Automatica 93 (2018) 321–332

Contents lists available at ScienceDirect

Automatica

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

System identification using kernel-based regularization: New insights on stability and consistency issues☆

Gianluigi Pillonetto

Department of Information Engineering, University of Padova, Padova, Italy

ARTICLE INFO

Article history:
Received 28 December 2016
Received in revised form 6 February 2018
Accepted 26 February 2018

Keywords:
Learning from examples
System identification
Reproducing kernel Hilbert spaces of dynamic systems
Kernel-based regularization
BIBO stability
Regularization networks
Generalization and consistency

ABSTRACT

Learning from examples is one of the key problems in science and engineering. It deals with function reconstruction from a finite set of direct and noisy samples. *Regularization in reproducing kernel Hilbert spaces* (RKHSs) is widely used to solve this task and includes powerful estimators such as regularization networks. Recent achievements include the proof of the statistical consistency of these kernel-based approaches. Parallel to this, many different system identification techniques have been developed but the interaction with machine learning does not appear so strong yet. One reason is that the RKHSs usually employed in machine learning do not embed the information available on dynamic systems, e.g. BIBO stability. In addition, in system identification the independent data assumptions routinely adopted in machine learning are never satisfied in practice. This paper provides some new results which strengthen the connection between system identification and machine learning. Our starting point is the introduction of RKHSs of dynamic systems. They contain functionals over spaces defined by system inputs and allow to interpret system identification as learning from examples. In both linear and nonlinear settings, it is shown that this perspective permits to derive in a relatively simple way conditions on RKHS stability (i.e. the property of containing only BIBO stable systems or predictors), also facilitating the design of new kernels for system identification. Furthermore, we prove the convergence of the regularized estimator to the optimal predictor under conditions typical of dynamic systems.

1. Introduction

Learning from examples is key in science and engineering, considered at the core of intelligence's understanding (Poggio & Shelton, 1999). In mathematical terms, it can be described as follows. We are given a finite set of training data (x_i, y_i), where x_i is the so called input location while y_i is the corresponding output measurement. The goal is then the reconstruction of a function with good prediction capability on future data: for a new pair (x, y), the prediction $g(x)$ should be close to y.

To solve this task, nonparametric techniques have been extensively used in the last years. Within this paradigm, instead of assigning to the unknown function a specific parametric structure, g is searched over a possibly infinite-dimensional functional space. The modern approach uses Tikhonov regularization theory (Bertero, 1989; Tikhonov & Arsenin, 1977) in conjunction with Reproducing Kernel Hilbert Spaces (RKHSs) (Aronszajn, 1950; Bergman, 1950). RKHSs possess many important properties, being in one to one correspondence with the class of positive definite kernels. Their connection with Gaussian processes is also described in Aravkin, Bell, Burke, and Pillonetto (2015), Bell and Pillonetto (2004), Kimeldorf and Wahba (1970) and Lukic and Beder (2001).

While applications of RKHSs in statistics, approximation theory and computer vision trace back to Bertero, Poggio, and Torre (1988), Poggio and Girosi (1990) and Wahba (1990), these spaces were introduced to the machine learning community in Girosi (1997). RKHSs permit to treat in a unified way many different regularization methods. The so called kernel-based methods (Evgeniou, Pontil, & Poggio, 2000; Schölkopf & Smola, 2001) include smoothing splines (Wahba, 1990), regularization networks (Poggio & Girosi, 1990), Gaussian regression (Rasmussen & Williams, 2006), and support vector machines (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; Vapnik, 1998). In particular, a regularization network (RN) has the structure

$$\hat{g} = \arg\min_{f \in \mathscr{H}} \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{N} + \gamma \|f\|^2_{\mathscr{H}} \qquad \text{RN} \tag{1}$$

where $\mathscr{H}$ denotes a RKHS with norm $\|\cdot\|_{\mathscr{H}}$.

☆ This research has been partially supported by the Progetto di Ateneo CPDA147754/14-New statistical learning approach for multi-agents adaptive estimation and coverage control. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström. The author would like to thank Isabella Vendramini for helpful discussions on the subject.

E-mail address: giapi@dei.unipd.it.

https://doi.org/10.1016/j.automatica.2018.03.065
0005-1098/

Thus, the function estimate minimizes an objective sum of two contrasting terms: the quadratic loss which measures the adherence to experimental data and the regularizer (the RKHS squared norm) which restores the well-posedness. Finally, the positive scalar γ is the regularization parameter which has to suitably trade off these two components.

The use of (1) has significant advantages. The choice of an appropriate RKHS, often obtained just including function smoothness information (Schölkopf & Smola, 2001), and a careful tuning of γ, e.g. by the empirical Bayes approach (Aravkin, Burke, Chiuso, & Pillonetto, 2012, 2014; Maritz & Lwin, 1989), can well balance bias and variance. One can thus obtain favorable mean squared error properties. Furthermore, even if $\mathscr{H}$ is infinite-dimensional, the solution $\hat{g}$ is always unique, belongs to a finite-dimensional subspace and is available in closed-form. This result comes from the representer theorem (Argyriou & Dinuzzo, 2014; Argyriou, Micchelli, & Pontil, 2009; Kimeldorf & Wahba, 1970; Schölkopf, Herbrich, & Smola, 2001). Building upon the work (Wahba, 1977), many new results have been also recently obtained on the statistical consistency of (1). In particular, the property of $\hat{g}$ to converge to the optimal predictor as the data set size grows to infinity is discussed e.g. in Mukherjee, Niyogi, Poggio, and Rifkin (2006), Poggio, Rifkin, Mukherjee, and Niyogi (2004), Smale and Zhou (2007), Wu, Ying, and Zhou (2006) and Yuan and Tony Cai (2010). Parallel to this, many system identification techniques have been developed in the last decades. In linear contexts, the first regularized approaches trace back to Akaike (1979), Kitagawa and Gersch (1996) and Schiller (1979), see also Goodwin, Gevers, and Ninness (1992) and Ljung, Goodwin, and Agero (2014) where model error is described via a nonparametric structure. More recent approaches, also inspired by nuclear and atomic norms (Chandrasekaran, Recht, Parrilo, & Willsky, 2012), can instead be found in Grossmann, Jones, and Morari (2009), Liu and Vandenberghe (2009), Mohan and Fazel (2010), Pillonetto, Chen, Chiuso, De Nicolao, and Ljung (2016) and Rojas, Toth, and Hjalmarsson (2014). In the last years, many nonparametric techniques have been proposed also for nonlinear system identification. They exploit e.g. neural networks (Lin, Horne, Tino, & Giles, 1996; Shun-Feng & Yang, 2002), Volterra theory (Franz & Schölkopf, 2006), kernel-type estimators (Leithead, Solak, & Leith, 2003; Pillonetto, Chiuso, & Quang, 2011b; Zhao, Chen, Bai, & Li, 2015) which include also weights optimization to control the mean squared error (Bai & Liu, 2007; Bai, 2010; Roll, Nazin, & Ljung, 2005). Important connections between kernel-based regularization and nonlinear system identification have been also obtained by the least squares support vector machines (Suykens, Alzate, & Pelckmans, 2010; Suykens, Van Gestel, Brabanter, De Moor, & Vandewalle, 2002) and using Gaussian regression for state space models (Frigola, Lindsten, Schon, & Rasmussen, 2013; Frigola & Rasmussen, 2013). Most of these approaches are inspired by machine learning, a fact not surprising since predictor estimation is at the core of the machine learning philosophy. Indeed, a black-box relationship can be obtained through (1) using past inputs and outputs to define the input locations (regressors). However, the kernels currently used for system identification are those conceived by the machine learning community for the reconstruction of *static* maps. RKHSs suited to linear system identification have been proposed only recently, e.g. in computer vision exploiting compound matrices built from system trajectories (Vishwanathan, Smola, & Vidal, 2007) or by introducing *stable spline kernels* which embed information on impulse response regularity and stability (Chen, Ohlsson, & Ljung, 2012; Pillonetto, Chiuso, & De Nicolao, 2011a; Pillonetto & De Nicolao, 2010). Furthermore, while stability of a RKHS (i.e. its property of containing only stable systems or predictors) is treated in Carmeli, De Vito, and Toigo (2006), Dinuzzo (2015) and Pillonetto, Dinuzzo, Chen, Nicolao, and Ljung (2014), the nonlinear scenario still appears unexplored. Beyond stability, we also notice that the most used kernels for nonlinear regression, e.g. the Gaussian (Schölkopf & Smola, 2001), do not include other important information on dynamic systems like the fact that output energy is expected to increase if input energy augments.

Another aspect that weakens the interaction between system identification and machine learning stems also from the (apparently) different contexts these disciplines are applied to. In machine learning one typically assumes that data (x_i, y_i) are i.i.d. random vectors assuming values on a bounded subset of the Euclidean space. But in system identification, even when the system input is white noise, the input locations are not mutually independent. Already in the classical Gaussian noise setting, the outputs are not even bounded, i.e. there is no compact set containing them with probability one. Remarkably, this implies that none of the aforementioned consistency results developed for kernel-based methods can be applied. Some extensions to the case of correlated samples can be found in Guo and Zhou (2013), Steinwart, Hush, and Scovel (2009), Vidyasagar (1997) and Wang and Zhou (2011) but still under conditions far from the system identification setting.

In this paper we provide some new insights on the interplay between system identification and machine learning in a RKHS setting. Our starting point is the introduction of what we call *RKHSs of dynamic systems* which contain functionals over input spaces $\mathscr{X}$ induced by system inputs u. More specifically, each input location $x \in \mathscr{X}$ contains a piece of the trajectory of u so that any $g \in \mathscr{H}$ can be associated to a dynamic system. When u is a stationary stochastic process, its distribution then defines the probability measure on $\mathscr{X}$ from which the input locations are drawn. Again, we stress that this framework has been (at least implicitly) used in previous works on nonlinear system identification, see e.g. Lin et al. (1996), Pillonetto, Chiuso, Quang, et al. (2011b), Sjöberg et al. (1995), Shun-Feng and Yang (2002) and Suykens et al. (2002). However, it has never been cast and studied in its full generality under a RKHS setting.

