Proceedings of the 41st IEEE
Conference on Decision and Control
Las Vegas, Nevada USA, December 2002

Interval Predictors for Unknown Dynamical Systems: an Assessment of Reliability[1]

Giuseppe Calafiore[2] M.C. Campi[3]

Abstract

This paper presents new results for the assessment of reliability of predictive interval maps constructed using a consistency criterion with respect to a finite number of observations. Given a regression vector, our predictive map returns an interval in which the output of the unknown system is likely to fall at the next time instant. The key question we address is then the following: if the map has been constructed based on N observations, what is the probability that the next (unseen) output will actually fall in the interval predicted by the map? We answer to this fundamental question in two different settings. In the first setting we assume that the observations are statistically independent and identically distributed (i.i.d.), while in the second setting we study the case when the observations are generated by a mixing remote process. This latter case is the most relevant in the applications, since it allows for statistical dependence between past and future observations.

Keywords: Prediction, Set-valued maps, Learning theory, Convex optimization.

1 Introduction

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 the standard prediction-error identification setting, [2], [5], 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 mechanism that generates the data are satisfied. However, assuming that we know the structure of the data generation system is often unrealistic.

In this paper, we follow a different 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 directly optimizing over the final result. Precisely, for a prespecified 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 [4] and [8]. In particular, in a previous paper [3] the authors of the present paper proposed an approach for the evaluation of the reliability of interval models based on the Vapnik and Chervonenkis (VC) learning framework. While satisfactory from a conceptual point of view, this theory provides lower bounds on the number of samples required to attain the desired reliability that may be too large to be of practical interest. The results of [3] were also limited to the i.i.d. case, an assumption which is not satisfied by generic dynamical systems possessing memory.

In this paper, we develop a different approach for the assessment of reliability, that provides a dramatic improvement in the bounds on the required number of samples with respect to the VC theory approach, see Section 3. Also, in Section 4, we extend the results for

[1]This work is supported in part by CNR Agenzia 2000 funds, and by the European Commission under the project HYBRIDGE IST-2001-32460

[2]Dipartimento di Automatica e Informatica, Politecnico di Torino. Tel.: +39-011-564.7071; Fax: +39-011-564.7099. E-mail: calafiore@polito.it

[3]Università di Brescia, Italy. E-mail: campi@ing.unibs.it

reliability under i.i.d. observations to the case of weakly dependent observations.

2 Interval predictors and data-consistency

In this section, we introduce the first two key elements of our approach: models that return an interval as output (Interval Predictor Models) and the notion of η-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 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 φ may contain past values of input and output measurements, thus behaving as a regression vector. Standard auto regressive structures AR(n)

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

give rise to (dynamic) IPMs by setting $\varphi(k) \doteq [y(k-1) \cdots y(k-n)]^T$, $q = [\theta^T \, e]^T \in \mathbb{R}^{n+1}$, 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 where variability is present in both an additive and multiplicative fashion

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

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. In our exposition, we assume in particular Δ to be a sphere with center θ and radius r

$$\Delta \doteq \{\theta + \delta : \ \theta, \delta \in \mathbb{R}^n, \|\delta\| \leq r\}. \tag{3}$$

For the current model structure (2)–(3), 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 $e(k)$. Given $\varphi(k)$, the output of the model is the interval

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

One thing that needs to be made clear at this point is that models like (2) 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 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 $k = 1, \ldots, N$, and that the observations are collected in the data sequence $D_N \doteq \{\varphi(k), y(k)\}_{k=1}^N$. We have the following definition.

Definition 1 *An interval model (1) is η-consistent with a given batch of observations D_N if, given $\eta \in [0, 1]$*

$$y(k) \in \mathcal{I}(\varphi(k)), \ \textit{for } k \in \mathcal{K}_\eta, \tag{4}$$

where $\mathcal{K}_\eta \subseteq \{1, \ldots, N\}$ is a set of cardinality $N_\eta \doteq \lfloor \eta N \rfloor$, where $\lfloor \cdot \rfloor$ denotes integer part.

In other words, the above definition requires that the assumed model is not falsified by (and therefore is consistent with) a fraction η of the observations. In particular, when $\eta = 1$ we will say that the model is consistent (or 1-consistent) with (all) the observations. Notice that, for IPMs described as in (1), the η-consistency condition means that there exists a feasible sequence $\{q(k) \in \mathcal{Q}\}_{k \in \mathcal{K}_\eta}$ that satisfies the model equations, i.e. $y(k) = \mathcal{M}(\varphi(k), q(k))$, for $k \in \mathcal{K}_\eta$.

