2020 European Control Conference (ECC)
May 12-15, 2020. Saint Petersburg, Russia

Enhanced kernels for nonparametric identification of a class of nonlinear systems

Mirko Mazzoleni, Matteo Scandella, Simone Formentin and Fabio Previdi

***Abstract*— This paper deals with nonparametric nonlinear system identification via Gaussian process regression. We show that, when the system has a particular structure, the kernel recently proposed in [1] for nonlinear system identification can be enhanced to improve the overall modeling performance. More specifically, we modify the definition of the kernel by allowing different orders for the exogenous and the autoregressive parts of the model. We also show that all the hyperparameters can be estimated by means of marginal likelihood optimization. Numerical results on two benchmark simulation examples illustrate the effectiveness of the proposed approach.**

I. Introduction

The introduction of nonparametric kernel-based methods for linear system identification provided a paradigm-shift in the field [2] and in several practical applications, see, e.g., [3]. Instead of adopting a-priori parameterized models to search for the best solution within a finite-dimensional space, Reproducing Kernel Hilbert Spaces (RKHSs) [4] can be addressed, thus allowing to reconstruct the full infinite-length impulse response. Such spaces are defined by, and, in turn, define, a proper kernel function. Exploiting the connection of RKHS with the framework of Gaussian Processes (GPs) [5], the unknown kernel hyperparameters can be estimated by empirical Bayes (marginal likelihood) or SURE procedures [6]. This is the counterpart of model order selection for standard parametric methods, but, in most cases, it turns out to be more effective in trading bias and variance of the model due to its *continuous* nature [7]. This is especially important when few data are available [8], since many properties of the parametric Prediction Error Approach (PEM) method are only asymptotic [9] (i.e., they hold only when the number of data tends to infinity).

Kernel-based methods can be effectively employed also for nonlinear system identification [1], [10], [11], [12], [13] when little prior information is available on the model structure. In [1], the authors proposed a kernel-based method where the unknown system is considered as a realization of a zero-mean Gaussian random field f. The minimum variance estimator of f is available in closed form and belongs to a RKHS defined by a proper kernel (covariance) function. The kernel devised in [1] is given by a weighted sum of Gaussian kernels, where each weight encodes the "fading memory" property of dynamical systems. A specific hyperparameter tunes the degree of interaction between past inputs and outputs that, however, is constrained to be *the same* for both the exogenous and autoregressive parts.

In this work, we propose an adaptation of the kernel function introduced in [1] for a specific class of nonlinear systems in which input and output regressors affect the system dynamics *separately*. Such systems include, e.g., control-affine systems like $y_{t+1} = a(y_t) + bu_t$, where y_t, u_t denote the output and input at time t, a is a generic function of the output [14] and b is a real constant. Notice that control-affine systems like the one above are widely employed, especially in robotic motion control, see [15] and reference therein.

The proposed kernel allows for *different degrees of interactions*, between previous inputs and outputs, to be included in the model. If the model is composed only by the exogenous (or the autoregressive) part, the proposed kernel restores to the one in [1]. The best interaction degrees are estimated from data, not requiring the user to select any critical variable such as regressors, or model, orders.

The remainder of the paper is organized as follows. Section II formulates the nonlinear system identification problem. Section III presents the proposed enhanced kernel for the identification of the considered class of systems. In Section IV, the algorithm for obtaining a Bayesian estimate of the dynamics is briefly reviewed. Section V characterizes the space of functions induced by the enhanced kernel. Section VI compares the performance of the proposed kernel with the benchmark. Lastly, Section VII is devoted to concluding remarks and future developments.

II. Problem statement

Suppose to have at disposal a set $\{y_t\}$ of noisy output measurements from a nonlinear dynamical system fed with an input $\{u_t\}$, with $t \in \mathbb{Z}$. The set of past inputs and outputs at time t is defined as

$$y^t \equiv \begin{bmatrix} y_{t-1} & y_{t-2} & \dots \end{bmatrix}^\top, \quad u^t \equiv \begin{bmatrix} u_{t-1} & u_{t-2} & \dots \end{bmatrix}^\top.$$

Let N be the number of measured data. The vector of measured outputs is defined as:

$$y^+ \equiv \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^\top,$$

while the vector of past outputs at time $t = 1$ is defined as $y^- \equiv y^1 = \begin{bmatrix} y_0 & y_{-1} & \dots \end{bmatrix}^\top$. Be $\mathbb{N}$ the set of natural numbers without 0. Given two time instants t and $\tau \in \mathbb{Z}$, we

M. Mazzoleni, M. Scandella and F. Previdi are with the Department of Management, Information and Production Engineering, University of Bergamo, Via G. Marconi 5, 24044 Dalmine (BG), Italy.

