POLITECNICO DI TORINO
Repository ISTITUZIONALE

Prediction error quantification through probabilistic scaling

Original
Prediction error quantification through probabilistic scaling / Mirasierra, V., Mammarella, M., Dabbene, F., Alamo, T.. - In: IEEE CONTROL SYSTEMS LETTERS. - ISSN 2475-1456. - ELETTRONICO. - 6:(2022), pp. 1118-1123. [10.1109/lcsys.2021.3087361]

Availability:
This version is available at: 11583/2915348 since: 2021-07-27T13:12:31Z

Publisher:
IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS

Published
DOI:10.1109/lcsys.2021.3087361

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

14 September 2026

Prediction error quantification through probabilistic scaling

Victor Mirasierra[1], Martina Mammarella[2], Fabrizio Dabbene[2,*], Teodoro Alamo[1]

***Abstract*— In this paper, we address the probabilistic error quantification of a general class of prediction methods. We consider a given prediction model and show how to obtain, through a sample-based approach, a probabilistic upper bound on the absolute value of the prediction error. The proposed scheme is based on a probabilistic scaling methodology in which the number of required randomized samples is independent of the complexity of the prediction model. The methodology is extended to address the case in which the probabilistic uncertain quantification is required to be valid for every member of a finite family of predictors. We illustrate the results of the paper by means of a numerical example.**

***Index Terms*— Randomized algorithms, Estimation, Machine learning, Model Validation, Uncertain systems.**

I. Introduction and Problem Formulation

QUANTIFYING the error related to the process of approximating a set of given data with a prescribed prediction method represents a fundamental requirement, which has given rise to an entire research area known as *Uncertainty Quantification*, see e.g. [1], [2] and references therein.

Motivated by this necessity, methods for directly constructing predictive models with prescribed robustness guarantees have recently gained popularity. For instance, [3] presents several methods based on interval analysis to construct intervals which are guaranteed to contain the true value, under the assumption of deterministically bounded noise. Similarly, data-based approaches exploiting the availability of random samples, providing probabilistic guarantees, are being developed. These methods extend classical quantile regression [4]. In particular, we point out the probabilistic interval predictions proposed in [5], [6].

All these methods require to design (or re-design) the estimator using a specific ad-hoc model. However, this approach may not result practical when data-analysts have already constructed a model exploiting a "preferred" technique (e.g., one based on support vector machines (SVM)) and they want to assess, before deployment, the actual uncertainty of their model.

For this reason, researchers have started investigating post-processing methods for quantifying the uncertainty of a *given* predictor. This means that no new methodology is proposed for constructing a regression model but that such a model (or a family of candidate ones) is given. This philosophy is exactly the one pursued in uncertainty quantification methods, see e.g. the recent approaches based on polynomial chaos [1], or the *conformal predictors* [7]. These methods typically use additional *validation* (or *calibration*) data to determine precise levels of confidence in new predictions [8].

In this paper, we move a step further in this direction and present sampling-based techniques for assessing the corresponding error in a *computationally* efficient way. Indeed, this novel approach, extending recent results on probabilistic scaling, e.g., [9], [10], requires a number of randomized samples independent of the complexity of the prediction model (i.e. the dimension of the regressor).

In particular, we consider that given $x \in \mathbb{R}^{n_x}$, an estimation $\hat{y}$ for $y \in \mathbb{R}$ is provided by operator $T : \mathbb{R}^{n_x} \to \mathbb{R}$. That is,

$$\hat{y} = T(x).$$

We assume that the operator T is a given predictive model that has been designed by means of any modelling methodology (first principles, linear regression, SVM regression, neural network, etc.).

We want to provide a probabilistic bound on the prediction error. More formally, we consider the random vector $w = (x, y) \in \mathbb{R}^{n_x} \times \mathbb{R} \subseteq \mathbb{W}$, with stationary probability distribution $\Pr_{\mathbb{W}}$, and we aim at constructing a function $\rho : \mathbb{R}^{n_x} \to \mathbb{R}$ such that, with probability no smaller than $(1 - \delta)$,

$$\Pr_{\mathbb{W}}\{|y - T(x)| \leq \rho(x)\} \geq 1 - \varepsilon.$$

The method relies on the possibility of accessing random data, that is, observations couples $w = (x, y)$. Note that these are new data not used to construct $T(\cdot)$. We show that the sample complexity of the proposed techniques (i.e. the number of observations required) does not depend on the chosen regression model but only on the desired probabilistic levels.

The remainder of the paper is structured as follows. In Section II we propose a first simple result, which allows to obtain an initial probabilistic bound on the prediction error via probabilistic maximization, given a predictive model. The obtained bound, which can be computed by means of a simple algorithm, is independent of the given x. Section III, focuses on including those situations in which the expected size of the error does depend on x, and propose a probabilistic bound conditioned to x. This approach is extended in Section IV to the case when a "family" of candidates estimators is considered. Finally, in Section V, we included the possibility of designing the predictors by means of kernel methods, providing at the same time also a measure of the expected size of the error. All these approaches are illustrated by means of

This work was supported by the Agencia Estatal de Investigación (AEI)-Spain under Grant PID2019-106212RB-C41/AEI/10.13039/501100011033 and partially funded by the Italian IIT and MIUR within the 2017 PRIN (N. 2017S559BB).

[1] Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, Sevilla, Spain.