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 has been discussed in [3]. There, the authors introduced a size measure $\mu_Q = \gamma + \alpha r$ for the interval map[1] resulting from the structure (2)–(3), and then constructed an optimal consistent model solving a Linear Programming problem:

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

$$\begin{aligned} &\text{minimize } \gamma + \alpha r, \text{subject to:} \\ &r, \gamma \geq 0 \end{aligned}$$

[1] Note that, if we choose $\alpha = E[\|\varphi(k)\|]$, then μ_Q measures the average amplitude of the output interval.

$$\varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq y(k)$$
$$-\varphi^T(k)\theta - r\|\varphi(k)\| - \gamma \leq -y(k)$$
$$k = 1, \ldots, N.$$

The second issue has also been tackled in [3] using a VC theory approach. The VC learning framework provides a theoretical answer to the reliability problem, but has two main drawbacks: first, the bounds on the number of observations required to obtain a desired reliability may be too high to be of practical interest, and second, this problem is aggravated when the theory is extended from i.i.d. processes to mixing (weakly dependent) processes.

The key objective of this paper is to propose a new framework for the assessment of reliability of optimal linear IPMs. This is done in Section 3 for i.i.d. processes, and it is extended to weakly dependent processes in Section 4. This new approach yields bounds on the required number of observations which are dramatically better than those predicted by the VC theory. In talks with Y. Oishi on occasion of the 15th IFAC conference, we discovered that similar results have been independently derived in [6].

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

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.*

$$R(\hat{\mathcal{I}}_N) \doteq \text{Prob}_P\{y \in \hat{\mathcal{I}}_N(\varphi)\}. \tag{5}$$

Before moving on to present the main result on reliability of interval models, we briefly illustrate geometrically the construction of an interval map. For ease of exposition, we fix $\gamma = 0$, and consider $n = 2$. In this case, at each time instant k, in order to be consistent with an observation, the parameter $\theta(k)$ must lie on the line $\{\theta(k) : \varphi^T(k)\theta(k) = y(k)\}$. When N observations are collected, the algorithm in Theorem 1 determines a minimal circle (in the space of parameter θ) that intersects *all* the observation lines, as shown in an example in Figure 1.

Once the optimal model is constructed, we may ask whether or not a new upcoming observations will still

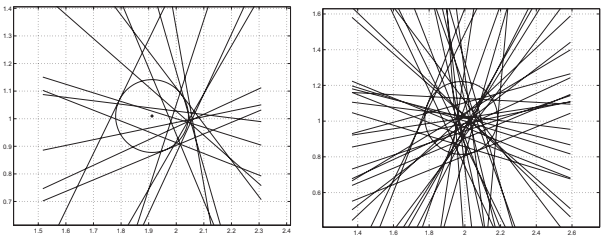

Figure 1: Optimal circle for $N = 15$ and 40 observations.

be consistent with our model. Reliability gives an answer to this question in terms of probability. The main result for i.i.d. observations is given in the following theorem.

Theorem 2 *Let* $D_N = \{x(k) = [\varphi(k)^T\ y(k)]^T\}_{k=1}^N$ *be observations extracted from an i.i.d. sequence with unknown probability measure* P*, and let* $\hat{\mathcal{I}}_N$ *be the optimal interval map computed according to Theorem 1. Then, for any* $\epsilon, \delta > 0$ *such that*

$$\epsilon\delta = \frac{n+2}{N+1} \tag{6}$$

it holds that

$$\text{Prob}_{P^N}\left\{R(\hat{\mathcal{I}}_N) \geq 1 - \epsilon\right\} \geq 1 - \delta. \tag{7}$$

Proof. Consider $N+1$ i.i.d. observations $D_{N+1} = \{z(1), \ldots, z(N+1)\}$, $z(k) \doteq [\psi^T(k)\ \eta(k)]^T$, $\psi(k) \in \mathbb{R}^n$, extracted according to the unknown probability measure P. 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

$$\hat{\bar{R}}_N \doteq \frac{1}{N+1}\sum_{k=1}^{N+1} v_k. \tag{8}$$

We have that

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

which yields

$$E_{P^{N+1}}[\hat{\bar{R}}_N] = \bar{R}_N. \tag{9}$$

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 $\theta \in \mathbb{R}^n, r, \gamma$