S. Formentin is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, via G. Ponzio 34/5, 20133 Milano, Italy.

Email to: mirko.mazzoleni@unibg.it.

define, for $i \in \mathbb{N}$, the variables $x_i, z_i \in \mathbb{R}^2$ to be

$$\begin{aligned} x_1 &\equiv [y_{t-1} \quad u_{t-1}], & z_1 &\equiv [y_{\tau-1} \quad u_{\tau-1}] \\ x_2 &\equiv [y_{t-2} \quad u_{t-2}], & z_2 &\equiv [y_{\tau-2} \quad u_{\tau-2}] \\ &\vdots & &\vdots \end{aligned} \tag{1}$$

with $x \in \mathbb{R}^{\infty\times 2}$ and $z \in \mathbb{R}^{\infty\times 2}$ sequences given by

$$x \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}, \quad z \equiv \begin{bmatrix} z_1 \\ z_2 \\ \vdots \end{bmatrix}. \tag{2}$$

We assume that the one-step ahead predictor

$$\hat{y}_{t|t-1} = F(y^t, u^t) \tag{3}$$

is time invariant and strictly causal, i.e. does not depend on u_t. Then, it is possible to describe the unknown nonlinear dynamical systems with the model

$$y_t = F(y^t, u^t) + e_t, \quad t = 1, \ldots, N \tag{4}$$

where $\{e_t\}$ is a white-noise sequence that represents the one-step ahead prediction error $e_t \equiv y_t - \hat{y}_{t|t-1}$, $e_t \perp e_s, t \neq s$, $e_t \perp u_s \ \forall t, s$. Furthermore, we assume that e_t is Gaussian with zero mean and unknown, but constant, variance, i.e. $e_t \sim \mathcal{N}\left(0, \eta^2\right)$.

The aim now is to estimate $F : \mathbb{R}^{\infty\times 2} \to \mathbb{R}$ in (4). In order to do so, F is interpreted as a realization of a zero-mean Gaussian random field f[1]. In virtue of the Gaussianity assumption, defining f requires only to specify the covariance between the random variables $f(x)$ and $f(z)$, where x, z are any couple of possible arguments for F. In our setting, the input locations (regressors), defined in (2), consist of past inputs and outputs (y^t, u^t). The training set is $\{(y^t, u^t), y_t\}$ for $t = 1, \ldots, N$.

Let the covariance of f be $\mathbb{E}[f(x)f(z)]$ (remember that f has zero mean), where $\mathbb{E}$ is the expectation operator. Then, the covariance function can be equivalently defined by the *kernel function*

$$G : \mathbb{R}^{\infty\times 2} \times \mathbb{R}^{\infty\times 2} \to \mathbb{R}. \tag{5}$$

The identification of F in (4) is pursed in a Gaussian processes framework [5], where (5) defines a prior over the space of possible functions to be learned. Once the hyperparameters of (5) have been estimated by marginal likelihood optimization [5], [16], the nonlinear model is obtained as the minimum variance estimate of the Gaussian process model. The estimate can also be interpreted as the solution of a Tikhonov-type variational problem, defined on the RKHS induced by (5), see [17] for a connection between Gaussian processes and RKHS.

This work focuses on the identification of systems in the following form:

$$\mathcal{S} : y_t = a\left(y^t\right) + b\left(u^t\right), \tag{6}$$

where a, b are nonlinear functions of only past outputs and inputs, respectively. If a or b are zero, we have the NFIR (Nonlinear Finite Impulse Response) and NAR (Nonlinear AutoRegressive) case.

In the following sections, we will first review the kernel function, of the type of (5), introduced in [1]. Then, we propose an adaptation of this previously defined kernel that is able to better handle specific system identification problems.

III. Kernels for nonlinear system identification

A. The mixture of Gaussian kernel

The kernel proposed in [1] consists in a weighted sum, for $t = 1, \ldots, \infty$, of Gaussian kernels of the form

$$\begin{aligned} K_t(x, z; p, \psi) &= \mathbb{E}[f_t(x) f_t(z)] \\ &\equiv \exp\left(-\frac{\sum_{j=1}^{p} \left\| x_{t+j-1}^{(:)} - z_{t+j-1}^{(:)} \right\|^2}{\sigma^2}\right) \\ p &\in \mathbb{N}, \sigma \in \mathbb{R}^+, \end{aligned} \tag{7}$$

where f_t are zero-mean independent Gaussian random fields, K_t is the covariance of f_t and $x, z \in \mathbb{R}^{\infty\times 2}$ are generic kernel arguments as defined in (2). The symbol $^{(:)}$ indicates that all the elements of the vectors x_{t+j-1}, z_{t+j-1} are taken into consideration, i.e. $\left\| x_1^{(:)} \right\|^2 = \left\| [y_{t-1} \quad u_{t-1}] \right\|^2$, where $\|\cdot\|^2$ is the Euclidean norm. The *Mixture of Gaussian kernel* is therefore defined as [1]:

$$K(x, z; p, \psi) \equiv \sum_{t=1}^{\infty} \beta_t \cdot K_t(x, z; p, \psi) \quad p \in \mathbb{N} \tag{8a}$$

$$\beta_t \equiv \lambda_1 e^{-t\cdot\lambda_2}, \quad \lambda_1, \lambda_2 \in \mathbb{R}^+. \tag{8b}$$

The function in (8) depends on the hyperparameters $\psi = [\lambda_1, \lambda_2, \sigma]$ and p. We have that:

- σ is the standard deviation of the Gaussian kernel, i.e. the kernel width;
- λ_1 and λ_2 define β_t, that weights the influence of past data on the output y_t. As we go back in time, it is assumed that past data have less and less influence on the current outcome;
- p accounts for the order of interaction between past input and past output data. Notice how, in the definition (7), both inputs and outputs have *the same* degree of interaction.

