The Quest for the Right Kernel in Bayesian Impulse Response Identification: The Use of OBFs ⋆

Mohamed A. H. Darwish [a,b], Gianluigi Pillonetto [c], Roland Tóth [a].

[a] *Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands.*

[b] *Electrical Engineering Department, Faculty of Engineering, Assiut University, 71515 Assiut, Egypt.*

[c] *Information Engineering Department, University of Padova, Padova 35131, Italy.*

Abstract

Kernel-based regularization approaches for impulse response estimation of Linear Time-Invariant (LTI) systems have received a lot of attention recently. The reason is that regularized least-squares estimators may achieve a favorable bias/variance trade-off compared with classical Prediction Error Minimization (PEM) methods. To fully exploit this property, the kernel function needs to capture relevant aspects of the data-generating system at hand. Hence, it is important to design automatic procedures for kernel design based on data or prior knowledge. The kernel models, so far introduced, focus on encoding smoothness and BIBO-stability of the expected impulse response while other properties, like oscillatory behaviour or the presence of fast and slow poles, have not been successfully implemented in kernel design. Inspired by the representation theory of dynamical systems, we show how to build stable kernels able to capture particular aspects of system dynamics via the use of Orthonormal Basis Functions (OBFs). In particular, desired dynamic properties can be easily encoded via the generating poles of OBFs. Such poles are seen as hyperparameters which are tuned via marginal likelihood optimization. Special cases of our kernel construction include Laguerre, Kautz, Generalized OBFs (GOBFs)-based kernel structures. Monte-Carlo simulations show that OBFs-based kernels perform well compared with stable spline/TC kernels, especially for slow systems with dominant poles close to the unit circle. Moreover, the capability of Kautz basis to model resonating systems is also shown.

Key words: Bayesian identification; System identification; Reproducing kernel Hilbert space; Orthonormal basis functions; Machine learning; Regularization.

1 Introduction

Data-driven modeling of *Linear Time-Invariant* (LTI) systems is a well-established field [1–3]. After choosing the form of the model for capturing the dynamics, e.g., state-space, transfer function or impulse response representation, the main stream model estimation methods either fall into the category of *Maximum Likelihood/Prediction Error Minimization* (ML/PEM) [1,2] or subspace [4] approaches. In ML/PEM, a finite-dimensional parametric model structure is first postulated and then parameter estimation is performed by minimizing the ℓ_2-loss of the prediction error. A main difficulty within these approaches is the choice of an adequate model structure with sufficiently low order and sufficiently high performance (most commonly the achievable prediction accuracy). This is related to the classical bias/variance trade-off of model estimation. The classical approach to resolve this trade-off is to resort to complexity criteria as *Akaike Information Criterion* (AIC) [5], *Bayesian Information Criterion* (BIC) [6] or *Cross-Validation* (CV) [1]. However, the performance of the ML/PEM equipped, e.g., with AIC, is not always satisfactory, especially for short and noisy observations [7].

A different approach to deal with the bias/variance dilemma is to resort to regularization. Popular approaches are, e.g., ℓ_1/LASSO [8], nuclear norm [9] and

⋆ This research has benefited from the financial support of the Student Mission, Ministry of Higher Education, Government of Egypt and the Netherlands Organization for Scientific Research (NWO, grant no.: 639.021.127). This paper was not presented at any IFAC meeting. Corresponding author Mohamed A. H. Darwish. Tel. +31-(0)65-939-7952.

Email addresses: `m.a.h.darwish@tue.nl` (Mohamed A. H. Darwish), `giapi@dei.unipd.it` (Gianluigi Pillonetto), `r.toth@tue.nl` (Roland Tóth).

Non-Negative Garotte (NNG) [10]. However, tuning of the regularization parameter of these methods in real world applications is often found to be a difficult task. In [11], an approach called SPARSEVA is proposed, which provides an automatic tuning of the amount of regularization to ensure consistency of the regularized estimator. However, ℓ_1 regularization is employed to perform parameter selection/order selection rather than optimizing the bias/variance trade-off. This points to the need for automatization of classical model order selection into a single step approach.