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

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 for that *one* single constraint relative to the k-th observation is removed in each problem. From Theorem 4 (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.[2] 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 (8) 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 (9) the expected value of the reliability is bounded as

$$\bar{R}_N = E_{P^{N+1}}[\hat{\bar{R}}_N] \geq 1 - \frac{n+2}{N+1}. \tag{10}$$

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

$$\begin{aligned}E_{P^N}[R(\hat{\mathcal{I}}_N)] \leq (1-\epsilon)&\text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) < 1-\epsilon\}\\ &+1 \cdot \text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) \geq 1-\epsilon\}.\end{aligned} \tag{11}$$

Letting $\bar{\delta} \doteq \text{Prob}_{P^N}\{R(\hat{\mathcal{I}}_N) < 1-\epsilon\}$, combining the bounds (10), (11) we obtain that $\epsilon\bar{\delta} \leq \frac{n+2}{N+1}$, from which the statement of the theorem immediately follows. □

We remark that the bound on the number of samples derived from (6), which is basically $N \geq O(n/\epsilon\delta)$ greatly improves with respect to the bounds obtained by means of the VC-theory approach. In particular, this bound no longer depends on the VC-dimension of the model class, and the dependence in ϵ is linear instead of quadratic.

[2] Whenever one of the problems $\mathcal{P}^k$ does not improve the objective, we select as its optimal solution the optimal solution of $\mathcal{P}$.

4 Reliability of IPMs for weakly dependent observations

The results derived in the previous section for the i.i.d. case are now extended to β-mixing processes.

Definition 3 (β-mixing coefficient, [1]) *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{B}$ and $\mathcal{C}$ be two sub σ-algebras of $\mathcal{F}$. The β-mixing coefficient of $\mathcal{B}$ and $\mathcal{C}$ is defined as*

$$\beta \doteq E\left[\sup_{C \in \mathcal{C}} |P(C) - P(C \mid \mathcal{B})|\right].$$

Suppose now that $\{x(k) \doteq [\varphi^T(k) \ y(k)]^T\}_{k=-\infty}^{\infty}$ is a strict-sense stationary random process, and let $\mathcal{F}_i^j$ denote the σ-algebra generated by the random variables $\{x(k), i \leq k \leq j\}$ (if $j = \infty$, it is intended that $i \leq k$, and similarly if $i = -\infty$). We have the following definition.

Definition 4 (β-mixing process) *Let*

$$\beta(T) \doteq E\left[\sup_{C \in \mathcal{F}_0^0} |P(C) - P(C \mid \sigma\{\mathcal{F}_{-\infty}^{-T}, \mathcal{F}_T^{\infty}\})|\right].$$

The function $\beta(T)$ is named the β-mixing function associated with $\{x(k)\}_{k=-\infty}^{\infty}$. If $\beta(T) \to 0$ as $T \to \infty$, the process is said to be β-mixing.

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

Theorem 3 *Let $D_N = \{x(k) = [\varphi(k)^T \ y(k)]^T\}_{k=1}^N$ be observations extracted from a stric-sense stationary sequence, and let $\hat{\mathcal{I}}_N$ be the optimal interval map computed according to Theorem 1. Further, let P_1^N be the N-dimensional probability of the stationary process and let, for ease of notation, $P = P_1^1$ (that is P is 1-dimensional marginal distribution). Finally, define $R(\hat{\mathcal{I}}_N)$ as in (5) 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*

$$\epsilon\delta = \inf_T \left\{\frac{(n+2)}{\lfloor N/T \rfloor} + \beta(T)\right\}, \tag{12}$$

where $\beta(T)$ is the β-mixing function associated with $\{x(k)\}_{k=-\infty}^{\infty}$, it holds that

$$\text{Prob}_{P_1^N}\left\{R(\hat{\mathcal{I}}_N) \geq 1-\epsilon\right\} \geq 1-\delta. \tag{13}$$

Before proving the theorem, we note that if the observation process is β-mixing, then $\beta(T) \to 0$ as $T \to \infty$ and, for any $\epsilon > 0$, the confidence parameter δ given by (12) goes to zero as the number of data points N tends to infinity.

