PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON
DYNAMICAL SYSTEMS AND DIFFERENTIAL EQUATIONS
May 24 – 27, 2002, Wilmington, NC, USA pp. **1–10**

# A LEARNING THEORY APPROACH TO THE CONSTRUCTION OF PREDICTOR MODELS

G. Calafiore* and M.C. Campi**

*Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi 24 – 10129 Torino, Italy
**Dipartimento di Elettronica per l'Automazione
Università di Brescia
Via Branze 38 – 25123 Brescia, Italy

**Abstract.** This paper presents new results for the identification of predictive models for unknown dynamical systems. The three key elements of the proposed approach are: i) an unknown mechanism that generates the observed data; ii) a family of models, among which we select our predictor, on the basis of past observations; iii) an optimality criterion that we want to minimize. A major departure from standard identification theory is taken in that we consider interval models for prediction, that is models that return output *intervals*, as opposed to output *values*. Moreover, we introduce a consistency criterion (the model is required to be consistent with observations) which act as a constraint in the optimization procedure. In this framework, the model has not to be interpreted as a faithful description of reality, but rather as an instrument to perform prediction. To the optimal model, we attach a *certificate of reliability*, that is a statement of the probability that the computed model will actually be consistent with future unknown data.

1. **Introduction.** In the standard prediction-error setting for identification of dynamical models, [2], [7], a parametric model structure is first selected, and the parameters of the model are then estimated using an available batch of observations. The identified model may then be used to determine a predicted value for the output of the system, together with probabilistic intervals of confidence around the prediction. A crucial observation on this approach is that the interval of confidence determined as above may poorly describe the actual probability that the future output will fall in the computed interval, if the (unknown) system that generates the observations is structurally different from what it is assumed in the parametric model. In other words, the standard approach provides reliable predictions only if strong hypotheses on the structure and order of the mechanism that generates the data are satisfied. However, assuming that one knows the structure of the data generation system is often unrealistic. Therefore, the following question about any identification approach to predictive models arises naturally: what can we say about the reliability of the estimated model? That is, can we quantify with precision the probability that the future output will belong to the confidence interval given by the model?

In this paper, we follow a novel 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. First, 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; and, second, the model structure selection can be performed by

*Key words and phrases.* Predictive models, learning theory, convex optimization.

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 inspired by recent works in which concepts from learning theory have been applied to the field of system identification, see for instance [5] and [8].

The paper is organized as follows: Section 2 introduces the family of models under study. In Section 3 we present the computational results for the construction of interval models, using linear regression structures. In Section 4, we develop a method based on leave-one-out estimation techniques, to assess the reliability of the constructed model as a function of the finite observation size, under the assumption that the observations are independent and identically distributed (iid). These results are then extended in Section 5 to the case of weakly dependent observations.

2. **Interval predictors and data-consistency.** In this section, we introduce two key elements of our approach: models that return an interval as output (Interval Predictor Models) and the notion of consistency with observed data.

Let $\Phi \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}$ be given sets, denoted respectively as the *instance* set and the *outcome* set. An interval predictor model (IPM) is a rule that assigns to each instance vector $\varphi \in \Phi$ a corresponding output interval. That is, an IPM is a set-valued map

$$\mathcal{I} : \varphi \to \mathcal{I}(\varphi) \subseteq Y.$$

Interval models may be described in parametric form as follows. First, a model 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 $\xi = \mathcal{M}(\varphi, q)$, for some parameter $q \in \mathcal{Q} \subseteq \mathbb{R}^{n_q}$. An IPM 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 IPM is defined through the relation

$$\mathcal{I}(\varphi) \doteq \{\xi : \ \xi = \mathcal{M}(\varphi, q), \, q \in \mathcal{Q}\}. \tag{1}$$

In this case, the IPM is also indicated by $\mathcal{M}_{\mathcal{Q}}$, and the corresponding output interval is $\mathcal{M}_{\mathcal{Q}}(\varphi)$. In a dynamic setting, at each time instant the instance vector $\varphi$ may contain past values of input and output measurements, thus behaving as a regression vector. Standard auto regressive structures AR($n$)

$$\xi(k) = \varphi^T(k)\theta + e(k), \ \ |e(k)| \leq \gamma, \tag{2}$$

where $\varphi(k) \doteq [y(k-1) \cdots y(k-n)]^T$, give rise to (dynamic) IPMs by setting $q = [\theta^T \, e]^T \in \mathbb{R}^{n+1}$, and $\mathcal{Q} = \{q : \, q[1:n] = \theta, q[n+1] = e \in [-\gamma, \gamma]\} = \{\theta\} \times [-\gamma, \gamma]$. ARX($p, m$) structures can be used 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 structures[1] where variability is present in both an additive and multiplicative fashion

$$\xi(k) = \varphi^T(k)\theta(k) + e(k), \ \ |e(k)| \leq \gamma. \tag{3}$$

Here, the regression parameter is considered to be time-varying, i.e. $\theta(k) \in \Delta \subseteq \mathbb{R}^n$, where $\Delta$ is some assigned bounded set. In our exposition, we assume in particular $\Delta$ to be a sphere with center $\theta$ and radius $r$

$$\Delta \doteq \{\theta + \delta : \ \theta, \delta \in \mathbb{R}^n, \|\delta\| \leq r\}. \tag{4}$$