The Gaussian random field f is therefore modeled as $f = \sum_{t=1}^{\infty} f_t$. As an example, consider $p = 2$. Then, the kernel in (8) reads as:

$$\begin{aligned} &K(x, z; 2, \psi) \\ &= \beta_1 \cdot \exp\left(-\frac{\left\| x_1^{(:)} - z_1^{(:)} \right\|^2 + \left\| x_2^{(:)} - z_2^{(:)} \right\|^2}{\sigma^2}\right) \\ &+ \beta_2 \cdot \exp\left(-\frac{\left\| x_2^{(:)} - z_2^{(:)} \right\|^2 + \left\| x_3^{(:)} - z_3^{(:)} \right\|^2}{\sigma^2}\right) + \ldots \end{aligned}$$

[1] A one-dimensional Gaussian random field is a Gaussian process.

In this case we have that f, with covariance $K(x,z;2,\psi)$, admits the representation

$$f(y^t,u^t) = f_1(y_{t-1},u_{t-1},y_{t-2},u_{t-2}) + f_2(y_{t-2},u_{t-2},y_{t-3},u_{t-3}) + \dots, \tag{9}$$

that is, a system were nonlinear interactions between inputs and outputs are present on variables at different times.

Again, notice how in (9) the past inputs and outputs are constrained to be included with the same degree, i.e. it is not possible to have, as an example, a random field of the form $f_1(y_{t-1},u_{t-1},y_{t-2})$.

B. Enhanced kernel for the considered class of systems

As showed in (9), the kernel (8) strives for interactions between u and y. Consequently, if the considered class of systems is that of (6), a more tailored kernel can be employed that does not consider those interactions and allows different order of past inputs and outputs to be included in the model. In this view, we propose the following covariance function S, that we call *enhanced kernel*:

$$\begin{aligned} S\left(x,z;p^y,p^u,\overline{\psi}\right) &\equiv \sum_{t=1}^{\infty}\beta_t^y\cdot K_t^y\left(x,z;p^y,\overline{\psi}\right) \\ &\quad+\sum_{t=1}^{\infty}\beta_t^u\cdot K_t^u\left(x,z;p^u,\overline{\psi}\right), \\ \beta_t^y \equiv \lambda_1^y e^{-t\cdot\lambda_2^y},&\quad \beta_t^u \equiv \lambda_1^u e^{-t\cdot\lambda_2^u}, \\ \lambda_1^y,\lambda_1^u,\lambda_2^y,\lambda_2^u\in\mathbb{R}^+,&\quad p^y,p^u\in\mathbb{N}, \end{aligned} \tag{10}$$

where p^y, p^u define the order of interaction between past inputs and past outputs, respectively, and $\lambda_1^y,\lambda_1^u,\lambda_2^y,\lambda_2^u$ have the same role as in (8b). The $K_t^y\left(x,z;p^y,\overline{\psi}\right)$ and $K_t^u\left(x,z;p^u,\overline{\psi}\right)$ are kernels defined as

$$K_t^y\left(x,z;p^y,\overline{\psi}\right) \equiv \exp\left(-\frac{\sum_{j=1}^{p^y}\left\|x_{t+j-1}^{(1)}-z_{t+j-1}^{(1)}\right\|^2}{\left(\sigma^y\right)^2}\right), \tag{11a}$$

$$K_t^u\left(x,z;p^u,\overline{\psi}\right) \equiv \exp\left(-\frac{\sum_{j=1}^{p^u}\left\|x_{t+j-1}^{(2)}-z_{t+j-1}^{(2)}\right\|^2}{\left(\sigma^u\right)^2}\right), \tag{11b}$$

where the apexes $^{(1)}$ and $^{(2)}$ indicate that only the first and second elements of the vectors x_{t+j-1}, z_{t+j-1} are used, i.e. only the y's or the u's enter in the kernel functions K_t^y and K_t^u, respectively. The hyperparameters, except p^y and p^u, are contained in the vector $\overline{\psi}=[\lambda_1^y,\lambda_1^u,\lambda_2^y,\lambda_2^u,\sigma^y,\sigma^u]$.

