

A Convex Analytic Approach to System Identification

Venkatesh Saligrama

Department of Electrical and Computer Engineering
Boston University, Boston MA 02215

Abstract—This paper introduces a new concept for system identification in order to account for random and non-random(deterministic/set-membership) uncertainties. While, random/stochastic models are natural for modeling measurement errors, non-random uncertainties are well-suited for modeling parametric and non-parametric components. The new concept introduced is distinct from earlier concepts in many respects. First, inspired by the concept of uniform convergence of empirical means developed in machine learning theory, we seek a stronger notion of convergence in that the objective is to obtain probabilistic uniform convergence of model estimates to the minimum possible radius of uncertainty. Second, the formulation lends itself to convex analysis leading to description of optimal algorithms, which turn out to be well-known instrument-variable methods for many of the problems. Third, we characterize conditions on inputs in terms of second-order sample path properties required to achieve the minimum radius of uncertainty. Finally, we present fundamental bounds and optimal algorithms for system identification for a wide variety of standard as well as non-standard problems that include special structures such as unmodeled dynamics, positive real conditions, bounded sets and linear fractional maps.

Keywords: complex systems, convex analysis, input design, linear algorithms, robust identification

I. INTRODUCTION

System identification deals with mathematical modeling of an unknown system in a model class from noisy data. The model class is non-random, i.e., specified without a probability distribution, while noise, which typically refers to measurement error is modeled as the realization of stochastic process. The model class is deterministic and can be specified in terms of a finite-dimensional parametrization or more generally as a non-parametric class of systems. This paper introduces a new concept for system identification, inspired by the idea of uniform convergence of empirical means (UCEM) from machine learning theory [24], [32], in order to account for both random and non-random(deterministic/set-membership) uncertainties. The main objective of our formulation is to minimize the probability of the worst-case estimation errors, i.e., *we seek probabilistic uniform convergence of model estimates*.

The question of dealing with both random and non-random aspects in an identification problem is arguably as old as system identification/estimation theory. Indeed, the well known

maximum-likelihood estimate picks the (deterministic) model that maximizes the likelihood of the observed data. In the minimum-prediction error(MPE) rule [13], which can be thought of as a generalization to the maximum likelihood, one chooses models that minimize the prediction error (or in general a loss function). The MPE scheme is quite general and can be used in a wide variety of contexts including situations where the real system does not belong to the chosen model class [3], [14], [11], [33]. Nevertheless, as pointed out in [5] it is in general difficult to assess the model quality of the estimates obtained through MPE methods with finite data. In general there is a significant discrepancy—when there is unmodeled error—between the estimate obtained by minimizing an empirical cost over finite data and the asymptotic theoretical estimate.

This latter issue of dealing with unmodeled dynamics over finite data together with the need for control-oriented models (i.e., with guarantees on uncertainty), has motivated a number of researchers to propose numerous non-mixed formulations. In [9] a stochastic embedding of unmodeled dynamics is described. Milanese [17], [18] initiated research on a purely worst-case paradigm with stochastic noise absorbed within a worst-case framework as uniformly bounded noise(UBN). This research led to rapid development of worst-case algorithms for identification [10], [16], [23], its time complexity [20], [6] and evaluation [8], [7] for a wide variety of problems. Nevertheless, the UBN model can be conservative particularly in accounting for stochastic noise. This has motivated various researchers [30], [27] to consider additional constraints on sample-paths of worst-case noise processes. There has also been related research on explicitly modeling feasible sets through ellipsoidal as well as polytopic constraints [34], [31]. Although, such constraints do lead to desirable behavior in many cases, it is in general difficult to explicitly model both stochastic noise and unmodeled dynamics within a purely worst-case setting. Consequently, while deficiencies in the MPE setup require stronger convergence notions, the purely worst-case approach can either be conservative or require extensive modeling in complex situations.

Motivated by these concerns, a mixed approach has been discussed in [30], [26], [27], where the approach has been to attempt a separation between unmodeled dynamics and noise and seek models that minimize the distance between the model class and the underlying system. In parallel mo-

This research was supported by the ONR Young Investigator Program and Presidential Early Career Award (PECASE) N00014-02-100362, NSF CAREER award ECS 0449194, and NSF Grant CCF 0430983

tivated by persistent identification, Wang [36] has proposed a new concept for mixed identification, wherein the idea is to pick model estimates so that the worst-case probability of “mismatch” error being larger than some threshold is minimized. Furthermore, a new notion of sample complexity is developed wherein the idea is to find the smallest time by which a confidence level for the probabilistic objective can be reached. These ideas resemble notions of minimax risk functions employed in the statistical and estimation literature [12]. There the idea is to choose parameters that minimize the worst-case (over all parameters) the expected value of the estimation error. Nevertheless, this problem (in terms of finding optimal algorithms, optimal inputs, minimum sample complexity etc.) is in general intractable in both the estimation and identification contexts. Wang [36] addresses this issue by deriving upper and lower bounds for a number of different cases. The upper bounds are usually obtained by using least-squares algorithms for periodic inputs while the lower bounds are obtained by considering the noiseless case. Although, these bounds can sometimes be suitable, they can be overly conservative for situations where Least Squares (LS) is not a convergent algorithm. This is particularly the case when there is a separation between model, unmodeled dynamics and noise as was shown in [27]. In summary, although there are a number of formulations that have been proposed for dealing with mixed scenarios, it is generally difficult to characterize optimal algorithms, inputs and fundamental bounds.

To address these issues we propose a new formulation for mixed components in Section II. Preliminary work along these lines has appeared in our earlier publications [25], where consistent identification of FIR models was described. This paper generalizes that formulation and deals with a wide variety of new problems. This formulation is inspired by notions of uniform convergence of empirical means and approximate learning in machine learning theory. Our objective is to seek models/estimates that minimize the probability of the worst-case errors, i.e., we seek uniform convergence in probability of the estimation error. Although, this notion is stronger than the earlier proposed formulations, we show in Section III, that it is also tractable through convex analysis particularly for problems defined by linear observations, convex subsets of systems and additive random noise and under the ℓ_∞ norm. This can be extended to other norms by approximating in a finite family of linear cost functions. Specifically, we show that optimal algorithms can be associated with hyperplane separation of convex sets, which can further be related to instrument-variable techniques. In Section IV the paper describes new insights for parameter estimation in stochastic noise. In particular, it is well known that a persistency of excitation condition is required to ensure consistency. By means of our mixed formulation we show that this condition is also necessary for achieving uniform consistency. This section also develops solution techniques for problems with convex parametric constraints. Section V then develops algorithms for optimal identification in the presence of unmodeled dynamics

and noise. The analysis is then extended to problems to cover situations where the desired solution can be expressed as a non-linear function of a linear operator. In particular, this strategy is employed to show that uniform convergence of parametric error to zero can be established in the presence of noise and unmodeled dynamics for stable systems in an ℓ_1 space. In Section VI we show how this approach also leads to solutions to problems described by LFT structures. Another aspect of the paper is in characterizing inputs required for system identification. We derive conditions on input sequences in terms of their second order sample path properties.

a) Notation: A^* denotes the complex-conjugate transpose of the matrix A . Z^+ is the set of positive integers. ℓ denotes real valued sequence on positive integers. ℓ_p , $p \geq 1$ denotes the space of sequences on Z^+ bounded in the ℓ_p norm (in [15] these concepts are defined on the space of natural numbers but we only consider sequences over positive integers and denote them as ℓ_p with an abuse in notation). For a signal, $x \in \ell_p$, P_m denotes the truncation operator: $P_m(x) = (x(0), \dots, x(m-1), 0, \dots)$, X_n is the column vector $(x(0), x(1), \dots, x(n-1))^*$. $\|x\|_p$ denotes the ℓ_p norm of the discrete time signal $x(\cdot)$. $\|A\|_{p,q} = \sup_{x \in \ell_p} \|Ax\|_q / \|x\|_p$ denotes the induced norm of operator or matrix.

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

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

The n -point cross-correlation between two signals

$$r_{xy}^n(\tau) = \sum_{i=0}^{n-\tau} x(\tau+i)y(i)$$

$N(m, \sigma)$ denotes the gaussian distribution with mean m and standard deviation σ . $\text{Prob}\{A\}$ denotes the probability of an event A and $E\{X\}$ denotes the expected value of the random variable X . Finally z^{-1} is the z -transform variable and corresponds to the usual shift operation. LTI refers to linear time invariant systems and BIBO refers to bounded-input-bounded-output systems, i.e., the space of systems defined by the disc algebra of analytic functions bounded in the closed unit disc. \mathbb{RH}_2 refers to stable real-rational LTI systems with bounded squared norm. The impulse response of an LTI system, H , is represented by $\{h(k)\}_{k \in Z^+}$. The output response, $y(t)$, to an input $u(t)$ is denoted by $y(t) = Hu(t) = (h \star u)(t)$, where \star denotes the convolution operation. For a real-rational LTI system, H , we often associate a minimal realization in state space:

$$H \equiv \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array} \right] \quad (1)$$

II. PROBLEM FORMULATION

In this section we formulate a mixed stochastic-deterministic problem. Mathematically, we are given a sequence of real-valued observations, $y(k)$, which are assumed to be generated according to an underlying model. The model

is uncertain in that it has both random and non-random aspects that provide information on how the sequence of observations are generated. More precisely, let

$$y(k) = F^k(\theta, \Delta, w), \quad k = 0, 1, \dots, n-1 \quad (2)$$

where, F^k is a time dependent real-valued regressor; $w(\cdot)$ is a random noise process; θ, Δ are non-random components with the tuple $(\theta, \Delta) \in \mathcal{S}$, where \mathcal{S} is a subset of a separable normed linear space \mathcal{H} . The finite dimensional component, θ , is to be estimated by means of an algorithm, $\hat{\theta}^n(y)$ from a sequence of n observations. We denote by $\hat{\theta}$ (with an abuse of notation) the infinite sequence of algorithms, i.e., $\{\hat{\theta}^n\}$. To clarify the notation, consider an LTI example with a 1-tap FIR with unmodeled dynamics, i.e.

$$\begin{aligned} y(t) &= \theta u(t) + \sum_{k=0}^t \delta(t-k)u(k) + w(t) \\ &\equiv F_u^t(\theta, \Delta, w); \quad (\theta, \Delta) \in \mathcal{S} = \mathbf{R} \times \ell_1; \quad u(\cdot) \in \ell_\infty \end{aligned}$$

where, $\delta(t)$ is the impulse response of the unmodeled component, Δ .

Remark: In general, note that the variables θ and Δ , may not be independent, i.e., the set of values θ and Δ can be jointly specified by the set $\mathcal{S} \subset \mathcal{H}$. This models situations involving optimal approximations, i.e., θ characterizes the best approximation in the model class, while Δ denotes the residual error.

We now introduce the mixed problem. Suppose, $\|\cdot\|$ is a suitable norm on \mathbf{R}^m . We denote by $\alpha(\hat{\theta}^n, \gamma, \mathcal{S})$ the probability that the worst-case estimation error is larger than γ for n observations, i.e.,

$$\alpha(\hat{\theta}^n, \gamma, \mathcal{S}) = \text{Prob} \left\{ \sup_{(\theta, \Delta) \in \mathcal{S}} \|\theta - \hat{\theta}^n(y)\| \geq \gamma \right\}$$

where the probability is taken with respect to the noise distribution. Observe that the set, \mathcal{S} is a separable space and so measurability is preserved under a supremum [1]. Note that the problem is meaningful even when the perturbation, Δ , does not exist. Indeed, this situation is analogous to the concept of uniform convergence of empirical means. For the simplest situation in learning theory, we are given a family of functions, $f(x) \in \mathcal{F}$. N i.i.d. samples of $\{x_i\}$ are drawn from a distribution defined by P and the problem is to show that,

$$\text{Prob} \left\{ \sup_{f \in \mathcal{F}} \left| E_P(f(x)) - \frac{1}{N} \sum_{i=1}^N f(x_i) \right| \geq \gamma \right\} \rightarrow 0$$

Extensions to this problem include consideration of general loss functions and function approximation with lower complexity function classes. The principle difficulty in applying learning theory to system identification is that the solutions there depend on observations being independent and do not deal with dynamical scenarios.

Returning to our context, we define the asymptotic radius of uncertainty, $R(\hat{\theta}, \mathcal{S})$, for the sequence of algorithms $\hat{\theta}$ as:

$$R(\hat{\theta}, \mathcal{S}) = \inf \left\{ \gamma \in \mathbf{R} \mid \limsup_n \alpha(\hat{\theta}^n, \gamma, \mathcal{S}) = 0 \right\}$$

This leads to the notion of consistency:

Definition 1: We say that the sequence of algorithms $\hat{\theta}$ is uniformly consistent if the radius of uncertainty is equal to zero.