More generally, an ellipsoidal parameter set may be considered:

$$\Delta \doteq \{\zeta \in \mathbb{R}^n : (\zeta - \theta)^T P^{-1} (\zeta - \theta) \leq 1\}, \tag{5}$$

where $P \succ 0$ is a positive definite matrix.

For the model structure (3), (4), the parameters describing the set $\mathcal{Q}$ are the center $\theta$ and radius $r$ of $\Delta$, and the magnitude bound $\gamma$ on the additive term $e(k)$. Given $\varphi(k)$, the output of the model is the interval

$$\mathcal{I}(\varphi(k)) = [\varphi^T(k)\theta - (r\|\varphi(k)\| + \gamma), \ \varphi^T(k)\theta + (r\|\varphi(k)\| + \gamma)]. \tag{6}$$

For the ellipsoidal model (3), (5) we instead have the interval

$$\mathcal{I}(\varphi(k)) = [\varphi^T(k)\theta - (\sqrt{\varphi^T(k) P \varphi(k)} + \gamma), \ \varphi^T(k)\theta + (\sqrt{\varphi^T(k) P \varphi(k)} + \gamma)]. \tag{7}$$

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 a '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

[1]Notice that assuming a structure for constructing the predictive model does *not* mean that we are assuming that the actual mechanism that generates the data actually has this structure.

defined the interval map $\mathcal{I}$ that assigns to each $\varphi(k)$ an interval $\mathcal{I}(\varphi(k))$, and this map is used for prediction.

2.1. **Model consistency.** Assume now that one realization of an unknown bivariate stationary process $\{x(k)\} = \{\varphi(k), y(k)\}$, $\varphi(k) \in \mathbb{R}^n$, $y(k) \in \mathbb{R}$ is observed over a finite time window $k = 1, \dots, N$, and that the observations are collected in the data sequence $D_N \doteq \{\varphi(k), y(k)\}_{k=1,\dots,N}$. We have the following definition.

**Definition 1.** An interval model (1) is consistent with a given batch of observations $D_N$ if

$$y(k) \in \mathcal{I}(\varphi(k)), \text{ for } k = 1, \dots, N.$$

In other words, the above definition requires that the assumed model is not falsified by the observations. Notice that, for IPMs described as in (1), the consistency condition means that there exists a feasible sequence $\{q(k) \in \mathcal{Q}\}$ that satisfies the model equations, i.e. $y(k) = \mathcal{M}(\varphi(k), q(k))$, for $k = 1, \dots, N$.

Two fundamental issues need to be addressed at this point. The first one concerns the algorithmic construction of data consistent models. The second issue pertains to the reliability properties of the constructed models. In particular, we can ask how large the probability is that a new unseen datum will still be consistent with the model. The first issue is discussed in the following section, while Section 4 and Section 5 address the second one.

3. **Interval Predictors with linear structure.** Consider first the model structure (3), (4), and introduce a size measure $\mu_Q = \gamma + \alpha r$ for this interval map. Notice that, if we choose $\alpha = E[\|\varphi(k)\|]$, then $\mu_Q$ measures the average amplitude of the output interval. In this case, the optimal model that minimize $\mu_Q$ can be efficiently computed solving a Linear Programming problem. The following theorem holds (see [4] for a proof).

**Theorem 1** (Linear IPM - spherical parameter set)**.** *Given an observed sequence $D_N = \{\varphi(k), y(k)\}$, $k = 1, \dots, N$, a model order $n$, and a 'size' objective $\mu_Q = \gamma + \alpha r$, where $\alpha$ is a fixed non-negative number, an optimal consistent linear IPM is computed solving the following linear programming problem in the variables $\theta \in \mathbb{R}^n, r, \gamma$*

$$\begin{aligned} &\text{minimize } \gamma + \alpha r, \text{subject to:} \\ &\quad r, \gamma \geq 0 \\ &\quad \varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq y(k) \\ &\quad -\varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq -y(k) \\ &\quad k = 1, \dots, N. \end{aligned}$$

Similarly, for the model structure (3), (5), the optimal model can be efficiently computed solving a semidefinite (convex) optimization problem, as detailed in the following theorem.

