Componentwise Hölder Inference for Robust Learning-Based MPC

José María Manzano, David Muñoz de la Peña, Jan-Peter Calliess, and Daniel Limon

***Abstract*—This article presents a novel learning method based on componentwise Hölder continuity, which allows one to consider independently the contribution of each input to each output of the function to be learned. The method provides a bounded prediction error, and its learning property is proven. It can be used to obtain a predictor for a nonlinear robust learning-based predictive controller for constrained systems. The resulting controller achieves better closed loop performance and larger domains of attraction than learning methods that only consider nonlinear set membership, as illustrated by a case study.**

***Index Terms*—Inference algorithms, machine learning, nonlinear systems, predictive control, robust stability.**

I. Introduction

Among the different techniques applied to learning-based predictive control, one of the most popular consists in using a machine learning algorithm to obtain a predictor from input–output data, such as direct weight optimization [1], [2], Gaussian processes [3], [4] or neural networks [5].

When using the model of a system, estimated from past observations, prediction errors are present. To guarantee a safe evolution of the system, the learning method must provide a description of the uncertainty between the real evolution of the plant and the estimated one.

Lipschitz interpolation methods [6], [7] may provide a bound on the prediction error. They have also been referred to as nonlinear set membership (NSM) [8]. However, knowledge of the true Lipschitz constant of the plant is required. If this constant is estimated stochastically [9], the deterministic feature of the framework is lost. However, there exists a class of learning rules, named *kinky inference* (KI) [10], which encompasses Lipschitz interpolation and NSM, that provides guaranteed bounds on the prediction error, with Lipschitz constants estimated from the data set.

This property has been taken into account to design safe learning-based predictive controllers [11], [12]. In [13], a data-based robust model predictive control (MPC) was presented using the *smooth projected KI* version of the KI class, for systems with constraints on the inputs. Later, in [14], a stabilizing robust version for systems subject to input and output constraints was developed.

The main drawback is the possible conservatism of the controller, due to the open-loop nature of the predictions, inherited from robust predictive controllers' design. Thus, the availability of tighter descriptions of the prediction error [15], or tractable procedures to estimate the reachability sets of the prediction models [16] can lead to an enhanced design of safe learning-based MPC. Therefore, the derivation of learning methods able to reduce the bound on the uncertainty is crucial.

The objective of this article[1] is to introduce a novel learning methodology that extends and improves KI, which is able to significantly reduce the prediction error. To this end, an extended version of continuity of functions is used, namely, *componentwise Hölder continuity*, in order to derive a novel learning method, named *componentwise Hölder kinky inference* (CHoKI).

In its core, this method generalizes the real-valued Hölder constant L and exponent p to matrices, in order to capture different variations of the target along different dimensions of the vector field that is to be learned. A nonparametric estimation method and the procedure to calculate its hyper-parameters from measured (possibly noisy) data is presented. It is first proven that the proposed learning method can estimate any Hölder continuous function over a compact domain. Then, it is also shown that the proposed method encompasses and extends existing methods, such as KI or NSM.

Besides, it is rigorously proven that the proposed learning method ensures a bounded estimation of the prediction error and that the method is a learning algorithm. This is a key aspect, since it allows one to relate the accuracy of the predictions to the density of the available observations. In addition, it renders the online version of such algorithm applicable, that is, makes it prone to decrease prediction errors as more points are taken into consideration by the predictor.

Another contribution of this article is the application of the novel estimation method to model nonlinear dynamical systems, and based on this, to enhance existing model predictive controllers. To this aim, the robust constrained MPC of [14] is extended to use the componentwise prediction model, implying a double benefit: Not only the prediction errors are decreased with the new method, but also the domain of attraction of the controllers is enlarged since the componentwise approach allows for tighter bounds on the propagation of the uncertainty.

Notation: A set of integers $[a, b]$ is denoted $\mathbb{I}_a^b$. The notation (v, w) implies $[v^T, w^T]^T$, and $v \leq w$ implies that the inequality holds for every component. $\|v\|$ stands for the Euclidean norm of v, and $|v| = \{w : w_i = |v_i|, \forall i\}$. Given two sets A, B, $A \oplus B$ denotes the Minkowski sum, and $A \ominus B$ the Pontryagin difference. Their Cartesian product is denoted $A \times B = \{(x, y)|x \in A, y \in B\}$. $\|A\|$ denotes its Lebesgue measure and its norm is denoted $\|A\|_\infty = \max_{a \in A} \|a\|_\infty$. The box $\mathbb{B}(v) \subset \mathbb{R}^{n_v}$ is defined as $\mathbb{B}(v) = \{y : |y| \leq v\}$, and the ball $\mathcal{B}(v) \subset \mathbb{R}^{n_v}$ is defined as $\mathcal{B}(v) = \{y : 0 \leq y \leq v\}$. The ball over a set A, $\mathcal{B}(A)$, represents the Cartesian closed topological hull of A. An n-dimensional

Manuscript received November 13, 2020; accepted January 15, 2021. Date of publication February 2, 2021; date of current version November 4, 2021. This work was supported in part by the Agencia Estatal de Investigación (AEI) under GrantPID2019-106212RB-C41/AEI/10.13039/501100011033 and by MINERCO-Spain and FEDERfunds under Grant DPI2016-76493-C3-1-R. Recommended by Associate Editor E. C. Kerrigan. *(Corresponding author: José María Manzano.)*

José María Manzano is with the Department of Engineering, Universidad Loyola Andalucía, 41704 Dos Hermanas, Spain (e-mail: jmanzano@uloyola.es).

David Muñoz de la Peña and Daniel Limon are with the Department of Systems Engineering, and Automation, Universidad de Sevilla, 41092 Sevilla, Spain (e-mail: dmunoz@us.es; dlm@us.es).

Jan-Peter Calliess is with the Department of Engineering Science, University of Oxford, OX30HW Oxford, U.K. (e-mail: jcalliess@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2021.3056356.

Digital Object Identifier 10.1109/TAC.2021.3056356

[1] A preliminary version of this article was presented in [17]

column vector of ones is denoted $\mathbb{1}_n$, and the n, m-dimensional matrix of ones is denoted $\mathbb{1}_{n\times m}$. Analogously for a vector (or matrix) of zeros, denoted $\mathbb{0}_{n(\times m)}$. The ith row of a matrix M is denoted M_i.

II. Componentwise Hölder Inference

Consider an unknown (target) function within two compact spaces $\mathcal{W} \in \mathbb{R}^{n_w}$ and $\mathcal{Y} \in \mathbb{R}^{n_y}$ (referred to as *input* and *output* spaces, respectively) such that $f : \mathcal{W} \to \mathcal{Y}$. The only information available *a priori* from this function f is a collection of (possibly) noisy observations, gathered in a data set of $N_\mathcal{D}$ points, defined as $\mathcal{D} = \{(w_i, \tilde{y}_i)\},\ i \in \mathbb{I}_1^{N_\mathcal{D}}$, where $\tilde{y}_i$ is the noise-corrupted measurement, assuming the noise is bounded by some known $\bar{\epsilon} \in \mathbb{R}^{n_y}$. The data set containing only inputs are denoted $\mathcal{W}_\mathcal{D}$.

The objective is to learn f and to predict unseen query points $q \in \mathcal{W} \setminus \mathcal{W}_\mathcal{D}$. To this end, Hölder continuity of the function is assumed, that is

$$\|f(w_1) - f(w_2)\| \leq L\|w_1 - w_2\|^p,\ \forall w_1, w_2 \tag{1}$$

where the exponent p is a scalar such that $0 < p \leq 1$. In particular, if $p = 1$ this property is called Lipschitz continuity, with constant L, which is also a scalar.

