Automatica 110 (2019) 108542

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

On a class of interval predictor models with universal reliability ☆

S. Garatti [a], M.C. Campi [b,*], A. Carè [b]

[a] Dipartimento di Elettronica, Informazione e Bioingegneria. Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy
[b] Department of Information Engineering. University of Brescia, via Branze 38, 25123 Brescia, Italy

ARTICLE INFO

Article history:
Received 13 September 2018
Received in revised form 15 May 2019
Accepted 25 July 2019
Available online xxxx

Keywords:
Set-valued models
Interval prediction
Convex optimization
Statistical learning
Minimax regression

ABSTRACT

An Interval Predictor Model (IPM) is a rule by which some observable variables (system inputs) are mapped into an interval that is used to predict an inaccessible variable (system output). IPMs have been studied in Campi et al. (2009), where the problem of fitting an IPM on a set of observations has been considered. In the same paper, upper-bounds on the probability that a future system output will fall outside the predicted interval (misprediction) have also been derived in a stationary and independent framework. While these bounds have the notable property of being valid independently of the unknown mechanism that has generated the data, in general the actual probability distribution of the misprediction does depend on the data generation mechanism and, hence, these bounds may introduce conservatism when applied to a specific case. In this paper, we study the reliability of an important class of IPMs, called *minimax layers*, and show that this class exhibits the special property that the probability distribution of the misprediction is *known exactly* and is *universal*, i.e., is always the same irrespective of the data generation mechanism. This result carries important consequences on the use of minimax layers in practice.

1. Introduction

An Interval Predictor Model (IPM) is a rule $I(\cdot)$ that assigns to a vector of *explanatory variables* $\boldsymbol{x} \in \mathbb{R}^p$ (system inputs) an interval $I(\boldsymbol{x}) \subseteq \mathbb{R}$, which is used to predict the system output. Often, the rule $I(\cdot)$ is constructed from observations: one collects a set of input–output data, $(\boldsymbol{x}_t, y_t)$, $t = 1, \ldots, N$,[1] and identifies an IPM guided by the following two principles: (i) the IPM is consistent with the data-set, that is, points in the data-set are correctly described by the IPM, (ii) the IPM width is minimized so as to obtain small and informative prediction intervals. In Campi, Calafiore, and Garatti (2009), the reliability of interval predictors identified along the above described scheme have been studied in a stationary and independent framework as specified by the following assumption.

Assumption 1 (*Stationarity and Independence*)**.** The process $(\boldsymbol{x}_t, y_t)$, $t = 1, 2, \ldots$, with $\boldsymbol{x}_t \in \mathbb{R}^p$ and $y_t \in \mathbb{R}$ is i.i.d. (independent and identically distributed). Its (*unknown*) distribution at any time t is denoted with $\mathbb{P}$.[2] ⋆

Under this assumption the reliability of an IPM $I(\cdot)$ is formally defined in Campi et al. (2009) as

$$\eta(I) = \mathbb{P}\{y \in I(\boldsymbol{x})\},$$

where $(\boldsymbol{x}, y)$ are distributed according to $\mathbb{P}$. The closer $\eta(I)$ to 1, the more reliable the predictor. When this definition is applied to the IPM $\hat{I}$ that is identified from the data-set $(\boldsymbol{x}_t, y_t)$, $t = 1, 2, \ldots, N$, one should note that $\eta(\hat{I})$ becomes a random variable because $\hat{I}$ depends on the observed data $(\boldsymbol{x}_t, y_t)$. Being a random variable, $\eta(\hat{I})$ is characterized by its probability distribution, and one would like this distribution to concentrate near the value 1 (high reliability). The main result proven in Campi et al. (2009), and then refined in Calafiore (2010) and Campi and Garatti (2011), is that ($\mathbb{P}^N$ refers to data $(\boldsymbol{x}_t, y_t)$, $t = 1, 2, \ldots, N$,

☆ M.C. Campi and A. Carè were partly supported by the H&W 2015 program of the University of Brescia under the project "Classificazione della fibrillazione ventricolare a supporto della decisione terapeutica" (CLAFITE). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Adrian George Wills under the direction of Editor Torsten Soderstrom.

* Corresponding author.
E-mail addresses: simone.garatti@polimi.it (S. Garatti), marco.campi@unibs.it (M.C. Campi), algo.care@unibs.it (A. Carè).

[1] Although we adopted the standard time index t, the results of this paper apply also to non-temporal sequences of data, e.g. spatially indexed data.

[2] Stationarity (that is, the distribution of $(\boldsymbol{x}_t, y_t)$ is the same for any t) says that the system is invariant in time. Independence, instead, rules out the presence of inter-time correlations. However, the results in Campi et al. (2009) are approximately applicable to correlated processes provided that the correlation pattern is estimated and compensated for according to a deconvolution process, see Campi et al. (2009) for more discussion.

https://doi.org/10.1016/j.automatica.2019.108542
0005-1098/

by which $\hat{I}$ has been constructed)

$$\mathbb{P}^N\{\eta(\hat{I}) < 1-\epsilon\} \leq \beta, \tag{1}$$

where β (*confidence parameter*) goes to zero exponentially fast with N, and can therefore be made very small (e.g., 10^{-6}) for data sample sizes N of practical interest. When β is so small to be negligible, one can think of $1-\epsilon$ as a "practically certain" lower-bound for $\eta(\hat{I})$. Importantly, β does not depend on $\mathbb{P}$ (i.e., the data generation system).

1.1. The result of this paper

Result (1) is valid for any data generation mechanism. Nonetheless, in general, the distribution of $\eta(\hat{I})$ does depend on the specific data generation mechanism and, therefore, the bound in (1) can be conservative for a specific data generation mechanism. In contrast, in this paper our goal is to investigate classes of IPMs for which the reliability $\eta(\hat{I})$ is independent of the data generation mechanism and can therefore be evaluated without conservatism. It turns out that the class of *minimax layers* which is well-known from the statistical literature (see Section 2.3) does have this property. This result is established and discussed in full extension in this article.

In more specific terms, minimax layers are obtained from linearly parameterized regression models by fitting the parameters according to a minimax criterion and then considering the smallest layer around the so-obtained model that contains the data points (all of them, or all but the exception of some of them). The main result of this paper is that, under very general assumptions, the distribution of the reliability $\eta(\hat{I})$ for minimax layers is always the same independently of the data generation mechanism. In other words, the value of $\mathbb{P}^N\{\eta(\hat{I}) < 1-\epsilon\}$ does not depend on how data are generated and, hence, it becomes a quantity known to the user who can employ it to certify the reliability of the prediction (e.g., by building *exact* confidence intervals for $\eta(\hat{I})$). We express this fact by saying that the reliability is *universal*. Another fundamental fact also proved in this paper is that the reliability distribution does not depend on the type of regressors (e.g., polynomial or trigonometric) that are used in the model. At a conceptual level, this result implies a separation principle: while the chosen regressors do impact on the width of the minimax layer, the reliability distribution is not influenced by them. Hence, any prior knowledge on the data generation mechanism can be used to properly design the regressors and, moreover, one can adjust the regressors by a-posteriori evaluating the layer width while the reliability is kept under control by the theoretical results established in this paper.

1.2. Discussion on related literature

The present paper is in the vein of the IPM theory introduced in Campi et al. (2009). Interval predictor models as descriptive tools existed before (Campi et al., 2009), and the reader can consult the theory of differential inclusions, set-valued dynamical systems and set prediction, (Jaulin, Kieffer, Braems & Walter, 2001; Jaulin, Kieffer, Didrit & Walter, 2001; Kieffer, Jaulin, & Walter, 2002; Milanese, Norton, Piet-Lahanier, & Walter, 2013; Walter & Pronzato, 1997). For a philosophical discussion on the probabilistic viewpoint as compared to bounding approaches (see e.g. Milanese et al. (2013)), the reader is instead referred to the position paper (Campi, Csáji, Garatti, & Weyer, 2012). See also Calafiore (2010), Crespo, Giesy, and Kenny (2014), Crespo, Kenny, and Giesy (2015, 2016) and Lacerda and Crespo (2017) for other recent contributions on IPMs, and Patelli, Broggi, Tolo, and Sadeghi (2013) and Carè, Garatti, and Campi (2014) for a software and a fast algorithm.