Alternatively, a novel kernel-based regularization approach, that merges ideas from machine learning, statistics and dynamical systems, has been introduced recently [12] and further developed in [7,13,14]. By this approach, an impulse response model structure is postulated and the model estimation is tackled by minimizing a regularized functional defined over a *Reproducing Kernel Hilbert Space* (RKHS). This approach overcomes ill-posedness and ill-conditioning by restricting the high degree-of-freedom offered by the nonparametric estimator via the inclusion of smoothness and stability information on the to-be-estimated impulse response function. To do so, a suitable kernel structure K, which uniquely defines the RKHS (estimation space) and depends on an unknown hyperparameter vector β, has to be designed. The tuning of β replaces the classical model order selection and can be efficiently accomplished by a Bayesian interpretation of regularization, where the unknown impulse response is a realization of a zero-mean stochastic Gaussian process [15] with covariance given by the kernel. In particular, one can employ an empirical Bayes approach where the marginal likelihood of β is maximized [16–18]. Some available kernel structures, inspired by machine learning literature, are, e.g., the *Stable Spline* (SS) kernels [12], *Diagonal/Correlated* (DC) kernels [13] and first order stable spline kernels known as *Tuned/Correlated* (TC) kernels [13], etc. Such kernel models include smoothness and exponential decay information on the impulse response, but do not embed explicitly other dynamical aspects.

By taking a look at system theory, *Orthonormal Basis Functions* (OBFs) have attractive properties in both system identification and series expansion representation of LTI systems [19,20]. OBFs can be generated by a cascaded network of stable inner transfer functions, i.e., *all-pass* filters, completely determined (modulo the sign) by their poles. In the frequency-domain, OBFs provide a complete orthonormal basis for the Hardy space $\mathscr{H}_2(\mathbb{E})$. This is the space of functions over $\mathbb{C}$ that are squared integrable on the unit circle and analytic outside of it. Furthermore, in the time-domain, their correspondents, i.e., their impulse responses, constitute a complete orthonormal basis for $\ell_2(\mathbb{N})$, i.e., the space of squared summable sequences, making them attractive for construction of kernels that can overcome the current challenges.

There have already been few attempts to introduce OBFs-based kernels for impulse response estimation in the Bayesian setting, e.g., [21]. However, the proposed OBFs-based kernels do not perform well compared, e.g., with the TC kernel, as shown in [21, Section V]. Indeed, in this paper, we will show that the formulation in [21] generates a kernel which does not imply the stability of the associated model set. Moreover, the number of basis functions to be used to construct such kernels, which is closely related to the used OBFs, is also an open question [21] which hampers the utilization of this idea in the Bayesian estimator. Indeed, when utilizing OBFs model structure or constructing a kernel function based on OBFs, we face two issues: i) the choice of an appropriate set of OBFs that has a wide representation capability; ii) the choice of an effective number of expansion coefficients to be estimated, i.e., the required number of basis repetition. As we will see, these two issues are related, i.e., with a "wrong" choice of the basis, a long expansion is needed while, with a "well-chosen" basis, a short expansion is sufficient to achieve the same prediction capability. We are aiming at a data-driven approach to decide on both issues.

In this paper, we tackle these problems by constructing implicitly stable OBFs-based kernels directly in the time-domain via the use of a decay term that weights the OBFs. We show that appropriate construction of this decay term not only ensures the stability of the impulse responses in the associated RKHS, but also circumvents the need for selecting the number of basis functions.

Preliminary work regarding this result can be found in [22]. However, novel aspects of the current paper include

(1) Derivation of the OBFs-based kernel from both system theoretic and machine learning perspectives;
(2) Derivation of the connection between regularized impulse response estimation with OBFs-based kernels and regularized OBFs expansion estimation presented in [21, Section IV];
(3) Stability analysis of kernels induced by general OBFs.

Since OBFs are determined in terms of the poles of the inner function that generates them, these poles can be interpreted as hyperparameters of the associated kernels. Hence, estimation of these poles can be performed in a Bayesian setting by maximizing the marginal likelihood with respect to the observed data. So, by tuning the generating poles together with the decay term through a marginal likelihood maximization, the abovementioned issues, i.e., i), ii), can be circumvented in a completely data-driven manner. As an illustration of the construction mechanism, three special cases of OBFs-based kernel structures, i.e., Laguerre, Kautz and Generalized OBFs basis [23,24], are introduced and compared to other structures such as the TC kernel. The paper is organized as follows. The problem statement is provided in Section 2. Then, in Section 3, important definitions related to the considered RKHSs are given. In Section 4, the concept of OBFs and their associated RKHS are introduced. Regularized impulse response estimation is explained in Section 5. In Section 6, regularized estimation with OBFs-based kernels is revealed. The proposed method is assessed by an extensive Monte Carlo simulation in Section 7, followed by the conclusions in Section 8.