**Theorem 2** (Linear IPM - ellipsoidal parameter set)**.** *Given an observed sequence $D_N = \{\varphi(k), y(k)\}$, $k = 1, \dots, N$, a model order $n$, and a 'size' objective $\mu_Q = \gamma + \text{Tr } PW$, where $W \succ 0$ is a given weight matrix, an optimal consistent linear IPM is computed solving the following semidefinite programming problem in the variables $P = P^T$, $\theta, \gamma$, and in the slack variables $\epsilon_k$*

$$\begin{aligned} &\text{minimize } \gamma + \text{Tr } PW, \text{subject to:} \\ &P \succ 0,\ \gamma \geq 0 \\ &\begin{bmatrix} \varphi^T(k)P\varphi(k) & y(k) - \varphi^T(k)\theta - \epsilon_k \\ y(k) - \varphi^T(k)\theta - \epsilon_k & 1 \end{bmatrix} \succeq 0, \\ &\quad \epsilon_k \leq \gamma,\ \epsilon_k \geq -\gamma, \\ &k = 1, \dots, N. \end{aligned}$$

4. **Reliability of IPMs for iid observations.** 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.

Suppose an optimal IPM of the form (3), (4) is determined using Theorem 1,[2] given a batch $D_N = \{x(k)\}_{k=1,\dots,N}$, $x(k) \doteq [\varphi^T(k)\ y(k)]^T$, of iid observations extracted according to an unknown probability measure $P$, and denote with $\hat{\mathcal{I}}_N$ the resulting optimal interval map.

[2]Notice that, in order to avoid repetitions, we discuss only the 'spherical' case in the sequel. Analogous results may be easily derived for the 'ellipsoidal' case as well.

**Definition 2.** The *reliability* $R(\hat{\mathcal{I}}_N)$ of the IPM $\hat{\mathcal{I}}_N$ is defined as the probability that a new unseen datum $x = [\varphi^T \; y]^T$ generated by the same process that produced $D_N$, is consistent with the computed model, i.e.

$$R(\hat{\mathcal{I}}_N) \doteq \text{Prob}_P\{y \in \hat{\mathcal{I}}_N(\varphi)\}. \tag{8}$$

The main result for iid observations is given in the following theorem.

**Theorem 3.** *Let $D_N = \{x(k) = [\varphi(k)^T \; y(k)]^T\}_{k=1,\ldots,N}$ be observations extracted from an iid sequence with unknown probability measure $P$, and let $\hat{\mathcal{I}}_N$ be the optimal interval map[3] computed according to Theorem 1. Then, for any $\epsilon, \delta > 0$ such that*

$$\epsilon\delta = \frac{n+2}{N+1} \tag{9}$$

*it holds that*

$$\text{Prob}_{P^N}\left\{R(\hat{\mathcal{I}}_N) \geq 1-\epsilon\right\} \geq 1-\delta. \tag{10}$$

**Proof.** Consider $N+1$ iid observations $D_{N+1} = \{z(1), \ldots, z(N+1)\}$, $z(k) \doteq [\psi^T(k) \; \eta(k)]^T$, $\psi(k) \in \mathbb{R}^n$, $\eta(k) \in \mathbb{R}$, extracted according to the unknown probability measure $P$. These are 'thought' (i.e. not actual) observations and serve the purpose of proving our result. Denote with $\hat{\mathcal{I}}_N^k$, $k = 1, \ldots, N+1$, the optimal interval map which is consistent with the $N$ observations

$$D_N^k \doteq \{z(1), \ldots, z(k-1), z(k+1), \ldots, z(N+1)\}.$$

Notice that $\hat{\mathcal{I}}_N^k$ is not necessarily consistent with the observation $z(k)$.

The idea of the proof is as follows: first we notice that $R(\hat{\mathcal{I}}_N)$ is a random variable belonging to the interval $[0, 1]$. Then, we show that the expected value of $R(\hat{\mathcal{I}}_N)$ is close to 1 and from this we infer a lower bound on the probability of having reliability not smaller than $1-\epsilon$. Define

$$\bar{R}_N \doteq E_{P^N}[R(\hat{\mathcal{I}}_N)],$$

where $E$ is the expectation operator, and, for $k = 1, \ldots, N+1$, let

$$v_k \doteq \begin{cases} 1, & \text{if } z(k) \text{ is consistent with } \hat{\mathcal{I}}_N^k \\ 0, & \text{otherwise,} \end{cases}$$

i.e. the random variable $v_k$ is equal to one, if $z(k)$ is consistent with the model obtained by means of the batch of the remaining observations $D_N^k$, and it is zero otherwise. Let also

$$\hat{\bar{R}}_N \doteq \frac{1}{N+1}\sum_{k=1}^{N+1} v_k. \tag{11}$$

We have that

$$\begin{aligned} E_{P^{N+1}}[v_k] &= E_{P^N}\left[E_P[v_k|D_N^k]\right] \\ &= E_{P^N}\left[\text{Prob}_P\{\eta(k) \in \hat{\mathcal{I}}_N^k(\psi(k))\}\right] = E_{P^N}[R(\hat{\mathcal{I}}_N^k)] = \bar{R}_N, \end{aligned}$$