[2] CNR-IEIIT; c/o Politecnico di Torino; Corso Duca degli Abruzzi 24, Torino; Italy.

[*] corresponding author *fabrizio.dabbene@ieiit.cnr.it* (F. Dabbene).

a running numerical example, and conclusions are drawn in Section VII.

Notation

Given an integer N, $[N]$ denotes the integers from 1 to N. Given $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the greatest integer no larger than x and $\lceil x \rceil$ the smallest integer no smaller than x. Given integers k, N, with $0 \leq k \leq N$, and parameter $\varepsilon \in [0,1]$, the Binomial cumulative distribution function is denoted as

$$\mathbf{B}(k; N, \varepsilon) \doteq \sum_{i=0}^{k} \left(\begin{array}{c} N \\ i \end{array} \right) \varepsilon^i (1-\varepsilon)^{N-i}.$$

Given the measurable function $g(x,y)$ and the probability distribution $\mathrm{Pr}_{\mathbb{W}}$, we denote by $\mathbb{E}_{\mathbb{W}}\{g(x,y)\}$ the expected value of the random variable $g(x,y)$ and by $\mathbb{E}_{\mathbb{W}}\{g(x,y)|x\}$ the expected value of $g(x,y)$ conditioned to x. The following definition is borrowed from the field of order statistics [9].

Definition 1 (Generalized Max): Given a collection of N scalars $Q = \{q_1, q_2, \ldots, q_N\} = \{q_i\}_{i=1}^N$, and an integer $r \in [N]$, we say that $q_r^+ \in Q$ is the r-largest value of Q if there is no more than $r-1$ elements of Q strictly larger than q_r^+.

Hence, q_1^+ denotes the largest value in Q, q_2^+ the second largest one, and so on until q_N^+, which is equal to the smallest one. We also use the alternative notation $q_r^+ = \max^{(r)}\{q_i\}_{i=1}^N$.

II. Uncertainty Quantification using Probabilistic Maximization

In this section we present an initial probabilistic bound for the error $y - T(x)$ based on probabilistic maximization. Suppose that we draw N independent and identically distributed (i.i.d.) samples $\{(x_i, y_i)\}_{i=1}^N$ according to distribution $\mathrm{Pr}_{\mathbb{W}}$, and we denote as

$$q_i \doteq |y_i - T(x_i)|, \quad i \in [N]$$

the absolute value of the corresponding prediction errors. A well established result [11] shows that the largest value in the sequence $\{q_i\}_{i=1}^N$, i.e., q_1^+, provides a probabilistic upper bound on the random variable $q = |y - T(x)|$. Formally, given $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, [11, Theorem 1] states that if

$$N \geq \frac{1}{\varepsilon} \log(\frac{1}{\delta}) \tag{1}$$

then, with probability no smaller than $1-\delta$,

$$\mathrm{Pr}_{\mathbb{W}}\left\{q > q_1^+\right\} \leq \varepsilon.$$

It is immediate to observe that this result provides a first simple probabilistic scheme for uncertainty quantification: If N i.i.d. samples $\{(x_i, y_i)\}_{i=1}^N$ are drawn according to $\mathrm{Pr}_{\mathbb{W}}$, with N satisfying (1), then with probability at least $1-\delta$

$$\mathrm{Pr}_{\mathbb{W}}\left\{|y - T(x)| \leq q_1^+\right\} \geq 1 - \varepsilon.$$

We notice that the required sample complexity (i.e. the number of samples N) depends only on ε and δ. Moreover, no specific assumptions are required on $T(x)$ or $\mathrm{Pr}_{\mathbb{W}}$.

However, we also note that this scheme may provide extremely conservative results, especially if the support of the random variable $q = |y - T(x)|$ is not finite and N is large. In fact, suppose that $y - T(x)$ is a zero mean Gaussian random variable. Then, the probabilistic upper bound obtained from q_1^+ will be too conservative if one of the samples $q_i = |y_i - T(x_i)|$ departs considerably from zero, which occurs with a probability that increases with N. We conclude that only relying on the largest observed value of $|y - T(x)|$ hinders the computation of sharp probabilistic bounds, especially for small values of ε and δ, leading to a large number of samples N.

In order to circumvent this issue, we resort to the following result [9, Property 3], which states how to obtain a probabilistic upper bound of a random scalar variable by means of the notion of generalized max (see Definition 1).

Property 1: Given $\varepsilon \in (0,1)$, $\delta \in (0,1)$ and $r \geq 1$, let $N \geq r$ be such that

$$\mathbf{B}(r-1; N, \varepsilon) \leq \delta. \tag{2}$$

Suppose that $q \in \mathbb{W} \subseteq \mathbb{R}$ is a random scalar variable with probability distribution $\mathrm{Pr}_{\mathbb{W}}$. Draw N i.i.d. samples $\{q_i\}_{i=1}^N$ from distribution $\mathrm{Pr}_{\mathbb{W}}$. Then, with a probability no smaller than $1-\delta$,

$$\mathrm{Pr}_{\mathbb{W}}\{q > \max^{(r)}\{q_i\}_{i=1}^N\} \leq \varepsilon. \tag{3}$$

Remark 1 (On Property 1): This result is proved in [9] using techniques from the field of order statistics [12]. As discussed in [9], this result may be alternatively derived by applying the scenario approach with discarded constraints [13], [14]. Adaptations of this result have been used in the context of chance constrained optimization [15], [16], and stochastic model predictive control [10], [17], [18].