The kernel function in (10) permits to treat in an independent way the contributions of the exogenous and the autoregressive part of the model. Furthermore, for models of type (6), it allows for different order of interaction between past inputs and outputs, via the hyperparameters p^u and p^y. The Gaussian random field f is therefore modeled as

$$f=\textstyle\sum_{t=1}^{\infty}f_t^y+\sum_{t=1}^{\infty}f_t^u,$$

where K_t^y, K_t^u are the covariances of the zero-mean independent Gaussian random fields f_t^y, f_t^u, respectively. As an example, considering the kernel $S\left(x,z;2,3,\overline{\psi}\right)$, we can impose the following representation for f:

$$\begin{aligned} f(y^t,u^t) &= f_1^y(y_{t-1},y_{t-2})+f_2^y(y_{t-2},y_{t-3})+\dots \\ &+ f_1^u(u_{t-1},u_{t-2},u_{t-3})+f_2^u(u_{t-2},u_{t-3},u_{t-4})+\dots \end{aligned} \tag{12}$$

It is interesting to notice that, when is known that there is dependence only on previous inputs or only on previous outputs, the kernel in (10) is analogous to that in (8).

IV. Bayesian nonparametric algorithm for nonlinear system identification

In this section, we briefly review the Bayesian nonparametric identification scheme outlined in [1]. This rationale will be applied also to the proposed enhanced kernel. Let θ denote the vector containing the standard deviation of the noise and the kernel hyperparameters, i.e. $\theta=[\eta,\psi]$ when the covariance of f is K in (8), while $\theta=[\eta,\overline{\psi}]$ when the kernel S in (10) is used. The vector θ is considered a random variable, independent from p, with poorly informative prior such that non-negativity of its components is ensured. The parameter p is modeled as a random variable as well: its distribution assigns equal probability to each of the (arbitrarily large) discrete values of p. Since, in practice, the values of y^- are not completely known, the unknown components are set to zero (as done when initializing linear parametric predictors, see Section 3.2 of [9]). In view of this, the density $\mathbf{p}\left(f,\theta,p|y^-,u\right)$ is approximated as $\mathbf{p}\left(f,\theta,p|u\right)$. The joint density of y^+,f,θ and p is therefore approximated as (we omit the dependence of each density on the input u):

$$\mathbf{p}\left(y^+,f,\theta,p|y^-\right)\approx\mathbf{p}\left(y^+|f,\theta,p,y^-\right)\mathbf{p}\left(f|\theta,p\right)\mathbf{p}\left(\theta,p\right). \tag{13}$$

By relying on approximation (13) and assuming (8) as covariance of f, it holds that the minimum variance estimate of f, for known y^+,y^-,θ,p,u is [1]

$$\begin{aligned} \hat{f}(x) &= \mathbb{E}\left[f(x)|y^+,y^-,\theta,p,u\right] \\ &= \sum_{t=1}^{N}c_t K\left(x,\left(y^t,u^t\right);p\right), \end{aligned} \tag{14}$$

where x is a generic input location, c_t is the t-th component of the vector

$$c=\left(\Sigma_y\left(p,\theta\right)\right)^{-1}y^+, \tag{15}$$

and $\Sigma_y\in\mathbb{R}^{N\times N}$ is an invertible matrix with (i,j)-entry given by

$$\left[\Sigma_y\right]_{i,j}=K\left(\left(y^i,u^i\right),\left(y^j,u^j\right);p\right)+\eta^2\delta_{ij}, \tag{16}$$

where δ_{ij} is the Kronecker delta. Furthermore, we have that:

$$\mathbf{p}(y^+|y^-,\theta,p,u)=\frac{\exp\left(-\frac{1}{2}\left(y^+\right)^\top\left(\Sigma_y\left(p,\theta\right)\right)^{-1}y^+\right)}{\sqrt{\det\left(2\pi\Sigma_y\left(p,\theta\right)\right)}}, \tag{17}$$

that is, the marginal likelihood of observed data y^+, conditioned on past outputs, the system input and hyperparameters, is Gaussian.

To provide an estimate for θ and p, the following rationale is adopted [1]. Define:

$$\theta_p = \arg\min_{\theta} J_p(\theta), \tag{18}$$

where

$$J_p(\theta) \equiv -\log \mathbf{p}\left(y^+|y^-, \theta, p, u\right). \tag{19}$$

Minimizing (19) we obtain an estimate $\hat{\theta}_p$ for each discrete value in the set $\{1, \dots, p\}$. Then, the best p is obtained as:

$$\hat{p} = \arg\min_{p} J_p\left(\hat{\theta}_p\right). \tag{20}$$

The estimator of the hyperparameters vector θ is therefore $\hat{\theta}_{\hat{p}}$. The prediction at a generic input location x is thus computed by using (14) with $\hat{\theta}_{\hat{p}}$ and $\hat{p}$.

Remark. The derivations in (14) - (17) are still valid if the kernel S is employed in place of K. The choice of the best interaction values is performed as in (18)-(20), for all values in the set given by the Cartesian product $\{1, \dots, p^y\} \times \{1, \dots, p^u\}$.

