15th Triennial World Congress, Barcelona, Spain

IDENTIFICATION OF RELIABLE PREDICTOR MODELS FOR UNKNOWN SYSTEMS: A DATA-CONSISTENCY APPROACH BASED ON LEARNING THEORY

Giuseppe Calafiore [*,1] **M.C. Campi** [**] **Laurent El Ghaoui** [***]

[*] *Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy. e-mail: calafiore@polito.it*
[**] *Dipartimento di Elettronica per l'Automazione, Università di Brescia, Italy. e-mail: campi@ing.unibs.it*
[***] *Department of Electrical Engineering and Computer Science, UC Berkeley, CA. e-mail: elghaoui@eecs.berkeley.edu*

Abstract: In this paper we present preliminary results for a new framework in identification of predictor models for unknown systems, which builds on recent developments of statistical learning theory. The three key elements of our approach are: the unknown mechanism that generates the observed data (referred to as the remote data generation mechanism – DGM), a selected *family of models*, with which we want to describe the observed data (the data descriptor model – DDM), and a *consistency criterion*, which serves to assess whether a given observation is compatible with the selected model. The identification procedure will then select a model within the assumed family, according to some given optimality objective (for instance, accurate prediction), and which is consistent with the observations. To the optimal model, we attach a *certificate* of reliability, that is a statement of probability that the computed model will be consistent with future unknown data.

Keywords: Identification, Set-valued maps, VC theory, Convex optimization.

1. INTRODUCTION

Dynamical models of systems and time series are constructed for different purposes. Among others, for the analysis of the system properties, for predicting future output values, and for designing suitable controllers to feedback-connect to the system, see the classical references (Box *et al.*, 1994), (Ljung, 1999).

A data descriptor model is intended as an analytic model which is able to explain the observed data, and which has additional desirable features, such as good prediction capabilities. In this paper we are mainly concerned with data descriptor models to be used for prediction. A model of this kind can be used in all applications where forming a reliable prediction of the system output is important. Typical examples are the forecast of future events, in which case the model is directly used to the final prediction purpose, and construction of predictive controllers, where the model is used as an instrument to foresee the future system output behavior in dependence of the selected input.

A good predictor model should always return a prediction value along with a statement on the reliability of such a prediction. Examples of such statements are: the future system output is spread around the prediction value according to a certain probabilistic distribution; or: the future system output belongs to a certain interval centered in the prediction value with given (high) probability. Without a reliability statement, a prediction is of little use. Thus, the following question about any identification approach to prediction models arises naturally: What can we say about the reliability of the estimated model? That is, can

[1] G. Calafiore acknowledges funding from the CNR Agenzia 2000 program.

we quantify with precision the probability that the future output will belong to the interval given by the model? This question would be easily answered if we could assume that the true data generation system has a given particular structure. However, assuming that we know the structure of the data generation system is often unrealistic.

The goal of the present paper is to introduce a new approach for the construction of predictor models. Instead of insisting to follow a standard identification route where one first constructs a parametric model by minimizing an identification cost, and then uses the model to work out the prediction interval, we directly consider interval models (that is, models returning an interval as output) and use data to ascertain the reliability of such models. In this way, the procedure for selecting the model is directly tailored to the final purpose for which the model is being constructed. We gain two fundamental advantages over the standard identification approach:

i) The reliability of the estimation can be quantified independently of the data generation mechanism. In other words, (under certain hypotheses to be discussed later) we are able to attach to a model a label certifying its reliability, whatever the true system is.

ii) The model structure selection can be performed by directly optimizing over the final result. Precisely, for a pre-specified level of reliability, we can choose the model structure that gives the smallest prediction interval.

The results of the present paper have been made possible by some recent developments in the statistical learning literature. In particular, our results build on Learning Theory results of (Vapnik and Chervonenkis, 1971) and (Vidyasagar, 1997), and on previous works in which concepts from learning theory have been applied to the field of system identification, see for instance (Campi and Kumar, 1998) and (Weyer, 2000). The purpose of the paper is to provide an introductory account of the theory of data-consistent models. The results are preliminary. In particular, the finite-sample reliability results given in Section 3 rest on the assumption that the DGM generates i.i.d. sequences. This assumption is not satisfied by generic dynamical systems possessing memory. The theory can however be extended to non i.i.d. processes (in particular to mixing processes), following the ideas introduced by (Nobel and Dembo, 1993) and (Yu, 1994). This extension goes beyond the scope of the present paper, and it is subject of ongoing research.