Several questions arise when trying to apply Property 1 to the probabilistic error quantification problem:

Choice of N: It was proved in [19, Corollary 1] that the constraint $\mathbf{B}(r-1; N, \varepsilon) \leq \delta$ holds if

$$\varepsilon N \geq r - 1 + \log\frac{1}{\delta} + \sqrt{2(r-1)\log\frac{1}{\delta}}. \tag{4}$$

Thus, given r, δ, and ε, the sample size N can be obtained as the smallest integer N satisfying (4). Another possibility is to compute, by means of a numerical procedure, the smallest integer N satisfying $\mathbf{B}(r-1; N, \varepsilon) \leq \delta$.

Choice of δ: Since $1-\delta$ determines the probability of the satisfaction of the probabilistic constraint (3), it is important to choose δ sufficiently close to zero. In view of (4), we have that N grows logarithmically with $\frac{1}{\delta}$. This implies that significantly small values of δ (say $\delta = 10^{-6}$) can be used without an excessive impact in the number of samples N.

Choice of r: If r is chosen to be too small, then the obtained probabilistic bounds might turn to be too conservative because the obtained upper bound would be determined by a reduced number of possible *extreme* values. We notice from (4) that the larger the value of r, the larger the number of required samples N. We also derive from (4) that $\frac{r-1}{N} < \varepsilon$. A reasonable choice for r with an appropriate trade off between sample complexity N and sharpness of the results is $r = \left\lfloor \frac{\varepsilon N}{2} \right\rfloor$.

Choice of ε: Parameter ε determines the size of the confidence interval in the uncertainty quantification process. In

uncertainty quantification, values of ε much smaller than 0.05 are not frequent.

We now state a result, which has been presented in a different context in [10] and [16], that shows how to obtain N in such a way that (2) is satisfied for the particular choice $r = \lfloor \frac{\varepsilon N}{2} \rfloor$.

Lemma 1: Given $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, suppose that $N \geq \frac{7.47}{\varepsilon} \ln \frac{1}{\delta}$ and $r = \lfloor \frac{\varepsilon N}{2} \rfloor$. Then $\mathbf{B}(r-1; N, \varepsilon) \leq \delta$.

Proof: See the appendix of [20], where it is proved that the claim holds if $N \geq \frac{(1+\sqrt{3})^2}{\varepsilon} \ln(\frac{1}{\delta})$. The result follows from $(1+\sqrt{3})^2 < 7.47$.

Property 1, along with the previous discussion on the choice of r, leads to Algorithm 1, which provides a simple procedure to compute a probabilistic bound on the prediction error $y - T(x)$.

Algorithm 1 Probabilistic Fixed-Size Bound on Error

1: Given a predictor $T : \mathbb{R}^{n_x} \to \mathbb{R}$, and probability levels $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, choose

$$N \geq \frac{7.47}{\varepsilon} \ln \frac{1}{\delta} \quad \text{and} \quad r = \left\lfloor \frac{\varepsilon N}{2} \right\rfloor. \tag{5}$$

2: Draw N i.i.d. samples $\{(x_i, y_i)\}_{i=1}^N$ according to $\Pr_{\mathbb{W}}$.
3: Compute $q_i = |y_i - T(x_i)|$, $i \in [N]$.
4: Return $\rho = \max^{(r)}\{q_i\}_{i=1}^N$ as the probabilistic upper bound for $|y - T(x)|$.

The probabilistic guarantees of the upper bound generated with Algorithm 1 are provided in the next corollary.

Corollary 1: The output ρ of Algorithm 1 satisfies, with probability no smaller than $1-\delta$, $\Pr_{\mathbb{W}}\{|y - T(x)| > \rho\} \leq \varepsilon$.

Proof: From Lemma 1, we have that the values for N and r obtained in step 1 of Algorithm 1 guarantee that $B(r-1; N, \varepsilon) \leq \delta$. Thus, we conclude from Property 1 that $\rho = \max^{(r)}\{|y_i - T(x_i)|\}_{i=1}^N$ satisfies, with probability no smaller than $1-\delta$, $\Pr_{\mathbb{W}}\{|y - T(x)| > \rho\} \leq \varepsilon$. □

Numerical example: Algorithm 1

Consider the function

$$y = f(x, n_1, n_2) = (10 + n_1)x + 10\sin(4x) + 5 + n_2. \tag{6}$$

We assume that x is a random scalar with uniform distribution in $[-2.5, 2.5]$ and n_1, n_2 are random scalars drawn from zero-mean Gaussian distributions with variances 7 and 3, respectively. Suppose that the optimal predictor $T(x) = 10x + 10\sin(4x) + 5$ for the random scalar $y = f(x, n_1, n_2)$ is available[1]. We fix the probabilistic levels to $\varepsilon = 0.05$ and $\delta = 10^{-6}$, which leads to $N = 2,065$ and $r = 51$ (see step 1 of Algorithm 1). We draw N i.i.d. samples $\{(x_i, y_i)\}_{i=1}^N$ and obtain $\rho = 10.77$. Thus, according to Corollary 1, with probability no smaller than $1-\delta$, $\Pr_{\mathbb{W}}\{|y - T(x)| > \rho\} \leq \varepsilon$. We notice that for this example it is not difficult to obtain the sharpest probabilistic bounds for $y - T(x)$ corresponding to a given x. It suffices to notice that given x, $y - T(x)$ is a zero-mean Gaussian random variable with variance $7x^2 + 3$.