The minimum radius of uncertainty is given by:

$$R_0(\mathcal{S}) = \inf_{\hat{\theta}} R(\hat{\theta}, \mathcal{S})$$

Finally, we define sample complexity and optimality of an algorithm:

Definition 2: The sample complexity of the algorithm $\hat{\theta}$ is the smallest number of observations, $L(\epsilon, \xi, \hat{\theta}, \mathcal{S})$, such that the probability that the worst-case estimation error is larger than $\gamma = R_0(\mathcal{S}) + \epsilon$, is at most α , i.e.,

$$L(\epsilon, \xi, \hat{\theta}, \mathcal{S}) = \min \left\{ n \in \mathbf{Z}^+ \mid \alpha(\hat{\theta}^n, \gamma, \mathcal{S}) \leq \xi \right\}$$

An algorithm, $\hat{\theta}_0$ is said to be *optimal* if its sample complexity is smaller than any other algorithm, i.e.,

$$L(\epsilon, \xi, \hat{\theta}_0, \mathcal{S}) \leq L(\epsilon, \xi, \hat{\theta}, \mathcal{S}), \quad \forall \xi > 0; \quad \epsilon > 0$$

These definitions are related to the mixed formulations proposed in [36] with important differences. Wang et. al. [36] consider a weaker version in that their formulation is related (but not exactly¹) to the worst-case probability of error, i.e., their notion is related to the left-hand-side, while our notion is related to the right-hand-side of the following inequality.

$$\sup_{(\theta, \Delta) \in \mathcal{S}} \text{Prob} \left\{ \|\theta - \hat{\theta}^n(y)\| \geq \gamma \right\} \leq \text{Prob} \left\{ \sup_{(\theta, \Delta) \in \mathcal{S}} \|\theta - \hat{\theta}^n(y)\| \geq \gamma \right\}$$

The two problems can be significantly different as the following examples indicate:

b) *Example:* Let X be a binary number taking values in the set $\{-1, 1\}$. Suppose W is a binary random variable taking values in the set $\{-1, 1\}$ each of which is equally likely. We immediately see that $\max_X \text{Prob}\{WX \geq 1/2\} = 1/2$; $\text{Prob}\{\max_X(WX) > 1/2\} = 1$

c) *Example:* Suppose, $Y = \theta + W$ with $\theta \in [-1, 1]$ and W uniformly distributed in $[-1, 1]$. It follows that, for a linear algorithm, $\hat{\theta}(Y) = \alpha Y$,

$$\min_{\alpha} \text{Prob}\{\max_{\theta} |\alpha Y - \theta| > 1/2\} \geq 1/2$$

where we have used the fact that the minimum value is equal to $\min_{\alpha} \text{Prob}\{\max_{\theta} |(1-\alpha)\theta + \alpha W| > 1/2\}$. The lower bound now follows by choosing, $\theta = W$. It turns out through detailed analysis that the optimal solution for the weaker notion is given by $\alpha = 3/4$, and we get,

$$\min_{\alpha} \max_{\theta} \text{Prob}\{|\alpha Y - \theta| \geq 1/2\} \leq 1/3$$

Although, the weaker notion may suffice in many cases, the main problem is that it is difficult to analyze exactly (as seen even for the above example). This issue is well

¹[36] considers simpler structures where the model and unmodeled dynamics are decoupled

known in statistical estimation [12] and recently the authors have analyzed the weaker notion [2], [29] in the context of composite hypothesis testing, and except for simple situations the problem is generally intractable. In [36] this issue is addressed by deriving upper and lower bounds, where the upper bound is computed with a least squares algorithm and periodic inputs. The lower bound is computed by considering the noiseless case. Nevertheless, for a large number of problems especially when there is a decomposition between model and unmodeled dynamics the upper and lower bounds are not tight. Indeed, [27] provides examples wherein neither periodic inputs nor LS algorithms lead to consistency, while there exist specialized inputs and linear algorithms that do result not only in consistency but also polynomial sample-complexity. One advantage of our formulation is that it lends itself to exact analysis through application of convexity theory and by extension design of optimal inputs and algorithms, particularly for problems defined by linear observations, convex subsets of systems and additive random noise for ℓ_∞ norm, as illustrated in the sequel. We also demonstrate extensions of this approach to other norms by approximating in a finite family of linear cost functions.

III. OPTIMAL ALGORITHMS FOR CONVEX PROBLEMS

In this section we state and prove a general theorem, which is applied in subsequent sections. The main result here is that under some technical conditions linear algorithms have optimal sample complexity. The theorem is a generalization of Smolyak's theorem [22] to observations with random noise. Consider a linear space, \mathcal{H} , and a convex and balanced subset, $\mathcal{S} \subset \mathcal{H}$. Suppose, elements of the set, \mathcal{S} are observed through a linear measurement structure with additive random noise, $w(\cdot)$, i.e.,

$$y(k) = \lambda_k(h) + w(k), \quad k = 1, \dots, n, \quad h \in \mathcal{S} \quad (3)$$

where, λ_k , is a linear functional on the space, \mathcal{H} . Thus the observations are linear over \mathcal{H} . The task is to compute a linear function $\Phi(h)$, where Φ maps h to \mathbb{R}^p , i.e., $\Phi(h) = (\phi_1(h), \dots, \phi_p(h))^T$. The linear function $\Phi(h)$ is often referred to as the solution operator.

d) Example: The measurement structure includes ARMA models as well. This is because, if $\psi(k)$ is the sequence of regressors formed for an ARMA structure (see [13] for details) then,

$$y(k) = \psi^T(k)\theta + w(k)$$

forms a linear measurement structure on the variable θ . ■
For the setup described by Equation 3 we have the following theorem.

Theorem 1: Suppose $w(\cdot)$ is i.i.d. Gaussian noise such that $w(k) \sim \mathcal{N}(0, \sigma^2)$. Then linear algorithms are optimal when the norm measure on $\Phi(h)$ is the ℓ_∞ norm.

Remark: In general the result holds for more general noise distributions, which are symmetric and monotone around zero. These properties can be used to argue the equivalence

between the probabilistic problem and an appropriate worst-case problem. In this paper we limit our attention to gaussian processes (the non i.i.d. case is established in the next section) for the sake of simplicity.

Remark: Although we have assumed Gaussian noise the theorem holds for other noise distributions as well. The main property required is that the noise distribution be symmetric and monotonic around zero. For instance, the uniform distribution satisfies these properties as well.

Proof: If the radius of uncertainty is infinite then linear algorithms are no worse than any other algorithm. Therefore, suppose $\gamma_0 = R_0(\mathcal{S}) < \infty$. For simplicity, consider estimation of a single linear functional, $\phi(h)$. Suppose, the sequence of algorithms, $\hat{\phi} = \{\hat{\phi}^n(y)\}$, achieves the sample complexity, $L(\epsilon, \xi, \hat{\phi}, \mathcal{S})$. This implies,

$$\text{Prob} \left\{ \sup_{h \in \mathcal{S}} |\phi(h) - \hat{\phi}^n(y)| \geq \gamma_0 + \epsilon \right\} \leq \xi, \quad \forall n \geq L(\epsilon, \xi, \hat{\phi}, \mathcal{S}) \quad (4)$$

We are left to establish that there is a linear algorithm that achieves the same performance. The proof of optimality of the linear algorithm would follow because $\hat{\phi}^n(y)$ is arbitrary and can be replaced by an optimal algorithm.

Now, Equation 4 is equivalent to existence of a set $\mathcal{A} \subset \mathbb{R}^n$ with $\text{Prob}\{\mathcal{A}\} \leq \xi$ such that,

$$\sup_{w \in \mathbb{R}^n - \mathcal{A}} \sup_{h \in \mathcal{S}} |\phi(h) - \hat{\phi}^n(y)| \leq \gamma_0 + \epsilon \quad (5)$$

where, w , is now a worst-case noise signal for which the error is uniformly bounded by $\gamma_0 + \epsilon$. Next we establish that the set $\mathbb{R}^n - \mathcal{A}$ can be chosen to be convex without impacting the above inequality. We state this in the form of a proposition below and the details can be found in the appendix,

Proposition 1: If Equation 5 is satisfied then it follows that there exists a convex balanced subset, $\mathcal{W}_0^n \subset \mathbb{R}^n$, with $\text{Prob}\{\mathcal{W}_0^n\} \geq 1 - \xi$ such that

$$\sup_{w \in \mathcal{W}_0^n} \sup_{h \in \mathcal{S}} |\phi(h) - \hat{\phi}^n(y)| \leq \gamma_0 + \epsilon \quad (6)$$

We are now left to prove that a linear algorithm is an alternative optimal estimator as well. For this we follow the proof of Smolyak's theorem [22]. The main idea is illustrated in Figure 1.

Consider the following set of \mathbb{R}^{n+1} :

$$\mathcal{V} = \{(\phi(h), \lambda_1(h) + w(1), \dots, \lambda_n(h) + w(n)) \mid w \in \mathcal{W}_0^n, h \in \mathcal{S}\}$$

From Equation 6, it follows that $(\gamma_0 + \epsilon, 0, \dots, 0)$ is not contained in the interior of the set \mathcal{V} . This implies the existence of a hyperplane such that,

$$c_0(\phi(h) - (\gamma_0 + \epsilon)) + \sum_{j=1}^n c_j y_j \leq 0, \quad \forall h \in \mathcal{S}, \quad w \in \mathcal{W}_0^n$$

Now, the linear functionals, $\{\lambda_1(\cdot) + e_1^T, \dots, \lambda_n(\cdot) + e_n^T\}$, where, e_j , is a column vector of length n with j th component equal to one and all other components equal to zero, are linearly independent. This in turn implies that $c_0 \neq 0$. Now

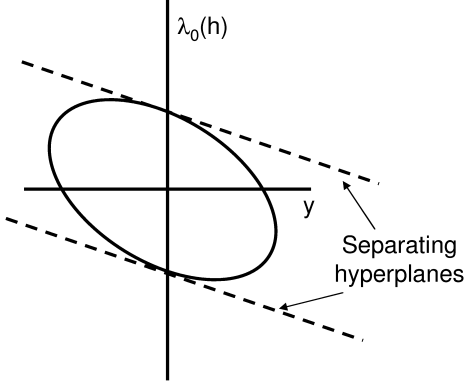


Fig. 1. Illustration of Convex Separation; X and Y axes denote range of measurement and parametric values respectively. With high probability they lie in a bounded convex set.

by using the fact that \mathcal{V} is convex and balanced we see that $(-\gamma - \epsilon, 0, \dots, 0)$ is also another boundary point. Therefore we obtain the complementary inequality, i.e.,

$$c_0(\phi(h) + (\gamma + \epsilon)) + \sum_{j=0}^n c_j y_j \geq 0, \quad \forall h \in \mathcal{S}, \quad w \in \mathcal{W}_0^n$$

Consequently, for every $\epsilon, \xi > 0$, it follows that there are a sequence of coefficients $\{q^n(k) = -c_k/c_0\}_{k=0}^n$ not all zero such that,

$$\sup_{w \in \mathcal{W}_0^n} \sup_{h \in \mathcal{S}} \left| \sum_{k=0}^n q^n(k) y(k) - \phi(h) \right| \leq \gamma_0 + \epsilon, \quad \forall n \geq N(\epsilon, \xi)$$

thus, establishing result for a single linear functional. To generalize it to the linear function, $\Phi(h)$, we note that,

$$\begin{aligned} \|\Phi(h) - \hat{\Phi}(h)\|_\infty &= \max_j \left| \phi_j(h) - \hat{\phi}_j^n(y) \right| \\ &\geq \left| \sum_{k=0}^n q_j^n(k, \phi_l) y(k) - \phi_j(h) \right| \end{aligned}$$

where, $q_j^n(k, \phi_l)$ are the optimal linear weights corresponding to the linear functionals $\phi_l(h)$. ■