The theoretical framework of this paper is grounded on the scenario approach of Calafiore and Campi (2006), Campi and Garatti (2008, 2018) and Garatti and Campi (2013). In the terminology of the scenario approach, Lemma 1 in Appendix A of the present paper states that estimating a minimax layer is a *fully supported problem* (see Definition 3 in Campi and Garatti (2008)). This is key to establishing the fundamental result here proved that the distribution of the reliability of minimax layers does not depend on the data generation mechanism and on the regressors. Moreover, in this paper we also introduce IPMs with tunable width and guarantee their reliability in the spirit of the results in Carè, Garatti, and Campi (2015). Part of the material here presented appeared in preliminary form in the conference paper (Garatti & Campi, 2009): specifically, when Theorem 1 of this paper is applied to describe the reliability of the widest layer ($\ell = 1$ in Theorem 1) the main theorem of Garatti and Campi (2009) is recovered. Importantly, Theorem 1 of the present paper lends itself to be used as a rigorous quantitative tool for the selection of the layer width, which is not possible from the result in Garatti and Campi (2009).

Minimax layers are grounded on the minimax criterion of best fit, also known as L_∞ criterion. We shall provide some references about the L_∞ criterion of best fit in Section 2.3 after rigorously defining the construction of minimax layers.

1.3. Structure of the paper

Minimax layers are formally introduced in the next Section 2. In Section 3, we focus on the universal reliability of minimax layers and provide a complete description of the corresponding reliability distribution. Conclusions are drawn in Section 4. All of the technical proofs are provided in Appendix A.

2. Minimax layer IPMs

We consider linearly parameterized regression models of a variable $y \in \mathbb{R}$ on a p-dimensional vector of explanatory variables $\boldsymbol{x} \in \mathbb{R}^p$. Precisely, given q *regressor functions* $f_j : \mathbb{R}^p \to \mathbb{R}$, $j = 1, \ldots, q$, the regression model is given by

$$\hat{y}(\boldsymbol{x}) = \sum_{j=1}^{q} \theta_j f_j(\boldsymbol{x}) = f(\boldsymbol{x})^T\theta, \tag{2}$$

where $f(\boldsymbol{x}) := [f_1(\boldsymbol{x}) \ \cdots \ f_q(\boldsymbol{x})]^T$ is the vector of regressor functions and $\theta = [\theta_1 \ \cdots \ \theta_q]^T$ is the vector of tunable parameters. As a simple example, (2) encompasses affine models in $\boldsymbol{x}$, that is $\hat{y}(\boldsymbol{x}) = \theta_1 x^{(1)} + \theta_2 x^{(2)} + \cdots + \theta_p x^{(p)} + \theta_{p+1}$, where superscript (i) indicates the ith component of vector $\boldsymbol{x}$.

Given a batch of N independent and identically distributed (i.i.d.) observations $(\boldsymbol{x}_t, y_t)$, $t = 1, \ldots, N$, the regression model is tuned according to the minimax, or L_∞, criterion of best fit, which amounts to selecting the parameters θ_j so as to minimize the maximum deviation of the observed y_t's from $\hat{y}(\boldsymbol{x}_t)$, namely,

$$\min_{\theta=[\theta_1 \ \cdots \ \theta_q]^T} \max_{t=1,\ldots,N} \left| y_t - f(\boldsymbol{x}_t)^T\theta \right|. \tag{3}$$

The optimal solution of (3) is denoted by $\theta^* = [\theta_1^* \ \cdots \ \theta_q^*]^T$, and the optimal value is $h^* := \max_{t=1,\ldots,N} |y_t - f(\boldsymbol{x}_t)^T\theta^*|$. The layer of vertical height $2h^*$ centered around the fitted model $\hat{y}(\boldsymbol{x}) := f(\boldsymbol{x})^T\theta^*$ is called the *minimax layer* (also known as the Chebyshev layer), see Fig. 1. The minimax layer provides a rule $I : \boldsymbol{x} \to I(\boldsymbol{x}) \subset \mathbb{R}$, where, to each $\boldsymbol{x}$, there corresponds the interval $I(\boldsymbol{x})$ given by the intersection of the vertical line departing from $\boldsymbol{x}$ with the minimax layer, i.e.,

$$\hat{I}(\boldsymbol{x}) = [f(\boldsymbol{x})^T\theta^* - h^*, f(\boldsymbol{x})^T\theta^* + h^*]. \tag{4}$$

This rule defines a so-called Interval Predictor Model (IPM), Campi et al. (2009).

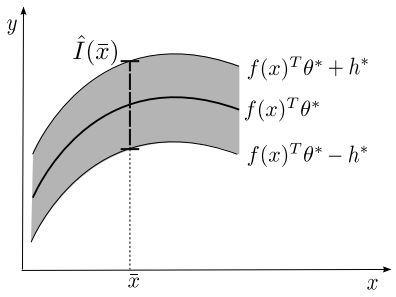

Fig. 1. The minimax layer and the corresponding interval prediction for a given $\bar{\boldsymbol{x}}$.

2.1. Minimax ℓ-th layer IPMs

The *minimax layer* described above is not the only IPM that can be of interest and other layers can be defined based on the fitted function $\hat{y}(\boldsymbol{x}) = f(\boldsymbol{x})^T\theta^*$; for instance, one can consider layers that include some but not all of the observations. More formally, consider the set of values $v_t = |y_t - f(\boldsymbol{x}_t)^T\theta^*|$, $t = 1, \ldots, N$, sort them in descending order with no repeats so that $v_{(1)} > v_{(2)} > \cdots$ and define h^*_ℓ to be the ℓ-th highest value, i.e., $h^*_\ell = v_{(\ell)}$. Note that, with this notation, $h^*_1 = h^*$. The *minimax ℓ-th layer* IPM (or, for short, just the *ℓ-th layer*) is defined by the rule

$$\hat{I}_\ell(\boldsymbol{x}) = [f(\boldsymbol{x})^T\theta^* - h^*_\ell, f(\boldsymbol{x})^T\theta^* + h^*_\ell] \tag{5}$$

(note that $\hat{I}_1(\boldsymbol{x}) = \hat{I}(\boldsymbol{x})$). While the minimax layer ($\ell = 1$) includes all the observations, for $\ell > 1$ the ℓ-th layer is not consistent with all of them, i.e., $y_t \notin \hat{I}_\ell(\boldsymbol{x}_t)$ for some values of $t \in \{1, \ldots, N\}$, and it is a remarkable fact, proved in Appendix A (see Lemma 1 and the discussion thereafter) that, under very mild assumptions, the ℓ-th layer is inconsistent with precisely $\ell + q - 1$ observations. The fact that the ℓ-th layer $\hat{I}_\ell(\boldsymbol{x})$ is thinner than $\hat{I}(\boldsymbol{x})$ comes at the price of a loss in reliability. As we shall see, an *exact* evaluation of this loss for all the values of ℓ is possible under very general conditions.

2.2. Notation

The *reliability* of the minimax ℓ-th layer (5) is defined as $\mathbb{P}\{(\boldsymbol{x}, y)$ such that $\left|y - f(\boldsymbol{x})^T\theta^*\right| \leq h^*_\ell\}$, and it will be indicated by $\eta(\hat{I}_\ell)$ or just η_ℓ for short.

2.3. Some historical remarks on L_∞ regression

L_∞ regression has a long history, see e.g. Harter (1975). It was introduced by Euler (1749), some half a century before least squares regression, although a first resolution method for particular cases was given only in the late 18th century by Laplace (1783, 1793), and then extended to a more general framework in the early 19th century by Fourier (1824). Since then, L_∞ regression has been further developed by many authors, notably by Chebyshev (1854) and Haar (1918). A surge of renewed interest for this method started in the 1950s, partly spurred by the development of linear programming techniques to compute the L_∞ regression solution, (Appa & Smith, 1973; Armstrong & Kung, 1979; Barrodale & Phillips, 1975; Karst, 1958; Planitz & Gates, 1991; Wagner, 1959; Zhang, 1993). See Arthanari and Dodge (1993), Birkes and Dodge (1993) and Cheney (1999) for comprehensive presentations of L_∞ regression. Paper Harter (1982) points out that the minimax method is a valuable alternative to least squares provided that the causes of variability of y are well-captured by the explanatory variables $\boldsymbol{x}$.

3. The universal reliability of minimax layers

Before stating the main theorem, we prove some preliminary results of independent interest that are instrumental to the derivation of the main theorem.

