Learning nonlinear systems via Volterra series and Hilbert-Schmidt operators

Filippo Cacace, *Member, IEEE*, Vittorio De Iuliis, *Member, IEEE*, Alfredo Germani, *Senior Member, IEEE* and Mario Merone, *Member, IEEE*

***Abstract*— This paper examines the application of regularization techniques and kernel methods in addressing the task of learning nonlinear dynamical systems from input-output data. Our assumption is that the estimator belongs to the space of polynomials composed of Hilbert-Schmidt operators, which ensures the ability to approximate nonlinear dynamics arbitrarily, even within bounded but non-compact data domains. By employing regularization techniques, we propose a finite-dimensional identification procedure that exhibits computational complexity proportional to the square of the size of the training set size. This procedure is applicable to a broad range of systems, including discrete and continuous time nonlinear systems on finite or infinite dimensional state spaces. We delve into the selection of the regularization parameter, taking into account the measurement noise, and also discuss the incorporation of causality constraints. Furthermore, we explore how to derive estimates of the Volterra series of the operator by selecting a parametric inner product between data trajectories.**

***Index Terms*— Reproducing Kernel Hilbert space, System Identification, Learning, Nonlinear systems, Volterra series**

I. Introduction

Black-box identification of unknown systems from observed input-output data is a central problem for systems theory since the beginning and it has recently become an active area of research for machine learning leading to an interesting cross fertilization between classical parameter identification and model selection methods [1], statistical learning techniques [2] and kernel methods [3]: see [4], [5] for comprehensive overviews. The impact of this cross fertilization is testified by the fact that estimation techniques that were typically associated to the machine learning community – such as artificial neural networks, support vector machines, regression trees, etc. – have now found a stable place in control theory [6]–[8], proving their accuracy in data-based control strategies on large-scale complex scenarios [9], [10]. Some of the key mathematical tools used in this new paradigm are reproducing kernel Hilbert spaces (RKHS) [11]–[16], kernel methods and regularization networks [3], [17]–[26], Gaussian process regression [27], [28], support vector machines [29], [30], representer theorems [25], [31], [32] and regularized solutions [33]–[36].

In an abstract setting the output of a system is generated by some nonlinear operator applied to the input, that may include the initial state of the system. The challenge is that the identification of this input-output operator is always an ill-posed problem since the operator cannot be uniquely inferred from a finite set of input and output trajectories. We recall that a problem is ill-posed if one of the three conditions of existence, uniqueness and continuous dependence on data fails to hold for its solution [33], [37]. Even when a solution exists in the solution space, the unavoidable presence of noise may cause existence to fail. This is where regularization comes into play. The essential idea is to find the best approximation of the solution in a space where the approximation exists, is unique and continuous with respect to the data. This introduces a trade-off between *accuracy*, i.e. good accordance of actual and predicted data, and the requirement that solutions are well-behaved. In the regularization approach this trade-off is usually controlled by a scalar *regularization parameter* [35], [38]. Kernel methods are then used to generate the best approximation in the infinite-dimensional operator space by projecting on the subspace of operators that can be represented from the available data. This approach has been widely exploited in the learning techniques mentioned above. The existing techniques have been applied to linear systems [23], [27], [39]–[43], linear discrete-time systems with outliers [44], discrete-time dynamics in Euclidean spaces [45]–[47], and nonlinear functions with discrete samples [22], [48]. It is also worth mentioning that the theoretical features of Volterra series and polynomial kernel regression have already been considered in several important works, for example [45], [47], [49], [50], and a huge literature exists on block-oriented approaches to nonlinear identification exploiting specific structures of nonlinear operators, such as Hammerstein models, Wiener models, and combinations of them (see [51], [52]).

Our work extends the approach presented in [53] for the linear case. The core idea is to extend existing results to the general problem of nonlinear maps over separable Hilbert spaces. Our task is to compute a finite-dimensional

Corresponding author: Mario Merone (e-mail: m.merone@unicampus.it).

F. Cacace, is with Research Unit of Computer Systems and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, Italy (e-mail: f.cacace@unicampus.it).

V. De Iuliis is with the Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica Università degli Studi dell'Aquila (e-mail: vittorio.deiuliis@univaq.it).

A. Germani is with Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica Università degli Studi dell'Aquila (e-mail: alfredo.germani@univaq.it).

M. Merone, is with Research Unit of Computer Systems and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, Italy (e-mail:m.merone@unicampus.it).

approximation of the unknown input-output map $\mathcal{H}_x \to \mathcal{H}_y$ starting from a *training set* of input-output samples $\{x_i, y_i\}$, $i = 1, \dots, N$, where $\mathcal{H}_x$ and $\mathcal{H}_y$ are generic Hilbert spaces. This training set may either be generated by user-defined probe inputs, or be collected from the autonomous dynamics of the system under consideration. This abstract framework encompasses many cases, for example x_i and y_i may be finite-dimensional vectors of input and output values but also vector-valued functions of time corresponding to the input and the output of a dynamical system, or functions on a spatial domain when the operator of interest is the solution of a PDE. Thus, the results obtained in this setting can be applied to discrete as well as continuous time systems evolving in $\mathbb{R}^n$, distributed parameter systems, nonlinear systems with delay, finite-dimensional sequences, etc. The aims and contributions of this work can be summarized as follows.

- We establish an abstract framework to generalize regularization and kernel-based techniques to a broader class of systems. We study which hypotheses are needed to ensure that essential properties of the solutions provided by these methods are not lost in the generalization.
- We show that with a suitable choice of the solution space regularization techniques can approximate arbitrarily well nonlinear systems with reasonable physical constraints.
- We show how the available prior knowledge can be plugged into the framework to restrict the solution space and improve the accuracy of the learning process. However, we show that solutions can be found in absence of any prior knowledge in a completely black-box setting.
- We show that solution spaces based on polynomial or Volterra series have computational complexity comparable to linear estimates and that solutions can incorporate causality requirements.

As for the first point we derive a closed form optimal approximation of the Volterra series (or of its truncated version up to a specified order) by extending the well known technique of minimizing a regularized quadratic performance index computed on the training set, and we show that this solution admits a finite-dimensional computation.

The second contribution is to identify the most general solution space that yields solution operators with a physical characterization. The unknown nonlinear input-output map is represented as a Volterra series of Hilbert-Schmidt operators (H-S operators in the following). This choice is motivated by the considerations reported in Section II-C.

As for the third point, we show that the general estimator can be specialized by defining the inner product that generates the estimator based on a specific kernel. In other words one can introduce the prior information on the unknown operator to be estimated by choosing the most appropriate kernel to define the inner product on the linear space of the input trajectories.

Finally, we prove that causal operators can be introduced in this framework without additional constraints on the solution space. Our results prove the existence of a solution without any *a priori* hypothesis on the structure of the operator.

Although most of the above points have been considered and studied in the previous literature mentioned above, we argue that providing a general and flexible framework has important benefits and implications. In the first place, it avoids the need of "reinventing the wheel" in each specific situation whereas at the same time it indicates where specific choices, for example the kernel design or the choice of the meta-parameters, come into play to tailor the solution to specific application needs.

The paper is organized as follows. In Section II we introduce the basic definitions and assumptions used throughout the paper. Section III contains the main results on the computation of the optimal polynomial and Volterra series approximation. Section V links our framework to the well established theory of reproducing kernel Hilbert spaces (RKHS), by showing that our polynomial operators with H-S terms constitute in fact a RKHS. Section VI discusses an example to illustrate the results and Section VII concludes the paper.

Notation: $\mathcal{L}_2(D; C)$ denotes the Hilbert space of square integrable functions $f : D \to C$. The scalar product between elements $x_1, x_2 \in \mathcal{L}_2(D; C)$ is denoted $[x_1, x_2]_{\mathcal{L}_2}$ and $\|x\|_{\mathcal{L}_2} = \sqrt{[x, x]_{\mathcal{L}_2}}$ is the norm of an element $x \in \mathcal{L}_2(D; C)$. Throughout the paper we denote $\mathrm{col}_{i=1}^{n}(x_i)$ the vector with n entries x_i, and $\mathrm{row}_{i=1}^{n}(x_i)$ the row vector with n entries x_i, that can be scalars, vectors, functions or matrices. The lower and upper bound of the index is sometimes omitted for brevity. The linear combination $\sum_{i=1}^{k} a_i x_i$ of a finite number k of elements $x_i \in \mathcal{L}_2(D; C)$ with coefficients $a_i \in \mathbb{R}$ is concisely denoted as $a^\top \mathrm{col}_i(x_i)$, where $a = \mathrm{col}_i(a_i)$. The same convention is used when $M \in \mathbb{R}^{n \times k}$ is a matrix with scalar entries m_{ij}, $M\mathrm{col}_{i=1}^{k}(x_i) := \mathrm{col}_{i=1}^{n}(\sum_{j=1}^{k} m_{ij} x_j)$.

II. Preliminary definitions and results

A. Problem statement and assumptions

The input-output map of an unknown nonlinear system $\mathcal{S}$ is modeled as a nonlinear operator from the input linear space $\mathcal{H}_x = \mathcal{L}_2(D; H_1)$ to the output linear space $\mathcal{H}_y = \mathcal{L}_2(D; H_2)$. The input and the output of the system are therefore square integrable functions on D, which is a bounded domain (for example $D = [0, T]$)[1]. When D is discrete, $\mathcal{H}_x = \ell_2(D; H_1)$ and $\mathcal{H}_y = \ell_2(D; H_2)$. H_1 and H_2 are real separable Hilbert spaces. The image spaces H_1 and H_2 are purposefully kept generic to include a variety of input and output data, such as those generated by finite-dimensional systems, distributed parameter systems, delay systems, etc. We will refer to $x_i \in \mathcal{H}_x$ and $y_i \in \mathcal{H}_y$ as *input and output trajectories*, bearing in mind the reference case where $D = [0, T]$ and $H_1 = \mathbb{R}^{n_x}$, $H_2 = \mathbb{R}^{n_y}$ that models linear or nonlinear systems with finite-dimensional input and output.

We consider the following problem. Given a training set of pairs (x_i, y_i), where $x_i \in \mathcal{H}_x$ is the input corresponding to $y_i \in \mathcal{H}_y$, we want to find a finite-dimensional nonlinear approximation in the least-squares sense of the unknown input-output map. To this end we shall make use of polynomial functions of given order of the input and of the corresponding Volterra series as the solution space. For the reasons explained in Section II-C, the existence of a polynomial approximation of the input-output operator rests on the following assumption.

[1]Throughout the paper we denote t the elements of D. However, D may be any domain for example a spatial domain.

Assumption 1: The operator $F : \mathcal{H}_x \to \mathcal{H}_y$ is uniformly S-continuous on the sets of bounded energy signals of interest. S-continuity is defined in Definition 4. We also require that the y_i are defined on all D, since their values are needed to compute the output of the estimated operator.

Assumption 2: The output signals y_i of the training set are defined in all D.

B. Volterra series and Hilbert-Schmidt operators

We briefly introduce a few formal notations to represent polynomials and Volterra series in a compact way. The following definition is for the case $H_1 = \mathbb{R}^n$ but it can be easily extended to separable Hilbert spaces.

Definition 1: [Kronecker product on Hilbert spaces]. Given $x, z \in \mathcal{H}_n = \mathcal{L}_2(D; \mathbb{R}^n)$ their Kronecker product $x \boxtimes z$ is the element of $\mathcal{L}_2(D^2; \mathbb{R}^{n^2})$ such that, $\forall \xi, \zeta \in D$,

$$(x \boxtimes z)(\xi, \zeta) := x(\xi) \otimes z(\zeta), \tag{1}$$

where $\otimes$ is the ordinary Kronecker product.

Remark 1: The value of $x \boxtimes z$ is the product of the values of x and z at different points of D. Consequently, the temporal or spatial correlation between functions, when D is respectively a temporal or spatial domain, can be expressed through operators on $x \boxtimes z$.

Let $\{\phi_k\}$ be an orthonormal basis of $\mathcal{H}_n$. Then $x \boxtimes z$ belongs to $\mathcal{H}_n^{\boxed{2}}$, a subspace of $\mathcal{L}_2(D^2; \mathbb{R}^{n^2})$ defined as

$$\mathcal{H}_n^{\boxed{2}} = \{w \in \mathcal{L}_2(D^2; \mathbb{R}^{n^2}) : w = \sum_{k,j} a_{k,j} \left(\phi_k \boxtimes \phi_j\right), \ \sum_{k,j} a_{k,j}^2 < \infty\}. \tag{2}$$

In essence, $\mathcal{H}_n^{\boxed{2}}$ is the span of the Kronecker product of versors in the basis of $\mathcal{H}_n$. When $x = z$, $x^{\boxed{2}} = x \boxtimes x$ defines the Kronecker power of elements of H_n. This can be generalized to higher-orders by defining $x^{\boxed{i}} = x \boxtimes x^{\boxed{i-1}} \in \mathcal{H}_n^{\boxed{i}}$, $i \in \mathbb{N}$, where $\mathcal{H}_n^{\boxed{i}} \subset \mathcal{L}_2(D^i; \mathbb{R}^{n^i})$ is the span of $\phi_{k_1} \boxtimes \ldots \boxtimes \phi_{k_i}$ with square summable coefficients $a_{k_1,\ldots,k_i}$. Furthermore, when the image of elements in $\mathcal{H}_n$ is not $\mathbb{R}^n$ but a space of functions with values in $\mathbb{R}^n$, Definition 1 is extended by replacing in (1) $x(\xi) \otimes z(\zeta)$ with $x(\xi) \boxtimes z(\zeta)$, where $\boxtimes$ is the Kronecker product in the image space. We now show how the inner product in $\mathcal{H}_n^{\boxed{m}}$ is related to the inner product in $\mathcal{H}_n$.