Proof. The proof extends that of Theorem 2. Here, we do not introduce any auxiliary sequence $\{z(k)\}$ (as we did in the i.i.d. case) since in a mixing context this is difficult to handle. We also note that even in the i.i.d. 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

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

and

$$\hat{\bar{R}}_N \doteq \frac{1}{N} \sum_{k=1}^{N} v_k. \tag{14}$$

Moreover, given an index $k \in \{1, \ldots, N\}$, let P_{k^c} be the $(N-1)$-dimensional distribution of D_N^k, and $P_{k|k^c}$ the conditional distribution of $x(k)$ given D_N^k. Now, we have

$$\begin{aligned} E_{P_1^N}[v_k] &= E_{P_{k^c}}\left[E_{P_{k|k^c}}[v_k|D_N^k]\right] \\ &= E_{P_{k^c}}\left[E_{P_{k|k^c}}[1(x(k) \in \hat{\mathcal{I}}_N^k)|D_N^k]\right] \\ &\leq E_{P_{k^c}}\left[E_P[1(x(k) \in \hat{\mathcal{I}}_N^k)]\right] + \beta(1) \\ &= E_{P_{k^c}}\left[R(\hat{\mathcal{I}}_N^k)\right] + \beta(1) \\ &\leq E_{P_1^N}\left[R(\hat{\mathcal{I}}_N)\right] + \beta(1) \\ &= \bar{R}_N + \beta(1), \end{aligned}$$

which yields

$$E_{P_1^N}[\hat{\bar{R}}_N] \leq \bar{R}_N + \beta(1). \tag{15}$$

Following the same rationale as in the proof of Theorem 2 where equation (9) is replaced by (15), 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 $x(1), x(T+1), x(2T+1), \ldots$ □

A Appendix

In this Section, we present the statement and proof of a key theorem (Theorem 4), which is used in the proof of the main result (Theorem 2). We first state two technical lemmas which are used in the proof. The first lemma is the well known Helly's result for the intersection of convex sets (see for instance [7]).

Lemma 1 (Helly) *Let $\{C_i\}_{i=1,\ldots,n}$ be a finite collection of convex sets in $\mathbb{R}^p$. If every subcollection consisting of $p+1$ or fewer sets has a non-empty intersection, then the entire collection has a non-empty intersection.*

The second technical result is contained in the following lemma.

Lemma 2 *Given a set S of $p+2$ points in $\mathbb{R}^p$, there exist two points among these, say ξ_1, ξ_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 $v_3 x_3 = 0$, it follows also that $\alpha_1\alpha_2 < 0$. We now have that

$$\begin{aligned} v_3^T x_1 &= (\alpha_1 v_1 + \alpha_2 v_2)^T x_1 = \alpha_2 v_2^T x_1 \\ v_3^T x_2 &= (\alpha_1 v_1 + \alpha_2 v_2)^T x_2 = \alpha_1 v_1^T x_2, \end{aligned}$$

where the first term has the same sign as α_2, and the second has the same sign as α_1, therefore $v_3^T x_1$ 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 main result of this Appendix. Consider the convex optimization problem $\mathcal{P}$ in the variable $\vartheta \in \mathbb{R}^d$

$$\mathcal{P}: \quad \text{minimize } s(\vartheta) \quad \text{subject to} \quad \vartheta \in \mathcal{X}_i,\ i = 1, \ldots, m,$$

where $s(\vartheta)$ is a linear objective, and $\mathcal{X}_i$, $i = 1, \ldots, m$ are closed convex sets. Let the convex problem $\mathcal{P}_k$, $k = 1, \ldots, m$ be obtained from $\mathcal{P}$, removing the k-th constraint

$$\mathcal{P}_k: \quad \text{minimize} s(\vartheta) \quad \text{subject to} \quad \vartheta \in \mathcal{X}_i,\ i = 1, \ldots, k-1, k+1, \ldots, m.$$

Let ϑ^* be any optimal solution of $\mathcal{P}$, and let ϑ_k^* be any optimal solution of $\mathcal{P}_k$. We say that the k-th constraint $\mathcal{X}_k$ is a *support* constraint for $\mathcal{P}$, if problem $\mathcal{P}_k$ has an optimal solution ϑ_k^* such that $s(\vartheta_k^*) < s(\vartheta^*)$. We have the following theorem.