2. DATA-CONSISTENT MODELS

Let $y(k) \in \mathbb{R}$ be the observed output of an unknown system, and let $\varphi(k) \in \mathbb{R}^n$ be an explanatory (or regression) vector constructed from past inputs and outputs. For the time being, we shall make no assumptions on the nature or structure of the unknown system, and refer to it as the remote *data generation mechanism* (DGM).

Assume we observe one realization of the unknown process over a finite time window $k = 1, \dots, N$, and collect the observations in the data sequence $D_N \doteq \{y(k), \varphi(k)\}_{k=1,\dots,N}$. Then, we seek to explain the observed data using a *data descriptor model* (DDM).

Our standpoint in this paper is that a DDM is a rule that assigns to each regression vector $\varphi(k)$ a certain value interval for the corresponding output. That is, a DDM is a set-valued map

$$\mathcal{I}(\cdot) : \varphi(k) \to \mathcal{I}(\varphi(k)) \subset \mathbb{R}.$$

For algorithmic reasons, in the sequel we will often consider DDM that are described in parametric form as follows. First, a system class $\mathcal{M}$ is considered (for instance a linear, auto-regressive class), such that the output of a system in the class is expressed as $\eta(k) = \mathcal{M}(\varphi(k), q)$, for some parameter $q \in \mathcal{Q} \subset \mathbb{R}^{n_q}$. A DDM is then obtained by selecting a particular feasible set $\mathcal{Q}$, and considering all possible outputs obtained for $q \in \mathcal{Q}$, i.e. the DDM is defined through the relation

$$\mathcal{I}(\varphi(k)) \doteq \{\eta : \ \eta = \mathcal{M}(\varphi(k), q), \ q \in \mathcal{Q}\}. \quad (1)$$

In this case, the DDM is also indicated by $\mathcal{M}_{\mathcal{Q}}$ and the corresponding output interval is $\mathcal{M}_{\mathcal{Q}}(\varphi(k))$.

A comprehensive theory on set-valued dynamical models in continuous time has been developed in (Aubin and Cellina, 1984), (Aubin, 1990). Here, we consider the problem of identification of set-valued maps in discrete time.

To the interval model introduced above, we associate a *consistency condition* which assesses whether a given observation is in agreement with the assumed model. We introduce the following definition.

De£nition 1. A DDM is *consistent* with the observed data sequence D_N if, for all $k = 1, \dots, N$, the following data-consistency condition holds

$$y(k) \in \mathcal{I}(\varphi(k)). \quad (2)$$

This means that the assumed model is not falsified by the observed data. In particular, for DDM described as in (1), this means that there exists a feasible sequence $\{q(k) \in \mathcal{Q}\}_{k=1,\dots,N}$ that satisfies the model equations, i.e. $y(k) = \mathcal{M}(\varphi(k), q(k))$, for $k = 1, \dots, N$.

2.1 Common model structures

The abstract model description introduced above specializes to some well-known model structures used in system identification. Linear AR(n) models with bounded noise

$$y(k) = \varphi^T(k)\theta + e(k), \ \ |e(k)| \le \gamma,$$

with $\varphi(k) \doteq [y(k-1) \cdots y(k-n)]^T$, are accommodated into the previous structure, taking $q = [\theta^T, e]^T \in \mathbb{R}^{n+1}$, and $\mathcal{Q} = \{q : q[1:n] = \theta, q[n+1] = e \in [-\gamma, \gamma]\} = \theta \times [-\gamma, \gamma]$. ARX($p, m$)

models can be accommodated similarly, considering $\varphi(k) \doteq [y(k-1)\cdots y(k-p)u(k-1)\cdots u(k-m)]^T$.