Lemma 1: If $x, y, z, w \in \mathcal{H}_n$, $\mathcal{H}_n$ separable, then

$$[x \boxtimes y, \ w \boxtimes z]_{\mathcal{H}_n^{\boxed{2}}} = [x, w]_{\mathcal{H}_n} [y, z]_{\mathcal{H}_n} \tag{3}$$

$$[x^{\boxed{m}}, y^{\boxed{m}}]_{\mathcal{H}_n^{\boxed{m}}} = [x, y]_{\mathcal{H}_n}^m \tag{4}$$

$$\|x^{\boxed{m}}\|_{\mathcal{H}_n^{\boxed{m}}} = \|x\|_{\mathcal{H}_n}^m. \tag{5}$$

Proof: It is sufficient to prove the first relationship. Since $\mathcal{H}_n$ is separable, given an orthonormal basis $\{\phi_k\}$ of $\mathcal{H}_n$,

$$x \boxtimes y = \sum_{k_1} \sum_{k_2} [x, \phi_{k_1}][y, \phi_{k_2}] \left(\phi_{k_1} \boxtimes \phi_{k_2}\right) \tag{6}$$

$$\begin{aligned}[x \boxtimes y, w \boxtimes z]_{\mathcal{H}_n^{\boxed{2}}} &= \sum_{k_1} \sum_{k_2} [x, \phi_{k_1}][y, \phi_{k_2}][w, \phi_{k_1}][z, \phi_{k_2}] \\ &= [x, w]_{\mathcal{H}_n} [y, z]_{\mathcal{H}_n}.\end{aligned} \tag{7}$$

■

In this paper we consider polynomials whose monomial terms are H-S operators. We first recall the definition of H-S operators.

Definition 2: [H-S operators]. A Hilbert-Schmidt operator on separable Hilbert spaces $L : \mathcal{H}_x \to \mathcal{H}_y$ is a linear operator such that

$$Lx = \sum_{k=1}^{\infty} [x, \phi_k]_{\mathcal{H}_x} L\phi_k \tag{8}$$

$$\|L\|_{\text{H.S.}}^2 = \sum_{k=1}^{\infty} \|L\phi_k\|_{\mathcal{H}_y}^2 < \infty, \tag{9}$$

where $\{\phi_j\}$ is any orthonormal basis of $\mathcal{H}_x$ (*i.e.* the norm does not depend on $\{\phi_j\}$). □

The definition implies that

$$\|Lx\|_{\mathcal{H}_y} \leqslant \|L\|_{\text{H.S.}} \|x\|_{\mathcal{H}_x}. \tag{10}$$

H-S operators from $\mathcal{H}_x$ to $\mathcal{H}_y$ are bounded finite-energy input-output operators, as it is apparent from (9). We recall that on infinite-dimensional Hilbert spaces the identity operator *is not* a H-S operator because $\|I\|_{\text{H.S.}}^2 = \sum_{k=1}^{\infty} \|\phi_k\|_{\mathcal{H}_y}^2 = \infty$, thus the input-output behavior is not represented by H-S operators when there is a direct term from the input to the output, *e.g.* $y_i = x_i$ cannot be represented through H-S operators when $\mathcal{H}_y$ is infinite-dimensional. Finally, the linear space of H-S operators $L : \mathcal{H}_x \to \mathcal{H}_y$ endowed with inner product

$$[L_1, \, L_2]_{\text{H.S.}} = \sum_{k=1}^{\infty} [L_1 \phi_k, \, L_2 \phi_k]_{\mathcal{H}_y} \tag{11}$$

is a Hilbert space $\mathcal{L}_{\text{H.S.}}(\mathcal{H}_x; \mathcal{H}_y)$.

Polynomial operators $\mathcal{H}_x \to \mathcal{H}_y$ are obtained by summing monomials that are H-S operators, $M_m : \mathcal{H}_x^{\boxed{m}} \to \mathcal{H}_y$, $m \geqslant 1$. A generic monomial term is written $M_m(x_1 \boxtimes x_2 \boxtimes \ldots \boxtimes x_m)$, where $x_i \in \mathcal{H}_x$, and its value in $\mathcal{H}_y$. M_m is multilinear with respect to its arguments x_i. Thus, a monomial $M_m(x^{\boxed{m}})$ is a linear operator with respect to $x^{\boxed{m}}$ but not to x.

Since M_m is a H-S operator, it is defined by specifying its value on some basis of $\mathcal{H}_x^{\boxed{m}}$, as prescribed by (8). Let $\{\phi_k\}$ be an orthonormal basis of $\mathcal{H}_x$ and $K_m = \{k_1, \ldots, k_m\}$ a multi-index with m elements. Then $\xi \in \mathcal{H}_x^{\boxed{m}}$ can be represented as

$$\xi = \sum_{K_m} a_{K_m} (\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}),$$

where $a_{K_m} = [\xi, \phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}]_{\mathcal{H}_x^{\boxed{m}}}$, and

$$M_m(\xi) = \sum_{K_m} a_{K_m} M_m \left(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}\right), \tag{12}$$

where $\sum_{K_m} \|M_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m})\|_{\mathcal{H}_y}^2 < \infty$ as required by (9).

In particular, it descends from (3) that when $\xi = x^{\boxed{m}}$ then $a_{K_m} = \prod_{j \in K_m} [x, \phi_j]$. Consequently,

$$M_m\left(x^{\boxed{m}}\right) = \sum_{K_m} \left(\prod_{j \in K_m} [x, \phi_j] \right) M_m \left(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}\right). \tag{13}$$

Definition 3: [H-S Polynomial operators]. Given $\mathcal{L}_2$ spaces $\mathcal{H}_x$, $\mathcal{H}_y$ and a finite sequence $\{M_m\}$, $m = 1, \ldots, \nu$ of H-S

operators $M_m : \mathcal{H}_x^{\boxed{m}} \to \mathcal{H}_y$, the operator $P_\nu : \mathcal{H}_x \to \mathcal{H}_y$ defined as

$$P_\nu(x) := \sum_{m=1}^{\nu} M_m(x^{\boxed{m}}) \tag{14}$$

is called a ν-degree polynomial operator. □

Remark 2: Since $P_\nu : \mathcal{H}_x \to \mathcal{H}_y$, by definition $y = P_\nu(x)$ is square integrable over D. Notice that however P_ν is neither a linear operator nor, consequently, a H-S operator.

Definition 3 does not include a constant term $M_0 \in \mathcal{H}_y$, that can be added as a separate term. For example, in the case of $\mathcal{L}_2(D;\mathbb{R}^n)$, a complete second-degree polynomial is

$$y = M_0 + M_1 x + M_2(x^{\boxed{2}}) \tag{15}$$

$$y(t) = y_0(t) + \int_D \kappa^{M_1}(t,\tau)x(\tau)\,\mathrm{d}\tau + \int_{D\times D} \kappa^{M_2}(t,\sigma,\tau)\,(x(\sigma)\otimes x(\tau))\,\mathrm{d}\tau\,\mathrm{d}\sigma, \tag{16}$$

and y is specified though the choice of $M_0 = y_0 \in \mathcal{H}_y$ and of the operator kernels $\kappa^{M_1} : D^2 \to \mathbb{R}^{n_y \times n_x}$, $\kappa^{M_2} : D^3 \to \mathbb{R}^{n_y \times n_x^2}$. Since $M_1 \in \mathcal{L}_{\text{H.S.}}(\mathcal{H}_x;\mathcal{H}_y)$ and $M_2 \in \mathcal{L}_{\text{H.S.}}(\mathcal{H}_x^{\boxed{2}};\mathcal{H}_y)$ are H-S operators, it holds that

$$\int_{D\times D} \|\kappa^{M_1}(t,\tau)\|^2\,\mathrm{d}\tau\,\mathrm{d}t < \infty, \tag{17}$$

$$\int_{D^3} \|\kappa^{M2}(t,\sigma,\tau)\|^2\mathrm{d}\tau\,\mathrm{d}\sigma\,\mathrm{d}t < \infty. \tag{18}$$

The Volterra series can be obtained as the limit $\lim_{\nu\to\infty} P_\nu(x)$ of a sequence of polynomial operators. It is worth remarking that this representation of the input-output map is not necessarily a causal one. For example, to introduce causality in (15) one needs to add the constraints $\kappa^L(t,\tau) = 0$ whenever $\tau > t$ and $\kappa^Q(t,\sigma,\tau) = 0$ whenever $\sigma \geqslant \tau$ or $\tau \geqslant t$.

The following theorem is proved in Appendix A.

Theorem 1: For any ν, the linear space of polynomial operators $P_\nu : \mathcal{H}_x \to \mathcal{H}_y$ on separable Hilbert spaces defined in (14) is an Hilbert space $\mathcal{L}_\nu^P$ with inner product

$$[P_\nu^1,\, P_\nu^2]_{\mathcal{L}_\nu^P} = \sum_{m=1}^{\nu} [M_m^1, M_m^2]_{\mathcal{L}_{\text{H.S.}}}. \tag{19}$$

In particular, with $\|M_i\|_{\text{H.S.}}^2 = \sum_{K_i} \|M_i(\phi_{k_1}\boxtimes\cdots\boxtimes\phi_{k_m})\|_{\mathcal{H}_x^{\boxed{i}}}^2$

$$\|P_\nu\|_{\mathcal{L}_\nu^P}^2 = \sum_{i=1}^{\nu} \|M_i\|_{\text{H.S.}}^2. \tag{20}$$

C. Polynomial approximations of nonlinear operators

The choice of H-S operators to approximate nonlinear dynamical systems has two main motivations. The first one is that H-S operators are bounded as one would expect from a physical system. The second one is that a polynomial based on H-S operators can approximate with arbitrary precision a large class of input-output maps. In fact, the approximation of nonlinear operators by means of polynomials requires some version of the Weierstrass approximation theorem suited to Hilbert or Banach spaces. Such extensions have been widely explored in the literature in the past decades [54]–[57] but they crucially depend on the compactness of the domain. This hypothesis is restrictive in the context of system identification, e.g. the ball of finite energy signals is bounded but not compact. A possibility to overcome this problem is to restrict ourselves to maps that are uniformly continuous with respect to the S-topology [58], since it was proved in [59] (Theorem 3) that given a map $F : \mathcal{H}_x \to \mathcal{H}_y$ uniformly continuous with respect to the S-topology on a bounded set $\Omega \subset \mathcal{H}_x$, $\forall \epsilon > 0$ there exists a continuous polynomial $P : \mathcal{H}_x \to \mathcal{H}_y$ such that

$$\sup_{x\in\Omega} \|F(x) - P(x)\| < \epsilon. \tag{21}$$

In other words, on bounded sets the continuous polynomials are dense with respect to the family of uniformly S-continuous functions.

Definition 4: [58], [59] A map $F : \mathcal{H}_x \to \mathcal{H}_y$ is said to be *uniformly continuous with respect to the S-topology* if, for any $\epsilon > 0$ there exists a self-adjoint non-negative definite trace-class operator $S_\epsilon : \mathcal{H}_x \to \mathcal{H}_y$ such that $[S_\epsilon(x_1 - x_2),\, x_1 - x_2]_{\mathcal{H}_x} < 1$ implies $\|F(x_1) - F(x_2)\|_{\mathcal{H}_y} < \epsilon$ for all $x_1,\, x_2 \in \mathcal{H}_x$.

The S-topology is weaker than the norm topology and maps that are uniformly continuous with respect to the S-topology are also compact (Theorem 1 in [59]). In addition, the functions in this class can be represented as continuous nonlinear functions on H-S operators [60] and in particular linear operators that are uniformly continuous with respect to the S-topology are H-S operators. This is not only a nice mathematical characterization but it also implies the very "physical" property of being an input-output map with a smoothing action. Summarizing, the choice of H-S operators as monomial terms of $P(x)$ is justified both on physical and mathematical terms.

III. Learning the Optimal Volterra Series

A. Learning the optimal linear approximation

We first consider the problem of estimating the best linear approximation (*i.e.* $\nu = 1$ in (14)), to introduce regularized estimates that will be extended to the polynomial case. The material in this section summarizes the results in [53] for the linear case. Given a training set $\{(x_i, y_i)\}$, the simplest way to compute a linear approximation $\hat{L}x = \hat{M}_1 x + \hat{M}_0$ of the input-output operator is the well known Least Squares Estimate (LSE) $\hat{L} = \arg\min_L J(L)$, $J(L) = \sum_i \|y_i - Lx_i\|_{\mathcal{H}_y}^2$. This approach has two main drawbacks. The first one is that when both $\{x_i\}$ and $\{y_i\}$ have N independent elements the LSE yields a $\hat{L}$ such that $J(\hat{L}) = 0$ because $\forall i\ y_i = \hat{L}x_i$, and consequently the approach is prone to over-fitting. On the other side, when there is no linear operator such that $\forall i\ y_i = \hat{L}x_i$, LSE yields an infinite norm estimate. More precisely, any sequence $\{\hat{L}_n\}$ of operators with finite-rank n that minimizes $J(L)$ is such that $\|\hat{L}_n\|_{\text{H.S.}} \to \infty$ [33], [53]. The problem is relevant in practice when the output y_i is affected by noise n_i, since the operator $\hat{L}$ tends to track the output noise in order to reduce the residual $y_i - \hat{L}x_i$.