V. Characterization of the RKHS generated by the enhanced kernel

The minimum variance estimate in (14) admits a deterministic counterpart that exploits the *representer theorem*. The estimate is given by the solution of the following variational problem [17] (assuming that the enhanced kernel S is employed):

$$\hat{f} = \arg\min_{s \in \mathcal{H}_S} \sum_{i=1}^{N} \left(y_i - s\left(y^i, u^i\right)\right)^2 + \eta^2 \|s\|^2_{\mathcal{H}_S} \tag{21}$$

where $\mathcal{H}_S$ is the RKHS induced by S in (10), and $s \in \mathcal{H}_S$ is an unknown function to be learned. Therefore, the properties of the function $\hat{f}$ are defined by those of the space of functions $\mathcal{H}_S$. In particular, $\mathcal{H}_S$ is the space of functions given by the completion, w.r.t. the inner product

$$\left\langle \sum_i m_i S(\cdot, x_i), \sum_j n_j S(\cdot, x_j) \right\rangle_{\mathcal{H}_S} = \sum_{i,j} m_i n_j S(x_i, z_j), \tag{22}$$

of the linear span $\sum_{i=1}^{l} m_i S(\cdot, x_i)$, for all choices of $l, \{m_i\}$ and $\{x_i\}$.

A. Characterization of the RKHS $\mathcal{H}_K$

To start with, we briefly review the characterization of the RKHS $\mathcal{H}_K$ associated with the kernel K in (8). Define $x \equiv [x_1, \dots, x_n]^\top \in \mathbb{R}^n$, $z \equiv [z_1, \dots, z_n]^\top \in \mathbb{R}^n$. The kernel $K(x, z)$ is assumed to be composed of n mixtures. Then, omitting the dependence on ψ, for $p = \{1, \dots, n\}$:

$$K(x, z;\, n, p) = \sum_{j=1}^{n-p+1} \beta_j \exp\left(-\frac{\sum_{i=1}^{p}\left(x_{i+j-1} - z_{i+j-1}\right)^2}{\sigma^2}\right). \tag{23}$$

Define also $(j) = (j, \dots, j+p-1)$ and the j-th component kernel as

$$K_j(x, z\,; n, p) = e^{-\|x_{(j)} - z_{(j)}\|^2/\sigma^2}, \tag{24}$$

so that

$$K(x, z;\, n, p) = \sum_{j=1}^{n-p+1} \beta_j K_j(x, z\,; n, p). \tag{25}$$

The characterization of the RKHS $\mathcal{H}_K$ is then given by the following theorem [1]:

Theorem 1: Let $X \subset \mathbb{R}^n$ any set with non-empty interior, $x, z \in X$, $K : X \times X \to \mathbb{R}$. Then, $\mathcal{H}_K$ is the direct orthogonal sum of the spaces $\mathcal{H}_{K_j}$ induced by the kernels K_j:

$$\mathcal{H}_K = \oplus_{j=1}^{n+p-1} \mathcal{H}_{K_j}. \tag{26}$$

In Theorem 1, we have that $\mathcal{H}_{K_j} = \otimes_{i=1}^{p} \mathcal{H}_{K_{ji}}$ is given by the tensor product of the spaces associated with the single univariate kernels $K_{ji} = e^{-(x_{i+j-1} - z_{i+j-1})^2/\sigma^2}$ that compose K_j through $K_j = \beta_j \prod_{i=1}^{p} K_{ji}$. It follows from [4] that each $r \in \mathcal{H}_K$ can be expressed as $r = \sum_{j=1}^{n-p+1} r_j$ with $r_j \in \mathcal{H}_{K_j}$, with the norm (see [1] for the exact definition of this norm):

$$\|r\|^2_{\mathcal{H}_K} = \sum_{j=1}^{n-p+1} \|r_j\|^2_{\mathcal{H}_{K_j}} < \infty. \tag{27}$$

The general case where x, z are defined as in (1)-(2) leads to completely analogous conclusion, with $\mathcal{H}_{K_j} = \otimes_{i=1}^{2p} \mathcal{H}_{K_{ji}}$, since now both u and y contribute to the multivariate Gaussian kernel K_j.

B. Characterization of the RKHS $\mathcal{H}_S$

As previously done, we assume that the kernel $S(x, z)$ is composed by only n mixtures, with $x, z \in \mathbb{R}^{n \times 2}$ defined as in (1)-(2) but truncated at length n. Analogously to (24), define

$$K^y_j(x, z\,; n, p^y) = e^{-\left\|x^{(1)}_{(j)} - z^{(1)}_{(j)}\right\|^2/(\sigma^y)^2}, \tag{28a}$$

$$K^u_j(x, z\,; n, p^u) = e^{-\left\|x^{(2)}_{(j)} - z^{(2)}_{(j)}\right\|^2/(\sigma^u)^2}, \tag{28b}$$

where K^y_j, K^u_j induce the RKHS spaces $\mathcal{H}^y_{K_j}, \mathcal{H}^u_{K_j}$, which functions depend on past outputs and inputs, respectively. The kernel in (10) can be rewritten in a way similar to (25):

$$S(x, z;\, n, p^y, p^u) = \sum_{j=1}^{n-p^y+1} \beta^y_j K^y_j(x, z;\, n, p^y) \tag{29a}$$