3.1. Existence and uniqueness of θ^*

The existence of θ^* immediately follows by the observation that the function to be minimized, $\max_{t=1,\ldots,N} |y_t - f(\boldsymbol{x}_t)^T\theta|$, is non-negative and piecewise-linear.

Uniqueness is more involved and is proven under the following conditions.

Condition 1. *The probability $\mathbb{P}$ according to which observations are generated admits density $p(\boldsymbol{x}, y)$.* ⋆

Condition 2. *For any $\overline{\theta} \in \mathbb{R}^q$, $\overline{\theta} \neq 0$, relationship $f(\boldsymbol{x})^T\overline{\theta} = 0$ holds at most on a zero Lebesgue measure set.* ⋆

Condition 2 says that the functions $f_j(\boldsymbol{x})$ are linearly independent on nonzero Lebesgue measure sets, and this corresponds to requiring that none of the regressor functions is superfluous over a set having nonzero Lebesgue measure. For example, this condition is not satisfied when the regressor functions are $n + 2$ polynomials of degree at most n, so that one regressor function is certainly a linear combination of the others. However, standard choices of regressor functions satisfy the condition (e.g., monomials of different degrees, orthonormal trigonometric terms, etc.).

Since $(\boldsymbol{x}, y)$ admits density, $\boldsymbol{x}$ also does, that is, the marginal probability $\mathbb{P}_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is absolutely continuous with respect to the Lebesgue measure, so that Condition 2 implies that

$$\mathbb{P}_{\boldsymbol{x}}\left\{f(\boldsymbol{x})^T\overline{\theta} = 0\right\} = 0, \quad \forall \overline{\theta} \in \mathbb{R}^q,\ \overline{\theta} \neq 0. \tag{6}$$

Uniqueness of the solution of (3) follows from (6), as established in the following proposition (the proof of which is in Appendix A.1).

Proposition 1. *Problem (3) with $N \geq q$ admits with probability 1 a unique solution if and only if (6) holds.* ⋆

3.2. The probability distribution of η_ℓ

The following theorem is the main result of this paper and states that the minimax layer ($\ell = 1$) and minimax ℓ-th layers ($\ell > 1$) have universal reliability.

Theorem 1. *Let $N \geq q + 1$, and assume that Conditions 1 and 2 hold. For any given $\ell \in \{1, \ldots, n\}$, where n is the number of distinct values in the set $\{|y_t - f(\boldsymbol{x}_t)^T\theta^*|,\ t = 1, \ldots, N\}$, the probability distribution of $\eta_\ell \in [0, 1]$ is*

$$F_{\eta_\ell}(z) := \mathbb{P}^N\{\eta_\ell \leq z\} = \sum_{i=0}^{q+\ell-1} \binom{N}{i} (1-z)^i z^{N-i}. \tag{7}$$

Note that $F_{\eta_\ell}(z)$ does not depend on the probability $\mathbb{P}$ according to which data are generated, nor does it depend on the regression functions f_j used. ⋆

In the theorem, $\mathbb{P}^N = \mathbb{P} \times \cdots \times \mathbb{P}$ refers to the product probability distribution of the N observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$. The proof is given in Appendix A.2; in the following we discuss the significance of the theorem.

In words, Theorem 1 says that η_ℓ is a random variable with a *Beta* distribution with parameters $(N - q - \ell + 1, q + \ell)$, irrespective of the probability with which $(\boldsymbol{x}_t, y_t)$ are extracted

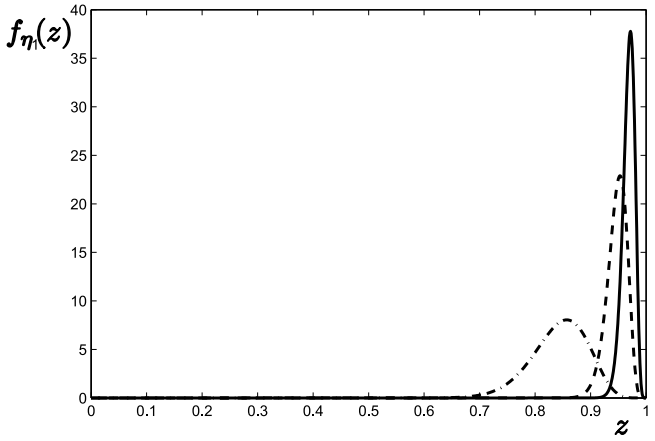

Fig. 2. $f_{\eta_1}(z)$ for $q = 7$ and $N = 50$ (dash-dotted line), $N = 150$ (dashed line), $N = 250$ (solid line).

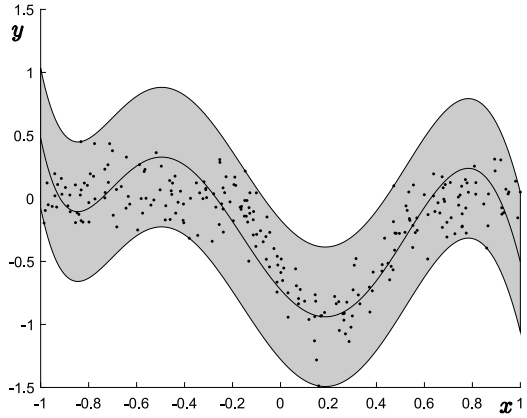

Fig. 3. Polynomial regression model and corresponding minimax layer. ($\theta^* = [-0.7233, -2.1691, 4.3338, 7.2844, -6.9323, -5.6068, 3.3085]$, $h^* = 0.5536$).

and of the functional form of the regressors f_j. The property that the distribution of η_ℓ does not depend on the distribution of the observations can be phrased by saying that "η_ℓ is a pivotal random variable". Note that, differently from results like (1) that are available in the IPM literature, Eq. (7) does not only provide a bound, it assigns the exact probability distribution of η_ℓ.[3] From Eq. (7) one can compute the probability density of η_ℓ, which is

$$f_{\eta_\ell}(z) := \frac{\mathrm{d}}{\mathrm{d}z} F_{\eta_\ell}(z) = (q+\ell)\binom{N}{q+\ell}(1-z)^{q+\ell-1} z^{N-q-\ell}.$$

Its expectation

$$\mathbf{E}[\eta_\ell] = \int_0^1 z f_{\eta_\ell}(z)\,\mathrm{d}z = \frac{N-q-\ell+1}{N+1} \tag{8}$$

is the mean value of the reliability η_ℓ.

The probability density function of η_1 is graphically visualized for different values of N in Fig. 2. As it appears, the distribution of η_1 tends to concentrate near 1 as N increases. By using this density, one can quantify exactly the reliability of the minimax layer for any finite N without availing of any knowledge of the data generation mechanism. For an approximate evaluation one can use Lemma 1 in Alamo, Tempo, Luque, and Ramirez (2015), so obtaining

$$\mathbb{P}^N\{\eta_1 \le z\} = \sum_{i=0}^{q}\binom{N}{i}(1-z)^i z^{N-i} \le e^q\left(\frac{1-z}{e}+z\right)^N,$$

which reveals that, for any fixed z, $\mathbb{P}^N\{\eta_1 \le z\}$ tends to zero exponentially fast as N increases.

Besides characterizing the minimax layer, Theorem 1 also quantifies the loss in reliability incurred for reducing the width of the minimax layer by taking larger values of ℓ. By inspecting (7), one observes that q and ℓ only appear one summed to the other in the upper limit of the summation, so that increasing ℓ to $\ell+1$ has the same effect as increasing by one the size of the parameter vector θ. In particular, it holds that $\mathbf{E}[\eta_{\ell+1}] = \mathbf{E}[\eta_\ell] - \frac{1}{N+1}$.

In the following, we provide an example and additional comments to help gain insight in all these results.

3.3. An example

Let $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}$,[4] and suppose that $N = 250$ independent points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{250}, y_{250})$ are available.[5]

A polynomial regression model $y(\boldsymbol{x}) = \theta_1 + \theta_2\boldsymbol{x} + \cdots + \theta_7\boldsymbol{x}^6$ is tuned according to the L_∞ criterion:

$$\min_{\theta=[\theta_1\ \cdots\ \theta_7]^T} \max_{t=1,\ldots,250} \left|y_t - (\theta_1 + \theta_2\boldsymbol{x}_t + \cdots + \theta_7\boldsymbol{x}_t^6)\right|, \tag{9}$$

and the corresponding minimax layer is shown in Fig. 3.