One method to cope with the above problems is to introduce a penalty for the norm of the estimate in the cost function. Intuitively, this regularization alleviates over-fitting because "too precise" solutions are discarded in favor of more compact

ones, and at the same time it yields finite norm estimates when the LSE estimate does not exist. In the case of the best affine estimate the regularization approach amounts to defining a cost functional $J_N : \mathbb{R}^+ \times \mathcal{L}_{\text{H.S.}} \times \mathcal{H}_y \to \mathbb{R}^+$ and, with $L\xi = M_1\xi + M_0$,

$$J_N(\lambda, L) = \sum_{i=1}^{N} \|y_i - M_1 x_i - M_0\|^2_{\mathcal{H}_y} + \lambda \|M_1\|^2_{\text{H.S.}}, \quad (22)$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter [13], [23], [61] (notice that the regularization parameter does not affect M_0). The solution is found by using variational calculus to find $\arg\min_L J_N(\lambda, L)$. Let $\epsilon > 0$ and $\Delta_L x = \Delta_{M_1} x + \Delta_{M_0}$, $M_0 \in \mathcal{H}_y$, $\Delta_{M_1} \in \mathcal{L}_{\text{H.S.}}$. Straightforward computations yield that in the minimum $\hat{L}$

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\epsilon} J_N(\lambda, \hat{L} + \epsilon\Delta_L)\Big|_{\epsilon=0} \\ &= \sum_{i=1}^{N} -2[y_i - \hat{L}x_i, \Delta_L x_i]_{\mathcal{H}_y} + 2\lambda[\hat{M}_1, \Delta_{M_1}]_{\text{H.S.}} = 0. \end{aligned}$$

The minimum is obtained by imposing that the derivative is null $\forall \Delta_{M_1}, \Delta_{M_0}$. In particular, choosing $\Delta_{M_1} = 0$ one obtains

$$\sum_{i=1}^{N} -2[y_i - \hat{M}_1 x_i - \hat{M}_0, \Delta_{M_0} x_i]_{\mathcal{H}_y} = 0$$

that can be satisfied $\forall \Delta_{M_0}$ only when

$$\hat{M}_0 = \sum_{i=1}^{N} y_i - \hat{M}_1 x_i = \bar{y} - \hat{M}_1 \bar{x}, \quad (23)$$

that yields $\hat{M}_0$ as a function of $\hat{M}_1$. We notice that the "affine" term $\hat{M}_0$ disappears when $\bar{y} = \bar{x} = 0$. For this reason, we can consider the *centered* training set

$$(\mathring{X}, \mathring{Y}) = (\text{col}_i(x_i - \bar{x}), \text{col}_i(y_i - \bar{y})), \quad (24)$$

that has $\bar{y} = \bar{x} = 0$, and compute the best linear and polynomial estimate with the tacit assumption that affine estimates are obtained by adding $\hat{M}_0$ given by (23). On a centered training set and for a given choice of the weight λ, the best estimate of L is now defined as

$$\hat{L} = \arg \min_{M_1 \in \mathcal{L}_{\text{H.S.}}} \sum_{i=1}^{N} \|\mathring{y}_i - M_1 \mathring{x}_i\|^2_{\mathcal{H}_y} + \lambda \|M_1\|^2_{\text{H.S.}}, \quad (25)$$

and it can be computed as follows [53].

Theorem 2: If Assumption 2 holds, then, given a centered training set $\{x_i, y_i\}$, $\mathbf{x} = \{x_i\}$ $i = 1, \dots, N$, and the matrix $A(\mathbf{x}) \in \mathbb{R}^{N \times N}$ with entries

$$A(\mathbf{x})_{(i,j)} = [x_i, x_j]_{\mathcal{H}_x}, \quad (26)$$

1) For any $\lambda \in \mathbb{R}$ the optimal linear estimator $\hat{M}_1$ defined by (25) exists and, denoting $Y = \text{col}_i(y_i) \in \mathcal{H}_y^N$, $\forall k$,

$$\hat{M}_1 \phi_k = \alpha_k^\top(\lambda, \mathbf{x}) Y = \sum_{i=1}^{N} \alpha_{ki}(\lambda, \mathbf{x}) y_i \quad (27)$$

$$\alpha_k^\top(\lambda, \mathbf{x}) = r_k^\top(\mathbf{x}) \left(\lambda I_N + A(\mathbf{x})\right)^{-1} \in \mathbb{R}^{1 \times N} \quad (28)$$

$$r_k^\top(\mathbf{x}) = \text{row}_{i=1}^{N}[x_i, \phi_k]_{\mathcal{H}_x} \quad (29)$$

where $\{\phi_k\}$ is an orthonormal basis of $\mathcal{H}_x$. Moreover, each $\hat{M}_1 \phi_k \in \mathcal{H}_y$ is defined $\forall t \in D$.

2) The optimal linear estimate of $\hat{M}_1 \xi$, $\forall \xi \in \mathcal{H}_x$, is

$$\hat{M}_1 \xi = \text{row}_i([\xi, x_i]_{\mathcal{H}_x}) \left(\lambda I_N + A(\mathbf{x})\right)^{-1} Y. \quad (30)$$

3) $\|\text{col}_i(y_i - \hat{M}_1 x_i)\|_{\mathcal{H}_y^N}$ is a non-decreasing function of λ and $\|\hat{M}_1\|_{\text{H.S.}}$ is a non-increasing function of λ.
4) If $\text{col}_i(y_i - \tilde{L}x_i) = 0$ for some $\tilde{L} \in \mathcal{L}_{\text{H.S.}}$, then $\lim_{\lambda \to 0^+} \hat{M}_1 = \arg\min_{L \in \mathcal{L}_{\text{H.S.}}} \|L\|_{\text{H.S.}} : \text{col}_i(y_i - Lx_i) = 0$. □

When the dataset is not centered, $\hat{M}_1$ is given by (27) where x_i and y_i are replaced by $x_i - \bar{x}$ and $y_i - \bar{y}$, and $\hat{M}_0$ is given by (23). For completeness a proof of the first point is reported in Appendix B. The second point follows from straightforward computations. For points 3)–4) see Thm. 3 and 5 in [53] or Thms. 3.12, 3.15 in [33].

The second point of Theorem 2 provides a finite dimensional closed-form expression of $\hat{M}_1 \xi$ for any $\xi \in \mathcal{H}_x$ as a linear combination of the y_i in the training set. Since $\left(\lambda I_N + A(\mathbf{x})\right)^{-1} \text{col}_i(y_i)$, which is a vector of functions with entries in $\mathcal{H}_y$, depends only on the training set, it can be computed once and for all. The computation of $\hat{M}_1 \xi$ for a generic ξ requires only to compute the vector $\text{row}_i([\xi, x_i]_{\mathcal{H}_x}) \in \mathbb{R}^N$. From the point of view of the computational complexity the most computationally intensive operation is the scalar product, that typically involves the numerical computation of an integral. In this sense, the computational complexity of (30) is quadratic in the size N of the training set, since $A(\mathbf{x})$ contains $N(N+1)/2$ distinct entries of the kind $[x_i, x_j]_{\mathcal{H}_x}$.

Example 1: In the case of continuous-time systems $\mathcal{H}_x = \mathcal{L}_2(D; \mathbb{R}^{n_x})$, $\mathcal{H}_y = \mathcal{L}_2(D; \mathbb{R}^{n_y})$, let $M_\lambda = \lambda I_n + A(\mathbf{x})$. Straightforward manipulations yield

$$\begin{aligned} \hat{y}(t) =& (\hat{M}_1 x)(t) = \\ & \int_D \left((x^\top(\tau) \text{row}_i(x_i(\tau)) M_\lambda^{-1}) \otimes I_{n_y} \right) \text{col}_i(y_i)(t)\, \mathrm{d}\tau \\ =& \int_D \text{row}_i(y_i(t)) {M_\lambda^{-1}}^\top \text{col}_i(x_i^\top(\tau)) x(\tau)\, \mathrm{d}\tau, \end{aligned} \quad (31)$$

where we recognize the underlying operator kernel of $\hat{M}_1$, $\kappa^{\hat{M}_1}(t, \tau) \in \mathbb{R}^{n_y \times n_x}$,

$$\kappa^{\hat{M}_1}(t, \tau) = \text{row}_i(y_i(t)) {M_\lambda^{-1}}^\top \text{col}_i(x_i^\top(\tau)). \quad (32)$$

B. Learning the optimal polynomial approximation

This section considers the problem of computing the ν-th degree polynomial that best approximates $F : \mathcal{H}_x \to \mathcal{H}_y$ from a training set of input and output signals $\{x_i, y_i\}$, $i = 1, \dots, N$. The setting and the notation are the same as in Section III-A. When $M_0 = 0$ our hypotheses model is that there exists a polynomial operator $P_\nu^\star$ such that

$$y_i = P_\nu^\star(x_i), \quad (33)$$

and we aim at computing $\hat{M}_m$, $m = 1, \dots, \nu$ so that

$$\hat{y}_i = \hat{P}_\nu(x_i) = \sum_{m=1}^{\nu} \hat{M}_m(x_i^{\boxed{m}}), \quad (34)$$

is the "best" approximation of y_i, where, as before, $x_i \in \mathcal{H}_x$, $y_i \in \mathcal{H}_y$, $\hat{M}_m \in \mathcal{L}_{\text{H.S.}}(\mathcal{H}_x^{\boxed{m}}; \mathcal{H}_y)$. Given an orthonormal basis $\{\phi_k\}$ of $\mathcal{H}_x$, $\hat{M}_m$ can be represented as in (12),

$$\hat{M}_m(x^{\boxed{j}}) = \sum_{K_m} a_{K_m}(x) \hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}), \tag{35}$$

where $a_{K_m}(x) = [x, \phi_{k_1}]_{\mathcal{H}_x} \cdot \ldots \cdot [x, \phi_{k_m}]_{\mathcal{H}_x} \in \mathbb{R}$ and

$$\sum_{K_m} (a_{K_m}(x_1) \cdot a_{K_m}(x_2)) = [x_1, x_2]^j_{\mathcal{H}_x}. \tag{36}$$

The optimal estimate $\hat{P}$ is obtained by minimizing the functional $J_N^\nu : \mathbb{R}^+ \times \mathcal{L}_\nu^P \to \mathbb{R}^+$,

$$J_N^\nu(\lambda, P_\nu) = \sum_{i=1}^N \|y_i - P_\nu(x)\|^2_{\mathcal{H}_y} + \lambda \|P_\nu\|^2_{\mathcal{L}_\nu^P}. \tag{37}$$

Theorem 3: If Assumption 2 holds, then, given a training set $\{x_i, y_i\}$, $i = 1, \ldots, N$ and the matrix $A_\nu(\mathbf{x}) \in \mathbb{R}^{N \times N}$ with entries

$$(A_\nu(\mathbf{x}))_{(i,j)} = \sum_{m=1}^\nu [x_i, x_j]^m_{\mathcal{H}_x}, \tag{38}$$

for any $\lambda \in \mathbb{R}$ the optimal polynomial estimator with respect to (37) exists and its monomials $\hat{M}_m$, are given by

$$\hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}) = \text{row}_i(a_{K_m}(x_i)) \left(\lambda I_N + A_\nu(\mathbf{x})\right)^{-1} Y \tag{39}$$

where $\{\phi_k\}$ is an orthonormal basis of $\mathcal{H}_x$ and $Y = \text{col}_i(y_i)$. Moreover, $\hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m})$ are defined $\forall t \in D$.