Even if in this context the input space can turn out unbounded (e.g. when the system input is Gaussian) and complex (e.g. $\mathscr{X}$ is a function space itself in continuous-time), it will be shown that our perspective is key to obtain the following achievements:

- linear and nonlinear system identification can be treated in a unified way in both discrete- and continuous-time. Thus, the estimator (1) can be used in many different contexts, relevant for the control community, just changing the RKHS. This is important for the development of a general theory which links regularization in RKHS and system identification;
- system input's role in determining the nature of the RKHS is made explicit. This will be also described in more detail in the linear system context, illustrating the distinction between the concept of RKHSs $\mathscr{H}$ of dynamic systems and that of RKHSs $\mathscr{I}$ of impulse responses;
- for linear systems we provide a new and simple derivation of the necessary and sufficient condition for RKHS stability (Carmeli et al., 2006; Dinuzzo, 2015; Pillonetto et al., 2014) that relies just on basic RKHS theory;
- in the nonlinear scenario, we obtain a sufficient condition for RKHS stability which has wide applicability. We also derive a new stable kernel for nonlinear system identification;
- consistency of the RN (1) is proved under assumptions suited to system identification, revealing the link between consistency and RKHS stability.

The paper is organized as follows. In Section 2 we provide a brief overview on RKHSs. In Section 3, the concept of RKHSs of dynamic systems is defined by introducing input spaces $\mathscr{X}$ induced by system inputs. The case of linear dynamic systems is then detailed via its relationship with linear kernels. The difference between the

RKHSs of dynamic systems and the RKHSs of impulse responses is also elucidated. Section 4 discusses the concept of stable RKHS. We provide a new simple characterization of RKHS stability in the linear setting, then giving a sufficient condition for the nonlinear scenario. We also introduce a new kernel for nonlinear system identification, testing its effectiveness on a benchmark problem. In Section 5, the consistency of the RN (1) is proved in the general framework of RKHSs of dynamic systems. Conclusions end the paper while proofs of the consistency results are gathered in the Appendix.

In what follows, the analysis is always restricted to causal systems and, to simplify the exposition, the input locations contain only past inputs so that output error models are considered. If an autoregressive part is included, the consistency analysis in Section 5 remains unchanged while the conditions developed in Section 4 guarantee predictor (in place of system) stability.

2. Brief overview on RKHSs

We use $\mathscr{X}$ to indicate a function domain, often referred to as the *input space* in machine learning. Its generic element is the *input location*, denoted by x or a in the sequel. All the functions are assumed real valued, so that $g : \mathscr{X} \to \mathbb{R}$.

Our goal is to estimate maps to make predictions over the whole $\mathscr{X}$. Thus, a basic requirement is to use an hypothesis space $\mathscr{H}$ with functions well defined pointwise for any $x \in \mathscr{X}$. In particular, assume that all the pointwise evaluators $g \to g(x)$ are linear and bounded over $\mathscr{H}$, i.e. $\forall x \in \mathscr{X}$ there exists $C_x < \infty$ such that

$$|g(x)| \leq C_x \|g\|_{\mathscr{H}}, \quad \forall g \in \mathscr{H}. \tag{2}$$

Definition 1 (*RKHS*)**.** A reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ over $\mathscr{X}$ is a Hilbert space containing functions $g : \mathscr{X} \to \mathbb{R}$ where (2) holds.

RKHSs are connected to the concept of positive definite kernel, a particular function defined over $\mathscr{X} \times \mathscr{X}$.

Definition 2 (*Positive Definite Kernel and Kernel Section*)**.** A symmetric function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called *positive definite kernel* if, for any integer p, it holds

$$\sum_{i=1}^{p}\sum_{j=1}^{p} c_i c_j \mathscr{K}(x_i, x_j) \geq 0, \quad \forall (x_k, c_k) \in (\mathscr{X}, \mathbb{R}), \quad k = 1, \ldots, p.$$

The *kernel section* $\mathscr{K}_x$ centered at x is the function from $\mathscr{X}$ to $\mathbb{R}$ defined by $\mathscr{K}_x(a) = \mathscr{K}(a, x) \quad \forall a \in \mathscr{X}$.

The following theorem provides the one-to-one correspondence between RKHSs and positive definite kernels (Aronszajn, 1950).

Theorem 3 (*Moore–Aronszajn and Reproducing Property*)**.** *To every RKHS $\mathscr{H}$ there corresponds a unique positive definite kernel $\mathscr{K}$ such that the so called* reproducing property *holds, i.e.*

$$\langle \mathscr{K}_x, g \rangle = g(x) \quad \forall (x, g) \in (\mathscr{X}, \mathscr{H}). \tag{3}$$

Conversely, given a positive definite kernel $\mathscr{K}$, there exists a unique RKHS of real-valued functions defined over $\mathscr{X}$ where (3) holds.

Theorem 3 shows that a RKHS $\mathscr{H}$ is completely defined by a kernel $\mathscr{K}$, also called the *reproducing kernel* of $\mathscr{H}$. More specifically, it can be proved that the RKHS contains all the finite linear combinations of kernel sections along with some particular infinite sums (Pillonetto et al., 2014, Part II). A consequence of this is that any $g \in \mathscr{H}$ inherits kernel properties, e.g. continuity of $\mathscr{K}$ implies that all the $g \in \mathscr{H}$ are continuous (Cucker & Smale, 2001, p. 35).

The kernel sections give also the closed-form solution of the RN (1), as illustrated in the famous representer theorem.

Theorem 4 (*Representer Theorem*)**.** *The solution of (1) is*

$$\hat{g} = \sum_{i=1}^{N} \hat{c}_i \mathscr{K}_{x_i}, \tag{4}$$

where the scalars $\hat{c}_i$ are the components of the vector

$$\hat{c} = (\mathbf{K} + \gamma N I_N)^{-1} Y, \tag{5}$$

Y is the column vector with ith element y_i, I_N is the $N \times N$ identity matrix and the (i, j) entry of $\mathbf{K}$ is $\mathscr{K}(x_i, x_j)$.

Another RKHS characterization that we will exploit later on is obtained when the kernel can be diagonalized as follows

$$\mathscr{K}(a, x) = \sum_{i=1}^{\infty} \zeta_i \rho_i(a) \rho_i(x), \quad \zeta_i > 0 \; \forall i. \tag{6}$$

The following result then holds (Evgeniou et al., 2000, p. 15) or (Cucker & Smale, 2001, p. 36).

Theorem 5 (*Spectral Representation of a RKHS*)**.** *Let (6) hold and assume that the ρ_i form a set of linearly independent functions on $\mathscr{X}$. Then, one has*

$$\mathscr{H} = \left\{ g \mid g(x) = \sum_{i=1}^{\infty} c_i \rho_i(x) \;\; s.t. \;\; \sum_{i=1}^{\infty} \frac{c_i^2}{\zeta_i} < \infty \right\}, \tag{7}$$

and

$$\langle f, g \rangle_{\mathscr{H}} = \sum_{i=1}^{\infty} \frac{b_i c_i}{\zeta_i}, \quad \|f\|_{\mathscr{H}}^2 = \sum_{i=1}^{\infty} \frac{b_i^2}{\zeta_i}, \tag{8}$$

where $f = \sum_{i=1}^{\infty} b_i \rho_i$ and $g = \sum_{i=1}^{\infty} c_i \rho_i$.

The expansion (6) can e.g. be obtained by the Mercer theorem (Hochstadt, 1973; Mercer, 1909). In particular, let μ_x be a nondegenerate σ-finite measure on $\mathscr{X}$. Then, under somewhat general conditions (Sun, 2005), the ρ_i and ζ_i in (6) can be set to the eigenfunctions and eigenvalues of the integral operator induced by $\mathscr{K}$, i.e.

$$\int_{\mathscr{X}} \mathscr{K}(\cdot, x) \rho_i(x) d\mu_x(x) = \zeta_i \rho_i(\cdot), \quad 0 < \zeta_1 \leq \zeta_2 \leq \ldots. \tag{9}$$

Such ρ_i form a complete orthonormal basis in the Lebesgue space $\mathscr{L}_2^{\mu_x}$ of functions square integrable under μ_x.[1]

3. RKHSs of dynamic systems and the linear scenario

3.1. RKHSs of dynamic systems

The *RKHSs of dynamic systems* rely on simple constructions of input spaces $\mathscr{X}$ induced by system inputs u.

Discrete-time. First, the discrete-time setting is considered. Assume we are given a system input $u : \mathbb{Z} \to \mathbb{R}$. Then, we think of any input location in $\mathscr{X}$ indexed by the time $t \in \mathbb{Z}$. Different cases arise depending on the postulated system model. For example, one can have

$$x_t = [u_t \;\; u_{t-1} \;\; \ldots \;\; u_{t-m+1}]^T, \tag{10}$$

where m is the system memory. This construction is connected to FIR or NFIR models and makes $\mathscr{X}$ a subset of the classical Euclidean space $\mathbb{R}^m$.

Another scenario is

$$x_t = [u_t \;\; u_{t-1} \;\; u_{t-2} \;\; \ldots]^T, \tag{11}$$

[1] Thus, the representation (7) is not unique since spectral maps are not unique. Eigendecompositions depend on the measure μ_x but lead to the same RKHS.

where any input location is a sequence (an infinite-dimensional column vector) and the input space $\mathscr{X}$ becomes a subset of $\mathbb{R}^\infty$. The definition (11) is related to infinite memory systems, e.g. IIR models in linear settings.

Continuous-time. The continuous-time input is the map $u : \mathbb{R} \to \mathbb{R}$. In this case, the input location x_t becomes the function $x_t : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$x_t(\tau) = u(t-\tau), \quad \tau \geq 0, \tag{12}$$

i.e. x_t contains the input's past up to the instant t. In many circumstances, one can assume $\mathscr{X} \subset \mathscr{P}^c$, where $\mathscr{P}^c$ contains piecewise continuous functions on $\mathbb{R}_+$. When the input is causal, and u_t is smooth for $t \geq 0$, the x_t is indeed piecewise continuous.