which yields

$$E_{P^{N+1}}[\hat{\bar{R}}_N] = \bar{R}_N. \tag{12}$$

The key point is now to determine a lower bound for $E_{P^{N+1}}[\hat{\bar{R}}_N]$. We proceed as follows: consider one fixed realization $z(1), \ldots, z(N+1)$, and build the optimal map which is consistent with *all* of this observations, $\hat{\mathcal{I}}_{N+1}$. This map results from the solution of the convex optimization problem $\mathcal{P}$ in the variables $\theta \in \mathbb{R}^n, r, \gamma$

$$\begin{aligned} \mathcal{P}: \; &\text{minimize } \gamma + \alpha r, \text{subject to:} \\ &r, \gamma \geq 0 \\ &\psi^T(k)\theta - r\|\psi(k)\| - \gamma \leq \eta(k), \\ &-\psi^T(k)\theta - r\|\psi(k)\| - \gamma \leq -\eta(k), \\ &k = 1, \ldots, N+1. \end{aligned}$$

[3]We assume that all problems, when feasible, have a unique optimal solution. Should this not be the case, suitable tie-break rules could be used, as explained in [3].

The other optimal maps $\hat{\mathcal{I}}_N^k$ result from optimization problems $\mathcal{P}^k$, $k = 1, \ldots, N+1$ which are identical to $\mathcal{P}$, except that *one* single constraint relative to the $k$-th observation is removed in each problem. From Theorem 5 (in the Appendix) we know that at most $d = n+2$ of the observations when removed from $\mathcal{P}$ will change the optimal solution and improve the objective. Therefore, at least $N+1-d$ of the problems $\mathcal{P}^k$ are equivalent to $\mathcal{P}$. From this it follows that there exist at least $N+1-d$ optimal maps $\hat{\mathcal{I}}_N^k$, such that $z(k)$ is indeed consistent with $\hat{\mathcal{I}}_N^k$. Hence, at least $N+1-d$ of the $v_k$'s must be equal to one, and from (11) we have that

$$\hat{\bar{R}}_N \geq \frac{N+1-d}{N+1} = 1 - \frac{n+2}{N+1}, \text{ almost surely.}$$

Therefore, from (12) the expected value of the reliability is bounded as

$$\bar{R}_N = E_{P^{N+1}}[\hat{\bar{R}}_N] \geq 1 - \frac{n+2}{N+1}. \tag{13}$$

Now, given $\epsilon > 0$, we can bound the expectation $E_{P^N}[R(\hat{\mathcal{I}}_N)]$ from above as

$$E_{P^N}[R(\hat{\mathcal{I}}_N)] \leq (1-\epsilon)\text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) < 1-\epsilon\} + 1 \cdot \text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) \geq 1-\epsilon\}. \tag{14}$$

Letting $\bar{\delta} \doteq \text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) < 1-\epsilon\}$, combining the bounds (13), (14) we obtain that

$$\epsilon\bar{\delta} \leq \frac{n+2}{N+1},$$

from which the statement of the theorem immediately follows. □

5. **Reliability of IPMs for weakly dependent observations.** The results derived in the previous section for the iid case are now extended to $\beta$-mixing processes.

Let $\{x(k)\}_{k=-\infty}^{\infty}$ be a strict sense stationary process with distribution $P_{(-\infty,\infty)}$ and, given a set $I$ of integers, let $x_I$ denote $\{x(k)\}_{k\in I}$ and $P_I$ be the marginal distribution associated with $x_I$. We have the following definition.

**Definition 3** ($\beta$-mixing coefficients, $\beta$-mixing process, [1])**.** The $\beta$-mixing coefficients of $\{x(k)\}_{k=-\infty}^{\infty}$ are defined as:

$$\beta(T) \doteq \{|P_{(-\infty,\infty)}(C) - (P_{(-\infty,-1]\cup[1,\infty)} \times P_0)(C)|,\ C \in \sigma(x_{(-\infty,-T]}, x(0), x_{[T,\infty)})\}.$$

Process $\{x(k)\}_{k=-\infty}^{\infty}$ is $\beta$-mixing if $\beta(T) \to 0$ as $T \to \infty$.

If a process is formed by a sequence of independent random variables, then $P_{(-\infty,\infty)} = P_{(-\infty,-1]\cup[1,\infty)} \times P_0$, so that $\beta(T) = 0$, $\forall T$, and hence an independent process is a trivial example of a $\beta$-mixing process. In general, $\beta(T)$ is a measure of independence between events separated by a time lag $T$. $\beta$-mixing processes are often used to describe the correlation among data in presence of dynamics.