Based on this property, Lipschitz interpolation methods learn the ground truth function f using the data set of samples [6], [7], as well as the constant L and the exponent p. Encompassing such methods, the standard KI is presented in [10]. If the parameters are known *a priori*, the method guarantees bounded prediction error and sample consistency. If they are unknown, [18] proposes a method to estimate L that maintains bounded errors and consistency.

A. Componentwise Hölder Continuity

The Hölder continuity condition relates the effect on the output of a variation on the input, bounding the worst case. In this article, we extend the Hölder property to a componentwise setting, where the contribution of the variation of each input on each output is taken into account separately.

Hence, we propose to use matrices (instead of scalars) as the parameters of the Hölder property: $\mathcal{L}$, $\mathcal{P} \in \mathbb{R}^{n_y\times n_w}$, yielding the following definition:

Definition 1 (Componentwise Hölder continuity): Given the matrices $\mathcal{L}$ and $\mathcal{P}$, a function $f : \mathcal{W} \to \mathcal{Y}$ is componentwise $\mathcal{L}$-$\mathcal{P}$-Hölder continuous if $\forall w_1, w_2 \in \mathcal{W}$ and $\forall i \in \mathbb{I}_1^{n_y}$

$$|f_i(w_1) - f_i(w_2)| \leq \sum_{j=1}^{n_w} \mathcal{L}_{ij}|w_{1,j} - w_{2,j}|^{\mathcal{P}_{ij}}. \tag{2}$$

In Hölder continuity, L aggregates the effect of the inputs on the outputs into a single constant. On the contrary, the proposed componentwise approach uses each $\mathcal{L}_{i,j}$ to take into account separately the effect of each input j on each output i.

This componentwise Hölder continuity condition may be rewritten in a more compact form, using the following notation:

Given a vector $w \in \mathbb{R}^{n_w}$ and two matrices $\mathcal{L}$, $\mathcal{P} \in \mathbb{R}^{n_y\times n_w}$, we define

$$\mathfrak{d}_\mathcal{L}^\mathcal{P}(w) := \left(a : a_i = \sum_{j=1}^{n_w} \mathcal{L}_{i,j} w_j^{\mathcal{P}_{i,j}}, \forall i \in \mathbb{I}_1^{n_y}\right). \tag{3}$$

Then, the componentwise Hölder continuity in (2) can be written as

$$|f(w_1) - f(w_2)| \leq \mathfrak{d}_\mathcal{L}^\mathcal{P}(|w_1 - w_2|),\ \forall w_1, w_2. \tag{4}$$

The following theorem states under which conditions Hölder continuity and componentwise Hölder continuity are equivalent. The proof of the theorem is presented in the appendix.

Theorem 1: Let $f : \mathcal{W} \subseteq \mathbb{R}^{n_w} \to \mathcal{Y} \subseteq \mathbb{R}^{n_y}$.

1) If f is Hölder continuous in $\mathcal{W}$, then f is componentwise Hölder continuous in $\mathcal{W}$.
2) If $\mathcal{W}$ is compact and f is componentwise Hölder continuous in $\mathcal{W}$, then f is Hölder continuous in $\mathcal{W}$.

Corollary 1: If $\mathcal{P} = p\mathbb{1}_{n_y\times n_w}$, the equivalence in Theorem 1 holds for any input space $\mathcal{W}$, even if it is not compact.

With a slight abuse of notation, in what is to follow the mapping in (3) may also be used for sets, such that for a given set $A \subset \mathbb{R}^{n_w}$, we denote $\mathfrak{d}_\mathcal{L}^\mathcal{P}(A) := \{\mathfrak{d}_\mathcal{L}^\mathcal{P}(x) | x \in A\}$.

Corollary 2: In the Lipschitz case and for every output component $i \in \mathbb{I}_1^{n_y}$, if the Lipschitz constant is such that $L = \|\mathcal{L}_i\|_\infty$, then $L\|w_i\|_\infty \geq \mathfrak{d}_{\mathcal{L}_i}(|w_i|)$.

B. Componentwise Hölder KI

Proceeding as in the standard KI approach, assuming that f is Hölder continuous, and given a data set of input–output observations, this article presents the *componentwise Hölder kinky inference* (CHoKI) predictor.

Provided that the componentwise Hölder condition (2) holds in virtue of Theorem 1, the proposed estimation method is

$$\begin{aligned}\hat{\mathfrak{f}}(q; \Theta, \mathcal{D}) = &\frac{1}{2} \min_{i=1,\ldots,N_\mathcal{D}} (\tilde{y}_i + \mathfrak{d}_\mathcal{L}^\mathcal{P}(|q - w_i|)) \\ &+ \frac{1}{2} \max_{i=1,\ldots,N_\mathcal{D}} (\tilde{y}_i - \mathfrak{d}_\mathcal{L}^\mathcal{P}(|q - w_i|))\end{aligned} \tag{5}$$

where $\Theta = \{\mathcal{L}, \mathcal{P}\}$.

The resulting prediction function is componentwise Hölder continuous, with the same parameters, as stated in the following lemma (whose proof can be found in the appendix).

Lemma 1: For a given $\Theta = \{\mathcal{L}, \mathcal{P}\}$, the CHoKI predictor $\hat{\mathfrak{f}}$ given by (5) is componentwise $\mathcal{L}$-$\mathcal{P}$-Hölder continuous.

In case that the parameters $\mathcal{L}$ and $\mathcal{P}$ are unknown *a priori*, they must be estimated from the available input–output data. In standard KI methods, the Hölder constant can be derived from $\mathcal{D}$ by a procedure based on sample consistency [18]. However, in the componentwise case, the method presented in [18] cannot be used, since $\|w_1 - w_2\|$ provides an aggregated measurement of the effect of the inputs on the outputs, and there is no direct information of the contribution of each input on a particular output. In order to infer this contribution, an optimization method is adopted, extending the results of [19] to obtain the matrices $\mathcal{L}$ and $\mathcal{P}$.

The parameters $\Theta = \{\mathcal{L}, \mathcal{P}\}$ are estimated by solving an optimization problem offline, which depends on a regularization parameter $\lambda \in \mathbb{R}^{n_y}$ and two data sets: the data set $\mathcal{D}$, used for estimation, and a data set $\mathcal{D}_{\text{test}}$ used for validation. In this optimization problem, a measure of the performance of the prediction over the data set $\mathcal{D}_{\text{test}}$, $g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}})$, plus a regularization term is minimized, subject to a constraint that ensures their consistency with the samples of the data set

$$\Theta = \arg\min_\Theta g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}}) + \tau\|\mathcal{L} - \mathcal{L}_0\|_1 \tag{6a}$$

$$\text{s.t. } |\tilde{y}_i - \tilde{y}_j| - \lambda \leq \mathfrak{d}_\mathcal{L}^\mathcal{P}(|w_i - w_j|) \tag{6b}$$

$$\forall w_i, w_j \in \mathcal{W}_\mathcal{D}, w_i \neq w_j,$$

$$0 < \mathcal{P}_{ij} \leq 1, i \in \mathbb{I}_1^{n_y}, j \in \mathbb{I}_1^{n_w} \tag{6c}$$