Note that (12) is the continuous-time counterpart of (11) while that of (10) can be obtained just zeroing part of the input location, i.e.

$$x_t(\tau) = u(t-\tau)\xi_T(\tau), \quad \tau \geq 0, \tag{13}$$

where ξ_T is the indicator function of the interval $[0, T]$. In linear systems, (13) arises when the impulse response support is compact.

RKHSs $\mathscr{H}$ of functions over domains $\mathscr{X}$, induced by system inputs u as illustrated above, are hereby called *RKHSs of dynamic systems*. Thus, if $g \in \mathscr{H}$, the scalar $g(x_t)$ is the noiseless output at t of the system fed with the input trajectory contained in x_t. Note that g in general is a functional: in the cases (11)–(13) the arguments x_t entering $g(\cdot)$ are infinite-dimensional objects.

3.2. The linear system scenario

RKHSs of linear dynamic systems are now introduced also discussing the structure of the resulting RN.

Linear system identification was faced in Pillonetto and De Nicolao (2010) and Pillonetto et al. (2014, Part III) by introducing *RKHSs of impulse responses*. These are spaces $\mathscr{I}$ induced by kernels K defined over subsets of $\mathbb{R}_+ \times \mathbb{R}_+$. They thus contain causal functions, each of them representing an impulse response θ. The RN which returns the impulse response estimate was

$$\hat{\theta} = \arg\min_{\theta \in \mathscr{I}} \sum_{i=1}^{N} \frac{\left(y_i - (\theta \otimes u)_{t_i}\right)^2}{N} + \gamma \|\theta\|^2_{\mathscr{I}}, \tag{14}$$

where $(\theta \otimes u)_{t_i}$ is the convolution between the impulse response and the input evaluated at t_i.

The *RKHSs of linear dynamic systems* here introduced are instead associated to (output) linear kernels $\mathscr{K}$ defined on $\mathscr{X} \times \mathscr{X}$ through convolutions of K with system inputs. In particular, if x_t and x_τ are as in (10)–(13), one has[2]

$$\mathscr{K}(x_t, x_\tau) = (u \otimes (K \otimes u)_\tau)_t$$

which e.g. in continuous-time becomes

$$\mathscr{K}(x_t, x_\tau) = \int_0^{+\infty} u(t-\alpha) \left(\int_0^{+\infty} u(\tau - \beta) K(\alpha, \beta) d\beta \right) d\alpha. \tag{15}$$

These kernels lead to the RN (1) which corresponds to (14) after the "reparametrization" $g(x_t) = (\theta \otimes u)_t$ so that, in place of the impulse response θ, the optimization variable becomes the functional $g(\cdot)$.

The kernels $\mathscr{K}$ arising in discrete- and continuous-time are described below in (16) and (19). The distinction between the RKHSs induced by K and $\mathscr{K}$ will be further discussed in Section 3.3.

IIR models. Consider the input locations defined by (11). The input space contains sequences and $\mathscr{X} \subseteq \mathbb{R}^\infty$. Interpreting any input location as an infinite-dimensional column, we can use ordinary algebra's notation to handle infinite-dimensional objects. For example, if $(a, x) \in (\mathscr{X}, \mathscr{X})$ then $a^T x = \langle a, x \rangle_2$, where $\langle \cdot, \cdot \rangle_2$ is the inner-product in the classical space ℓ_2 of squared summable sequences.

Let K be symmetric and positive semidefinite infinite-dimensional matrix K (the nature of K is discussed also in Section 3.3). Then, the function

$$\mathscr{K}(x, a) = x^T K a, \quad (x, a) \in \mathbb{R}^\infty \times \mathbb{R}^\infty \tag{16}$$

defines a linear kernel on $\mathscr{X} \times \mathscr{X}$. Following arguments similar to those developed in the FIR case, one can see that the RKHS associated to such $\mathscr{K}$ contains linear functions of the form $g(x) = a^T K x$ with $a \in \mathbb{R}^\infty$. Note that each $g \in \mathscr{H}$ is a functional defined by the sequence $a^T K$ which represents an impulse response. In fact, one can deduce from (11) that $g(x_t)$ is the discrete-time convolution, evaluated at t, between u and $a^T K$.

The RN with $\mathscr{H}$ induced by (16) now implements regularized IIR estimation. The solution is given by the representer theorem (4) and turns out

$$\hat{g}(x) = \sum_{i=1}^{N} \hat{c}_i \mathscr{K}_{x_i}(x) = \hat{\theta}^T x, \tag{17}$$

where $x_i := x_{t_i}$ and the $\hat{c}_i$ are the components of (5). From the nature of the input locations (11) one immediately sees that the infinite-dimensional column vector

$$\hat{\theta} := \sum_{i=1}^{N} \hat{c}_i K x_i \tag{18}$$

contains the impulse response coefficients estimates. Finally, note that the FIR scenario (11) is obtained just thinking of K as a truncated kernel, i.e. $K(i, j) = 0$ for $i > m$ or $j > m$.

Continuous-time. The continuous-time scenario arises considering the input locations defined by (12) or (13). The input space $\mathscr{X}$ now contains causal functions. Considering (12), given a positive-definite kernel $K : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$, the linear kernel $\mathscr{K}$ is

$$\mathscr{K}(x, a) = \int_{\mathbb{R}_+ \times \mathbb{R}_+} K(t, \tau) x(t) a(\tau) dt d\tau \tag{19}$$

which coincides with (15) when $x = x_t$ and $a = x_\tau$. Each kernel section $\mathscr{K}_x(\cdot)$ is a continuous-time linear system with impulse response $\theta(\cdot) = \int_{\mathbb{R}_+} K(\cdot, t) x(t) dt$. Thus, the corresponding RKHS contains linear functionals and (1) now implements regularized system identification in continuous-time. Using the representer theorem, the solution of (1) is

$$\hat{g}(x) = \sum_{i=1}^{N} \hat{c}_i \mathscr{K}_{x_i}(x) = \int_{\mathbb{R}_+} \hat{\theta}(\tau) x(\tau) d\tau \tag{20}$$

where $\hat{c}$ is still defined by (5) while $\hat{\theta}$ is the impulse response estimate given by

$$\hat{\theta}(\tau) := \sum_{i=1}^{N} \hat{c}_i \int_{\mathbb{R}_+} K(\tau, t) x_i(t) dt. \tag{21}$$

3.3. Relationship between RKHSs of impulse responses and RKHSs of dynamic systems

In (16), the infinite-dimensional matrix K represents a kernel over $\mathbb{N} \times \mathbb{N}$. Then, let $\mathscr{I}$ be the corresponding RKHS which contains infinite-dimensional column vectors $\theta = [\theta_1 \; \theta_2 \ldots]^T$. We will now see that $\mathscr{I}$ is the RKHS of impulse responses associated to $\mathscr{H}$, i.e. each $\theta \in \mathscr{I}$ is the impulse response of a linear system $g \in \mathscr{H}$. In

[2] Translated in a stochastic setting, $\mathscr{K}$ can be seen as the covariance of a causal random process of covariance K filtered by u. It was also called output kernel in Pillonetto et al. (2014, Part III).

particular, let K admit the following expansion in terms of linearly independent infinite-dimensional (column) vectors ψ_i:

$$K = \sum_{i=1}^{\infty} \zeta_i \psi_i \psi_i^T.$$

According to Theorem 5, the span of the ψ_i provides all the $\theta \in \mathscr{I}$. Moreover, if $\theta = \sum_{i=1}^{\infty} c_i \psi_i$, then $\|\theta\|_{\mathscr{I}}^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\zeta_i}$. The equality

$$\begin{aligned}\mathscr{K}(a, x) = a^T K x &= a^T \left(\sum_{i=1}^{\infty} \zeta_i \psi_i \psi_i^T \right) x \\ &= \sum_{i=1}^{\infty} \zeta_i \underbrace{\left(a^T \psi_i\right)}_{\rho_i(a)} \underbrace{\left(\psi_i^T x\right)}_{\rho_i(x)},\end{aligned}$$

also provides the expansion of $\mathscr{K}$ in terms of functionals $\rho_i(\cdot)$ defined by $\rho_i(x) := \psi_i^T x$. Assuming that such functionals are linearly independent, it comes from Theorem 5 that each dynamic system $g \in \mathscr{H}$ has the representation $g(\cdot) = \sum_{i=1}^{\infty} c_i \rho_i(\cdot)$. It is now obvious that such system is associated to the impulse response $\theta = \sum_{i=1}^{\infty} c_i \psi_i$ and the two spaces are isometrically isomorphic since

$$\|g\|_{\mathscr{H}}^2 = \sum_{i=1}^{\infty} \frac{c_i^2}{\zeta_i} = \|\theta\|_{\mathscr{I}}^2.$$

This result holds also in continuous-time where $\mathscr{I}$ is now the RKHS associated to the kernel $K : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$. Letting ψ_i be real-valued functions on $\mathbb{R}_+$, this comes from the same arguments adopted in discrete-time but now applied to the expansions $K(t, \tau) = \sum_{i=1}^{\infty} \zeta_i \psi_i(t) \psi_i(\tau)$ and

$$\begin{aligned}\mathscr{K}(x, a) &= \int_{\mathbb{R}_+ \times \mathbb{R}_+} K(t, \tau) x(t) a(\tau) dt d\tau \\ &= \sum_{i=1}^{\infty} \zeta_i \underbrace{\left(\int_{\mathbb{R}_+} \psi_i(t) x(t) dt \right)}_{\rho_i(x)} \underbrace{\left(\int_{\mathbb{R}_+} \psi_i(\tau) a(\tau) d\tau \right)}_{\rho_i(a)}.\end{aligned}$$