More interestingly, we can consider ARX model structures where variability is present in both an additive and multiplicative fashion

$$y(k) = \varphi^T(k)\theta(k) + e(k), \quad |e(k)| \leq \gamma. \quad (3)$$

Here, the regression parameter is considered to be time-varying, i.e. $\theta(k) \in \Delta \subseteq \mathbb{R}^n$, where Δ is some assigned bounded set. This model structure is obtained from the general one by setting $q = [\theta,\, e] \in \mathbb{R}^{n+1}$, and $\mathcal{Q} = \{q : q[1:n] = \theta \in \Delta, q[n+1] = e \in [-\gamma, \gamma]\} = \Delta \times [-\gamma, \gamma]$. The computation of the output interval for these model structures is reported in Section 4.1, for the case when Δ is a sphere or an ellipsoid.

One thing that needs to be made clear at this point is that models like (3) are not intended to be a parametric representation of the true system. In particular, $\theta(k)$ has not to be interpreted as an estimate of a true time-varying parameter. It is merely an instrument through which we defined a map $\mathcal{I}(\cdot)$ that assigns to each $\varphi(k)$ an interval $\mathcal{I}(\varphi(k))$, and this map is used for prediction.

2.2 Optimal data-consistent models

Consider now a family of interval models $\mathcal{M}_{\mathcal{Q}}$, where $\mathcal{Q}$ can be any member of a specified family of sets. Among all members of this interval model family which are consistent with the observed data, we are interested in those models providing the most "informative" (i.e. the smallest) output prediction map $\mathcal{I}(\cdot)$. Assume a data descriptor model structure is given, and let $\mu_{\mathcal{Q}}$ be a scalar parameter that defines the "size" of the output map associated with the model identified by $\mathcal{Q}$. Clearly, the choice of a suitable size measure $\mu_{\mathcal{Q}}$ depends on (and is suggested by) the specific problem at hand. Then, we have the following definition.

De£nition 2. An optimal data-consistent model (ODCM) $\mathcal{M}_{Q^*}$ is such that

$$\mathcal{M}_{Q^*} = \text{arg min } \mu_Q \text{ subject to } y(k) \in \mathcal{M}_{\mathcal{Q}}(\varphi(k)), \text{ for } k = 1, \ldots, N.$$

Two fundamental issues remain now to be discussed. The first one concerns the reliability properties of models constructed using the data consistency approach. In particular, we can ask how large the probability is that a new unseen datum will be consistent with the model. The second issue pertains to the algorithmic construction of optimal data consistent models. The first issue is discussed in Section 3, while Section 4 is concerned with the second issue. There, we in particular develop efficient polynomial-time algorithms for computing the ODCM for the ARX structure (3). Many other issues remain open, however, and will be discussed in the concluding Section 5.

3. A LEARNING THEORY APPROACH TO MODEL RELIABILITY

In this section, we tackle the fundamental issue of assessing the *reliability* of a data-consistent model, with respect to its ability to predict the future behavior of the unknown system. To keep the exposition as clear as possible, we develop our analysis assuming that the sequence $\{y(k), \varphi(k)\}$ is an i.i.d. sequence generated by a stationary remote process. The stationarity of the process basically means that the DGM is operating in steady state. As for the i.i.d. assumption, it is indeed a strong one in the context of system identification. However, this hypothesis is not critical for our developments, and can be relaxed by resorting to mixing processes.

Let a family of interval models $\mathcal{M}_{\mathcal{Q}}$ be given, where the set $\mathcal{Q}$ can be any element of a specified family of sets, and let $x(k) \doteq [\varphi(k)^T\ y(k)]^T \in X \subseteq \mathbb{R}^{n+1}$, for all k. For a fixed model in the class (i.e. fixed $\mathcal{Q}$), we introduce the consistency function $h(x(k), \mathcal{Q})$ such that

$$h(x(k), \mathcal{Q}) \doteq \begin{cases} 1, \text{ if } y(k) \in \mathcal{M}_{\mathcal{Q}}(\varphi(k)) \\ 0, \text{ otherwise.} \end{cases} \quad (4)$$