Proof: The minimum $\hat{P}_\nu$ of J_N^ν is the solution of the system

$$\forall m = 1, \ldots, \nu, \ \forall \Delta_{M_m} \in \mathcal{L}_{\text{H.S.}}(\mathcal{H}_x^{\boxed{m}}; \mathcal{H}_y) : \left. \frac{\mathrm{d} J_N^\nu(\lambda, \hat{P}_\nu + \epsilon \Delta_{M_m})}{\mathrm{d}\epsilon} \right|_{\epsilon=0} = 0. \tag{40}$$

Proceeding as in the proof of Theorem 2 in Appendix B we obtain that, $\forall m$, $\forall \Delta_{M_m} \in \mathcal{L}_{\text{H.S.}}(\mathcal{H}_x^{\boxed{m}}; \mathcal{H}_y)$,

$$\sum_{i=1}^N [y_i - \hat{P}_\nu(x_i), \ \Delta_{M_m}(x_i)]_{\mathcal{H}_y} = \lambda [\hat{M}_m, \ \Delta_{M_m}]_{\mathcal{L}_{\text{H.S.}}}, \tag{41}$$

that is, $\forall m = 1, \ldots, \nu$, $\forall K_m$,

$$\hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}) = \frac{1}{\lambda} \sum_{i=1}^N a_{K_m}(x_i)(y_i - \hat{P}_\nu(x_i)). \tag{42}$$

Since, by definition,

$$\hat{P}_\nu(x_i) = \sum_{m=1}^\nu \sum_{K_m} a_{K_m}(x_i) \hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}), \tag{43}$$

replacing (42) in (43) yields

$$\begin{aligned} \hat{P}_\nu(x_i) =& \frac{1}{\lambda} \sum_{m=1}^\nu \sum_{K_m} a_{K_m}(x_i) \sum_{\ell=1}^N a_{K_m}(x_\ell)(y_\ell - \hat{P}_\nu(x_\ell)) \\ =& \frac{1}{\lambda} \left(\sum_{\ell=1}^N y_\ell \left(\sum_{m=1}^\nu [x_i, \, x_\ell]^m_{\mathcal{H}_x} \right) \right. \\ & \left. - \sum_{\ell=1}^N \hat{P}_\nu(x_\ell) \left(\sum_{m=1}^\nu [x_m, \, x_\ell]^m_{\mathcal{H}_x} \right) \right) \end{aligned} \tag{44}$$

Thus, by stacking $\hat{P}_\nu(x_i)$,

$$\text{col}_i(\hat{P}_\nu(x_i)) = \left(\lambda I_N + A_\nu(\mathbf{x})\right)^{-1} A_\nu(\mathbf{x}) \, \text{col}_i(y_i). \tag{45}$$

Finally, we replace (45) into (42) to obtain

$$\begin{aligned} \hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}) =& \frac{1}{\lambda} \sum_{i=1}^N a_{K_m}(x_i)(y_i - \hat{P}_\nu(x_i)) \\ =& \frac{1}{\lambda} \text{row}_i(a_{K_m}(x_i) \Big(I_N - \left(\lambda I_N + A_\nu(\mathbf{x})\right)^{-1} A_\nu(\mathbf{x}) \Big) \text{col}_i(y_i) \\ =& \text{row}_i(a_{K_m}(x_i) \left(\lambda I_N + A_\nu(\mathbf{x})\right)^{-1} \text{col}_i(y_i). \end{aligned} \tag{46}$$

Again, Assumption 2 together with (46) guarantees that $\hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m})$ exists $\forall t \in D$ and $\forall (k_1, \ldots, k_m)$. ■

Corollary 1: In the hypotheses of Theorem 3 the optimal polynomial estimate $\hat{P}_\nu \xi$, $\xi \in \mathcal{H}_x$, is given by

$$\hat{P}_\nu(\xi) = \sum_{m=1}^\nu \text{row}_i([\xi, x_i]^m_{\mathcal{H}_x}) \left(\lambda I_N + A_\nu(\mathbf{x})\right)^{-1} \text{col}_i(y_i). \tag{47}$$

□

The proof is entirely analogous to Theorem 2 in Section III-A. Remarkably, the polynomial estimate (47) has complexity proportional to N^2 just as the linear estimate (30). In other words, the computational complexity of (47) is insensitive to the degree ν of the polynomial, since $A_\nu(\mathbf{x})$ is obtained by summing the entry-wise powers of $A(\mathbf{x})$.

Remark 3: When one considers the full polynomial including the affine term, *i.e.*

$$y_i = P^\star_\nu(x_i) + M_0, \tag{48}$$

it is easy to prove that the estimation procedure of Theorem 3 and Corollary 1 is exactly the same and that the best estimate $\hat{M}_0$ of M_0 is given by

$$\hat{M}_0 = \frac{1}{N} \sum_{i=1}^N \left(y_i - \hat{P}_\nu(x_i) \right). \tag{49}$$

Example 2: For vector-valued continuous-time systems consider the case $\nu = 2$. The generic quadratic operator $P_2 = M_1(x) + M_2(x^{\boxed{2}}) \in \mathcal{L}_2^P$ can be represented as

$$\begin{aligned} (P_2 x)(t) =& y(t) = \int_D \kappa^{M_1}(t, \tau) x(\tau) \, \mathrm{d}\tau \\ & + \int_{D \times D} \kappa^{M_2}(t, \tau, \sigma)(x(\tau) \otimes x(\sigma)) \, \mathrm{d}\tau \mathrm{d}\sigma. \end{aligned} \tag{50}$$

Let $M_{\lambda,2} = \lambda I_n + A_2$. If we are interested in the explicit expression of the operator kernels of the optimal quadratic estimator (47), for $\kappa^{\hat{M}_1}$ we can use (32) of Example 1, with M_λ replaced by $M_{\lambda,2}$. To express $\kappa^{\hat{M}_2}$ we use (47), with $j = 2$. Straightforward manipulations yield

$$\begin{aligned} \hat{M}_2(x^{\boxed{2}}) = \int_{D^2} & \text{row}_i(y_i(t))(M^{-1}_{\lambda,2})^\top \text{col}_i(x_i(\tau) \otimes x_i(\sigma))^\top \\ & \cdot (x(\tau) \otimes x(\sigma)) \, \mathrm{d}\tau \mathrm{d}\sigma, \end{aligned} \tag{51}$$

and the underlying kernel $\kappa^{\hat{M}_2}(t, \tau, \sigma) \in \mathbb{R}^{n_y \times n_x^2}$ is

$$\kappa^{\hat{M}_2}(t, \tau, \sigma) = \text{row}_i(y_i(t))(M^{-1}_{\lambda,2})^\top \text{col}_i(x_i(\tau) \otimes x_i(\sigma))^\top. \tag{52}$$

C. Convergence of the approximated Volterra series

Having obtained the best polynomial estimate for any given order ν our next step is logically to derive the best regularized estimate of the Volterra series of the system by letting $\nu \to \infty$. Before doing so it is appropriate to offer some comments to put these results in the framework of recent kernel-based identification methods (see, among many others, [4], [23], [27], [62]). The results of the previous sections emphasize the prominent role of the matrices $A(\mathbf{x})$ (linear case) and $A_\nu(\mathbf{x})$, that have entries of the form, respectively, $[x_i, x_j]_{\mathcal{H}_x}$, $[x_i, x_j]^m_{\mathcal{H}_x}$, and encode the "similarity" of the elements x_i and x_j. The inner product may seem a very specific choice of similarity, but the framework described so far is abstract and the inner product in $\mathcal{H}_x$ is actually a design parameter that may lead to different estimates. In the continuous time case with finite-dimensional inputs/outputs the standard inner product $[x_i, x_j] = \int_D x_i(s)^\top x_j(s)\,\mathrm{d}s$ is an instance of the more general case $[x_i, x_j]_\kappa = \int_D x_i(s)^\top \kappa(s) x_j(s)\,\mathrm{d}s$, where $\kappa : D \to \mathbb{R}^{n_x \times n_x}$ is symmetric and positive definite. By changing κ we obtain different regularized estimates. As discussed in [53], κ corresponds to the notion of "kernel" of the regularized estimate (not to be confused with the kernel of a RKHS introduced in Section V), and its choice, discussed for example in [18] for the case of LTI systems, should crucially incorporate prior knowledge about the system to identify, for example stability or frequency content. Although the solutions described in sections III-A, III-B apply beyond LTI systems, they may incorporate this flexibility by customizing the inner product in $\mathcal{H}_x$ to reflect the prior knowledge about the system. Clearly, in a totally black-box approach one may not possess sufficient prior knowledge to make this choice. In this case it is still possible to estimate the unknown operator by choosing any "natural" inner product in the input space.

We now apply this flexibility in the choice of the inner product to estimate the Volterra series as the limit of a polynomial approximation. To our knowledge, this problem has been studied for the first time in [49]. In our case the problem is whether the polynomial approximations in (47) converge for $\nu \to \infty$. From the definition of $A_\nu(\mathbf{x})$ in (38) it is immediate to see that as $\nu \to \infty$ the entries of $A_\nu(\mathbf{x})$ diverge when $|[x_i, x_j]| > 1$. In order to obtain a convergent series of polynomials we may re-define the norm (20) with a weight function $\mathbf{w}$ so that higher-order monomials have a larger weight. Given a positive and unbounded function $\mathbf{w} : \mathbb{N} \to \mathbb{R}_+$, $\lim_{m\to\infty} \mathbf{w}(m) = \infty$, let us re-define the inner product and the norm in $\mathcal{L}^P_\nu$ as

$$[P^1_\nu,\, P^2_\nu]_{\mathbf{w}} = \sum_{m=1}^{\nu} \mathbf{w}(m)[M^1_m, M^2_m]_{\mathcal{L}_{\text{H.S.}}}, \tag{53}$$

$$\|P_\nu\|^2_{\mathbf{w}} = \sum_{m=1}^{\nu} \mathbf{w}(m)\|M_m\|^2_{\mathcal{L}_{\text{H.S.}}}. \tag{54}$$

By repeating the steps in the proof of Theorem 3 we easily obtain the representation of the optimal polynomial estimator with respect to a functional based on this new norm.

Theorem 4: If Assumption 2 holds, then, given a positive and unbounded function $\mathbf{w} : \mathbb{N} \to \mathbb{R}_+$, a training set $\{x_i, y_i\}$, $i = 1, \dots, N$ and the functional

$$J^\nu_N(\lambda, P_\nu) = \sum_{i=1}^{N} \|y_i - P_\nu(x)\|^2_{\mathcal{H}_y} + \lambda \|P_\nu\|^2_{\mathbf{w}} \tag{55}$$

then for any $\lambda \in \mathbb{R}$ the optimal polynomial estimator with respect to (55) exists and its monomials $\hat{M}_m$ are given by

$$\hat{M}_m(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}) = \mathrm{row}_i \frac{a_{K_m}(x_i)}{\mathbf{w}(m)} \left(\lambda I_N + A_{\mathbf{w},\nu}(\mathbf{x})\right)^{-1} Y \tag{56}$$

where $\{\phi_k\}$ is an orthonormal basis of $\mathcal{H}_x$, $Y = \mathrm{col}_i(y_i)$ and the matrix $A_{\mathbf{w},\nu}(\mathbf{x}) \in \mathbb{R}^{N\times N}$ has entries

$$\left(A_{\mathbf{w},\nu}(\mathbf{x})\right)_{(i,j)} = \sum_{m=1}^{\nu} \frac{[x_i, x_j]^m_{\mathcal{H}_x}}{\mathbf{w}(m)}. \tag{57}$$

Moreover, $\hat{M}(\phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_j})$ are defined $\forall t \in D$ and

$$\hat{P}_{\mathbf{w},\nu}(\xi) = \sum_{m=1}^{\nu} \mathrm{row}_i \left(\frac{[\xi, x_i]^m_{\mathcal{H}_x}}{\mathbf{w}(m)}\right) \left(\lambda I_N + A_{\mathbf{w},\nu}(\mathbf{x})\right)^{-1} Y. \tag{58}$$

□

At this point we can let $\nu \to \infty$, provided that $\mathbf{w}$ is suitably chosen. As already remarked in [49] different choices of $\mathbf{w}$ yield different possible polynomial kernels. The step forward with respect to [49] is that, in the light of the results mentioned in Section I, all these kernels are *universal*, in the sense that they are capable of uniformly approximating all S-continuous functions on bounded sets of $\mathcal{H}_x$. For example, if we choose $\mathbf{w}(m) = w^m$ with $w > 1$ we obtain the following limit for the Volterra series.

Corollary 2: In the hypotheses of Theorem 4, if $\mathbf{w}(m) = w^m$ with $w > \max_{i,j}([x_i, x_j]_{\mathcal{H}_x})$, then the function $\hat{P}_{\mathbf{w},\infty}(\xi) = \lim_{\nu\to\infty} \hat{P}_{\mathbf{w},\nu}(\xi)$ is well defined for all ξ such that $[\xi, x_i]_{\mathcal{H}_x} < w$, $i = 1, \dots, N$, and, with $Y = \mathrm{col}_i(y_i)$,

$$\hat{P}_{\mathbf{w},\infty}(\xi) = r^\top_{\mathbf{w},\infty}(\xi, \mathbf{x}) \left(\lambda I_N + A_{\mathbf{w},\infty}(\mathbf{x})\right)^{-1} Y \tag{59}$$

$$r^\top_{\mathbf{w},\infty}(\xi, \mathbf{x}) = \mathrm{row}_i \left(\frac{[\xi, x_i]_{\mathcal{H}_x}}{w - [\xi, x_i]_{\mathcal{H}_x}}\right) \tag{60}$$

$$\left(A_{\mathbf{w},\infty}(\mathbf{x})\right)_{(i,j)} = \frac{[x_i, x_j]_{\mathcal{H}_x}}{w - [x_i, x_j]_{\mathcal{H}_x}}. \tag{61}$$

□

It is natural to identify (59) as the estimate of the Volterra series of the unknown system based on the rational kernel (61). Notice that, since $[x_1, x_2] \leqslant \|x_1\|\|x_2\|$, the constraint on w of Corollary 2 translates into a bound on the energy of the input signals. Another interesting choice of $\mathbf{w}$, inspired to [49], is

$$\mathbf{w}(m) = m!\kappa(t)^m, \tag{62}$$

with $\kappa(t) > 0$ a positive and monotonic function of $t \in D$. The choice (62) yields the exponential kernel

$$\left(A_{\mathbf{w},\infty}(\mathbf{x})\right)_{(i,j)} = e^{\frac{1}{\kappa(t)}[x_i, x_j]_{\mathcal{H}_x}} - 1. \tag{63}$$

The function $\kappa(t)$ is chosen monotonically increasing with t in order to normalize the (potentially large) values $[x_i, x_j]_{\mathcal{H}_x}$ and to prevent numerical issues.

Remark 4: The learning procedure described in Section III-B requires to fix two meta-parameters, (λ, ν), whereas the procedure of Section III-C requires $(\lambda, \mathbf{w})$, where $\mathbf{w}$ is a function. The choice of λ is discussed in Section III-D. Other approaches are possible since, as discussed at the beginning of this section, the inner product itself can be considered a design choice. For example, if the inner product is defined so that $|[x_i, x_j]| < 1$ the function $\mathbf{w}$ is no longer necessary. In all cases the choice of the kernel for $[x_i, x_j]$ determines the kernel $(A_{\mathbf{w},\infty})$ of the Volterra series.