A. Extensions

We can generalize the above theorem to squared error norms as well. Although, it is possible to bound the squared norm with the infinity norm this can be conservative (for example, for a vector $x \in \mathbb{R}^p$ we have $\|x\|_\infty \leq \sqrt{p}\|x\|_2$). Instead, our idea is to express squared norm value as the maximum over all unit magnitude linear functionals—the existence of which is guaranteed through duality—and then find estimates for each of these linear functionals. Specifically, let $x \in \mathbb{R}^p$ endowed with the squared norm metric and let $v(\cdot)$ be a linear functional on \mathbb{R}^p . Therefore, $v(\cdot)$ can be associated with a vector $(v_1, \dots, v_p) \in \mathbb{R}^p$. We then have,

$$\max_{\|v\|_2 \leq 1} v(x) = \|x\|_2$$

Now to draw a direct correspondence with the theorem we proceed as follows. We consider the problem of estimating the linear function $\Phi(h) = (\phi_1(h), \phi_2(h), \dots, \phi_p(h))$, with an algorithm $\hat{\Phi}^n(\cdot)$ with the squared norm as the error measure. For ease of exposition we let $x_j = \phi_j(h)$. We then select linear functionals,

$$v(x) = \sum_{k=1}^p v_k x_k$$

Corresponding to each choice of $v(\cdot) \in \mathbb{R}^p$ we have an optimal linear algorithm, $q_v^n(k)$ with the minimum radius of uncertainty, $\gamma_0 + \epsilon$ (with respect to the squared norm), i.e.,

$$\left| \sum_k q_v^n(k) y(k) - \sum_{k=1}^p v_k x_k \right| \leq \gamma_0 + \epsilon \quad (7)$$

Clearly, $v(\cdot)$ consists of an infinite family and needs discretization. Let v^k be such a discretization with k belonging to some finite index set. We then obtain a finite set of linear inequalities. A feasible point, $(x_1, \dots, x_p) = \Phi(h) = (\phi_1(h), \dots, \phi_p(h))$ in this finite set forms a good estimate for the squared norm. Specifically, let $v^k = (v_1^k, \dots, v_p^k) \in \mathbb{R}^p$, $k \in \mathcal{I}$, $\|v^k\|_2 = 1$. Suppose, the minimum angle of separation for the collection of vectors is smaller than $\beta < \pi/4$. Consider a feasible point, x belonging to the set defined by Equation 7 with $v \in \{v^k\}_{k \in \mathcal{I}}$ and let $\hat{\Phi}^n(y) = x$ be a non-linear estimator that maps the observations to a feasible point. It then follows that,

Corollary 1: If there exists an algorithm that achieves a radius of uncertainty $\gamma_0 + \epsilon$ for a data length n with probability α then:

$$\|\Phi(h) - \hat{\Phi}^n(y)\|_2 \leq \frac{\gamma_0 + \epsilon}{1 - \beta}$$

Proof: See the appendix.

These theorems are readily generalized to stationary gaussian noise setting with bounded power spectral density.

Corollary 2: Let, $Y_n = \Lambda_n(h) + Q_n W_n$, where Y_n, W_n are column vectors of length n and with $\|Q_n\|_2 \leq 1$ and $\Lambda_n = (\lambda_1(\cdot), \lambda_2(\cdot), \dots, \lambda_n(\cdot))$ is a sequence of linear functionals. Then a linear algorithm achieves the minimum diameter of uncertainty for estimating a linear function $\Phi(h)$ mapping \mathcal{H} to \mathbb{R}^p from data under the ℓ_∞ and squared norm error measures.

Remark: Note that the results do not provide an explicit expression for computing the minimum radius of uncertainty (and indeed this problem is . However, they do provide a basis for determining this uncertainty through linear algorithms.

IV. PARAMETER ESTIMATION REVISITED

In this section we provide a complementary perspective on parameter estimation in random noise. The motivation for revisiting this well studied topic is threefold: (1) First, this topic has not been studied under the new criterion of uniform convergence. In the prediction-error paradigm [13], which unifies many of the results in stochastic identification, the objective is to analyze prediction errors, i.e., $V(\theta) = \frac{1}{n} \sum_{t=1}^n E(y(t) - \hat{y}(t/\theta))^2$ as a function of the predictor.

Our objective differs in two ways: (a) It directly deals with the estimation error, i.e., $\|\theta - \hat{\theta}^n(y)\|$ (b) we seek uniform convergence, i.e., $\text{Prob}\{\sup_{\theta} \|\theta - \hat{\theta}^n(y)\|\}$. (2) It turns out that under this new criterion the well known input requirements become transparent and can be shown to be both necessary as well as sufficient for achieving consistency. We point out that most results that exist in the literature in this connection are sufficient conditions and necessity is not easy to establish in many cases. (3) We can describe tight lower bounds for error for finite data without the need for a consistent estimator, which is required typically for Cramer-Rao bounds. (4) Finally, a new aspect is that optimal algorithms for identification problems with convex parametric constraints, which model parametric dependencies, can also be developed.

To describe our problem we need the notions of persistent input and persistency of excitation. A persistent input is an ℓ_∞ sequence with unbounded energy, i.e.,

Definition 3: An ℓ_∞ bounded sequence, $(x(0), x(1), \dots, x(k), \dots)$, is said to be persistent if

$$\liminf_n \|P_n x\|_2 \rightarrow \infty$$

We point out that this definition is weaker than the conventional notion, where the input signal is required to be quasi-stationary and have finite power, i.e., $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^n x^2(t) = C$. We next describe persistency of excitation, which also is weaker than the conventional notion.

Definition 4: An input $u(\cdot)$ is said to be persistently exciting of order, m , if there exists m instruments (discrete-time signals), $v_1(\cdot), v_2(\cdot), \dots, v_m(\cdot)$ such that

- 1) The input $u(\cdot)$ is persistent.
- 2) The $m \times m$ matrix, Σ_n , with $[\Sigma_n]_{jk} = r_{v_j u}^n(k)$ satisfies, $\Sigma_n \geq \rho I_m$, for all $n \geq N_0$, for some $N_0 \in \mathbb{Z}^+$ and $\rho > 0$.

We first prove the result for FIR parameterizations, i.e.,

$$y(t) = \sum_{k=0}^{p-1} h(k)u(t-k) + w(t)$$

We assume that $w(t)$ is i.i.d. gaussian noise. The following result follows:

Theorem 2: For the setup above the radius of uncertainty is zero if and only if the input $u(\cdot)$ is persistently exciting of order p .

Proof: The sufficiency part of the result can be derived along the lines of [13] and is omitted. The necessity part to the best of author's knowledge is new. Since norms on finite dimensional spaces are all equivalent we employ the ℓ_∞ norm here. We are then given that there exists an algorithm, $\hat{h}_k^n(y)$ such that,

$$\text{Prob} \left\{ \sup_{h \in \mathbf{R}^p} \|h(k) - \hat{h}_k^n(y)\|_\infty \right\} \rightarrow 0$$

To apply Theorem 1 we need to define λ_i s first:

$$\lambda_j(h) = \sum_{k=0}^{p-1} h(k)u(j-k), \quad j = 1, 2, \dots, n$$

With this association we have a linear observation structure and since the objective is to estimate the FIR taps, $\Phi(\cdot)$ in this case is the identity operator. Therefore, the assumptions of Theorem 1 are satisfied. Consequently, based on our hypothesis, there exists a linear algorithm that achieves uniform consistency as well. This implies existence of an instrument q_k^n corresponding to each coefficient $h(k)$ so:

$$\begin{aligned} \left| \sum_{j=0}^n q_k^n(j)w(j) \right| + \sum_{j=0}^n (q_k^n(j)u(j-k) - 1)h(k) & \quad (8) \\ + \sum_{j=0, j \neq k}^{p-1} \sum_{l=0}^n q_k^n(l)u(l-j)h(j) & \xrightarrow{N \rightarrow \infty} 0 \end{aligned}$$

Since, $h \in \mathbf{R}^p$ is arbitrary it follows that,

$$\sum_{l=0}^n q_k^n(l)u(l-k) = 1; \quad \sum_{l=0}^n q_k^n(l)u(l-j) = 0, \quad \forall j \neq k \quad (9)$$

and

$$\text{Prob} \left\{ \left| \sum_{j=0}^n q_k^n(j)w(j) \right| \geq \epsilon \right\} \xrightarrow{N \rightarrow \infty} 0$$

Since, $w(\cdot)$ is an i.i.d. gaussian random process, $\sum_{j=0}^n q_k^n(j)w(j)$ is gaussian with mean zero and variance equal to $\|q_k^n\|_2^2$. Therefore, we need

$$\limsup_n \|q_k^n\|_2 \rightarrow 0$$

Now from Equation 9 we have from Cauchy-Schwartz inequality,

$$1 = \left| \sum_{l=0}^n q_k^n(l)u(l-k) \right| \leq \|q_k^n\|_2 \left(\sum_{l=0}^n u^2(l-k) \right)^{1/2}$$

This implies that $\|(P_n - P_{n-1})u\|_2 \rightarrow \infty$. Therefore, the input must be persistently exciting and from Equation 9 it follows that the persistency of excitation must be of order p as stated. ■

The proof is readily extended to the more general case with linear parameterizations, i.e.,

Corollary 3: Let the input-output equation be given by

$$y(t) = \psi^T(t)\theta + w(t), \quad \theta \in \mathbf{R}^p \quad (10)$$

where, $\psi^T(t) = [y(t-1), y(t-2), \dots, y(t-m_1), u(t), u(t-1), \dots, u(t-m_2)]$ is the usual regression vector formed from previous inputs and outputs, where $m_1 + m_2 = p$. The radius of uncertainty is equal to zero if and only if the input is persistently exciting of order p .

These ideas can also be extended to more general convex parameterizations which account for parametric dependencies. We consider finitely many parametric constraints modeled as:

$$|Z\theta| \leq b \quad (11)$$

where, $Z \in \mathbf{R}^{l \times p}$. Note that the parametric set is balanced. We describe some relevant situations where such convex constraints naturally arise. Our first example is the case of

a positive real condition, which can be modeled by convex constraints:

$$\sum_k \theta_k \cos(\omega k) \geq 0, \quad \omega \in [-\pi, \pi]$$

To utilize these constraints we can discretize the infinite system of inequalities and through translation consider convex and balanced subsets of the parameter space. Another example where these results are applicable is when the set of parameters belong to a linear manifold, i.e., $Z\theta = b$. We have the following corollary for convex parameter constraints:

Corollary 4: Consider the setup of Equations 10, 11. Furthermore, let $\Psi_n = [\psi(0), \psi(1), \dots, \psi(n)]^T$ be the data regression matrix. It follows that for a data length n the radius of uncertainty (in the ℓ_∞ norm) is smaller than δ if and only if there exists matrices $Q \in \mathbb{R}^{p \times n}$ and Γ such that,

$$Q\Psi_n + \Gamma Z = I, \quad |\Gamma b| + \|Q\|_{2,\infty} \leq \delta, \quad \Gamma \geq 0 \quad (12)$$

where, all of the inequalities are to be interpreted pointwise.

Proof: The proof of the result follows from convex duality and in particular a straightforward application of Farkas lemma, which is stated here for the sake of completion [19]:

Lemma 1 (Farkas Lemma): Consider a set of linear functions, $\{\mu_j\}$ on the parameter θ . Then it follows that,

$$\begin{aligned} (\mu_0^T \theta - b_0 > 0 \text{ whenever } \mu_i^T \theta - b_i > 0, \quad i = 1, 2, \dots, m) \\ \Updownarrow \\ \mu_0 = \sum_{k=1}^m \gamma_k \mu_k, \quad \sum_{k=1}^m \gamma_k b_k \leq b_0, \quad \gamma_k \geq 0 \end{aligned}$$

We now return to the proof of Corollary 4. First, note that the problem setup conforms to the convex setup. Indeed, the input-output data is given by a linear measurement structure, the parameter set is convex and balanced. The desired solution, $\Phi(\theta) = \theta$, is linear. Therefore, it follows similar to the arguments used in the preceding theorems that if the radius of uncertainty is smaller than δ , then there must exist a matrix Q and an integer n , for a given $\epsilon > 0, \xi > 0$, such that

$$|(Q\Psi_n - I)\theta + Qw| \leq \delta + \epsilon, \quad \forall |Z\theta| \leq b; \quad w \in \mathcal{W}_0^n$$

where, \mathcal{W}_0^n is the same convex set with probability mass larger than $1 - \xi$ as described in Theorem 1. The proof now follows by a direct application of Farkas lemma. ■

The solution consists of finding a Q that satisfies Equation 12, which can be solved through convex programming. An estimate of θ is then given by Qy , where y is the observation vector of length n .

Remark: Observe that when the constraints are given by $Z\theta = 0$, we should expect that the persistency of excitation condition can be relaxed. This is indeed the case as seen from Equation 12. This is because in this case, in the dual setting, the positive constraint $\Gamma \geq 0$ no longer exists. Therefore, we are only left to satisfy $Q\Psi_n + \Gamma Z = I$ for an arbitrary Γ . It follows that if Z has rank m then Φ_n need only have rank $p - m$ to ensure that the equation can be satisfied. Therefore,

the input needs to be persistently exciting of a smaller order than the parametric dimension.

