On System Identification of Complex Systems from Finite Data

Saligrama R. Venkatesh and Munther A. Dahleh

***Abstract*—System identification deals with computation of mathematical models from an *a priori* chosen model-class, for an unknown system from finite noisy data. The popular maximum-likelihood principle is based on picking a model from a chosen model-parameterization that maximizes the likelihood of the data. Most other principles including set-membership identification can be thought of as extensions of this principle in so far as the concept of choosing a model to fit the data is concerned. Although these principles have been extremely successful in addressing several problems in identification and control, they have not been completely effective in addressing the question of identification in the context of uncertainty in the model class/parameterization. We introduce a new principle for identification in this paper. The principle is based on choosing a model from the model-parameterization which best approximates the unknown real system belonging to a more complex space of systems which do not lend themselves to a finite-parameterization. The principle is particularly effective for robust control as it leads to a precise notion of parametric and nonparametric error and the identification problem can be equivalently perceived as that of robust convergence of the parameters in the face of unmodeled errors. The main difficulty in its application stems from the interplay of noise and unmodeled dynamics and requires developing novel two-step algorithms that amount to annihilation of the unmodeled error followed by averaging out the noise. The principle contributions of the paper are in establishing: 1) robust convergence for a large class of systems, topologies, and unmodeled errors; 2) sample path based finite-time polynomial rate of convergence; and 3) annihilation-correlation algorithms, for linearly parameterized model structures, thus, illustrating significant improvements over prediction-error and set-membership approaches.**

***Index Terms*—Control-oriented identification, polynomial sample-complexity, robust control, robust learning, statistics, undermodeling.**

I. Introduction

SYSTEM identification deals with choosing mathematical models from a *known model set* to characterize the input–output behavior of an unknown system from finite noisy data. Noise, finite length of data, and time variation are some of the issues that limits the choice of a complex model set. There are many instances when this limitation is significant enough that it becomes necessary to deal with situations where no model in the model set can adequately describe the real system behavior. One common situation encountered is in the context of high-order time-varying dynamics where rate of variation precludes choosing models of high complexity. Another situation encountered is when lumped parameter models are used to characterize a partial differential equation (PDE). The former situation arises for instance while identifying acoustic transfer functions in a changing cabin environment (see [38]) and the latter situation arises while modeling the vertical dynamics of a ultra-high-rise elevator (see [3] and [33]).

In this paper, we limit our attention to linear systems and describe the means of dealing with system identification in instances when the real system cannot be adequately described by a chosen model set. In these instances, it is useful to conceptualize a larger, more complex set to which the real system belongs, and the idea in identification is to understand how well, how fast, and how easily can we approximate any arbitrary element in the complex set with some element in the chosen model set. A useful notion to describe complexity in this context is that of Kolmogorov n-width (see [25] and [30]). We deal with such systems where Kolmogorov n-width is, uniformly over all $n \in Z^+$, bounded away from zero. This means that the real systems/instances under consideration are inherently complex in that a choice of a larger dimensional model set does not *a priori* guarantee better representation of the real system.

At this point, it is worth pointing out that the origins of using simplified models for complex systems dates back to [41] where notions of n-widths, ϵ-entropy, and complexity were introduced in the context of identification and control. The notion of ϵ-entropy was used to characterize the degree of approximation in modeling a given set of possible systems within the class of finite-dimensional models. More recently, in [10] and [11], the idea of using restricted-complexity models, where the real-system may not belong to the model class, has been advocated for identification in order to guard against over parameterization of the estimated models.

We deal with the problem of system identification of simplified models of possibly more complex systems by appealing to the following principle: choose that model from the model set which best approximates the real system. This principle, which we call minimize-unmodeled dynamics (MUD), requires making explicit the idea of conceptualizing the class/set to which the real system belongs. The novelty of the formulation lies in its effectiveness in interfacing identification with robust control design. Every formulation in system identification has a natural set of variables associated with it. These variables form the basis for analyzing hypothetical situations, validating the assumptions, and quantifying the quality of the empirical model estimates. In our formulation, these variables are precisely, the parametric and nonparametric error, and the associated sample

Manuscript received May 5, 1997; revised October 21, 1998 and March 3, 2000. Recommended by Associate Editor A. Tesi. This work was supported in part by the National Science Foundation under Grants 9157306-ECS and AFOSR F49620-95-0219, and in part by Siemens AG.

The authors are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139–4307 USA.

Publisher Item Identifier S 0018-9286(01)01153-9.

complexity. The parametric error corresponds to the distance between the estimate and the best approximation measured in the parametric space. The nonparametric error corresponds to the distance between the system and the best approximation in the parametric space. These variables that characterize our identification problem naturally fall in line with robust control framework.

In this paper, our attention is restricted to the identification of stable linear time-invariant (LTI) systems from finite-noisy measurements of inputs and outputs. We are interested in both asymptotic as well as finite-data situations. Application of the principle to the class of LTI systems reduces to three questions. 1) Does there exist an estimator for the model parameters that converge asymptotically to the best approximation? 2) Can the parametric and nonparametric errors be quantified? 3) How long a data length does it take to reduce the parametric error to a prespecified tolerance bound? We require that the answers to these questions be independent of which element of the class/set the real system belongs. Thus, a stringent requirement of uniform convergence over the class of systems is desired.

The principal difficulty arises because of an interplay between noise and unmodeled dynamics. In the absence of noise the problem, in most situations, reduces to a function approximation problem for which well-known solutions already exist. In the absence of residual error, the problem reduces to a familiar situation dealt within classical identification (see [16], [28], [35], and [29]): given a finite-dimensional LTI process, find the best set of parameters from noisy data. However, under the influence of both, the problem is nontrivial even in the simplest of situations to answer the questions we posed earlier. Tackling the problem requires us to resort to a sequence of strategies. We develop a novel two-step algorithm, where, in the first step, the unmodeled dynamics is annihilated, except for transients. Application of the first step, therefore, nearly reduces the identification of model parameters to the familiar problem dealt with in the classical identification setting (see [16] and [28]). Thus, it remains to average out the noise and the transients. These steps require application of inputs that are persistently exciting of order infinity. Higher-order chirp inputs are constructed for this express purpose. The two-step procedure, along with such special inputs, are extremely effective for identification of model parameters and estimation of parametric and nonparametric error. We show that our goals can not only be attained asymptotically, but also that, for any prespecified error, a relatively short length of data is required. Some of these ideas have already appeared in our earlier publications (see [36], [32]).

The organization of the paper is as follows. Section IV introduces the framework and in the context of a simple finite-impulse response (FIR) model-parameterization to motivate the problem formulation to follow later. This section is also used to informally introduce other approaches—classical identification and set-membership identification—to discuss results pertaining to this example. This serves to motivate the problem setup and the general purpose of the paper. In Section V, the FIR example is used as the basis to understand the key requirements of the input. The subsequent sections then present LTI identification of complex systems in limited-complexity parameterizations for several different topologies.

For the sake of brevity, we postpone discussion of related work until later. We only note the fact that this topic has received wide attention. It is widely perceived that system-identification and robust-control pose a fundamental dichotomy. There is wide acceptance among researchers that the gap between robust-control design and traditional system identification (see, for instance, [16] and [28], which are standard texts on the subject) is yet unresolved. Consequently, the subject matter has received wide attention from as early as 1980s with an entire issue of the IEEE TRANSACTIONS on AUTOMATIC CONTROL devoted to this topic (see [8]). It is our view that the question remains as to how one can derive descriptions suitable for robust control design from finite corrupted data, and this forms the focus of this paper.

II. NOTATION

An * denotes the complex-conjugate transpose of the matrix A. Z^+ is the set of positive integers. ℓ denotes the space of sequences on Z^+, ℓ_p, and $p \geq 1$ denotes the space of sequences on Z^+ bounded in the ℓ_p norm (see [17]). For a signal $x \in \ell$, P_n denotes the truncation operator: $P_m(x) = (x(0), \ldots, x(m-1), 0, \ldots)$, X_n is the column vector $(x(0), x(1), \ldots, x(n-1))^*$, and $\Gamma_n(x)$ is the toeplitz matrix

$$\Gamma_n(x) = \begin{bmatrix} x(0) & 0 & \cdots & \cdots \\ x(1) & x(0) & 0 & \cdots \\ \vdots & \ddots & \ddots & \vdots \\ x(n-1) & x(n-2) & \cdots & x(0) \end{bmatrix}. \quad (1)$$

The n-point autocorrelation $r_x^n(\cdot)$ of a signal x is

$$r_x^n(\tau) = \frac{1}{n} \sum_{i=0}^{n-\tau} x(\tau+i)x(i); \quad r_x^n(-\tau) = r_x^n(\tau); \quad \tau \in [0, 1, \ldots, n].$$

$S_x^n(k)$ is the corresponding DFT of the autocorrelation sequence, i.e.,

$$S_x^n(k) = \sum_{t=0}^{n-1} r_x^n(t) \exp\left(-j\frac{2\pi}{n}kt\right).$$

$r_{xy}^n(\tau) = (1/n)\sum_{i=0}^{n-\tau} x(\tau+i)y(i)$ is the n-point cross correlation between two signals, and $S_{xy}^n(\cdot)$ is the corresponding cross-spectral density. $\langle x, y \rangle$ is the inner product of two signals x and y. For any infinite sequence $x(\cdot)$, $\hat{x}$ denotes the fourier transform. $N(m, \sigma)$ denotes the Gaussian distribution with mean m and standard deviation σ. $\mathcal{P}_w\{A\}$, $\Xi_w\{A\}$ denotes the probability and expected value of an event A. A real-valued function g, satisfying $g(n) \leq Cn$, is said to be $g \cong \mathcal{O}(n)$. For an LTI stable system, T, $\hat{T}$ denotes the λ-transform (or the z^{-1}-transform).

III. PRELIMINARIES

This section broadly describes the framework that will be used in the rest of the paper. We assume that the real system is a causal shift-invariant operator T belonging to a normed linear space $\mathcal{T}$ that takes inputs in ℓ to system outputs in ℓ. In several instances, we will further restrict the system T to belong

to an *a priori* set $\mathcal{I}$. However, the *a priori* set $\mathcal{I}$ is such that it does not lend itself to finite-parameterizations, a notion that will be made precise shortly. The objective is to "identify" the real process from input–output data. In most instances, we assume that we are free to choose the input u as long as we constrain the amplitude in time, i.e., $\|u\|_\infty \leq 1$. This constraint could arise from practical considerations to limit the input toa regime where assumptions of linearity hold. We suppose that noise, w, enters additively at the output of the process and corrupts the measurements. Noise is modeled by means of temporal constraints (probabilistic or set-valued) on noise sample-paths. In the set-valued case there is some set $\mathcal{W}_n$ from which the noise can take any sample path, i.e., $(w(0), w(1), \ldots, w(n)) \in \mathcal{W}_n$ and in the probabilistic situation, noise is assumed to be a stationary stochastic process. As a point of digression it is worth pointing out that it is possible to describe stochastic white noise with set-valued descriptions (see [23], [34], and [35]). These models have been used in the context of identification in [34], [35], [29], and [24] and have been shown to be "equivalent" to using stochastic models. In this paper, we will extensively use such set-valued white-noise descriptions to model noise. In summary, the measured output y and the input u are related by the following equation.

$$y(t) = Tu(t) + w(t), \qquad t = 0, 1, \ldots, n, \quad T \in \mathcal{I} \tag{2}$$

where $\mathcal{I}$, as alluded to before, is a complex *a priori* known set. We will now precisely define the notion of a complex set and the noise models that will be used in the paper.

A. Complex Priors

In this section, a notion of complexity based on Kolmogorov n-width is defined. We suppose that the real system T belongs to some set, $\mathcal{I}$. We associate the notion of complexity with the set, $\mathcal{I}$. Let $\mathcal{G}_\kappa$ denote any κ-dimensional subspace of $\mathcal{T}$. We say that the set, $\mathcal{I}$, is complex if it has a Kolmogorov κ-width which is asymptotically bounded away from zero irrespective of the size of κ, i.e.,

$$\limsup_{\kappa} \inf_{\mathcal{G}_\kappa} \sup_{T \in \mathcal{I}(\gamma)} \inf_{G \in \mathcal{G}_\kappa} \|T - G\| \geq \gamma > 0. \tag{3}$$

In particular, notice that the definition implies that choosing a more complex model parameterization does not *a priori* guarantee any reduction in the residual error. This definition is not far-fetched as we will see in the example below. The principle reason for such a definition stems from the need to confront residual unmodeled dynamics as an intrinsic aspect in system identification. In other words, our attempt is to disallow model reduction after identifying the entire system as part of the solution methodology. Although there are other ways to enforce this requirement, such as restricting the length of data, this definition is more natural. In the following example we show that complexity depends both on the topology on the space of systems and the "cardinality" of $\mathcal{I}$.

Example 1: Consider, the infinite-dimensional subset of LTI operators $\mathcal{T}$ with the kernel $\{t(k)\}$ satisfying

$$\mathcal{I} = \{T \in \mathcal{T} \mid |t(k)| \leq L\rho^k \qquad \forall k \in Z^+\} \tag{4}$$

where L and $0 \leq \rho < 1$ are known constants. This space is not complex if ℓ_1 norm is used as the distance measure. To see this we pick an FIR model-parameterization, $\mathcal{G}_{FIR}$ of order $\epsilon/L\log(\rho)$. Then, for every system $T \in \mathcal{I}$, there is a corresponding element in $\mathcal{G}_{FIR}$ which is no further than ϵ in the ℓ_1 norm. However, if we define the norm by $\sum_{k=0}^{\infty} |t(k)/\rho^k|$ the class of systems becomes complex. Similarly the ℓ_1 unit ball, Δ, is a class of complex systems as long as the norm measure is the ℓ_1 norm. However, with $\sum_{k=0}^{\infty} |\delta(k)\rho^k|$, $\rho < 1$ ($\delta(k)$ is the impulse response sequence of Δ), as the norm measure, the ℓ_1 unit ball is no longer complex.

B. Noise Models

In this subsection, we present stochastic and set-valued models for noise to be used in the rest of the paper. Typically, the principle feature of noise is that it is persistent and independent (uncorrelated) of the input. Our goal is to describe stochastic and set-valued models for noise which have these features of independence and persistency.

In the stochastic setting, these requirements are minimal and do not pose any problem and, in fact, almost any filtered i.i.d. process is admissible. For the sake of simplicity, we allow only filtered (stable-linear) white-Gaussian models for noise in this paper. Such noise processes are persistent and it remains to show that a strong notion of independence can also be established between an arbitrary input and any noise sample path. This follows from a large-deviations based probabilistic bound on the correlation between two signals. As alluded to before, the noise w can be characterized as white-Gaussian noise filtered through a stable linear filter H, i.e.,

$$w = Hv, \quad v \in N(0, \sigma), \quad \|H\|_1 \leq \eta$$

Lemma 1: Suppose, $\{w(\cdot)\}$ is as above and $u(t) \in [-1, 1]$, $t = 1, \ldots, n$ is a fixed vector, then

$$\mathcal{P}\left\{\frac{1}{n}\left|\sum_{t=1}^{n} u(t)w(t)\right| \geq \alpha\right\} \leq n \exp\left(-\frac{n\alpha}{\eta\sigma^2}\right). \tag{5}$$

Proof: We first notice the following set of inequalities:

$$\begin{aligned}
&\mathcal{P}\left\{\frac{1}{n}\left|\sum_{t=1}^{n} u(t)w(t)\right| \geq \alpha\right\} \\
&\quad = \mathcal{P}\left\{\frac{1}{n}\left|\sum_{t=1}^{n} h(t)\sum_{k=t}^{n} u(k)v(k-t)\right| \geq \alpha\right\} \\
&\quad \leq \mathcal{P}\left\{\frac{1}{n}\max_{1\leq t\leq n}\left|\sum_{k=t}^{n} u(k)v(k-t)\right| \geq \alpha/\eta\right\} \\
&\quad \leq n\mathcal{P}\left\{\frac{1}{n}\left|\sum_{k=1}^{n} u(k)v(k)\right| \geq \alpha/\eta\right\}.
\end{aligned}$$

Now, the last expression can be evaluated by means of a large deviations-type bound (see [40]). ∎

We now present a similar model in the set-valued setting. We do this by means of *a priori* known set, $\mathcal{W}$. A sample path for noise is any element that belongs to this set. The main problem with this approach is that it is difficult to enforce independence of the input from noise while still maintaining persistency. If

the set is too large, for instance, belonging to unit ball in ℓ_∞, independence between noise and input is lost. If the set is too small, for instance, belonging to unit ball in ℓ_2, we sacrifice persistency. Thus, any set-based model for noise has to balance these the two extremes. Nevertheless, set-valued models are useful for several reasons. In our case, we desire a uniform convergence property over all systems belonging to some set $\mathcal{I}$. The analysis of such algorithms becomes more streamlined with set-valued noise because in this case the problem formulation will require uniform convergence over both noise and real systems. Also, this allows for obtaining guaranteed error bounds which is aligned with robust control framework. Besides, it has been recently shown in [35], [29] that set-valued models for noise (when modeled appropriately) do not result in anymore conservatism than their stochastic counterparts. With these issues in mind, we will present the following model, which is appropriate and balances the extremes of independence from input and persistency:

$$\mathcal{W}_n = \left\{ w \in \mathbb{R}^n | \sup_{q \in \mathcal{Q}^m} \left| \frac{1}{\sqrt{n}\log(n)} \sum_{t=0}^{n} w(t) e^{iq(t)} \right| \leq 1 \right\} \tag{6}$$

where $\mathcal{Q}^m$ is the class of polynomials in t of order m over the field of reals. We verify that the noise model is rich enough to contain typical sample paths generated by an i.i.d. process.