Remark 6. The functionals ρ_i could turn out linearly dependent even if the ψ_i composing K are linearly independent. This depends on the nature of the input space. For instance, let $\mathscr{X}$ contain only the input locations induced by $u_t = \sin(\omega t)$. Then, if ψ_i is a rational transfer function with zeros $\pm j\omega$, the functional ρ_i associated to ψ_i vanishes over $\mathscr{X}$. In these cases, there is no isometry between $\mathscr{H}$ and $\mathscr{I}$: the same dynamic system g could be defined by different impulse responses $\theta^i \in \mathscr{I}$. In particular, results on RKHSs induced by sums of kernels reported in Aronszajn (1950, Section 6 on p. 352) allow us to conclude that $\|g\|_{\mathscr{H}} = \min_i \|\theta^i\|_{\mathscr{I}}$. Thus, among all the possible equivalent representations in $\mathscr{I}$ of the dynamic system $g \in \mathscr{H}$, the complexity of g is quantified by that of minimum norm. This is illustrated through the following simple continuous-time example. Assume that

$$K(t, \tau) = \zeta_1 \psi_1(t) \psi_1(\tau) + \zeta_2 \psi_2(t) \psi_2(\tau),$$

where the Laplace transforms of ψ_1 and ψ_2 are given, respectively, by the rational transfer functions

$$W_1(s) = \frac{2s}{s + 1 + \sqrt{2}}, \quad W_2(s) = \frac{s + 1 - \sqrt{2}}{s + 1},$$

which satisfy $W_1(\sqrt{-1}) = W_2(\sqrt{-1})$. Let the input space contain only the input locations induced by $u_t = \sin(t)$. Then, the functionals (ρ_1, ρ_2) associated, respectively, to (ψ_1, ψ_2) coincide over the entire $\mathscr{X}$. Thus, the two impulse responses ψ_1 and ψ_2 induce the same system $g \in \mathscr{H}$. Using Theorem 5, one has $\|\psi_1\|_{\mathscr{I}}^2 = \frac{1}{\zeta_1}$ and $\|\psi_2\|_{\mathscr{I}}^2 = \frac{1}{\zeta_2}$, which implies $\|g\|_{\mathscr{H}}^2 = \min\left(\frac{1}{\zeta_1}, \frac{1}{\zeta_2}\right)$.

4. Stable RKHSs

BIBO stability of a dynamic system is a familiar notion in control. In the RKHS context, we have the following definition.

Definition 7 (*Stable Dynamic System*)**.** Let u be any bounded input, i.e. satisfying $|u_t| < M_u < \infty \, \forall t$. Then, the dynamic system $g \in \mathscr{H}$ is said to be (BIBO) stable if there exists a constant $M_y < \infty$ such that $|g(x_t)| < M_y$ for any t and any input location x_t induced by u.

Note that, for $g \in \mathscr{H}$ to be stable, the above definition implicitly requires the input space of $\mathscr{H}$ to contain any x_t induced by any bounded input.

Definition 8 (*Stable RKHS*)**.** Let $\mathscr{H}$ be a RKHS of dynamic systems induced by the kernel $\mathscr{K}$. Then, $\mathscr{H}$ and $\mathscr{K}$ are said to be stable if each $g \in \mathscr{H}$ is stable.

To derive stability conditions on the kernel, let us first introduce some useful Banach spaces. The first two regard the discrete-time setting:

- the space ℓ_1 of absolutely summable real sequences $a = [a_1 \; a_2 \ldots]$, i.e. such that $\sum_{i=1}^{\infty} |a_i| < \infty$, equipped with the norm
$$\|a\|_1 = \sum_{i=1}^{\infty} |a_i|;$$
- the space ℓ_∞ of bounded real sequences $a = [a_1 \; a_2 \ldots]$, i.e. such that $\sup_i |a_i| < \infty$, equipped with the norm
$$\|a\|_\infty = \sup_i |a_i|.$$

The other two are their continuous-time counterparts, i.e. $\mathscr{L}_1$ which is the space of functions $a : \mathbb{R}_+ \to \mathbb{R}$ equipped with the norm $\|a\|_1 = \int_{\mathbb{R}_+} |a(t)| dt$ and $\mathscr{L}_\infty$ whose norm is $\|a\|_\infty = \inf\{M \text{ s.t. } |a(t)| \le M \text{ almost everywhere}\}$.

4.1. The linear system scenario

We start studying the stability of RKHSs of linear dynamic systems. for the IIR and continuous-time kernels (16) and (19). First it is useful to recall the classical result linking BIBO stability and impulse response summability.

Proposition 9 (*BIBO Stability and Impulse Response Summability*)**.** *Let θ be the impulse response of a linear system. Then, the system is BIBO stable iff $\theta \in \ell_1$ in discrete-time or $\theta \in \mathscr{L}_1$ in continuous-time.*

The next proposition provides the necessary and sufficient condition for RKHS stability in the linear scenario.

Proposition 10 (*RKHS Stability in the Linear Case*)**.** *Let $\mathscr{H}$ be the RKHS of dynamic systems $g : \mathscr{X} \to \mathbb{R}$ induced by the IIR kernel (16). Then, the following statements are equivalent*

(1) *$\mathscr{H}$ is stable;*
(2) *The input space $\mathscr{X}$ contains ℓ_∞ so that*
$$g(a) < \infty \text{ for any } (g, a) \in (\mathscr{H}, \ell_\infty);$$
(3) $\sum_{i=1}^{\infty} \left| \sum_{j=1}^{\infty} K(i, j) a_j \right| < \infty$ *for any* $a \in \ell_\infty$.

Let instead $\mathscr{H}$ be the RKHS induced by the continuous-time kernel (19). The following statements are then equivalent

(1) *$\mathscr{H}$ is stable;*
(2) *The input space $\mathscr{X}$ contains $\mathscr{L}_\infty$ so that*
$$g(a) < \infty \text{ for any } (g, a) \in (\mathscr{H}, \mathscr{L}_\infty);$$
(3) $\int_{\mathbb{R}_+} \left| \int_{\mathbb{R}_+} K(t, \tau) a(\tau) d\tau \right| dt < +\infty$ *for any* $a \in \mathscr{L}_\infty$.

Proof. The proof is developed in discrete-time. The continuous-time case follows exactly by the same arguments with minor modifications.

(1) $\rightarrow$ (2) Recalling Definition 7 and subsequent discussion, this is a direct consequence of the BIBO stability assumption of any $g \in \mathscr{H}$.

(2) $\rightarrow$ (1) Given any $g \in \mathscr{H}$, let $\theta = [\theta_1\ \theta_2\ \ldots]^T$ its associated impulse response and define

$$x_t = [\text{sign}(\theta_1)\ \text{sign}(\theta_2)\ \ldots]^T. \tag{22}$$

The assumption $g(x_t) = \theta^T x_t = \|\theta\|_1 < \infty$ implies that $\theta \in \ell_1$ and the implication follows by Proposition 9.

(2) $\rightarrow$ (3) By assumption, the kernel is well defined over the entire $\ell_\infty \times \ell_\infty$. Hence, any kernel section $\mathscr{K}_a$ centered on $a \in \ell_\infty$ is a well defined element in $\mathscr{H}$ and corresponds to a dynamic system with associated impulse response $\theta = Ka$. With x_t still defined by (22), one has $\mathscr{K}_a(x_t) = \|\theta\|_1 < \infty$ which implies $Ka \in \ell_1$ and proves (3).

(3) $\rightarrow$ (2) By assumption, any impulse response associated to any kernel section centered on $a \in \ell_\infty$ belongs to ℓ_1. This implies that the kernel $\mathscr{K}$ associated to $\mathscr{H}$ is well defined over the entire $\ell_\infty \times \ell_\infty$. Recalling Definition 1 and Eq. (2), RKHS theory then ensures that any $g \in \mathscr{H}$ is well defined pointwise on ℓ_∞ and $g(a) < \infty\ \forall a \in \ell_\infty$. ■

Point (3) contained in Proposition 10 was also cited in Dinuzzo (2015) and Pillonetto et al. (2014) as a particularization of a quite involved and abstract result reported in Carmeli et al. (2006). The stability proof reported below turns instead out surprisingly simple. The reason is that, with the notation adopted in (16) and (19), the outcomes in Carmeli et al. (2006) were obtained starting from spaces $\mathscr{I}$ of impulse responses induced by K. Our starting point is instead the RKHSs $\mathscr{H}$ of dynamic systems induced by $\mathscr{K}$ (in turn defined by K). This different perspective permits to greatly simplify the analysis: kernel stability can be characterized just combining basic RKHS theory and Proposition 9.

Proposition 10 shows that RKHS stability is implied by the absolute integrability of K, i.e. by

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|K(i,j)| < \infty \ \text{ or } \ \int_{\mathbb{R}_+\times\mathbb{R}_+}|K(t,\tau)|dtd\tau < \infty \tag{23}$$

in discrete- and continuous-time, respectively. The condition (23) is also necessary for nonnegative-valued kernels (Pillonetto et al., 2014, Section 13). Then, considering e.g. the continuous-time setting, the popular Gaussian and Laplacian kernels, which belong to the class of radial basis kernels $K(t,s) = h(|s-t|)$ for $t, s \geq 0$, are all unstable. Stability instead holds for the stable spline kernel which was obtained in Pillonetto, Chiuso, and De Nicolao (2010) by applying an exponential change of coordinates to Green functions (Poggio & Girosi, 1990). It is given by

$$K(t,s) = e^{-\beta\max(t,s)} \quad t, s \geq 0 \tag{24}$$

where $\beta > 0$ is related to the impulse response decay rate.

4.2. The nonlinear system scenario

Let us now consider RKHSs of nonlinear dynamic systems with input locations (10)–(13). A very simple sufficient condition for RKHS stability is reported below.

Proposition 11 (*RKHS Stability in the Nonlinear Case*)**.** *Let $\mathscr{H}$ be a RKHS of dynamic systems induced by the kernel $\mathscr{K}$. Let B^r_∞ denote the closed ball of radius r induced by $\|\cdot\|_\infty$, contained in ℓ_∞ or $\mathbb{R}^m$ in discrete-time, or in $\mathscr{L}_\infty$ in continuous-time. Assume that, for any r, there exists C_r such that*

$$\mathscr{K}(x,x) < C_r < \infty, \quad \forall x \in B^r_\infty.$$

Then, the RKHS $\mathscr{H}$ is stable.