Our next task is to comment on the issues of optimality and choice of optimal inputs. From Theorem 1 it follows that linear algorithms are optimal for the ℓ_∞ norm. From the results in Corollary 4 we have for a convex parametric set \mathcal{S} that,

$$\begin{aligned} \text{Prob} \left\{ \max_{\theta \in \mathcal{S}} \|\theta - \theta^n(y)\|_\infty \geq \epsilon \right\} &\geq \text{Prob} \{ \|Qw\|_\infty \geq \epsilon - \|\Gamma b\|_\infty \} \\ &= \text{erfc} \left(\frac{\epsilon - \|\Gamma b\|_\infty}{\|Q\|_{2,\infty}} \right) \end{aligned}$$

where, $\text{erfc}(\cdot)$, denotes the gaussian error function [12] and we have assumed that radius of uncertainty is equal to zero (for simplicity of exposition). These results in turn help in loosely characterizing inputs that would lead to optimal identification. This would follow by first minimizing, $\|Q\|_{2,\infty}$ subject to constraints of Equation 12. The optimal cost of this optimization problem would provide an index $J(\Phi_n)$ which is a function of the regressor Φ_n . One can then attempt to minimize the index by choosing different inputs.

A. Uniform Bounded Noise

Although, the UBN setup does not strictly fall within the framework described in Section II the proof technique of Theorem 1 can be used to establish that the radius of uncertainty is bounded away from zero. Let,

$$y(t) = \sum_{k=0}^{p-1} h(k)u(t-k) + w(t), \quad w(t) \in [-\gamma, \gamma], \quad |u(\cdot)| \leq 1 \quad (13)$$

In the context of UBN we are interested in an estimate $\hat{h}^n(y)$ based on input-output data of length n so the worst-case error over all FIRs and admissible noise realizations are minimized:

$$\min_{\hat{h}^n} \mapsto \sup_{\mathcal{W}^n} \sup_{h \in \mathbb{R}^p} \|h - \hat{h}^n\|_1$$

where, $\mathcal{W}^n = [-\gamma, \gamma]^n$ is the set of all admissible noise sequences up to time length n . Similarly, let \mathcal{U}^n be the set of all unit-amplitude-bounded input sequences of length n . It is well known [23] that regardless of the input, estimator and data-length the estimation error is always bounded from below by γ . We provide a new proof for this result based on Theorem 1 as stated below.

Theorem 3: The minimum radius of uncertainty is larger than γ for any algorithm, $\hat{h}^n(y)$.

Proof: We note that,

$$\sup_{\mathcal{W}^n} \sup_h \|h - \hat{h}^n\|_1 \geq \sup_{\mathcal{W}^n} \sup_h \|h - \hat{h}^n\|_\infty$$

Now, both \mathcal{W}^n and h are convex balanced sets, therefore it follows from the steps used in the proof of Theorem 1 that there must exist a linear algorithm that achieves optimal error for the latter (RHS of inequality above) problem. Thus there is a linear instrument, $q^n(\cdot)$, with the error estimate, E , given

by,

$$E = \max_{0 \leq m \leq p-1} \left| \sum_{t=0}^n q_m^n(t)w(t) + \sum_{j=0}^n (q_m^n(j)u(j) - 1)h(0) + \sum_{j=1}^{p-1} \sum_{l=0}^n q_m^n(l)u(l-j)h(j) \right|$$

This implies that if the error, $E < \infty$ then,

$$1 = \sum_{j=0}^n q_m^n(j)u(j) \leq \|q_m^n\|_1 \|u\|_\infty = \|q_m^n\|_1, \forall 0 \leq m \leq p-1$$

Now, we see that,

$$E \geq \sup_w \left| \sum_{t=0}^n q_m^n(t)w(t) \right| \geq \|q_m^n\|_1 \|w\|_\infty \geq \gamma, \forall m$$

thus establishing the result. \blacksquare

Theorem 1 provides a situation under which the minimum radius of uncertainty is zero in that if a small subset of noise sequences of vanishingly small measure were rendered inadmissible then the radius of uncertainty is equal to zero. This can also be seen indirectly from our results in an earlier paper [30].

V. UNMODELED DYNAMICS WITH STOCHASTIC NOISE

In this section we deal with the problem of asymptotic consistency of systems with mixed uncertainties consisting of stochastic measurement noise and worst-case (deterministic) unmodeled dynamics. The problem arises in system identification whenever restricted complexity models are used for identification. System identification with restricted complexity models have been explored in [27], [8], [35] with deterministic worst-case noise and in mixed situations in [30], [27], [36], [4]. Here we study mixed situations and derive fundamental conditions required to achieve uniform convergence of the model estimates to their optimal approximations in the model class. Our setup will deal with the following input-output equation:

$$y(t) = Hu(t) + w(t) = G(\theta)u(t) + \Delta u(t) + w(t); \quad (14)$$

where, u is a bounded input; $H = G(\theta) + \Delta \in \mathcal{S}$ is the subset of systems; $G(\theta) \in \mathcal{G}$ is a model class of interest; $w(t)$ is a Gaussian stochastic process. The main goal is to estimate $G(\theta)$ from input output data in a uniform manner as described in Section II. More specifically, we consider model classes that are subspaces, i.e.,

$$\mathcal{G} = \left\{ G \in \mathbb{RH}_2 \mid G = \sum_{k=1}^m \theta_k G_k, G_k \in \mathbb{RH}_2, \theta_k \in \mathbb{R}^m \right\} \quad (15)$$

In the following section we discuss issues that arise when the perturbations Δ are specified independent of the model class.

A. Overbounding Unmodeled Dynamics

The main goal of this section is to motivate the need for an optimal decomposition of the system into model and unmodeled dynamics. We do this by computing the radius of uncertainty for identifying an FIR model class with unmodeled dynamics. We then generalize these results to the more general case. Suppose, \mathcal{G}_{FIR} , is the class of finite impulse response sequences (FIR) of order m . Let,

$$y(t) = Hu(t) + w(t) = Gu(t) + \Delta u(t) + w(t) \quad (16)$$

where, $H \in \ell_1$, $G \in \mathcal{G}_{FIR}$, $\Delta \in \mathbf{\Delta}$ and $w(0), w(1), \dots$ is a sequence of i.i.d. jointly gaussian random variables; inputs are bounded, i.e., $|u(k)| \leq 1$, $k \in \mathbb{Z}^+$; the unmodeled class $\mathbf{\Delta}$ is given by

$$\mathbf{\Delta} = \{\Delta \in \ell_1 \mid \|\Delta\|_1 \leq \gamma\} \quad (17)$$

which indicates the fact that the m th order FIR account for the underlying system within a worst-case error of size γ . This is clearly a redundant decomposition of the underlying system. The first m -taps of the true system are accounted for both by the model and the unmodeled dynamics. This results in a minimum radius of uncertainty larger than γ .

Proposition 2: For the setup above it follows that the minimum radius of uncertainty is greater than γ .

Proof: The convexity of the set of systems, linear observations as well as linearity of the solution operator $\Phi(h) = (h_0, \dots, h_{m-1})$ can be readily verified. Now the radius of uncertainty can only be smaller if we only consider a smaller subset of systems, specifically, perturbations that satisfy $\delta(k) = 0$ for all $k \geq m$. By Theorem 1 there must exist instruments $q_l^n(\cdot)$ that are optimal for estimating the FIR model class. Consider estimation of the first tap. It follows that if the radius of uncertainty is bounded then,

$$\sum_{k=0}^n q_0^n(k)u(k) = 1$$

This implies that,

$$\left| \sum_{k=0}^n q_0^n(k)(G + \Delta)u(k) + \sum_{k=0}^n q_0^n(k)w(k) - h(0) \right| \geq \left| \sum_{k=0}^n q_0^n(k)u(k)\delta(0) \right| = \gamma$$

This proof can be generalized for other model classes as well, which we state below without proof. \blacksquare

Proposition 3: Suppose the system decomposition is convex and balanced such that for some $\eta > 0$, the model ηG is an element of the unmodeled set $\mathbf{\Delta}$, then the radius of uncertainty must be larger than η .

The main problem with these results is the apparent gap between practical reality and the minimum radius of uncertainty achieved through the decomposition of Equation 16.

For instance, consider the case when noise is relatively small. In this case the decomposition suggests that the radius of uncertainty still does not go to zero. However, in practice the first m -taps of the impulse response of H can be identified accurately for the noiseless case. In this light, the parametrization in Equation 17 over bounds the existing input-output relationship and leads to a conservative result. This issue was acknowledged by the author in [27] and it was pointed out that once a normed space of systems is chosen, there is an optimal decomposition in the sense that it minimizes the unmodeled error. Once such a decomposition is chosen the objective would be to estimate the optimal model component. For the FIR case, such a decomposition would lead to the following description for the unmodeled error:

$$\Delta = \{z^{-m}\Delta \mid \|\Delta\| \leq \gamma\} \quad (18)$$

which is optimal for ℓ_1 and \mathcal{H}_2 norms. This is because, this decomposition leads to the minimum unmodeled error.

B. Identification in Hilbert Spaces

Motivated by the discussion in the previous section we derive identification results for more general model spaces. The basic structure used in the proof (as seen from Equations 5, 6) is that we require the set \mathcal{S} after the decomposition to remain convex. On a Hilbert space this convex decomposition is preserved and we should expect Theorem 1 to apply in this setting as well. The main problem is that all elements belonging to unit balls in \mathbb{RH}_2 are not necessarily stable. Therefore, there are hardly any inputs that satisfy the requirements for uniform consistency. Therefore, the unmodeled perturbations have to be further constrained in some manner. These issues are discussed in this section.

We are now left to describe the unmodeled error. Motivated by Proposition 3 we again seek to ascribe modeled and unmodeled components to different aspects of data. This can be done naturally since the dual is also a Hilbert space and orthogonal separation between modeled and unmodeled components can be easily established. Indeed, using Cauchy's theorem we can do more as in the following proposition:

Proposition 4: Consider the model class given by Equation 15. Then, it follows that, every real-rational system, H , can be written as:

$$H = G_0 + F\Delta \quad (19)$$

for some $G_0 = G(\theta_0)$, a real-rational all pass-function F and an arbitrary perturbation Δ . The zeros of F^* are pole locations of the G_1, G_2, \dots, G_m .

Proof: The proof follows by applying Cauchy's theorem [21]. ■

e) Example: There are m poles at zero for an FIR model space of order m , and therefore the system F is a delay of order m . This implies that the unmodeled component is given by the set of all stable real-rational functions of the form $z^{-m}\Delta$, with Δ being an arbitrary element in \mathbb{RH}_2 . ■

As before we restrict the size of the perturbation and consider first the following subsets of real systems:

$$\Delta_2 = \{\Delta \mid \|\Delta\|_2 \leq \gamma\}, \quad \Delta_1 = \{\Delta \mid \|\Delta\|_1 \leq \gamma\} \quad (20)$$

As we remarked earlier the second subset is considered since the first (although natural) contains unstable systems. Consequently, as it turns out it is difficult to find inputs that lead to uniform consistency for the first set. For notational simplicity we define:

$$\begin{aligned} \mathcal{S}_2 &= \{H \in \mathbb{RH}_2 \mid H = G + F\Delta, G \in \mathcal{G}, \Delta \in \Delta_2\} \\ \mathcal{S}_1 &= \{H \in \mathbb{RH}_2 \mid H = G + F\Delta, G \in \mathcal{G}, \Delta \in \Delta_1\} \end{aligned} \quad (21)$$

We now state our main result in this section.

Theorem 4: Let $\hat{u} = Fu$ be the filtered input through the all pass filter, F and $u^l(\cdot) = G_l u(\cdot)$. A necessary and sufficient condition for uniform consistency over the set \mathcal{S}_2 is that the input be persistent and there exist a sequence of instruments, $q^n(\cdot)$ such that,

$$\sum_{k=0}^m r_{q^n u}(k) g_l(k) = r_{q^n u^l}(0) = 1, \quad l = 1, 2, \dots, m \quad (22)$$

$$\sum_{\tau=0}^n \left| r_{q^n \hat{u}}(\tau) \right|^2 \xrightarrow{n \rightarrow \infty} 0; \quad |r_{q^n w}(0)| \xrightarrow{n \rightarrow \infty} 0 \text{ in probability}$$

On the other hand if \mathcal{S}_1 is considered then the conditions for uniform consistency remains the same except that the second condition is weaker, i.e.,

$$\max_{1 \leq \tau \leq n} \left| r_{q^n \hat{u}}(\tau) \right| \xrightarrow{n \rightarrow \infty} 0 \quad (23)$$

Furthermore, these conditions imply that the input must be persistently exciting of infinite order.