[1]We address the problem of determining predictor $T(\cdot)$ in Section V.

Thus, using standard confidence interval analysis for a scalar Gaussian variable, we obtain that

$$\Pr_{\mathbb{W}}\{|y - T(x)| > 1.96\sqrt{7x^2 + 3}\} \leq 0.05.$$

Figure 1 shows, for a new validation set of N i.i.d. samples, the (fixed size) probabilistic bounds for y provided by Algorithm 1 (i.e. $\Pr_{\mathbb{W}}\{y \in [T(x) - \rho, T(x) + \rho]\} \geq 1 - \varepsilon$), along with the exact probabilistic bounds. We notice that Algorithm 1 fails to capture the varying size of the exact probabilistic bounds. We address this issue in the next sections.

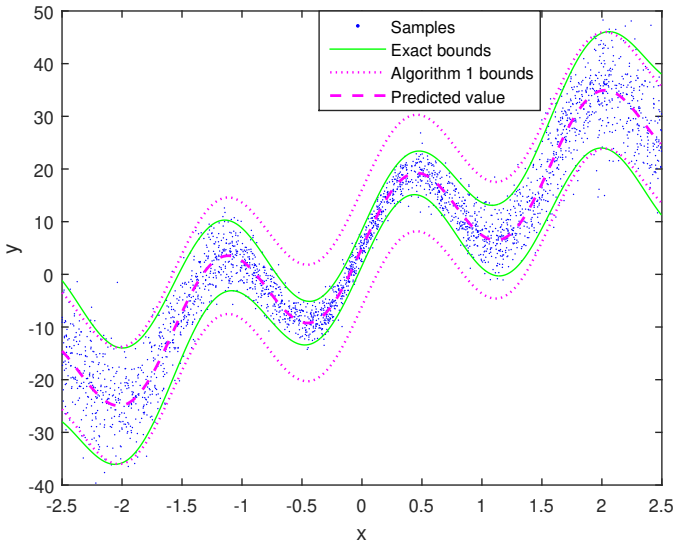

Fig. 1. Numerical example: comparison between the probabilistic bounds obtained with Algorithm 1 with the exact ones. We notice that the size of the interval bounds provided by Algorithm 1 are independent of $\boldsymbol{x}$.

III. Conditioned Uncertainty Quantification

The simplicity of Algorithm 1 comes at a price: *the obtained upper bound does not depend on* x. Clearly, this is not an issue if the error $e = y - T(x)$ is independent of x. However, in many situations, the expected size of the error does depend on x. For example, the prediction errors are often correlated with the size of the predicted variable, which in turn is correlated with x. From here, we infer that information on the expected error can often be obtained from x.

Under some strong assumptions, the probability distribution of $y - T(x)$ conditioned to x, can be computed in an explicit way. This is the case, for example, when $T(x)$ is obtained by means of Gaussian process regression [21, §2] or when exponential models are employed [22, §4]. However, we notice that, although these kernel-based approaches can indeed provide estimations of the conditioned expectation

$$\sigma^2(x) = \mathbb{E}_{\mathbb{W}}\{(y - T(x))^2 \,|\, x\},$$

the accuracy of the estimations will depend on the satisfaction of the underlying assumptions (i.e. Gaussian process and exponential model, respectively) and the adequate selection of the kernels (along with their hyper-parameters) used to obtain $T(x)$. There are other possibilities to obtain conditioned error quantification, like sensitivity analysis, techniques based on Fisher information matrix, bootstrapping, etc. [2], [1].

We also mention here Parzen method [23], which serves to estimate the probability density function of a random variable. More general multivariate kernel-based generalizations are also available (see e.g., [24]). In these methods, an estimation $\hat{\sigma}(x)$ of $\sigma(x)$ is obtained from

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{M}(y_i - T(x_i))^2\Gamma(x, x_i)}{\sum_{i=1}^{M}\Gamma(x, x_i)}, \tag{7}$$

where $\Gamma : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$ is an appropriately chosen function and $\{(x_i, y_i)\}_{i=1}^{M}$ are i.i.d. samples drawn from $\mathrm{Pr}_{\mathbb{W}}$. Under non very restrictive constraints [23], [24], the provided estimation $\hat{\sigma}(x)$ converges to the actual value $\sigma(x)$ as M tends to infinity.

For a fixed value of x, $d = (y - T(x))^2$ is a random non-negative variable with expectation $\sigma^2(x)$. Thus, we can resort to the Markov inequality [25], [26], to obtain

$$\mathrm{Pr}_{\mathbb{W}}\{d \geq \xi\sigma^2(x)\} \leq \frac{1}{\xi},\ \forall \xi > 0.$$

Thus, choosing $\xi = \frac{1}{\varepsilon}$, we obtain $\mathrm{Pr}_{\mathbb{W}}\left\{d \geq \frac{\sigma^2(x)}{\varepsilon}\right\} \leq \varepsilon$. Equivalently,

$$\mathrm{Pr}_{\mathbb{W}}\left\{|y - T(x)| \geq \frac{\sigma(x)}{\sqrt{\varepsilon}}\right\} \leq \varepsilon.$$

The obtained probabilistic upper bound suffers from the following two limitations: (i) generally, $\sigma(x)$ is unknown and only a rough estimation is available (as the ones commented before), and (ii) Markov inequality yields overly conservative results in many situations [26]. A meaningful exception to this is when the errors $y - T(x)$ are of Gaussian nature. In this case, using the Chi-squared distribution [25], sharp probabilistic bounds of the form $\mathrm{Pr}_{\mathbb{W}}\{|y - T(x)| \geq \gamma_\varepsilon\sigma(x)\} = \varepsilon$, can be obtained.

