Worst-Case Identification of Nonlinear Fading Memory Systems

Munther A. Dahleh*
Lab. for Information & Decision Systems
M.I.T.
Cambridge, MA 02139

Eduardo D. Sontag[‡]
Dept. of Mathematics, SYCON
Rutgers University
New Brunswick, NJ 08903

David N.C. Tse[†]
Lab. for Information & Decision Systems
M.I.T.
Cambridge, MA 02139

John N. Tsitsiklis[§]
Lab. for Information & Decision Systems
M.I.T.
Cambridge, MA 02139

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

1 Introduction

Recently, there has been an 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 [2, 5]. This problem is of special importance when the data are corrupted with bounded noise. The case where the objective is to optimize prediction for a fixed input was analyzed by many researchers (see [10] and the references therein). The problem is more interesting when the objective is to approximate the original system as an operator. a problem extensively discussed in [20], especially when the plant's order is not known a priori. For linear time invariant plants, such 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 ([8, 6, 7]). When the topology is induced by the ℓ_1 norm, a complete study of asymptotic identification was furnished in [18] for arbitrary inputs, and the question of optimal input design was addressed. Input design has been addressed in stochastic settings (e.g. [11, 21]), but not in worst-case settings. Related work on the worst-case identification problem was also reported in [13, 14, 12, 15, 3, 9].

In this paper, the work of Tse *et al* [18] is extended to analyse the worst-case asymptotic identification of nonlinear fading memory systems. As in [18], the study is done in two steps. The first step 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. An *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 which contains the possible inputs that can be used for performing the identification experiments. We consider the set of models $\mathcal{X}$ as discrete-time, causal functions from $\mathcal{U}$ to $\Re^\infty$; a plant $h \in \mathcal{X}$ 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 $h \in \mathcal{X}$ further satisfies the following properties:

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

2. h has equilibrium-initial behavior:

$$h_{n+1}(0u) = h_n(u) \text{ for all } n,$$

where $0u$ is the input $0, u_0, u_1, \ldots$. (In general, we will 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 the first n elements of which are given by the finite sequence w.)

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 sequences $v \in [-1, 1]^k$, $w \in [-1, 1]^t$,

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

To measure the identification error, we shall consider the metric ρ to be the one corresponding to the operator gain:

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

It can be seen that systems in $\mathcal{X}$ satisfying the above property necessarily must 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 apriori. In the above definition, this bound is normalized to one.

Examples of FM Systems:

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.

Hammerstein systems.
These are systems which are 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 function $g : \Re \to \Re$. It is easily 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 [1, 16].

3   Identification Setup

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

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

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

Given an identification algorithm and a chosen input, we would like to consider the limiting situation when longer and longer of the 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 $h \in \mathcal{M}$ and for all disturbances $\mathbf{d}$ with $\|\mathbf{d}\|_\infty \leq \delta$,

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

Equivalently,

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

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 $N(d, h, \varepsilon)$ so that

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

whenever $n \geq N$. We say that the convergence is uniform if $N(y, h, \varepsilon)$ depends only on ε. For more motivations and discussions on these definitions, see [18].

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 will obtain a general characterization of the asympotically optimal algorithms and the resulting optimal error, for any given input u. We will 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 $\mathcal{M}$ which are consistent with the observed data up until time n:

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

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

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

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

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

4

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

$$D(u, \delta) = \sup_{h \in \mathcal{M}} \sup_{||d||_\infty \leq \delta} \operatorname{diam}(S_\infty(u, u * h + d, \delta))$$

Under appropriate topological conditions on the model set, this quantity characterizes the optimal asymptotic worst-case error. The following result is a generalization of Proposition 3.3, Theorem 3.4 and Proposition 3.9 in the LTI case ([18]) to our present setting.

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

$$\frac{D(u, \delta)}{2} \leq E_\infty(u, \delta) \leq D(u, \delta)$$

Furthermore, if $\mathcal{M}$ 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 $\mathcal{M} = \cup_i \mathcal{M}_i$, where the $\mathcal{M}_i$'s are compact and increasing. This decomposition gives a complexity index to each plant in $\mathcal{M}$, as the index of the smallest M_i containing the plant. 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 $\mathcal{M}$ is compact, one can use an algorithm which simply returns the plant in $\mathcal{M}$ which 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 $\mathcal{M}$ is separable. The proof is along the same lines as the proofs of Lemma 4.5 and Proposition 4.6 in [18]. The optimal algorithm has roughly the same structure as that for the σ-compact case.