For each $\mathcal{Q}$, let us define the set $A_{\mathcal{Q}}$ of all $x(k)$ which are consistent with the assumed model

$$A_{\mathcal{Q}} \doteq \{x \in X : h(x, \mathcal{Q}) = 1\}, \quad (5)$$

and the collection $\mathcal{A}$ of subsets of X, $\mathcal{A} \doteq \{A_{\mathcal{Q}}\}$, where $\mathcal{Q}$ ranges over its specified feasible family of sets. Let further P be some unknown probability measure on X, and suppose that, for a fixed $\mathcal{Q}$, it is desired to compute an empirical estimate of $P(A_{\mathcal{Q}})$. To this end, we collect N i.i.d. samples of $x(k) \in X$ (drawn according to the density P) in the multisample $\mathbf{x} \doteq [x^T(1)\, x^T(2) \cdots x^T(N)]^T \in X^N$, and define the *empirical probability* of $A_{\mathcal{Q}}$ as

$$\hat{P}(A_{\mathcal{Q}}; \mathbf{x}) \doteq \frac{1}{N} \sum_{j=1}^{N} I_{A_{\mathcal{Q}}}(x(j)), \quad (6)$$

where $I_{A_{\mathcal{Q}}}$ is the *indicator function* of the set $A_{\mathcal{Q}}$, i.e.

$$I_{A_{\mathcal{Q}}}(x(k)) \doteq \begin{cases} 1, \text{ if } x(k) \in A_{\mathcal{Q}} \\ 0, \text{ otherwise.} \end{cases}$$

The following theorem is a direct application of a fundamental result of learning theory (see (Vidyasagar, 1997)), and provides an assessment of the reliability of the empirical estimate $\hat{P}(A_{\mathcal{Q}}; \mathbf{x})$, for *any* $A_{\mathcal{Q}} \in \mathcal{A}$.

Theorem 1. (Vapnik and Chervonenkis, 1971). Let all symbols be defined as above, and let

$$p(N, \epsilon, \mathcal{A}) \doteq \text{Prob}\{\mathbf{x} \in X^N : \exists A_{\mathcal{Q}} \in \mathcal{A} \text{ such that } |\hat{P}(A_{\mathcal{Q}}; \mathbf{x}) - P(A)| > \epsilon\}.$$

If the collection of sets $\mathcal{A}$ has finite Vapnik-Chervonenkis dimension, say VC-dim $(\mathcal{A}) \leq d$, then, for $N \geq d$, $\epsilon > 0$

$$p(N, \epsilon, \mathcal{A}) \leq 4 \left(\frac{2eN}{d}\right)^d \exp(-N\epsilon^2/8), \quad (7)$$

for all probability measures P.

Interpreted in our model-consistency setting, the above theorem means the following: given any model in the considered family (i.e. given any $\mathcal{Q}$), and given a data multisample $\mathbf{x} \in X^N$, the empirical probability of consistency converges (for increasing N) to the true probability, irrespective of the underlying distribution P, i.e. the underlying DGM. In this case, we say that the collection of sets $\mathcal{A}$ has the property of distribution-free uniform convergence of empirical probabilities (UCEP), (Vidyasagar, 1997).

It is evident form the above result that it is of paramount importance to assess wether a given family $\mathcal{A}$ has finite VC-dimension, and, if this is the case, to determine an upper bound d on this quantity. In this paper, we will study in particular the case (relevant in the identification applications discussed in the sequel) when the consistency condition (2) can be expressed in terms of the satisfaction of t polynomial inequalities, i.e.

$$h(x, \mathcal{Q}) = \begin{cases} 1, \text{ if } \tau_1(x,\omega) > 0, \ldots, \tau_t(x,\omega) > 0 \\ 0, \text{ otherwise,} \end{cases} \quad (8)$$

where $\tau_1(x,\omega), \ldots, \tau_t(x,\omega)$ are polynomials in x, ω, of maximum degree g, and where $\omega \in \mathbb{R}^{n_\omega}$ represents a vector of variables needed to describe the consistency conditions by means of polynomial inequalities. In this case a result of (Karpinsky and Macintyre, 1997), proved in the refined form below in (Vidyasagar, 1997) Chapter 10, provides an upper bound for the VC-dimension of $\mathcal{A}$ (see also the introductory exposition from (Sontag, 2000) for general results on the computation of the VC-dimension of classes of concepts).