What is the confidence we have in the claim that a next, still unseen, point falls in the layer with probability at least 90%? This question is the same as asking for the probability that $\eta_1 \ge 0.9$, and the answer can be found in Theorem 1: this probability is equal to $1 - \sum_{i=0}^{7}\binom{250}{i}(1-0.9)^i 0.9^{250-i} \approx 1 - 10^{-5}$. In other words, it is extremely likely that the obtained minimax layer contains at least 90% of the probability mass with which data are generated. From Eq. (8) we also see that the mean value of the reliability η_1 for minimax layers constructed based on (9) is exactly $\frac{250-7}{250+1} \approx 0.968$. Knowing the exact distribution of η_1 as given by Theorem 1 makes it possible to compute an upper-bound to the reliability η_1 as well, and therefore to provide an exact confidence interval for η_1. For example, it holds true that $\mathbb{P}^N\{\eta_1 \le 0.996\} = \sum_{i=0}^{7}\binom{250}{i}(1-0.996)^i 0.996^{250-i} \approx 1 - 10^{-5}$, from which $\mathbb{P}^N\{\eta_1 \in [0.9, 0.996]\} \approx 1 - 2\cdot 10^{-5}$.

Upon inspection of Fig. 3, it appears that the constructed layer is not tight around the data points and in fact the layer contains wide empty portions. Considering instead a trigonometric regression model, we can set out to solve the minimization

[3] For a more specific comparison with the results of Campi and Garatti (2011), note that equality (7) holds for the class of IPMs considered in this paper, while the results in Theorem 2.1 of Campi and Garatti (2011) are valid in wider generality. It can be observed that the distribution η_1 as computed in the present paper achieves exactly the bound given in Campi and Garatti (2011), Theorem 2.1 (equation (3) with $k = 0$). Hence, this paper proves that the bound in Campi and Garatti (2011) is tight for the class of IPMs at hand in this paper. On the other hand, the bound in Campi and Garatti (2011) is looser than (7) when the IPM is allowed not to be consistent with some of the data points ($\ell > 1$ in this paper; $k > 0$ in Campi and Garatti (2011)), as it is clear from the extra binomial coefficient term in equation (3) of Campi and Garatti (2011). Moreover, the construction in this paper and in Campi and Garatti (2011) are different: ℓ in this paper determines the width reduction of the minimax layer around the fitted model $\hat{y}(\mathbf{x})$, while the parameter k in Campi and Garatti (2011) accounts for the removal of data points from the data-set according to a generic scheme, see Campi and Garatti (2011) for more details.

[4] We consider a toy example with $\boldsymbol{x} \in \mathbb{R}$ to allow visualization of the results.

[5] For the sake of completeness, we let the reader know that the points $(\boldsymbol{x}_t, y_t)$ were generated according to the equation

$$y_t = -e^{-15(\boldsymbol{x}_t - 0.2)^2} + n_t,$$

where the $\boldsymbol{x}_t$'s are i.i.d., uniformly distributed over $[-1, 1]$, and the n_t's are defined by $n_t = 0.09\log\left(\frac{1+u_t}{1-u_t}\right)$, with u_t independently and uniformly distributed over $[-1, 1]$.

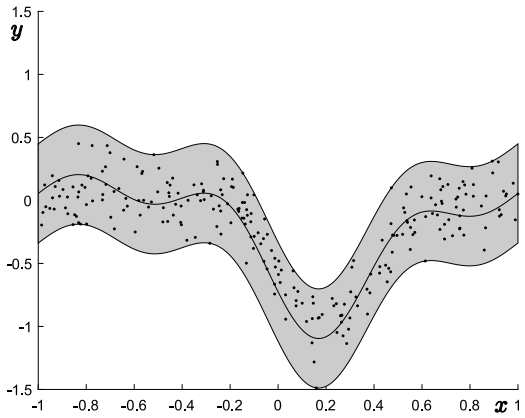

Fig. 4. Trigonometric regression model and corresponding minimax layer. ($\theta^* = [-0.2311, -0.2818, -0.3758, -0.1876, -0.1037, -0.1839, -0.0144]$, $h^* = 0.3936$).

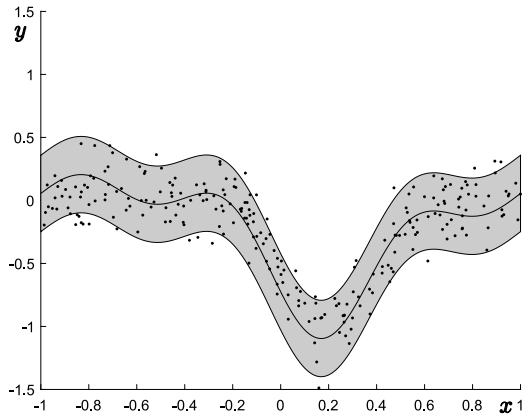

Fig. 5. Trigonometric regression model and corresponding minimax 19-th layer ($h^*_{19} = 0.3032$). We recall that the minimax ℓ-th layer is not consistent with $\ell + q - 1$ observations, and therefore in this case 25 data points lie outside the layer.

problem

$$\min_{\theta=[\theta_1 \ \cdots \ \theta_7]^T} \max_{t=1,\ldots,250} \Big| \ y_t - (\theta_1 + \theta_2 \sin(\pi \boldsymbol{x}_t) + \theta_3 \cos(\pi \boldsymbol{x}_t) + \cdots + \theta_6 \sin(3\pi \boldsymbol{x}_t) + \theta_7 \cos(3\pi \boldsymbol{x}_t)) \Big|.$$

The obtained layer is in Fig. 4, and it more tightly wraps the observations (which is also clear from the optimal value $h^* = 0.3936$ against the optimal value with a polynomial regression that was $h^* = 0.5536$). As already noted, Theorem 1 holds irrespective of the chosen regression functions, so that we can make in this case the same claims as before; in particular that a layer constructed around the trigonometric regression model satisfies condition $\eta_1 \geq 0.9$ with probability $1 - 10^{-5}$.

In closing, we note that in this toy example data $(\boldsymbol{x}_t, y_t)$ are in $\mathbb{R}^2$, should $\boldsymbol{x}$ be of higher dimension, all the considerations here exposed would remain the same since the results in this paper do not depend on the dimension of $\boldsymbol{x}$ but only on the number of regressor functions used.

3.4. A-posteriori selection of the regressor functions

The example in the previous section shows that the layer width h^* depends on the chosen regressor functions. In selecting the regression functions, one uses the prior information available on the problem. Moreover, the value of h^* becomes known to the user at the end of the optimization procedure. This suggests that one can try different choices of regressor functions and a-posteriori select the one that gives the highest accuracy, that is, the lowest value of h^*. For instance, in the example of Section 3.3, the user can inspect the result represented in Fig. 3 against that in Fig. 4 and decide in favor of the second construction since it provides a tighter description of the observations. This way of proceeding, however, involves a choice that requires some attention as explained in what follows. The fact that a polynomial layer like the one in Fig. 3 is reliable at level 90% with confidence $1 - 10^{-5}$ means that in one experiment out of 10^5 the layer has reliability less than 90%. Similarly, a trigonometric layer like the one in Fig. 4 has reliability less than 90% in one out of 10^5 cases. If one layer is chosen from the two types of layers after that they have been constructed from data, it is possible that, every time a layer (either polynomial or trigonometric) with reliability below 90% is constructed, this layer is chosen by the user. This fact may increase above 10^{-5} the probability of selecting a layer with reliability below 90%. This probability, however, can be taken under control by a rigorous union bound: if a polynomial layer can have reliability below 90% with probability 10^{-5}, and so does a trigonometric layer, then both layers certainly have reliability not less than 90% with probability at least $1 - 2 \cdot 10^{-5}$ and, hence, whichever criterion is used to select the layer, with probability $1 - 2 \cdot 10^{-5}$, the chosen layer has the desired level of reliability.