Proof. Let $u \in \ell_\infty$ in discrete-time or $u \in \mathscr{L}_\infty$ in continuous-time. Then, we can find a closed ball B^r_∞ containing, for any t, all the input locations x_t induced by u as defined in (10)–(13). For any $g \in \mathscr{H}$ and $x_t \in B^r_\infty$, from the reproducing property and the Cauchy–Schwarz inequality one obtains

$$\begin{aligned}|g(x_t)| &= |\langle g, \mathscr{K}_{x_t}\rangle_{\mathscr{H}}| \leq \|g\|_{\mathscr{H}}\|\mathscr{K}_{x_t}\|_{\mathscr{H}}\\ &= \|g\|_{\mathscr{H}}\sqrt{\mathscr{K}(x_t,x_t)} \leq \|g\|_{\mathscr{H}}\sqrt{C_r},\end{aligned}$$

hence proving the stability of $\mathscr{H}$. ■

The following result exploits the fact that kernels products (sums) induce RKHSs of functions which are products (sums) of the functions induced by the single kernels (Aronszajn, 1950, p. 353 and 361).

Proposition 12 (*RKHS Stability with Kernels Sum and Product*)**.** *Let $\mathscr{K}_1$ and $\mathscr{K}_2$ be stable kernels. Then, the RKHSs induced by $\mathscr{K}_1 \times \mathscr{K}_2$ and $\mathscr{K}_1 + \mathscr{K}_2$ are both stable.*

The two propositions above allow to easily prove the stability of a very large class of kernels, as discussed in discrete-time in the rest of this section.

Radial basis kernels. First, consider the input locations (10) contained in $\mathscr{X} \subseteq \mathbb{R}^m$. As already mentioned in Section 4.1, radial basis kernels $\mathscr{K}(x,a) = h(|x-a|)$, with $|\cdot|$ now to indicate the Euclidean norm, are widely adopted in machine learning.

An important example is the *Gaussian kernel*

$$\mathscr{K}(x,a) = \exp\left(-\frac{|x-a|^2}{\eta}\right), \quad \eta > 0 \tag{25}$$

whose stability immediately derives from Proposition 11.

More in general, all the radial basis kernels are stable[3] since they are constant along their diagonal ($\mathscr{K}(x,x) = h(0)$).

Despite its stability, some drawbacks affect the use of (25) in system identification. First, the fact that $\mathscr{K}(x,x)$ is constant implies that these models do not include the information that output energy is likely to increase if input energy augments. Second, similarity among input locations is measured without using the information that $u_{t-\tau}$ is expected to have less influence on the prediction of y_t as the positive lag τ augments. Such limitation is also in some sense hidden by the finite-dimensional context. In fact, if the input locations are now defined by (11), i.e. the system memory is infinite, the Gaussian kernel becomes

$$\mathscr{K}(x,a) = \exp\left(-\frac{\|x-a\|_2}{\eta}\right). \tag{26}$$

This model is not reasonable: it is not continuous around the origin of $\mathbb{R}^\infty$ and, out of the diagonal, is null for many input locations.

New kernel for nonlinear system identification. We now show how stable spline kernels, which embed exponential stability of linear systems (Pillonetto & De Nicolao, 2010), can be useful also to define nonlinear models. Specifically, let K_α be a stable spline kernel, e.g. diagonal with $K_\alpha(i,i) = \alpha^i$ or given by $K_\alpha(i,j) = \alpha^{\max(i,j)}$

[3] This statement should not be confused with the result discussed in Section 4.1 in the linear system scenario. There, we have seen that radial basis kernels lead to unstable linear kernels $\mathscr{K}$ when used to define K in the IIR (16) and continuous-time (19) case. Here, the Gaussian and Laplace kernels are instead used to define directly $\mathscr{K}$ in the nonlinear system scenario.

for any integer i and j. Then, define the *nonlinear stable spline* (NSS) kernel as

$$\mathscr{K}(a,x) = a^T K_\alpha x \times \exp\left(-\frac{(a-x)^T K_\alpha (a-x)}{\eta}\right) \quad \text{NSS}, \tag{27}$$

which corresponds to the product between a linear kernel and a modified version of (25). Such kernel defines a new infinite-dimensional RKHS suited for identification of nonlinear output error models. Being no more constant along the diagonal, it embeds the information that output energy augments if input energy increases. In addition, Propositions 11 and 12 ensure that this new feature is obtained preserving BIBO stability. Note also that, letting the dimensionality m of the regression space go to infinity, the difficult selection of the discrete dimension of the regressors x has been eliminated. In fact, input locations similarity is regulated by the hyperparameter α that includes the information that the influence of $u_{t-\tau}$ on y_t goes to zero as the time lag τ increases.

4.3. Numerical experiment

The following nonlinear system is taken from Spinelli, Piroddi, and Lovera (2005):

$$\begin{aligned} f(x_t) = {} & u_t + 0.6u_{t-1} + 0.35\,(u_{t.2} + u_{t-4}) - 0.25u_{t-3}^2 \\ & + 0.2(u_{t-5} + u_{t-6}) + 0.9u_{t-3} + 0.25u_t u_{t-1} + 0.75u_{t-2}^3 \\ & - u_{t-1}u_{t-2} + 0.5\left(u_t^2 + u_t u_{t-2} + u_{t-1}u_{t-3}\right). \end{aligned}$$

Then, consider the identification of the following two systems, called (S1) and (S2):

$$y_t = f(x_t) + e_t \quad \text{(S1)}, \qquad y_t = \sum_{k=1}^{\infty} \theta_k u_{t-k} + f(x_t) + e_t \quad \text{(S2)},$$

where all the u_t and e_t are independent Gaussian noises of variance 4. Note that (S2) contains the sum of a linear time invariant system (details on the impulse response θ are given below) and the nonlinear FIR in (S1). Our aim is to identify the two systems from 1000 input–output pairs (x_t, y_t) via (1). The performance is measured by the percentage fit on a test set of 1000 noiseless system outputs contained in y^{test}, i.e.

$$100\%\left(1 - \frac{|y^{test} - \hat{y}^{test}|}{|y^{test} - \bar{y}^{test}|}\right), \tag{28}$$

where $\bar{y}^{test}$ is the mean of the components of y^{test} while $\hat{y}^{test}$ is the prediction from an estimated model. We will display MATLAB boxplots of the 100 fits achieved by (1) after a Monte Carlo of 100 runs, using different kernels. At any run, independent noises are drawn to form new identification and test data. The impulse response h in (S2) also varies. It is a 10-th order rational transfer function with ℓ_2 norm equal to 10 (this makes similar the contribution to the output variance of the linear and nonlinear system components) and poles inside the complex circle of radius 0.95, randomly generated as detailed in Pillonetto et al. (2016, section 7.4).

First, we use the Gaussian kernel (25) over an m-dimensional input space. Plugged in (1), it defines the hyperparameter vector $[m\ \eta\ \gamma]$. For tuning the Gaussian kernel hyperparameters, the regressor vector dimension m is chosen by an oracle not implementable in practice. Specifically, at any run, for each $m \in \{1, \ldots, 50\}$ the pair (η, γ) is determined via marginal likelihood optimization (De Nicolao, Sparacino, & Cobelli, 1997) using only the identification data. Multiple starting points have been adopted to mitigate the effect of local minima. The oracle has then access to the test set to select, among the 50 couples, that maximizing the prediction fit (28). The two left boxplots in Fig. 1 report the prediction fits achieved by this procedure applied to identify the first (top) and the second (bottom) system. Even if the Gaussian kernel is equipped with the oracle, its performance is satisfactory only in the (S1) scenario while the capability of predicting outputs in the (S2) case is poor during many runs. In place of the marginal likelihood, a cross-validation strategy has been also used for tuning (η, γ), obtaining results similar to those here displayed.

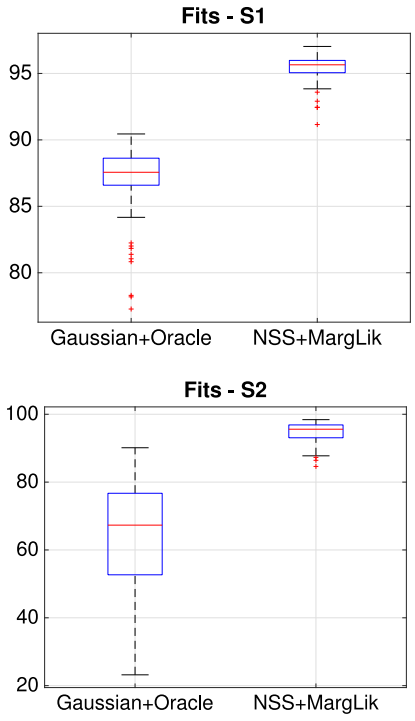

Fig. 1. Boxplots of the 100 test set fits achieved in the scenarios S1 (top panel) and S2 (bottom) by (1) exploiting either the Gaussian kernel (25) with the oracle selecting the regressors space dimension (left boxplots) or the new kernel NSS (27) with hyperparameters tuned via marginal likelihood optimization (right boxplots).

Now, we show that in this example model complexity can be better controlled avoiding the difficult and computationally expensive choice of discrete orders. In particular, we set $m = \infty$ and use the new NSS kernel (27) with diagonal K defined by $K_\alpha(i,i) = \alpha^i$. For tuning the NSS hyperparameter vector (α, η, γ), no oracle having access to the test set is employed but just a single continuous optimization of the marginal likelihood that uses only the identification data. The fits achieved by NSS after the two Monte Carlo studies are in the right boxplots of the two figures above: in both the cases the new estimator behaves very nicely.

5. Consistency of regularization networks for system identification

5.1. The regression function

In what follows, the system input u is a stationary stochastic process over $\mathbb{Z}$ in discrete-time or $\mathbb{R}$ in continuous-time. The distribution of u induces on $\mathscr{X}$ the (Borel nondegenerate) probability measure μ_x, from which the input locations x_i are drawn. In view of their dependence on u, the x_i are in general correlated each other and unbounded, e.g. for Gaussian u no bounded set contains x_i with probability one. Such peculiarities, inherited by the system identification setting, already violate the data generation assumptions routinely adopted in machine learning.