In order to avoid these limitations, we can again resort to probabilistic maximization. Suppose that an estimation $\hat{\sigma}(x)$ of $\sigma(x)$ is available. Suppose also that $\hat{\sigma}(x) > 0$, for all $x \in \mathbb{R}^{n_x}$. We could define the scaling factor γ as

$$\gamma = \frac{|y - T(x)|}{\hat{\sigma}(x)}.$$

With this definition, any probabilistic upper bound $\bar{\gamma}$ on γ would provide a probabilistic upper bound on $|y - T(x)|$. That is,

$$\mathrm{Pr}_{\mathbb{W}}\{\gamma > \bar{\gamma}\} \leq \varepsilon \ \Rightarrow \mathrm{Pr}_{\mathbb{W}}\{|y - T(x)| > \bar{\gamma}\hat{\sigma}(x)\} \leq \varepsilon.$$

This means that we could slightly modify Algorithm 1 to obtain a novel algorithm capable of obtaining a probabilistic upper bound conditioned by the value of x. This idea is implemented in Algorithm 2.

Algorithm 2 Conditioned Probabilistic Bound on Error

1: Given a predictor $T : \mathbb{R}^{n_x} \to \mathbb{R}$, an estimator $\hat{\sigma} : \mathbb{R}^{n_x} \to (0, \infty)$, of $\sqrt{\mathbb{E}_{\mathbb{W}}\{(y - T(x))^2 \mid x\}}$, probability levels $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$, choose N and r according to (5).
2: Draw N i.i.d. samples $\{(x_i, y_i)\}_{i=1}^{N}$ according to $\mathrm{Pr}_{\mathbb{W}}$.
3: Compute $\gamma_i = \frac{|y_i - T(x_i)|}{\hat{\sigma}(x_i)}$, $i \in [N]$.
4: Return $\bar{\gamma} = \max^{(r)}\{\gamma_i\}_{i=1}^{N}$, as probabilistic upper bound for $\frac{|y - T(x)|}{\hat{\sigma}(x)}$.

The following Corollary states the probabilistic guarantees of the output $\bar{\gamma}$ of Algorithm 2.

Corollary 2: The output $\bar{\gamma}$ of Algorithm 2 satisfies, with probability no smaller than $1 - \delta$,

$$\mathrm{Pr}_{\mathbb{W}}\{|y - T(x)| > \bar{\gamma}\hat{\sigma}(x)\} \leq \varepsilon.$$

Proof: The proof follows the same lines as the proof of Corollary 1. That is, we infer from Property 1 and Lemma 1 that the proposed choice of N and r guarantees that, with probability no smaller than $1 - \delta$,

$$\mathrm{Pr}_{\mathbb{W}}\left\{\gamma = \frac{|y - T(x)|}{\hat{\sigma}(x)} > \bar{\gamma}\right\} \leq \varepsilon.$$

Thus, we conclude $\mathrm{Pr}_{\mathbb{W}}\{|y - T(x)| > \bar{\gamma}\hat{\sigma}(x)\} \leq \varepsilon$. □

Remark 2 (On normalization of $\hat{\sigma}(x)$): We notice that the upper bound obtained by means of Algorithm 2 provides identical results when the estimator $\hat{\sigma}(x)$ is replaced by a scaled version $\hat{\sigma}_\xi(x) = \xi\hat{\sigma}(x)$, where $\xi > 0$. Thus, multiplicative errors in the estimation of $\sigma(x)$ are corrected in an implicit way by the algorithm.

Remark 3 (Difference with convex scenario approaches): Scenario approaches (see e.g. [5], [6]) obtain both the estimator and probabilistic guarantees in a single optimization problem that requires a number of samples that increases both with the dimension of the regressor used in the predictive model and the number of samples that are allowed to violate the interval predictions. Our approach can be applied to any given predictor $T(\cdot)$ and has a sample complexity that does not depend on the dimension of the regressor. This allows us to consider kernel approaches in a possible infinite dimensional lifted space (see Section V).

IV. Uncertainty quantification for finite families of estimators

The probabilistic bounds proposed for error $e = y - T(x)$ depend not only on the intrinsic random relationship between x and y (joint probability distribution), but also on the choice of the estimators $T(\cdot)$ and $\hat{\sigma}(\cdot)$. Since there exists a myriad of possibilities for choosing $T(\cdot)$ and $\hat{\sigma}(\cdot)$, we now analyze the problem of choosing among a finite family $\mathcal{F}$ of possible pairs $(T(\cdot), \hat{\sigma}(\cdot))$ the one that minimizes the size of the obtained probabilistic bounds. The following result states the relationship between the cardinality $n_{\mathcal{F}}$ of $\mathcal{F}$, and the probabilistic specifications (ε, δ), with the number of samples required to obtain the corresponding bounds.