More in general, many regression models can be compared and one of them can be chosen while preserving high confidence. Suppose e.g. that we insist to have confidence $1 - 10^{-5}$ that the reliability is at least 90% while choosing an IPM from a set of 100 IPMs. Using Theorem 1, one can draw the conclusion that this result can be achieved by an increase of the number of observations from 250 (as it was in the example in Section 3.3) to 309. In fact, formula (7) in Theorem 1 ensures that, with 309 observations, a given IPM is guaranteed to have reliability $\eta_1 \geq 0.9$ with probability $1 - 10^{-7}$. Thus, the probability that none of the 100 candidate IPMs has reliability less than 90% is at least $1 - 100 \cdot 10^{-7} = 1 - 10^{-5}$. The fact that the confidence can be increased from $1 - 10^{-5}$ to $1 - 10^{-7}$ and yet the number of observations remains moderate is due to the fact that the Beta distribution is thin-tailed. This can be expressed by saying that *confidence is cheap*.

3.5. A-posteriori selection of ℓ

The user might be willing to accept a reduction in reliability to favor accuracy, i.e., to obtain a thinner layer. To this end, minimax ℓ-th layers can be built for various values of ℓ and the reliability of the chosen one can be taken under control by the same union bound argument that was used in Section 3.4. For example, we have seen that the construction of the minimax layer in Fig. 4 is reliable at level 90% with confidence $1 - 10^{-5}$. The 19-th layer for the same data-set is shown in Fig. 5 and is reliable at level 80% with confidence $1 - 10^{-5}$. Thus, with probability $1 - 2 \cdot 10^{-5}$ it holds that the construction in Fig. 4 returns a layer with reliability at least 90% and, simultaneously, the construction in Fig. 5 returns a layer with reliability at least 80%. Therefore, an a-posteriori selection leaves the user with a confidence of $1 - 2 \cdot 10^{-5}$ that the reliability is either 90% or 80% depending on the choice.

4. Conclusions

In this paper we have shown that minimax layers form a class of Interval Predictor Models that achieve *universal reliability*, i.e., their reliability has the same probability distribution independently of the data generation mechanism under very general

assumptions. We have discussed the implications of this universal reliability property in terms of complete separation between accuracy and reliability. We have also shown that a union bound argument allows one to guarantee the reliability of a model that is selected a-posteriori among several ones. Future work will concentrate on improving the union bound so as to possibly remove any conservatism in it contained. Moreover, one can devise more sophisticated schemes to generate various models to choose from. For example, a regularization approach similar to Campi and Carè (2013) can be employed in order to slim down the number of regressors, while the theory here developed can be used to keep control on the reliability of the chosen layer.

Appendix A. Proofs

A.1. Proof of Proposition 1

(if) Suppose that (6) holds, so that the probability that the vector $f(\boldsymbol{x}) = [f_1(\boldsymbol{x}) \ \cdots \ f_q(\boldsymbol{x})]^T$ belongs to a given subspace of $\mathbb{R}^q$ of dimension less than q is zero. We show that this implies that the condition

$$\text{for every choice of } q \text{ different indexes } t_1, t_2, \ldots, t_q \text{ from } 1, \ldots, N, \text{ the vectors } f(\boldsymbol{x}_{t_1}), f(\boldsymbol{x}_{t_2}), \ldots, f(\boldsymbol{x}_{t_q}) \text{ are linearly independent} \tag{A.1}$$

holds with probability 1. Since (A.1) is the well-known Haar's condition for the uniqueness of the solution of (3) (see Cheney, 1999; Haar, 1918), the "if" part of the proposition is then established.

To show (A.1), note that (6) implies that the probability that $f(\boldsymbol{x}_{t_1}) = 0$, i.e. that $f(\boldsymbol{x}_{t_1})$ falls at the origin, is zero. Hence, $f(\boldsymbol{x}_{t_1}) \neq 0$ with probability 1, and consider the 1-dimensional subspace containing $f(\boldsymbol{x}_{t_1})$. The probability that $f(\boldsymbol{x}_{t_2})$ belongs to this subspace is zero again by (6), so that $f(\boldsymbol{x}_{t_1})$ and $f(\boldsymbol{x}_{t_2})$ form a subspace of dimension 2 with probability 1. Proceeding the same way with all the q vectors $f(\boldsymbol{x}_{t_1}), f(\boldsymbol{x}_{t_2}), \ldots, f(\boldsymbol{x}_{t_q})$, we arrive to the conclusion that (A.1) holds with probability 1.

(only if) Suppose instead that (6) does not hold, that is, $\mathbb{P}_{\boldsymbol{x}}\{f(\boldsymbol{x})^T\bar{\theta} = 0\} > 0$ for some given $\bar{\theta} \neq 0$. Then, there is a non-zero probability that $f(\boldsymbol{x}_t)^T\bar{\theta} = 0$ for all $t = 1, \ldots, N$. In this case, denoting by θ^* a solution to (3), we have that $\max_{t=1,\ldots,N} |y_t - f(\boldsymbol{x}_t)^T(\theta^* + \bar{\theta})| = \max_{t=1,\ldots,N} |y_t - f(\boldsymbol{x}_t)^T\theta^*|$, i.e. $\theta^* + \bar{\theta}$ attains the same optimal value as θ^*. This shows that the solution is not unique with non-zero probability. □

A.2. Proof of Theorem 1

Preliminary results.

To establish the result, we have to enlarge our viewpoint and, instead of considering minimax problems with N independent observations, we need to consider any number M, $M \geq q + 1$, of independent observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_M, y_M)$ generated according to $\mathbb{P}$:

$$\min_{\theta} \max_{t=1,\ldots,M} |y_t - f(\boldsymbol{x}_t)^T\theta|. \tag{A.2}$$

Throughout this part on preliminary results, (θ^*, h^*) denotes the (unique with probability 1) solution and optimal value of (A.2).

Definition 1 (*Observation of Support*)**.** An observation $(\boldsymbol{x}_k, y_k)$, $k \in \{1, 2, \ldots, M\}$, is of *support* for (A.2) if

$$\min_{\theta} \max_{\substack{t\in\{1,\ldots,M\} \\ t\neq k}} \left|y_t - f(\boldsymbol{x}_t)^T\theta\right| < \min_{\theta} \max_{t=1,\ldots,M} \left|y_t - f(\boldsymbol{x}_t)^T\theta\right|,$$

i.e., if its removal improves the solution. ⋆

Thanks to convexity, it is clear that an observation of support $(\boldsymbol{x}_k, y_k)$ must be also active, that is, $|y_k - f(\boldsymbol{x}_k)^T\theta^*| = h^*$. In the present setup, it also holds true with probability 1 that an active observation is of support, so that with probability 1 the observations of support coincide with the active observations.

To show this, suppose that there is an observation, say $(\boldsymbol{x}_M, y_M)$, that is active but not of support. Because it is active, it holds that

$$|y_M - f(\boldsymbol{x}_M)^T\theta^*| = h^*. \tag{A.3}$$

On the other hand, (θ^*, h^*) is with probability 1 also the solution and the optimal value of the problem

$$\min_{\theta} \max_{t=1,\ldots,M-1} |y_t - f(\boldsymbol{x}_t)^T\theta|, \tag{A.4}$$

because (A.4) attains the same optimal value of (A.2) since $(\boldsymbol{x}_M, y_M)$ is not of support and the solution to (A.4) is unique with probability 1 (note that $M - 1 \geq q$). Thus, (θ^*, h^*) depends on $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{M-1}, y_{M-1})$ only. For any given $\boldsymbol{x}_M$ there are just two values of y_M such that (A.3) holds true and since $(\boldsymbol{x}_M, y_M)$ is independent of $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{M-1}, y_{M-1})$, and thereby of θ^*, the probability that y_M takes one of these two values is zero, because $(\boldsymbol{x}_M, y_M)$ is generated according to $\mathbb{P}$, a probability that has density. This gives the sought result that the probability that $(\boldsymbol{x}_M, y_M)$ is active but not of support is zero.

To proceed, we need the following lemma.

Lemma 1. *For any $M \geq q + 1$, the number of observations of support for (A.2) is $q + 1$ with probability 1.* ⋆

Proof of Lemma 1. We first show that the number of observations of support can be less than $q + 1$ with probability zero only.