Remark 5: A property of estimator in Section III-B is that its complexity does not depend on the degree chosen for the polynomial estimator. Moreover, the estimate of the Volterra series (59) only requires finite-dimensional operations. In general, the complexity of the learning procedure is quadratic in the size of the training set due to the need of computing $[x_i, x_j]_{\mathcal{H}_x}$ for all possible pairs. The computational cost becomes linear in N if it is possible to design a training set of orthogonal input functions, since this makes the matrix $A_\nu(\mathbf{x})$ block diagonal.

Remark 6: Since $\hat{P}_\nu : \mathcal{H}_x \to \mathcal{H}_y$, the framework described in this section guarantees that the estimated trajectories are square integrable on D for any input x. For example, when $\mathcal{H}_y = \mathcal{L}_2(\mathbb{R}_+; \mathbb{R}^n)$ it is obvious that $\lim_{t\to\infty} \hat{P}(x)(t) = 0$. When D is a bounded temporal domain, it may be of interest to extend the estimate $\hat{y}$ to a larger domain. In other words, given a training set in $[0, T]$, one might want to estimate the output of the system for an input in $[0, \infty)$. This interesting extension is a theme of further research.

D. Choice of the regularization parameter and noisy data

In Section V we shall see that the space of the estimators is a reproducing kernel Hilbert space, and thus the choice of λ may exploit results in that area [27], [42]. In this section we show how to choose λ in the case of noisy measurements.

The results in Theorem 2 are easily extended to the linear space of polynomial operators. In particular, we have that whenever there exists a "true" polynomial operator $P^\star_\nu$ such that $y_i = P^\star_\nu(x_i)$, for all i, and the data $\{(x_i, y_i)\}$ are exact, then the estimate (47) is such that

$$\lim_{\lambda\to 0^+} \hat{P}_\nu = \arg \min_{P_\nu \in \mathcal{L}^P_\nu} \left\{ \|P_\nu\|^2_{\mathcal{L}^P_\nu} : \; y_i = P_\nu(x_i), \right\} \tag{64}$$

that is, $\lim_{\lambda\to 0^+} \hat{P}_\nu$ is the minimum norm polynomial operator of degree ν that yields a null residual on the training set. Notice that $\lim_{\lambda\to 0^+} \hat{P}_\nu = P^\star_\nu$ is not ensured (the true operator could be not minimum norm). In presence of noise it is in general $\sum_i \|y_i - P^\star_\nu(x_i)\|^2 > 0$, that is, the residual is not null even for the true operator. It is therefore useless to decrease λ to reduce $\sum_i \|y_i - P^\star_\nu(x_i)\|^2$ too much, and we obtain the following maximum likelihood rule to tune λ:

Optimal Rule - Choose λ so that $\sum_i \|y_i - \hat{P}_\nu(x_i)\|^2$ equals the expected value of the residual of $P^\star_\nu$.

The optimal rule can be implemented when the expected value does not depend on $P^\star_\nu$ itself, for example additive output noise, $y_i = P^\star_\nu(x_i) + n^y_i$, where the expected value of the residual is $\sum_i \mathbb{E}[\|n^y_i\|^2_{\mathcal{H}_y}]$. Conversely, the optimal rule cannot be implemented when the the residual depends on $P^\star_\nu$, that is not available. This is for example the case of additive noise in the input, $y_i = P^\star_\nu(x_i + n^x_i) + n^y_i$. The following heuristic rule is commonly used in these cases [33]:

Heuristic Rule - Choose λ so that $\sum_i \|y_i - \hat{P}_\nu(x_i)\|^2$ equals its expected value.

The rule can be justified as follows. In the first place, it is statistically satisfied by $P^\star_\nu$ itself. In the second place, when the "true" operator belongs to $\mathcal{L}^P_\nu$, the rule yields a unique value of λ, since the sample residual of $\hat{P}_\nu$ increases with λ (Theorem 2) and its expected value depends on $\|\hat{P}_\nu\|$ that decreases with λ (Theorem 2). Finally, it is worth mentioning that the application of Heuristic Rule can be implemented by extending criteria that hold for the finite-dimensional case based on the concept of equivalent degrees of freedom of regularized estimators [63], [64].

When more hyper-parameters are added to the scheme, for example the function $\mathbf{w}$, the degree ν of the polynomial, or parameters that specialize the inner product, it is necessary to resort to more powerful hyper-parameter estimation techniques, like empirical Bayes and marginal likelihood [65], C_p statistic [66], k-fold cross validation [67], generalized cross validation [68], etc.

IV. Recursive computation of causal estimates

In the estimation of dynamical system D may be a temporal domain and the variables x, y in $y = F(x)$ are functions of time, *i.e.* $x \in \mathcal{L}_2([0, T]; H_1)$, $y \in \mathcal{L}_2([0, T]; H_2)$. It makes sense to introduce in the estimate of F a causality constraint, *i.e.* restrict ourselves to causal operators.

Definition 5: An operator $F : \mathcal{L}_2([0, T]; H_1) \to \mathcal{L}_2([0, T]; H_2)$ is said to be *causal* if, whenever $y(t)$ exists, for some $t \in [0, T]$, $y(t)$ depends only on $x(\tau)$, $\tau \in [0, t]$. □

The optimal polynomial estimate (47) is not causal, because $\hat{P}_\nu(\xi)(t)$ is obtained as a linear combination of $y_i(t)$ with coefficients that depends on $\xi \in [0, T]$ via the inner products $[\xi, x_i]_{\mathcal{H}_x}$. However, the estimate is causal for $t = T$. Consequently, a simple method to obtain a causal estimate is to define two parametric spaces $\mathcal{H}^t_x = \mathcal{L}_2([0, t]; H_1)$, $\mathcal{H}^t_y = \mathcal{L}_2([0, t]; H_2)$ and let the estimates evolve with t. This estimate can be computed at any t by integrating a differential representation. Let us rewrite (47) as

$$\hat{y}(t) = \hat{P}_\nu(\xi)(t) = r_\nu(\xi, \mathbf{x}, t)^\top M^{-1}_{\lambda,\nu}(\mathbf{x}, t) \mathrm{col}_i(y_i(t)) \tag{65}$$

$$r_\nu(\xi, \mathbf{x}, t)^\top = \sum_{m=1}^{\nu} \mathrm{row}_i \left(\left([\xi, x_i]_{\mathcal{H}^t_x} \right)^m \right) \tag{66}$$

$$M_{\lambda,\nu}(\mathbf{x}, t) = (\lambda I_N + A_\nu(\mathbf{x}, t)) \tag{67}$$

$$(A_\nu(\mathbf{x}, t))_{i,j} = \sum_{m=1}^{\nu} \left([x_i, x_j]_{\mathcal{H}^t_x} \right)^m, \tag{68}$$

where $r_\nu(\xi, \mathbf{x}, 0) = 0$, $A_\nu(\mathbf{x}, 0) = 0$, $M_{\lambda,\nu}(\mathbf{x}, 0) = \lambda I_N$. Clearly, $\hat{y}(t)$ is a causal estimate, because it depends only on past values of x_i and y_i. We can easily obtain a differential representation by recalling that

$$\frac{\mathrm{d}}{\mathrm{d}t} [x_i, x_j]_{\mathcal{H}^t_x} = [x_i(t), x_j(t)]_{H_1} \tag{69}$$

thus the terms in (65) can be computed by integrating the scalar products from a null initial condition. (65) provides the sought causal and recursive estimate of the optimal $\hat{P}_\nu(\xi)(t)$. This causal estimate can be easily adapted to the case of discrete-tine systems by replacing the differential equation with a difference equation.

V. Connection with RKHS

In this Section we interpret the results of Section III in the light of the theory of reproducing kernel Hilbert spaces (RKHS) in order to establish a connection and shed more light on the properties of the estimation framework previously described. We specialize the abstract definition of RKHS [39], [40] to our framework.

Definition 6: Given a set X and a Hilbert space $\mathcal{H}_y$, a $\mathcal{H}_y$-valued RKHS on X is a Hilbert space $\mathcal{H}$ such that the elements of $\mathcal{H}$ are functions $f:\ X \to \mathcal{H}_y$ and $\forall x \in X$ there exists a positive constant C_x such that

$$\|f(x)\|_{\mathcal{H}_y} \ \leqslant\ C_x \|f\|_{\mathcal{H}}. \tag{70}$$

Definition 7: A $\mathcal{H}_y$-valued *kernel* of positive type on $X \times X$ is a map

$$K:\ X \times X \to \mathcal{B}(\mathcal{H}_y; \mathcal{H}_y) \tag{71}$$

where $\mathcal{B}(\mathcal{H}_y; \mathcal{H}_y)$ is the Banach space of bounded operators $\mathcal{H}_y \to \mathcal{H}_y$ with the uniform norm, such that $\forall N \in \mathbb{N}$, $x_i \in X$, $c_i \in \mathbb{C}$, $i = 1, \dots, N$, $\forall y \in \mathcal{H}_y$,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i \bar{c}_j \left[K(x_i, x_j) y, y\right]_{\mathcal{H}_y} \ \geqslant\ 0. \tag{72}$$

□

A $\mathcal{H}_y$-valued RKHS on X canonically defines a $\mathcal{H}_y$-valued *kernel* of positive type on $X \times X$ as follows [39], [40]. Given $x \in X$, define the **evaluation** operator $\mathbf{ev}_x : \mathcal{H} \to \mathcal{H}_y$ such that, for any $f \in \mathcal{H}$

$$\mathbf{ev}_x(f) = f(x). \tag{73}$$

(70) guarantees that the evaluation map $\mathbf{ev}$ is a bounded operator with the conjugate $\mathbf{ev}_x^* : \mathcal{H}_y \to \mathcal{H}$. The *reproducing kernel* $K:\ X \times X \to \mathcal{B}(\mathcal{H}_y; \mathcal{H}_y)$ associated to $\mathcal{H}$ is

$$K(x_1, x_2) = \mathbf{ev}_{x_1}\, \mathbf{ev}_{x_2}^*. \tag{74}$$

Lemma 2: [39] K defined by (74) is of positive type. □

Conversely, if $\mathcal{H}$ is a real vector space, any *symmetric* $\mathcal{H}_y$-valued kernel of positive type (*i.e.* $K(x_1, x_2) = K(x_2, x_1)$) defines a unique $\mathcal{H}_y$-valued RKHS $\mathcal{H}$ whose reproducing kernel is K (see Proposition 2.3 in [39]). If $\mathcal{H}$ is a complex vector space, a kernel of positive type is always hermitian.

The first property that descends from the definition (74) of the reproducing kernel is that $\forall x_1, x_2 \in X$ and $y \in \mathcal{H}_y$,

$$K(x_1, x_2) y = \mathbf{ev}_{x_1}\, \mathbf{ev}_{x_2}^*\, y = (\mathbf{ev}_{x_2}^*\, y)(x_1)\ \in \mathcal{H}_y \tag{75}$$

and consequently, $\forall x \in X$, $y \in \mathcal{H}_y$,

$$\mathbf{ev}_x^*\, y = K(., x) y\ \in \mathcal{H}, \tag{76}$$

that defines the conjugate $\mathbf{ev}_x^*$ of $\mathbf{ev}_x$ from the reproducing kernel K. The second property that immediately descends from the definition (74) is the *reproducing property*

$$[f(x),\, y]_{\mathcal{H}_y} = [\mathbf{ev}_x\, f, y]_{\mathcal{H}_y} = [f, \mathbf{ev}_x^*\, y]_{\mathcal{H}} \tag{77}$$

that holds $\forall f \in \mathcal{H}$, $x \in X$, $y \in \mathcal{H}_y$.

We now return to the operators $\mathcal{H}_x \to \mathcal{H}_y$ considered in Section III. Given two Hilbert spaces $\mathcal{H}_x$ and $\mathcal{H}_y$, it is easy to prove that the Hilbert space $\mathcal{L}_{\text{H.S.}}$ of H-S operators $\mathcal{H}_x \to \mathcal{H}_y$ is a RKHS (with $X = \mathcal{H}_x$).

Theorem 5: The Hilbert space $\mathcal{L}_{\text{H.S.}}$ of H-S operators $L : \mathcal{H}_x \to \mathcal{H}_y$ is a RKHS with kernel $K(x_1, x_2) = [x_1,\, x_2]_{\mathcal{H}_x} I_{\mathcal{H}_y}$, where $I_{\mathcal{H}_y}$ denotes the identity in $\mathcal{H}_y$.