The identification data $\{x_i, y_i\}_{i=1}^{\infty}$ are assumed to be a stationary stochastic process. In particular, each couple (x, y) has joint

probability measure $\mu_{yx}(y,x) = \mu_{y|x}(y|x)\mu_x(x)$ where $\mu_{y|x}$ is the probability measure of the output y conditional on a particular input location x.

Given a function (dynamic system) f, the least squares error associated to f is

$$\mathscr{E}(y-f(x))^2 = \int_{\mathscr{X}\times\mathbb{R}} (y-f(x))^2 d\mu_{yx}(y,x). \tag{29}$$

The following result is well known and characterizes the minimizer of (29) which goes under the name of *regression function* in machine learning.

Theorem 13 (*The Regression Function*). *We have*

$$f_\rho = \arg\min_f \mathscr{E}(y-f(x))^2,$$

where f_ρ is the regression function *defined for any $x \in \mathscr{X}$ by*

$$f_\rho(x) = \int_{\mathbb{R}} y\, d\mu_{y|x}(y|x). \tag{30}$$

In our system identification context, f_ρ is the dynamic system associated to the optimal predictor that minimizes the expected quadratic loss on a new output drawn from μ_{yx}.

5.2. Consistency analysis

Consider a scenario where $\mu_{y|x}$ (and possibly also μ_x) is unknown and only N samples $\{x_i, y_i\}_{i=1}^N$ from μ_{yx} are available. We study the convergence of the RN in (1) to the optimal predictor f_ρ as $N \to \infty$ under the input-induced norm

$$\|f\|_x^2 = \int_{\mathscr{X}} f^2(x) d\mu_x(x).$$

This is the norm in the classical Lebesgue space $\mathscr{L}_2^{\mu_x}$ already introduced at the end of Section 2.

We assume that the reproducing kernel of $\mathscr{H}$ admits the expansion $\mathscr{K}(x,a) = \sum_{i=1}^\infty \zeta_i \rho_i(x)\rho_i(a), \;\; \zeta_i > 0\ \forall i$, with ζ_i and ρ_i defined via (9) and the probability measure μ_x.

Exploiting the regression function, the measurements process can be written as

$$y_i = f_\rho(x_i) + e_i, \tag{31}$$

where the errors e_i are zero-mean and identically distributed. They can be correlated each other and also with the input locations x_i. Given any $f \in \mathscr{H}$ and the ℓ-th kernel eigenfunction ρ_ℓ, we define the random variables $\{v_{\ell i}\}_{i\in\mathbb{Z}}$ by combining f, ρ_ℓ and the errors e_i defined by (31) as follows

$$v_{\ell i} = (f(x_i) + e_i)\, \rho_\ell(x_i). \tag{32}$$

Let $\mathscr{H}$ be stable so that the $\{v_{\ell i}\}$ form a stationary process. In particular, note that each $v_{\ell i}$ is the product of the outputs from two stable systems: the first one, $f(x_i) + e_i$, is corrupted by noise while the second one, $\rho_\ell(x_i)$, is noiseless.

Now, by summing up over ℓ the cross covariances of lag k using kernel eigenvalues ζ_ℓ as weights one obtains,[4]

$$c_k := \sum_{\ell=1}^\infty \zeta_\ell Cov\left(v_{\ell i}, v_{\ell,i+k}\right) \tag{33}$$

where $Cov(\cdot,\cdot)$ is the covariance operator. The next proposition shows that summability of the c_k is key for consistency.

Proposition 14 (*RN Consistency in System Identification*). *Let $\mathscr{H}$ be the RKHS with kernel*

$$\mathscr{K}(x,a) = \sum_{i=1}^\infty \zeta_i \rho_i(x)\rho_i(a), \;\; \zeta_i > 0\ \forall i, \tag{34}$$

where (ζ_i, ρ_i) are the eigenvalues/eigenfunctions pairs defined by (9) under the probability measure μ_x. Assume that, for any $r > 0$ and f s.t. $\|f\|_{\mathscr{H}} \le r$, there exists a constant C_r such that

$$\sum_{k=0}^\infty |c_k| < C_r < \infty, \tag{35}$$

with c_k defined in (33). Let also

$$\gamma \propto \frac{1}{N^\alpha}, \tag{36}$$

where α is any scalar in $(0, \frac{1}{2})$. Then, if $f_\rho \in \mathscr{H}$ and $\hat{g}_N$ is the RN in (1), as $N \to \infty$ one has

$$\|\hat{g}_N - f_\rho\|_x \longrightarrow_p 0, \tag{37}$$

where $\longrightarrow_p$ denotes convergence in probability.

The fact that $\sum_{\ell=1}^\infty |\zeta_\ell| < \infty$ is already an indication that (35) is not so hard to be satisfied. Indeed, to the best of our knowledge, (35) is the weakest RKHS condition currently available which guarantees RN consistency. In fact, previous works, like (Sun & Wu, 2010; Zou, Li, & Xu, 2009), beyond considering only noises with densities of compact support use mixing conditions (which rule out infinite memory systems) with fast mixing coefficients decay. These assumptions largely imply (35) as it can e.g. be deduced by Lemma 2.2 in Dehling and Philipp (1982).

From (33) one can see that (35) essentially reduces to studying summability of the covariances of the $v_{\ell i} = (f(x_i) + e_i)\, \rho_\ell(x_i)$. In particular, the covariance of such product sequences depends on the first four moments of the component sequences, see eq. 31 in Wecker (1978). The stability of $\mathscr{H}$ is thus crucial to ensure the existence of the moments of the $v_{\ell i}$. This is e.g. connected with the use of kernels like (16), (19) and (27) where the influence of past input locations on the output decays exponentially to zero as time progresses.

The relevance of (35) further emerges if more specific experimental conditions are considered. An example is given below by specializing Proposition 14 to the continuous-time linear setting. Here, the aim is to reconstruct the continuous-time impulse response of a linear system fed with a stationary input process from a sampled and noisy version of the output. Below, one can e.g. think of the sampling instants as $t_i = i\Delta + \delta_i$ where δ_i are identically distributed random variables with support on $[0, \Delta]$. Then, it is shown that, for Gaussian u, a weak condition on the input covariance's decay rate already guarantees consistency. This outcome can also be seen as a non trivial extension to the dynamic context of studies on functional linear regression like e.g. that illustrated in Yuan and Tony Cai (2010) under independent data assumptions.

Proposition 15 (*RN Consistency in Continuous-Time Linear System Identification*). *Let $\mathscr{H}$ be the RKHS with kernel $\mathscr{K}$ defined by the continuous-time stable spline kernel $K(t,s) = e^{-\beta \max(t,s)}$ with support restricted to any compact set of $\mathbb{R}_+ \times \mathbb{R}_+$. Assume that*

- *the regression function f_ρ is a continuous-time and time -invariant linear system with impulse response θ satisfying $\int \theta^2(t)dt < \infty$;*

[4] As shown in the Appendix the c_k in (33) are invariant w.r.t. the particular spectral decomposition used to obtain the pairs (ζ_i, ρ_i).

- *the system input u is a stationary Gaussian process with $Cov(u(t+\tau), u(t))$ decaying to zero as $1/\tau^{1+\delta}$ for $\delta > 0$;*
- *the errors e_i in (31) are white, independent of u.*

Then, letting γ satisfy (36), the RN in (1) is consistent, i.e.

$$\|\hat{g}_N - f_\rho\|_x \longrightarrow_p 0. \tag{38}$$

Remark 16 (*Convergence to the True Impulse Response*)**.** The optimal predictor f_ρ is a dynamic system, unique as map $\mathscr{X} \to \mathbb{R}$. However, considering e.g. the linear scenario in Proposition 15, such functional can be associated to different impulse responses (recall the relationship between the space $\mathscr{H}$ of linear dynamic systems and the space $\mathscr{I}$ of impulse responses illustrated in Section 3.3). Convergence to θ can be guaranteed under persistently exciting conditions related to u and the sampling instants t_i (Ljung, 1999). One can then wonder which kind of impulse response estimate is obtained if such conditions are not satisfied. The answer is given by Pillonetto et al. (2014, Theorem 3 on p. 371) which reveals that the impulse response estimate (21) associated to $\hat{g}^N$ coincides with (14). Thus, among all the possible impulse responses defining the optimal predictor f_ρ, the estimator (21) will asymptotically privilege that of minimum norm in $\mathscr{I}$.

6. Conclusions

We have discussed RKHSs whose elements are functionals associated to dynamic systems. This perspective establishes a solid link between system identification and machine learning, with focus on the problem of *learning from examples*. Such framework has led to simple derivations of RKHSs BIBO stability conditions. In the nonlinear scenario, including in the regularized estimator also other stability notions, e.g. in the Sontag's sense (Sontag, 2008), appears an important future research direction.

Finally, RN convergence to the optimal predictor has been proved under assumptions tailored to system identification, also pointing out the link between consistency and RKHS stability.

Appendix

Proof of Proposition 14

We start reporting three useful lemmas instrumental to the main proof. First, define

$$\hat{f} = \arg\min_{f \in \mathscr{H}} \|f - f_\rho\|_x^2 + \gamma \|f\|_{\mathscr{H}}^2 \tag{39}$$

and

$$\eta_i(\cdot) = \left[y_i - \hat{f}(x_i) \right] \mathscr{K}(x_i, \cdot). \tag{40}$$

In addition, the notation $S_x : \mathscr{H} \to \mathbb{R}^N$ is the sampling operator defined by $S_x f = [f(x_1) \dots f(x_n)]$ while S_x^* is its adjoint given by

$$S_x^* c = \sum_{i=1}^N c_i \mathscr{K}_{x_i} \quad \forall\, c \in \mathbb{R}^N. \tag{41}$$

The first lemma below involves the definitions of $\eta_i, \hat{f}, S_x$ and S_x^* given above. It is derived from Smale and Zhou (2007) and Proposition 1 in Smale and Zhou (2005).

Lemma 17. *It holds that*

$$\mathscr{E}\eta_i = \gamma \hat{f}. \tag{42}$$

Furthermore, if $v \in \mathscr{H}$ and f satisfies

$$\left(\frac{S_x^* S_x}{N} + \gamma I \right) f = v,$$

where I denotes the identity operator, one has

$$\|f\|_{\mathscr{H}} \le \frac{1}{\gamma} \|v\|_{\mathscr{H}}. \tag{43}$$

The second lemma states a bound between the expected RKHS distance between $\hat{g}_N$ and $\hat{f}$.