With probability 1 we have that

$$h^* = \min_{\theta} \max_{t=t_1,\ldots,t_d} \left|y_t - f(\boldsymbol{x}_t)^T\theta\right|, \tag{A.5}$$

where $(\boldsymbol{x}_{t_1}, y_{t_1}), \ldots, (\boldsymbol{x}_{t_d}, y_{t_d})$ are the observations of support. Indeed, with probability 1 the observations of support are the active observations and these latter alone determine the solution (θ^*, h^*) to (A.2) by convexity. Consider now data-sets $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_M, y_M)$ for which d, the number of observations of support, is less than $q + 1$. Since θ has dimension q and $d \leq q$, θ has at least as many components as there are observations of support. Hence, Eq. (A.5) implies that $h^* = 0$ whenever $f(\boldsymbol{x}_{t_1}), \ldots, f(\boldsymbol{x}_{t_d})$ are linearly independent, a situation that occurs with probability 1 (see the proof of Proposition 1). On the other hand, h^* is also given by $h^* = \max_{t=1,\ldots,M} |y_t - f(\boldsymbol{x}_t)^T\theta^*|$, so that $h^* = 0$ implies that

$$y_t = f(\boldsymbol{x}_t)^T\theta^*, \ \text{for all } t = 1, \ldots, M. \tag{A.6}$$

The first part of the proof is now completed by showing that (A.6) can happen with probability zero only. To this end, suppose that the observations of support are the first d observations; then, θ^* in (A.6) depends on observations $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_d, y_d)$ only. The next observation $(\boldsymbol{x}_{d+1}, y_{d+1})$ (this observation is in the set of M observations since $M \geq q+1 > d$) is independent of the first d observations, and thereby of θ^*, and has to satisfy the relation in (A.6), that is, $y_{d+1} = f(\boldsymbol{x}_{d+1})^T\theta^*$. For any value of $\boldsymbol{x}_{d+1}$, only one value of y_{d+1} satisfies this relation, and this happens with probability zero only because $(\boldsymbol{x}_{d+1}, y_{d+1})$ is generated according $\mathbb{P}$, a probability that has a density.

We next show that the number of observations of support cannot be more than $q + 1$.

For the sake of contradiction, suppose that the number of observations of support is greater than $q + 1$, and consider the following $M + 1$ regions in $\mathbb{R}^{q+1}$:

$$F_t = \left\{(\theta, h) \in \mathbb{R}^{q+1} : \ |y_t - f(\boldsymbol{x}_t)^T\theta| \leq h\right\},$$

for $t = 1, \ldots, M$, and

$$F_{M+1} = \left\{(\theta, h) \in \mathbb{R}^{q+1} : \ h < h^*\right\}.$$

For any choice $\{t_1, t_2, \ldots, t_{q+2}\}$ of $q+2$ indexes from the set $\{1, 2, \ldots, M+1\}$, we have that

$$\bigcap_{t=t_1,\ldots,t_{q+2}} F_t \neq \emptyset. \tag{A.7}$$

Indeed, if $\{t_1, t_2, \ldots, t_{q+2}\} \in \{1, 2, \ldots, M\}$, then (θ^*, h^*) is a point in $\bigcap_{t=t_1,\ldots,t_{q+2}} F_t$ and hence (A.7) holds. Suppose instead that one of the indexes $t_1, t_2, \ldots, t_{q+2}$ is $M+1$, say $t_{q+2} = M+1$. Then, we certainly have $\min_\theta \max_{t=t_1,\ldots,t_{q+1}} |y_t - f(\boldsymbol{x}_t)^T\theta| < h^*$ because at least one observation of support is missing in the list of $q+1$ observations with respect to which max is taken (recall that we have supposed that the number of observations of support is greater than $q+1$). This means that $\bigcap_{t=t_1,\ldots,t_{q+1}} F_t$ contains a point $(\bar\theta, \bar h)$ with $\bar h < h^*$. But then, this point is also in F_{M+1} and (A.7) remains proven in this case too.

Since (A.7) holds and since all sets F_t, $t = 1, \ldots, M+1$, are convex, resorting to Helly's theorem (see e.g. Rockafellar, 1970) now yields

$$\bigcap_{t=1,\ldots,M+1} F_t \neq \emptyset.$$

This last relation means that we can find a point (θ^{**}, h^{**}) which is simultaneously in all F_t, $t = 1, \ldots, M$, so that it satisfies $|y_t - f(\boldsymbol{x}_t)^T\theta^{**}| \leq h^{**}$, $t = 1, \ldots, M$, and that is also in F_{M+1}, so that $h^{**} < h^*$. But then this (θ^{**}, h^{**}) would outperform (θ^*, h^*), the optimal solution, and this is a contradiction. This concludes the proof of the lemma. □

As a consequence of the previous results, with probability 1 there are exactly $q+1$ observations, those of support, that lie on the boundary of the optimal minimax layer ($|y_t - f(\boldsymbol{x}_t)^T\theta^*| = h^*$) and that univocally determine it. All other $M-(q+1)$ observations, which are not of support, are strictly inside the optimal layer, and, since they are generated independently of the observations of support, and thereby of θ^*, Condition 1 straightforwardly gives the following property.

Proposition 2 (*Non-Degeneracy Property*). *The probability that for some $t, \tau \in \{1, \ldots, M\}$, $t \neq \tau$, $|y_t - f(\boldsymbol{x}_t)^T\theta^*| = |y_\tau - f(\boldsymbol{x}_\tau)^T\theta^*|$ and $(\boldsymbol{x}_t, y_t)$, $(\boldsymbol{x}_\tau, y_\tau)$ are not both of support for (A.2) is zero.* ⋆

This non-degeneracy property, together with the fact that all the observations of support attain the same value, entails that the number of distinct values in $\{|y_t - f(\boldsymbol{x}_t)^T\theta^*|, \ t = 1, \ldots, M\}$, which is the number of distinct minimax ℓ-th layer IPMs that can be obtained, is equal to $M-q$ with probability 1. When $M = N$, this gives n (see the statement of Theorem 1) equal to $N-q$.

Main derivations.

In this part of the proof, (θ^*, h^*) denotes the solution and optimal value of (3).

Fix a value of $\ell \in \{1, \ldots, N-q\}$. Let $\mathbf{E}[\eta_\ell^k]$ be the kth order moment of the reliability η_ℓ. The proof of Theorem 1 is based on evaluating $\mathbf{E}[\eta_\ell^k]$, for $k = 1, 2, \ldots$, and then deducing the probability distribution F_{η_ℓ} of η_ℓ from the resulting moment problem.

By definition, η_ℓ is the probability that, for fixed (θ^*, h_ℓ^*), one more observation falls in the minimax layer so that, by the independence of observations, η_ℓ^k is the probability that k more observations fall in the layer. Thus, letting $(\boldsymbol{x}_{N+1}, y_{N+1}), \ldots, (\boldsymbol{x}_{N+k}, y_{N+k})$ be k extra observations, η_ℓ^k can be written as

$$\eta_\ell^k = \mathbb{P}^k\left\{ (\boldsymbol{x}_{N+1}, y_{N+1}), \ldots, (\boldsymbol{x}_{N+k}, y_{N+k}) \text{ such that } \left|y_t - f(\boldsymbol{x}_t)^T\theta^*\right| \leq h_\ell^*, t = N+1, \ldots, N+k \right\}.$$

Now, we compute the expectation of η_ℓ^k when θ^* and h^* vary in dependence of the first N random observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ that are used to construct the layer. With the notation $z_m^n = (\boldsymbol{x}_m, y_m), (\boldsymbol{x}_{m+1}, y_{m+1}), \ldots, (\boldsymbol{x}_n, y_n)$ and $Z_m^n = \mathbb{R}^{(p+1)} \times \cdots \times \mathbb{R}^{(p+1)} = \mathbb{R}^{(p+1)\cdot(n-m+1)}$ = domain for z_m^n, it holds that

$$\begin{aligned}
\mathbf{E}[\eta_\ell^k] &= \int_{Z_1^N} \eta_\ell^k \, \mathrm{d}\mathbb{P}^N \\
&= \int_{Z_1^N} \mathbb{P}^k\left\{z_{N+1}^{N+k} \text{ such that } \left|y_t - f(\boldsymbol{x}_t)^T\theta^*\right| \leq h_\ell^*, \right. \\
&\qquad \left. t = N+1, \ldots, N+k\right\} \mathrm{d}\mathbb{P}^N \\
&= [\mathbb{I}_A = \text{indicator function of set } A] \\
&= \int_{Z_1^N} \left[\int_{Z_{N+1}^{N+k}} \mathbb{I}_{\{|y_t - f(\boldsymbol{x}_t)^T\theta^*| \leq h_\ell^*, t=N+1,\ldots,N+k\}} \, \mathrm{d}\mathbb{P}^k \right] \mathrm{d}\mathbb{P}^N \\
&= \int_{Z_1^{N+k}} \mathbb{I}_{\{|y_t - f(\boldsymbol{x}_t)^T\theta^*| \leq h_\ell^*, t=N+1,\ldots,N+k\}} \, \mathrm{d}\mathbb{P}^{N+k}.
\end{aligned}$$

Now, let $S = \{t_1, \ldots, t_N\}$ be a generic subset of N indexes from $\{1, 2, \ldots, N+k\}$ and let $\mathcal{S}$ be the family of all possible choices of S ($\mathcal{S}$ contains $\binom{N+k}{N}$ elements). Moreover, define $\bar S = \{1, 2, \ldots, N+k\} - S$.