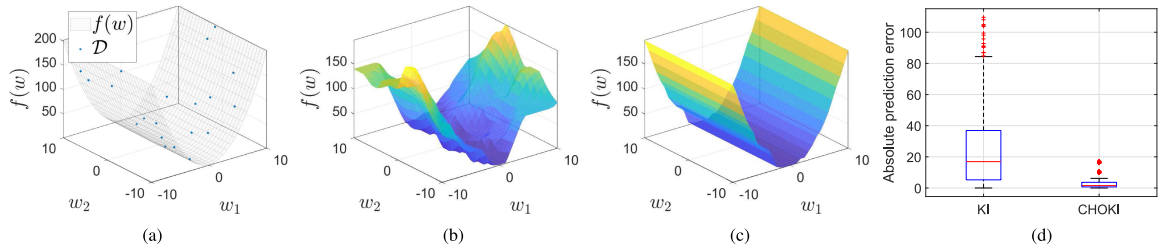


Fig. 1. Prediction of $f(w) = 2w_1^2 + 0.5\sqrt{|w_2|}$ using both the standard KI and the CHoKI method, given $N_{\mathcal{D}} = 20$ random data points. (a) Ground truth and data set. (b) KI prediction. (c) CHoKI prediction. (d) Prediction error (absolute value).

where τ is a design regularization hyper-parameter used to ensure boundedness of $\mathcal{L}$, and $\mathcal{L}_0$ stands for any possible prior guess of $\mathcal{L}$. Besides, the cost function g must be positive and bounded, for any size of $\mathcal{D}$. A possible choice of the performance measure is the mean squared prediction error

$$g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}}) = \frac{1}{N_{\mathcal{D}}} \sum_{w_i \in \mathcal{W}_{\mathcal{D}_{\text{test}}}} \|\hat{\mathfrak{f}}(w_i; \Theta, \mathcal{D}) - \tilde{y}_i\|^2. \tag{7}$$

Remark 1: For an analysis on the effect of the regularization hyper-parameter λ, the reader is referred to [18], where $\lambda \geq 2\bar{\mathfrak{e}}$ is taken for each component to effectively smooth out the effect of the noise in the prediction.

Remark 2: The regularization term of the cost function, $\tau \| \mathcal{L} - \mathcal{L}_0 \|_1$, prevents the problem from overfitting the noise, while ensuring boundedness of $\mathcal{L}$. If $\mathcal{D}_{\text{test}}$ is separate from $\mathcal{D}$ and such that g is bounded for all $\mathcal{L}$, then it can be removed, setting $\tau = 0$.

Remark 3: In practice, it may be easier to fix $\mathcal{P}$ a priori, and to optimize over $\mathcal{L}$, provided that the assumptions hold for the chosen $\mathcal{P}$.

Next, based on T. Mitchell's definition of a learning algorithm [20], the following theorem states that CHoKI is a learning method, which is a key contribution of this article.

Theorem 2: Let $\Theta = \{\mathcal{L}, \mathcal{P}\}$ be obtained as the solution of (6) for a data set $\mathcal{D}$ with $\lambda \geq 2\bar{\mathfrak{e}}$, and assume that the function f satisfies the componentwise Hölder condition (4) for the pair $\{\mathcal{L}_f, \mathcal{P}\}$ in $\mathcal{W}$. Then, $\mathcal{L}$ is bounded and

$$|f(w) - \hat{\mathfrak{f}}(w; \Theta, \mathcal{D})| \leq \mathfrak{d}^{\mathcal{P}}_{\mathcal{L}_f + \mathcal{L}}(\mathcal{R}_{\mathcal{D}}) + \frac{\lambda}{2} + \bar{\mathfrak{e}} \tag{8}$$

where $\mathcal{R}_{\mathcal{D}} = \max_{w \in \mathcal{W}} \min_{w_j \in \mathcal{W}_{\mathcal{D}}} (|w_j - w|)$ measures the maximum radius between a possible query and the data set.

The proof of this theorem can be found in the appendix. Based on this theorem, it can be derived that the worst-case prediction error is bounded for all queries q in a compact space $\mathcal{W}$. Besides, it proves that as more observations are added to the data set, the prediction error decreases, vanishing up to $\lambda/2 + \bar{\mathfrak{e}}$ for infinitely dense data sets, when $\mathcal{R}_{\mathcal{D}} \to 0$.

Corollary 3: If the real Hölder parameters $\mathcal{L}_f$ and $\mathcal{P}$ were known, the worst-case prediction error is bounded by

$$|f(w) - \hat{\mathfrak{f}}(w; \Theta, \mathcal{D})| \leq \mathfrak{d}^{\mathcal{P}}_{\mathcal{L}_f}(\mathcal{R}_{\mathcal{D}}) + 2\bar{\mathfrak{e}}$$

following the proof in [18] for the standard KI approach.

Finally, it is demonstrated that CHoKI enhances the existing methods based on Hölder continuity. This is achieved by proving that KI is a particular case of CHoKI for a certain parameter setting, as it is stated in the following lemma.

Lemma 2: If $\mathcal{L} = L\mathbb{1}_{n_w}^T$ and the Lipschitz case (i.e., $\mathcal{P} = \mathbb{1}_{n_y \times n_w}$) is applied using the one norm in the standard KI, then both methods are equal.

Note that the method proposed to obtain $\mathcal{L}$, $\mathcal{P}$ adds degrees of freedom with respect to the scalar KI case. So in general, the CHoKI predictor will perform equal or better than standard KI over $\mathcal{D}_{\text{test}}$, since the latter is a particular case of the former.

The overall performance of the proposed method is illustrated in the following example.

Example 1: Consider the function $f : \mathcal{W} \subset \mathbb{R}^2 \to \mathbb{R}$:

$$f(w) = 2w_1^2 + \frac{\sqrt{|w_2|}}{2}$$

within the input space $\mathcal{W} = \mathbb{B}([10, 10])$. Note that f is not Lipschitz continuous in the origin. Fig. 1 depicts the prediction errors generated by CHoKI and KI trained on a set of $N_{\mathcal{D}} = 20$ random sample data, over a grid of 900 query points. The Hölder parameters were obtained as per (6) with $\lambda = 0$, yielding $L = 34.6$, $p = 1$ and $\mathcal{L} = [37.5,\ 0.2]$, $\mathcal{P} = [1,\ 0.75]$. Note that with CHoKI the pointwise prediction error decreases up to 84%.

III. CHoKI-Based Robust MPC

This section presents an extension of the learning-based MPC presented in [14], using the CHoKI method to learn the plant dynamics and to derive a prediction model. A discrete plant whose manipulable inputs are $u \in \mathbb{R}^{n_u}$ and whose measured controlled outputs are $y \in \mathbb{R}^{n_y}$ is considered. These signals must be limited to the constraint compact sets $u \in \mathcal{U}$ and $y \in \mathcal{Y}$. The measured output can be modeled as a NARX regression of previous inputs and outputs [13]

$$y(k+1) = f(x(k), u(k)) + e(k) \tag{9}$$

where

$$x(k) = (y(k), \ldots, y(k - n_a), u(k-1), \ldots, u(k - n_b)) \tag{10}$$

is the regression state $x \in \mathbb{R}^{n_x}$, for some memory horizons n_a and $n_b \in \mathbb{N}_0$.