Proof: First we consider the set \mathcal{S}_2 . We note that the projection ΠH of the system H onto the subspace \mathcal{G} is linear and of finite rank. Therefore, the solution operator $\Phi(H) = (\theta_1, \dots, \theta_m)^T$ is linear. The subset \mathcal{S}_2 is convex and the observation given by Equation 14 is linear. For the sake of exposition we next consider one-parameter model class. The general case follows in an identical manner. Now by Theorem 1 if uniform consistency holds there must exist a number n and an instrument, $q^n(\cdot)$ so that, for any $\epsilon > 0$, $\xi > 0$

$$\left| \sum_{k=0}^n q^n(k) w(k) + \sum_{j=0}^n \sum_{k=0}^{n-j} q^n(k+j) u(k) h(j) - \theta_1 \right| \leq \epsilon, \quad \forall \theta_1 \in \mathbb{R}$$

with $w \in \mathcal{W}_0^n$ and $\text{Prob}\{\mathcal{W}_0^n\} \geq 1 - \xi$. As before, this implies the condition that the instrument must de-correlate the noise, i.e., $|\sum_{k=0}^n q^n(k) w(k)| \rightarrow 0$. We are now left with an instrument that must satisfy,

$$\begin{aligned} & \left| \sum_{j=0}^n \sum_{k=0}^{n-j} q^n(k+j) u(k) h(j) - \theta_1 \right| \\ &= \left| \sum_{j=0}^n \sum_{k=0}^{n-j} q^n(k+j) u(k) (\theta_1 g_1(j) + (f_1 * \delta)(j)) - \theta_1 \right| \leq \epsilon \end{aligned}$$

where, $g_1(k), f_1(k), \delta(k)$ are the impulse responses for G_1, F_1, Δ respectively. Observe that we have applied Proposition 4 and substituted $F_1\Delta$ for the perturbation and so F_1 and G_1 are perpendicular. By arbitrarily varying θ, Δ we see that,

$$\sum_{j=0}^n r_{q^n u}^n(j) g_1(j) = r_{q^n u^1}(0) = 1; \quad \sup_{\Delta \in \Delta_2} \left| \sum_{j=0}^n r_{q^n \hat{u}}^n(j) \delta(j) \right| \leq \epsilon,$$

where, for the first expression $u^1 = G_1 u$, and in the second expression we have absorbed the filter F_1 into the input, i.e., $\hat{u} = F_1 u$. The condition arising from the noise and the first expression together imply that the input must be persistently exciting. This follows from the fact that,

$$1 = \left| \sum_{j=0}^n q^n(j) u^1(j) \right| \leq \|q^n\|_2 \|G_1\|_{\mathcal{H}_\infty} \|P_n u\|_2$$

where, $\|\cdot\|_{\mathcal{H}_\infty}$ is the usual \mathcal{H}_∞ norm. But the noise condition implies that $\|q^n\| \rightarrow \infty$. Therefore, $\|P_n u\| \rightarrow \infty$, since G_1 is bounded in \mathcal{H}_∞ because it is stable real rational transfer function. Next for the perturbation we apply Cauchy-Schwartz inequality to obtain:

$$\sup_{\Delta \in \Delta_2} \left| \sum_{j=0}^n r_{q^n \hat{u}}^n(j) \delta(j) \right|^2 = \gamma \left| \sum_{j=0}^n (r_{q^n \hat{u}}^n(j))^2 \right| \rightarrow 0$$

For the set Δ_1 we similarly have,

$$\sup_{\Delta \in \Delta_1} \left| \sum_{j=0}^n r_{q^n \hat{u}}^n(j) \delta(j) \right| = \gamma \max_{j \leq n} |r_{q^n \hat{u}}^n(j)| \rightarrow 0$$

This establishes the necessity. We will not describe the sufficiency conditions since the argument is similar to the argument for sufficiency for FIR model classes, which is dealt in the next section. ■

Remark: To see how the convex separation results in achieving uniform consistency note that the conditions on the input from the modeled part implies that $G_1 u(\cdot)$ is aligned with the instrument $q^n(\cdot)$. Similarly, the unmodeled part implies that $F_1 u(\cdot)$ is perpendicular to $q^n(\cdot)$. This is possible because G_1 and F_1 are orthogonal and so $G_1 u$ and $F_1 u$ are uncorrelated when driven by white noise.

Remark: We point out that the conditions for \mathcal{S}_2 are more demanding than \mathcal{S}_1 . For the set \mathcal{S}_1 we only require that the normalized worst-case cross-correlation between input and instrument to go to zero. This condition can easily be accomplished for inputs that have small autocorrelation coefficients, i.e., inputs that are white in their second order properties. Then an appropriately filtered input can be selected as the instrument. In the \mathbb{RH}_2 case we need a more stringent condition since we need the sum of the squares of the normalized cross-correlation also go to zero. This is understandable on account of the fact that \mathbb{RH}_2 forms a larger class of perturbations than ℓ_1 . ■

We next discuss issues of optimality and choice of inputs. First, notice that one could incorporate additional parametric and non-parametric constraints as we did in Section IV and Equation 12. For these set of problems the fact that the optimal algorithm is linear follows from Theorem 1. The problem now reduces to looking for inputs satisfying Equation 22. The constraints on the input (disregarding the noise constraint) imply that the optimal instrument is the solution to:

$$\min(q^n)^T \Psi_n \Psi_n^T (q^n) \quad \text{subject to} \quad \sum_{j=0}^n q^n(j) u(j) = 1$$

where, $\Psi_n = [\psi(0), \psi(1), \dots, \psi(n)]^T$, with $\psi(t) = [\hat{u}(t), \hat{u}(t-1), \dots, \psi(0)]$. Let, $\mathbf{u} = [u(0), \dots, u(n)]^T$. If the value of the solution to the quadratic optimization problem has a value greater β (this is the opposite of what we want) then:

$$(q^n)^T \Psi_n \Psi_n^T (q^n) = \frac{1}{\mathbf{u}^T (\Psi_n \Psi_n^T)^{-1} \mathbf{u}} \geq \beta \iff \begin{bmatrix} 1/\beta & \mathbf{u}^T \\ \mathbf{u} & \Psi_n \Psi_n^T \end{bmatrix} \geq 0$$

where, we have used Schur's complementation to derive the last expression above. We now argue the difficulty of finding inputs that satisfy this condition by applying it to a 1-tap FIR model class. For this situation the above condition implies that,

$$\sum_{j=1}^n (r_u^n(j))^2 \geq \beta$$

If the input is chosen i.i.d. then the expected value of the expression on the left is bounded away from zero. Therefore, for a $\beta > 0$ (independent of n) the above condition is met with probability one leading us to believe that few inputs satisfy the opposite condition, i.e., $\sum_{j=1}^n (r_u^n(j))^2 \ll \beta$.

On the other hand the weaker condition for Δ_1 can be met by a number of inputs. In particular we only need to establish that there exist inputs with small autocorrelations with high probability. Then, by choosing an appropriately filtered input as the instrument the conditions of Equation 23 can be satisfied.

Lemma 2: Suppose $u(t)$ is a discrete i.i.d. bernoulli random process with mean zero and variance 1, then:

$$\text{Prob} \left\{ \max_{0 < k \leq n} \|r_u^n(k)\|_\infty \geq \alpha \right\} \leq n \exp(-n\beta(\alpha)) \quad (24)$$

where $\beta(\alpha) = 1 + \frac{1-\alpha}{2} \log_2 \frac{1-\alpha}{2} + \frac{1+\alpha}{2} \log_2 \frac{1+\alpha}{2}$

Proof: The proof of this result can be found in [6].

In summary we have characterized conditions for uniform consistency and optimality for Hilbert spaces, where convex separation between model class and unmodeled dynamics proved to be critical. In the next section we consider a situation where such a convex separation does not exist.

C. Identification in ℓ_1 and non-linear functionals

We now turn to the problem of identification in the ℓ_1 norm. In the previous section the Hilbert space structure led to convex decomposition of the unmodeled error and model class and Theorem 1 could be applied in a straightforward manner.

For FIR model classes, under the ℓ_1 norm, the decomposition is still convex. This is due to the fact that optimal FIR approximation in ℓ_1 and \mathbb{RH}_2 are identical. However for more general classes under the ℓ_1 norm the decomposition through optimal approximations is not convex. In the next section, we present identification of FIR model class in the ℓ_1 norm. The results developed for FIR case will be used to compute uniformly consistent estimators for more general model classes.

D. FIR Model Class

In this section we derive necessary and sufficient conditions for identification of FIR model classes in the presence of unmodeled dynamics under the ℓ_1 norm. To this end, consider the optimal decomposition:

$$\mathcal{S} = \{H \in \ell_1 \mid H = G + \Delta, G \in \mathcal{G}_{FIR}, \Delta \in \mathbf{\Delta}\} \quad (25)$$

where, $\mathbf{\Delta} = \{z^{-m}\Delta \mid \|\Delta\|_1 \leq \gamma\}$; \mathcal{G}_{FIR} is the class of m-tap FIR sequences.

In the following theorem we derive necessary and sufficient conditions for achieving uniform consistency for estimating the FIR coefficients.

Theorem 5: The radius of uncertainty for the above setup is equal to zero if and only if there is a persistent input and a corresponding family of instrument variables $q^n(\cdot)$ which uniformly de-correlates the input and noise, i.e., a number $N(\rho)$ for every $\rho > 0$ such that for all $n > N(\rho)$:

$$\max_{1 \leq \tau \leq n} |r_{q^n u}^n(\tau)| \xrightarrow{n \rightarrow \infty} 0, \quad |r_{q^n u}^n(0)| = 1 \quad (26)$$

$$|r_{q^n w}^n(0)| \xrightarrow{n \rightarrow \infty} 0 \text{ in probability} \quad (27)$$

Proof: (\implies) We observe that all norms are equivalent on finite dimensional spaces [15]. Consequently, uniform consistency in the ℓ_1 norm for the first m FIR taps is identical to uniform consistency in the ℓ_∞ norm. This implies by hypothesis that, the minimum radius of uncertainty is zero for the ℓ_∞ norm as well. Therefore, given $\epsilon > 0$, $\xi > 0$ there is an algorithm, \hat{h} and a number $L(\epsilon, \xi, \hat{h}, \mathcal{S})$ such that $\text{Prob}\left\{\sup_{H \in \mathcal{S}} \|P_m(h - \hat{h}^n)\|_\infty \geq \epsilon\right\} \leq \xi$, $n \geq L(\epsilon, \xi, \hat{h}, \mathcal{S})$. The setup now mirrors Theorem 1 since the set, \mathcal{S} is convex; $\Phi(H) = (h(0), \dots, h(m-1))^T$; w is i.i.d. gaussian noise; measurement structure is linear, i.e., $\lambda_t(h) = Hu(t)$; norm measure on $\Phi(\cdot)$ is the ℓ_∞ norm. Consequently, by hypothesis, there is a linear algorithm with weights, $q_l^n(\cdot)$, for each FIR tap, $0 \leq l \leq m$, such that,

$$\sup_{\mathbf{w} \in \mathcal{W}_0^n} \sup_{H \in \mathcal{S}} \left| \sum_{k=0}^n q_l^n(k) y(k) - h(l) \right| \leq \epsilon; \quad \forall n \geq N(\epsilon, \xi)$$

and $\text{Prob}\{\mathcal{W}_0^n\} \geq 1 - \xi$ Upon substitution it follows that,

$$\left| \sum_{k=0}^n q_l^n(k) w(k) + \sum_{k=0}^n (q_l^n(k) u(k) - 1) h(k) + \sum_{j=m}^n \sum_{k=0}^{n-j} q_l^n(k+j) u(k) h(j) \right| \xrightarrow{n \rightarrow \infty} 0$$

Next, by noting that Δ is arbitrary with ℓ_1 norm less than γ and $w(\cdot)$ is an arbitrary element in the convex set \mathcal{W}_0^n , which also contains the zero element we have,

$$\sum_{k=0}^n q_l^n(k) u(k) = 1; \quad \max_{1 < j < n} \left| \sum_{k=0}^{n-j} q_l^n(k+j) u(k) \right| \leq \epsilon \\ \left| \sum_{k=0}^n q_l^n(k) w(k) \right| \leq \epsilon, \quad \mathbf{w} \in \mathcal{W}_0^n$$

The proof now follows by first observing that for the hypothesis to be true we must have, $\text{Prob}\{\mathcal{W}_0^n\} \geq 1 - \xi$. The persistency of the input follows along similar lines as in Theorem 2.