Proof: Hilbert-Schmidt operators are bounded and (10) implies that (70) holds with $C_x = \|x\|_{\mathcal{H}_x}$. Moreover, since

$$[K(x, x)y, y]_{\mathcal{H}_y} = \|x\|^2_{\mathcal{H}_x} \|y\|^2_{\mathcal{H}_y} \geqslant 0 \tag{78}$$

it is easy to verify that K is a $\mathcal{H}_y$-valued kernel of positive type and that $K(\cdot, x)y = [\cdot,\, x]_{\mathcal{H}_x} y$ is a H-S operator $\mathcal{H}_x \to \mathcal{H}_y$ that satisfies the reproducing property (77) since, $\forall L \in \mathcal{L}_{\text{H.S.}}$, $x \in \mathcal{H}_x$, $y \in \mathcal{H}_y$, one has

$$\begin{aligned}[L, \mathbf{ev}_x^*\, y]_{\mathcal{L}_{\text{H.S.}}} &= [L, [.,\, x]_{\mathcal{H}_x}\, y]_{\mathcal{L}_{\text{H.S.}}} \\ &= \sum_{k=1}^{\infty} [L\phi_k,\, [\phi_k, x]_{\mathcal{H}_x} y]_{\mathcal{H}_y} = [Lx,\, y]_{\mathcal{H}_y}.\end{aligned} \tag{79}$$

Finally, the completion of the linear space of H-S operators in the form $\sum_{k=1}^{N} K(., \phi_k) y_k = \sum_{k=1}^{N} [.,\, \phi_k]_{\mathcal{H}_x} y_k$, with arbitrary N and $\{y_k\}$ is exactly $\mathcal{L}_{\text{H.S.}}$, since it follows from (8) that,

$$Lx = \sum_{k=1}^{\infty} [x, \phi_k]_{\mathcal{H}_x} L\phi_k = \sum_{k=1}^{\infty} K(x, \phi_k) y_k, \tag{80}$$

with $y_k = L\phi_k$. Since K is symmetric, uniqueness of the kernel follows from Proposition 2.3 in [39]. ■

Theorem 5 shows that $K(x_1, x_2)$ generates *any* operator $L \in \mathcal{L}_{\text{H.S.}}$. H-S operators are, in this sense, the most "natural" space of linear operators between $\mathcal{L}_2$ spaces that constitute a RKHS.

Example 3: In the case of continuous-time systems $\mathcal{H}_x = \mathcal{L}_2(D; \mathbb{R}^{n_x})$, $\mathcal{H}_y = \mathcal{L}_2(D; \mathbb{R}^{n_y})$, the kernel of the RKHS, $K(x_1, x_2) = [x_1, x_2]_{\mathcal{H}_x} I_{\mathcal{H}_y}$, is

$$\begin{aligned}[K(x_1, x_2)y](t) =& [x_1, x_2]_{\mathcal{H}_x} y(t) \\ =& \left(\int_D x_1^\top(\tau) x_2(\tau)\, \mathrm{d}\tau \right) y(t),\end{aligned}$$

and the corresponding RKHS is the linear space of H-S operators $L_x : \mathcal{H}_x \to \mathcal{H}_y$ defined as $L_x \xi = [\xi, x]_{\mathcal{H}_x} y$ where $\xi,\, x \in \mathcal{H}_x$, $y \in \mathcal{H}_y$.

The essential property that connects RKHS and Volterra series based on H-S monomials is that the Hilbert space $\mathcal{L}_\nu^P$ of polynomial operators $P_\nu :\ \mathcal{H}_x \to \mathcal{H}_y$ introduced in Section II-B turns out to be a RHKS (with $X = \mathcal{H}_x$).

Theorem 6: The linear space $\mathcal{L}_\nu^P$ of polynomial operators $P_\nu :\ \mathcal{H}_x \to \mathcal{H}_y$ with H-S monomials of Theorem 1 is a RKHS with kernel $K_\nu(x_1, x_2) = \sum_{m=1}^{\nu} [x_1, x_2]^m_{\mathcal{H}_x} I_{\mathcal{H}_y}$, where $I_{\mathcal{H}_y}$ is the identity operator in $\mathcal{H}_y$.

Proof: We already know that $\mathcal{L}_\nu^P$ is a Hilbert space of functions $\mathcal{H}_x \to \mathcal{H}_y$ thus we need only to prove (70). This is straightforward to prove, since

$$\begin{aligned}\|P_\nu(x)\|_{\mathcal{H}_y} =& \left\| \sum_{m=1}^{\nu} M_m(x^{\boxed{m}}) \right\| \leqslant \sum_{m=1}^{\nu} \|M_m\|_{\text{H.S.}} \left\|x^{\boxed{m}}\right\|_{\mathcal{H}_x^{\boxed{m}}} \\ \leqslant& \sum_{m=1}^{\nu} \|M_m\|_{\text{H.S.}} \|x\|_{\mathcal{H}_x}^m \\ \leqslant& \left(\sum_{m=1}^{\nu} \|x\|_{\mathcal{H}_x}^{2m} \right)^{\frac{1}{2}} \left(\sum_{m=1}^{\nu} \|M_m\|_{\text{H.S.}}^2 \right)^{\frac{1}{2}} \\ =& C_x \|P_\nu\|_{\mathcal{L}_\nu^p}, \end{aligned} \tag{81}$$

where we have used Theorem 1. We have thus proved that $\mathcal{L}_\nu^P$ is a RKHS. The kernel K_ν is clearly of positive type. Notice that from (4) it follows that $K_\nu(x_1, x_2) = \sum_{m=1}^{\nu} [x_1^{\boxed{m}}, x_2^{\boxed{m}}]_{\mathcal{H}_x^{\boxed{m}}} I_{\mathcal{H}_y}$ and thus $K_\nu(., x)y$ is an operator $\mathcal{H}_x \to \mathcal{H}_y$ in $\mathcal{L}_\nu^P$. We can easily verify that K_ν satisfies the reproducing property (77) since, $\forall P \in \mathcal{L}_\nu^P$, $x \in \mathcal{H}_x$, $y \in \mathcal{H}_y$,

$$\begin{aligned}[P, \mathbf{ev}_x^* y]_{\mathcal{L}_\nu^p} =& \sum_{m=1}^{\nu} \left[M_m, [., x^{\boxed{m}}]_{\mathcal{H}_x^{\boxed{m}}} y \right]_{\mathcal{L}_{\text{H.S.}}(\mathcal{H}_x^{\boxed{m}}; \mathcal{H}_y)} \\ =& \sum_{m=1}^{\nu} \sum_{K_m} \left[M_m\left(\phi_{K_m}\right), \left[\phi_{s(m)}, x^{\boxed{m}}\right]_{\mathcal{H}_x^{\boxed{m}}} y \right]_{\mathcal{H}_y} \\ =& \sum_{m=1}^{\nu} \sum_{K_m} \left[M_m\left(\phi_{K_m}\right), \prod_{j \in K_m} [x, \phi_j]_{\mathcal{H}_x} y \right]_{\mathcal{H}_y} \\ =& \sum_{i=m}^{\nu} \left[\sum_{K_m} \prod_{j \in K_m} [x, \phi_j]_{\mathcal{H}_x} M_m\left(\phi_{K_m}\right), y \right]_{\mathcal{H}_y} \\ =& \left[\sum_{m=1}^{\nu} M_m(x^{\boxed{m}}), y \right]_{\mathcal{H}_y} = [P_\nu(x), y]_{\mathcal{H}_y}, \end{aligned} \tag{82}$$

where we have used (12), and $\phi_{K_m} := \phi_{k_1} \boxtimes \cdots \boxtimes \phi_{k_m}$. ■

Example 4: In the case of continuous-time systems $\mathcal{H}_x = \mathcal{L}_2(D; \mathbb{R}^{n_x})$, $\mathcal{H}_y = \mathcal{L}_2(D; \mathbb{R}^{n_y})$, the kernel of the RKHS of polynomial operators, $K_\nu(x_1, x_2)$, is

$$\begin{aligned}[K_\nu(x_1, x_2)y](t) =& \sum_{m=1}^{\nu} \left([x_1, x_2]_{\mathcal{H}_x}\right)^m y(t) \\ =& \sum_{m=1}^{n} \left(\int_D x_1^\top(\tau) x_2(\tau) \, \mathrm{d}\tau \right)^m y(t), \end{aligned}$$

and the corresponding RKHS is the linear space of polynomial operators $P_\nu^x : \mathcal{H}_x \to \mathcal{H}_y$ defined as $P_\nu^x \xi = \sum_{m=1}^{\nu} [\xi, x]_{\mathcal{H}_x}^m y$ where ξ, $x \in \mathcal{H}_x$, $y \in \mathcal{H}_y$.

As a consequence of Theorem 6 we can apply the results concerning RKHS to $\mathcal{L}_\nu^P$. The optimal polynomial estimator (47) can be rewritten, by using the notation of the RKHS $\mathcal{L}_\nu^P$

$$\begin{aligned}\hat{P}_\nu =& \text{row}_i(K_\nu(., x_i)) M_{\lambda,\nu}^{-1} \text{col}_i(y_i) \\ =& \sum_{i=1}^{N} K_\nu(., x_i) (M_{\lambda,\nu}^{-1})_i \text{col}_i(y_i) \end{aligned} \tag{83}$$

where we have denoted $M_{\lambda,\nu} = (\lambda I_N + A_\nu(\mathbf{x})) \in \mathbb{R}^{N \times N}$, and $(M_{\lambda,\nu}^{-1})_i$ the i-th row of $M_{\lambda,\nu}$. This result is a consequence of the representer theorem for RKHS (Theorem 4.2 in [61]), see also [2], [32], [43].

Theorem 7: [61] Given a set $\{(x_i, y_i)\}$, $x_i \in \mathcal{H}_x$, $y_i \in \mathcal{H}_y$, $i = 1, \ldots, N$, and the functional J_N^ν in (37), the minimizer of J_N^ν, $\hat{P}_\nu = \arg\min_{\mathcal{L}_\nu^P} J_N^\nu(P_\nu)$ can be represented as

$$\hat{P}_\nu = \sum_{i=1}^{N} K_\nu(., x_i) \hat{c}_i, \quad \hat{c}_i \in \mathcal{H}_y. \tag{84}$$

It is immediate to verify that (84) coincides with (47) with $M_{\lambda,\nu} \text{col}_i(\hat{c}_i) = \text{col}_i(y_i)$. In other words, the optimal polynomial estimation derived in Theorem 3 and Corollary 1 coincides with the solution arising from the theory of RKHS.

VI. Numerical example

In this example we show that a good estimate of the input-output behavior of a nonlinear system can be obtained with a moderate computational burden when the training set is based on input functions that are similar to the ones to be estimated. This similarity condition is reasonable, since in the nonlinear case the only way to "learn" the behavior of the system is from similar cases, a restriction that can be lifted for linear operators. We also illustrate how to set the regularization parameter λ from prior information of the measurement noise.

We consider a Volterra-Lotka system with polynomial nonlinearities and measurement noise, described by

$$\dot{z}_1 = -a_{11} z_1 + a_{12} z_1 z_2 \tag{85}$$

$$\dot{z}_2 = a_{21} z_1 - a_{22} z_1 z_2 + x \tag{86}$$

$$y_1 = z_1 + n_1(\omega) \tag{87}$$

$$y_2 = z_2 + n_2(\omega), \tag{88}$$

where $x \in \mathcal{H}_x = \mathcal{L}([0, T]; \mathbb{R})$ is a non-negative input, the parameters are $a_{11} = a_{22} = 0.5$, $a_{12} = a_{21} = 0.25$, and we assume $y = (y_1, y_2) \in \mathcal{H}_y = \mathcal{L}([0, T]; \mathbb{R}^2)$ are available noisy measurements of the state (z_1, z_2). The measurement noise $n = (n_1, n_2) \in \mathcal{H}_y$ chosen in the example is a colored stochastic process generated as

$$\dot{n}_h(t) = -a_n n_h(t) + b_n \omega_h(t), \tag{89}$$

for $h = 1, 2$, with $a_n = -2$; $b_n = 0.1$ and ω_h mutually independent white-noise processes. Notice that above, in accordance with standard notation, the subscripts $h = 1, 2$ denote state, output and noise components, and not their realizations, which will be denoted with subscript letter i in what follows. Clearly, if $F(x)$ is the true nonlinear operator of the Volterra-Lotka system, $\mathbb{E}[\|y - F(x)\|_{\mathcal{H}_y}^2] = \mathbb{E}[\|n\|_{\mathcal{H}_y}^2]$. The noise can be represented as $n = F_N(\omega_1, \omega_2)$, where F_N is the H-S operator described by the equations (89). From the white-noise theory it is known that $\mathbb{E}[\|n\|_{\mathcal{H}_y}^2] = \|F_N\|_{\text{H.S.}}^2$ (see for example Lemma 4, [53]). Therefore, the knowledge of the noise parameters allows to apply the Optimal Rule of Section III-D to choose the optimal value of λ. The norm of F_N can be computed as

$$\|F_N\|_{\text{H.S.}}^2 = \int_0^T \int_0^t \left\| e^{a_n (t-\tau)} b_n \right\|_{\mathbb{R}}^2 \mathrm{d}\tau \, \mathrm{d}t, \tag{90}$$

and, with $T = 250$ and the parameters listed above we obtain $\|F_N\|_{\text{H.S.}}^2 = 1.25$.