$$+ \sum_{j=1}^{n-p^u+1} \beta^u_j K^u_j(x, z;\, n, p^u). \tag{29b}$$

It is immediate to notice that the $\mathcal{H}_{K^y_j}, \mathcal{H}_{K^u_j}$ are completely analogous to the $\mathcal{H}_{K_j}$, i.e. $\mathcal{H}_{K^y_j} = \otimes_{i=1}^{p^y} \mathcal{H}_{K^y_{ji}}$ and $\mathcal{H}_{K^u_j} = \otimes_{i=1}^{p^u} \mathcal{H}_{K^u_{ji}}$, with $\mathcal{H}_{K^y_{ji}}, \mathcal{H}_{K^u_{ji}}$ the RKHS defined by the single univariate kernels $K^y_{ji} = e^{-\left(x^{(1)}_{i+j-1} - z^{(1)}_{i+j-1}\right)^2/(\sigma^y)^2}$

and $K^u_{ji} = e^{-\left(x^{(2)}_{i+j-1} - z^{(2)}_{i+j-1}\right)^2/(\sigma^u)^2}$, respectively. From Theorem 1, it follows that

$$\mathcal{H}_{K^y} = \oplus_{j=1}^{n+p^y-1} \mathcal{H}_{K^y_j}, \tag{30a}$$

$$\mathcal{H}_{K^u} = \oplus_{j=1}^{n+p^u-1} \mathcal{H}_{K^u_j}, \tag{30b}$$

where $\mathcal{H}_{K^y}, \mathcal{H}_{K^u}$ are the RKHS induced by the components (29a) and (29b) respectively.

Since the kernels K^y_j, K^u_j depend on different and non-overlapping domains, we have that $\mathcal{H}_{K^y} \cap \mathcal{H}_{K^u} = \emptyset$. Therefore, the space generated by the kernel S in (10) is:

$$\mathcal{H}_S = \mathcal{H}_{K^y} \oplus \mathcal{H}_{K^u} \tag{31}$$

Each $s \in \mathcal{H}_S$ can be expressed as $s = \sum_{j=1}^{n-p^y+1} s^y_j + \sum_{j=1}^{n-p^u+1} s^u_j$ with $s^y_j \in \mathcal{H}_{K^y_j}, s^u_j \in \mathcal{H}_{K^u_j}$, with the norm:

$$\|s\|^2_{\mathcal{H}_S} = \sum_{j=1}^{n-p^y+1} \|s^y_j\|^2_{\mathcal{H}_{K^y_j}} + \sum_{j=1}^{n-p^u+1} \|s^u_j\|^2_{\mathcal{H}_{K^u_j}}, \tag{32}$$

$$= \|s^y\|^2_{\mathcal{H}_{K^y}} + \|s^u\|^2_{\mathcal{H}_{K^u}} < \infty, \tag{33}$$

with $\|s^y\|^2_{\mathcal{H}_{K^y}}, \|s^u\|^2_{\mathcal{H}_{K^u}}$ defined as in (27).

In order to clarify the difference between the spaces $\mathcal{H}_K$ in (26) and $\mathcal{H}_S$ in (31), consider the following example. Suppose that $p^y = p^u = p$, and x, z defined as in (28). It follows that (29) can be written as

$$S(x, z;\, n, p) = \sum_{j=1}^{n-p+1} \beta^y_j K^y_j(x, z;\, n, p) + \beta^u_j K^u_j(x, z;\, n, p). \tag{34}$$

Assume now that $n = p^y = p^u = 2$, from (25) and (29). Then, $j = 1$ and, respectively:

$$K(x, z;\, 2, 2) = \beta_1 e^{-\frac{(y_{t-1}-y_{\tau-1})^2+(y_{t-2}-y_{\tau-2})^2}{\sigma^2}} \cdot \beta_1 e^{-\frac{(u_{t-1}-u_{\tau-1})^2+(u_{t-2}-u_{\tau-2})^2}{\sigma^2}} \tag{35}$$

$$S(x, z;\, 2, 2, 2) = \beta^y_1 e^{-\frac{(y_{t-1}-y_{\tau-1})^2+(y_{t-2}-y_{\tau-2})^2}{(\sigma^y)^2}} + \beta^u_1 e^{-\frac{(u_{t-1}-u_{\tau-1})^2+(u_{t-2}-u_{\tau-2})^2}{(\sigma^u)^2}} \tag{36}$$

so that, for the case (35) we have that

$$\mathcal{H}_{K_1} = \mathcal{H}_{K^y_{11}} \otimes \mathcal{H}_{K^y_{12}} \otimes \mathcal{H}_{K^u_{11}} \otimes \mathcal{H}_{K^u_{12}}, \tag{37}$$

while for the case (36) it holds

$$\mathcal{H}_{S_1} = \left(\mathcal{H}_{K^y_{11}} \otimes \mathcal{H}_{K^y_{12}}\right) \oplus \left(\mathcal{H}_{K^u_{11}} \otimes \mathcal{H}_{K^u_{12}}\right), \tag{38}$$

where $\mathcal{H}_{K_j}$ is the RKHS generated by the kernel K_j in (24), $\mathcal{H}_{K^y_{ji}}, \mathcal{H}_{K^u_{ji}}$ are, as previously defined, the RKHS induced by the single univariate Gaussian kernels on y and u, respectively, and $\mathcal{H}_{S_j}$ is the RKHS associated with the j-th component of (34). Again, notice that if only the exogenous or autoregressive part is present, the space (38) is identical to (37), so that the kernel S reduces to kernel K.