(\Leftarrow) To prove sufficiency we construct an IV technique that is uniformly consistent. Basically we let, $q^n = q_1^n$ and its delayed copies up to m-delays serve as instruments. In particular, for length n of data, we let, Q_n be the matrix composed of the instrument q^n and its m delays, U_n a corresponding matrix of input sequence with its m delays and R_n the residual delays of the input sequence.

$$Q_n = \begin{bmatrix} q^n(0) & 0 & \dots & 0 \\ q^n(1) & q^n(0) & 0 & \dots \\ \vdots & \vdots & \ddots & q^n(0) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad (28)$$

$$U_n = \begin{bmatrix} u(0) & 0 & \dots & 0 \\ u(1) & u(0) & 0 & \dots \\ \vdots & \vdots & \ddots & u(0) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}; \quad R_n = \begin{bmatrix} 0 & 0 & \dots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ u(0) & 0 & \dots \\ \vdots & \ddots & \ddots \end{bmatrix}$$

Suppose, \mathbf{g} is the column vector composed of the first m impulse response coefficients, i.e., $\mathbf{g} = P_m H$ and δ is the column vector of impulse response of the unmodeled dynamics, i.e., $\delta = (P_n - P_m)H$. Let \mathbf{y} , \mathbf{w} be the output and noise signal vectors of length N respectively. Then,

$$\mathbf{y} = U_n \mathbf{g} + R_n \delta + \mathbf{w} \implies Q_n^T \mathbf{y} = Q_n^T U_n \mathbf{g} + Q_n^T R_n \delta + Q_n^T \mathbf{w} \quad (29)$$

It follows from hypothesis that $Q_n^T U_n$ is an invertible matrix with bounded inverse for large enough N since it converges to a lower triangular matrix with identical elements on the diagonal. Again from our hypothesis together with the fact that that $\|\Delta\| \leq \gamma$, it follows that, $Q_n^T R_n \delta$ converges to zero. To show this, note that the k th row of $Q_n^T R_n$ is given by,

$$(Q_n^T R_n)_k = \begin{bmatrix} \sum_{j=0}^{n+1-(m+k)} q^n(m+(k-1)+j)u(j), \\ \sum_{j=0}^{n-(m+k)} q^n(m+k+j)u(j), \dots, q^n(m+(k-1)+j)u(0) \end{bmatrix}$$

Consequently, by hypothesis it follows that,

$$|(Q_n^T R_n)_k \delta| \leq \|\Delta\|_1 \max_{\tau} \left| \sum_{j=0}^{n+1-(m+k+\tau)} q^n(m+(k-1)+j+\tau)u(j) \right| \leq \gamma \epsilon$$

Finally, $Q_n^T \mathbf{w}$ also converges to zero in expectation by hypothesis and noting that $(w(0), w(1), \dots, w(n))$ are i.i.d. random variables. ■

We next discuss the issue of sample complexity. The sample-complexity of an algorithm, $\hat{\Phi}(y)$ is said to polynomial [32] in learning theory if

$$L(\epsilon, \xi, \hat{\Phi}, \mathcal{S}) \leq \left(\frac{1}{\epsilon}\right)^d \log\left(\frac{1}{\xi}\right)$$

for some integer d . We establish polynomial sample complexity of linear algorithms for the FIR model class here. First, we choose instruments $q^n(k) = u(k)/n$, where the input sequence is a bernoulli process. It is then easy to establish that,

$$\begin{aligned} & \text{Prob} \left\{ \sup_{H \in \mathcal{S}} \|Q_n^T \mathbf{y} - P_m(H)\|_{\infty} \geq \epsilon \right\} \leq \\ & \text{Prob} \left\{ \max_{0 < \tau \leq n} \frac{|r_u^n(\tau)| + |r_{uw}^n(\tau)|}{n} \geq \epsilon \right\} \leq \\ & \text{Prob} \left\{ \max_{0 < \tau \leq n} \frac{|r_u^n(\tau)|}{n} \geq \frac{\epsilon}{2} \right\} + \text{Prob} \left\{ \max_{0 < \tau \leq n} \frac{|r_{uw}^n(\tau)|}{n} \geq \frac{\epsilon}{2} \right\} \leq \\ & n \exp(-n\beta(\epsilon)) + n \exp(-n\epsilon^2/\sigma^2) \end{aligned}$$

where we have taken the probability over both the input and the noise process. The first term in the last expression uses Lemma 2 and the second term uses the fact that w is gaussian. Now, it follows directly by substitution that, for some $C > 0$,

$$L(\epsilon, \xi, \hat{\Phi}, \mathcal{S}) \leq C \left(\frac{1}{\epsilon}\right)^2 \log\left(\frac{1}{\xi}\right)$$

implying polynomial sample-complexity of convergence.

E. General Model Classes

In this section we derive algorithms for achieving uniform consistency for general model classes (that are subspaces) in the ℓ_1 norm. The main difficulty is that the decomposition achieved through optimal approximation is not convex. Hence, Theorem 1 cannot be directly applied. We proceed by first describing the optimal approximation as a known non-linear function of linear functions on the underlying system space. Since linear functions of the system can be efficiently estimated the non-linear function can be estimated as well. Our setup consists of,

$$y(t) = Hu(t) + w(t) = Gu(t) + \Delta u(t) + w(t), \quad G \in \mathcal{G}, \quad \Delta \in \mathbf{\Delta}$$

where, $\mathcal{G} = \left\{ \sum_{j=1}^p \alpha_j G_j \mid \alpha_j \in \mathbb{R}, G \in \mathbb{RH}_2 \right\}$ and $\mathbf{\Delta} = \{ \|\Delta\| \leq \gamma \mid \Delta \in \mathcal{G}^{\perp} \}$, where \mathcal{G}^{\perp} is the space of annihilators

of the model space \mathcal{G} . The decomposition follows from well-known results in ℓ_1 approximation theory [15]. Noise, $w(\cdot)$ is assumed to be i.i.d. gaussian noise as before. Our task is to determine the model G_0 that minimizes the unmodeled error, i.e., $G_0 = \text{argmin}_{G \in \mathcal{G}} \|H - G\|$. The following example illustrates the issue of non-convexity of ℓ_1 decompositions.

f) *Example:* Consider, the following convex space of systems,

$$\mathcal{S} = \{H \in \ell_1 \mid H_{\alpha} = \alpha H_1 + (1 - \alpha)H_2; \alpha \in [0, 1]\}$$

where $H_1(z^{-1}) = 1$ and $H_2(z^{-1}) = z^{-1} + \left(1 - \frac{z^{-1}}{2}\right)^{-1}$. Our model class is given by $\mathcal{G} = \left\{ \theta \left(1 - \frac{z^{-1}}{2}\right)^{-1} \mid \theta \in \mathbb{R} \right\}$. From the alignment conditions for optimality [15] it follows that the unique optimal approximation, $G(H_1)$, for the system H_1 is $G(H_1) = 0$. Similarly, we deduce that $G(H_2) = \left(1 - \frac{z^{-1}}{2}\right)^{-1}$. Moreover, $G(H_{\alpha}) = G(H_2)$ for all $\alpha \neq 1$ is the unique optimal approximation as well. Therefore, the set of optimal approximants for the convex set of systems \mathcal{S} is not convex. ■

Motivated by these issues we solve the problem of ℓ_1 identification indirectly. The idea is that ℓ_1 minimizer can be arbitrarily closely approximated by a non-linear function of linear functions of the system impulse response, i.e.,

$$G_0(H) = \psi(\mathcal{L}(H))$$

where, $\psi(\cdot)$ is non-linear continuous map, which maps to the model subspace and \mathcal{L} is a finite rank linear operator mapping the system H to an appropriate finite dimensional space. Uniform consistency in ℓ_1 would then follow by estimating the linear functions, whose consistency conditions are in turn characterized by Theorem 5. To this end we have the following theorem.

Theorem 6: The input characterization as in Theorem 5 is sufficient for uniform consistency in the ℓ_1 norm.

The proof of the theorem relies on a number of results that are provided below.

Proposition 5: The optimal ℓ_1 approximation $G_0(H)$ for a given system, H , can be re-written as,

$$G_0(H) = \Pi(H) + \text{argmin}_{G \in \mathcal{G}} \|H - \Pi H - G\| = \lim_{m \rightarrow \infty} G_0(P_m H)$$

where, Π is the ℓ_2 projection on to the space \mathcal{G} .

Proof: The proof follows from the following equality:

$$\min_{G \in \mathcal{G}} \|H - G\|_1 = \min_{G \in \mathcal{G}} \|H - \Pi H - G\|_1$$

The limiting argument follows from the fact that $P_m H \rightarrow H$ and continuity of projection operation and the ℓ_1 norm. ■

For notational simplicity we denote the modified problem on the RHS by G_r , i.e.,

$$G_r(H) = \text{argmin}_{G \in \mathcal{G}} \|H - \Pi H - G\|$$

and the modified system by, H_r ,

$$H_r = H - \Pi H$$

The main task is to now show how to further simplify the problem to an instantiation of Theorem 1. First, note that ΠH is a linear and the range of the projection is a finite dimensional space. Therefore, under the conditions derived in Theorem 4, we can consistently estimate the projection in a uniform manner. The problem reduces to estimating $G_r(H)$. In this regard, we have the following proposition:

Proposition 6: If there is $G \in \mathcal{G}$ such that $\|H - G\|_1 \leq \gamma$ then there is some constant L such that, $\|H - \Pi H\|_1 \leq L\gamma$

Proof: By hypothesis we have that $\|H - \Pi H\|_2 \leq \|H - G\|_2 \leq \|H - G\|_1 \leq \gamma$. We now use the state-space form for ΠH , G , as described by Equation 1. Since, \mathcal{G} is a subspace, the state-transition matrix, A , is fixed. We assume that control vector B is the free parameter that needs to be estimated. Let B_1, B_2 be the corresponding control vectors for $G, \Pi H$ respectively. It follows,

$$\|G - \Pi H\|_2 \leq \left\| \begin{bmatrix} C & CA & CA^2 & \dots & CA^n \end{bmatrix}^T (B_2 - B_1) \right\|_2 \leq 2\gamma \quad (30)$$

Since (A, C) is observable the matrix $\begin{bmatrix} C & CA & CA^2 & \dots & CA^n \end{bmatrix}^T$ is invertible. Therefore, $\|B_2 - B_1\|_2 \leq K_0\gamma$ for some constant $K > 0$ that depends on the system matrices, A, C . Since, B_2, B_1 are finite dimensional and norms on finite dimensional spaces are equivalent, we also have $\|B_2 - B_1\|_1 \leq K_1\gamma$ for some $K_1 > 0$. This implies that $\|G - \Pi H\|_1 \leq K_2\gamma$ for some $K_2 > 0$. It now follows that,

$$\|H - \Pi H\|_1 \leq \|H - G\|_1 + \|G - \Pi H\|_1 \leq L\gamma$$

Our final proposition states that,

Proposition 7: Let, G_m denote the best ℓ_1 approximation of the FIR system $P_m H_r$, i.e., $G_m = \operatorname{argmin}_{G \in \mathcal{G}} \|P_m(H_r - G)\|_1$. Then it follows that,

$$\lim_{m \rightarrow \infty} \|P_m(H_r - G_m)\|_1 = \|H_r - G_r\|_1$$

Proof: The proof follows by the following sequence of inequalities:

$$\begin{aligned} \|H_r - G_r\|_1 &\leq \|H_r - G_m\|_1 \\ \|P_m(H_r - G_m)\|_1 &\leq \|P_m(H_r - G_r)\|_1 \end{aligned}$$

Therefore,

$$\begin{aligned} 0 &\leq \|P_m(H_r - G_r)\|_1 - \|P_m(H_r - G_m)\|_1 \\ &\leq \|(I - P_m)(H_r - G_m)\|_1 - \|(I - P_m)(H_r - G_r)\|_1 \rightarrow 0 \end{aligned}$$

where the final limit follows from the fact that $\|G_m\|_1$ and $\|G_r\|_1$ are bounded and have exponentially decaying tails. ■ The theorem now follows from the fact that, G_0 can be decomposed as a known non-linear operator acting on a collection of linear functionals on the system H , i.e., $\psi(\Pi H, P_m(H - \Pi H))$ where: 1) Estimate of an FIR model $P_m(H - \Pi H)$ for the modified system $H - \Pi H$ can be obtained consistently if conditions in Theorem 5 are satisfied; 2) ΠH can be obtained if conditions of Equation 23 are satisfied, which are equivalent to conditions in Theorem 5. ■

VI. UNCERTAIN LFT STRUCTURES

We consider several different LFT structures for which the tools developed in Theorem 1 can be applied to characterize conditions on inputs, which achieve uniform consistency. To focus attention on structured perturbations we limit our attention to real-rational space of systems endowed with the \mathcal{H}_2 norm.