Theorem 2. With all symbols defined as above, and being e the Neper constant, we have

$$\text{VC-dim}\,(\mathcal{A}) \leq d, \text{ with } d = 2n_\omega \log_2(4egt). \quad (9)$$

With these premises, we now state a first key result on the reliability of optimal data-consistent models.

Theorem 3. Let $\{y(k), \varphi(k)\}$ be an i.i.d. sequence generated by a stationary remote process (our DGM), and let $D_N \doteq \{y(k), \varphi(k)\}_{k=1,\ldots,N}$ be the sequence of data observed over the time window $k = 1, \ldots, N$. Consider a data descriptor model (DDM) for which the consistency condition can be checked via feasibility of a set of t polynomial inequalities

$$\tau_1(x,\omega) > 0, \ldots, \tau_t(x,\omega) > 0,$$

where $\tau_1(x,\omega), \ldots, \tau_t(x,\omega)$ are a polynomials in x, ω of maximum degree g. Assume an ODCM $\mathcal{M}_\mathcal{Q}$ is constructed based on the observation sequence D_N. Define the *reliability* $R(\mathcal{M}_\mathcal{Q})$ of the ODCM as the probability of consistency of the model with "future" unknown data

$$R(\mathcal{M}_\mathcal{Q}) \doteq \text{Prob}\{\mathbf{x}(k) = [\varphi^T(k)\ y(k)]^T : h(x(k), \mathcal{Q}) = 1, \text{ for } k > N\}. \quad (10)$$

For any fixed *accuracy* $0 < \epsilon < 1$ it may be asserted with *confidence* $1 - \delta(N, \epsilon, d)$ that

$$R(\mathcal{M}_\mathcal{Q}) > 1 - \epsilon, \quad (11)$$

being

$$\delta(N, \epsilon, d) = 4 \left(\frac{2eN}{d}\right)^d \exp(-N\epsilon^2/8), \quad (12)$$

$$d = 2n_\omega \log_2(4egt). \quad (13)$$

The proof of this theorem is an immediate consequence of Theorem 1 and Theorem 2.

4. ODCM FOR ARX STRUCTURES WITH PARAMETRIC AND ADDITIVE NOISE

In this section, we discuss in detail the construction and reliability analysis for a class of ARX model structures with additive and multiplicative time-varyability. We assume the following descriptor model

$$y(k) = \varphi^T(k)\theta(k) + e(k), \quad (14)$$

where the regression parameter is considered to be time-varying, i.e. $\theta(k) \in \Delta$, being Δ a sphere with center θ and radius r, i.e.

$$\Delta \doteq \{\xi : \xi = \theta + \delta,\ \|\delta\| \leq r\}. \quad (15)$$

Subsequently, we will also consider the more general case when the bounding set for the parameters is an ellipsoid. The additive noise $e(k)$ is an unknown-but-bounded sequence, i.e. $|e(k)| \leq \gamma$, where $\gamma \geq 0$ is a model parameter. Thus, the parameters describing the set $\mathcal{Q}$ are the center θ and radius r of Δ, and the magnitude bound γ on the additive term $e(k)$. Models of this type have already been considered by the authors in (El Ghaoui and Calafiore, 2000).

In the following section we present efficient algorithms for computing the ODCM in this setting, and provide explicit results for their reliability.

4.1 Computing the ODCM for time-varying ARX structures

The objective is to determine the parameters $\theta \in \mathbb{R}^n$, $r, \gamma > 0$ of an ARX model, that minimize a "size" measure μ_Q, under the model consistency constraints. Here, we take $\mu_Q = \gamma + \alpha r$, $\alpha \geq 0$. Note that, if $\alpha = E[\|\varphi(k)\|]$, μ_Q measures the average amplitude of the output interval. The optimal model can be computed by means of Linear Programming, as detailed in the following theorem.