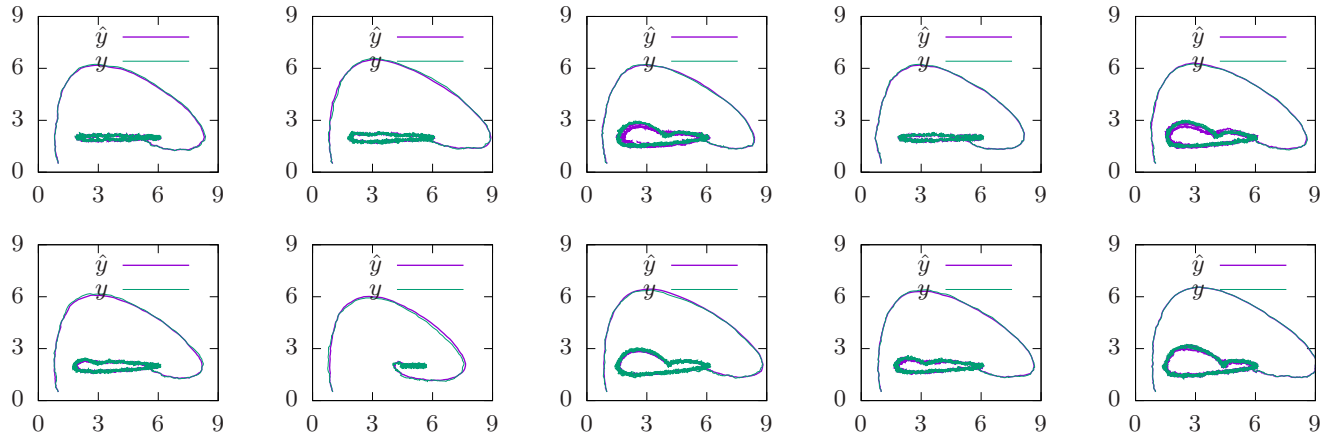

Fig. 1. Noisy output $\boldsymbol{y_i}$ on $\mathbf{10}$ samples of the training set compared to $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ generated by the estimate (59) of the Volterra series for $\boldsymbol{\lambda} = \mathbf{2} \cdot \mathbf{10^{-4}}$.

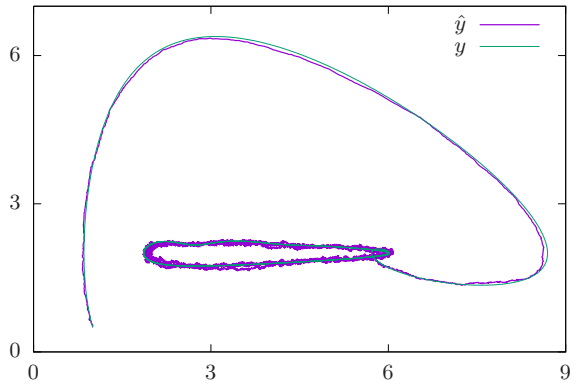

Fig. 2. Noise-less output $\boldsymbol{y}$ of the test function compared to (noisy) $\hat{\boldsymbol{y}}$ generated by the estimate (59) of the Volterra series for $\boldsymbol{\lambda} = \mathbf{2} \cdot \mathbf{10^{-4}}$. The value of $\boldsymbol{\lambda}$ has been chosen according to the Optimal Rule of Section III-D.

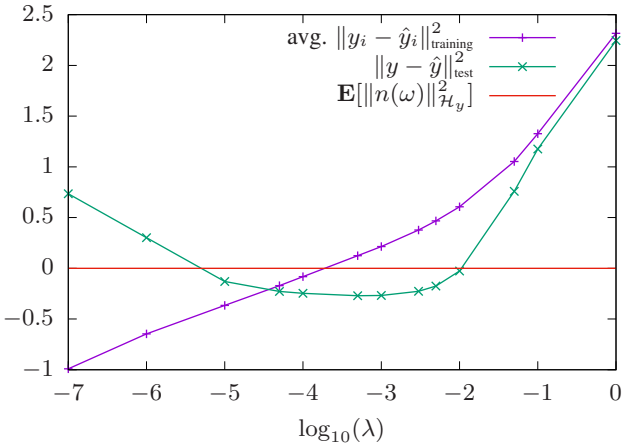

Fig. 3. The average residual of the training set for the Volterra-Lotka system, $\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \|\boldsymbol{y(t)} - \hat{\boldsymbol{y}}\boldsymbol{(t)}\|^{\mathbf{2}}$, increases with $\boldsymbol{\lambda}$ as predicted by the theory (log scale on the $\boldsymbol{y}-$axis). In presence of noise, the error on the test input has a minimum when the residual equals the expected value of the norm of the noise, at $\boldsymbol{\lambda} \simeq \mathbf{10^{-3}}$, in accordance with the Optimal Rule of Section III-D.

We used a training set of $N = 100$ harmonic functions having form $x_i(t) = 1+\sin(\theta_i t+\phi_i)$, where $\theta_i \in [0, \frac{15\pi}{T}]$, $\phi_i \in [0, 1]$ are random variables uniformly distributed, $T = 250$ is the time horizon of the input trajectory, and $i = 1, \ldots, N$. The estimate was computed by using the Volterra series estimate (59) with the rational kernel (61)

$$\left(A_{\mathbf{w},\infty}\right)_{(i,j)}(\mathbf{x}) = \frac{[x_i, x_j]_{\mathcal{H}_x}}{w - [x_i, x_j]_{\mathcal{H}_x}}.$$

with $w = 4\max_{i,j}\{[x_i,\, x_j])_{\mathcal{H}_x}$ and $\hat{M}_0 = \bar{y} - \hat{P}_{w,\infty}\bar{x}$.

A sample of 10 noisy output trajectories y_i in the training set is plotted in the phase state for $t \in [0, T]$ in Fig. 1, together with the corresponding estimates obtained by the estimator (59)–(61) for $\lambda = 2 \cdot 10^{-4}$ and w.

The same estimator is used on the input test function $x(t) = 1+\sin(\frac{8\pi}{T}t+0.5)$. Fig. 2 shows the true and estimated output for this test function. Notice that the true output is plotted without measurement noise, whereas the estimated output is not as smooth because of the noise in the training set.

The choice $\lambda = 2 \cdot 10^{-4}$ for these plots was made in accordance to the Optimal Rule of Section III-D, based on the knowledge of the measurement noise. Fig. 3 shows that $\lambda = 2 \cdot 10^{-4}$ is the value that makes the average norm of the residuals $\frac{1}{N}\sum_{i=1}^{N} \|y_i(t) - \hat{y}_i(t)\|^2$ equal to $\mathbb{E}[\|n\|^2_{\mathcal{H}_y}] = \|F_N\|^2_{\text{H.S.}} = 1.25$ (the vertical axis in this plot is in logaritmic scale). We notice that this value of λ allows to obtain an error norm in the test function very close to the minimum that is obtained approximately for $\lambda \in [10^{-4}, 10^{-3}]$. We remark that this choice of the regularization parameter is based on some *a priori* knowledge of the measurement noise.

Next, we compare causal and non-causal estimates for this system. The difference between the two cases is more evident with larger levels of measurement noise, thus the plots in Fig. 4 have been obtained with $b_n = 0.4$ that yields $\|F_N\|^2_{\text{H.S.}} = 19.98$. In this case we used the exponential kernel (63), $\left(A_{\mathbf{w},\infty}\right)_{(i,j)}(\mathbf{x}) = e^{\frac{1}{\kappa(t)}[x_i,x_j]_{\mathcal{H}_x}} - 1$, with

$$\kappa(t) = \max_{i,j}\left\{[x_i, x_j]_{\mathcal{H}_x^t}\right\} \tag{91}$$

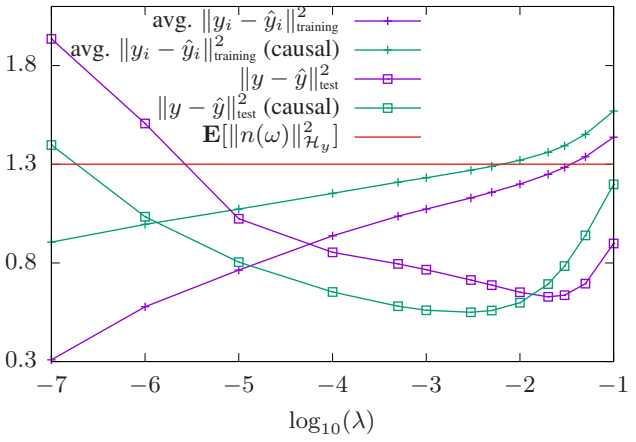

Fig. 4. Average residuals and errors at varying $\boldsymbol{\lambda}$ for causal and non causal estimates (log scale on the $\boldsymbol{y}-$axis). Causal estimates have larger residuals but smaller errors. Notice that in both cases the minimum is reached when the residual equals the norm of the noise.

as a normalization factor to avoid numerical instabilities. Fig. 4 shows that causal estimates have larger error on the training set, which is not surprising because of the additional causality constraint, but smaller errors on the test set. The error reduction when using the causal estimate, although not so evident from the logarithmic plot, is significant (around 15%). We notice that also with the exponential kernel for both estimates the minimum error on the test set is obtained when the error on the training set equals the norm of the measurement noise, in accordance with the Optimal Rule of Section III-D.

VII. Conclusions

In this work, we have presented a general estimator for nonlinear unknown input-output operators. There are several crucial aspects that warrant further investigation, including the selection of the regularization parameter in the presence of input noise, the determination of the Volterra series kernel based on prior information, and the potential extension of the estimator to a domain beyond the training set, such as unbounded time intervals.

Appendix

A. Proof of Theorem 1

It is clear that the linear combination of polynomials (14) is still a polynomial. It is also immediate to check that (19) is an inner product and (20) is a norm. The non trivial point to prove is that the operators M_m, that by definition form a linear space of H-S linear operators $\mathcal{H}_x^{\boxed{m}} \to \mathcal{H}_y$, have a H-S norm which does not depend on the choice of an orthonormal basis $\{\phi_k\}$ of $\mathcal{H}_x$. In other words, we need to prove that $\|M_i\|^2_{\text{H.S.}}$ does not depend on $\{\phi_k\}$. This can be done similarly to the linear case (see for example [69], Lemma 3.4.2) thanks to property (3) in Lemma 1. In fact, given two multi-indexes K_i and L_i and denoting $\Phi_{K_i} = \phi_{k_1} \boxtimes \ldots \boxtimes \phi_{k_i}$, $\Phi_{L_i} = \phi_{l_1} \boxtimes \ldots \boxtimes \phi_{l_i}$, (3) implies that

$$[\Phi_{K_i}, \Phi_{L_i}]_{\mathcal{H}_x^{\boxed{i}}} = [\phi_{k_1}, \phi_{l_1}]_{\mathcal{H}_x} \cdot \ldots \cdot [\phi_{k_i}, \phi_{l_i}]_{\mathcal{H}_x} \tag{92}$$

which is 1 if and only if $K_i = L_i$. In plain words, Φ_{K_i} is an orthonormal basis of $\mathcal{H}_x^{\boxed{i}}$ and the standard proof applies. □

B. Proof of Theorem 2

Since the training set is normalized we can assume $\hat{M}_0 = 0$. Let $\Delta_L \in \mathcal{L}_{\text{H.S.}}$ and $\epsilon \in \mathbb{R}$.

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\epsilon} J_N(\lambda, \hat{L} + \epsilon\Delta_L)\Big|_{\epsilon=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\sum_{i=1}^{N}\Big[y_i - \hat{L}x_i - \epsilon\Delta_L x_i, y_i - \hat{L}x_i - \epsilon\Delta_L x_i\Big]_{\mathcal{H}_y}\right. \\
&\left.+\lambda\Big[\hat{M}_1 + \epsilon\Delta_L, \hat{M}_1 + \epsilon\Delta_L\Big]_{\text{H.S.}}\right)_{\epsilon=0} \qquad (93) \\
&= \sum_{i=1}^{N} -2[y_i - \hat{L}x_i, \Delta_L x_i]_{\mathcal{H}_y} + 2\lambda[\hat{M}_1, \Delta_L]_{\text{H.S.}}. \qquad (94)
\end{aligned}$$

From (93) it follows that $\forall \Delta_L \in \mathcal{L}_{\text{H.S.}}$, the functional J_N is continuous with respect to L and convex, in fact

$$\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2}\frac{J_N(\lambda, \hat{M}_1 + \epsilon\Delta_L)}{2}\Big|_{\epsilon=0} = \sum_{i=1}^{N}\|\Delta_L x_i\|^2_{\mathcal{H}_y} + \lambda\|\Delta_L\|^2_{\text{H.S.}} > 0, \tag{95}$$

and from (94) we obtain that the condition:

$$\forall \Delta_L \in \mathcal{L}_{\text{H.S.}}: \quad \sum_{i=1}^{N}[y_i - \hat{M}_1 x_i, \Delta_L x_i]_{\mathcal{H}_y} = \lambda[\hat{M}_1, \Delta_L]_{\text{H.S.}} \tag{96}$$

yields the global minimum of $J_N(\lambda, L)$. Let $\{\phi_k\}$ be an orthonormal basis of $\mathcal{H}_x$. By using (11) in (96) we obtain with simple derivations

$$\sum_{k=1}^{\infty}\left[\sum_{i=1}^{N}[x_i, \phi_k]_{\mathcal{H}_x}(y_i - \hat{M}_1 x_i) - \lambda\hat{M}_1\phi_k, \Delta_L\phi_k\right]_{\mathcal{H}_y} = 0. \tag{97}$$

Since (97) holds $\forall \Delta_L \in \mathcal{L}_{\text{H.S.}}$, the condition becomes

$$\forall k: \quad \hat{M}_1\phi_k = \frac{1}{\lambda}\sum_{i=1}^{N}[x_i, \phi_k]_{\mathcal{H}_x}(y_i - \hat{M}_1 x_i) \tag{98}$$

In order to solve (98) we start by solving for $\hat{M}_1 x_i$. From (8) we obtain

$$\lambda\hat{M}_1 x_i = \sum_{j=1}^{N}[x_i, x_j]_{\mathcal{H}_x} y_j - \sum_{j=1}^{N}[x_i, x_j]_{\mathcal{H}_x}\hat{M}_1 x_j. \tag{99}$$

Let $a_{ij} = [x_i, x_j]_{\mathcal{H}_x}$, $A := \text{col}_{i=1}^{N}\text{row}_{j=1}^{N}(a_{ij}) \in \mathbb{R}^{N\times N}$, $M(\lambda) := (\lambda I_N + A)$. (99) can be rewritten as

$$\begin{aligned}
\lambda\,\text{col}_i(\hat{M}_1 x_i) &= A\,\text{col}_i(y_i) - A\,\text{col}_i(\hat{M}_1 x_i), \\
\text{col}_i(y_i - \hat{M}_1 x_i) &= \lambda M(\lambda)^{-1}\text{col}_i(y_i). \qquad (100)
\end{aligned}$$

Denoting $r_k^\top = \text{row}_{i=1}^{N}[x_i, \phi_k]_{\mathcal{H}_x}$, and , by replacing (100) in (98) we get

$$\forall k: \ \hat{M}_1\phi_k = r_k^\top M(\lambda)^{-1}\text{col}_i(y_i), \tag{101}$$

that provides the sought expression of the best linear estimator. Finally, Assumption 2 together with (101) guarantees that $(\hat{M}_1\phi_k)(t)$ exists $\forall t \in D$ and $\forall k$. □

REFERENCES

[1] L. Ljung, *System identification*. Wiley Online Library, 1999.