Definition 3 is a two-sided definition of $\beta$-mixing. More often, a one-sided definition is adopted where

$$\beta(T) \doteq \{|P_{(-\infty,\infty)}(C) - (P_{(-\infty,0]} \cup P_{[1,\infty)}(C)|,\ C \in \sigma(x_{(-\infty,0]}, x_{[T,\infty)})\}.$$

Here, we have preferred to adopt a two-sided definition since it is more handy in the present context and, as it can be verified, is not more restrictive than the one-sided definition (i.e. if $\beta(T) \to 0$ in the one-sided definition, this also occurs in the two-sided definition, though with different $\beta(T)$'s).

The key result for the reliability of optimal interval models constructed using dependent observations is contained in the following theorem.

**Theorem 4.** *Let $D_N = \{x(k) = [\varphi(k)^T\ y(k)]^T\}_{k=1,\ldots,N}$ be observations extracted from a strict-sense stationary sequence, and let $\hat{\mathcal{I}}_N$ be the optimal interval map computed according to Theorem 1. Define $R(\hat{\mathcal{I}}_N)$ as in (8) where $[\varphi^T\ y]^T$ is independent of $D_N$ (that is, $R(\hat{\mathcal{I}}_N)$ is a measure of accuracy of the interval predictor for unseen data, independent of the observations through which the predictor has been constructed). Then, for any $\epsilon, \delta > 0$ such that*

$$\epsilon\delta = \inf_T \left\{ \frac{(n+2)}{\lfloor N/T \rfloor} + \beta(T) \right\}, \tag{15}$$

*where $\beta(T)$ is the $\beta$-mixing function associated with $\{x(k)\}_{k=-\infty}^{\infty}$ and $\lfloor\cdot\rfloor$ denotes integer part, it holds that*

$$\text{Prob}_{P_{[1,N]}}\left\{ R(\hat{\mathcal{I}}_N) \geq 1-\epsilon \right\} \geq 1-\delta. \tag{16}$$

Before proving the theorem, we note that if the observation process is $\beta$-mixing, then $\beta(T) \to 0$ as $T \to \infty$ and, for any $\epsilon > 0$, the confidence parameter $\delta$ given by (15) goes to zero as the number of data points $N$ tends to infinity faster than $T$.

**Proof.** The proof extends that of Theorem 3. Here, we do not introduce any auxiliary sequence $\{z(k)\}$ (as we did in the iid case) since in a mixing context this is difficult to handle. We also note that even in the iid case we could have used the original sequence $\{x(k)\}$ in place of $\{z(k)\}$ with a little loss in the final result: $N+1$ would have been replaced by $N$. Define

$$\bar{R}_N \doteq E_{P_{[1,N]}}[R(\hat{\mathcal{I}}_N)],$$

and, for $k = 1, \ldots, N$, let

$$v_k \doteq \begin{cases} 1, & \text{if } x(k) \text{ is consistent with } \hat{\mathcal{I}}_N^k \\ 0, & \text{otherwise,} \end{cases}$$

where $\hat{\mathcal{I}}_N^k$ is the optimal map which is consistent with

$$D_N^k \doteq \{x(1), \ldots, x(k-1), x(k+1), \ldots, x(N)\},$$

and

$$\hat{\bar{R}}_N \doteq \frac{1}{N} \sum_{k=1}^{N} v_k. \tag{17}$$

Let $C_k$ be the support of $v_k$, i.e. $1(C_k) = v_k$. Now, we have

$$\begin{aligned} E_{P_{[1,N]}}[v_k] &= P_{(-\infty,\infty)}(C_k) \\ &\leq (P_{(-\infty,k-1]\cup[k+1,\infty)} \times P_k)(C_k) + \beta(1) \\ &= E_{P_{(-\infty,k-1]\cup[k+1,\infty)}} \left[ E_{P_k}[1(x(k) \in \hat{\mathcal{I}}_N^k)] \right] + \beta(1) \\ &= E_{P_{[1,k-1]\cup[k+1,N]}} \left[ R(\hat{\mathcal{I}}_N^k) \right] + \beta(1) \\ &\leq E_{P_{[1,N]}} \left[ R(\hat{\mathcal{I}}_N) \right] + \beta(1) \\ &= \bar{R}_N + \beta(1), \end{aligned}$$

which yields

$$E_{P_{[1,N]}}[\hat{\bar{R}}_N] \leq \bar{R}_N + \beta(1). \tag{18}$$

Following the same rationale as in the proof of Theorem 3 where equation (12) is replaced by (18), it is easy to conclude that the result of the theorem holds true for $\epsilon, \delta > 0$ such that

$$\epsilon\delta = \frac{n+2}{N} + \beta(1).$$

The result for a general $T$ is obtained by considering the data subsequence $x(1), x(T+1), x(2T+1), \ldots$ □

6. **Numerical example.** We propose a simple numerical example to illustrate the nature of the presented results. For the purpose of the example, we assumed that the 'unknown' mechanism generating the data is

$$y(k) = u(k-1)(1 + w_1(k)) + 0.1u(k-2)w_2(k), \tag{19}$$

where $w(k) = [w_1(k)\ w_2(k)]^T$ is Gaussian with zero mean and $E[w(k)w^T(j)] = I\delta_{kj}$, being $\delta_{kj}$ the Kronecker delta function, and $u(k) = \sin(k)$.