In cases where $p^y = p^u = p = 2, n > p$, the complete space generated by the kernels K and S can be written as

$$\mathcal{H}_K = \oplus_{j=1}^{n-p+1} \mathcal{H}_{K_j}, \tag{39}$$

$$\mathcal{H}_S = \oplus_{j=1}^{n-p+1} \mathcal{H}_{S_j}, \tag{40}$$

where the difference between (39) and (40) is only in the definition of the spaces that are generated by the j-th component of the mixture, see (35) and (36).

VI. Numerical experiments

The *enhanced kernel* introduced in (10) is compared to the *more general kernel* (8) on the two benchmarks systems (M1) and (M2) taken from [1] and reported below, such that both an exogenous and an autoregressive parts are present.

$$\begin{aligned} y_t =& 0.5y_{t-1} - 0.05y^2_{t-2} + u^2_{t-1} + 0.8u_{t-2} + e_t \\ e_t \sim& \text{WGN}\left(0, 0.22^2\right) \end{aligned} \tag{M1}$$

$$\begin{aligned} y_t =& 0.8y_{t-1} + u_{t-1} - 0.3u^3_{t-1} + 0.25u_{t-1}u_{t-2} \\ & -0.3u_{t-2} + 0.24u^3_{t-2} - 0.2u_{t-2}u_{t-3} - 0.4u_{t-3} + e_t \\ e_t \sim& \text{WGN}\left(0, 0.14^2\right) \end{aligned} \tag{M2}$$

For each system, we perform $M = 100$ Monte Carlo runs. The initial condition is null for each system. The used input is $u \sim \text{WGN}\left(0, 1^2\right)$, where WGN stands for White Gaussian Noise. We compared the identification performances of the two kernels with $N = 50, 200, 400.$ training data. For practical reason, we set the number of mixtures in (8)-(10) to $n = 20$. The interaction orders p, p^y, p^u assume values in the set $\{1, 2, 3, 4, 5\}$. The best hyperparameters vector θ and p are then determined from data using marginal likelihood optimization. All hyperparameters were initialized with the value of 1, and a constrained optimization interior-point method was used. The constraints guarantee the positivity of hyperparameters. We tested the performance of the methods on a separate test dataset $\{u^{test}_t, y^{test}_t\}_{t=1}^{10000}$ generated in the same way as the training one. The performance of the j-th simulation, $j = 1, \ldots, M$, is measured via the RMSE (Root Mean Square Error):

$$\text{RMSE}_j = \sqrt{\frac{\sum_{t=1}^{10000}\left(\hat{y}^{test}_t - y^{test}_t\right)^2}{10000}}, \tag{41a}$$

$$\hat{y}^{test}_t = \hat{f}_j\left(y^{t,test}, u^{t,test}\right), \tag{41b}$$

where $(y^{t,test}, u^{t,test})$ is the test set up to time $t-1$, and $\hat{f}_j$ is the estimate obtained in the j-th run. We choose to use RMSE in order to compare with the results in [1].

The results in Fig. 1 - 3 compare the boxplots of RMSE_j for the kernels in (8)-(10), respectively, with different numbers of training data. The horizontal dashed line represents the standard deviation of the noise e_t, i.e. the best possible expected prediction error for the specific system considered.

As it is possible to appreciate, the proposed kernel is able to better exploit the structure of the specific class of models (6). The results improved for both types of

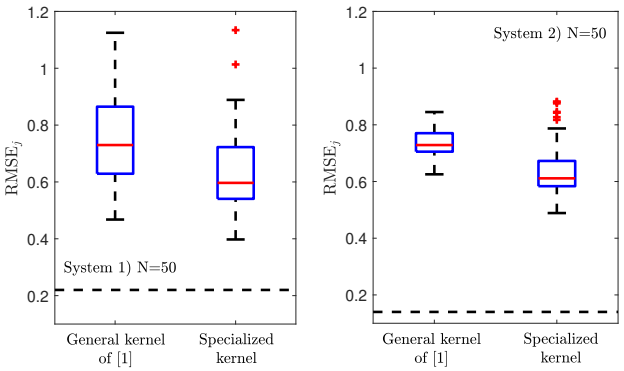

Fig. 1: Identification results with $N = 50$ data for M1 (left) and M2 (right).