Theorem 4 *The number of support constraints for problem $\mathcal{P}$ is at most d.*

Proof. If $m \leq d$ the result is obvious, therefore we consider the case $m > d$, and prove the statement by contradiction. Suppose then that problem $\mathcal{P}$ has $n_s > d$ support constraints; in particular we start assuming $n_s = d + 1$, the case $n_s > d + 1$ will easily follow as shown below.

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

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

Now, if there exist an index k such that $\bar{\vartheta}_k^*$ belongs to the convex hull $\text{co}\{\bar{\vartheta}_1^*, \ldots, \bar{\vartheta}_{k-1}^*, \bar{\vartheta}_{k+1}^*, \ldots, \bar{\vartheta}_{d+1}^*\}$, then a-priori $\bar{\vartheta}_k^*$ satisfies all constraints except possibly for the k-th, but $\bar{\vartheta}_1^*, \ldots, \bar{\vartheta}_{k-1}^*$, $\bar{\vartheta}_{k+1}^*, \ldots, \bar{\vartheta}_{d+1}^*$ all satisfy the k-th constraint, therefore all points in $\text{co}\{\bar{\vartheta}_1^*, \ldots, \bar{\vartheta}_{k-1}^*, \bar{\vartheta}_{k+1}^*, \ldots, \bar{\vartheta}_{d+1}^*\}$ satisfy the k-th constraint, hence $\bar{\vartheta}_k^*$ satisfies the k-th constraint, and therefore it satisfies *all* constraints. From this it follows that $\bar{\vartheta}_k^*$ is a feasible solution for problem $\mathcal{P}$, and has an objective value $s(\vartheta_k^*) < s(\vartheta^*)$, therefore ϑ^* is not optimal for P (contradiction), and we are done.

Otherwise (i.e. if there does not exist a $\bar{\vartheta}_k^* \in \text{co}\{\bar{\vartheta}_1^*, \ldots, \bar{\vartheta}_{k-1}^*, \bar{\vartheta}_{k+1}^*, \ldots, \bar{\vartheta}_{d+1}^*\}$) we can always find two points, say $\bar{\vartheta}_1^*, \bar{\vartheta}_2^*$, such that the line segment $\overline{\bar{\vartheta}_1^* \bar{\vartheta}_2^*}$ intersects at least one hyperplane passing through the remaining $d - 1$ points $\bar{\vartheta}_3^*, \ldots, \bar{\vartheta}_{d+1}^*$. Such couple of points always exist, by virtue of Lemma 2. Denote with $\bar{\vartheta}_{1,2}^*$ a point in this intersection. Notice that $\bar{\vartheta}_{1,2}^*$ certainly satisfies all constraints, except possibly the first and the second. Now, $\bar{\vartheta}_{1,2}^*, \bar{\vartheta}_3^*, \ldots, \bar{\vartheta}_{d+1}^*$ are d points in a flat of dimension $d - 2$. 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 $d - 1$ points in a flat of dimension $d - 3$.

If we go on like this, either we will stop the process at a certain step (and then we are done), or we will proceed 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 for three points in a flat of dimension one, there is always one which lies in the convex hull of the other two. We therefore proved that problem $\mathcal{P}$ cannot have $d + 1$ support constraints.

From a geometric point of view, we proved the following: for $k = 1, \ldots, d+1$, let $\mathcal{C}_k^{\mathcal{H}}$ denote the convex hull generated by the points $\bar{\vartheta}_1^*, \ldots, \bar{\vartheta}_{k-1}^*, \bar{\vartheta}_{k+1}^*, \ldots, \bar{\vartheta}_{d+1}^*$. Then, the convex sets $\mathcal{C}_k^{\mathcal{H}}$, $k = 1, \ldots, d + 1$ have at least one point in common.

Suppose now that $n_s > d + 1$, i.e. there are more than $d + 1$ points $\bar{\vartheta}_k^*$ all lying in the same open half-space. By the previous reasoning, for any subset composed of $d + 1$ of these points, the corresponding collection of convex sets $\mathcal{C}_k^{\mathcal{H}}$ has a non-empty intersection, therefore by Helly's theorem (Lemma 1) the whole collection $\mathcal{C}_k^{\mathcal{H}}$, $k = 1, \ldots, n_s$ has at least a point in common, and this completes the proof. □

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

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

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

[6] Y. Oishi and H. Kimura. Model-set identification based on learning-theoretic inequalities. In *15th IFAC World Congress*, Barcelona, Spain, July 2002.

[7] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1972.

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