We set $\varphi(k) = u(k-1)$, and seek an explanatory interval model of the form (3), (4), with $n = 1$:

$$y(k) = \varphi(k)\theta(k) + e(k).$$

In order to fit this explanatory model to the data, we collected $N = 200$ observations $\varphi(k), y(k)$ of the remote process (19) in the data sequence $D_N = \{\varphi(k), y(k)\}$, $k = 1, \ldots, N$.

Setting for instance $\mu_Q = \gamma + 0.6r$, and solving the linear program in Theorem 1 on the basis of the collected observations, we obtained an optimal center $\theta = 0.9612$ with variation radius $r = 2.158$, and level of additive noise $\gamma = 0.1022$. The resulting interval model is shown in Figure 1, together with the observed values of $\varphi(k), y(k)$.

Theorem 3 then states *a-priori* that the reliability level inequality (10) holds with $\epsilon\delta = 3/(N+1)$, for any optimal model of the considered type.

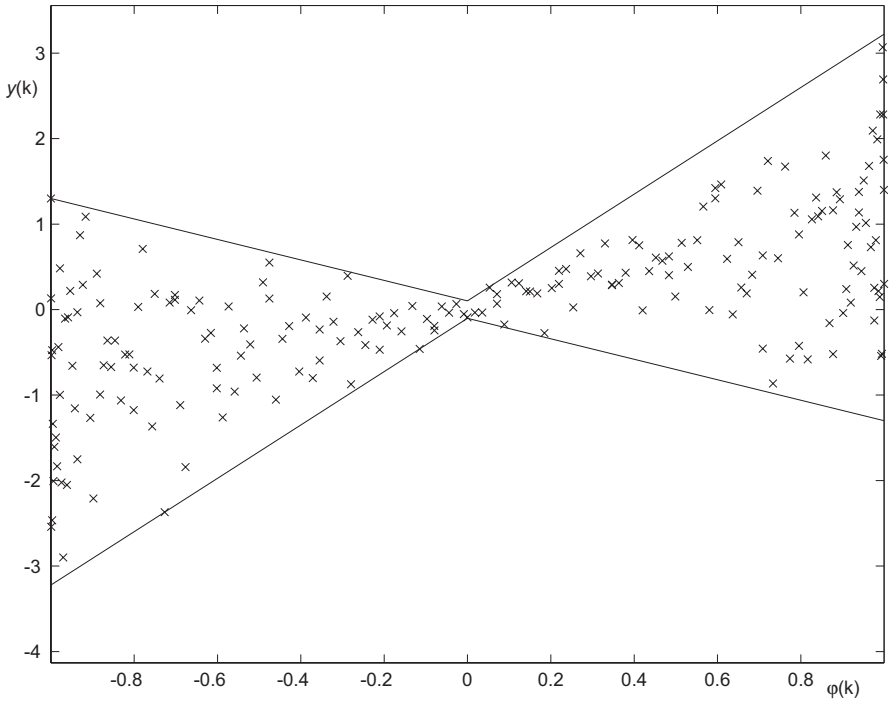


FIGURE 1. For given $\varphi(k)$, the figure shows the resulting interval of possible outputs $y(k)$, as predicted by the optimal interval model constructed on the basis of $N = 200$ observations. The figure also shows the observed points used to construct the model.

We also verified *a-posteriori* (i.e. after the model has been constructed) the reliability of the computed model, by generating $\tilde{N} = 10000$ new observations according to the 'unknown' mechanism, and testing whether they are or not consistent with the constructed model. The result was an estimated empirical reliability of $\tilde{R} = 0.978$ for the above optimal model. In the a-posteriori test the model is fixed, and one can hence apply the standard Hoeffding inequality [6] to qualify the empirical estimate with accuracy $\tilde{\epsilon}$ and confidence $\tilde{\delta}$. In particular, for $\log(2\tilde{\delta})/2\tilde{\epsilon}^2 < \tilde{N}$, we have that

$$\text{Prob}\{|\tilde{R} - R| \geq \tilde{\epsilon}\} \leq \tilde{\delta}.$$

7. **Conclusions.** In this paper, we have studied dynamical models that return a prediction interval for the output of an unknown remote system. From a computational point of view, interval predictors with linear structure can be efficiently constructed numerically, on the basis of a finite number $N$ of past observations, by means of convex programming.

For the more fundamental issue of determining the reliability of such predictors, we derived bounds on the sample complexity $N$ that grow as the inverse of the required probabilistic levels of confidence. These bounds improve by orders of magnitude upon similar bounds derived in [4] that were obtained by means of the Vapnik-Chervonenkis probability inequality.

**Appendix** A. We next present the statement and proof of a key theorem (Theorem 5 below), which is used in the proof of the main result (Theorem 3).