Lemma 18. *Let $r = 2\|f_\rho\|_{\mathscr{H}}$. Then, for any $\gamma > 0$ one has*

$$\mathscr{E}\|\hat{g}_N - \hat{f}\|_x \le \frac{1}{\gamma} \sqrt{\max\left(1, \max_i\, \zeta_i\right)} \sqrt{\frac{2C_r}{N}}. \tag{44}$$

Proof. It comes from the representer theorem and (41) that

$$\hat{g}_N = S_x^* (\mathbf{K} + N\gamma I_N)^{-1} Y.$$

Then, we have

$$\begin{aligned} &\left(\frac{S_x^* S_x}{N} + \gamma I \right) \hat{g}_N \\ &= \frac{S_x^*}{N} \left(\mathbf{K}(\mathbf{K} + N\gamma I_N)^{-1} + N\gamma (\mathbf{K} + N\gamma I_N)^{-1} \right) Y \\ &= \frac{S_x^*}{N} Y. \end{aligned}$$

Hence, one has

$$\begin{aligned} \hat{g}_N - \hat{f} &= \left(\frac{S_x^* S_x}{N} + \gamma I \right)^{-1} \left(\frac{S_x^* Y}{N} - \frac{S_x^* S_x \hat{f}}{N} - \gamma \hat{f} \right) \\ &= \left(\frac{S_x^* S_x}{N} + \gamma I \right)^{-1} \frac{1}{N} \sum_{i=1}^N (\eta_i - \mathscr{E}[\eta_i]), \end{aligned}$$

where we used the equality $S_x^* Y - S_x^* S_x \hat{f} = \sum_{i=1}^N \eta_i$ and (42). Using (43) in Lemma 17, we then obtain

$$\|\hat{g}_N - \hat{f}\|_{\mathscr{H}} \le \frac{1}{\gamma} \left\| \frac{1}{N} \sum_{i=1}^N (\eta_i - \mathscr{E}[\eta_i]) \right\|_{\mathscr{H}}. \tag{45}$$

Now, let $f := f_\rho - \hat{f}$. With the e_i defined in (31), we can write

$$\begin{aligned} &\eta_i(\cdot) - \mathscr{E}\eta_i(\cdot) \\ &= [f(x_i) + e_i]\, \mathscr{K}(x_i, \cdot) - \mathscr{E}\,[f(x_i) + e_i]\, \mathscr{K}(x_i, \cdot) \\ &= \sum_{\ell=1}^\infty \zeta_\ell \left[(f(x_i) + e_i)\rho_\ell(x_i) - \mathscr{E}(f(x_i) + e_i)\rho_\ell(x_i) \right] \rho_\ell(\cdot) \\ &= \sum_{\ell=1}^\infty \zeta_\ell \left[v_{\ell i} - m_\ell \right] \rho_\ell(\cdot) \end{aligned}$$

where $v_{\ell i} = (f(x_i) + e_i)\, \rho_\ell(x_i)$ and $m_\ell = \mathscr{E} v_{\ell i}$. Now, the structure of the RKHS norm outlined in (8) allows us to write

$$\begin{aligned} &\langle \eta_i(\cdot) - \mathscr{E}\eta_i, \eta_j(\cdot) - \mathscr{E}\eta_j \rangle_{\mathscr{H}} \\ &= \langle \sum_{\ell=1}^\infty \zeta_\ell \left(v_{\ell i} - m_\ell \right) \rho_\ell(\cdot), \sum_{\ell=1}^\infty \zeta_\ell \left(v_{\ell j} - m_\ell \right) \rho_\ell(\cdot) \rangle_{\mathscr{H}} \\ &= \sum_{\ell=1}^\infty \zeta_\ell \left(v_{\ell i} - m_\ell \right) \left(v_{\ell j} - m_\ell \right). \end{aligned}$$

So, using definition (33)

$$\mathscr{E}\langle \eta_i - \mathscr{E}\eta_i, \eta_j - \mathscr{E}\eta_j \rangle_{\mathscr{H}} = \sum_{\ell=1}^\infty \zeta_\ell Cov\left(v_{\ell i}, v_{\ell, j} \right) = c_{|i-j|}.$$

Comparing the values of the objective in (39) at the optimum $\hat{f}$ and at f_ρ, one finds $\|\hat{f}\|_{\mathscr{H}} \leq \|f_\rho\|_{\mathscr{H}}$ so that

$$\|f_\rho - \hat{f}\|_{\mathscr{H}} = \|f\|_{\mathscr{H}} \leq 2\|f_\rho\|_{\mathscr{H}}.$$

This, combined with (35), implies that for any $\gamma > 0$

$$\sum_{k=0}^{\infty} |c_k| < C_r < \infty, \quad r = 2\|f_\rho\|_{\mathscr{H}}.$$

Hence, we obtain

$$\mathscr{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\eta_i - \mathscr{E}[\eta_i])\right\|_{\mathscr{H}}^2\right] \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|c_{|i-j|}| \leq \frac{2C_r}{N}.$$

Now, recall that, if $f = \sum_{i=1}^{\infty} a_i\rho_i$, then $\|f\|_{\mathscr{H}}^2 = \sum_{i=1}^{\infty}\frac{a_i^2}{\zeta_i}$ while $\|f\|_x^2 = \sum_{i=1}^{\infty} a_i^2$. Thus, for any $f \in \mathscr{H}$, one has

$$\|f\|_x^2 \leq \max\left(1, \max_i \zeta_i\right)\|f\|_{\mathscr{H}}^2.$$

Jensen's inequality and (45) then completes the proof. ■

Now, we need to set up some additional notation. Following Smale and Zhou (2005, 2007), given the integral operator

$$L_{\mathscr{K}}[\rho_i] := \int_{\mathscr{X}} \mathscr{K}(\cdot, u)\rho_i(u)d\mu_x(u) = \zeta_i\rho_i,$$

for $r > 0$ we define

$$L_{\mathscr{K}}^{-r}[f] = \sum_{i=1}^{\infty}\frac{c_i}{\zeta_i^r}\rho_i.$$

The third lemma reported below contains the inequality (46) which was also derived in Smale and Zhou (2005) assuming a compact input space. However, the bound holds just assuming the validity of the kernel expansion (34). To see this, recalling Theorem 5, first note that $L_{\mathscr{K}}^{-r}f_\rho \in \mathscr{L}_2^{\mu_x}$ for any $0 \leq r \leq 1/2$. So, there exists $g \in \mathscr{L}_2^{\mu_x}$, say $g = \sum_{i=1}^{\infty} d_i\rho_i$, such that $f_\rho = \sum_{i=1}^{\infty}\zeta_i^r d_i\rho_i$. After simple computations, one obtains $\hat{f} - f_\rho = -\sum_{i=1}^{\infty}\frac{\gamma}{\zeta_i+\gamma}\zeta_i^r d_i\rho_i$ and the same manipulations contained in the proof of Theorem 4 in Smale and Zhou (2005, p. 295) lead to the following result.

Lemma 19. *For any $0 < r \leq 1/2$, one has*

$$\|\hat{f} - f_\rho\|_x \leq \gamma^r\|L_{\mathscr{K}}^{-r}f_\rho\|_x. \tag{46}$$

Combining (44) and (46), for any $0 < r \leq 1/2$ it holds that

$$\mathscr{E}\|\hat{g}_N - f_\rho\|_x \leq \gamma^r\|L_{\mathscr{K}}^{-r}f_\rho\|_x + \frac{1}{\gamma}\sqrt{\max\left(1, \max_i \zeta_i\right)}\sqrt{\frac{2C_r}{N}}. \tag{47}$$

Hence, when γ is chosen according to (36), $\mathscr{E}\|\hat{g}_N - f_\rho\|_x$ converges to zero as N grows to ∞. Using the Markov inequality, (37) is finally obtained.

Proof of Proposition 15

Let $\mathscr{I}$ be the space of impulse responses with compact support e.g. on $[0, T]$ induced by the stable spline kernel K. Then, it comes from Pillonetto et al. (2010) that $\|\theta\|_{\mathscr{I}}^2 \propto \int_0^T \dot{\theta}^2(t)e^{\beta t}dt$. So, finite energy of the first derivative of θ ensures that the optimal predictor f_ρ belongs the RKHS $\mathscr{H}$ defined by the linear kernel $\mathscr{K}$ induced by K.

Now, we have just to prove that condition (35) holds. Recall that the c_k are invariant w.r.t. the particular kernel expansion of $\mathscr{K}$ adopted. For the stable spline kernel K we choose the expansion $K(t, \tau) = \sum_{\ell=1}^{\infty}\zeta_\ell\psi_\ell(t)\psi_\ell(\tau)$ derived in Pillonetto et al. (2010) where

$$\psi_\ell(t) = \sqrt{2}\sin\left(\frac{e^{-\beta t}}{\sqrt{\zeta_\ell}}\right), \quad \zeta_\ell = \frac{1}{(\ell\pi - \pi/2)^2}.$$

The eigenfunctions thus satisfy

$$|\psi_\ell(t)| \leq \sqrt{2} \ \forall(\ell, t). \tag{48}$$

Now, let f be any dynamic system satisfying $\|f\|_{\mathscr{H}} \leq r$ and let θ be the associated impulse response of minimum norm living in $\mathscr{I}$.[5] From the arguments discussed in Section 3.3 one then has $\|\theta\|_{\mathscr{I}} \leq r$. It then holds that

$$|\theta(t)| = |\langle\theta, K_t\rangle_{\mathscr{I}}| \leq r\sqrt{K(t, t)} \Longrightarrow \max_{t\in[0,T]}|\theta(t)| \leq A_r < \infty \tag{49}$$

with A_r independent of the particular f chosen inside the ball of radius r of $\mathscr{H}$.