Theorem 1: Consider the finite family of candidate estimators

$$\mathcal{F} = \{\ (T_j(\cdot), \hat{\sigma}_j(\cdot))\ :\ j \in [n_{\mathcal{F}}]\ \},$$

where $T_j : \mathbb{R}^{n_x} \to \mathbb{R}$ and $\hat{\sigma}_j : \mathbb{R}^{n_x} \to (0, \infty)$ for every $j \in [n_{\mathcal{F}}]$. Given $\varepsilon \in (0,1)$, $\delta \in (0,1)$ and $r \geq 1$, let $N \geq r$ be such that $\mathbf{B}(r-1; N, \varepsilon) \leq \frac{\delta}{n_{\mathcal{F}}}$. Draw N i.i.d. samples $\{(x_i, y_i)\}_{i=1}^N$ from distribution $\Pr_{\mathbb{W}}$ and denote

$$\bar{\gamma}_j \doteq \max{}^{(r)} \left\{ \frac{|y_i - T_j(x_i)|}{\hat{\sigma}_j(x_i)} \right\}_{i=1}^N,\ j \in [n_{\mathcal{F}}]. \tag{8}$$

Then, with a probability no smaller than $1-\delta$,

$$\Pr_{\mathbb{W}}\{|y - T_j(x)| > \bar{\gamma}_j \hat{\sigma}_j(x)\} \leq \varepsilon,\ j \in [n_{\mathcal{F}}].$$

Proof: Denote $\delta_{\mathcal{F}}$ the probability that at least one of the randomly obtained scalars $\{\bar{\gamma}_j\}_{j=1}^{n_{\mathcal{F}}}$, obtained from the random multi-sample $\{(x_i, y_i)\}_{i=1}^N$, does not satisfy the constraint

$$E_j(\bar{\gamma}_j) \doteq \Pr_{\mathbb{W}} \left\{ \frac{|y - T_j(x)|}{\hat{\sigma}_j(x)} > \bar{\gamma}_j \right\} \leq \varepsilon. \tag{9}$$

Thus,

$$\begin{aligned} \delta_{\mathcal{F}} &= \Pr_{\mathbb{W}^N}\{\varepsilon < \max_{j \in [n_{\mathcal{F}}]} E_j(\bar{\gamma}_j)\} \\ &\leq \sum_{j=1}^{n_{\mathcal{F}}} \Pr_{\mathbb{W}^N}\{\varepsilon < E_j(\bar{\gamma}_j)\} \leq \sum_{j=1}^{n_{\mathcal{F}}} \frac{\delta}{n_{\mathcal{F}}} = \delta. \end{aligned}$$

We notice that the last inequality is due to the assumption $B(r-1; N, \varepsilon) \leq \frac{\delta}{n_{\mathcal{F}}}$, (8) and Property 1. Thus, with probability no smaller than $1 - \delta_{\mathcal{F}} \geq 1 - \delta$, inequality (9) is satisfied for every $j \in [n_{\mathcal{F}}]$. □

Remark 4 (Sample complexity for finite families): In view of Lemma 1, it suffices to draw $N = \left\lceil \frac{7.47}{\varepsilon} \log \frac{n_{\mathcal{F}}}{\delta} \right\rceil$ i.i.d. samples from $\Pr_{\mathbb{W}}$ to obtain a probabilistic uncertainty quantification for the complete finite family $\mathcal{F}$. In order to select the best pair $(T_j(\cdot), \hat{\sigma}_j(\cdot))$ in $\mathcal{F}$, one could choose the index $j \in [n_{\mathcal{F}}]$ providing the sharpest probabilistic uncertainty bounds. That is, the one minimizing $\sum_{i=1}^N \bar{\gamma}_j \hat{\sigma}_j(x_i)$. Since $n_{\mathcal{F}}$ enters in a logarithmic way in the sample complexity bound, large values for $n_{\mathcal{F}}$ are affordable. In this case, the search for the most appropriate pair $(T_j(\cdot), \hat{\sigma}_j(\cdot))$ does not need to be exhaustive, and sub-optimal search in the finite family $\mathcal{F}$ could be envisaged (since the probabilistic bounds provided are valid for every member of the family $\mathcal{F}$).

V. Kernel Central Prediction and Uncertainty Quantification

Suppose that M i.i.d. samples $\{(x_i, y_i)\}_{i=1}^M$ are available. We now address the design of the predictor $T(x)$ by means of kernel methods while guaranteeing that the procedure also provides us with an estimation of $\sigma(x)$. Given x, let us define the loss functional

$$J(\theta; x) = \theta^T \Sigma_\theta \theta + \sum_{i=1}^M (y_i - \theta^T \varphi(x_i))^2 \Gamma(x, x_i),$$

where $\varphi : \mathbb{R}^{n_x} \to \mathbb{R}^{n_\theta}$ is the regressor function and $\theta^T \Sigma_\theta \theta$ is a regularization term. A possible choice is $\Sigma_\theta = \tau \mathrm{I}$, where $\tau > 0$. Finally, $\Gamma : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$ is an appropriately chosen weighting function. We assume that $\Gamma(x, z)$ is a decreasing function of $\|x - z\|$, where $\|\cdot\|$ is a given norm. For example, $\Gamma(x, z) = \exp(-\lambda \|x - z\|)$, where $\lambda > 0$.