Theorem 4.2 *If $\mathcal{M}$ is separable, then*

$$\frac{D(u, \delta)}{2} \leq E_\infty(u, \delta) \leq \lim_{x \downarrow \delta} D(u, x)$$

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:

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 every finite sequences $v \in [-1, 1]^k, w \in [-1, 1]^t$, $f_{t+k}(vw) = f_t(w)$. 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 $g : [-1, 1]^p \to \Re$ such that for all time n, and all input u,

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

5

We call g the *memory function* for f. Hence we have that $\|f\| = \|g\|_\infty$, 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 system, 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. Let

$$\rho(\mathcal{M}) := \sup\{r \mid 0 < \delta < r \ \Rightarrow \ \exists\, g, h \in \mathcal{M} \text{ with } \|g - h\| = r\} \ .$$

Note that if $\mathcal{M}$ is path connected, then $\rho(\mathcal{M}) = \text{diam}(\mathcal{M})$.

Lemma 5.1 If $2\delta < \rho(\mathcal{M})$, then $D(u, \delta) \geq 2\delta$ for all $u \in \mathcal{U}$.

Proof. See [18] ∎

Since $\rho(\mathcal{M}) > 2\delta$ for most of the reasonable model sets, the above result gives a general lower bound. We now investigate how to choose an input which 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 [17] that the worst case diameter is equal to the diameter of the uncertainty set when the output is identically equal to 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:

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

The following result says that persistently exciting inputs are optimal.

Theorem 5.1 *Assume $\mathcal{M}$ 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 $0 < \delta \leq \frac{\text{diam}(\mathcal{M})}{2}$.*

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

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

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

It follows from Theorem 4.2, Theorem 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 we use a persistently exciting input.

Corollary 5.3 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 which can achieve an asymptotic error of less than 2δ.

A natural question which arises at this point is whether persistently exciting inputs exist. In the stable LTI case, this was shown to be the case [18]. 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 $\omega_0 \in [-1, 1]^\infty$ which 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 is 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 so 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 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 that

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

(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 $K = \|h\| \geq \|h(\omega_0)\| \geq K - 2\varepsilon$ 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, one 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 [13], 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 $I = [-1, 1]^p \subset \Re^p$. 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 $f : I \to \Re$ such that $f(x_i) = y_i, i = 1, 2, \ldots, m$, where the y_i's 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 $\Re^p$ 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 the opposite sides of the $p - 1$ dimensional hyperplane containing F.

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 (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 $P = \cup_{i=1}^d S_i$, then for each facet F of some S_k which 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 $f(x) \equiv \sum_i \alpha_i f(u_i)$, where u_i's are the vertices of S_j and $x = \sum_i \alpha_i u_i$. 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 $P = \cup_j S_j$, 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 guarantee

the consistency of the estimates. A similar situation is encountered in linear system identification from frequency response data [8], where 1 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 which 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 which monotonically goes to zero.

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

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

in the input as data points in the cube $[-1, 1]^p$. We maintain a simplex structure in $[-1, 1]^p$ with these data points as vertex set, and the structure is incrementally modified more or less according to the procedure discussed earlier, with a slight twist. Let $C_n = \cup_j S_j$ be the union of the simplices at time n, and 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 simplex 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 p-th 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 $[-1, 1]^p$ which does not contain any blocks in u, and we can construct a p-step finite memory system with a continuous memory function $f : \Re^p \to \Re$ 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 seen that the diameter of the largest simplex must go to zero. Hence, $n(p)$ is finite.

We now claim that:

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

9

for all d, $\|d\|_\infty \le \delta$. 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

$$\|h - h^\epsilon\| \le \epsilon \tag{2}$$

Let g^ϵ be the (q-step) memory function for h^ϵ. For $p \ge q$, $\phi_{n(p)}(h(u)+d)$ 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 $[-1,1]^p$ which depends only on the last q coordinates. Now,

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

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

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

Now pick p sufficiently large such 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 $x = \sum_i \alpha_i x_i$, where x_i are the vertices of the simplex containing x. Since g^p agreees with the noiseless output data at the vertices, by Eqn. 2, for each i,