Theorem 1 [Richness]: Suppose $x(0)$, $x(1)$, $x(2), \ldots$ is a discrete-time random process (white-Gaussian process or Bernoulli process) with mean zero and bounded variance. Then,

$$\mathcal{P}(P_n x(t) \in \mathcal{W}_n) \overset{n \to \infty}{\longrightarrow} 1 \tag{7}$$

where $\mathcal{W}_n$ is given by (6).

Proof: The proof is presented in the Appendix. ∎

The set-valued models are convex and balanced, i.e.,

$$w_1, w_2 \in \mathcal{W}_n \Longrightarrow \alpha w_1 + (1-\alpha) w_2 \in \mathcal{W}_n, \quad \alpha \in [0,1]. \tag{8}$$

This can be seen from the fact that the constraints imposed are all linear, i.e., the set $\mathcal{W}$ can be alternatively characterized as

$$\mathcal{L}_n w \leq 1 \tag{9}$$

where $\mathcal{L}_n$ is some linear operator. By straightforward calculation, it follows that the noise model is invariant with respect to filtering with a filter of ℓ_1 norm less than one.

IV. Problem Formulation and Discussion

The objective of this section is to formulate the problem of system identification for complex priors in instances where finite data limits choice of a complex model set. To realize our goal, we first present a simple example as a means of motivating the problem. Consider the example of an FIR model-parameterization, $\mathcal{G}_{FIR}$ of order m, i.e.,

$$\mathcal{G}_{FIR} = \left\{ G(\lambda) = \sum_{k=0}^{m-1} g(k) \lambda^k \right\} \tag{10}$$

as a means to model an LTI stable system, $T \in \mathcal{T}$ of Section III. In general, we will study identification for the following model sets:

$$\mathcal{G} = \left\{ G \in \mathcal{T} | G \equiv \left[\begin{array}{c|c} A & B \\ \hline C & 0 \end{array} \right], B \in \mathbb{R}^{m \times q} \right\}. \tag{11}$$

A and C are fixed *a priori* with m states, q inputs, and p outputs, and $\kappa = mq$, parameters of B to be estimated.

For the present, our aim is to study the behavior of the FIR estimates in relation to the real process in a way that is meaningful for robust control. To this end, we choose the ℓ_1 norm as the topology on the space of LTI systems to assess the performance of the estimate. From the perspective of robust control, the best model is one that is closest to the real process in the ℓ_1 norm, as this model has the smallest unmodeled dynamics associated with it. Thus, the best FIR sequence $G(T)$ is

$$G(T) = \underset{G \in \mathcal{G}_{FIR}}{\operatorname{argmin}} \|T - G\|_1 = \{t(k)\}_{k=0}^{m-1} \tag{12}$$

where $t(k)$ is the impulse-response sequence of the system T. Associated with this FIR model is the nonparametric error γ

$$\gamma = \|T - G(T)\|_1. \tag{13}$$

However, since we do not have access to the real system T, a more meaningful means of computing them from input–output data is necessary. We observe that every $T \in \mathcal{T}$ can be decomposed as a linear sum of the best model $G(T) \in \mathcal{G}_{FIR}$ and the residual error $\Delta(\lambda)$

$$\begin{aligned} T &= G(T) + \lambda^m \Delta, \quad \|\Delta\|_1 \leq \gamma, \\ G(T) &= (t(0), t(1), \ldots, t(m-1)) \end{aligned} \tag{14}$$

with $G(T)$ and $\lambda^m \Delta(\lambda)$ being "orthogonal" to each other. An equivalent notion of a best approximation holds for the general case as in the following proposition.

Proposition 1: Suppose $\mathcal{T}$ is a normed linear vector space $\mathcal{G}$, a subspace of $\mathcal{T}$, and $\mathcal{G}^\perp$ the annihilator of $\mathcal{G}$ in the dual $\mathcal{T}^*$ of $\mathcal{T}$. We let the notation argmin denote the set of all minimizers. Then, the following statements are equivalent:

1) $G \in \operatorname{argmin}_{G' \in \mathcal{G}} \|T - G'\|$;
2) $T = G + \Delta$ for some $G \in \mathcal{G}$, and $\langle \phi, \Delta \rangle = \|\Delta\|$ for some $\phi \in \mathcal{G}^\perp$, $\|\phi\| \leq 1$.

Thus, the duality theorem above characterizes the residual $\Delta = T - G(T)$ as an element that is aligned with the annihilator of $\mathcal{G}$. ∎

Now, returning to the FIR example, based on decomposition in (14), the input–output relation of (2) is rewritten as

$$\begin{aligned} y(t) &= G(T)u(t) + \lambda^m \Delta(\lambda) u(t) + w(t), \\ & t = 0, 1, \ldots, n \end{aligned} \tag{15}$$

with $\gamma = \|\Delta(\lambda)\|_1$. Suppose $\|G(T) - G^n(y,u)\|$ is the parametric error with G^n the estimate of $G(T)$ and, γ^n, the estimate for $\gamma(T)$ based on observations up to time n. Then, the

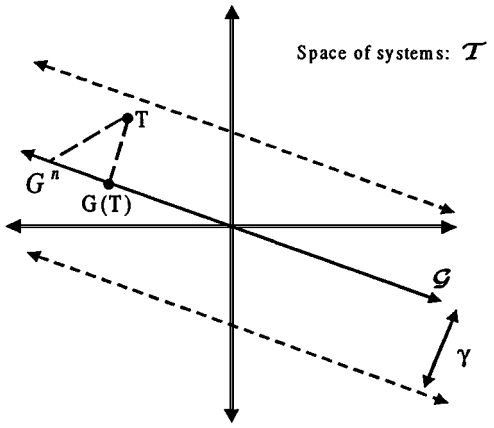

Fig. 1. Illustration of the real system T as an element belonging to an infinite-strip of width γ in the system space, $\mathcal{T}$.

parametric error is a function of the residual error and noise. To see this, observe that if the residual error is unknown and extremely large for the FIR model-parameterization, the measurements from time m onwards are not meaningful and we will have to rely on the first m measurements to estimate the parametric part $G(T)$. However, there are m parameters and m noisy measurements. It is well-known (see [16] and [28]) that, for this situation, the parametric estimates will also be unreliable even for a modest amount of noise. It is to be noted that with set-based model for noise, the estimation error will be large, but bounded. A way around the difficulty of obtaining poor estimates for the model is to assume that the chosen model-parameterization (in this instance the FIR parameterization) is "good," i.e.,

$$\mathcal{I}(\gamma) = \{T \in \text{LTI stable} \mid \|T - G\| \leq \gamma, \text{ for some } G \in \mathcal{G}\} \tag{16}$$

with the norm being the ℓ_1 norm and $\mathcal{G}$ is $\mathcal{G}_{FIR}$ in this particular instance. Note that this argument holds in the general situation too. The relationship between the real-process, T, and the model-subspace $\mathcal{G}$ is shown in Fig. 1. In a typical practical context one often has a good idea, either from physics, or through experience that a model-set characterizes the dominant dynamics well, although, there is some nonzero residual error. In such situations, the above assumption is reasonable. Also, barring pathological cases a good estimate of the residual error can be obtained for large enough data (this will be shown in an example at the end of the paper). Therefore, we can always choose another model-set if we are not happy with the residual error for the chosen model-set. We point out that this assumption is required on account of the need for error bounds based on finite data. The length of data required to "guarantee" that the estimates are within a pre-specified error bound is called sample-complexity. Clearly, if the "sample-complexity" were finite, then it will still be finite for any multiple of γ. In particular, for the asymptotic case, if we had finite-sample complexity, then we would converge in the parametric space for every $\gamma \in \mathbb{R}$, which is the set of all LTI stable systems.

We are now left to define the parametric error. The definition suggested earlier depends on the exact knowledge of the system T which is not meaningful when we do not know it. Therefore, the only alternative left is to define the parametric error as follows. There are two definitions corresponding to the set-based/probabilistic noise models

$$\begin{aligned}\alpha^n &= \sup_{w \in \mathcal{W}_n} \sup_{T \in \mathcal{I}(\gamma)} \|G(T) - G^n(y,u)\| \\ &= \sup_{w \in \mathcal{W}_n} \sup_{\|\Delta\| \leq \gamma} \|G(T) - G^n(y,u)\| \end{aligned} \tag{17}$$

or, equivalently, in the probabilistic case we can define as that number α^n such that

$$\mathcal{P}_w \left\{ \sup_{T \in \mathcal{I}(\gamma)} \|G(T) - G^n(y,u)\| \geq \alpha^n \right\} \leq \delta \tag{18}$$

for a pre-specified confidence level δ. *An algorithm is said to be consistent if the parametric error converges to zero.* Note that we require uniform convergence over all systems admissible in $\mathcal{I}(\gamma)$ which will be an important point when we consider minimum prediction error techniques. As a point of digression a related notion of convergence and consistency has been discussed in the set-membership literature (see [19] and references therein). Informally, an algorithm is robustly convergent if the parametric estimates converge to the "correct" parameters in the limit of vanishing residual-error and noise. Our notion of convergence is stronger, in the LTI context, in that we require the estimates to uniformly converge to the "correct" parameters even in the presence of nonzero residual-dynamics and noise.

We are now left to define the notion of sample-complexity. Again there are two definitions corresponding to set-valued and probabilistic models for noise. In the former case, the sample complexity is defined as that number $\mathcal{N}(\epsilon, \gamma, \kappa) \in Z^+$ such that, for every $\epsilon, \gamma > 0$ and model-set dimension κ, the following holds:

$$\begin{aligned}&\sup_{w \in \mathcal{W}_n} \sup_{T \in \mathcal{I}(\gamma)} \big| \|T - G^n\| - \|T - G(T)\| \big| \leq \epsilon, \\ &\quad \forall n \geq \mathcal{N}(\epsilon, \gamma, \kappa). \end{aligned} \tag{19}$$

Equivalently, for the probabilistic situation, the sample complexity is defined as the number $\mathcal{N}(\epsilon, \delta, \gamma, \kappa)$ such that

$$\begin{aligned}&\mathcal{P}_w \left\{ \sup_{T \in \mathcal{I}(\gamma)} \big| \|T - G^n\| - \|T - G(T)\| \big| \geq \epsilon \right\} \leq \delta, \\ &\quad \forall n \geq \mathcal{N}(\epsilon, \delta, \gamma, \kappa). \end{aligned} \tag{20}$$

Sample complexity is said to be polynomial if $\mathcal{N}(\epsilon, \gamma, \kappa) \cong \mathcal{O}((\gamma/\epsilon)^k)\kappa^l)$, $k, l \in Z^+$ for the set-valued situation and $\mathcal{N}(\epsilon, \delta, \gamma, \kappa) \cong \mathcal{O}((\gamma/\epsilon)^k)\kappa^l \log(\kappa/\delta))$, $k, l \in Z^+$ in the probabilistic situation.

We will now briefly discuss the classical minimum-prediction-error and the set-membership principles to see how they fit in and address the identification problem formulated in (19), (20). We will use the FIR example to illustrate some of the deficiencies in applying these principles.

A. Minimum Prediction Error (MPE) Principle

The MPE principle operates in the context of a stochastic noise models and this falls well within the experimental setup that we have in Section III. Typically, the MPE principle is used in situations where some model in the model set, $\mathcal{G}$, does adequately describe the real system, T. However, the MPE principle can be generalized to the case where the real system does not necessarily belong to the model-set $\mathcal{G}$. This situation is termed as *approximate system modeling* in the statistical identification literature (see [2] and the references therein). The principle remains the same and the suggestion is to pick that model from the model set, $\mathcal{G}$, which minimizes the prediction error (see [16] and [28]). This notion is also related to the ML principle (see [16] and [28]) where the model that maximizes the likelihood of data is chosen. The usual notion of consistency and convergence for the approximate modeling situation is defined in the sense of ML principle, in that, an estimator is said to be consistent if the estimates converge to that model which maximizes the likelihood function for the set $\mathcal{G}$. To simplify the discussion we use the FIR example with ℓ_1 topology on the system space, $\mathcal{T}$. We notice that the approximation in the sense of $\mathcal{H}^2$ and ℓ_1 are identical for FIR model-class. The corresponding MPE in this situation amounts to minimizing the squared-sum prediction error, i.e.,

$$G^n = \operatorname*{argmin}_{\mathcal{G}_{FIR}} \mathcal{V}(G) = \operatorname*{argmin}_{\mathcal{G}_{FIR}} \frac{1}{n}\sum_{t=0}^{n}\left|y(t) - \sum_{k=0}^{m-1} g(k)u(t-k)\right|^2. \quad (21)$$

This is the familiar least-squares algorithm, and $\mathcal{V}(\cdot)$ is the loss function. It is possible to show (see [16]) that the estimates do converge for stationary stochastic noise and with stronger assumptions on the allowable class of systems (in [15], exponential memory bounds are assumed for the residual error). It follows that:

$$\lim_{n\to\infty} G^n \to G^* = \operatorname*{argmin}_{G\in\mathcal{G}_{FIR}} \int_{\omega} (T-G)^*(\omega)\Phi_u(\omega) \cdot (T-G)(\omega)\, d\omega, \quad w.p.\ 1 \quad (22)$$

where Φ_u is the power spectral density of the input u. Thus, if the input is white, it follows that $G^* = G(T)$. It is now tempting to conclude that our goal has been accomplished. However, quite to the contrary there are several shortcomings in the MPE results which we briefly outline below.

1) Pointwise Convergence: The parametric convergence, as defined in the classical identification literature, is weaker and not taken uniformly over all the unmodeled errors, i.e., for every fixed system $T \in \mathcal{I}$ it is shown that

$$\limsup_n \sup_{G\in\mathcal{G}} |\mathcal{V}^n(G) - \mathcal{V}(G)| \longrightarrow 0.$$

Thus, the convergence is defined for each fixed "real system," $T \in \mathcal{T}$, and, in this sense, is pointwise. This is particularly important because in the absence of additional information and with finite data, we do not have access to the real system, and the unmodeled error has to be considered in the worst case.

2) Proof Technique: The convergence proof in the minimum-prediction error approach works on the following principle: the prediction error $\mathcal{V}^n(G)$ converges uniformly to $\mathcal{V}(G)$, and, therefore, the argmin $\mathcal{V}(G)$ converges as well. In view of the above uniform convergence issue, it is possible to generalize the formulation (use $\sup_{\|\Delta\|_1\le\gamma}\sup_{G\in\mathcal{G}_{FIR}}|\mathcal{V}^n(G)-\mathcal{V}(G)|$). However, convergence of the loss function is not uniform in the ℓ_1 topology

$$\limsup_n \sup_{\|\Delta\|_1\le\gamma}\sup_{G\in\mathcal{G}_{FIR}} |\mathcal{V}^n(G) - \mathcal{V}(G)| \not\longrightarrow 0.$$

Indeed, it can be easily shown that, even with white inputs, we will have

$$\sup_{\|\Delta\|_1\le\gamma}\sup_{G\in\mathcal{G}} |\mathcal{V}^n(G) - \mathcal{V}(G)| \ge \|\Delta\|_2.$$

Therefore, at the bare minimum, we will need to use a different proof technique to prove parametric convergence.

3) Error Bounds: Bias error and variance is typically used to characterize residual and parametric errors. Although these are related to parametric and nonparametric errors, they are not the same, and it is, in general, difficult to derive a direct correspondence. In general, MPE-based approaches have not satisfactorily addressed computation of such error bounds.

4) Sample Complexity: Perhaps the most important practical limitation of the minimum-prediction error approach is that, in the presence of unmodeled errors, the rate of convergence depends on the rate of convergence of the unmodeled error to its corresponding spectrum, i.e., we need to know the behavior of

$$\left|\frac{1}{n}\sum_{t=0}^{n}(\Delta u(t))^2 - \lim_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n}(\Delta u(t))^2\right|.$$

This can be seen in the convergence argument used. One uses the convergence of the loss function to characterize the behavior of the estimates. At any rate, this is a serious drawback in many real-time applications where one cannot wait for the asymptotic to set in order to derive a model for the system.