We start with a a technical lemma.

**Lemma 1.** *Given a set $S$ of $p+2$ points in $\mathbb{R}^p$, there exist two points among these, say $\xi_1, \xi_2$, such that the line segment $\overline{\xi_1\xi_2}$ intersects the hyperplane (or one of the hyperplanes if indetermination occurs) generated by the remaining $p$ points $\xi_3, \ldots, \xi_{p+2}$.*

**Proof.** Choose any set $S'$ composed of $p-1$ points from $S$, and consider the bundle of hyperplanes passing through $S'$. If this bundle has more than one degree of freedom, augment $S'$ with additional arbitrary points, until the bundle has exactly one degree of freedom. Consider now the translation which brings one point of $S'$ to coincide with the origin, and let $S''$ be the translated point set. The points in $S''$ lie now in a subspace $\mathcal{F}$ of dimension $p-2$, and all the hyperplanes of the (translated) bundle are of the form $v^T x = 0$, where $v \in \mathcal{V}$, being $\mathcal{V}$ the subspace orthogonal to $\mathcal{F}$, which has dimension 2.

Call $x_4, \ldots, x_{p+2}$ the points belonging to $S''$, and $x_1, x_2, x_3$ the remaining points. Consider three fixed hyperplanes $H_1, H_2, H_3$ belonging to the bundle generated by $S''$, which pass through $x_1, x_2, x_3$, respectively; these hyperplanes have equations $v_i^T x = 0$, $i = 1, 2, 3$. Since $\dim \mathcal{F} = 2$, one of the $v_i$'s (say $v_3$) must be a linear combination of the other two, i.e. $v_3 = \alpha_1 v_1 + \alpha_2 v_2$.

Suppose that one of the hyperplanes, say $H_1$, leaves the points $x_2, x_3$ on the same open half-space $v_1^T x > 0$ (note that assuming $v_1^T x > 0$, as opposed to $v_1^T x < 0$ is a matter of choice since the sign of $v_1$ can be arbitrarily selected). Suppose that also another hyperplane, say $H_2$, leaves the points $x_1, x_3$ on the same open half-space $v_2^T x > 0$. Then, it follows that $v_1^T x_3 > 0$, and $v_2^T x_3 > 0$. Since $v_3^T x_3 = 0$, it follows also that $\alpha_1 \alpha_2 < 0$. We now have that

$$v_3^T x_1 = (\alpha_1 v_1 + \alpha_2 v_2)^T x_1 = \alpha_2 v_2^T x_1$$
$$v_3^T x_2 = (\alpha_1 v_1 + \alpha_2 v_2)^T x_2 = \alpha_1 v_1^T x_2,$$

where the first term has the same sign as $\alpha_2$, and the second has the same sign as $\alpha_1$. Thus, $v_3^T x_1$ and $v_3^T x_2$ do not have the same sign. From this reasoning it follows that not all the three hyperplanes can leave the complementary two points on the same open half-space, and the result is proved. □

We now come to the key instrumental result. Consider the convex optimization program

$$\mathcal{P} : \min_{x \in \mathbb{R}^n} \; c^T x \;\; \text{subject to} \;\; x \in \mathcal{X}_i, \; i = 1, \ldots, m,$$

where $\mathcal{X}_i$, $i = 1, \ldots, m$ are closed convex sets. Let the convex programs $\mathcal{P}_k$, $k = 1, \ldots, m$, be obtained from $\mathcal{P}$ by removing the $k$-th constraint:

$$\mathcal{P}_k : \min_{x \in \mathbb{R}^n} \; c^T x \;\; \text{subject to} \;\; x \in \mathcal{X}_i, \; i = 1, \ldots, k-1, k+1, \ldots, m.$$

Let $x^*$ be any optimal solution of $\mathcal{P}$ (assuming it exists), and let $x_k^*$ be any optimal solution of $\mathcal{P}_k$ (again, assuming it exists). We have the following definition.

**Definition 4** (Support constraints)**.** The $k$-th constraint $\mathcal{X}_k$ is a *support* constraint for $\mathcal{P}$ if problem $\mathcal{P}_k$ has an optimal solution $x_k^*$ such that $c^T x_k^* < c^T x^*$.

The following theorem holds.

**Theorem 5.** *The number of support constraints for problem $\mathcal{P}$ is at most $n$.*

**Proof.** We prove the statement by contradiction. Suppose then that problem $\mathcal{P}$ has $n_s > n$ support constraints and choose any $(n+1)$-tuple of constraints among these.

Then, there exist $n+1$ points (say, without loss of generality, the first $n+1$ points) $x_k^*$, $k = 1, \ldots, n+1$, which are optimal solutions for problems $\mathcal{P}_k$, and which lie all in the same open half-space $\{x : c^T x < c^T x^*\}$. We show next that, if this is the case, then $x^*$ is not optimal for $\mathcal{P}$, which constitutes a contradiction.