As it is usual in machine learning, for given x, a *central* estimation for y is provided by $T(x) = \theta_c^T(x)\varphi(x)$, where $\theta_c(x)$ is given by $\theta_c(x) = \arg\min_\theta J(\theta; x)$. We notice that the proposed estimator is a weighted least square estimator with a ridge regression regularization term [27], [28].

There exists two possibilities to obtain predictor $T(x)$ and local estimations $\hat{y}_i(x) = \theta_c^T(x)\varphi(x_i)$, $i \in [M]$ (which will be needed to compute the Parzen estimator for $\sigma(x)$):

Based on $\varphi(\cdot)$: Since $J(\theta; x)$ is a strictly convex quadratic function of θ, the optimal value $\theta_c(x)$ can be obtained determining the value of θ for which the gradient of $J(\theta; x)$ with respect to θ vanishes.

Based on a kernel formulation: Defining the kernel function $K(\cdot, \cdot)$ as $K(x_a, x_b) = \varphi^T(x_a)\Sigma_\theta^{-1}\varphi^T(x_b)$, the estimation $T(x)$, along with the local estimations $\hat{y}_i(x)$, $i \in [M]$, can be obtained in an explicit way by means of well-known kernel tricks (see e.g., [29], [30, §14.4.3] and references therein). In this case, the kernel formulation allows to approach the regression problem in a possibly infinite dimensional lifted space [28].

Once the local estimations $\{\hat{y}_i(x)\}$, $i \in [M]$ have been computed, the estimation for $\sigma(x)$ follows from the following local Parzen estimator (see also equation (7)):

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^M (y_i - \hat{y}_i(x))^2 \Gamma(x, x_i)}{\sum_{i=1}^M \Gamma(x, x_i)}. \tag{10}$$

See the Appendix of the extended arXiv version of this document [31] for a detailed description on how to obtain predictors $T(x)$ and $\hat{\sigma}(x)$ for both considered possibilities (i.e. based on a regressor $\varphi(\cdot)$ or based on a kernel formulation). As commented before, the Parzen estimator converges, under non very restrictive assumptions, to the actual value $\sigma(x)$ as M tends to infinity ([23], [24]). A too reduced number of samples M, or a non appropriate choice for weighting factors $\Gamma(\cdot, \cdot)$, may translate into a degraded estimation of $\sigma(x)$, which will not affect the probabilistic properties of the obtained bounds (that are guaranteed by Theorem 1), but will lead to more conservative bounds. We also notice that an additional set of N i.i.d samples is required to compute the scaling factor $\bar{\gamma}$ in Algorithm 2.

VI. Numerical example: Kernel finite families

We revisit now the numerical example proposed in Section II. For the predictor $T(\cdot)$ we consider a radial basis function kernel $k(x_a, x_b) = 50\exp(-\frac{|x_a - x_b|^2}{0.2})$, and for the estimator $\hat{\sigma}(x)$ the Parzen estimator in (10), where $M = 2,065$ and the pairs $\{(\tilde{y}_i, \tilde{x}_i)\}_{i=1}^M$ are i.i.d. samples from $\Pr_{\mathbb{W}}$. We consider a family of weighting functions $\Gamma(x, z) = \exp(-\lambda|x - z|)$, where $\lambda \in [10]$. Thus, the finite family $\mathcal{F}$ consists of each of

the $n_{\mathcal{F}} = 10$ possible pairs $(T_j(\cdot), \hat{\sigma}_j(\cdot))$ that can be obtained with the $n_{\mathcal{F}}$ values considered for the hyper-parameter λ using the methodology proposed in Section V. Setting $\varepsilon = 0.05$, $\delta = 10^{-6}$ and $n_{\mathcal{F}} = 10$, we obtain from Theorem 1 and Lemma 1 that the choice $N = 2407$ and $r = 60$ is sufficient to obtain a probabilistic uncertainty quantification valid for all the members of the family. The value of λ minimizing the size of the obtained probabilistic bounds is attained at $\lambda = 1$. The resulting scaling parameter is $\bar{\gamma} = 2.15$. See Figure 2 for a comparison of the results obtained for the same validation set that was used to generate Figure 1. The ratio of violation in the validation set for the proposed finite family approach was 0.0332, whereas it was 0.0511 for the exact probabilistic bounds.

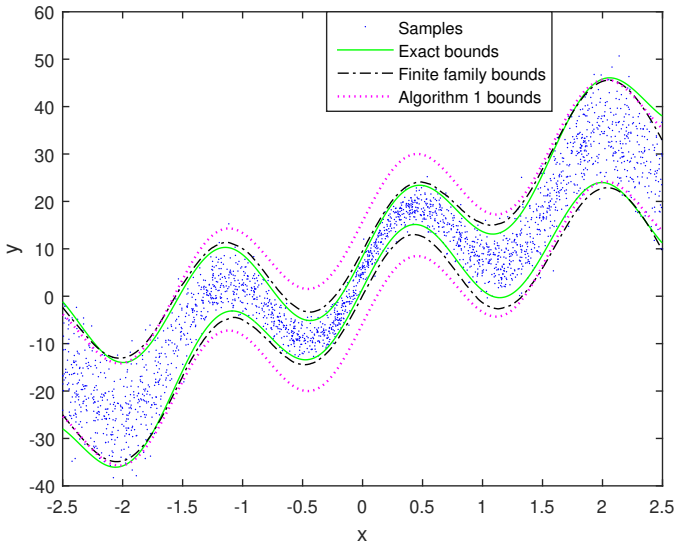

Fig. 2. Probabilistic upper bounds obtained by means of a finite family of kernel estimators, Algorithm 1 bounds, and exact bounds. We notice that the new probabilistic bounds are modulated by the estimated value $\hat{\sigma}(x)$.