5) Colored Input: In (22), $G^* = G(T)$ if and only if the input were white. To see this, consider

$$y = G(Wu) + \Delta(Wu) + w$$

where the input u is a white noise process and W is a LTI filter. In this situation, the model output Gu is correlated with the output error, which in this case is given by $\Delta(Wu) + w$. The typical means of dealing with such situations in the minimum prediction error formulation effectively reduces to estimating ΔW. Alternatively, this can be thought of as whitening the output error before using the least-squares algorithm. However, this procedure requires estimating Δ, resulting in inflating sample-complexity, which we plan to avoid.

6) Other Topologies: Although $\mathcal{H}^2$ norm is very useful as a measure for systems, it is not particularly useful for describing unmodeled dynamics in the context of robust control. There are some special topological properties, with induced norms as an important case, that lend themselves easily to robust control applications. The minimum prediction error paradigm is focused on finding good descriptions that have good data-fit, while we need to find good models to fit systems. From this perspective, it is difficult to see how one can formulate a "system"-fitting problem for other topologies with a MPE based approach.

B. Set-Membership Identification

In previous work, significant effort has been focused on system-identification for robust control, and the formulation has come to be known as set-membership identification. This line of research goes back to [27] and [7]. The formulation has received much attention in [14], [8], [9], [21], [20], [18] and [31] (also, see references the therein) where both ℓ_1- and $\mathcal{H}_\infty$-error criterion were adopted. The premise behind the set-membership approach is that data is generated from one of several sets of possible behaviors. The goal is to choose the set of all those behaviors that could have generated the data. To fix the idea, we consider a simple situation which is well-known in the literature. The set of behaviors is given as follows:

$$y = G_0 u + v, \quad \|v\|_\infty \leq \delta, \quad G_0 \in \mathcal{G}_{FIR}. \tag{23}$$

From input–output data, which is a single sample path, the objective is to determine the set of all consistent behaviors, characterized by the above equation, i.e.,

$$G^n \in \mathcal{S}^n = \{G \in \mathcal{G}_{FIR} |\ |(y - Gu)(k)| \leq \delta, k = 0, 1, \ldots, n\}. \tag{24}$$

Thus, in a sense it is very similar to the ML principle in that seeking the "most likely" model from a chosen model class is replaced with computing all the models which are consistent with data. However, even this difference is understandable in view of the set-valued noise models. Historically, the interpretation of the above formulation is that of embedding the uncertainty in to the data generation process. To understand this point of view in the above instance, consider "identification" of the process T with an input, u, whose amplitude is bounded by unity. Consider the case when we know, *a priori*, that the process T is no more than at a distance δ from the space of FIR model-parameterization in the ℓ_1 norm. In this case, the "model" for the data-generation process contains the input–output behavior of the real-system. In other words, if $|u(t)| \leq 1$, we then observe that

$$\begin{aligned}\{(y,u) | y = Tu, \|T - G\| \leq \delta, \text{for some } G \in \mathcal{G}FIR\} \\ \subset \{(y,u) | y = Gu + v, \|v\|_\infty \leq \delta, \text{for some } G \in \mathcal{G}FIR.\}\end{aligned}$$

Thus, the unmodeled dynamics in the original problem can be embedded within an appropriate noise model. Be that as it may be, such an embedding has severe drawbacks. Fundamentally, the embedding is extremely coarse. It turns out that parametric error $\text{diam}(\mathcal{S}^n)$ is 2δ (see [31]) no matter what the input and its length and topologies and model-sets used. This is not very pleasing. For one, we have an intuitive feeling that, at least in the case of LTI stable systems, it should be possible to identify not only an FIR model, but the entire system if we had infinite data. One of the reasons for the conservatism comes from the fact that the data-generation process includes effects that could arise from time-varying and nonlinear behavior. The principal reason, however, is that there is a redundancy in the parameterization of the behaviors. That is, the uncertainty due to the noise v admits FIR models that are of norm smaller than δ. This accounts for the fact that the diameter of uncertainty can be no smaller than 2δ. It is fair to say that set-membership literature has not satisfactorily resolved the issue of uncertainty with respect to the model-parameterization and data effectively.

The principal difference between our formulation and set-membership identification is that, by resolving to minimize the uncertainty between the model-parameterization and real system, we are forced to explicitly incorporate the class (or space) in which the real system belongs. Furthermore, minimization of residual dynamics sets up a decomposition between the model parameterization and the residual dynamics. To summarize, we enforce a decomposition of the data into a triple: model, residual, and noise. To avoid redundancy in the description of data, the residual is separated from noise effects. Noise is independent of the process while the residual has a special structure that is associated with a chosen model set. Each of these elements account for different aspects in the experiment.

V. FIR Model-Sets and Input Design

In this section, we will derive several conditions that are sufficient to guarantee robust convergence of model parameters as defined by (19) and (20). We will derive these conditions in the context of FIR model set and, later, it will turn out that these conditions hold for identification of general model sets. As a point of reference, it has been well established in statistical identification that consistency of model estimates demands a persistency of excitation equal to the model dimension on the input. In contrast, in set-membership identification, a much more stringent condition is required. In particular, it is shown in [5] and [26] that in order to identify an FIR of order m, the input sequence will need to have every element of the set, $\{-1, 1\}^m$. We will see that in our particular situation we will need something far short of the latter requirement but much more than merely satisfying a persistency of excitation condition.

We will arrive at our conditions by testing several different inputs. First, we rewrite (2) in expanded notation by separating the terms into model and residual error, and see that

$$\begin{aligned} y(s) &= Tu(s) + w(s) \\ &= \overbrace{\sum_{k=0}^{m} t(k)u(s-k)}^{\text{FIR Model}} + \overbrace{\sum_{k=m+1}^{s} t(k)u(s-k)}^{\text{Residual Error}} + w(s). \end{aligned} \tag{25}$$

In the absence of noise, it is clear that a *pulse input* of unit amplitude is sufficient to recover the FIR model exactly. On the other hand, persistent noise can only be averaged out by a persistent input. However, a persistent input will also "excite" the unmodeled error. In particular, for the above equation, whenever $s \geq m+1$, the data also contains contributions from the unmodeled dynamics. For a *periodic input* u, with $u(s+l) = u(s)$, the data for $s = 0, l, 2l, \ldots$ can be written as

$$y(s) = \sum_{k=0}^{s} t(k)u(s-k) + w(s) = (t(0) + t(l) + \cdots)u(0) + (t(1) + t(l+1) + \cdots)u(1) + \cdots + w(s).$$

Thus, one obtains information only on linear combinations of the $t(k)$s, and not individual coefficients. It is, therefore, impossible to determine only the model coefficients, no matter how long the input signal and how large the length of the period. Therefore, no matter how large the period, in the worst case, the unmodeled error will couple with the model-set dynamics.

To get a better handle on the situation, we appeal to identification of one-parameter models. The situation when observations equations are linear and the residual error and noise belong to convex and balanced sets has been described in detail in [30]. It can be shown that, for the simple case of one parameter model, a linear algorithm is optimal in achieving the minimum parametric error. The least-squares algorithm is a linear algorithm, although not necessarily the optimal one. However, it does lead us to understand what conditions on input need to be satisfied. Suppose we apply the least squares algorithm (for the one-parameter FIR model). Then, the resulting unmodeled error contribution is given by

$$\left(\sum_{s=0}^{n}(u^2(s))^{-1}\sum_{s=0}^{n}u(s)\sum_{k=1}^{s}t(k)u(s-k)\right) = (r_u^n(0))^{-1}\sum_{s=1}^{n}r_u^n(s)t(s).$$

Therefore, for consistency to hold, we need the autocorrelation terms of the input to vanish uniformly. In principal, we can also use filtered inputs Fu where F is a causal stable filter and u has vanishing auto-correlation terms. This is based on the fact that

$$\left|(r_u^n(0))^{-1}\sum_{k=0}^{n}f(k)r_u^n(s-k)\right| \leq \|F\|_1 \max_{0<k\leq n}\left|\frac{r_u^n(k)}{r_u^n(0)}\right|.$$

However, the fact remains that, at a fundamental level, there be a signal u, such that the autocorrelation terms uniformly decay to zero. It follows that our objective is to search for such inputs. Periodic inputs do not satisfy the uniformly decaying correlation, and pulse inputs are not persistent. Therefore, we examine the *swept-sine* input as the next candidate

$$u(t) = \exp(i\alpha t^2), \quad \alpha \in \mathbb{R}, \qquad t = 0, 1, 2, \ldots \tag{26}$$

which, as the following Lemma shows, fails the test of uniformly decaying correlation.

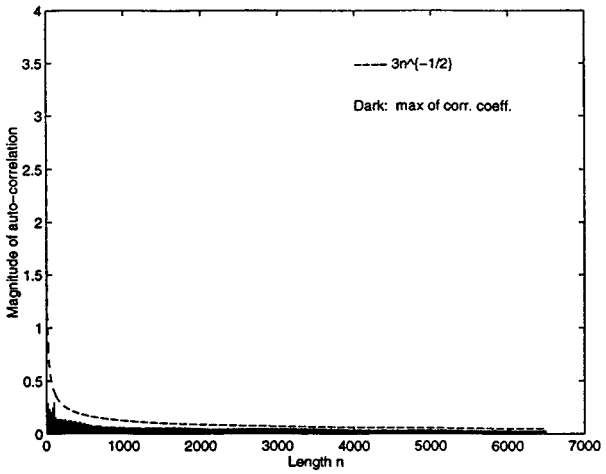

Fig. 2. Rate of decay of $\max_{0<\tau\leq n}|r_u^n(\tau)|$.

Lemma 2: For almost all real numbers $\alpha \in \mathbb{R}$, the worst case auto correlation of the sine sweep in (26) is bounded away from zero, i.e.,

$$\limsup_n \max_{0<\tau\leq n}|r_u^n(\tau)| \geq 1/2. \tag{27}$$

Proof: See the Appendix.

This means that in the worst case we will have the following property for the least-squares algorithm:

$$\limsup_n \sup_{T\in\mathcal{I}(\gamma)}\|G(T) - G^n\| \geq \gamma/2$$

Since, a chirp doesn't suffice, we introduce a *higher-order chirp*, i.e.,

$$u(t) = \exp(i\alpha t^3), \quad \alpha \in \mathbb{R}, \qquad t = 0, 1, 2, \ldots. \tag{28}$$

We have the following important property for such class of signals.

Theorem 2: The signal $u(\cdot)$ of (28) satisfies

$$\max_{0<\tau\leq n-1}|r_u^{n-1}(\tau)| \leq L(\alpha)\frac{\log(n)}{\sqrt{n}}, \qquad L(\alpha) > 0 \tag{29}$$

for almost all $\alpha > 0$, except for a set of Lebesgue measure zero.

Remark: The higher-order chirp is persistently exciting of infinite order.

Proof: See the Appendix for the proof.

Fig. 2 shows a plot of uniform decay of the autocorrelation terms with length of the higher-order chirp. Another signal that satisfies the uniformly decaying auto-correlation property is the *random i.i.d.* process, as the following lemma shows.

Lemma 3: Suppose $u(t)$ is a discrete i.i.d. Bernoulli random process with mean zero and variance 1, then

$$\mathcal{P}\left\{\max_{0<k\leq n}\|r_u^n(k)\|_\infty \geq \alpha\right\} \leq n\exp(-n\beta(\alpha)) \tag{30}$$

where $\beta(\alpha) = 1 + (1-\alpha/2)\log_2(1-\alpha/2) + (1+\alpha/2)\log_2(1+\alpha/2)$.

TABLE I

MPE	p.e. of order m
Set-membership	Galois sequence
MUD	Continuous spectrum

Application of the higher-order chirp input is equivalent to a random input in the following sense. If with a higher-order chirp (noise model is white Gaussian)

$$\mathcal{P}_w\left\{\sup_{T\in\mathcal{I}(\gamma)}||G^n-G(T)||\geq\epsilon\right\}\leq\delta \tag{31}$$

then, with a random input, we will have that the probability there is a random input u such that the same event

$$\left\{\mathcal{P}_w\left\{\sup_{T\in\mathcal{I}(\gamma)}||G^n-G(T)||\geq\epsilon\right\}\leq\delta\right\} \tag{32}$$

occurs has very high probability. In other words, almost every input sample drawn from a Bernoulli distribution will also typically satisfy uniform convergence of the parameter estimates. We summarize the input conditions in three different setting in Table I.

Revisiting the FIR Example: Equipped with these inputs, we revisit the FIR example and apply the familiar least-squares algorithm to form the estimates. We will decompose the least-squares algorithm in a novel way—as a combination of annihilation of unmodeled dynamics followed by averaging the noise—to derive a general method to target other cases. Consider the LS algorithm

$$\begin{aligned}G^n&=\operatorname*{argmin}_{g(\cdot)}\sum_{t=0}^{n}\left|y(t)-\sum_{k=0}^{m-1}g(k)u(t-k)\right|^2\\&=\frac{1}{n}(\Gamma_n(u)^*\Gamma_n(u))^{-1}\Gamma_n(u)^*Y_n.\end{aligned}$$

As a special example consider the problem of identifying the second impulse response coefficient of a system T using a least-squares estimator. Consider

$$Y_n=\begin{bmatrix}t(0)&0&\cdots&\cdots\\t(1)&t(0)&0&\cdots\\\vdots&\vdots&\vdots&\vdots\\t(n)&t(n-1)&\cdots&t(0)\end{bmatrix}U_n+W_n. \tag{33}$$

The impulse-response sequence of the optimal model is given by $(0,t(1),0,\ldots)$, and that of the unmodeled error is given by $(t(0),0,t(2),\ldots)$. Therefore, the annihilator for the unmodeled error is given by $v=(0,1,0,0\ldots)$. We attempt to annihilate the unmodeled error by using v as the realization of an anticausal system and filter the output, y with this system. This amounts to multiplying (33) by the following matrix:

$$Q=\begin{bmatrix}0&1&0&\cdots&\cdots\\0&0&1&0&\cdots\\\vdots&\vdots&\vdots&\vdots&\vdots\end{bmatrix}.$$

The resulting output upon algebraic manipulations can be written as the equation at the bottom of the page where $\Lambda(U_n)=(0,u(0),u(1),\ldots,u(n-2))^*$, $\Lambda^k(U_n)=\Lambda(\Lambda^{k-1}(U_n))$ and $\Lambda^{-1}(U_n)=(u(1),u(2),\ldots,u(n-1),0)$. Because, the input is chosen such that U_n and $\Lambda^k(U_n)$ has vanishing correlation property, it follows that the residual error is perpendicular to the model output. Therefore, we see that the minimizer of the following problem:

$$(U_n^*U_n)^{-1}U_nQY_n=\operatorname*{argmin}_{g\in\mathbb{R}}||QY_n-gU_n||\overset{n\to\infty}{\longrightarrow}t(1)$$

will converge to $t(1)$ which is the model in the model set that minimizes the unmodeled error. In summary, the anti-causal filtering results in separating the measured output into two components—model output and error—which are approximately orthogonal to each other with an approximate selection of the input. In this way the second step is the usual parametric estimation for the exact case. Fig. 3 illustrates the approach. The extension of this approach to other cases will become clear at the end of next section. For now, we outline the following steps:

1) obtain a decomposition of the system in to a model and unmodeled dynamics;
2) determine the annihilator (filter) of the unmodeled part;
3) pass the output through filter;
4) cross correlate the filtered output with the input.

Observe that the cross-correlation step and the annihilator step can be interchanged in the following sense. Instead of filtering the output y, we may multiply the output, Y_n, by the upper toeplitz matrix, $\Gamma_n(u)^*$ [see (1)], and then annihilate the resulting vector, $\Gamma_n(u)^*Y_n$ with the annihilator. This latter approach is particularly well suited for non-Hilbert spaces, such as banach spaces, for which an annihilator cannot be characterized explicitly. We only know that the residual-error is aligned with $\mathcal{G}^\perp$. In this situation, after denoising ($\Gamma_n(u)^*Y_n$) it is possible to formulate an optimization problem based on the alignment condition to compute the model-estimate, thus indirectly applying the annihilation step. We will see how this is done when we discuss ℓ_1 identification.