Notation

The following notation will be used throughout the paper: $\mathbb{C}$ denotes the complex plane, $\mathbb{D}$ is the interior of the unit disc, i.e., $\{z \in \mathbb{C} \mid |z| < 1\}$, $\mathbb{T}$ is the unit circle, i.e., $\{z \in \mathbb{C} \mid |z| = 1\}$ and $\mathbb{E}$ is the exterior of the unit disc, i.e., $\mathbb{E} = \mathbb{C} \setminus (\mathbb{D} \cup \mathbb{T})$. z^* denotes the complex conjugate of the complex number $z \in \mathbb{C}$, $\mathbb{R}$ stands for real numbers, while $\mathbb{Z}$ denotes the set of integer numbers and $\mathbb{N}$ is all positive integers. $\|a\|_2$ represents the Euclidean norm of a vector a and $|A|$ is the determinant of a square matrix A. I_n denotes the n-dimensional identity matrix and δ_{ij} denotes the Kronecker delta, i.e., $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Finally, $G(z) : \mathbb{C} \to \mathbb{C}$ denotes the transfer function of a discrete-time, LTI, causal, *Single-Input Single-Output* (SISO) system, q is the forward time-shift operator, i.e., $qx(t) = x(t+1)$, $t \in \mathbb{Z}$ is the discrete time and with $G(q)$ we denote the time-domain transfer operator corresponding to $G(z)$.

2 Problem Statement

Consider a SISO, finite order, asymptotically stable and LTI discrete-time data-generating system described by

$$y(t) = G_0(q)u(t) + v(t), \tag{1}$$

where $y : \mathbb{Z} \to \mathbb{R}$ is the output, $u : \mathbb{Z} \to \mathbb{R}$ is the input of the system and v is a white Gaussian noise process with variance σ^2, i.e., $v(t) \sim \mathscr{N}(0, \sigma^2)$, independent of the input u. Here v is considered to be white for the sake of simplicity. The case when v is colored can be handled in a straightforward way as shown in [25, Section 5.3]. Furthermore,

$$G_0(q) = \sum_{k=1}^{\infty} \mathrm{g}_k q^{-k}, \tag{2}$$

is the transfer operator of the deterministic part of the system (1), where $\mathrm{g} = \{\mathrm{g}_k\}_{k=1}^{\infty}$ is the unknown impulse response associated with $G_0(q)$. In (2), it is assumed that G_0 does not have a feedthrough term, i.e., $\mathrm{g}_0 = 0$. Hence, the data-generating system (1) can be written as

$$y(t) = \sum_{k=1}^{\infty} \mathrm{g}_k u(t-k) + v(t) = (\mathrm{g} \circledast u)(t) + v(t), \tag{3}$$

where $(\mathrm{g} \circledast u)(t)$ denotes the convolution between the impulse response g and the input u at time t. Given N data points $\mathcal{D}_\mathrm{N} = \{u(t), y(t)\}_{t=1}^{N}$ generated by (3), our goal is to estimate g, as well as possible. The corresponding identification criterion to achieve this goal will be defined later.

3 Basis functions and Hilbert spaces

This section is included to make the paper self-contained and readers who are familiar with Hilbert spaces and kernel functions can skip it without losing the line of reasoning in the sequel of the discussion. Note that, in the next sections, we will discuss both real- and complex-valued spaces. Therefore, in this section, we will treat the general case of complex spaces from which the special case of real-valued spaces follows immediately.

3.1 Kernel functions and their RKHS

Let us first recall the definition of a positive definite kernel.