VII. Conclusions

In this paper, we proposed a methodology to obtain a probabilistic upper bound on the absolute value of the prediction error via a sample-based approach. We provided a series of approaches of increasing complexity. All the proposed techniques share the desirable characteristic of requiring a number of observations which is independent of the prediction model complexity. This is made possible by the exploitation of a probabilistic scaling scheme.

References

[1] R. G. McClarren, *Uncertainty quantification and predictive computational science*. Springer, 2018.

[2] A. F. Villaverde, E. Raimúndez, J. Hasenauer, and J. R. Banga, "A comparison of methods for quantifying prediction uncertainty in systems biology." *8th Conference on Foundations of Systems Biology in Engineering FOSBE 2019*, vol. 52, no. 26, pp. 45–51, 2019.

[3] P. J. Meyer, A. Devonport, and M. Arcak, *Interval reachability analysis*. Springer, 2021.

[4] R. Koenker, *Quantile regression*. Cambridge University Press, 2005.

[5] M. C. Campi, G. Calafiore, and S. Garatti, "Interval predictor models: Identification and reliability," *Automatica*, vol. 45, no. 2, pp. 382–392, 2009.

[6] S. Garatti, M. Campi, and A. Carè, "On a class of interval predictor models with universal reliability," *Automatica*, vol. 110, pp. 1–9, 2019.

[7] V. Balasubramanian, S.-S. Ho, and V. Vovk, *Conformal prediction for reliable machine learning: Theory, adaptations and applications*. Newnes, 2014.

[8] G. Shafer and V. Vovk, "A tutorial on conformal prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, pp. 371–421, 2008.

[9] T. Alamo, J. Manzano, and E. Camacho, "Robust design through probabilistic maximization," in *Uncertainty in Complex Networked Systems. In Honor of Roberto Tempo*. Birkhäuser, 2018, pp. 247–274.

[10] M. Mammarella, T. Alamo, F. Dabbene, and M. Lorenzen, "Computationally efficient stochastic MPC: A probabilistic scaling approach," in *Proc. of 4th IEEE Conference on Control Technology and Applications (CCTA)*, 2020, pp. 25–30.

[11] R. Tempo, E. Bai, and F. Dabbene, "Probabilistic robustness analysis: explicit bounds for the minimum number of samples," *Systems & Control Letters*, vol. 30, pp. 237–242, 1997.

[12] M. Ahsanullah, V. Nevzorov, and M. Shakil, *An introduction to order statistics*. Atlantis Press, 2013.

[13] M. Campi and S. Garatti, "The exact feasibility of randomized solutions of robust convex programs," *SIAM Journal of Optimization*, vol. 19, pp. 1211–1230, 2008.

[14] G. Calafiore, "Random convex programs," *SIAM Journal of Optimization*, vol. 20, pp. 3427–3464, 2010.

[15] T. Alamo, V. Mirasierra, F. Dabbene, and M. Lorenzen, "Safe approximations of chance constrained sets by probabilistic scaling," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 1380–1385.

[16] M. Mammarella, V. Mirasierra, M. Lorenzen, T. Alamo, and F. Dabbene, "Chance constrained sets approximation: A probabilistic scaling approach – EXTENDED VERSION," *arXiv preprint 2101.06052*, 2021.

[17] B. Karg, T. Alamo, and S. Lucia, "Probabilistic performance validation of deep learning-based robust NMPC controllers," *arXiv preprint 1910.13906*, 2019.

[18] M. Mammarella, T. Alamo, S. Lucia, and F. Dabbene, "A probabilistic validation approach for penalty function design in stochastic model predictive control," in *Proc. of the IFAC World Congress*, vol. 53, no. 2, 2020, pp. 516–521.

[19] T. Alamo, R. Tempo, A. Luque, and D. Ramirez, "Randomized methods for design of uncertain systems: Sample complexity and sequential algorithms," *Automatica*, vol. 52, pp. 160–172, 2015.

[20] M. Mammarella, T. Alamo, F. Dabbene, and M. Lorenzen, "Computationally efficient stochastic MPC: A probabilistic scaling approach," *arXiv preprint 2005.10572*, 2020.

[21] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning (Adaptive computation and machine learning)*. The MIT Press, 2005.

[22] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[23] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[24] B. Pelletier, "Kernel density estimation on Riemannian manifolds," *Statistics & probability letters*, vol. 73, no. 3, pp. 297–304, 2005.

[25] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th ed. Mc Graw Hill, 2002.

[26] M. Huber, "Halving the bounds for the Markov, Chebyshev, and Chernoff inequalities using smoothing," *The American Mathematical Monthly*, vol. 126, no. 10, pp. 915–927, 2019.

[27] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[29] K. P. Murphy, *Machine learning: A probabilistic perspective*. The MIT press, 2012.

[30] V. Vovk, "Kernel ridge regression," in *Empirical inference*. Springer, 2013, pp. 105–116.

[31] V. Mirasierra, M. Mammarella, F. Dabbene, and T. Alamo, "Prediction error quantification through probabilistic scaling," *arXiv preprint 2105.14187*, 2021.