Theorem 4. (Spherical parameter set). Given an observed sequence $D_N = \{y(k), \varphi(k)\}$, and a "size" objective $\mu_Q = \gamma + \alpha r$, where α is a fixed non-negative number, an optimal consistent ARX model is computed solving the following linear programming problem in the variables θ, r, γ

$$\text{minimize } \gamma + \alpha r, \text{ subject to:} \quad (16)$$

$$r, \gamma \geq 0 \quad (17)$$

$$\varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq y(k) \quad (18)$$

$$-\varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq -y(k) \quad (19)$$

$$k = 1, \ldots, N.$$

Proof. For each k, the model equation (14) with the additive noise bound $|e(k)| \leq \gamma$ defines a slab of allowable parameters $\theta(k)$

$$\{\theta(k) : \ |y(k) - \varphi^T(k)\theta(k)| \leq \gamma\}. \qquad (20)$$

In turn, the parameter $\theta(k)$ is bound in the sphere $\theta(k) = \theta + \delta(k), \ \|\delta(k)\| \leq r$, therefore, the k-th observation is consistent with the assumed model if and only if the slab intersects the sphere. Geometrically, for fixed γ, the problem amounts to finding the sphere of minimum radius that intersects *all* the slabs (20), for $k = 1, \ldots, N$.

The intersection condition between the slab and the sphere can be expressed as

$$|y(k) - \varphi^T(k)\theta| \leq \gamma + r\|\varphi(k)\|, \qquad (21)$$

which are indeed linear constraints on the decision variables, and the statement of the theorem immediately follows. □

As a generalization of the previous theorem, we next consider ARX models where the parameter is bound in an ellipsoid, i.e.

$$\Delta = \{\xi : \xi \in \mathcal{E}(\theta, P)\}, \qquad (22)$$

where $\mathcal{E}(\theta, P)$ denotes the ellipsoid of center in θ and "shape matrix" $P \succeq 0$

$$\mathcal{E}(\theta, P) = \left\{ x : \ \begin{bmatrix} P & (x-\theta) \\ (x-\theta)^T & 1 \end{bmatrix} \succeq 0 \right\}.$$

In this case, computing the optimal model amounts to determine the parameters θ, P, γ describing $\mathcal{Q}$, such that the consistency conditions are fulfilled, and a "size" objective μ_Q is minimized. In particular, we choose $\mu_Q = \gamma + \text{Tr}\, PW$, being $W \succeq 0$ a weight matrix. We show in the next theorem that the optimal model in this case can be computed efficiently, solving a semidefinite programming problem, i.e. a convex optimization problem with linear objective and LMI constraints, see (Vandenberghe and Boyd, 1996) for details.

Theorem 5. (Ellipsoidal parameter set). Given an observed sequence $D_N = \{y(k), \varphi(k)\}$, a model order n, a weight matrix $W \succeq 0$, and parameter set described as in (22), an optimal consistent ARX model is computed solving the following semidefinite programming problem in the variables P, θ, γ, and in the slack variables ϵ_k

$$\text{minimize } \gamma + \text{Tr}\, PW, \text{subject to:} \qquad (23)$$

$$P \succeq 0, \ \gamma \geq 0 \qquad (24)$$

$$\begin{bmatrix} \varphi^T(k)P\varphi(k) & y(k) - \varphi^T(k)\theta - \epsilon_k \\ y(k) - \varphi^T(k)\theta - \epsilon_k & 1 \end{bmatrix} \succeq 0, \qquad (25)$$

$$\epsilon_k \leq \gamma, \ \epsilon_k \geq -\gamma, \qquad (26)$$

$$k = 1, \ldots, N.$$

Proof. The proof follows the same line as for Theorem 4: For each k, the model equation (14) with the additive noise bound $|e(k)| \leq \gamma$ defines a slab of allowable parameters $\theta(k)$

$$\{\theta(k) : \ |y(k) - \varphi^T(k)\theta(k)| \leq \gamma\}. \qquad (27)$$

In turn, the parameter $\theta(k)$ is bound in the ellipsoid $\mathcal{E}(\theta, P) = \{\theta(k) : \ \theta(k) = \theta + \delta(k), \ \delta(k) \in \mathcal{E}(0, P)\}$.