VI. Identification in $\mathcal{H}^2$

We consider system identification in an $\mathcal{H}^2$ space largely for historical reasons. In many instances, optimal estimates on other spaces can be derived using parametric estimates obtained by assuming a $\mathcal{H}^2$ topology. Finally, we believe the analysis that will follow has implications in the context of detection and estimation in signal processing, a line of thought which we do not

$$QY_n=\underbrace{t(1)U_n}_{\text{Model Output}}+\overbrace{(t(2)\Lambda(U_n)+t(3)\Lambda^2(U_n)+\cdots)+t(0)\Lambda^{-1}(U_n)+QW_n}^{\text{Error}}$$

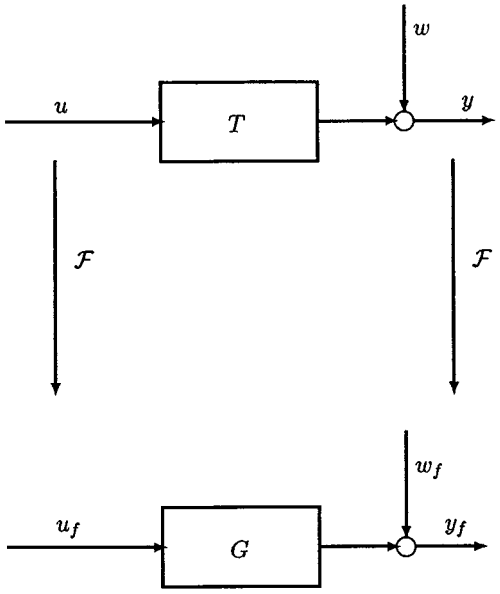

Fig. 3. Process of Identification.

explore here. With these prelimnaries, consider the space $\mathcal{H}^2$ which consists of the space of infinite sequences satisfying

$$\mathcal{H}^2 = \left\{ T \in \mathcal{T} |\ ||T||_2^2 = \sum_{k=0}^{\infty} \text{trace}(t(k)t(k)^*) < \infty \right\}$$

Suppose $T_1, T_2 \in \mathcal{H}^2$.The inner product is then defined as

$$\langle T_1, T_2 \rangle = \text{trace} \left(\sum_{k=0}^{\infty} t_1(k) t_2^*(k) \right). \tag{34}$$

On account of the fact that $\mathcal{H}^2$ is larger than the space of BIBO stable operators we restrict our *a priori* set as follows:

$$\mathcal{I}(\gamma) = \left\{ T \in \mathcal{H}^2 | \min_{G \in \mathcal{G}} ||T - G||_{\ell_1} \leq \gamma \right\} \tag{35}$$

where $\mathcal{G}$, as in (11), is a space of linearly parameterized stable finite-dimensional operators. Let

$$G(T) = \underset{G \in \mathcal{G}}{\text{argmin}} ||T - G||_2. \tag{36}$$

Since T belongs to $\mathcal{H}^2$, $G(T)$ is unique. We next discuss SISO and MIMO cases separately in the sequel.

A. SISO Systems and One-Parameter Models

The SISO identification problem follows by straightforward extension from the case of one—dimensional (1-D) parameterized models. To this end, let

$$\mathcal{G} = \{ \theta(1 - a\lambda)^{-1}; \theta \in \mathbb{R}, |a| < 1 \}$$

be the one dimensional subspace of SISO stable systems. The following proposition will enable us to characterize the unmodeled dynamics.

Proposition 2: Every $T \in \mathcal{I}(\gamma)$ can uniquely be written as

$$\hat{T}(\lambda) = \hat{\Delta}(\lambda) + \frac{\theta(T)}{1 - a\lambda}, \quad ||\Delta||_1 \leq L(a)\gamma, \quad \hat{\Delta}(a) = 0 \tag{37}$$

where $L(a)$ is a constant depending only on the pole location a.

Proof: Since $\hat{T}(\lambda)$ is an element of a closed Hilbert space $\mathcal{H}^2$ we know that $\hat{T}$ can be decomposed as

$$\hat{T}(\lambda) = \frac{\theta(T)}{1 - a\lambda} + \hat{\Delta}(\lambda), \quad \hat{\Delta} \in \mathcal{G}^{\perp}. \tag{38}$$

From Cauchy's integral formula, we have that

$$\hat{\Delta}(a) = \int \frac{\hat{\Delta}(\lambda)}{(1 - a\lambda)^*} d\lambda = 0. \tag{39}$$

Hence, now using Wiener's theorem on analytic functions and by hypothesis, we know that

$$\hat{\Delta}(\lambda) = (\lambda - a)\hat{\Delta}_a(\lambda) \tag{40}$$

for an arbitrary analytic function $\hat{\Delta}_a(\cdot) \in \ell_1$. It remains to show that $||\Delta||_1 \leq L(a)\gamma$. To prove this, consider an element $\theta^* \in \{\theta |\ ||T - \theta(1 - a\lambda)^{-1}||_1 \leq \gamma\}$, which exists by hypothesis. Then, we observe that $|\theta(T) - \theta^*| \leq 2\gamma$. Therefore, it follows that

$$\begin{aligned} &\left| \, ||T - G(T)||_1 - ||T - \theta^*(1 - a\lambda)^{-1}||_1 \right| \\ &\quad \leq \frac{L(a)}{2} |\theta(T) - \theta^*| \leq L(a)\gamma \end{aligned}$$

where $L(a)$ is some constant which only depends on the parameter, a. ∎

We will now present the application of the two step algorithm, at the end of which the estimator will be precisely described. As illustrated before, it consists of two steps—annihilation and cross correlation with the input—which we present below.

Step 1) Annihilation

Consider the input–output equation which we write elaborately for the sake of transparency as follows:

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n) \end{bmatrix} = \begin{bmatrix} \delta(0) & 0 & \cdots & 0 \\ \delta(1) & \delta(0) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \delta(n) & \cdots & \cdots & \delta(0) \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(n) \end{bmatrix}$$
$$+ \theta \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ a & 1 & 0 & \cdots & \cdots & 0 \\ a^2 & a & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ a^n & a^{n-1} & a^{n-2} & \cdots & \cdots & 1 \end{bmatrix}$$
$$\cdot \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(n) \end{bmatrix} + \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(n) \end{bmatrix}.$$

In the annihilation step, this equation is premultiplied by the following annihilation matrix $\mathcal{A}_n$:

$$\mathcal{A}_n = \begin{bmatrix} 1 & a & a^2 & \cdots & a^n \\ 0 & 1 & a & \cdots & a^{n-1} \\ \cdots & \ddots & \ddots & \ddots & \vdots \\ \cdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix}.$$

Denoting the resulting output by the symbol z, we get

$$\begin{bmatrix} z(0) \\ z(1) \\ \vdots \\ z(n) \end{bmatrix} = \mathcal{A}_n \begin{bmatrix} u(0) & 0 & \cdots & 0 \\ u(1) & u(0) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u(n) & u(n-1) & \cdots & u(0) \end{bmatrix} \begin{bmatrix} \delta(0) \\ \delta(1) \\ \vdots \\ \delta(n) \end{bmatrix} + \theta \mathcal{A}_n \mathcal{A}_n^T \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(n) \end{bmatrix} + \mathcal{A}_n \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(n) \end{bmatrix} \tag{41}$$

where, for the first term we have used the fact that convolution is commutative. Therefore, the convolution of Δ with u can be equivalently written as the convolution of u with Δ. We are now ready for our second step.

Step 2) Cross correlation

The second step consists of cross correlating the output z by the input u as follows:

$$\frac{1}{n} \sum_{k=0}^{n} z(k)u(k). \tag{42}$$

The above summation can be broken into three terms: the first corresponding to the unmodeled error, the second corresponding to the model output, and the third corresponding to the noise contribution. We first focus on the unmodeled term Λ_{umd}, which can be simplified in a straightforward manner to read

$$\Lambda_{umd} = \frac{1}{n}(u(0), \cdots, u(n))\mathcal{A}_n \cdot \begin{bmatrix} u(0) & 0 & \cdots & 0 \\ u(1) & u(0) & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ u(n) & u(n-1) & \cdots & u(0) \end{bmatrix} \begin{bmatrix} \delta(0) \\ \delta(1) \\ \vdots \\ \delta(n) \end{bmatrix}.$$

Again, since $(u(0), u(1), \ldots, u(n))\mathcal{A}_n$ is a convolution, we can interchange their order. This allows us to rewrite the unmodeled contribution as

$$\Lambda_{umd} = (1, a, a^2, \ldots, a^n) \cdot \operatorname{diag}\left(1, \frac{n}{n+1}, \ldots, \frac{n-j+1}{n+1}, \ldots, \frac{1}{n+1}\right) \cdot \begin{bmatrix} r_u^n(0) & r_u^n(1) & \cdots & r_u^n(n) \\ r_u^{n-1}(1) & r_u^{n-1}(0) & \cdots & r_u^{n-1}(n-1) \\ \vdots & \ddots & \ddots & \vdots \\ r_u^0(n) & r_u^0(n-1) & \cdots & r_u^0(0) \end{bmatrix} \cdot \begin{bmatrix} \delta(0) \\ \delta(1) \\ \vdots \\ \delta(n) \end{bmatrix}.$$

This equation, upon algebraic manipulations, simplifies to

$$\begin{aligned} \Lambda_{umd} &= \sum_{j=0}^{n} a^j \sum_{k=0}^{n} \frac{n-j+1}{n+1} r_u^{n-j}(k-j)\delta(k) \\ &= \sum_{j=0}^{n} a^j r_u^{n-j}(0)\delta(j) \\ &\quad + \sum_{\substack{j,k=0 \\ k \neq j}}^{n} \frac{n-j+1}{n+1} a^j r_u^{n-j}(k-j)\delta(k). \end{aligned} \tag{43}$$

The second term in the last expression can be easily disposed of by appealing to the fact that the autocorrelation coefficients for any lag not equal to zero uniformly decay to zero, i.e.,

$$\max_{0<\tau\leq n} \frac{n-j+1}{n+1} |r_u^{n-j}(\tau)| \leq C_0 \frac{\log(n-j)}{\sqrt{n-j}} \frac{n-j}{n} \leq C_0 \frac{\log(n)}{n}, \quad j = 0, 1, \ldots, n.$$

Therefore, the second term can be bounded as follows:

$$\left| \sum_{\substack{j,k=0 \\ k \neq j}}^{n} a^j r_u^{n-j}(k-j)\delta(k) \right| \leq C_0(1 + |a| + |a|^2 + \cdots + |a|^n)\|\Delta\|_1 \frac{\log(n)}{\sqrt{n}} \leq C\gamma \frac{\log(n)}{\sqrt{n}.}$$

Now, as far as the first term in last expression of (43) is concerned, we observe that

$$\frac{n-j+1}{n+1} r_u^{n-j}(0) = r_u^n(0) - \frac{1}{n+1} \sum_{k=n-j+1}^{n} u^2(k), \quad n \geq j > 0.$$

Therefore, the first term in last expression of (43) can be rewritten as

$$\sum_{j=0}^{n} a^j r_u^{n-j}(0)\delta(j) = r_u^n(0) \sum_{k=0}^{n} \delta(k)a^k - \frac{1}{n+1} \sum_{k=1}^{n} (k+1)a^k \delta(k)u^2(n-k+1).$$

The first term of the above expression approaches $\hat{\Delta}(a) = 0$ and, since $\|\Delta\|_1 \leq \gamma$, it follows that this term approaches zero at an exponential rate equal to $|a|^{n+1}$. The second term of the above expression is handled as follows:

$$\frac{1}{n+1} \left| \sum_{k=1}^{n} (k+1)a^k \delta(k)u^2(n-k+1) \right| \leq \frac{1}{n+1} \left\| \frac{d}{d\lambda} \frac{1}{1-a\lambda} \right\|_2 \|\Delta\|_2 \leq C_2 \frac{\gamma}{n}.$$

Putting all of the above computations together, we have

$$|\Lambda_{umd}| \leq C\gamma \frac{\log(n)}{\sqrt{n}}$$

It now remains to analyze contributions from model and noise terms in (42). The noise term can be handled in a straightforward manner

$$\frac{1}{n}(u(0),\ldots,u(n))\mathcal{A}_n \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(n) \end{bmatrix} = \sum_{k=0}^{n} a^k r_{uw}^n(k) \approx \mathcal{O}\left(\frac{\log(n)}{\sqrt{n}}\right).$$

Finally, we note that the model output $||(1/n)(u(0),\ldots,u(n))\mathcal{A}_n||_2^2$ (except for a scaling by the parameter $\theta(T)$) satisfies

$$\liminf_n \left|\left|\frac{1}{n}(u(0),\ldots,u(n))\mathcal{A}_n\right|\right|_2^2 > \sigma > 0.$$

In this way, all the terms except the model output vanish, and we are now ready to state our result.

Theorem 3: The least-squares estimator given by

$$\theta^n = (||P_n(1-a\lambda)^{-1}u||_2^2)^{-1}\sum_{i=0}^{n}\sum_{k=0}^{n-i} a^k y(i+k)u(i)$$

with the chirp input of (29) satisfies

$$\sup_{T\in\mathcal{I}(\gamma)}\sup_{w\in\mathcal{W}_n} ||\theta^n - \theta(T)||_2 \leq (L_0(a)\gamma + L_1(a))\left[\frac{\log(n)}{\sqrt{n}}\right] \tag{44}$$

where $L_0(a)$, $L_1(a)$ is a constant depending only on a.

Remark 1: The second term in (44) arises on account of noise, and is the same as when there is no unmodeled error. The first term arises on account of unmodeled dynamics. We see that the error uniformly approaches zero and scales linearly with the size of the unmodeled error. Upon closer reflection, we see that the error is completely independent of the norm of the system, i.e., $||T||$, which is a real improvement because the convergence in the parameters only depends on how good the approximation of the real system, with model set g, was in the first place. We also notice that the convergence of the parameters does not depend on the convergence of the residual errors in any manner. These answer some of the questions that we raised in Section IV in the context of discussion of MPE. ■

Remark 2: A similar result as in Theorem 3 holds in the stochastic setting too. This can be described for the LS estimator as follows:

$$\mathcal{P}_{w,u}\left\{\sup_{T\in\mathcal{I}(\gamma)} ||\theta^n - \theta(T)||_2 \geq \epsilon\right\} \leq \delta, \quad n \geq \left(\frac{\gamma}{\epsilon}\right)^2 \log\frac{1}{\delta}.$$

The proof presented here generalizes to any finitely parameterized family of stable models. At this point, it is interesting to compare the algorithm presented here with MPE and IV techniques. Although we have separated the annihilation and correlation into two distinct steps, the combined effect is essentially a least-squares algorithm. The question arises as to whether there is any advantage in executing the identification problem in two different steps. The results obtained here address some of the drawbacks pointed out in Section IV. We have shown that the convergence rate does not depend on the convergence rate of residual error. Also, the convergence result holds for a larger class of unmodeled dynamics (ℓ_1 here as opposed to exponentially decaying (M,ρ) type errors typically assumed in [15]). However, these differences, although significant, does not point to any advantage in a two-step algorithm. There are two principle contexts in which such algorithms gain importance. The first situation arises when the input is colored the second arises when the topology on the system space is a general Hilbert–Banach space. We will present, in this section, an example of the first situation. This situation arises commonly in the context of closed loop identification and will be elaborated in a forthcoming paper. The second situation is discussed in detail in subsequent sections.

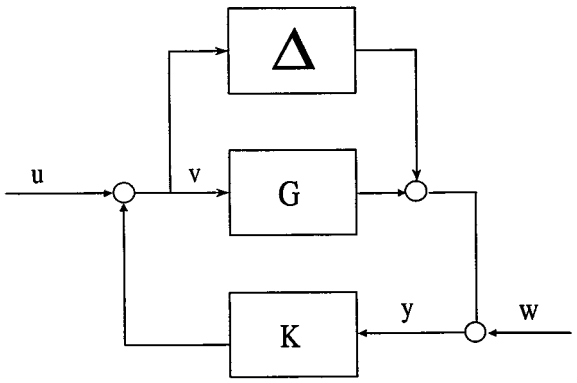

Fig. 4. Identification of model, G, in a closed loop situation by means of a dither signal, u.

Example 2: Consider the following numerical example where the system input is a filtered white noise signal u

$$y(t) = \overbrace{\frac{\theta}{1-.2\lambda}Wu(t)}^{\text{model output}} + \overbrace{(\lambda-.2)\Delta_1 Wu(t) + w(t)}^{\text{Residual Error}}$$
$$\theta = 1, \quad ||\Delta_1||_1 \leq 1, \quad W = \frac{1}{1-.9\lambda}, \qquad w \in \mathcal{N}(0,\sigma)$$

where the real system T has been decomposed into the model, $G = 1/(1-.2\lambda)$ and $\lambda - .2(\Delta_1)$ as before. We further postulate that the filter, W, is unknown except for the fact that $||W||_1 \leq \mu$. However, the input Wu into the system is known. This situation arises in the context of closed-loop identification, as shown in Fig. 4. The input u can be thought of as a dither signal employed for identification purposes. The second term in the above equation will then be replaced by $(\Delta v)(t)$, where v, is the input to the real system as shown in Fig. 4. The filter W is generally unknown in this situation, since it is an LFT of the controller with the unknown system.

