grant AFOSR-91-0368 and by the NSF under grant 9157306-ECS. E. D. Sontag's research was supported by AFOSR under Grant AFOSR-91-0346. D. N. C. Tse's research was supported by an NSERC Fellowship from the government of Canada, and by the NSF under Grant ECS-8552419. J. N. Tsitsiklis's research was supported by the NSF under Grant ECS-8552419 and by the AFOSR under Grant AFOSR-91-0368.

References

Boyd, S. and L. O. Chua (1985). Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Trans. Circuits Syst.*, **CAS-32,** 1150–1161.

Dahleh, M. A. and M. H. Khammash (1990). Controller design in the presence of structured uncertainty. *Automatica,* **29,** 37–56.

Dahleh, M. A., T. V. Theodosopoulos and J. N. Tsitsiklis (1993). The sample complexity of worst-case identification of FIR linear systems. *Syst. Control Lett.*, **20,** 157–166.

Doyle, J. C. (1982). Analysis of feedback systems with structured uncertainty. *IEEE Proc.*, **129,** 242–250.

Grotschel, M., L. Lovasz and A. Schrijver (1987). *Geometric Algorithms and Combinatorial Optimization.* Springer, Berlin.

Gu, G., P. P. Khargonekar and Y. Li (1992). Robust convergence of two stage nonlinear algorithms for identification in $\mathcal{H}_\infty$. *Syst. Control Lett.*, **18,** 253–264.

Gu, G. and P. P. Khargonekar (1992). Linear and nonlinear algorithms for identification in $\mathcal{H}_\infty$ with error bounds. *IEEE Trans. Autom. Control.*, **AC-37,** 953–963.

Helmicki, A. J., C. A. Jacobson and C. N. Nett (1991). Control-oriented system identification: a worst-case/deterministic approach in H_∞. *IEEE Trans. Autom. Control,* **AC-36,** 1163–1176.

Jacobson, C. A. and C. N. Nett (1991). Worst-case system identification in ℓ_1: optimal algorithms and error bounds. In *Proc.* 1991 *American Control Conf.*, Boston, MA.

Kacewicz, B. and M. Milanese (1992) On the optimal experiment design in the worst-case ℓ_1 system identification. In *Proc. Conf. Decision Control,* Tucson, AZ.

Ljung, L. (1987). *System Identification—Theory for the User.* Prentice-Hall, Englewood Cliffs, N.J.

Makila, P. M. (1991). Robust identification and Galois sequences. Technical Report 91-1, Process Control Laboratory, Swedish University of Abo.

Makila, P. M. and J. R. Partington (1991). Robust approximation and identification in H_∞. In *Proc.* 1991 *American Control Conf.*, Boston, MA.

Mehra, R. K. (1974). Optimal input signals for parameter estimation in dynamic systems—a survey and new results. *IEEE Trans. Autom. Control,* **AC-19,** 753–768.

Poolla, K. and A. Tikku (1994). On the time complexity of worst-case system identification. *IEEE Trans. Autom. Control,* **AC-39,** 944–950.

Shamma, J. S. and R. Zhao, (1990) Fading memory feedback systems and robust stability. *Automatica,* **29.**

Traub, J. F. and H. Wozniakowski (1980). *A General Theory of Optimal Algorithms.* Academic Press, New York.

Tse, D., M. A. Dahleh and J. N. Tsitsiklis (1993). Optimal

asymptotic identification under bounded disturbances. *IEEE Trans. Autom. Control*, **AC-38**, 1176–90.

Tse, D. N. C. and J. N. Tsitsiklis (1990). Sample complexity of worst-case system identification in the presence of bounded random noise. Unpublished manuscript.

Zames, G. (1979). On the metric complexity of casual linear systems: ϵ-entropy and ϵ-dimension for continuous-time. *IEEE Trans. Autom. Control*, **AC-24**, 222–230.

Zarrop, M. (1979). *Optimal Experimental Design for Dynamic System Identification*. Springer, New York.