Therefore, the k-th observation is consistent with the assumed model if and only if the slab intersects the ellipsoid. Geometrically, for fixed γ, the problem amounts to finding the ellipsoid of minimum size (in the sense of the weighted measure Tr PW) that intersects *all* the slabs (27), for $k = 1, \ldots, N$.

The intersection condition between the slab and the ellipsoid can be expressed as

$$|y(k) - \varphi^T(k)\theta| \leq \gamma + \|\varphi^T(k)E\|,$$

where E is a symmetric matrix square root of P, i.e. $P = EE^T$. This condition is equivalent to: $\exists \epsilon_k$ such that $|\epsilon_k| \leq \gamma$, and $|y(k) - \varphi^T(k)\theta - \epsilon_k| \leq \|\varphi^T(k)E\|$. Taking the square of this condition we get

$$(y(k) - \varphi^T(k)\theta - \epsilon_k)^2 \leq \varphi^T(k)EE^T\varphi(k), \qquad (28)$$

and using Schur complements, we easily obtain the LMI constraints stated in the theorem. □

Once a descriptor model of the form (14) has been identified using the discussed approach, it is straightforward to obtain an interval of prediction. Since the parameter vector $\theta(k)$ lies in the ellipsoid $\mathcal{E}(\theta, P)$, and $|e(k)| \leq \gamma$, the possible values for the model output at time $N + 1$ lie in the interval $\mathcal{I}_{N+1} = [y^+(N + 1) \ y^-(N + 1)]$, where

$$y^{\pm}(N+1) = \varphi^T(N+1)\theta \pm \left((\varphi^T(N+1)P\varphi(N+1))^{1/2} + \gamma \right).$$

Of course, the actual output of the system is guaranteed to lie in the computed interval, up to the reliability and confidence of the identified model.

4.2 Reliability analysis for ARX ODCMs

The reliability of ARX optimal models under i.i.d. hypotheses on the data generation mechanism can be studied by direct application of Theorem 3. The results for the spherical and ellipsoidal noise cases are reported in the following corollaries.

Corollary 1. (Reliability of "spherical" ARX).
Let $\mathcal{M}_\mathcal{Q}$ be the optimal ARX model with spherical parameter noise, computed according to Theorem 4, using N observations. Then, for any $0 < \epsilon < 1$, it can be asserted with confidence greater than $1 - \delta(N, \epsilon, d)$ that the reliability $R(\mathcal{M}_\mathcal{Q})$ of the computed model is ϵ-close to one, being $\delta(N, \epsilon, d)$ defined in (12), and

$$d = 2(n + 2)\log_2(16e), \ N \geq d.$$

Proof. From Theorem 4, the consistency condition for this class of models may be checked by means of the four linear inequalities (17)–(19), in the vector of variables $\omega = [\theta, \ r, \ \gamma]$, therefore we have $t = 4$ (number of polynomial inequalities), $n_\omega = n + 2$ (number of variables entering the consistency inequalities) and $g = 1$ (maximum degree of the

polynomial inequalities). From Theorem 2 we then determine $d = 2(n+2)\log_2(16e)$. The statement of the theorem then follows from direct application of Theorem 3. □

Corollary 2. (Reliability of "ellipsoidal" ARX). Let $\mathcal{M}_{\mathcal{Q}}$ be the optimal ARX model with ellipsoidal parameter noise, computed according to Theorem 5, using N observations. Then, for any $0 < \epsilon < 1$, it can be asserted with confidence greater than $1-\delta(N,\epsilon,d)$ that the reliability $R(\mathcal{M}_{\mathcal{Q}})$ of the computed model is ϵ-close to one, being $\delta(N,\epsilon,d)$ defined in (12), and