Returning to the example at hand, we can rewrite the above equation in a manner suitable for formulating a prediction error principle. To this end, let, $\psi(t) = (1-.2\lambda)^{-1}Wu(t)$, $s(t) = (\lambda-.2)\Delta_1 Wu(t) + w(t)$. Then, we have that

$$y(t) = \psi(t)\theta + s(t).$$

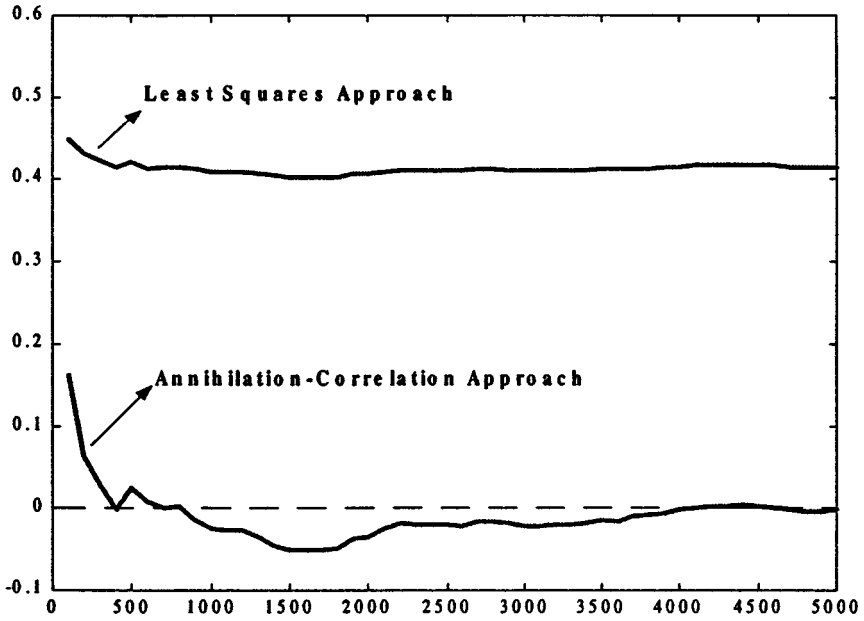

Fig. 5. Comparison of least squares and annihilation-correlation approaches for correlated inputs.

Now, the least squares works only when $\Xi\{\psi(t)s(t)\} = 0$. However, when the input to the system, Wu, is colored, this is no longer true. In this case, the residual error is correlated with the input. In point of fact, Fig. 5 shows the behavior of least squares estimates with the length of data and it is seen that the estimates do not converge to one (which is the right answer). The typical means by which this situation is handled in the MPE paradigm is to form a predictor, $\hat{y}(t/t-1;\theta)$, for $y(t)$, where the notation implies that the predictor should form the estimates based on previous measurements of inputs and outputs and the model parameter, θ. The parameters θ are then picked such that the mean-squared prediction error is minimized, i.e.,

$$\theta = \text{argmin}\frac{1}{n}\sum_{t=0}^{n}|y(t) - y(t/t-1;\theta)|^2.$$

Without going into the details, it suffices to say that such a procedure will amount to asking for identification of the entire system, i.e., the model along with the unmodeled error. Apart from the fact that this question will lead to an explosion in sample-complexity the solution is generally difficult to obtain. On the other hand, our objective is to identify only the model. In this regard, the annihilation followed by correlation step is extremely useful. We rewrite these steps for the sake of completion here. Let, $z(t)$, $\phi(t)$ be given by filtering the output, y, ψ, respectively, with the annihilating filter, as in(41). Then, the estimate for θ^n based on length n of data is given by

$$\hat{\theta}^n = \left(\sum_{k=0}^{n} u(k)\phi(k)\right)^{-1}\sum_{k=0}^{n} z(k)u(k).$$

The result is shown in Fig. 4, where it is seen that the estimates rapidly converge to the right solution. We will not ponder here as to the implication of this result, but only note that the annihilation-correlation principle has a scope outside the domain of traditional algorithms encountered in classical identification.

B. MIMO Systems and Finite-Dimensional Stable Models

The MIMO case turns out to be very similar to the one-parameter case, except for minor complications arising out of the need to be compatible with vectors as opposed to scalars. To simplify the notation, we introduce the following decomposition of the matrix B:

$$B = \sum_{j=1}^{q}\sum_{i=1}^{m}\alpha_{ij}H_{ij}, \qquad \alpha_{ij} \in \mathbb{R} \tag{45}$$

and H_{ij} is a $m \times q$ matrix with its ijth element equal to one, and all other elements equal to zero. Following on the lines of Section VI-A from Proposition 1, we have the following result.

Proposition 3: Every $T \in \mathcal{H}^2$ can be written as

$$T = \Delta(T) + G(T), \quad G(T) = \underset{G\in\mathcal{G}}{\text{argmin}}\,\|T - G\|_2,$$
$$\sum_{k=0}^{\infty} H_{ij}^*(A^*)^k C^*\delta(k) = 0, \quad \forall i,j.$$

The algorithm will be described next.

1) First, we filter the output y through the adjoint system as follows:

$$x_0(k-1) = A^*x_0(k) + C^*y(k), \quad x_0(n) = y(n)$$
$$z_{ij}^y(k-1) = u(k-1)^*(H_{ij}x_0(k-1)) + z_{ij}(k)$$
$$\forall i,j. \tag{46}$$

2) Next, the input is processed as follows, first through the model subspace:

$$x_{ij}(k) = Ax_{ij}(k-1) + H_{ij}u(k), \quad x_{ij}(0) = 0, i, j$$
$$y_{ij}(k) = Cx_{ij}(k)$$

and the output y_{ij} is then processed through the adjoint filter

$$v_{ij}(k-1) = A^*v_{ij}(k) + C^*y_{ij}(k), \quad x_0(n) = y_{ij}(n)$$
$$z_{ijkl}^u(k-1) = u(k-1)^*H_{kl}v_{ij}(k-1) + z_{ijkl}^u(k)$$
$$\forall k,l,i,j. \tag{47}$$

3) Determine $B^n = \sum_{i,j} \alpha_{ij} H_{ij}$ where α_{ij} is the solution to following set of mq equations:

$$\sum_{i=1}^{m}\sum_{j=1}^{q} \alpha_{ij} z^u_{ijkl}(0) = z^y_{ij}(0), i, j.$$

The following theorem states that the estimate B^n converges to the optimal.

Theorem 4: The estimator B^n satisfies

$$\sup_{T\in\mathcal{I}(\gamma)} \sup_{w\in\mathcal{W}_n} \|B^n - B\|_2 \le L\gamma pq \frac{\log(n)}{\sqrt{n}}$$

where L is some constant that depends only on the matrices A and C.

Proof: The proof is a direct extension of the 1-D SISO case, and is omitted.

VII. Hardy–Sobolov Spaces

As we have seen, the $\mathcal{H}^2$ metric lends to designing efficient algorithms for identification. Tractable robustness analysis problems are usually those that can be reduced to analyzing the stability of a system, G, that is perturbed by an element belonging to an uncertain unit ball in the topology under consideration. However, unit balls in $\mathcal{H}^2$ allow unstable systems and it is not clear how one can analyze problems of this nature. To deal with this problem we introduce a class of topologies that have the Hilbert–Space structure and yet satisfy the requirements of robust control. The robustness analysis and control for such topologies have been dealt with in [39], [32] and we deal with the identification problem here.

Suppose $\mathcal{T}$ is a normed linear vector space defined by

$$\mathcal{H}(r) = \left\{ T \in \mathcal{T} \left| \sum_{k=0}^{\infty} r(k)\, \text{trace}(t(k)t^*(k)) < \infty \right.\right\} \quad (48)$$

with the inner product defined by

$$\langle T_1, T_2\rangle = \sum_{k=0}^{\infty} r(k)\, \text{trace}(t_1(k)t_2^*(k)) \quad (49)$$

where $r(k)$ is a positive-weighting function that is monotonically increasing and satisfying the inequality

$$r(k) \ge k\log(k) + 1.$$

Then, the class of systems will all be in ℓ_1. This gives a Hilbert-space structure with norm denoted by $\mathcal{H}(r)$. This structure is useful in the context of identification. As a point of digression, observe that by setting $r(k) = \rho^k$, $\rho > 1$ we obtain the familiar class of functions that are analytic on the disc of radius $1/\rho$. The salient feature of these class of topologies is that they are stronger than both $\mathcal{H}_\infty$ and ℓ_1 as the following proposition shows.

Proposition 4: Suppose $T \in \mathcal{H}(\mathbf{r})$. Then,

$$\|T\|_{\mathcal{H}_\infty} \le \|T\|_1 \le C(r)\|T\|_{\mathcal{H}(r)}. \quad (50)$$

Proof: The proof follows by the usual Cauchy–Schwartz inequality, and is omitted. ∎

The norm may be motivated in the frequency domain for the case when $r(k) = k^2+1$. We denote such spaces by the symbol, $\mathcal{H}^{2,1}$, which is also known as the Hardy–Sobolov norm. The notation will become clear shortly. Consider systems T_1 and $T_2 \in \mathcal{H}^{2,1}$. By Parseval's theorem, we have

$$\langle T_1, T_2\rangle_{\mathcal{H}(r)} = \frac{1}{2\pi}\int_0^{2\pi} \hat{T}_1^*(\exp(i\theta))\hat{T}_2(\exp(i\theta))\, d\theta + \int_0^{2\pi} \frac{d\hat{T}_1^*(\exp(i\theta))}{d\theta}\, \frac{d\hat{T}_2(\exp(i\theta))}{d\theta} d\theta. \quad (51)$$

The reason for the notation should now be clear: the norm is defined by summing the $\mathcal{H}^2$ norm of the operator with the $\mathcal{H}^2$ norm of its derivative [see (51)]. We will mainly concern ourselves with such spaces for simplicity in notation. Also, in [39] and [32], we have shown that we can go even further by synthesizing nonconservative robust controllers against $\mathcal{H}^{2,1}$ type uncertainties. This follows from the fact that the image of an $\mathcal{H}^{2,1}$ ball is an ellipse at each frequency. Based on this fact, a novel IQC approach is developed for robustness analysis for structured perturbations. Furthermore, the set of all robustly stabilizing controllers is derived. These factors sufficiently justify developing identification results.

With these prelimnaries we next discuss the identification problem. As before, the class of systems belong to following prior:

$$\mathcal{I}(\gamma) = \{T \in \text{LTI} |\ \|T - G\|_{\mathcal{H}^{2,1}} \le \gamma, \text{ for some } G \in \mathcal{G}.\} \quad (52)$$

As in the previous section, the identification is carried out in two steps. The first step is the annihilation step. In order to do this, we need to characterize the separation between unmodeled dynamics and the model parameterization $\mathcal{G}$. In the familiar one-parameter model parameterization $\mathcal{G} = \theta/(1-a\lambda)$, this separation is particularly easy given by

$$T = \frac{\theta(T)}{1-a\lambda} + \Delta(T), \quad \Delta(T) = \sum_{k=0}^{\infty} \frac{\delta(k)}{k^2+1}\lambda^k$$
$$\sum_{k=0}^{\infty} \delta(k)\lambda^k = (\lambda - a)z, \qquad z \in \mathcal{H}^2.$$

In order to perform the annihilation step efficiently, we need to define the adjoint system. As the reader may recall from the $\mathcal{H}^2$ situation, the annihilation of the unmodeled error is obtained by running the input and the output of the system backward through the adjoint system. We now construct the adjoint system $\mathcal{F}$. Since, the model parameterization is finite the adjoint, $\mathcal{F}$ will also be finite and is the adjoint of the state-space realization of the impulse response function given by $(1, 2a, 5a^2, \ldots, (k^2+1)a^k, \ldots)$

$$\mathcal{F} \simeq \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right] = \left[\begin{array}{ccc|c} 3/a & -3/a^2 & 1/a^3 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \hline -2/a & 1/a^2 & -1/a^3 & 0 \end{array}\right]. \quad (53)$$

Following along the lines of $\mathcal{H}^2$ case, we are now ready to describe the algorithm which involves annihilation and subsequent averaging with the input as shown schematically in Fig. 3. Specifically, we have

$$x_1(k-1) = A_1 x_1(k) + B_1 y(k), \quad x_1(n) = 0 \tag{54}$$

$$z_1(k-1) = u(k-1)(C_1 x_1(k-1) + D_1 y(k-1)) + z_1(k) \tag{55}$$

$$x_2(k-1) = A_1 x_2(k) + B_1 u_{fl}(k), \quad x_2(n) = 0 \tag{56}$$

$$z_2(k-1) = u(k-1)(C_1 x_2(k-1) + D_1 u_{fl}(k-1)) + z_2(k) \tag{57}$$

where $u_{fl} = (1 - a\lambda)^{-1} u$. The estimate for $\theta(T)$ is given by

$$\theta^n = \frac{z_1(0)}{z_2(0).}$$

We can now extend the algorithm for the MIMO case as well. The following theorem characterizes the decomposition.

Proposition 5: Consider the model space $\mathcal{G}$ in (11). Every $T \in \mathcal{H}^{2,1}$ can be written uniquely as

$$T = G(T) + \Delta(T),$$
$$\sum_{k=0}^{\infty} (k^2+1) H_{ij} (A_1^*)^k C_1^* \delta(k)) = 0, \quad \forall i, j$$

where H_{ij} is as in (45).

Proof: The proof follows from the projection theorem on Hilbert spaces, and is omitted. ∎

We follow the steps for the $\mathcal{H}^2$ identification problem and obtain a similar estimator except for minor changes corresponding to the processing. First, we determine a state-space characterization, A_1, B_1, C_1, D_1 of the adjoint operator as we did in (53). Next, replace the matrices A, B, C, and$D$ by A_1, B_1, C_1, and D_1 in (46), (47). We have the following theorem for identification on $\mathcal{H}^{2,1}$.

Theorem 5: Consider the setup given in (2), and the estimator B^n derived as described above. Then,

$$\sup_{||\Delta||_{\mathcal{H}^{2,1}} \le \gamma} \sup_{w \in \mathcal{W}_n} ||B^n - B||_2 \le L\gamma pq \frac{\log(n)}{\sqrt{n}} \tag{58}$$

where L is a constant depending on matrices A and C.

Proof: The proof is identical except for the fact that an ℓ_1 bound on the unmodeled dynamics is unnecessary for $||\Delta||_{\mathcal{H}^{2,1}} \le \gamma$ by Proposition 4 suffices. ∎

A. Analysis of the Weighted Least-Squares Algorithm

Recall that, in Section IV, we had pointed out that there are several situations where the annihilation-correlation algorithm gains significance. We showed one situation when the inputs are correlated in Section V. We will present a second situation here where we show that for the general hilbert space topology presented in the previous section estimates based on generalized least-squares algorithms do not converge to the right solution.

To this end, consider the 1-D case, i.e., given $T \in \mathcal{H}$, the objective is to determine $\theta(T)$ where,

$$\theta(T) = \operatorname*{argmin}_{\theta \in \mathbb{R}} \left\| T - \frac{\theta}{1 - a\lambda} \right\|_{\mathcal{H}^{2,1}} \tag{59}$$

from inputs y and u. For simplicity we consider the noiseless case here, i.e.,

$$y(s) = Tu(s), \qquad s = 0, 1, \ldots.$$

The question arises whether there is a weighted least-squares criterion such as

$$\theta^n = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{t=0}^{n} \left(v^2(t) \left| y(t) - \frac{\theta}{1 - a\lambda} u(t) \right|^2 \right) \tag{60}$$

where $0 \le v(t)$, so that $\theta^n \longrightarrow \theta(T)$ uniformly over all $T \in \mathcal{I}(\gamma)$. The input, in this case, is assumed to be a Bernoulli process. To simplify the notation, let $A_n = (1, a, a^2, \ldots, a^n)^*$, $V_n = \operatorname{diag}\{v(0), v(1), \ldots, v(n)\}$, $R_n = \operatorname{diag}\{1, 2, \ldots, (n+1)^2 + 1\}$, $\Delta_n = (\delta(0), \delta(1), \ldots, \delta(n))^*$. The weighted least squares algorithm can be characterized using the projection theorem. The optimal estimate at time n satisfies

$$\sum_{s=0}^{n} v^2(s) \left(y(s) - \theta^n \left(\frac{1}{1 - a\lambda} u \right)(s) \right) \left(\frac{1}{1 - a\lambda} u \right)(s) = 0.$$

This implies that

$$\sum_{s=0}^{n} v^2(s) \left(\Delta(T)u(s) + (\theta(T) - \theta^n) \left(\frac{1}{1 - a\lambda} u \right)(s) \right) \cdot \left(\frac{1}{1 - a\lambda} u \right)(s) = 0.$$

Therefore, the expected value of parametric error $\theta^n - \theta(T)$ can be characterized by

$$\Xi\{(||V_n \Gamma_n(u) A_n||_2^2) |\theta(T) - \theta^n|\} = \Xi\{(\Gamma_n(u) A_n)^* V_n^2 \Gamma_n(u) \Delta_n\}. \tag{61}$$

We concentrate on the LHS of the above equality. It follows that, for the Bernoulli input, there is a constant $C \ge 1$ such that

$$C \left(\frac{||V_n \Gamma_n(u) A_n||_2^2}{\sum_{t=0}^{n} v^2(t)} \right) \le 1.$$

Therefore, the expression on the LHS of (61) can be simplified as

$$\Xi\{(||V_n \Gamma_n(u) A_n||_2^2 |\Theta(T) - \Theta^n|)\} \le \left(\frac{\sum_{t=0}^{n} v^2(t)}{C} \right) \Xi\{|\Theta(T) - \Theta^n|\}$$

Now, it follows in a straightforward manner that the RHS in (61) can be simplified to

$$\Xi\{(\Gamma_n(u) A_n)^* V_n^2 \Gamma_n(u) \Delta_n\} = A_n^* \operatorname{diag}\left\{ \sum_{s=0}^{n} v^2(t), \sum_{s=1}^{n} v^2(t), \ldots, v^2(n) \right\} \Delta_n = A_n^* H_n(v) \Delta_n$$

where $H_n(v) = \operatorname{diag}\{\sum_{s=0}^{n} v^2(t), \sum_{s=1}^{n} v^2(t), \ldots, v^2(n)\}$. Now, we also know that, for any ϵ, there is a length n such that

$$|A_n^* R_n \Delta_n| \le C_2 (n^2 + 1) |a|^{n+1} = \epsilon \longrightarrow 0$$

Therefore, letting $\sigma_n(v) = (\sum_{t=0}^{n} v^2(t))$, we have

$$\begin{aligned}\Xi\{|\theta(T) - \theta^n|\} &\geq \max_{|A_n^* R_n \Delta_n| \leq \epsilon} C_1 A_n^* \frac{H_n(v)}{\sigma_n(v)} \Delta_n \\ &\geq C_1 \max_{|A_n^* R_n \Delta_n| = 0} A_n^* \frac{H_n(v)}{\sigma_n(v)} \Delta_n \\ &\geq \beta \min_{v(\cdot)} \left\| A_n^* \left(R_n - \beta \frac{H_n(v)}{\sigma_n(v)} \right) \right\|_2 ||\Delta_n||_2 \geq C_3 ||\Delta_n||_2\end{aligned}$$

In this way, it can be seen that no matter what weights $v(\cdot)$ are chosen, the parametric error upon application of any weighted least squares algorithm will always be bounded away from zero. We have, therefore, established that the annihilation-correlation algorithm is essential for robust convergence in general hilbert spaces. In this way, we have shown two instances where the annihilation-correlation methodology succeeds while the traditional MPE paradigm fails.