In order to use the proposed estimation method, the arguments of f are aggregated into $w = (x, u) \in \mathbb{R}^{n_w}$. Given some historical trajectories of $u(k)$ and $y(k)$, it is possible to construct a data set $\mathcal{D} = \{(w_k, y_{k+1})\}$ for $k = 1, \ldots, N_{\mathcal{D}}$ and to predict a new output $y(k+1)$ given $\Theta = \{\mathcal{L}, \mathcal{P}\}$ as in (5)

$$\hat{y}(k+1) = \hat{\mathfrak{f}}(w(k); \Theta, \mathcal{D}). \tag{11}$$

The prediction model can be formulated in state-space as follows:

$$\hat{x}(k+1) = \hat{F}(x(k), u(k)) \tag{12a}$$

$$\hat{y}(k) = M\hat{x}(k) \tag{12b}$$

where $\hat{F}(x(k), u(k)) = (\hat{\mathfrak{f}}(x(k), u(k)), y(k), \ldots, y(k - n_a + 1), u(k), \ldots, u(k - n_b + 1))$ and $M = [I_{n_y}, 0, \ldots, 0]$.

We propose to use a robust model predictive controller which is stable by design, extending [14] to the enhanced learning method proposed in this article. Similarly to [14], the predictive controller is derived from

the following optimization problem, denoted $P_N(x(k);\Theta,\mathcal{D})$:

$$\min_{\mathbf{u}} \sum_{i=0}^{N-1} \ell(\hat{x}(i|k),u(i)) + \eta V_f(\hat{x}(N|k)) \tag{13a}$$

$$\text{s.t. } \hat{x}(0|k) = x(k) \tag{13b}$$

$$\hat{x}(j+1|k) = \hat{F}(\hat{x}(j|k),u(j)),\ j \in \mathbb{I}_0^{N-1} \tag{13c}$$

$$\hat{y}(j|k) = M\hat{x}(j|k) \tag{13d}$$

$$u(j) \in \mathcal{U} \tag{13e}$$

$$\hat{y}(j|k) \in \mathcal{Y}_j. \tag{13f}$$

Here, N stands for the prediction horizon, $\ell(\cdot,\cdot)$ is the stage cost, $\eta \geq 1$ is a design parameter and $V_f(\cdot)$ is the terminal cost. Notice that this robust MPC does not require a terminal constraint, but only a (weighted) suitable terminal cost function [14]. The sets of constraints is given by

$$\mathcal{Y}_j = \mathcal{Y}_{j-1} \ominus \mathcal{R}_j \tag{14}$$

with $\mathcal{Y}_0 = \mathcal{Y}$, where the reachability sets $\mathcal{R}_j$ account for the possible deviation of the nominal predictions $\hat{y}(j|k)$ from the real system, j steps ahead.

In this article, CHoKI is used to derive a more accurate prediction model (13c), and to get better estimations of the reachable sets $\mathcal{R}_j$, as it is presented next.

A. Reachability Sets

Although the method presented in [14] could be used to obtain reachability sets, in this section we present a procedure based on the CHoKI predictor to calculate these sets, which in general provides less conservative results. For its calculation, we will make use of the following lemma.

Lemma 3: Consider a sequence of future inputs $u(k+i), i \in \mathbb{I}_0^{N-1}$. Let $c_1 \in \mathbb{R}^{n_y}$ be a vector such that

$$|y(k+1) - \hat{y}(1|k)| \leq c_1. \tag{15}$$

The mismatch between a prediction at time step $k+j$ given the measurement at time step k, $\hat{y}(j|k)$, and the prediction at that time step given the measurement at time $k+1$, $\hat{y}(j-1|k+1)$, for the same sequence of control inputs is bounded by the sets

$$|\hat{y}(j|k) - \hat{y}(j-1|k+1)| \in \mathcal{M}_j \subseteq \mathbb{R}^{n_y} \tag{16a}$$

$$|\hat{w}(j|k) - \hat{w}(j-1|k+1)| \in \mathcal{G}_j \subseteq \mathbb{R}^{n_w}. \tag{16b}$$

The sets $\mathcal{M}$ and $\mathcal{G}$ can be obtained from the recursion

$$\mathcal{M}_j = \mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{G}_{j-1})), \tag{17a}$$

$$\mathcal{G}_j = \mathcal{M}_j \times \ldots \times \mathcal{M}_{\sigma(j)} \times \{0\} \times \ldots \times \{0\} \tag{17b}$$

with $\sigma(j) = \max(1, j - n_a)$, and $\mathcal{M}_1 = \mathcal{B}(c_1)$.

Proof. Given $w(k) = (x(k), u(k))$ and the sequence of future inputs $u(k+i)$ for $i \in \mathbb{I}_0^{N-1}$, provided that w contains predicted values if $j > n_a$, and real measurements if not, the definition of $|\hat{w}(j|k) - \hat{w}(j-1|k+1)|$ translates into (17b), provided that if $\hat{y}(j|k)$ is a real measurement, then the difference $|y(j|k) - y(j-1|k+1)|$ equals $\mathbb{0}_{n_y}$, and belongs to $\mathcal{M}_j$ otherwise.

To obtain $\mathcal{M}_j$ we make use of the componentwise Hölder continuity of the predictor $\hat{\mathfrak{f}}$. Given (15), we have that $\mathcal{M}_1 = \mathcal{B}(c_1)$. Note that $\hat{y}(j+1|k) = \hat{\mathfrak{f}}(\hat{x}(j|k), u(k+j); \Theta, \mathcal{D})$, and that $|\hat{\mathfrak{f}}_i - \hat{\mathfrak{f}}_j| \leq \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_i - w_j|)$. Hence, it follows that $|\Delta y| \leq \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{G}_j)$, so the minimum set that contains all possible values of $|\Delta y|$ is the Cartesian closed topological hull, $\mathcal{M}_{j+1} = \mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{G}_j))$. ■

This lemma proved that the smallest sets that contain all possible values of the differences $|\hat{y}(j|k) - \hat{y}(j-1|k+1)|$ and $|\hat{w}(j|k) - \hat{w}(j-1|k+1)|$ are $\mathcal{M}_j$ and $\mathcal{G}_j$, respectively.

The set $\mathcal{R}_j$ is defined as

$$\mathcal{R}_j = \{y : |y| \in \mathcal{M}_j\} \tag{18}$$

for all $j \in \mathbb{I}_1^N$, where $\mathcal{M}_j$ is calculated from (17) using $\mathcal{M}_1 = \mathcal{B}(c_1)$, with $c_1 = \mu$.

The following lemma proves that the sets $\mathcal{M}_j$ and $\mathcal{G}_j$ are boxes that can be calculated by a simple recursion.

Lemma 4: Given Lemma 3, let $c_j \in \mathbb{R}^{n_y}$ and $d_j \in \mathbb{R}^{n_w}$ be such that $\mathcal{M}_j = \mathcal{B}(c_j)$ and $\mathcal{G}_j = \mathcal{B}(d_j)$. The sets $\mathcal{M}_j$ and $\mathcal{G}_j$ can be calculated using the recursion $c_j = \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(d_{j-1})$ and $d_j = (c_j, \ldots, c_{\sigma(j)}, 0, \ldots, 0)$. Besides, $\mathcal{R}_j = \mathbb{B}(c_j)$.

Proof: The Cartesian hull of the map $\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{B}(v))$, (the tightest ball containing it) is a ball given by $\mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(v))$, i.e., $\mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{B}(v))) = \mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(v))$. This is inferred by noticing that $\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{B}(v)) = \{\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(x) | x \in \mathcal{B}(v)\}$, and that for every component $i \in \mathbb{I}_1^{n_y}$ it can be bounded that for all $j \in \mathbb{I}_1^{n_w}$, $0 \leq \mathcal{L}_{ij} x_j^{\mathcal{P}_{ij}} \leq \mathcal{L}_{ij} v_j^{\mathcal{P}_{ij}}$, for all $x \in \mathcal{B}(v)$. Besides, for $j > 1$, given that $\mathcal{G}_{j-1} = \mathcal{B}(d_{j-1})$, we have that

$$\mathcal{M}_j = \mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(\mathcal{G}_{j-1})) = \mathcal{B}(\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(d_{j-1})) = \mathcal{B}(c_j)$$

$$\mathcal{G}_j = \mathcal{M}_j \times \ldots \times \mathcal{M}_{\sigma(j)} \times \{0\} \times \ldots \times \{0\} = \mathcal{B}(d_j).$$

■

If the set $\mathcal{Y}$ is a polytope, as it is customary, then the resulting tightened constraints are also polytopes. Notice that these calculations are done only once, offline.