[2] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[3] B. Schölkopf and A. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[4] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[5] A. Chiuso and G. Pillonetto, "System identification: A machine learning perspective," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 281–304, 2019.

[6] Z. Lu, J. Sun, and K. Butts, "Multiscale asymmetric orthogonal wavelet kernel for linear programming support vector learning and nonlinear dynamic systems identification," *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 712–724, 2013.

[7] F. Smarra, G. D. Di Girolamo, V. De Iuliis, A. Jain, R. Mangharam, and A. D'Innocenzo, "Data-driven switching modeling for MPC using regression trees and random forests," *Nonlinear Analysis: Hybrid Systems*, vol. 36, p. 100882, 2020.

[8] V. De Iuliis, F. Smarra, C. Manes, and A. D'Innocenzo, "Stability analysis of switched ARX models and application to learning with guarantees," *Nonlinear Analysis: Hybrid Systems*, vol. 46, p. 101250, 2022.

[9] J. Drgoňa, D. Picard, M. Kvasnica, and L. Helsen, "Approximate model predictive building control via machine learning," *Applied Energy*, vol. 218, pp. 199–216, 2018.

[10] V. De Iuliis, G. D. Di Girolamo, F. Smarra, and A. D'Innocenzo, "A comparison of classical identification and learning-based techniques for cyber-physical systems," in *2021 29th Mediterranean Conference on Control and Automation (MED)*. IEEE, 2021, pp. 179–185.

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

[12] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin (New Series) of the American Mathematical Society*, 2001.

[13] M. Bisiacco and G. Pillonetto, "On the mathematical foundations of stable RKHSs," *Automatica*, vol. 118, p. 109038, 2020.

[14] F. Dinuzzo, "Kernels for linear time invariant system identification," *SIAM Journal on Control and Optimization*, vol. 53, no. 5, pp. 3299–3317, 2015.

[15] S. Saitoh, "Theory of reproducing kernels and its applications," *Longman Scientific & Technical*, 1988.

[16] ——, "Best approximation, Tikhonov regularization and reproducing kernels," *Kodai Mathematical Journal*, vol. 28, no. 2, pp. 359–367, 2005.

[17] F. Carli, T. Chen, and L. Ljung, "Maximum entropy kernels for system identification," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1471–1477, 2016.

[18] T. Chen, "Continuous-time DC kernel—a stable generalized first-order spline kernel," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4442–4447, 2018.

[19] G. De Nicolao and G. Ferrari Trecate, "Consistent identification of NARX models via regularization networks," *IEEE Transactions on Automatic Control*, vol. 44, no. 11, pp. 2045–2049, 1999.

[20] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in computational mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

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

[22] E. Maddalena, P. Scharnhorst, and C. Jones, "Deterministic error bounds for kernel-based learning techniques under bounded noise," *Automatica*, vol. 134, p. 109896, 2021.

[23] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[24] G. Ramaswamy, K.R.and Bottegal and P. Van den Hof, "Learning linear modules in a dynamic network using regularized kernel-based methods," *Automatica*, vol. 129, p. 109591, 2021.

[25] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *International conference on computational learning theory*. Springer, 2001, pp. 416–426.

[26] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[27] 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.

[28] M. Schürch, D. Azzimonti, A. Benavoli, and M. Zaffalon, "Recursive estimation for sparse Gaussian process regression," *Automatica*, vol. 120, p. 109127, 2020.

[29] Y.-F. Li, L.-J. Li, H.-Y. Su, and J. Chu, "Least squares support vector machine based partially linear model identification," *Lecture notes in computer science*, pp. 775–781, 2006.

[30] J. Suykens, T. Van Gestel, and J. De Brabanter, *Least squares support vector machines*. World scientific, 2002.

[31] A. Argyriou, C. Micchelli, and M. Pontil, "When is there a representer theorem? Vector versus matrix regularizers," *The Journal of Machine Learning Research*, vol. 10, pp. 2507–2529, 2009.

[32] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: a necessary and sufficient condition," *Advances in neural information processing systems*, vol. 25, pp. 189–196, 2012.

[33] M. S. Gockenbach, *Linear inverse problems and Tikhonov regularization*. American Mathematical Soc., 2016, vol. 32.

[34] P. Hansen, "The truncatedSVD as a method for regularization," *BIT Numerical Mathematics*, vol. 27, no. 4, pp. 534–553, 1987.

[35] A. Tikhonov and V. Arsenin, *Solutions of ill-posed problems*. John Wiley, 1977.

[36] C. Wang and D.-X. Zhou, "Optimal learning rates for least squares regularized regression with unbounded sampling," *Journal of Complexity*, vol. 27, no. 1, pp. 55–67, 2011.

[37] R. Courant and D. Hilbert, *Methods of mathematical physics*. Interscience Publishers, New York, 1966, vol. 2.

[38] S. Saitoh, "Hilbert spaces induced by Hilbert space valued functions," *Proceedings of the American Mathematical Society*, vol. 89, no. 1, pp. 74–78, 1983.

[39] C. Carmeli, E. De Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Analysis and Applications*, vol. 4, no. 04, pp. 377–408, 2006.

[40] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá, "Vector valued reproducing kernel Hilbert spaces and universality," *Analysis and Applications*, vol. 8, no. 01, pp. 19–61, 2010.

[41] A. Melkman and C. Micchelli, "Optimal estimation of linear operators in Hilbert spaces from inaccurate data," *SIAM Journal on Numerical Analysis*, vol. 16, no. 1, pp. 87–105, 1979.

[42] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Prediction error identification of linear systems: a nonparametric gaussian regression approach," *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.

[43] M. Yuan and T. Cai, "A reproducing kernel Hilbert space approach to functional linear regression," *The Annals of Statistics*, vol. 38, no. 6, pp. 3412–3444, 2010.

[44] M. Lindfors and T. Chen, "Regularized LTI system identification in the presence of outliers: A variational EM approach," *Automatica*, vol. 121, p. 109152, 2020.

[45] G. Birpoutsoukis, A. Marconato, J. Lataire, and J. Schoukens, "Regularized nonparametric Volterra kernel estimation," *Automatica*, vol. 82, pp. 324–327, 2017.

[46] A. Dalla Libera, R. Carli, and G. Pillonetto, "A novel multiplicative polynomial kernel for Volterra series identification," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 316–321, 2020.

[47] T. Dodd and R. Harrison, "A new solution to Volterra series estimation," *IFAC Proceedings Volumes*, vol. 35, no. 1, pp. 67–72, 2002.

[48] G. Torres Mendonça, J. Pongratz, and C. Reick, "Identification of linear response functions from arbitrary perturbation experiments in the presence of noise–part 1: Method development and toy model demonstration," *Nonlinear Processes in Geophysics*, vol. 28, no. 4, pp. 501–532, 2021.

[49] M. Franz and B. Schölkopf, "A unifying view of Wiener and Volterra theory and polynomial kernel regression," *Neural computation*, vol. 18, no. 12, pp. 3097–3118, 2006.

[50] T. Poggio, "On optimal nonlinear associative recall," *Biological Cybernetics*, vol. 19, no. 4, pp. 201–209, 1975.

[51] J. Li, X. Li, H.-T. Zhang, G. Chen, and Y. Yuan, "Data-driven discovery of block-oriented nonlinear models using sparse null-subspace methods," *IEEE Transactions on Cybernetics*, 2020.

[52] F. Giri and E.-W. Bai, *Block-oriented nonlinear system identification*. Springer, 2010, vol. 1.

[53] F. Cacace and A. Germani, "Learning causal estimates of linear operators from noisy data," *IEEE Trans. on Automatic Control*, 2022.

[54] M. Chaika and S. Perlman, "A Weierstrass theorem for a complex separable Hilbert space," *Journal of Approximation Theory*, vol. 15, no. 1, pp. 18–22, 1975.

[55] W. Porter, T. Clark, and R. De Santis, "Causality structure and the Weierstrass theorem," *Journal of Mathematical Analysis and Applications*, vol. 52, no. 2, pp. 351–363, 1975.

[56] P. Prenter, "A Weierstrass theorem for real, separable Hilbert spaces," *Journal of Approximation Theory*, vol. 3, no. 4, pp. 341–351, 1970.

[57] W. Root, "On the modeling of systems for identification. I: ε-representations of classes of systems," *SIAM Journal on Control*, vol. 13, no. 4, pp. 927–944, 1975.
[58] V. Sazonov, "A remark on characteristic functionals," *Theory of Probability & Its Applications*, vol. 3, no. 2, pp. 188–192, 1958.
[59] A. Bertuzzi, A. Gandolfi, and A. Germani, "A Weierstrass-like theorem for real separable Hilbert spaces," *Journal of Approximation Theory*, vol. 32, no. 1, pp. 76–81, 1981.
[60] A. De Santis, A. Gandolfi, A. Germani, and P. Tardelli, "Polynomial approximation for a class of physical random variables," *Proceedings of the American Mathematical Society*, vol. 120, no. 1, pp. 261–266, 1994.
[61] C. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural computation*, vol. 17, no. 1, pp. 177–204, 2005.
[62] T. Chen, "On kernel design for regularized LTI system identification," *Automatica*, vol. 90, pp. 109–122, 2018.
[63] G. De Nicolao, G. Sparacino, and C. Cobelli, "Nonparametric input estimation in physiological systems: problems, methods, and case studies," *Automatica*, vol. 33, no. 5, pp. 851–870, 1997.
[64] G. Wahba, "Bayesian "confidence intervals" for the cross-validated smoothing spline," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 45, no. 1, pp. 133–150, 1983.
[65] J. Berger, *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.
[66] C. Mallows, "More comments on Cp," *Technometrics*, vol. 37, no. 4, pp. 362–372, 1995.
[67] K. Eggensperger *et al.*, "Towards an empirical foundation for assessing Bayesian optimization of hyperparameters," in *NIPS workshop on Bayesian Optimization in Theory and Practice*, vol. 10, 2013.
[68] G. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.
[69] A. Balakrishnan, *Applied Functional Analysis*. Springer Science & Business Media, 2012, vol. 3.

Filippo Cacace received the Laurea degree in electronic engineering in 1988, and the PhD degree in computer science in 1992 both at Politecnico di Milano. In 2003 he joined University Campus Bio-Medico of Rome, where he is currently an associate professor. His current research interests include nonlinear systems and observers, stochastic and delay systems, system identification, and applications to systems biology.

Vittorio De Iuliis received the M.Sc. degree in Computer and Systems Engineering in 2014 and the Ph.D. degree in Information and Communication Technology in 2018, both from the University of L'Aquila, Italy, where he currently serves as Assistant Professor. From February 2021 to July 2021 he has been visiting researcher at the ICTEAM Institute at the Université Catholique de Louvain, Belgium. His research interests include delay systems, positive systems, filtering and system identification, machine learning methods for system identification.

Alfredo Germani received the Laurea degree in physics and the postdoctoral degree in computer and system engineering from University of Rome La Sapienza, Italy, in 1972 and 1974, respectively. Since 1987, he has been full professor of system theory at University of L'Aquila, Italy. He has published more than 160 research papers in the fields of systems theory, systems identification, and data analysis; nonlinear, stochastic, and optimal control theory; distributed and delay systems; finitely additive white noise theory; approximation theory; optimal polynomial filtering; image processing and restoring; and mathematical modeling for biological processes.

Mario Merone, Ph.D, is a Assistant Professor at the University Campus Bio-Medico di Roma. He received the Ph.D in Bioengineering and Bioscience in 2017. He has extensive experience in the field of artificial intelligence (Machine Learning and Deep Learning for Time-series Analysis). His research has led to 40 publications, all focused on AI for healthcare. He is currently the scientific leader of several projects involving the application of artificial intelligence in the medical field. He is a founding partner of an innovative start-up called BPCOMEDIA S.R.L., focused on technology transfer of research results. Mario Merone is a member of the IEEE, IEEE Computer Society, IEEE Computational Life Sciences and IEEE Sensors Council.