Due to the i.i.d. nature of the observations, each group of N observations has identical statistical properties as any other group. Therefore, if we indicate by θ_S^* and h_S^* the optimal solution and the optimal value of problem

$$\min_\theta \max_{t\in S} \left|y_t - f(\boldsymbol{x}_t)^T\theta\right|,$$

and by $h_{\ell,S}^*$ the ℓ-th highest value in the set $\{|y_t - f(\boldsymbol{x}_t)^T\theta_S^*|, \ t \in S\}$, we have that

$$\begin{aligned}
&\int_{Z_1^{N+k}} \mathbb{I}_{\{|y_t - f(\boldsymbol{x}_t)^T\theta^*| \leq h_\ell^*, t=N+1,\ldots,N+k\}} \, \mathrm{d}\mathbb{P}^{N+k} \\
&= \int_{Z_1^{N+k}} \mathbb{I}_{\left\{|y_t - f(\boldsymbol{x}_t)^T\theta_S^*| \leq h_{\ell,S}^*, t\in\bar S\right\}} \, \mathrm{d}\mathbb{P}^{N+k}, \quad \forall S \in \mathcal{S}.
\end{aligned} \tag{A.9}$$

From (A.8) and (A.9) we obtain that

$$\begin{aligned}
&\mathbf{E}[\eta_\ell^k] \\
&= \frac{1}{\binom{N+k}{N}} \sum_{S\in\mathcal{S}} \int_{Z_1^{N+k}} \mathbb{I}_{\left\{|y_t - f(\boldsymbol{x}_t)^T\theta_S^*| \leq h_{\ell,S}^*, t\in\bar S\right\}} \, \mathrm{d}\mathbb{P}^{N+k} \\
&= \frac{1}{\binom{N+k}{N}} \int_{Z_1^{N+k}} \sum_{S\in\mathcal{S}} \mathbb{I}_{\left\{|y_t - f(\boldsymbol{x}_t)^T\theta_S^*| \leq h_{\ell,S}^*, t\in\bar S\right\}} \, \mathrm{d}\mathbb{P}^{N+k}.
\end{aligned} \tag{A.10}$$

The computation of $\mathbf{E}[\eta_\ell^k]$ is now completed by showing that the integrand in (A.10) is with probability 1 constant and equal to $\binom{N+k-(q+\ell)}{k}$ so that

$$\mathbf{E}[\eta_\ell^k] = \frac{\binom{N+k-(q+\ell)}{k}}{\binom{N+k}{N}}, \quad k = 1, 2, \ldots \tag{A.11}$$

For fixed observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{N+k}, y_{N+k})$, the quantity $\sum_{S\in\mathcal{S}} \mathbb{I}_{\left\{|y_t - f(\boldsymbol{x}_t)^T\theta_S^*| \leq h_{\ell,S}^*, t\in\bar S\right\}}$ counts the number of choices of S such that the ℓ-th layer constructed from the observations with indexes in S contains all the remaining observations with indexes in $\bar S$. These choices of S are those such that (θ_S^*, h_S^*) is equal to the solution and optimal value $(\theta_{S\cup\bar S}^*, h_{S\cup\bar S}^*)$ of the problem with all $N+k$ observations,

$$\min_\theta \max_{t=1,\ldots,N+k} \left|y_t - f(\boldsymbol{x}_t)^T\theta\right|, \tag{A.12}$$

and $h_{\ell,S}^*$ is equal to $h_{\ell,S\cup\bar S}^*$, which is defined as the ℓ-th highest value in the set $\{|y_t - f(\boldsymbol{x}_t)^T\theta_{S\cup\bar S}^*|, \ t = 1, \ldots, N+k\}$. The event

that $(\theta^*_S, h^*_S) = (\theta^*_{S\cup\bar{S}}, h^*_{S\cup\bar{S}})$ and $h^*_{\ell,S} = h^*_{\ell,S\cup\bar{S}}$ happens if and only if $\bar{S}$ does not contain any of the observations such that

$$|y_t - f(\boldsymbol{x}_t)^T\theta^*_{S\cup\bar{S}}| \geq h^*_{\ell,S\cup\bar{S}}, \tag{A.13}$$

i.e., the observations of support for (A.12), which are all needed to determine $(\theta^*_{S\cup\bar{S}}, h^*_{S\cup\bar{S}})$, and the observations strictly inside the minimax layer that are needed to determine $h^*_{\ell,S\cup\bar{S}}$. Using Lemma 1 and Proposition 2, we immediately conclude that with probability 1 (A.13) is satisfied by $q+\ell$ observations in $S\cup\bar{S}$. Thus, the subset of k indexes $\bar{S}$ has to be chosen from a set of $N+k-(q+\ell)$ indexes, and the number of choices is $\binom{N+k-(q+\ell)}{k}$. Thus

$$\sum_{S\in\mathcal{S}} \mathbb{I}_{\left\{|y_t - f(\boldsymbol{x}_t)^T\theta^*_{\ell,S}| \leq h^*_{\ell,S}, t\in\bar{S}\right\}} = \binom{N+k-(q+\ell)}{k}$$

with probability 1, and (A.11) remains proven.

To conclude the proof, note now that $\mathbf{E}[\eta_\ell^k]$ can also be written as $\int_0^1 z^k \, \mathrm{d}F_{\eta_\ell}(z)$, where F_{η_ℓ} is the probability distribution of η_ℓ. Hence, (A.11) becomes

$$\int_0^1 z^k \, \mathrm{d}F_{\eta_\ell}(z) = \frac{\binom{N+k-(q+\ell)}{k}}{\binom{N+k}{N}}, \quad k = 1, 2, \ldots \tag{A.14}$$

The distribution function

$$F_{\eta_\ell}(z) = \sum_{i=0}^{q+\ell-1} \binom{N}{i} (1-z)^i z^{N-i}$$

(which gives $\mathrm{d}F_{\eta_\ell}(z) = (q+\ell)\binom{N}{q+\ell}(1-z)^{q+\ell-1}z^{N-q-\ell}\,\mathrm{d}z$) satisfies the infinite system of Eqs. (A.14) and the theorem statement is finally proved by noting that (A.14) defines a so called moment problem that admits a unique solution (see e.g. Corollary 1, §12.9, Chapter II of Shiryaev, 1996). □

References

Alamo, T., Tempo, R., Luque, A., & Ramirez, D. R. (2015). Randomized methods for design of uncertain systems: Sample complexity and sequential algorithms. *Automatica*, *52*, 160–172.

Appa, G., & Smith, C. (1973). On L_1 and Chebyshev estimation. *Journal of Mathematical Programming*, *5*, 73–87.

Armstrong, R., & Kung, D. (1979). Min-max estimates for a linear multiple regression problem. *Applied Statistics*, *28*, 93–100.

Arthanari, T., & Dodge, Y. (1993). *Wiley Classics Library, Mathematical Programming in Statistics*. New York, NY: John Wiley and Sons.

Barrodale, I., & Phillips, C. (1975). Solution of an over-determined system of linear equations in the Chebyshev norm. *ACM Transactions on Mathematical Software*, *1*, 264–270.

Birkes, D., & Dodge, Y. (1993). *Alternative Methods of Regression*. New York, NY: John Wiley and Sons.

Calafiore, G. C. (2010). Learning noisy functions via interval models. *Systems & Control Letters*, *59*(7), 404–413. http://dx.doi.org/10.1016/j.sysconle.2010.05.003.

Calafiore, G. C., & Campi, M. C. (2006). The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, *51*(5), 742–753.

Campi, M. C., Calafiore, G., & Garatti, S. (2009). Interval predictor models: Identification and reliability. *Automatica*, *45*(2), 382–392.