B. Stabilizing Conditions for CHoKI-MPC

In order to recover the safe-by-design properties of the controller in [14], the following assumptions must hold true.

Assumption 1: f is Hölder continuous.

Assumption 2: The prediction error is bounded by some $\mu \in \mathbb{R}^{n_y}$. This bound is known for all admissible x and u, i.e.,

$$|\hat{\mathfrak{f}}(x,u;\Theta,\mathcal{D}) - f(x,u) - e| \leq \mu. \tag{19}$$

Notice that this bound exists given Assumption 1, Theorem 2, and the compactness of the admissible sets.

Assumption 3: The stage cost function $\ell(x,u)$ is a componentwise Hölder continuous, positive definite function such that $\ell(x,u) \geq \alpha(\|x\|)$, where α is a $\mathcal{K}$-function. Given $\mathcal{P}$, its Hölder constant is $\mathcal{L}_\ell$.

Assumption 4: V_f is a componentwise Hölder continuous, positive definite function, with Hölder parameters $\mathcal{L}_f$, $\mathcal{P}$, such that Assumption 4 in [14] holds.

Assumption 5: Assumption 5 in [14] holds redefining ν as

$$\nu(\mu) = \sum_{j=1}^{N} \|\mathfrak{d}_{\mathcal{L}_\ell}^{\mathcal{P}}(\mathcal{G}_j)\|_\infty + \lambda \|\mathfrak{d}_{\mathcal{L}_f}^{\mathcal{P}}(\mathcal{G}_{N+1})\|_\infty. \tag{20}$$

Assumption 6: The prediction horizon N and the estimation error bound μ are such that the set $\mathcal{Y}_N$ is nonempty.

Next, we present the stability result for the CHoKI-MPC, which follows the same line of reasoning as the stability proof presented in [14] for KI predictors.

Theorem 3 (ISS stability): Suppose that assumptions 1–6 hold. Let $\kappa_N(x)$ be the control law derived from the solution of $P_N(x)$ (13) applied using a receding horizon policy. Then, for any $x(0) \in \Gamma$, the system to be controlled by the control law $u(k) = \kappa_N(x(k))$ is input-to-state stable with respect to the estimation error μ, $x(k) \in \Gamma$, and the constraints are always satisfied, i.e., $y(k) \in \mathcal{Y}, \forall k$.

Proof: The proof of this theorem follows the same steps as the proof of Theorem 1 in [14]. Lemma 2 in [14] has to be modified to take into account the CHoKI predictor as follows.

Lemma 5: For all $y \in \mathcal{Y}_j$ and for all Δy such that $|\Delta y| \in \mathcal{M}_j$, the sets $\mathcal{Y}_j$ are such that $y + \Delta y \in \mathcal{Y}_{j-1}$.

Proof: By definition, we have that if $|\Delta y| \in \mathcal{M}_j$ then $\Delta y \in \mathcal{R}_j$. Thus, since the origin is contained in $\mathcal{R}_j$

$$y + \Delta y \in \mathcal{Y}_j \oplus \mathcal{R}_j = \mathcal{Y}_{j-1} \ominus \mathcal{R}_j \oplus \mathcal{R}_j \subseteq \mathcal{Y}_{j-1}.$$

■

It is important to remark that the benefits of the robust predictive controller based on the componentwise Hölder approach is two-fold, compared to the KI-based MPC of [14]. First, the enhanced learning method potentially provides significantly less conservative estimation errors (see Fig. 1), yielding more accurate predictions. This leads to an improvement of the closed-loop performance of the controlled system. Second, note that the recursion used for the set of tightened constraints [c.f., (14)] is obtained using the componentwise Hölder metric, in contrast to the standard version presented in [14]. Hence, recalling Corollary 2, even if the maximum prediction errors were the same (which in general they are not), the back-off of the set of tightened constraints are less conservative ($\mathcal{R}_N^{\text{CHoKI}} \subseteq \mathcal{R}_N^{\text{KI}}$), thus yielding larger regions of the feasibility of the proposed controller. This double benefit of the proposed method will be illustrated in the following case study.

IV. Case Study

The system considered is the quadruple-tank process described in [21] and [22], which consists of four tanks, where the two on top discharge on the inferior ones. The tanks are fed with two pumps, whose flows enter two three-ways valves, which divide each flow into two branches, determined by the fractions γ_a and γ_b.

There are two control inputs, the flows q_a and q_b (m^3/h). The heights of the tanks are denoted as $h_i(\mathrm{m}), i \in \mathbb{I}_1^4$. The outputs of the system are the heights of the two lower tanks, i.e., h_1 and h_2. The model and its parameters can be found in [22]. Note that the model is only used to emulate the real plant. It is also assumed that the height sensors have a 2% measuring error. The error is generated randomly for each measurement using a uniform distribution.

The constraints in the inputs are $0 \leq q_{a,b} \leq 2.6\ \mathrm{m}^3/\mathrm{h}$, and the constraints in the heights are given by $0 \leq h_1 \leq 1.25\ \mathrm{m}$ and $0 \leq h_2 \leq 1.42\ \mathrm{m}$. The sampling time is 5 s, and the reference operating point is $h^{\text{ref}} = [0.65\ 0.65](\mathrm{m})$, $q^{\text{ref}} = [1.63\ 1.99](\mathrm{m}^3/\mathrm{h})$.

A data set with $N_{\mathcal{D}} = 36\,000$ is obtained as in [13], scaling the values of all signals between 0 and 1. The prediction horizon is set to $N = 5$. Separate cross validation tests are used to estimate $\mathcal{L}$ using $\lambda = 2\bar{\epsilon}$, such that the sum of the areas of $\mathcal{Y}_j$ is maximized, i.e.,

$$g = -\sum_{j=1}^{N} \|\mathcal{Y}_j\|. \tag{21}$$

This occurs for $n_a = 3$ and $n_b = 1$, yielding $\mu = [0.031,\ 0.032]$.[2] The set of tightened constraints is obtained considering the Lipschitz case.

The solver chosen for the optimization problem is MATLAB's *fmincon*. The stage and terminal costs are defined as

$$\ell(x,u) = \|x - x^{\text{ref}}\|_Q^2 + \|u - u^{\text{ref}}\|_R^2$$
$$V_f(x) = \|x - x^{\text{ref}}\|_P^2$$

where $Q = 100\ I_2$, $R = I_2$, and $\lambda = 10$. The terminal matrix P is obtained by solving an LQR for a linearisation of the CHoKI model around the reference point.

The initial state is set to $h_i = 0.45\ \mathrm{m}, i \in \mathbb{I}_1^4$, and the proposed CHoKI-based controller is applied to the system. In order to compare the results, the same setup is applied to the other two controllers, whose models are 1) the ideal state-feedback set of ODEs and 2) the standard KI proposed in [14].

[2]Recall that the signals are scaled between 0 and 1.