VIII. Identification in ℓ_1

We now turn to the problem of identification in the ℓ_1 norm. As before, the process T is an element of $\mathcal{I}(\gamma)$ given by:

$$\mathcal{I}(\gamma) = \{T \in \text{LTI} \mid ||T - G||_{\ell_1} \leq \gamma, \text{ for some } G \in \mathcal{G}\} \quad (62)$$

and we want to find an estimator, $G^n \in \mathcal{G}$ based on a data record of length n that converges uniformly to the best-approximation [as stated in (19)]. In the previous situation, the Hilbert-space structure readily allows the construction of annihilators for the residual dynamics. Unfortunately, this is not possible in the ℓ_1 situation. To see this, we characterize the decomposition in the form of a proposition below:

Proposition 6: The following statements are equivalent:

1) $\theta(T) \in \text{argmin}_{\theta \in \mathbb{R}} ||T - \theta(1 - a\lambda)^{-1}||$;
2) $T = G + \Delta$ for some $G \in \mathcal{G}$ and $\langle v, \Delta \rangle = ||\Delta||_1$ where $v \in \{(\lambda - a)z | z \in \ell_\infty\} \cap \mathcal{B}\ell_\infty$;

Close observation reveals that the residual error is aligned with some element of $\mathcal{G}^\perp$. However, $\mathcal{G}^\perp$ being infinite dimensional, it is not possible to "narrow" down the "set" of all residual dynamics so that a finite set of annihilators for the "set" could be constructed. More importantly, the decomposition is not convex. A convex set in ℓ_1 when decomposed into the tuple of model and unmodeled dynamics is no longer convex as a tuple, as seen from the following example.

Example 3:

$$\mathcal{I} = \{T \in \ell_1 | T = \alpha T_1 + (1 - \alpha) T_2; \alpha \in [0, 1], \hat{T}_1(\lambda) = 1, \hat{T}_2(\lambda) = \lambda + (1 - \lambda/2)^{-1}\}$$

and $\mathcal{G} = \{\theta(1 - \lambda/2)^{-1} | \theta \in \mathbb{R}\}$. We see from Proposition 6 and the fact that T_1 is aligned with $\mathcal{G}^\perp$ that $G(T_1) = 0$. Similarly, we deduce that $G(T_2) = (1 - \lambda/2)^{-1}$. However, now

$$G(T(\alpha)) = (1 - \lambda/2)^{-1}, \quad \forall \alpha \neq 1.$$

It can also be verified that these minimizers are unique. So, the set of minimizers $\mathcal{G}_{\min}$ is given by

$$\mathcal{G}_{\min} = \cup G(T) = \{(1 - \lambda/2)^{-1}, 0\}$$

which is not a convex set.

Motivated by these issues, we solve the problem of ℓ_1 identification indirectly. However, we will see that there is a direct correspondence with the Hilbert-space case and the two-step algorithm. For the sake of simplicity, we only consider the one-pole case in this paper, i.e., $\mathcal{G} = \theta/1 - a\lambda$. The general case extends in straightforward way and is omitted. The estimator is obtained by solving the following convex optimization problem:

$$\theta^n = \underset{\theta \in \mathbb{R}}{\text{argmin}} \sum_{\tau=0}^{m} \left| \frac{n}{n - \tau} r_{yu}^n(\tau) - \theta \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right| \quad (63)$$

where $m \equiv \mathcal{O}(\log(n))$. The optimization problem can be readily converted to a linear programming problem by standard manipulations (see [22]). The reader will recognize that the formulation of the optimization problem is a consequence of the de-noising procedure discussed in Section V, i.e., is $\Gamma_n(u)^*(Y_n - \theta\Gamma_n(u)A_n)/n$ where A_n is as defined in Section VII-A and $\Gamma_n(u)$ is given by (1). The second step is approximately a "disguised" version of the annihilation step. To see this notice that,

$$\begin{aligned}\theta(T) = \underset{\theta \in \mathbb{R}}{\text{argmin}} ||T - \theta(1 - a\lambda)^{-1}||_{\ell_1} \\ \iff \sum_{k=0}^{\infty} sgn(\delta(k)) a^k = 0, \\ \delta(k) = t(k) - \theta(T) a^k, \qquad k = 0, 1, \ldots.\end{aligned}$$

The fact that the Banach-space optimization above is equivalent to the optimization problem of (63) will be proved next. The main point to notice is that the second expression in the above equation is exactly the "orthogonality" condition that was imposed between the residual error and the model subspace in the Hilbert-space situation. We have the following theorem for the behavior of the estimator.

Theorem 6: The estimate given by the solution to the optimization problem in (63) satisfies (64), shown at the bottom of the page, where c_1 and c_2 are constants resulting from unmodeled error and noise respectively.

Before proving the theorem, we wish to point out several implications of the above result. The last inequality follows by choosing $m = \log(n)$, thus, the sample complexity is $\mathcal{O}(1/\epsilon^2)$ *implying that* ℓ_1 *identification has polynomial sample complexity*. The sample complexity for ℓ_1 derived here has no bearing on the parallel result for ℓ_1 in the context of set-membership identification (see [5] and [26]), where the sample complexity was shown to be exponential. However, it is still worth pondering the difference between the two approaches. As the reader may recall, in set-membership identification, the uncertainty due to unmodeled error has a redundancy with the

$$\sup_{T \in \mathcal{I}(\gamma)} \sup_{w \in \mathcal{W}_n} \left| \, ||T - \theta^n(1 - a\lambda)^{-1}||_1 - ||T - \theta(T)(1 - a\lambda)^{-1}||_1 \right| \leq \frac{(c_1\gamma + c_2)\log^2(n)}{\sqrt{n}} \quad (64)$$

parameterization. This accounts for the fact that the minimum achievable diameter of uncertainty is 2δ and this results in exponential sample complexity. These factors prove to be critical in obtaining vastly different results.

Proof: The theorem is proved in several steps.

Lemma 4: Let θ_0^m be the solution to the following optimization problem:

$$\theta_0^m = \text{argmin}||P_m(T - \theta(1-a\lambda)^{-1})||_1.$$

Then,

$$\left| \, ||T - \theta_0^m(1-a\lambda)^{-1}||_1 - ||T - \theta(T)(1-a\lambda)^{-1}||_1 \right| \leq C_0\gamma a^m.$$

Proof: The following set of inequalities follows by definition:

$$||P_m(T - \theta_0^m(1-a\lambda)^{-1})||_1 \leq ||P_m(T - \theta(T)(1-a\lambda)^{-1})||_1$$

and

$$||T - \theta(T)(1-a\lambda)^{-1}||_1 \leq ||T - \theta_0^m(1-a\lambda)^{-1}||_1.$$

Together, these result in the following sequence of inequalities:

$$\begin{aligned} &||P_m(T - \theta_0^m(1-a\lambda)^{-1})||_1 \\ &\quad + ||(I-P_m)(T - \theta(T)(1-a\lambda)^{-1})||_1 \\ &\quad \leq ||T - \theta(T)(1-a\lambda)^{-1}||_1 \leq ||T - \theta_0^m(1-a\lambda)^{-1}||_1. \end{aligned}$$

Therefore, we have that

$$\begin{aligned} 0 &\leq ||P_m(T - \theta(T)(1-a\lambda)^{-1})||_1 \\ &\quad - ||P_m(T - \theta_0^m(1-a\lambda)^{-1})||_1 \\ &\leq ||(I-P_m)(T - \theta_0^m(1-a\lambda)^{-1}||_1 \\ &\quad - ||(I-P_m)(T - \theta(T)(1-a\lambda)^{-1})||_1 \\ &\leq a^m||(\theta_0^m - \theta(T))(1-a\lambda)^{-1}||_1. \end{aligned}$$

Now, from the definition of $\mathcal{I}(\gamma)$ (see (62), it follows that θ_0^m and $\theta(T)$ should satisfy $|\theta_0^m - t(0)| \leq \gamma$ and $|\theta(T) - t(0)| \leq \gamma$ respectively. Therefore, $|\theta_0^m - \theta(T)| \leq 2\gamma$. The result now follows. ∎

Our next step will be to relate θ_0^m to θ^n of (63). To do this, we rewrite θ^n in (63) as

$$\theta^n = \theta_0^m + \partial\theta^n.$$

With this notation, we have the following lemma.

Lemma 5:

$$|\partial\theta^n| \leq c\gamma$$

where c is some constant.

Proof: First, we observe that

$$\begin{aligned} &\overbrace{|\partial\theta^n| \sum_{\tau=0}^{m} \left| \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right|}^{\approx c(a)|\partial\theta^n|} \\ &\quad \leq \sum_{\tau=0}^{m} \left| \frac{n}{n-\tau} r_{yu}^n(\tau) - (\theta_0^m + \partial^n) \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right| \\ &\qquad + \sum_{\tau=0}^{m} \left| \frac{n}{n-\tau} r_{yu}^n(\tau) - \theta_0^m \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right| \\ &\quad \leq 2 \sum_{\tau=0}^{m} \left| \frac{n}{n-\tau} r_{yu}^n(\tau) - \theta_0^m \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right|. \end{aligned}$$

The second inequality follows from the fact that the first term in the middle expression is the optimal solution to the optimization problem in (63). Next, notice that

$$r_{yu}^n(k) = \sum_{k=-\tau}^{n-\tau} t(k+\tau) r_u^{n-\tau}(k) + r_{wu}^n(k).$$

Now, from Lemma 4, Theorem 2 and (6), it follows that

$$\begin{aligned} &\sum_{\tau=0}^{m} \left| \frac{n}{n-\tau} r_{yu}^n(\tau) - \theta \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right| \\ &\qquad \leq c_0\gamma a^m + c_1\gamma + \frac{m(c_2\gamma + c_w)\log(n)}{n} \end{aligned}$$

(where, c_w, appears on account of the noise contribution). It now follows from the above equation and Lemma 4 that

$$||T - \theta^n(1-a\lambda)^{-1}||_1 \leq c\gamma. \tag{65}$$

With these prelimnaries, we are now ready to prove the theorem. Consider

$$\begin{aligned} &\min_{\theta\in\mathbb{R}} \sum_{\tau=0}^{m} \left| \frac{n}{n-\tau} r_{yu}^n(\tau) - \theta \sum_{k=-\tau}^{n-\tau} a^{k+\tau} r_u^{n-\tau}(k) \right| \\ &\quad \leq \min_{\theta\in\mathbb{R}} \left(\sum_{k=0}^{m} \left| P_m(T - \theta(1-a\lambda)^{-1}) \right. \right. \\ &\qquad \left. \left. + \sum_{\tau=0,\tau\neq k}^{n} (t(\tau) - \theta a^\tau) r_u^{n-k}(\tau - k) \right| \right) \\ &\qquad + \sum_{k=0}^{m} |r_{wu}^n(k)| \\ &\quad \leq \frac{m(l_1\gamma + l_2)\log(n)}{\sqrt{n}} + \min_{\theta\in\mathbb{R}} ||P_m(T - \theta(1-a\lambda)^{-1})|| \end{aligned}$$

where l_1, l_2 are constants. The last inequality follows from (65). Setting $m \approx \log(n)$, the result now follows in a straightforward manner. ∎

The important fact to be noticed is that the sample-complexity is independent of the size of the parameter. Thus even though the prior $\mathcal{I}(\gamma)$ is unbounded it does not affect the identification process. As in the $\mathcal{H}^{2,1}$ case, we wish to understand whether an appropriate signal space optimization will lead to the optimal set of parameters. Two types of error metrics are of interest—one based on minimizing the sum of the absolute values

$$\min_{\theta} \sum_{t=0}^{n} |y(t) - \theta((1-a\lambda)^{-1}u)(t)|$$

and the other based on minimizing the maximum of the absolute value

$$\min_{\theta} \max_{0\leq t\leq n} |y(t) - \theta((1-a\lambda)^{-1}u)(t)|.$$

These problems are hard to analyze and we resorted to simulating the three different approaches (MUD and the above two).

Example 4: For simplicity, we chose the system T to have an impulse response equal to $-1, 1, 0\ldots$. The model parameterization used was $\theta(1-\lambda/3)^{-1}$. For this parameterization, the optimal parameter in the sense of ℓ_1 is unique and turns out to be $\theta(T) = 1$. We applied two different inputs: the random input

TABLE II

Process T			
Algorithm	θ_r^n	θ_c^n	Estimated-Error
MUD	1.00	1.00	.11
$\|\cdot\|_1$	.6	.5	—
$\|\cdot\|_\infty$	$\approx .5$	.8	—

and the following input which is the real part of the higher order chirp, i.e.,

$$u(t) = \Re\left(\exp\left(i\frac{\sqrt{5}+1}{2}t^3\right)\right).$$

The output y was obtained by adding random noise of standard deviation equal to 0.2, i.e.,

$$y(t) = Tu(t)+w(t), \qquad t = 0, 1, \ldots, n; \quad w(\cdot) \equiv N(0, 0.2).$$

The input length n was chosen to be 3000. MUD was implemented by letting $m = 5$ in (63). The estimate corresponding to the random input is denoted θ_r^n and that from chirp input is denoted θ_c^n. The following table enumerates the results where estimated error corresponds to the computation of the error based on (64).

We observed that minimizing the maximum of the absolute value was not stable with random excitation in that the parameter values kept oscillating. The value in Table II corresponds to the average value obtained.

Although the numerical example cannot be equated with a proof, it should point to the fact that signal-space optimization will not necessarily result in parameters that are close to optimal ones in terms of minimizing the ℓ_1 norm distance between the model parameterization and the real system.

IX. Identification in Practice: Estimation of Residual Dynamics

Real-world applications demand that there is a reasonable way to validate and estimate the parameters used in the prior information. In fact, it is usually quire hard to characterize a mathematical linear space to which the real process belongs, as a consequence, it is hard to verify the validity of the prior information. In practice, however, such principles are adapted to the application at hand with the hope of creating reasonable models of the process. The effectiveness of these principles stem from the ability to provide models with uncertainty descriptions. In this section, we will describe how the error bounds derived in the last section allow us to estimate both the parametric error (error in the space $\mathcal{G}$) and the nonparametric error (an estimate of the prior γ).