Without loss of generality, the input u is now assumed zero-mean so that the $f(x_i)$ and $\rho_\ell(x_i)$ become zero-mean Gaussian processes. Using eq. 3.2 in Wecker (1978), for $k > 1$ one obtains

$$\begin{aligned} Cov\left(v_{\ell i}, v_{\ell,i+k}\right) &= Cov\left(f(x_i)\rho_\ell(x_i), f(x_{i+k})\rho_\ell(x_{i+k})\right) \\ &= Cov\left(f(x_i), f(x_{i+k})\right) Cov\left(\rho_\ell(x_i), \rho_\ell(x_{i+k})\right) \\ &+ Cov\left(f(x_i), \rho_\ell(x_{i+k})\right) Cov\left(f(x_i), \rho_\ell(x_{i-k})\right). \end{aligned} \tag{50}$$

Now, let h be any of the four covariances in the r.h.s. of (50). Combining (48), (49) and classical integral formulas for covariances computations, as e.g. reported in Papoulis (1991, p. 308–313), it is easy to obtain a constant B_r independent of ℓ such that $|h(k)| \leq B_r/k^{1+\epsilon}$. Condition (35) thus holds true and this completes the proof.

References

Akaike, H. (1979). Smoothness priors and the distributed lag estimator. Technical report, Department of Statistics, Stanford University.

Aravkin, A., Bell, B. M., Burke, J. V., & Pillonetto, G. (2015). The connection between Bayesian estimation of a Gaussian random field and RKHS. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(7), 1518–1524.

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012). On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs Minimum mse. *IFAC Proceedings Volumes*, *45*(16), 125–130.

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2014). Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ard and glasso. *Journal of Machine Learning Research (JMLR)*, *15*, 217–252.

Argyriou, A., & Dinuzzo, F. (2014). A unifying view of representer theorems. In *Proceedings of the 31th international conference on machine learning, Vol. 32* (pp. 748–756).

Argyriou, A., Micchelli, C. A., & Pontil, M. (2009). When is there a representer theorem? vector versus matrix regularizers. *Journal of Machine Learning Research*, *10*, 2507–2529.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Bai, E., & Liu, Y. (2007). Recursive direct weight optimization in nonlinear system identification: a minimal probability approach. *IEEE Transactions on Automatic Control*, *52*(7), 1218–1231.

Bai, E. W. (2010). Non-parametric nonlinear system identification: An asymptotic minimum mean squared error estimator. *IEEE Transactions on Automatic Control*, *55*(7), 1615–1626.

Bell, B. M., & Pillonetto, G. (2004). Estimating parameters and stochastic functions of one variable using nonlinear measurement models. *Inverse Problems*, *20*(3), 627.

Bergman, S. (1950). The Kernel function and conformal mapping. *Mathematical surveys and monographs*. AMS.

[5] The minimum norm impulse response is chosen without loss of generality since any other θ associated with f would induce the same input–output relationship.

Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, *75*, 1–120.

Bertero, M., Poggio, T., & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, 869–889.

Carmeli, C., De Vito, E., & Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, *4*, 377–408.

Chandrasekaran, V., Recht, B., Parrilo, P. A., & Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, *12*(6), 805–849.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, *48*(8), 1525–1535.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

De Nicolao, G., Sparacino, G., & Cobelli, C. (1997). Nonparametric input estimation in physiological systems: problems, methods and case studies. *Automatica*, *33*, 851–870.

Dehling, H., & Philipp, W. (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *The Annals of Probability*, *10*(3), 689–701.

Dinuzzo, F. (2015). Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, *53*(5), 3299–3317.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*.

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*, 1–50.

Franz, M. O., & Schölkopf, B. (2006). A unifying view of Wiener and Volterra theory and polynomial kernel regression. *Neural Computation*, *18*, 3097–3118.

Frigola, R., Lindsten, F., Schon, T. B., & Rasmussen, C. E. (2013). Bayesian inference and learning in Gaussian process state-space models with particle mcmc. In *Advances in neural information processing systems*.

Frigola, R., & Rasmussen, C. E. (2013). Integrated pre-processing for Bayesian nonlinear system identification with Gaussian processes. In *Proceedings of the 52nd annual conference on decision and control*.

Girosi, F. (1997). An equivalence between sparse approximation and support vector machines. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.

Goodwin, G. C., Gevers, M., & Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, *37*(7), 913–928.

Grossmann, C., Jones, C. N., & Morari, M. (2009). System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Proceedings of the 48th IEEE conference on decision and control* (pp. 4692–4697).

Guo, Z. C., & Zhou, D. X. (2013). Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, *38*(1), 207–223.

Hochstadt, H. (1973). *Integral equations*. John Wiley and Sons.

Kimeldorf, G., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*(2), 495–502.

Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series*. Springer.

Leithead, W. E., Solak, E., & Leith, D. J. (2003). Direct identification of nonlinear structure using Gaussian process prior models. In *Proceedings of European control conference*.

Lin, T., Horne, B. G., Tino, P., & Giles, C. L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, *7*, 1329–1338.

Liu, Z., & Vandenberghe, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, *31*(3), 1235–1256.

Ljung, L. (1999). *System identification - theory for the user* (2nd ed.). Upper Saddle River, N.J.: Prentice-Hall.

Ljung, L., Goodwin, G. C., & Agero, J. C. (2014). Stochastic embedding revisited: a modern interpretation. In *53rd IEEE conference on decision and control* (pp. 3340–3345).

Lukic, M. N., & Beder, J. H. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, *353*, 3945–3969.

Maritz, J. S., & Lwin, T. (1989). *Empirical Bayes method*. Chapman and Hall.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, *209*(3), 415–446.

Mohan, K., & Fazel, M. (2010). Reweighted nuclear norm minimization with application to system identification. In *American control conference* (pp. 2953–2959).

Mukherjee, S., Niyogi, P., Poggio, T., & Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, *25*(1), 161–193.

Papoulis, A. (1991). *Probability, random variables and stochastic processes*. Mc Graw-Hill.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., & Ljung, L. (2016). Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, *69*, 137–149.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2010). Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proceedings of the IEEE American cont. conf., Baltimora, USA*.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011). Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, *47*(2), 291–305.

Pillonetto, G., Chiuso, A., & Quang, Minh Ha (2011). A new kernel-based approach for nonlinear system identification. *IEEE Transactions on Automatic Control*, *56*(12), 2825–2840.

Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., Nicolao, G. De, & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: a survey. *Automatica*, *50*(3), 657–682.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, *78*, 1481–1497.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*(6981), 419–422.

Poggio, T., & Shelton, C. R. (1999). Machine learning, machine vision, and the brain. *AI Magazine*, *20*(3), 37–55.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Rojas, C. R., Toth, R., & Hjalmarsson, H. (2014). Sparse estimation of polynomial and rational dynamical models. *IEEE Transactions on Automatic Control*, *59*(11), 2962–2977.

Roll, J., Nazin, A., & Ljung, L. (2005). Nonlinear system identification via direct weight optimization. *Automatica*, *41*(3), 475–490.

Schiller, R. J. (1979). A distributed lag estimator derived from smoothness priors. *Economics Letters*, *2*(3), 219–223.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *Neural Networks and Computational Learning Theory*, *81*, 416–426.

Schölkopf, B., & Smola, A. J. (2001). *(Adaptive computation and machine learning). Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Shun-Feng, S., & Yang, F. Y. P. (2002). On the dynamical modeling with neural fuzzy networks. *IEEE Transactions on Neural Networks*, *13*, 1548–1553.

Sjöberg, J., Zhang, Q., Ljung, L., Delyon, B., Benveniste, A., Glorennec, P., et al. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, *31*(12), 1691–1724.

Smale, S., & Zhou, D. X. (2005). Shannon sampling II: connections to learning theory. *Applied and Computational Harmonic Analysis*, *19*, 285–302.

Smale, S., & Zhou, D. X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*, 153–172.

Sontag, E. D. (2008). Input to state stability: basic concepts and results. In *Nonlinear and Optimal Control Theory* (pp. 163–220).

Spinelli, W., Piroddi, L., & Lovera, M. (2005). On the role of prefiltering in nonlinear system identification. *IEEE Transactions on Automatic Control*, *50*(10), 1597–1602.

Steinwart, I., Hush, D., & Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, *100*(1), 175–194.

Sun, Hongwei (2005). Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, *21*(3), 337–349.

Sun, H., & Wu, Q. (2010). Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, *32*(2), 175–189.

Suykens, J. A. K., Alzate, C., & Pelckmans, K. (2010). Primal and dual model representations in kernel-based learning. *Statistical Surveys*, *4*, 148–183.

Suykens, J. A. K., Van Gestel, T., Brabanter, J. De, De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. Washington, D.C.: Winston/Wiley.

Vapnik, V. (1998). *Statistical learning theory*. New York, NY, USA: Wiley.

Vidyasagar, M. (1997). *A theory of learning and generalization: with applications to neural networks and control systems*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Vishwanathan, S. V. N., Smola, Alexander J., & Vidal, René (2007). Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, *73*(1), 95–119.
Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, *14*(4), 651–667.
Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.
Wang, C., & Zhou, D. X. (2011). Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, *27*(1), 55–67.
Wecker, W. E. (1978). A note on the time series which is the product of two stationary time series. *Stochastic Processes and their Applications*, *8*(2), 153–157.
Wu, Q., Ying, Y., & Zhou, D. X. (2006). Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, *6*, 171–192.
Yuan, M., & Tony Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, *38*, 3412–3444.
Zhao, W., Chen, H. F., Bai, E., & Li, K. (2015). Kernel-based local order estimation of nonlinear nonparametric systems. *Automatica*, *51*, 243–254.
Zou, B., Li, L., & Xu, Z. (2009). The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, *75*, 275–295.

Gianluigi Pillonetto was born on January 21, 1975 in Montebelluna (TV), Italy. He received the Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and the Ph.D. degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was Research Associate at the Department of Information Engineering, University of Padova, becoming an Assistant Professor in 2005. He is currently an Associate Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, estimation and machine learning.

He currently serves as Associate Editor for Automatica and IEEE Transactions on Automatic Control.

In 2003 he received the Paolo Durst award for the best Italian Ph.D. thesis in Bioengineering, and he was the 2017 recipient of the Automatica Prize.