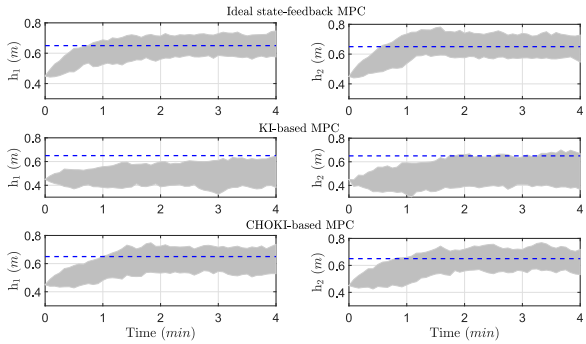


Fig. 2. Hundred closed-loop simulations of the quadruple-tank process, with three MPCs whose models are (top) the ideal state-feedback set of ODEs, (middle) the standard KI [14], and (bottom) the proposed CHoKI-based MPC.

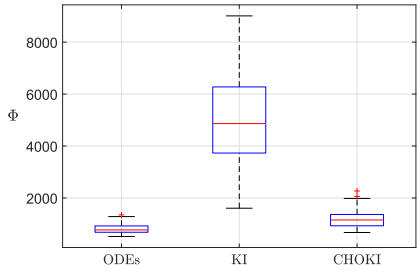


Fig. 3. Performance index comparison of the three controllers of Fig. 2.

The results are shown in Fig. 2, for 100 simulations subject to random noise. Besides, the performance of these simulations is measured according to

$$\Phi = \sum_{i=1}^{t_{\text{sim}}} \ell(x(i), u(i)) \tag{22}$$

which is compared in Fig. 3. Note that the data-based control problem is able to perform in a similar way to the ideal MPC, whereas the standard KI exhibits a worse performance, illustrating the main properties of CHoKI. Following the procedure in [13] results in $\gamma = 8100$, $\nu(\mu) = 49.72$, and $\phi = 4.45e5$.

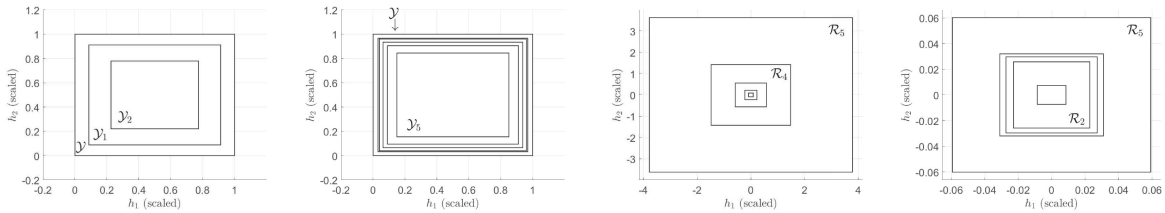
Note that the main advantage of CHoKI with respect to standard KI is not only the improvement of the prediction, which in general leads to better closed-loop performance results, but also the enlargement of the tightened constraints presented in Section III-A, which implies that the multivariate bound is less conservative. If we used the obtained $\mathcal{D}$ in the standard KI approach presented in [14], the prediction error would be $\mu^{\text{KI}} = [0.087\ 0.088]^2$ (which is a reduction of 36%). Then, the maximum prediction horizon such that Assumption 3 in [14] holds, i.e., that $\mathcal{Y}_N$ is not empty, is $N^{\text{KI}} = 2$, while $N^{\text{CHoKI}} = 8$.[3] The sets $\mathcal{Y}_j$ and $\mathcal{R}_j$ are represented in Fig. 4,[2] both for the CHoKI and the KI approaches. Note that $\|\mathcal{R}_5^{\text{CHoKI}}\| = 0.014$, while $\|\mathcal{R}_5^{\text{KI}}\| = 54.96$.[2]

Appendix

Proof of Theorem 1.1: Consider two parameters L and p such that $f : \mathcal{W} \to \mathcal{Y}$ is Hölder continuous. Provided that for a given $w \in \mathbb{R}^{n_w}$, $\|w\| \leq \sqrt{n_w}\|w\|_\infty$, define $L_\infty = L n_w^{(p/2)}$. Then, we have that

$$\|y_1 - y_2\| \leq L\|w_1 - w_2\|^p \tag{23a}$$

[3]$N = 5$ is chosen instead of $N = 8$ in order to not increase computational complexity in excess.



(a) (b) (c) (d)

Fig. 4. Sets $\mathcal{Y}_j$ and $\mathcal{R}_j$, both for the KI and the CHoKI approaches [3] in the quadruple-tank case. Note that $\mathcal{Y}_j$ is empty for $j \geq 3$ in the KI setup, and notice the different scales between the sets $\mathcal{R}_j$ in figures (c) and (d). (a) Sets $\mathcal{Y}_j$ with KI. (b) Sets $\mathcal{Y}_j$ with CHoKI. (c) Sets $\mathcal{R}_j$ with KI. (d) Sets $\mathcal{R}_j$ with CHoKI.

$$\leq L_\infty \|w_1 - w_2\|_\infty^p \tag{23b}$$
$$= L_\infty \max_j (|w_{1,j} - w_{2,j}|)^p \tag{23c}$$
$$\leq L_\infty \sum_{j=1}^{n_w} |w_{1,j} - w_{2,j}|^p \tag{23d}$$

where inequality (23d) holds provided that $|w_{1,j} - w_{2,j}|^p \geq 0$, $\forall j$. Then, defining $\mathcal{L} = L_\infty \mathbb{1}_{n_y \times n_w}$ and $\mathcal{P} = p\mathbb{1}_{n_y \times n_w}$, the function f is componentwise Hölder continuous. ■

Proof of Theorem 1.2: Note that for any $w_1, w_2, j \in \mathbb{I}_1^{n_w}$, we have that $|w_{1,j} - w_{2,j}| \leq \|w_1 - w_2\|_\infty$. Then

$$|y_{i,1} - y_{i,2}| \leq \sum_{j=1}^{n_w} \mathcal{L}_{i,j} |w_{1,j} - w_{2,j}|^{\mathcal{P}_{i,j}} \leq \sum_{j=1}^{n_w} \mathcal{L}_{i,j} \|w_1 - w_2\|_\infty^{\mathcal{P}_{i,j}}.$$

If $\mathcal{W}$ is a compact space, there exists a c such that $\|w\|_\infty \leq c$ for all $w \in \mathcal{W}$, so $\|w_1 - w_2\| \leq 2c$, $\forall w_1, w_2 \in \mathcal{W}$. Besides, if $x \in [0,\ 1]$ and $p_1 \geq p_2$, with $p_1, p_2 \in (0, 1]$, then $x^{p_1} \leq x^{p_2}$.