We will only concern ourselves with the validation and estimation of parametric and nonparametric errors for the identification of systems defined on Hardy–Sobolov spaces. It is, in this space, that the problem reduces to a quadratic optimization problem. To this end, we define

$$z_w(t) = y(t) - Gu(t) - \Delta u(t), \qquad t = 0, \ldots, n,$$
$$\langle G, \Delta\rangle_{\mathcal{H}^{2,1}} = 0$$

with the subscript w denoting the dependence on noise w. Since G and Δ are convolution operations in the time-domain it is natural to look at the frequency domain. In the frequency domain the orthogonality between G and Δ will also have an appropriate transformation. As we will soon see that this condition will be redundant for the problem we are about to solve. Taking the corresponding DFT of the above equation, we get

$$\hat{z}_w(k) = \hat{y}(k) - \hat{G}(k)\hat{u}(k) - \hat{\Delta}(k)\hat{u}(k),$$
$$k = 0, 1, \ldots, n$$

where $\hat{g}(k)$, $\hat{\Delta}(k)$, $\hat{u}(k)$, $\hat{y}(k)$ are the n-point DFTs of G, Δ, $u(\cdot)$, $y(\cdot)$, respectively. A point of concern is that the DFT operation will not diagonalize causal operations as above as written above. For large enough data, the error is insignificant and the above equation holds. We now need a good frequency domain estimate for the Sobolov norm. Recall, that the Hardy–Sobolov norm for Δ is given by

$$\|\Delta\|_{\mathcal{H}^{2,1}} = \int_0^{2\pi} \hat{\Delta}(\omega)\hat{\Delta}(\omega)^*\, d\omega + \int_0^{2\pi} \frac{d\hat{\Delta}(\omega)}{d\omega}\frac{d\hat{\Delta}(\omega)}{d\omega}^*\, d\omega.$$

It can be readily shown that, by using Cauchy–Schwartz inequality, the estimate γ^n given by

$$\gamma^n(\hat{\Delta}(\cdot)) = \frac{1}{n}\sum_{k=0}^{n-1}(\hat{\Delta}(k))^2 + \frac{n}{4\pi^2}\sum_{k=0}^{n-1}(\hat{\Delta}(k) - \hat{\Delta}(k+1))^2$$

converges to the $\mathcal{H}^{2,1}$ norm as $\mathcal{O}(1/n)$. We organize the data in to J frequency bins by filtering them through ideal filter banks of size $2\pi/J$. It can be shown that for white noise, the following holds (see [23]):

$$\langle V_j, S_{z_w}^n \rangle \leq 1 + \beta_n, \qquad j = 0, 1, \ldots, J-1$$

where $S_{z_w}^n$ is the power spectrum of the signal z_w. Each of these constraints and the cost function γ are quadratic expressions of their variables. We now state the following optimization problem:

$$\gamma^n = \min \frac{1}{n}\sum_{k=0}^{n-1}(\hat{\Delta}(k))^2 + \frac{n}{4\pi^2}\sum_{k=0}^{n-1}(\hat{\Delta}(k) - \hat{\Delta}(k+1))^2$$

subject to

$$\langle V_j, S_{z_w(t)}^n\rangle = \frac{1}{nk}\sum_{i=jk}^{j(k+1)-1}(\hat{y}(i) - \hat{g}(i)\hat{u}(i)\theta - \hat{\Delta}(i)\hat{u}(i))^2 \leq 1 + \beta_n, \qquad j = 0, 1, \ldots, J-1.$$

It should be clear that the minimizing solution will satisfy the orthogonal property. If not, let the solution be $\hat{\Delta}$ and G^n respectively. By hypothesis, the minimizer $\hat{\Delta}$ will have a component in the direction $\alpha P_n G_0$, $G_0 \in \mathcal{G}$, $\alpha \in \mathbb{R}$. Consider now $\hat{\Delta} - \alpha G_0$ and $G^n + \alpha G_0$ as a candidate solution. These satisfy the constraints of the optimization problem and moreover $\hat{\Delta} - \alpha G_0$ has a norm smaller than $\hat{\Delta}$.

The proposed optimization problem can be readily converted to an LMI (see [1] for details). An alternative option is to extend

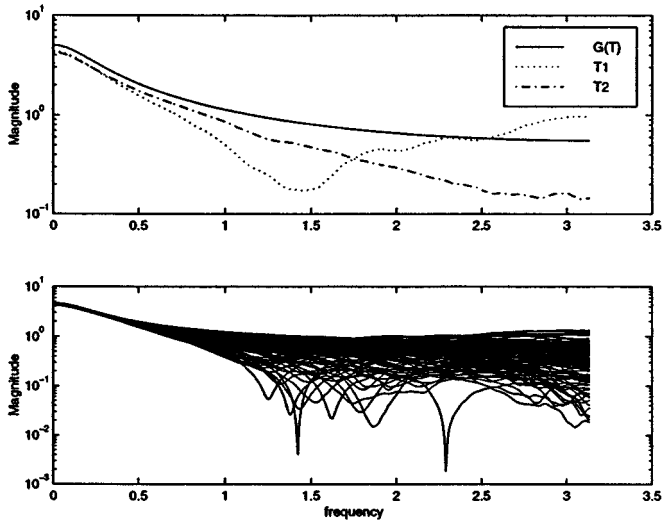

Fig. 6. Frequency response of T_1 and T_2 used in the example and randomly generated sample-ball of systems that are at a unit magnitude from $1/1-0.8\lambda$.

the algorithm in [12] for the current situation. In order to illustrate the ideas involved in identification of the model and estimation of unmodeled dynamics we consider a simple example.

Example 5: Consider the system, T, whose magnitude response in the frequency domain is shown in Fig. 6. This system has been chosen as follows:

$$T(\lambda) = \frac{1}{1-0.8\lambda} + \Delta(\lambda), \quad \Delta(\lambda) = \sum_{k=0}^{200} \delta(k)\lambda^k$$

where,

$$\delta(k) = \left(\sum_{k=0}^{200} \delta_1^2(k)(k^2+1) \right)^{-1/2} \frac{\delta_1(k)}{k^2+1}$$

where $\delta_1(k) = (\lambda - a)v(k)$ and v, a vector of length 200, was selected using a random number generator. Notice that, in so doing, we have normalized $\Delta(\lambda)$ to be of norm 1. The specific example of a randomly generated unmodeled dynamics is of no significance. We have done so here for the sake of simplicity. The results hold for any Δ of norm smaller than 1. We have also chosen the optimal value of $\theta(T)$ to be 1 for ease of exposition. The class of all such systems has been shown in Fig. 6.

We apply a random Gaussian input of mean 0 and standard deviation 1. Noise is simulated as a white Gaussian process of mean 0 and standard deviation of 0.3. The input–output data is of length 1000, i.e., we have,

$$y(t) = Tu(t) + w(t), \quad y = 0, 1, \ldots, n;$$
$$w(\cdot) \equiv N(0, 0.3)$$

From the input–output data $\{y, u\}$ we wish to pick a model in the class $\mathcal{G} = \{\theta/1-0.8\lambda, \theta \in \mathbb{R}\}$ that minimizes the unmodeled-error, i.e.,

$$\theta(T) = \underset{\theta}{\operatorname{argmin}} \left\| T - \frac{\theta}{1-0.8\lambda} \right\|_{\mathcal{H}^{2,1}}$$

Of course we do not have knowledge of the process, T, in order to compute $\theta(T)$ and only have access to finite data. Moreover, we want the procedure to work uniformly over all systems T that are within a bounded distance from the model-class, $\mathcal{G}$. To illustrate this fact, we solve the problem for two cases: when the real process is T_1 and when the real process is in T_2 which have magnitude response as shown in Fig. 6.

a) Identification of model-parameters: To simplify the notation we denote the annihilator in state-space as follows:

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{ccc|c} 3.75 & -4.6875 & 1.9531 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline -2.5 & -1.5625 & -1.9531 & 1 \end{array}\right] \quad (66)$$

With this notation we apply the following algorithm of Section VII:

$$x_1(k-1) = Ax_1(k) + By(k), \quad x(n) = 0 \quad (67)$$
$$z_1(k-1) = u(k-1)(Cx_1(k-1) + Dy(k-1)) + z_1(k) \quad (68)$$
$$x_2(k-1) = Ax_2(k) + Bu_{fl}(k), \quad x(n) = 0 \quad (69)$$
$$z_2(k-1) = u(k-1)(Cx_2(k-1) + Du_{fl}(k-1)) + z_2(k) \quad (70)$$

where $u_{fl} = (1-0.8\lambda)^{-1}u$. The estimate for $\theta(T)$ is given by

$$\theta^n = \frac{z_1(0)}{z_2(0)}.$$

These results were compared with the least squares and the weighted least-squares approach. The actual error is the difference $\theta(T) - \theta^n$, and the estimated error is that based on data and *a priori* assumptions. To apply the least-squares algorithm we first prefiltered the input with $1/(1-0.8\lambda)$. We then estimated the best θ^n that minimized the least-squares error between the filtered input and the measured output. We experimented with several different weights for the weighted least-squares problem and found them to be worse than uniform weighting. Recall that we made observations to this effect in Section VII-B. As a sample, we have used the weight $(1, 2a, 5a^2, \ldots)$ in the Table III. The procedure applied in this paper is denoted MUD, for minimizing-unmodeled-dynamics, the least squares by LS and the weighted least squares by WLS.

We see two issues at stake: the parametric error estimates and the unmodeled error. LS and WLS provide poor parametric-error-estimates in addition to increasing the level of unmodeled dynamics. In other words, it is extremely unlikely that the real-process can be realistically accounted for within the error bounds.

b) Estimating unmodeled error: In order to estimate the unmodeled error for when the process is T_1 we solve the optimization problem presented in this section. For simplicity of exposition we only use a single-filter bank for our constraints. In such a situation the upper-bound scales as the noise-energy.

TABLE III

Process T_1					
Algorithm	θ	% Error	Actual Error	Estimated-Error	unmodeled-error
MUD	.97	3	.03	.06	1.0
LS	.67	33	.33	.0002	1.9
WLS	.60	40	.4	.0002	2.2
Process T_2					
Algorithm	θ	% Error	Actual Error	Estimated-Error	unmodeled error
MUD	.99	1	.01	.06	1.01
LS	.64	36	.36	.0002	1.95
WLS	.58	42	.42	.0001	2.3

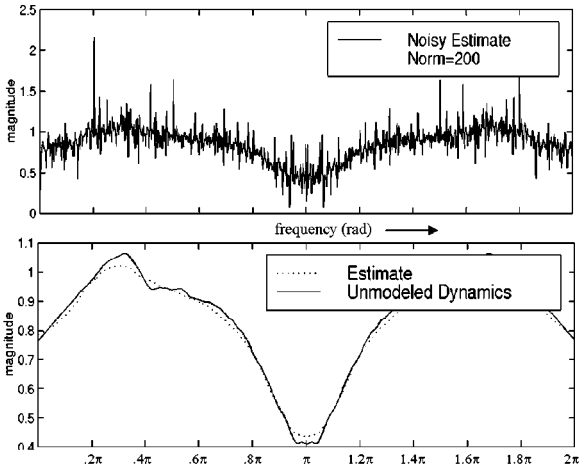

Fig. 7. Noisy estimate of unmodeled dynamics and frequency response of actual and estimated unmodeled error.

We also fix the parametric estimate θ to be θ^n obtained using the MUD algorithm. This simplifies our problem greatly and the solution can be obtained by applying the technique presented in [12]. We first write the cost function using Lagrange multipliers

$$\sum_{k=0}^{n-1}(\hat{\Delta}(k))^2+\frac{n}{4\pi^2}\sum_{k=0}^{n-1}(\hat{\Delta}(k)-\hat{\Delta}(k+1))^2 + \eta\left(\frac{1}{n^2}\sum_{j=0}^{n-1}(\hat{y}(j)-\hat{g}(j)\hat{u}(j)\theta^n-\hat{\Delta}(j)\hat{u}(j))^2-\alpha_n^2\right). \tag{71}$$

We make the following notations for ease of exposition:

$$S_{yu}^n=\text{diag}\{S_{yu}^n(0),S_{yu}^n(1),\dots S_{yu}^n(n-1)\},$$
$$S_u^n=\text{diag}\{S_u^n(0),S_u^n(1),\dots S_u^n(n-1)\},$$
$$\hat{b}=(S_{yu}^n-\text{diag}\{\hat{g}(0),\hat{g}(1),\dots\hat{g}(n-1)\}S_u^n\theta^n)$$

and

$$Q_n=\frac{n}{4\pi^2}\begin{bmatrix}1&-1&0&\cdots&\cdots\\-1&2&1&0&\cdots\\\vdots&\vdots&\vdots&\vdots&\vdots\\0&\cdots&\cdots&-1&2\end{bmatrix}+\frac{1}{n}I_n.$$

The reader may recognize the expression for Q_n as encoding the computation of the norm with the first term representing the derivative component. We are led to the following first order conditions:

$$(Q_n+\eta\,\text{diag}\{S_u^n\})\hat{\Delta}=\eta\hat{b}. \tag{72}$$

For each fixed η we can solve for $\hat{\boldsymbol{\Delta}}(\eta)$ using Cholesky factorization noting that the matrices are sparse. This computation turns out to be $\mathcal{O}(n\log(n))$. We now need to determine that value of η such that

$$\phi(\eta)=\|\hat{y}-\hat{u}\hat{G}\theta^n-\hat{u}\hat{\boldsymbol{\Delta}}(\eta)\|_2^2=\|w\|_2^2. \tag{73}$$

This turns out to be particularly simple as it can be shown that $\phi(\eta)$ is a monotonically decreasing function of η. The solution to the problem is shown in the bottom of in Fig. 7. The first subplot shows the noisy estimate obtained by setting

$$\hat{\boldsymbol{\Delta}}(k)=\frac{\hat{y}(k)-\theta^nG(k)u(k)}{u(k)}$$

and predictably has a norm of 200, while the second plot shows the denoised by the optimization problem above. The estimated unmodeled error is 0.95 which differs from the true value by 5%. This is surprisingly close to the true value of 1 especially because we only have one filter-bank.

c) Obtaining the model-parameterization: This is a hard problem and is outside the scope of this paper. However, since we can estimate the unmodeled error, it can form as a basis for updating the model-structure so that it minimizes the estimated unmodeled error. Thus for the examples under consideration it turns out that if we change the model structure to say, $\theta/1-0.7\lambda$ the "optimal" estimate using MUD results in a value of 1.64 and the unmodeled-error is 2.02. As a reality check, we also compute $\theta(T_1)$ with this model-structure and find it to be close to parametric estimate of 1.64. In this way iterating over such model-structures it is possible to show that the optimal model structure is indeed $\theta/(1-0.8\lambda)$. Although, computationally cumbersome, such a procedure can be fundamentally used to find a good candidate model-structure in the sense of minimal unmodeled dynamics.

This shows how the proposed formulation provides a tradeoff between parametric and nonparametric error estimates using measured data and the noise model. The computations are based on convex analysis.

X. Conclusion

In this paper, a new principle for system identification was introduced. In contrast to MPE and set-membership techniques, which prescribe picking a model from a model set that best fits or is consistent with the data respectively, the principle presented recommends picking a model that minimizes the unmodeled error. The principle formulated is meaningful when one has a clear idea about the original system as an element belonging to a complex prior, while the chosen model-set has relatively

limited complexity. Such problems arise naturally in many instances such as identification of time-varying systems where the time variation prohibits estimation of high order dynamics and in identification of lumped parameter models for systems governed by PDEs.

The formulation leads to a crisp definition for parametric and nonparametric components, and, in general, helps streamline identification methodology with robust control. The identification problem reduces to robust convergence of the parameters in the parametric space in the presence of residual dynamics and noise. We overcome the difficulties arising from residual dynamics and noise by developing novel two-step algorithms, with the first step annihilating the residual dynamics and the second step amounting to denoising the data. This methodology is successfully applied in a number of settings and for different topologies. The algorithm developed computationally bears similarities with recursive estimation techniques. However, the techniques developed here are distinct in a number of instances such as identification of limited complexity models in closed-loop setting and identification of low order models when systems are described in general hilbert spaces. We also show that the algorithms have polynomial sample complexity in the number of parameters that describe the model-set for a large number of instances. Unlike the analysis results in MPE where the convergence is pointwise and asymptotic, the results developed here are based on finite-time sample path analysis and hold equally well in set-valued as well as stochastic settings.

We forsee a rich set of extensions of these results and a significant set of open problems that need to be resolved. We point to a few of them here. First of all, it is unclear how the techniques can be generalized to rational model structures. One possible approach is to formulate the problem in a behavioral framework, however, this is currently laden with many technical difficulties. Another direction is in exploring the application of these ideas in estimation problems. Traditionally, Kalman filtering and estimation literature has assumed that the model has no uncertainty, and one could fruitfully employ some of the techniques developed in the paper to address estimation problems in the situations where there is uncertainty in the dynamics. Finally, we perceive that these techniques can be immediately generalized to slowly time-varying systems with little difficulty.

APPENDIX

Proof of Theorem 1: We discuss the case of Bernoulli process below.

Step 1) For a fixed $k \leq m$, let

$$\tilde{X}(\exp(i\alpha)) = \left(\sum_{t=0}^{n} x(t)\exp(i\alpha t^k) \right. \tag{74}$$

Next, notice that we can realize the above expression as a limit of an analytic function inside the unit disc. In fact, it is a polynomial of order n^k:

$$\tilde{X}(\lambda) = \sum_{t=0}^{n} x(t)\lambda^{t^k}, \quad \lambda = \exp(i\alpha). \tag{75}$$

Following on the lines of Lemma 3 and [5], we can show that, for any finite subset Ω of $\{\lambda| \ |\lambda| = 1\}$

$$\mathcal{P}\left\{\frac{1}{N}\max_{\lambda\in\Omega}\left|\sum_{t=0}^{n} x(t)\lambda^{t^k}\right| \geq \alpha\right\} \leq \text{Card}(\Omega)\exp(-nf(\alpha)) \tag{76}$$

where $f(\alpha)$ is as in Lemma 3.