Campi, M. C., & Carè, A. (2013). Random convex programs with L1-regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, *51*(5), 3532–3557.

Campi, M. C., Csáji, B. Cs., Garatti, S., & Weyer, E. (2012). Certified system identification: towards distribution-free results. In *16th IFAC Symposium on System Identification IFAC Proceedings Volumes*, *45*(16), 245–255. http://dx.doi.org/10.3182/20120711-3-BE-2027.00428.

Campi, M. C., & Garatti, S. (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, *19*(3), 1211–1230. http://dx.doi.org/10.1137/07069821X.

Campi, M. C., & Garatti, S. (2011). A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, *148*(2), 257–280.

Campi, M. C., & Garatti, S. (2018). *Introduction to the scenario approach, Vol. 26*. SIAM.

Carè, A., Garatti, S., & Campi, M. (2014). FAST - Fast algorithm for the scenario technique. *Operations Research*, *62*(3), 662–671.

Carè, A., Garatti, S., & Campi, M. C. (2015). Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization*, *25*(4), 2061–2080.

Chebyshev, P. (1854). *Mémoires présentés a l'Académie Impériale des Sciences de St. Pétersbourg par divers savants*, *7*, 539–568.

Cheney, E. (1999). *Introduction to Approximation Theory*. Providence, RI: AMS.

Crespo, L. G., Giesy, D. P., & Kenny, S. P. (2014). Interval predictor models with a formal characterization of uncertainty and reliability. In *53rd IEEE conference on decision and control* (pp. 5991–5996). http://dx.doi.org/10.1109/CDC.2014.7040327.

Crespo, L. G., Kenny, S. P., & Giesy, D. P. (2015). Random predictor models for rigorous uncertainty quantification. *International Journal for Uncertainty Quantification*, *5*(5), 469–489.

Crespo, L. G., Kenny, S. P., & Giesy, D. P. (2016). Interval predictor models with a linear parameter dependency. *Journal of Verification, Validation and Uncertainty Quantification*, *1*(2), 1–10.

Euler, L. (1749). Pièce qui a remporté le prix de l'Académie Royale des Sciences en 1748, sur l'inégalités du movement de Saturn et de Jupiter. In *Commentationes Astronomicae I: Vol. II 25, Leonhardi Euleri Opera Omnia* (pp. 45–157). Turici, 1960: Orel Fussli.

Fourier, J. (1824). *Histoire de l'Academie Royale des Sciences Paris, 29ff*, 47–55.

Garatti, S., & Campi, M. C. (2009). L-infinity layers and the probability of false prediction. In *Proceedings of the 15th IFAC symposium on system identification*.

Garatti, S., & Campi, M. C. (2013). Modulating robustness in control design: Principles and algorithms. *IEEE Control Systems*, *33*(2), 36–51. http://dx.doi.org/10.1109/MCS.2012.2234964.

Haar, A. (1918). Die Minkowskische Geometrie und die Annäherung an stetige Funktionen. *Mathematische Annalen*, *78*, 294–311.

Harter, H. (1975). The method of least squares and some alternatives – part III. *International Statistical Reviews*, *43*, 1–44.

Harter, H. (1982). Minimax method. In *Encyclopedia of Statistical Sciences, Vol. 4* (pp. 514–516). John Wiley & Sons.

Jaulin, L., Kieffer, M., Braems, I., & Walter, É. (2001). Guaranteed non-linear estimation using constraint propagation on sets. *International Journal of Control*, *74*(18), 1772–1782.

Jaulin, L., Kieffer, M., Didrit, O., & Walter, É. (2001). *Applied interval analysis with examples in parameter and state estimation, robust control and robotics*. Springer London Ltd.

Karst, O. (1958). Linear curve fitting using least deviation. *Journal of the American Statistical Association*, *53*, 118–132.

Kieffer, M., Jaulin, L., & Walter, É. (2002). Guaranteed recursive non-linear state bounding using interval analysis. *International Journal of Adaptive Control and Signal Processing*, *16*(3), 193–218.

Lacerda, M. J., & Crespo, L. G. (2017). Interval predictor models for data with measurement uncertainty. In *2017 American control conference* (pp. 1487–1492). http://dx.doi.org/10.23919/ACC.2017.7963163.

Laplace, P. (1783). *Mémoires de l'Académie Royale des Sciences*, *1783*, 17–46.

Laplace, P. (1793). *Mémoires de l'Académie Royale des Sciences*, *1789*, 1–87.

Milanese, M., Norton, J., Piet-Lahanier, H., & Walter, É. (2013). *Bounding approaches to system identification*. Springer Science & Business Media.

Patelli, E., Broggi, M., Tolo, S., & Sadeghi, J. (2013). Cossan software: A multidisciplinary and collaborative software for uncertainty quantification. In *Proceedings of the 2nd ECCOMAS thematic conference on uncertainty quantification in computational sciences and engineering, UNCECOMP 2017*.

Planitz, M., & Gates, J. (1991). Strict discrete approximation in the L_1 and L_∞ norms. *Applied Statistics*, *40*, 113–122.

Rockafellar, R. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Shiryaev, A. (1996). *Probability*. New York, NY, USA: Springer.

Wagner, H. (1959). Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, *54*(285), 206–212.

Walter, É., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. Springer Verlag.

Zhang, Y. (1993). Primal-dual interior point approach for computing L_1 solutions and L_∞ solutions of over-determined systems. *Journal of Optimization Theory and Applications*, *77*, 323–341.

Simone Garatti is Associate Professor at the Dipartimento di Elettronica ed Informazione of the Politecnico di Milano, Milan, Italy. He received the Laurea degree and the Ph.D. in Information Technology Engineering in 2000 and 2004, respectively, both from the Politecnico di Milano. In 2003, he was visiting scholar at the Lund University of Technology, Lund, Sweden, in 2006 at the University of California San Diego (UCSD), San Diego, CA, USA, and in 2007 at the Massachusetts Institute of Technology and the Northeastern University, Boston, MA, USA. He is member of the IFAC Technical Committee on Modeling, Identification and Signal Processing and of the IEEE Technical Committees on Robust and Complex Systems and on Systems Identification and Adaptive Control. He is also member of both the EUCA Conference Editorial Board and of the IEEE-CSS Conference Editorial Board. His research interests include data-driven optimization and decision-making, stochastic optimization for problems in systems and control, system identification, model quality assessment, and uncertainty quantification.

Marco Claudio Campi is Professor of Automatic Control at the *University of Brescia*, Italy. He has held visiting and teaching appointments at the *Australian National University*, Canberra, Australia; the *University of Illinois at Urbana–Champaign*, USA; the *Centre for Artificial Intelligence and Robotics*, Bangalore, India; the *University of Melbourne*, Australia; the *Kyoto University*, Japan; the *Texas A&M University*, USA; the *NASA Langley Research Center*, Hampton, Virginia. USA. Marco Campi is the chair of the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP), and has been in various capacities on the Editorial Board of Automatica, Systems and Control Letters and the European Journal of Control. He is a recipient of the "Giorgio Quazza" prize, and, in 2008, he received the IEEE CSS George S. Axelby outstanding paper award for the article The Scenario Approach to Robust Control Design. He has delivered plenary and semi-plenary addresses at major conferences including SYSID, MTNS, and CDC. Currently he is a distinguished lecturer of the Control Systems Society. Marco Campi is a Fellow of IEEE, a member of IFAC, and a member of SIDRA. The research interests of Marco Campi include: *system identification*, *stochastic systems*, *randomized methods, adaptive and data-based control*, *robust optimization*, and *learning theory*.

Algo Carè received the Ph.D. degree in informatics and automation engineering from the University of Brescia, Brescia, Italy, in 2013. He is currently a Research Fellow with the Department of Information Engineering, University of Brescia. He spent two years at The University of Melbourne, Melbourne, VIC, Australia, as a Research Fellow in system identification with the Department of Electrical and Electronic Engineering. Dr. Carè was a recipient of a two-year ERCIM Fellowship in 2016 that he spent at the Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences (MTA), Budapest, Hungary, and at the Multiscale Dynamics Group, National Research Institute for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. He received the triennial Stochastic Programming Student Paper Prize by the Stochastic Programming Society for the period 2013–2016. His current research interests include data-driven decision methods, system identification, and learning theory.