Let $p = \min_{i,j} \mathcal{P}_{i,j}$, and let $L = \max_i \sum_{j=1}^{n_y} \mathcal{L}_{i,j} (2c)^{\mathcal{P}_{i,j} - p}$. Then

$$\begin{aligned} \mathcal{L}_{i,j} \|w_1 - w_2\|_\infty^{\mathcal{P}_{i,j}} &\leq \mathcal{L}_{i,j} (2c)^{\mathcal{P}_{i,j}} \left(\tfrac{\|w_1 - w_2\|}{2c} \right)^{\mathcal{P}_{i,j}} \\ &\leq \mathcal{L}_{i,j} (2c)^{\mathcal{P}_{i,j}} \left(\tfrac{\|w_1 - w_2\|_\infty}{2c} \right)^{p} \\ &= \mathcal{L}_{i,j} (2c)^{\mathcal{P}_{i,j} - p} \|w_1 - w_2\|_\infty^p. \end{aligned}$$

Hence

$$\begin{aligned} |f_i(w_1) - f_i(w_2)| &\leq \sum_{j=1}^{n_y} \mathcal{L}_{i,j} \|w_1 - w_2\|_\infty^{\mathcal{P}_{i,j}} \\ &\leq \sum_{j=1}^{n_y} \mathcal{L}_{i,j} (2c)^{\mathcal{P}_{i,j} - p} \|w_1 - w_2\|_\infty^p \\ &\leq L \|w_1 - w_2\|_\infty^p. \end{aligned}$$

And since $\exists i^* : |f_{i^*}(w_1) - f_{i^*}(w_2)| = \|f(w_1) - f(w_2)\|_\infty$, we have that $\|f(w_1) - f(w_2)\|_\infty \leq L \|w_1 - w_2\|_\infty^p$. ■

Proof of Corollary 2: Using the Cauchy–Schwarz inequality

$$\mathfrak{d}_{\mathcal{L}}(|w|) = \langle \mathcal{L}, |w| \rangle \leq \|\mathcal{L}\|_\infty \|w\|_\infty = L \|w\|_\infty$$

where $\langle \cdot, \cdot \rangle$ denotes the canonical inner product. ■

Proof of Lemma 1: Given two scalars $x \geq 0$ and $0 < p \leq 1$, x^p is concave, so $\|x_1\|^p - \|x_2\|^p \leq \|x_1 - x_2\|^p$. Then, given $\mathcal{L}$, $\mathcal{P}$, $\mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w|)$ is Hölder continuous, since for each ith component

$$\sum_{j=1}^{n_w} \mathcal{L}_{i,j} |w_{1,j}|^{\mathcal{P}_{i,j}} - \sum_{j=1}^{n_w} \mathcal{L}_{i,j} |w_{2,j}|^{\mathcal{P}_{i,j}} \leq \sum_{j=1}^{n_w} \mathcal{L}_{i,j} (|w_{1,j} - w_{2,j}|)^{\mathcal{P}_{i,j}}.$$

The sum of two componentwise $\mathcal{L}$- $\mathcal{P}$-Hölder continuous functions f, g is also componentwise Hölder, since

$$\begin{aligned} |f(w_1) + g(w_1) - f(w_2) - g(w_2)| &\leq |f(w_1) - f(w_2)| + |g(w_1) - g(w_2)| \\ &\leq \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_1 - w_2|) + \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_1 - w_2|). \end{aligned}$$

Finally, the minimum (or equivalently, the maximum) of componentwise Hölder functions is also componentwise Hölder: Let denote $h(w) = \min(f, g)$ and assume, w.o.l.g., that $f(w_1) > g(w_2)$. Then, if $f(w_1) \leq g(w_1)$

$$\begin{aligned} |h(w_1) - h(w_2)| = |f(w_1) - g(w_2)| &= f(w_1) - g(w_2) \\ &\leq g(w_1) - g(w_2) \leq \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_1 - w_2|). \end{aligned}$$

Hence, $\hat{\mathfrak{f}}$ [c.f., (5)] is componentwise Hölder continuous, with Hölder parameters $\mathcal{L}$ and $\mathcal{P}$. ■

Lemma 6 (Sample consistency of CHoKI): If $\mathcal{L}$ and $\mathcal{P}$ are obtained as in (6), the CHoKI predictor (5) is sample-consistent (up to $\lambda/2$)

$$|\hat{\mathfrak{f}}(w_k) - f(w_k)| \leq \frac{\lambda}{2} + \bar{\mathfrak{e}},\ \forall w_k \in \mathcal{W}_{\mathcal{D}}. \tag{24}$$

Proof: Denote the indexes $i = \arg\min_n (\tilde{y}_n + \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_n|))$, and $j = \arg\max_n (\tilde{y}_n - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_n|))$. Then

$$\hat{\mathfrak{f}}(w_k) = \frac{1}{2} \underbrace{\left(\tilde{y}_i + \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_i|) \right)}_{B} + \frac{1}{2} \underbrace{\left(\tilde{y}_j - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_j|) \right)}_{A}.$$

It is first proven that $A \geq \tilde{y}_k$. If $j = k$ this is immediate. Otherwise

$$A = \tilde{y}_j - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_j|) \geq \tilde{y}_k - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_k|) = \tilde{y}_k.$$

Then, it is proven that $A \leq \tilde{y}_k + \lambda$. From (6b) we have that

$$|\tilde{y}_k - \tilde{y}_j| \leq \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_j|) + \lambda.$$

Provided that $\tilde{y}_j \geq A \geq \tilde{y}_k$

$$\begin{aligned} A &= \tilde{y}_j - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_j|) \\ &\leq \tilde{y}_k + |\tilde{y}_j - \tilde{y}_k| - \mathfrak{d}_{\mathcal{L}}^{\mathcal{P}}(|w_k - w_j|) \leq \tilde{y}_k + \lambda. \end{aligned}$$

The same procedure is applied to prove that $\tilde{y}_k - \lambda \leq B \leq \tilde{y}_k$. Then, $\hat{\mathfrak{f}}(w_k)$ is such that

$$\frac{1}{2}(\tilde{y}_k - \lambda) + \frac{1}{2}\tilde{y}_k \leq \hat{\mathfrak{f}}(w_k) \leq \frac{1}{2}(\tilde{y}_k + \lambda) + \frac{1}{2}\tilde{y}_k$$

or equivalently

$$\tilde{y}_k - \frac{\lambda}{2} \leq \hat{\mathfrak{f}}(w_k) \leq \tilde{y}_k + \frac{\lambda}{2},\ |\hat{\mathfrak{f}}(w_k) - \tilde{y}_k| \leq \frac{\lambda}{2}.$$

Finally, since $\tilde{y}_k \leq f(w_k) + \bar{\mathfrak{e}}$, $|\hat{\mathfrak{f}}(w_k) - f(w_k)| \leq \frac{\lambda}{2} + \bar{\mathfrak{e}}$. ■

Proof of Theorem 2: First, it is proven that $\mathcal{L}$ as a solution of (6) is bounded (i.e., not infinity). To this end, it is proven that $\mathcal{L}_f$ is a possible bounded solution that satisfies the constraint (6b). Indeed, note that since f is componentwise Hölder continuous for $(\mathcal{L}_f, \mathcal{P})$, $\mathcal{L}_f$ is bounded and for all $w_i \neq w_j \in \mathcal{W}$

$$|\tilde{y}_i - \tilde{y}_j| - \lambda \leq |y_i - y_j| \leq \mathfrak{d}^{\mathcal{P}}_{\mathcal{L}_f}(|w_i - w_j|)$$

which satisfies the condition.

Next, it is proven that the solution of (6), i.e.,

$$\Theta^* = \arg\min_{\Theta} g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}}) + \tau \| \mathcal{L} - \mathcal{L}_0 \|_1$$

is bounded even for infinitely dense data sets, as $N_{\mathcal{D}} \to \infty$.

To this end, $g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}})$ must be bounded for all $w \in \mathcal{W}$. For example consider

$$g(\Theta, \mathcal{D}, \mathcal{D}_{\text{test}}) = \frac{1}{N_{\mathcal{D}}} \sum_{w_i \in \mathcal{W}_{\mathcal{D}_{\text{test}}}} \|\hat{\mathfrak{f}}(w_i; \Theta, \mathcal{D}) - \tilde{y}_i\|^2.$$

Since $\mathcal{W}$ is compact, the noise is bounded and f is Hölder, then $\tilde{y}_i$ is bounded. Besides, Hölder continuity of $\hat{\mathfrak{f}}$ ensures that $\hat{\mathfrak{f}}(w, \Theta, \mathcal{D})$ is bounded for any Θ. Then, $g(\cdot)$ is upper-bounded by some k_1, irrespective of the number of data points.