Step 2) By Bernstein's inequality, we know that

$$\sup_{|\lambda|\leq 1}\left|\frac{d}{d\lambda}\sum_{t=0}^{n} x(t)\lambda^{t^k}\right| \leq n^k \sup_{|\lambda|\leq 1}\left|\sum_{t=0}^{n} x(t)\lambda^{t^k}\right|. \tag{77}$$

Suppose $\tilde{X}(\lambda)$ achieves its maximum at λ_0 on the boundary of the unit disc. By the mean-value theorem, we have

$$|\tilde{X}(\lambda_0)| - |\tilde{X}(\lambda)| \leq n^k|\lambda_0 - \lambda| \ |\tilde{X}(\lambda_0)|, \quad \forall\lambda \in \Omega. \tag{78}$$

This immediately implies that

$$|\tilde{X}(\lambda_0)| \leq \frac{|\tilde{X}(\lambda)|}{(1 - |\lambda_0 - \lambda|n^k).} \tag{79}$$

Thus a uniform grid of $2n^k$ on the unit-disc will guaranty that $|\lambda_0 - \lambda| \leq 1/(2n^k)$ for some $\lambda \in \Omega$. Therefore as a result we will have that:

$$\mathcal{P}\left\{\frac{1}{n}\sup_{|\lambda|\leq 1}\left|\sum_{t=0}^{n} x(t)\lambda^{t^k}\right| \geq \alpha\right\} \leq 2n^k\exp(-nf(\alpha/2)) \tag{80}$$

Since, $f(\alpha) = \mathcal{O}(1/\alpha^2)$ the result is easily verified.

Step 3) Observe that $\sum_{t=0}^{n} x(t)\exp(iq(t))$ can be expressed as a multivariate polynomial in m variables of order n^m, i.e.,

$$\begin{aligned}\sum_{t=0}^{n} x(t)\exp(iq(t)) &= \sum_{t=0}^{n} x(t)\prod_{j=1}^{n}\exp(i\alpha_j t^j) \\ &= \sum_{t=0}^{n} x(t)\prod_{j=1}^{n}\lambda_j^{t^j}, \qquad \lambda_j = \exp(i\alpha_j).\end{aligned} \tag{81}$$

Now, pretending that all variables except for λ_j are fixed we obtain a result similar to (80). Extension to the multivariate case is straightforward and will lead us to (7).

Proof of Lemma 2: We first simplify the expression for $r_u^n(\tau)$.

$$\begin{aligned}|r_u^n(\tau)| &= \left|\frac{1}{n}\sum_{t=0}^{n-\tau}\exp(i\alpha t^2)\exp(-i\alpha(t+\tau)^2)\right| \\ &= \left|\exp(-i\alpha\tau^2)\frac{1}{n}\sum_{t=0}^{n-\tau}\exp(-i2\alpha\tau t)\right| \\ &= \left|\frac{1 - \exp(-i2\alpha\tau(n-\tau+1))}{1-\exp(-i2\alpha\tau)}\right| \\ &= \frac{1}{n}\left|\frac{\sin(\alpha(n-\tau+1)\tau)}{\sin(\alpha\tau)}\right| \\ &\cong \frac{1}{n}\left|\frac{\sin(\alpha(n-\tau)\tau)}{\sin(\alpha\tau)}\right|.\end{aligned} \tag{82}$$

Now, if we let $\tau = n/2$ and $\beta_n = \alpha n/2$, we obtain

$$\text{RHS} = \frac{1}{n}\left|\frac{\sin(\beta_n n/2)}{\sin(\beta_n)}\right|. \tag{83}$$

Suppose α is a rational multiple of π. The expression above will be equal to one for a large enough $n \in Z^+$. For αs that are irrational multiples of π, the argument is a little more involved. We define the type of an irrational number as follows. Let the distance between a real number u and the nearest integer be denoted by $\prec \alpha \succ$, i.e.,

$$\prec u \succ = \min_{p \in Z} |p - u| = \min(u(\mathrm{mod}(1)), 1 - u(\mathrm{mod}(1))) \tag{84}$$

then, the type $\eta(u)$ of an irrational number u is defined as

$$\eta(u) = \sup \left\{ h | \liminf_{n \to \infty} n^h \prec nu \succ = 0 \right\}.$$

It is well known (see [6]) that the type of almost all irrational numbers, except on a set of Lebesgue-measure zero, is equal to one. Therefore, there is a sequence $\{n_k\}$ such that

$$n_k \prec \beta_{n_k} \succ \longrightarrow 0.$$

Now, for such a sequence, we immediately see that the RHS is close to 1/2 (using the property that $\sin(kx)/\sin(x)$ grows as k for x close to zero). ■

Proof of Theorem 2:

Step 1) We first simplify the autocorrelation coefficient $r_u^n(\tau)$. Pick any irrational $\alpha \in [0, 2\pi]$ and we have the following lemma.

Lemma 6:

$$|r_u^n(\tau)|^2 = \frac{1}{n^2} \left(n - \tau + \sum_{k=1}^{n-\tau} \exp(i\alpha k\tau(n+k)) \cdot \left(\frac{\sin(\alpha k\tau(n-\tau+1))}{\sin(\alpha k\tau)} \right) \right). \tag{85}$$

It now follows that:

$$\begin{aligned} |r_u^n(\tau)|^2 &= \frac{1}{n^2} \left(n - \tau + \sum_{k=1}^{n-\tau} \exp(i\alpha k\tau(n+k)) \cdot \frac{\sin(\alpha k\tau(n-\tau+1))}{\sin(\alpha k\tau)} \right) \\ &\leq \frac{1}{n} + \frac{1}{n^2} \sum_{k=1}^{n-\tau} \left| \frac{\sin(\alpha k\tau(n-\tau+1))}{\sin(\alpha k\tau)} \right| \\ &\leq \frac{1}{n} + \frac{1}{n^2} \sum_{j=1}^{n-\tau} |\cot(\alpha\tau j)| \end{aligned} \tag{86}$$

where $\cot(\cdot) = \cos(\cdot)/\sin(\cdot)$. It is, therefore, sufficient to prove that the second term goes to zero uniformly w.r.t. τ, i.e.,

$$\max_{0<\tau\leq n} \sqrt{\frac{1}{n^2} \sum_{j=1}^{n-\tau} |\cot(\alpha\tau j)|} \leq \epsilon \Longrightarrow \max_{0<\tau\leq n} |r_u^n(\tau)| \leq \epsilon. \tag{87}$$

The maximization in the above problem can be gotten rid of in the following straightforward way:

$$\max_{0<\tau\leq n} \frac{1}{n^2} \sum_{j=1}^{n-\tau} |\cot(\alpha\tau j)| \leq \frac{1}{n} \sum_{j=1}^{n} |\cot(\alpha\tau j)|. \tag{88}$$

Step 2) The sine function can be bounded from below by its argument, i.e.,

$$\begin{aligned} &\sin(t) \geq c_1 t, \quad \forall t \in [0, \pi/2] \quad \text{and} \\ &\sin(t) \geq c_2(\pi - t), \quad \forall t \in [\pi/2, \pi], \qquad c_1, c_2 \in \mathbb{R}. \end{aligned} \tag{89}$$

Now, for $t > 2\pi$, we know that

$$|\sin(t)| = |\sin(t \bmod(\pi))| = |\sin(\pi - t \bmod(\pi)|. \tag{90}$$

Now, from (89), it follows that:

$$|\cot(\alpha j)| \leq \left| \frac{1}{\sin(\alpha j)} \right| \leq \frac{C}{\prec j\beta \succ}, \qquad \beta = \alpha/\pi \tag{91}$$

where the notation is as in (84). Therefore, we need to only prove that

$$\frac{1}{n} \sum_{j=1}^{n} \frac{1}{\prec j\beta \succ} \longrightarrow 0. \tag{92}$$

Step 3) We now employ the Hardy–Littlewood theorem for this purpose, which is restated here in our words for the sake of completion (see [13] and [6]).

Proposition 7: For almost all irrational numbers, $\alpha \in [0, 2\pi]$ except on a set of Lebesgue-measure zero:

$$\sum_{k=1}^{\infty} \frac{1}{k^2 \prec k\alpha \succ} < \infty. \tag{93}$$

■

We now state Kroneckers lemma, which we employ to conclude the proof.

Lemma 7: Let a_k, b_k be sequences such that a_k is positive and decreasing to zero. Then, $\sum_{k=0}^{\infty} a_k b_k < \infty$ implies

$$\lim_{n \to \infty} a_n \sum_{k=0}^{n} b_k = 0 \tag{94}$$

(see [4] for a proof of Kronecker's lemma). By applying the above lemma, we are done. In our case, $1/k$ is precisely the positive decreasing sequence. The hypothesis of the lemma is satisfied by the Hardy–Littlewood theorem. Hence, we get

$$\sum_{k=1}^{\infty} \frac{1}{k \prec k\alpha \succ} < \infty \Longrightarrow \lim \frac{1}{n^2} \sum_{k=1}^{n^2} \frac{1}{\prec k\alpha \succ} \to 0. \tag{95}$$

In conclusion, we have, for almost any irrational number $\alpha \in [0, 2\pi]$ except on a set of Lebesgue-measure zero

$$\max_{0<\tau\leq n} |r_u^n(\tau)| \overset{n \to \infty}{\longrightarrow} 0 \tag{96}$$

and, therefore, the higher-order chirp is a robust input. The proof of the rate is outside the scope of the paper, and is proved in [32] and [37].

ACKNOWLEDGMENT

The authors are grateful to S. Mitter, F. Paganini, and G. Verghese for their valuable discussions and comments.

REFERENCES

[1] *Linear Matrix Inequalities in System and Control Theory*, vol. 15, 1994.
[2] P. E. Caines, *Linear Stochastic Systems*. New York: Wiley, 1988.
[3] R. M. Chi and H. T. Shu, "Longitudinal vibration of a hoist rope coupled with the vertical vibration of an elevator car," *J. Sound Vibrat.*, vol. 18, July 1991.

[4] K. L. Chung, *A Course in Probability Theory*. New York: Academic, 1974.
[5] M. A. Dahleh, T. V. Theodosopoulos, and J. N. Tsitsiklis, "The sample complexity of worst case identification of linear systems," *Syst. Control Lett.*, vol. 20, Mar. 1993.
[6] E. Engel, *A Road to Randomness in Physical Sciences*. New York: Springer-Verlag, 1992.
[7] E. Fogel and Y. F. Huang, "On the value of information in system identification-bounded noise case," *Automatica*, vol. 18, pp. 229–238, Mar. 1982.
[8] *IEEE Trans. Automat. Contr.*, vol. 37, July 1992.
[9] M. Gevers, "Toward a joint design of identification and control," in *Essays on Control: Perspectives in Theory and Its Applications*, H. L. Trentlman and J. C. Willems, Eds. Cambridge, MA: Birkhaüser, 1993.
[10] L. Giarre, B. Z. Kacewicz, and M. Milanese, "Model quality evaluation in set-membership identification," *Automatica*, vol. 33, June 1997.
[11] L. Giarre, M. Milanese, and M. Taragna, "Model-quality evaluation in $\mathcal{H}_\infty$ identification," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 691–698, May 1997.
[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1990.
[13] G. H. Hardy and J. E. Littlewood, "Some problems of diophantine approximation: The lattice points of a right angled triangle," *Abh. Math. Sem. Hamburg.*, vol. 1, pp. 212–249, 1922.
[14] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control-oriented system identification: A worst-case/deterministic approach in $\mathcal{H}_\infty$," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1163–1176, Oct. 1991.
[15] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 770–783, 1978.
[16] ——, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[17] D. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[18] P. M. Makila, "Robust control-oriented identification," in *Proc IFAC Symp System Identification*, Kitakyushu, Fukuoka, Japan, 1997.
[19] P. M. Makila and J. R. Partington, "On robustness in system identification," Automatica, vol. 35, no. 5, to be published.
[20] P. M. Makila, J. R. Partington, and T. K. Gustafsson, "Worst-case control-relevant identification," *Automatica*, vol. 31, pp. 1799–1819, Dec. 1995.
[21] M. Milanese and A. Vicino, "Information-based complexity and nonparametric worst-case identification," *J. Complexity*, vol. 9, pp. 427–446, Dec. 1993.
[22] K. G. Murty, *Linear Programming*. New York: Wiley, 1993.
[23] F. Paganini, "White noise rejection in a deterministic setting," in *Proc. IEEE Conf. Decision Control*, vol. 39, New York, May 1993, pp. 3658–3663.
[24] J. R. Partington and P. M. Makila, "Analysis of linear methods for robust identification in ℓ_1," *Automatica*, vol. 31, pp. 755–758, 1995.
[25] A. Pinkus, *N-Widths in Approximation Theory*. New York: Springer-Verlag, 1985.
[26] P. Poolla and A. Tikku, "On the time complexity of worst-case system identification," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 944–950, May 1994.
[27] F. C. Schweppe, *Uncertain Dynamical Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[28] T. Soderstrom and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[29] A. Tikku and L. Ljung, "Worst-case identification in ℓ_1 for fir linear systems," in *Proc. 34th IEEE Conf. Decision Control*, vol. 3, New York, 1995, pp. 2998–3003.
[30] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, *Information Based Complexity*. New York: Academic, 1998.
[31] D. N. C. Tse, M. A. Dahleh, and J. N. Tsitsiklis, "Optimal identification under bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1176–1190, Aug. 1993.
[32] S. R. Venkatesh, "System Identification for Complex Systems," Ph.D. dissertation, Massachusetts Institute of Technology, 1997.
[33] S. R. Venkatesh and Y. M. Cho, "Identification and control of high rise elevators," in *Proc. Amer. Control Conf.*, Philadelphia, PA, 1998.
[34] S. R. Venkatesh and M. A. Dahleh, "Classical identification in a deterministic setting," in *Proc. 35th IEEE Conf. Decision Control*, vol. 3, New York, 1995, pp. 2921–2926.
[35] ——, "Identification in the presence of unmodeled dynamics and noise," *IEEE Trans. Automat. Contr.*, vol. 42, Dec. 1997.
[36] ——, "System identification for complex-systems: Problem formulation and results," in *Proc. 36th IEEE Conf. Decision Control*, vol. 3, New York, 1997, pp. 2441–2446.
[37] ——, "Inputs with uniformly decaying auto-correlation coefficients," in *Proc. IFAC SYSID Symp.*, Santa Barbara, CA, June 2000.
[38] S. R. Venkatesh and A. M. Finn, ""A robust and reliable acoustic echo and noise cancellation system for cabin communications"," U.S. Pat. 09/692,531, 1999.
[39] S. R. Venkatesh, A. Megretski, and M. A. Dahleh, "On robust control-synthesis and analysis on a hilbert space," *Syst. Control Lett.*, vol. 39, Jan. 2000.
[40] M. Vidyasagar, *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. New York: Springer-Verlag, 1997.
[41] G. Zames, "On the metric complexity of causal linear systems: Estimates of ϵ-entropy and ϵ-dimension for continuous time," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 222–230, Apr. 1979.

Saligrama R. Venkatesh received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1997.

He worked as a Research Engineer in the Signal Processing Group at the United Technologies Research Center, Hartford, CT. Currently, he is a Visiting Scientist in the Department of Electrical Engineering at the Massachusetts Institute of Technology. His interests are in the areas of reliable learning and control, robust signal processing, and large and complex systems. He holds several patents in the areas of acoustic echo-cancellation and voice enhancement.

Dr. Venkatesh was the recipient of the Outstanding Research Award at the United Technologies Research Center in 1998.

Munther A. Dahleh was born in 1962. He received the B.S. degree from Texas A & M university, College Station, and the Ph.D. degree from Rice University, Houston, TX, both in electrical engineering, in 1983 and 1987, respectively.

He was been a Visiting Professor at the Department of Electrical Engineering, the California Institute of Technology, Pasadena, CA, during spring 1993. He has held consulting positions with several companies in the United States and abroad, and he is a Co-Founder of Crescent Technologies. Since 1987, he has been with the Department of Electrical Engineering and Computer Science, the Massachusetts Institute of Technology, Cambridge, MA, where he is currently a Full Professor. He is the coauthor (with Ignacio Diaz-Bobillo) of the book *Control of Uncertain Systems: A Linear Programming Approach* (Englewood Cliffs, NJ: Prentice-Hall, 1995), and the coauthor (with Nicola Elia) of the book *Computational Methods for Controller Design* (New York: Springer-Verlag, 1998).

Dr. Dahleh has been the recipient of the Ralph Budd award in 1987 for the best thesis at Rice University, the George Axelby outstanding paper award (paper coauthored with J.B. Pearson) in 1987, an National Science Foundation presidential young investigator award in 1991, the Finmeccanica career development chair in 1992, the Donald P. Eckman award from the American Control Council in 1993, and the Graduate Students Council teaching award in 1995. He was a plenary speaker at the 1994 American Control Conference. He served as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL and *Systems and Control Letters*.