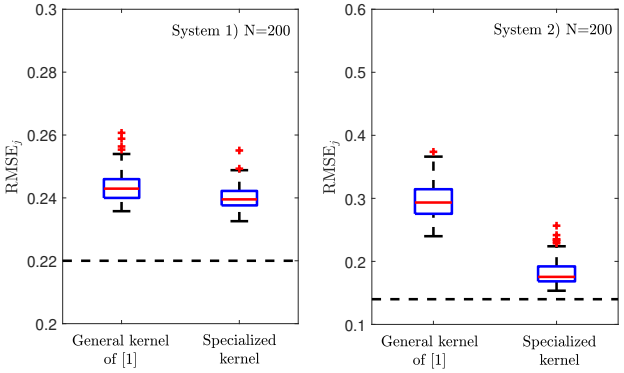

Fig. 2: Identification results with $N = 200$ data for M1 (left) and M2 (right).

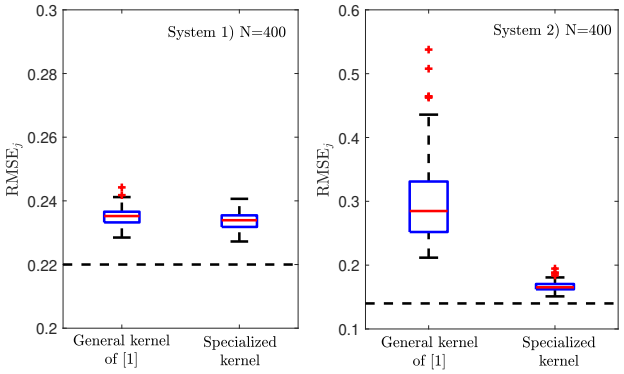

Fig. 3: Identification results with $N = 400$ data for M1 (left) and M2 (right).

benchmark models M1 and M2, for a small amount of data ($N = 50$, Fig. 1) as well as when more data are available ($N = 200, N = 400$, Fig. 2-3). In particular, focusing on Fig. 2-3, for the system 2), the proposed approach is able to better employ the additional data: in fact, for the approach of [1], the availability of $N = 400$ data instead of $N = 200$ observations does not change considerably the identification performance; instead, the proposed kernel attains notable improvements. These performance can be explained by considering that the proposed kernel (10) is taylored to the class of systems in (6), since it is designed to represent functions that are nonlinear in past input and past outputs in a disjoint manner. Furthermore, the enhanced kernel overcomes the limitation of fixed order of interaction p in (8), allowing for different orders of interaction p^y, p^u for the autoregressive and exogenous parts, respectively.

VII. Conclusions and Future Work

In this paper, we presented a modification of the state of the art kernel for nonparametric identification of a specific class of nonlinear systems. Such systems are separately nonlinear in inputs and outputs, such that these two components are additive. In the modified strategy, we allow for different degrees of interaction between previous inputs and outputs. Moreover, we characterize the Reproducing Kernel Hilbert Space of functions generated by the enhanced kernel. Simulation results on two different benchmark systems with different sizes of training data have shown how the proposed approach better exploits the structure of the underlying model. Future work will be devoted to the analysis of computational aspects and optimal hyperparameter tuning.

References

[1] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinearsystem identification," *IEEE Transactions on Automatic Control*, vol. 56, pp. 2825–2840, Dec 2011.

[2] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Survey kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, pp. 657–682, Mar. 2014.

[3] M. Mazzoleni, M. Scandella, S. Formentin, and F. Previdi, "Classification of light charged particles via learning-based system identification," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 6053–6058, IEEE, 2018.

[4] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[5] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[6] B. Mu, T. Chen, and L. Ljung, "On asymptotic properties of hyperparameter estimators for kernel-based regularization methods," *Automatica*, vol. 94, pp. 381 – 395, 2018.

[7] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian processes - revisited," *Automatica*, vol. 48, no. 8, pp. 1525 – 1535, 2012.

[8] A. Carè, B. C. Csàji, M. C. Campi, and E. Weyer, "Finite-sample system identification: An overview and a new correlation method," *IEEE Control Systems Letters*, vol. 2, pp. 61–66, 2018.

[9] L. Ljung, ed., *System Identification (2Nd Ed.): Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.

[10] E. Bai, "Non-parametric nonlinear system identification: An asymptotic minimum mean squared error estimator," *IEEE Transactions on Automatic Control*, vol. 55, pp. 1615–1626, July 2010.

[11] M. Mazzoleni, M. Scandella, S. Formentin, and F. Previdi, "Identification of nonlinear dynamical system with synthetic data: a preliminary investigation," *18th IFAC Symposium on System Identification (SYSID)*, 2018.

[12] M. Mazzoleni, S. Formentin, M. Scandella, and F. Previdi, "Semi-supervised learning of dynamical systems: a preliminary study," *16th European Control Conference (ECC)*, 2018.

[13] S. Formentin, M. Mazzoleni, M. Scandella, and F. Previdi, "Nonlinear system identification via data augmentation," *Systems & Control Letters*, vol. 128, pp. 56 – 63, 2019.

[14] E. D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*, vol. 6. Springer Science & Business Media, 2013.

[15] A. Zuyev and V. Grushkovskaya, "Motion planning for control-affine systems satisfying low-order controllability conditions," *International Journal of Control*, vol. 90, no. 11, pp. 2517–2537, 2017.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006.

[17] G. Wahba, *Spline models for observational data*, vol. 59. Siam, 1990.