Hence, assuming that $g(\cdot)$ is bounded by k_1, and taking into account that $\Theta_f = \{\mathcal{L}_f, \mathcal{P}\}$ is a feasible solution of the optimization problem, we have that

$$\begin{aligned} g(\Theta^*, \mathcal{D}, \mathcal{D}_{\text{test}}) \ &+ \ \tau \| \mathcal{L}^* - \mathcal{L}_0 \|_1 \\ &\leq \ g(\Theta_f, \mathcal{D}, \mathcal{D}_{\text{test}}) + \tau \| \mathcal{L}_f - \mathcal{L}_0 \|_1 \\ &\leq \ k_1 + \tau \| \mathcal{L}_f - \mathcal{L}_0 \|_1 = k_2 \end{aligned}$$

for certain constant k_2. Given that $g(\cdot)$ is positive, we have that

$$\tau \| \mathcal{L}^* - \mathcal{L}_0 \|_1 \leq g(\Theta^*, \mathcal{D}, \mathcal{D}_{\text{test}}) + \tau \| \mathcal{L}^* - \mathcal{L}_0 \|_1 \leq k_2.$$

Therefore, $\mathcal{L}$ is bounded for any size of the data set.

Finally, the bound given in the Theorem is proven. Denote $w_n = \arg\min_{w_j \in \mathcal{W}_{\mathcal{D}}}(|w_j - w|)$ the closest data point to the query w. The following operation is decomposed into three addends:

$$\begin{aligned} |f(w) - \hat{\mathfrak{f}}(w)| = \ & |f(w) - f(w_n)| \\ & + |\hat{\mathfrak{f}}(w_n) - \hat{\mathfrak{f}}(w)| + |f(w_n) - \hat{\mathfrak{f}}(w_n)|. \end{aligned}$$

The first term is less than or equal to $\mathfrak{d}^{\mathcal{P}}_{\mathcal{L}_f}(|w - w_n|)$, which is bounded by $\mathfrak{d}^{\mathcal{P}}_{\mathcal{L}_f}(\mathcal{R}_{\mathcal{D}})$. The second term is less than or equal to $\mathfrak{d}^{\mathcal{P}}_{\mathcal{L}}(|w - w_n|) + \lambda$, which is bounded by $\mathfrak{d}^{\mathcal{P}}_{\mathcal{L}}(\mathcal{R}_{\mathcal{D}}) + \lambda$. The third term is less than or equal to $\lambda/2 + \bar{\mathfrak{e}}$, as proven in Lemma 6. Then, the three bounds add together, proving (8). ■

Proof of Lemma 2: The proof follows developing $\mathfrak{d}^{\mathcal{P}}_{\mathcal{L}}(|q - w_i|)$ in (5) using the definition in (3) and bearing in mind that $\mathcal{L}_j = L$, for all column $j \in \mathbb{I}_1^{n_w}$. Hence, using the one norm for the standard KI

$$L\|q - w_i\| = L\sum_{j=1}^{n_w} |q - w_i| = \sum_{j=1}^{n_w} \mathcal{L}_j |q - w_i| = \mathfrak{d}^{\mathcal{P}}_{\mathcal{L}}(|q - w_i|).$$

■

REFERENCES

[1] J. R. Salvador, D. R. Ramirez, T. Alamo, and D. Muñoz de la Peña, "Offset free data driven control: Application to a process control trainer," *IET Control Theory Appl.*, vol. 13, no. 18, pp. 3096–3106, 2019.

[2] J. R. Salvador, D. Muñoz de la Peña, D. Ramirez, and T. Alamo, "Predictive control of a water distribution system based on process historian data," *Optimal Control Appl. Methods*, vol. 41, no. 2, pp. 571–586, 2020.

[3] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2737–2752, Jul. 2019.

[4] M. Maiworm, D. Limon, J. M. Manzano, and R. Findeisen, "Stability of gaussian process learning based output feedback model predictive control," *IFAC-PapersOnLine*, vol. 51, no. 20, pp. 455–461, 2018.

[5] D. Yu and J. Gomm, "Implementation of neural network predictive control to a multivariable chemical reactor," *Control Eng. Pract.*, vol. 11, no. 11, pp. 1315–1323, 2003.

[6] G. Beliakov, "Interpolation of lipschitz functions," *J. Comput. Appl. Math.*, vol. 196, no. 1, pp. 20–44, 2006.

[7] A. Sukharev, "Optimal method of constructing best uniform approximations for functions of a certain class," *USSR Comput. Math. Math. Phys.*, vol. 18, no. 2, pp. 21–31, 1978.

[8] M. Canale, L. Fagiano, and M. C. Signorile, "Nonlinear model predictive control from data: A set membership approach," *Int. J. Robust Nonlinear Control*, vol. 24, no. 1, pp. 123–139, 2014.

[9] J.-P. Calliess, "Bayesian lipschitz constant estimation and quadrature," in *Proc. Workshop Probabilistic Integration 29th Conf. Neural Inf. Process. Syst.*, 2015., pp. 1–5.

[10] J.-P. Calliess, "Conservative decision-making and inference in uncertain dynamical systems," Ph.D. dissertation, Dept. Eng. Sci., Univ. Oxford, Oxford, U.K., 2014.

[11] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *Proc. IEEE Conf. Decis. Control*., 2018, pp. 6059–6066.

[12] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annu. Rev. Control, Robot. Auton. Syst.*, vol. 3, pp. 269–296, 2020.

[13] J. M. Manzano, D. Limon, D. Muñoz de la Peña, and J.-P. Calliess, "Output feedback MPC based on smoothed projected kinky inference," *IET Control Theory Appl.*, vol. 13, no. 6, pp. 795–805, 2019.

[14] J. M. Manzano, D. Limon, D. Muñoz de la Peña, and J.-P. Calliess, "Robust learning-based MPC for nonlinear constrained systems,"*Automatica*, vol. 117,2020, Art. no. 108948.

[15] M. Lorenzen, F. Dabbene, R. Tempo, and F. Allgöwer, "Stochastic MPC with offline uncertainty sampling," *Automatica*, vol. 81, pp. 176–183, 2017.

[16] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious model predictive control using gaussian process regression," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2736–2743, Nov. 2020.

[17] J. M. Manzano, D. Limon, D. Muñoz de la Peña, and J.-P. Calliess, "Data-based robust MPC with componentwise hölder kinky inference," in *Proc. IEEE 58th Conf. Decis. Control*., 2019, pp. 6449–6454.

[18] J.-P. Calliess, S. J. Roberts, C. E. Rasmussen, and J. Maciejowski, "Lazily adapted constant kinky inference for nonparametric regression and model-reference adaptive control," *Automatica*, vol. 122, 2020, Art. no. 109216.

[19] J.-P. Calliess, "Lipschitz Optimisation for Lipschitz Interpolation," *Amer. Control Conf.*, 2017, pp. 3141–3146, 2017.

[20] T. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill Higher Education, 1997.

[21] K. H. Johansson, "The quadruple-tank process: A multivariable laboratory process with an adjustable zero," *IEEE Trans. Control Syst. Technol.*, vol. 8, no. 3, pp. 456–465, May 2000.

[22] I. Alvarado *et al.*, "A comparative analysis of distributed MPC techniques applied to the HD-MPC four-tank benchmark," *J. Process Control*, vol. 21, no. 5, pp. 800–815, 2011.