A. Multiplicative Uncertainty

We consider first the multiplicative uncertainty as shown in Figure 2(a) with $H \in \mathcal{H}_2$. The input-output equation is given by:

$$y(t) = Hu(t) + w(t) = G(1 + \Delta)u(t) + w(t)$$

where, $\mathcal{G} = \left\{ \sum_{j=1}^p \alpha_j G_j \mid \alpha_j \in \mathbb{R} \right\}$, the noise w is random i.i.d. gaussian noise and Δ is a multiplicative perturbation with bounded ℓ_1 norm and accounts for the undermodeling of H . Upon substitution we have,

$$\begin{aligned} y(t) &= \sum_{j=1}^p \alpha_j G_j (1 + \Delta)u(t) + w(t) \\ &= \sum_{j=1}^p \alpha_j (1 + \Delta)G_j u(t) + w(t) \\ &= \sum_{j=1}^p \alpha_j \psi_j(t) + \Delta \sum_{j=1}^p \alpha_j \psi_j(t) + w(t) \end{aligned}$$

where we have used the commutativity of the convolution operation and substituted $\psi_j(t) = G_j u(t)$. The parametrization is non-convex as shown in the following example.

g) Example: Consider the two elements $H_1 = -(1 - \Delta)G$ and $H_2 = (1 + \Delta)G$ in the setup above. Let the set of perturbations be bounded, i.e., $\|\Delta\| \leq \epsilon$ then $(H_1 + H_2)/2 = \Delta G$ cannot be expressed as $(1 + \tilde{\Delta})G$ for some $\|\tilde{\Delta}\| \leq \epsilon$. Therefore, the decomposition is not convex. However, if we restrict $\alpha_j \geq 0$ the decomposition turns out to be convex. In particular, the following subset of systems is convex.

$$\mathcal{S}^+ = \left\{ G(1 + \Delta) \mid G = \sum_j \alpha_j G_j, \alpha_j \geq 0, \|\Delta\| \leq \epsilon \right\}$$

Theorem 7: A necessary condition for uniform consistency is that Δ be orthogonal to the space \mathcal{G} and the inputs satisfy conditions of Theorem 4.

Proof: We first consider a convex subset, \mathcal{S}_0 , of systems. Specifically, constrain the parameters $\alpha_j \in [1, 2]$. Then $\mathcal{S}_0 = \sum_j (\alpha_j G_j + \tilde{\Delta} G_j) \subset \sum_j \alpha_j G_j (1 + \Delta) = \mathcal{S}$, where $\|\tilde{\Delta}\|, \|\Delta\|$ are both smaller than γ . Now translating the set \mathcal{S}_0 by modifying the variables $\tilde{\alpha}_j = \alpha_j - 3/2$ we get a new set \mathcal{S}_1 which is both balanced and convex. We can now apply Proposition 3 and Theorem 1 to argue that Δ and \mathcal{G} must be orthogonal for achieving uniform consistency. Furthermore, from Theorem 4 it follows that the relevant input conditions must also be satisfied. Now, there is a one-to-one mapping from the collection \mathcal{S}_1 to the collection \mathcal{S}_0 and therefore the

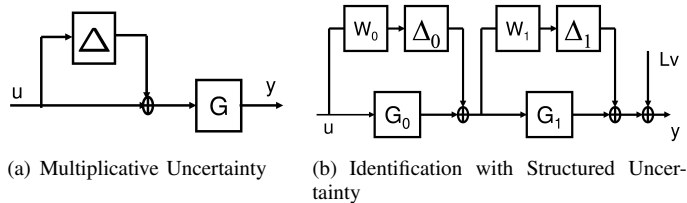


Fig. 2. Different Classes of Linear Fractional Structures; The second figure illustrates a serial connection of two components one of which is approximately known, while the other is unknown and must be estimated from input-output data

same conditions are necessary for \mathcal{S}_0 as well. Since, $\mathcal{S}_0 \subset \mathcal{S}$ the necessity holds for the general case as well. ■

We next apply the instrument derived for the set \mathcal{S}_1 to the general problem. It turns out that the notion of uniform consistency must be modified suitably. Indeed, since the perturbation is scaled by the parameter the error bound also has a multiplicative structure, i.e., we can only say that

$$\|\alpha - \alpha^n\|_\infty \leq \beta \|\alpha\|_\infty$$

with high probability for a large enough n . This is weaker than the conditions derived in the previous sections.

B. Serial Connections of Uncertain Systems

In this problem, shown in Figure 2(b), two uncertain systems, H_0, H_1 are connected in series. The system H_1 is approximately known, i.e., its best approximation G_1 (in the class \mathcal{G}_1) is known, and the system H_0 must be approximately identified in the model class, \mathcal{G}_0 from data. The interpretation of the setup is that we are given a set of uncertain components connected in series, where it is not possible to isolate any single component. The question arises as to when an approximate model for the uncertain component can be derived. We first write the input-output equation as follows:

$$y(k) = (G_0 + \Delta_0)(G_1 + \Delta_1)u(k) + w(k) \quad (31)$$

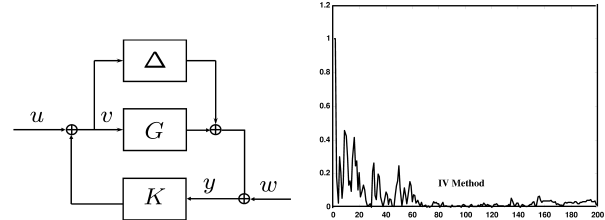
It is clear from the setup that the unmodeled dynamics enters the equations in a non-linear non-convex fashion. Nevertheless, necessary and sufficient conditions can still be established.

Theorem 8: A necessary and sufficient condition for uniform consistency of the above setup is that Δ be orthogonal to both model classes \mathcal{G}_0 and \mathcal{G}_1 and the input satisfy conditions of Theorem 4.

Proof: (necessity) Upon expanding Equation 31 we have,

$$y(t) = G_0 G_1 u(t) + G_1 \Delta_0 u(t) + G_0 \Delta_1 u(t) + \Delta_0 \Delta_1 u(t) + w(t) \quad (32)$$

Now, uniform consistency of G_0 implies uniform consistency of the above setup with $\Delta_1 = 0$ or $\Delta_0 = 0$. The former case reduces to an additive perturbation for which $\Delta \perp \mathcal{G}_0$ is a necessary condition, while the latter case is a multiplicative perturbation for which $\Delta \perp \mathcal{G}_1$ is necessary according to Theorem 7. The conditions on the input follow from Theorems 4.



(a) Block Diagram of Uncertain System in a Feedback Loop

(b) Identification error

Fig. 3. The unknown system has both parametric and non-parametric components with the parametric part to be identified from closed loop data; The setup is unconventional in that the system input is also measured; This arises in a number of applications involving indoor echo-cancellation Systems.

(sufficiency) Comparing Equation 32 with Equation 14 we see a potential difference only in the penultimate term, i.e., $\Delta_0 \Delta_1 u(t)$. Now since Δ_0 is perpendicular to \mathcal{G}_0 , it can be written as $F_0 \tilde{\Delta}$, for an arbitrary bounded perturbation $\tilde{\Delta}$, where F_0 is as described in Proposition 4. By absorbing $\tilde{\Delta} \Delta_1$ into a single perturbation, Δ , we can write, $\Delta_0 \Delta_1 = F_0 \Delta$. The new expression obtained is now consistent with conditions of Theorem 4, which then establishes sufficiency. ■

h) Example: As an instantiation of the theorem consider the case when G_0 and G_1 are FIRs of order m and n respectively. Then no consistent algorithm exists when the order, m , of G_0 is larger than n . This is because the unmodeled error $G_0 \Delta_1$ and the class of FIR models of order m are no longer orthogonal.

C. Identification of Uncertain Systems in Feedback Loops

Consider the feedback setup shown in Figure 3(a). The goal is to estimate the optimal approximation to $H = G + \Delta$ in the model class $G \in \mathcal{G}$ from input-output data. Our task is to design the input signal u to guarantee robust convergence to the optimal approximation.

Such problems arise in the context of design of adaptive acoustic echo-cancellation systems that operate in a feedback loop [28]. In general the high order acoustic dynamics coupled with significant time variation severely limits the choice of echo-canceller order. Moreover, the feedback filter, typically used to enhance the quality of speech, can impact the stability of the echo-cancellation system. In this light, a lower order echo-cancellation system is designed in the hope that robust performance can be ensured at the cost of sacrificing some performance. A dither input signal, u , is employed as an input to the loudspeaker and the output, v , of the loudspeaker is recorded in addition to the microphone signals, y . Thus the input output data consists of (u, v, y) as illustrated in Figure 3(a). Our setup consists of:

$$y(t) = H v(t) + w(t) = \sum_{j=1}^m \alpha_j G_j v(t) + F \Delta v(t) + w(t) \quad (33)$$

where, w is i.i.d. gaussian noise and we have decomposed $H \in \mathbb{RH}_2$ into orthogonal subspaces. Furthermore, we assume that the closed loop system is stable and the exogenous input $u(\cdot)$

and the internal signal, $v(\cdot)$, and the output signal $y(\cdot)$ are both measured. As the following theorem shows that a necessary and sufficient condition for uniform consistency remains the same, i.e.,

Theorem 9: A necessary and sufficient condition for uniform consistency of the optimal restricted complexity model $G \in \mathcal{G}$ is that the input satisfy conditions in Theorem 4.

Proof: By straightforward algebraic manipulations of the feedback loop we obtain that:

$$v(t) = W_1 u(t) + W_2 w(t), \quad W_1 = (1 - HK)^{-1}, \quad W_2 = KW_1$$

Although the system, W , is generally unknown (since it is a LFT of the controller with the unknown system), the fact that the controller is stabilizing implies W_1, W_2 are stable transfer functions. Therefore, $v(t)$ is bounded. Now consider Equation 33, which is exactly in the form of Equation 14 with the unmodeled dynamics orthogonal to \mathcal{G} . These are now consistent with assumptions of Theorem 4 with $u(t)$ replaced with $v(t)$. We are left to establish the conditions on the input signal. Uniform consistency holds if and only if there exists an instrument, $q^n(\cdot)$ that decorrelates the internal signal $v(t)$ and noise $w(t)$, i.e.,

$$\left| \sum_{k=0}^n q^n(k)v(k) \right| = 1, \quad \left| \sum_{k=0}^n q^n(k+j)v(k) \right| \rightarrow 0$$

and

$$\left| \sum_{k=0}^n q^n(k)w(k) \right| \rightarrow 0$$

This in turn implies that the filtered instrument, $\tilde{q}^n(\cdot) = W_1 q^n(\cdot)$ decorrelates the input signal and noise:

$$\left| \sum_{k=0}^n \tilde{q}^n(k)u(k) \right| = 1, \quad \left| \sum_{k=0}^n \tilde{q}^n(k+j)u(k) \right| \rightarrow 0$$

$$\left| \sum_{k=0}^n \tilde{q}^n(k)w(k) \right| \rightarrow 0$$

Note that in the above example since W_1 is unknown a convenient instrument for $v(t)$ is the filtered dither signal $u^1(t) = Gu(t)$. This is because when $u(t)$ is a i.i.d. dither signal:

$$\frac{1}{n} \left| \sum_{k=0}^n u^1(k)v(k) \right| \geq \beta > 0; \quad \frac{1}{n} \left| \sum_{k=0}^n u^1(k+j)v(k) \right| \rightarrow 0$$

and

$$\frac{1}{n} \left| \sum_{k=0}^n u^1(k)w(k) \right| \rightarrow 0$$

i) Example: The model space is given by $\mathcal{G} = \frac{\alpha}{1-0.3z^{-1}}$. The unmodeled dynamics is generated by normalizing a zero mean random gaussian sequence of length 1000 and normalizing it to have an ℓ_1 norm equal to one. The unmodeled dynamics is then filtered with the all-pass filter $F = (z^{-1} - 0.3)/(1-0.3z^{-1})$. This ensures orthogonal separation between the unmodeled error and the model subspace. The system H is generated by summing the model for $\alpha = 1$ with the unmodeled dynamics. Next, a controller K is chosen so that the closed loop system is stable and such that the sensitivity function W_1 has ℓ_1 norm smaller than one. A uniformly distributed i.i.d. input in the interval $[-0.5, 0.5]$ of length 300 is applied and the measurements $v(t)$ and $y(t)$ are collected in i.i.d. gaussian noise $w(k)$ with mean zero and variance 0.1. For the IV technique we use $(1 - 0.3z^{-1})^{-1}u(t)$ as the instrument in Equation 33 and identify α . Figure 3(b) shows the parametric error plotted as a function of data length for the IV scheme. We see that the IV technique converges, while it is well known that the least squares approach leads to large bias. This is not surprising considering the fact that the unmodeled output is correlated with the model output and so while the least squares approach leads to a large bias, the IV technique is chosen suitably to decorrelate both the uncertainty as well as the noise.