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

We have:

$$
\begin{aligned}
& |g^p(x) - g^\epsilon(x)| \\
= \ & \left| \sum_i \alpha_i g^p(x_i) - g^\epsilon(x) \right| \\
\le \ & \left| \sum_i \alpha_i g^\epsilon(x_i) - g^\epsilon(x) \right| + \epsilon \qquad \text{by Eqn. 4} \\
\le \ & \sum_i \alpha_i |g^\epsilon(x_i) - g^\epsilon(x)| + \epsilon \\
\le \ & 2\epsilon
\end{aligned}
$$

by Eqn. 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 $C_{n(p)}$ which is closest to x. By definition of $n(p)$, the distance of x' from x is at most $\delta_p < \epsilon_1$. Hence:

$$
\begin{aligned}
& |g^p(x) - g^\epsilon(x)| \\
= \ & |g^p(x') - g^\epsilon(x)| \qquad \text{by definition of the interpolant} \\
\le \ & |g^p(x') - g^\epsilon(x')| + \epsilon \qquad \text{by Eqn. 2} \\
\le \ & 3\epsilon \qquad \text{from above.}
\end{aligned}
$$

10

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. ∎

We would like to make a comment about the time complexity of this identification problem. It can be easily seen that in general, the time needed to identify a system to a prescribed accuracy grows exponentially as 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 $(\frac{1}{\epsilon})^p$, 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 of the order, whether or not there is noise. For the linear case, while it takes only linear time to identify a FIR system *exactly* when there is no noise, it has been shown [3, 15] 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 [19]. 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 modelled.

7 Conclusions

A framework for the analysis of asymptotic worst-case identification 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 which minimize this diameter are characterized. It is also shown that accurate asymptotic identification can be achieved by an optimal input, using an untuned algorithm based on spline interpolation.

References

[1] S. Boyd and L.O. Chua, "Fading Memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems*, Vol. CAS–32, 1985, pp. 1150–1161.

[2] M.A. Dahleh and M.H. Khammash, "Controller Design in the Presence of Structured Uncertainty," to appear in *Automatica* special issue on robust control.

[3]

[4] M.A.Dahleh, T.V.Theodosopoulos, J.N. Tsitsiklis " The Sample Complexity of Worst-Case Identification of FIR Linear Systems", *Systems and Control Letters*, March, 1993.

[5] J. C. Doyle, "Analysis of feedback systems with structured uncertainty," *IEEE Proceedings* 129, 242-250,1982.

[6] G. Gu, P.P. Khargonekar and Y. Li, "Robust convergence of two stage nonlinear algorithms for identification in $\mathcal{H}_\infty$, *Systems and Control Letters*, Vol 18, No. 4, April 1992.

[7] G. Gu and P.P. Khargonekar, " Linear and nonlinear algorithms for identification in $\mathcal{H}_\infty$ with error bounds, *IEEE Trans. A-C*, Vol 37, No. 7, July 1992, pp 953-963.

[8] A.J. Helmicki, C.A. Jacobson and C.N. Nett, "Control-oriented System Identification: A Worst-case/deterministic Approach in H_∞'," *IEEE Trans. A-C*, Vol 36, No. 10, October 1991, pp 1163-1176.

[9] C.A. Jacobson and C.N. Nett, "Worst-case system identification in ℓ_1: Optimal algorithms and error bounds," in *Proc. of the 1991 American Control Conference*, June 1991.

[10] L. Ljung, *System Identification - Theory for the User*, Prentice-Hall, 1987.

[11] R.K. Mehra, "Optimal input signals for parameter estimation in dynamic systems–A survey and new results", em IEEE Trans. Automatic Control, vol AC-19, pp.753-768, 1974.

[12] B. Kacewicz and M. Milanese, "On the Optimal Experiment Design in the Worst-Case ℓ_1 System Identification, submitted to *CDC*.

[13] P.M. Makila, "Robust Identification and Galois Sequences", Technical Report 91-1, Process Control Laboratory, Swedish University of Abo, January, 1991.

[14] P.M. Makila and J.R. Partington, "Robust Approximation and Identification in H_∞", Proc. 1991 American Control Conference, June, 1991.

[15] K. Poolla and A. Tikku, "On the time complexity of worst–case system identification," preprint, 1992.

[16] J.S. Shamma and R. Zhao, "Fading memory feedback systems and robust stability," to appear in *Automatica*.

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

[18] D. Tse, M.A. Dahleh and J.N. Tsitsiklis. Optimal Asymptotic Identification under bounded disturbances. To appear in the *IEEE Trans. Automat. Contr.*

[19] D.N.C. Tse and J.N. Tsitsiklis, " Sample Complexity of Worst-Case System Identification in the Presence of Bounded Random Noise", in preparation.

[20] G. Zames, "On the metric complexity of casual linear systems: ϵ-entropy and ϵ-dimension for continuous-time", *IEEE Trans. on Automatic Control*, Vol. 24, April 1979.

[21] M. Zarrop, *Optimal Experimental Design for Dynamic System Identification*, Springer-Verlag, New York, 1979.