Consider the line segments connecting $x^*$ with each of the $x_k^*$, $k = 1, \ldots, n+1$, and consider a hyperplane $\mathcal{H} \doteq \{c^T x = \alpha\}$ with $\alpha < c^T x^*$, such that $\mathcal{H}$ intersects all the line segments. Let $\bar{x}_k^*$ denote the point of intersection between $\mathcal{H}$ and the segment $\overline{x^* x_k^*}$. Notice that, by convexity, the point $\bar{x}_k^*$ certainly satisfies the constraints $\mathcal{X}_1, \ldots, \mathcal{X}_{k-1}, \mathcal{X}_{k+1}, \ldots, \mathcal{X}_{n+1}$, but it does not necessarily satisfy the constraint $\mathcal{X}_k$.

Suppose first that there exists an index $k$ such that $\bar{x}_k^*$ belongs to the convex hull $\text{co}\{\bar{x}_1^*, \ldots, \bar{x}_{k-1}^*, \bar{x}_{k+1}^*, \ldots, \bar{x}_{n+1}^*\}$. Then, since $\bar{x}_1^*, \ldots, \bar{x}_{k-1}^*, \bar{x}_{k+1}^*, \ldots, \bar{x}_{n+1}^*$ all satisfy the $k$-th constraint, so do all points in $\text{co}\{\bar{x}_1^*, \ldots, \bar{x}_{k-1}^*, \bar{x}_{k+1}^*, \ldots, \bar{x}_{n+1}^*\}$ and hence $\bar{x}_k^* \in \text{co}\{\bar{x}_1^*, \ldots, \bar{x}_{k-1}^*, \bar{x}_{k+1}^*, \ldots, \bar{x}_{n+1}^*\}$ satisfies the $k$-th constraint. On the other hand, as it has been mentioned above, $\bar{x}_k^*$ satisfies all other constraints $\mathcal{X}_1, \ldots, \mathcal{X}_{k-1}, \mathcal{X}_{k+1}, \ldots, \mathcal{X}_{n+1}$, and therefore $\bar{x}_k^*$ satisfies *all* constraints. From this it follows that $\bar{x}_k^*$ is a feasible solution for problem $\mathcal{P}$, and has an objective value $c^T x_k^* = \alpha < c^T x^*$, showing that $x^*$ is not optimal for $\mathcal{P}$. Since this is a contradiction, we are done.

Consider now the complementary case in which there does not exist a $\bar{x}_k^* \in \text{co}\{\bar{x}_1^*, \ldots, \bar{x}_{k-1}^*, \bar{x}_{k+1}^*, \ldots, \bar{x}_{n+1}^*\}$. Then, we can always find two points, say $\bar{x}_1^*, \bar{x}_2^*$, such that the line segment $\overline{\bar{x}_1^* \bar{x}_2^*}$ intersects at least one hyperplane passing through the remaining $n-1$ points $\bar{x}_3^*, \ldots, \bar{x}_{n+1}^*$. Such couple of points always exist by virtue of Lemma 1. Denote with $\bar{x}_{1,2}^*$ the point of intersection (or any point in the intersection, in case more than one exists). Notice that $\bar{x}_{1,2}^*$ certainly satisfies all constraints, except possibly the first and the second. Now, $\bar{x}_{1,2}^*, \bar{x}_3^*, \ldots, \bar{x}_{n+1}^*$ are $n$ points in a flat of dimension $n-2$. Again, if one of these points belongs to the convex hull of the others, then this point satisfies all constraints, and we are done. Otherwise, we repeat the process, and determine a set of $n-1$ points in a flat of dimension $n-3$.

Proceeding this way repeatedly, either we stop the process at a certain step (and then we are done), or we proceed all way down until we determine a set of three points in a flat of dimension one. In this latter case we are done all the same, since out of three points in a flat of dimension one there is always one which lies in the convex hull of the other two.

Thus, in any case we have a contradiction and this proves that $\mathcal{P}$ cannot have $n+1$ or more support constraints. □

## REFERENCES


[1] D. Bosq. *Nonparametric Statistics for Stochastic Processes*. Springer, New York, 1998.

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

[3] G. Calafiore and M.C. Campi. Uncertain convex problems: randomized solutions and confidence levels. *Working report, submitted for publication*, 2003.

[4] G. Calafiore, M.C. Campi, and L. El Ghaoui. Identification of reliable predictor models for unknown systems: a data-consistency approach based on learning theory. In *15$^{th}$ IFAC World Congress*, Barcelona, Spain, July 2002.

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

[6] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[7] L. Ljung. *System identification: theory for the user*. Prentice Hall, Englewood Cliffs, N.J., 1999.

[8] E. Weyer. Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, 36:1291–1299, 2000.


*E-mail address*: giuseppe.calafiore@polito.it, campi@ing.unibs.it