$$d = (n^2+3n+4)\log_2(24e),\ N \geq d.$$

Proof. From Theorem 5 (and from the relative proof), the consistency condition for this class of models may be checked by means of the two quadratic inequalities (28), $\epsilon_k^2 \leq \gamma^2$, and the linear inequality $\gamma \geq 0$, Therefore the total number of polynomial inequalities is $t = 3$, and their maximum degree is $g = 2$. The variables ω involved in the consistency inequalities are the $n(n+1)/2$ independent entries of E, the n entries of θ, and γ, ϵ_k, therefore $n_\omega = n(n+3)/2 + 2$, and from Theorem 2 we determine $d = (n^2+3n+4)\log_2(24e)$. The statement of the theorem then follows from direct application of Theorem 3. □

5. DISCUSSION AND FUTURE DIRECTIONS

In this paper we presented preliminary results for the identification of predictive set-valued models for unknown systems. The result of identification is a set-valued map which associates to the regressor a predicted output interval. To the predicted interval, we attach a reliability statement guaranteeing that the actual future output will fall in the computed interval, with a certain (high) probability.

The rationale behind the proposed approach is to derive a *direct* procedure for going from data to prediction intervals, assuming as little as possible on the remote mechanism that generates the data. Learning Theory seems to be the most natural framework for this purpose.

Many directions of research are still open. In particular, the i.i.d. assumption on the DGM can be replaced with a weak dependence (β-mixing) assumption. Based on the results of (Nobel and Dembo, 1993) and (Yu, 1994), we can anticipate that the key ideas and results presented in this paper remain valid also for mixing processes, but the reliability bounds $\delta(N,\epsilon,d)$ should be recomputed accordingly.

Another aspect which is mentioned but not treated in this paper is model structure optimization. Given a certain level of reliability, and for fixed confidence, we need to study which model structure and order gives, for instance, the smallest prediction interval, for a certain number N of observations. Also, we remark that all the theory could be restated in terms of "partially consistent" models, i.e. models which are consistent only with a given *fraction* of the observations. Of course, this makes a lot of sense if we are to derive models that are insensitive to possible outliers in the data.

As a final comment, we point out that a general criticism of VC-dimension approaches is that the number of observations required to attain a reasonable level of reliability is usually very high. This situation is aggravated if we depart from the i.i.d. assumption and move to mixing processes. At the time of this writing, we are developing a different approach to the problem, which is not based on the VC-theory, and which seems to be very promising in the direction of achieving the desired level of reliability using a much smaller number of data.

REFERENCES

Aubin, J.P. (1990). *Set-valued analysis*. Birkhäuser. Boston, MA.

Aubin, J.P. and A. Cellina (1984). *Differential inclusions: set valued maps and viability theory*. Springer-Verlag. New York.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994). *Time series analysis: forecasting and control*. Prentice Hall. Englewood Cliffs, N.J.

Campi, M.C. and P.R. Kumar (1998). Learning dynamical systems in a stationary environment. *Sys. Control Letters* **34**, 125–132.

El Ghaoui, L. and G. Calafiore (2000). Recursive identification of models with time-varying, structured uncertainty. In: *IFAC SYSID Conference*. Santa Barbara, California.

Karpinsky, M. and A.J. Macintyre (1997). Polynomial bounds for VC dimension of sigmoidal general pfaffian neural networks. *Journal of Computer System and Science* **54**, 169–176.

Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall. Englewood Cliffs, N.J.

Nobel, A. and A. Dembo (1993). A note on uniform laws of averages for dependent processes. *Stat. and Prob. Lett.* **17**, 169–172.

Sontag, E. (2000). VC dimension of neural networks. Available on-line at the www address `citeseer.nj.nec.com/191558.html`.

Vandenberghe, L. and S. Boyd (1996). Semidefinite programming. *SIAM Rev.* **38**(1), 49–95.

Vapnik, V.N. and A.Y. Chervonenkis (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and its Applications* **16**(2), 264–280.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization*. Springer-Verlag. London.

Weyer, E. (2000). Finite sample properties of system identification of arx models under mixing conditions. *Automatica* **36**, 1291–1299.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability* **22**(1), 94–116.