Remark: We contrast the IV approach presented in this paper with the well-known MPE approach. The idea in the MPE approach is to find models that minimize the so called prediction error. In specific circumstances this implies a preference for those models, where the prediction error is uncorrelated with model output. In the presence of complex uncertainties, especially when the error arises due to under modeling, this approach can lead to significant bias. IV methods can overcome this problem through suitable instruments. These instruments in specific circumstances can first decorrelate the uncertainty arising from correlated sources and then proceed to cancel stochastic uncorrelated noise as well.

VII. CONCLUSIONS

We have introduced a new concept for system identification with mixed stochastic and deterministic uncertainties that is inspired by machine learning theory. In comparison to earlier formulations our concept requires a stronger notion of convergence in that the objective is to obtain probabilistic uniform convergence of model estimates to the minimum possible radius of uncertainty. Our formulation lends itself to convex analysis leading to description of optimal algorithms, which turn out to be well-known instrument-variable methods for many of the problems. We have characterized conditions on inputs in terms of second-order sample path properties required to achieve the minimum radius of uncertainty. We have derived optimal algorithms for system identification for a wide variety of standard as well as non-standard problems that include special structures such as unmodeled dynamics, positive real conditions, bounded sets and linear fractional maps.

VIII. APPENDIX

(Proof of Proposition 1) We deal with the case when $\gamma_0 = 0$ for simplicity. The argument for arbitrary γ_0 proceeds in a similar manner. Let Y_n denote the column vector of the output signal of length n and $\mathcal{D}(Y_n)$ be the set of all $h \in \mathcal{S}$ that are consistent with the input-output data and the noise set, i.e.,

$$\mathcal{D}(Y_n) = \{h \in \mathcal{S} : y(k) = \lambda_k(h) + w(k); w \in \mathcal{W}_n, 1 \leq k \leq n\}$$

The local diameter of uncertainty is given by:

$$\text{diam}(Y_n) = \sup_{h_1, h_2 \in \mathcal{D}(Y_n)} |\lambda_0(h_1 - h_2)|$$

From [22] it follows that the estimation error for any estimator is bounded from below by one-half of the diameter of uncertainty. By hypothesis of our proposition and the preceding argument there exists a noise set, \mathcal{W}_n^0 , such that the diameter of uncertainty is smaller than 2ϵ . We are now left to produce a convex and balanced set that has the same properties. To this end, consider the following noise set:

$$\mathcal{W}_n^* = \{w \in \mathbb{R}^n \mid \lambda_k(h) + w = 0, |\lambda_0(h)| \leq \epsilon, 1 \leq k \leq n\}$$

On account of the fact that the measurements are linear and the constraints are convex and balanced, it follows that the set, \mathcal{W}_n^* is also convex and balanced. We are left to establish two facts:

- (A) The noise set \mathcal{W}_n^* is feasible, i.e., the global diameter of uncertainty over all feasible outputs Y_n is bounded by 2ϵ .
- (B) The probability measure of the noise set \mathcal{W}_n^* is larger than the feasible set, \mathcal{W}_n^0 .

These two facts together will establish the proposition since we have then found an alternative convex and balanced set to \mathcal{W}_n^0 that has a larger probability measure and has the same worst-case global diameter of uncertainty.

To establish (A) we observe from [22] that with a linear/convex measurement structure, the worst-case diameter of uncertainty is realized when the output is identically zero. Consequently, for all other output realizations the worst-case uncertainty can only be smaller. Now, by construction the set \mathcal{W}_n^* corresponds to the zero output signal and therefore by the preceding argument is feasible.

(B) is established by the following sequence of inequalities:

$$\begin{aligned} \text{Prob}\{\mathcal{W}_n^0\} &\leq \sup_{\mathcal{W}_n} \left\{ \text{Prob}\{\mathcal{W}_n\} \mid \sup_{h \in \mathcal{S}, w \in \mathcal{W}_n} \text{diam}\{Y_n\} \leq 2\epsilon \right\} \\ &\stackrel{(a)}{\leq} \sup_{\mathcal{W}_n} \left\{ \text{Prob}\{\mathcal{W}_n\} \mid \sup_{h=0, w \in \mathcal{W}_n} \text{diam}\{Y_n\} \leq 2\epsilon \right\} \\ &\stackrel{(b)}{\leq} \sup_{\mathcal{W}_n, 0 \in \mathcal{W}_n} \left\{ \text{Prob}\{\mathcal{W}_n\} \mid \sup_{h=0, w \in \mathcal{W}_n} \text{diam}\{Y_n\} \leq 2\epsilon \right\} \\ &\stackrel{(c)}{\leq} \sup_{\mathcal{W}_n, 0 \in \mathcal{W}_n} \left\{ \text{Prob}\{\mathcal{W}_n\} \mid \sup_{h=0, w \in \mathcal{W}_n} \text{diam}\{Y_n = 0\} \leq 2\epsilon \right\} \\ &\stackrel{(d)}{\leq} \text{Prob}\{\mathcal{W}_n^*\} \end{aligned}$$

Inequality (a) follows from the fact that the output Y_n in the RHS is restricted to the case when $h = 0$, which

amounts to a relaxation of the constraint. Inequality (c) follows from a similar argument. Inequality (b) follows from the fact that the worst-case diameter of uncertainty is invariant under translations of noise set \mathcal{W}_n . However, the probability measure of the noise set can change. Using the underlying symmetry and monotonicity of the noise distribution around zero, translation of the feasible set \mathcal{W}_n so that it contains the zero element leads to increasing the probability measure. Finally, (d) follows by definition of \mathcal{W}_n^* since it includes all possible noise elements for which the output is identically zero and the diameter of uncertainty is constrained to be bounded by 2ϵ . Consequently, no other noise set containing zero element can have a larger probability measure. ■

(Proof of Corollary 1) Consider an element $z \in \mathcal{H}_2$ with $\langle \cdot, \cdot \rangle$ denoting the inner product. Let $z_0 = z/\|z\|_2$ be a unit vector aligned with z ; z_α be any other unit vector. It follows that,

$$\|z\|_2 = \langle z_0, z \rangle \leq |\langle z_0 - z_\alpha, z \rangle| + |\langle z_\alpha, z \rangle| \leq \|z_0 - z_\alpha\|_2 \|z\|_2 + |\langle z_\alpha, z \rangle|$$

where the first expression in the last inequality follows cauchy-schwartz inequality. Denoting $v_\alpha(z) = \langle z_\alpha, z \rangle$ and letting α take on values in a finite index set \mathcal{I} such that there is at least an α for which $\|z_0 - z_\alpha\|_2 < 1$ we have

$$\|z\|_2 \leq \min_{\|z_0 - z_\alpha\|_2 < 1} \frac{|v_\alpha(z)|}{1 - \|z_0 - z_\alpha\|_2}$$

To relate it to Corollary 1 we let, $z = x_1 - x_2$, where x_1, x_2 are any two feasible elements satisfying Equation 7. It follows that,

$$\|z\|_2 = \|x_1 - x_2\|_2 \leq \min_{\|z_0 - z_\alpha\|_2 < 1} \frac{|v_\alpha(x_1 - x_2)|}{1 - \|z_0 - z_\alpha\|_2}$$

The result now follows by direct substitution. ■

REFERENCES

- [1] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., 1995.
- [2] V. Borkar, S. Mitter, and S. Venkatesh. Variations on a neyman pearson theme. *Journal of the Indian Statistical Institute*, 66:292–305, 2004.
- [3] P. Caines. *Linear Stochastic Systems*. New York, Wiley, 1988.
- [4] M. C. Campi. Exponentially weighted least squares identification of time-varying systems with white disturbances. *IEEE Trans. on Signal Processing*, 42:2906–2914, 1994.
- [5] M. C. Campi and E. Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47:1329–1333, 2002.
- [6] M. A. Dahleh, T. V. Theodosopoulos, and J. N. Tsitsiklis. The sample complexity of worst case identification of linear systems. *System and Control Letters*, 20:157–166, 1993.
- [7] A. Garulli and W. Reinelt. On model error modeling in set membership identification. In *Proceedings IFAC Symposium on System Identification, Santa Barbara (USA)*, pages WeMD1–3, 2000.
- [8] L. Giarre, B. Z. Kacewicz, and M. Milanese. Model quality evaluation in set-membership identification. *Automatica*, 33:1133–1139, 1997.
- [9] G. C. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model-order selection. *IEEE Transactions on Automatic Control*, 37:1343–1354, 1992.
- [10] G. Gu and P. P. Khargonekar. Linear and nonlinear algorithms for identification in h with error bounds. *IEEE Transactions on Automatic Control*, 37:953–963, 1992.

- [11] P. Van Den Hof, P. Heuberger, and J. Bokor. Identification with generalized orthonormal basis functions—statistical analysis and error bounds. *Special Topics in Identification Modelling and Control*, pages 39–48, 1993.
- [12] E. L. Lehmann. *Testing Statistical Hypotheses*. 2nd edn. Springer Verlag: New York, 1997.
- [13] L. Ljung. *System Identification: Theory for the user*. Prentice Hall Englewood Cliffs, NJ, 1987.
- [14] L. Ljung and Z. D. Yuan. Asymptotic properties of the least squares method for estimating transfer functions. *IEEE Transactions on Automatic Control*, pages 514–530, 1985.
- [15] D. Luenberger. *Optimization by vector space methods*. John Wiley and Sons, Inc, 1969.
- [16] P. M. Makila. Robust identification and galois sequences. *International Journal of Control*, 54:1189–1200, 1991.
- [17] M. Milanese and G. Belforte. Estimation theory and uncertainty intervals evaluation in the presence of unknown but bounded errors: Linear families of models and estimators. *IEEE Transactions on Automatic Control*, 27:408–414, 1982.
- [18] M. Milanese and A. Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27:997–1009, 1991.
- [19] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization*. Prentice Hall Englewood Cliffs, NJ, 1997.
- [20] P. Poolla and A. Tikku. On the time complexity of worst-case system identification. *IEEE Transactions on Automatic Control*, 39:944–50, 1994.
- [21] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill International Editions, 1976.
- [22] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski. *Information, uncertainty, complexity*. Addison-Wesley Pub. Co., 1983.
- [23] D. N. C. Tse, M. A. Dahleh, and J. N. Tsitsiklis. Optimal identification under bounded disturbances. *IEEE Transactions on A-C*, 38:1176–90, 1993.
- [24] V. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [25] S. Venkatesh. Necessary conditions for identification of uncertain systems. *Systems and Control Letters*, 53:117–125, 2004.
- [26] S. Venkatesh and M. A. Dahleh. System identification for complex-systems: Problem formulation and results. In *Proceedings of the 36th IEEE Conf. on Decision and Control, San Diego, CA*, 1997.
- [27] S. Venkatesh and M. A. Dahleh. On system-identification of complex-processes with finite data. *IEEE Transactions on Automatic Control*, 46:235–257, 2001.
- [28] S. Venkatesh and A. M. Finn. Cabin communication system without acoustic echo cancellation. *US Patent #6748086*, 2004.
- [29] S. Venkatesh and S. K. Mitter. Statistical modeling and estimation with limited data. In *International Symposium on Information Theory, Lausanne, Switzerland*, 2002.
- [30] S. R. Venkatesh and M. A. Dahleh. Identification in the presence of unmodeled dynamics and noise. *IEEE Transactions on Automatic Control*, 42:1620–1635, 1997.
- [31] A. Vicino and G. Zappa. Sequential approximation of feasible parameter sets for identification with set membership identification. *IEEE Transactions on Automatic Control*, 41:774–785, 1996.
- [32] M. Vidyasagar. *A Theory of Learning and Generalization*. New York: Springer-Verlag, 1997.
- [33] B. Wahlberg. System identification using laguerre models. *IEEE Transactions on Automatic Control*, pages 551–562, 1991.
- [34] B. Wahlberg and L. Ljung. Hard frequency-domain model error bounds from least-squares like identification techniques. *IEEE Transactions on Automatic Control*, 37:900–912, 1992.
- [35] L. Y. Wang. Persistent identification of time-varying systems. *IEEE Transactions on Automatic Control*, 42:66–82, 1997.
- [36] L. Y. Wang and G. G. Yin. Persistent identification of systems with unmodeled dynamics and exogenous disturbances. *IEEE Transactions on Automatic Control*, 45:1246–1256, 2000.