Definition 1 [Positive definite kernel]. *Let $\mathscr{X}$ be a metric space. A complex-valued function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ is called a positive definite kernel if it is continuous, symmetric and satisfies $\sum_{i,j=1}^{m} a_i a_j^* K(x_i, x_j) \geq 0$ for any finite set of points $\{x_1, \ldots, x_m\} \subset \mathscr{X}$ and $\{a_1, \ldots, a_m\} \subset \mathbb{C}$.* □

Definition 2 [Reproducing kernel]. *Let $\mathscr{H}$ be a Hilbert space of complex-valued functions on $\mathscr{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. A complex-valued function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ is a reproducing kernel for $\mathscr{H}$ if and only if*

(1) $\forall x \in \mathscr{X}, K_x = K(x, \cdot) \in \mathscr{H}$, where K_x is the so-called kernel section centered at x;
(2) The reproducing property holds, such that $f(x) = \langle f(\cdot), K(x, \cdot) \rangle_{\mathscr{H}}$, $\forall x \in \mathscr{X}, \forall f \in \mathscr{H}$. □

A Hilbert space of complex-valued functions which possesses a reproducing kernel is called an RKHS [26]. Moreover, due to the Moore-Aronszajn theorem [27], there is a one-to-one correspondence between an RKHS $\mathscr{H}$ and its reproducing kernel K, i.e., to every positive definite kernel K, there is a unique RKHS $\mathscr{H}$ with K as its reproducing kernel and vice versa.

Definition 3 [RKHS]. *Let K be a positive definite kernel function and $\mathscr{H}$ is the associated RKHS. Then, $\mathscr{H}$ is defined to be the closure of* $\mathrm{Span}\{K_x := K(x, \cdot) : x \in \mathscr{X}\}$, *i.e., the functions in $\mathscr{H}$ can be written as*

$$\mathscr{H} = \Big\{ f : \mathscr{X} \to \mathbb{C} \mid f(\cdot) = \sum_{i=1}^{\infty} a_i K_{x_i}(\cdot), \\ x_i \in \mathscr{X},\ a_i \in \mathbb{C},\ \|f\|_{\mathscr{H}} < +\infty \Big\},$$

where $\|f\|_{\mathscr{H}} = \sqrt{\langle f, f\rangle_{\mathscr{H}}}$ is the norm in $\mathscr{H}$ induced by the inner product defined in $\mathscr{H}$, which in terms of K satisfies that

$$\langle g, h\rangle_{\mathscr{H}} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j^* K(x_i, x_j),$$

for

$$g = \sum_{i=1}^{\infty} a_i K_{x_i}, \quad h = \sum_{j=1}^{\infty} b_j K_{x_j}. \qquad \square$$

Since the reproducing kernel K completely characterizes the associated RKHS $\mathscr{H}$, therefore, in the sequel, we shall denote that RKHS as $\mathscr{H}_K$ and its inner product as $\langle\cdot,\cdot\rangle_K$ with the associated norm $\|\cdot\|_K$.

3.2 An orthonormal basis viewpoint on kernels

Let us start by presenting the definition of the orthonormal basis of a Hilbert space.

Definition 4 [Orthonormal basis of a Hilbert space]. *A sequence $\{\vartheta_k\}_{k=1}^{\infty}$ in a Hilbert space $\mathscr{H}$ is said to be a complete orthonormal basis if the following conditions are satisfied:*

- $\langle\vartheta_i, \vartheta_j\rangle_{\mathscr{H}} = \begin{cases} 0, & \text{for } i \neq j \\ 1, & \text{for all } i = j \geq 1. \end{cases}$
- *For any $f \in \mathscr{H}$, $f = \sum_{k=1}^{\infty} c_k\vartheta_k$, where $c_k = \langle f, \vartheta_k\rangle_{\mathscr{H}}$ are the expansion coefficients of f under the basis $\{\vartheta_k\}_{k=1}^{\infty}$.* $\square$

Let $L_2(\mathscr{X})$ be the Hilbert space of squared integrable functions on $\mathscr{X}$. Furthermore, suppose that the kernel K is squared integrable, i.e.,

$$\int_{\mathscr{X}}\int_{\mathscr{X}} K^2(x, x')\, dx\, dx' < +\infty.$$

Then, there exists an orthonormal sequence of continuous eigenfunctions $\{\vartheta_i\}_{i=1}^{\infty} \in L_2(\mathscr{X})$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ with

$$\lambda_k\vartheta(x) \triangleq \int_{\mathscr{X}} K(x, x')\vartheta_k(x')\, dx',\ x \in \mathscr{X},$$

where λ_k is the eigenvalue associated with ϑ_k [26, Page 3]. Under these conditions, Mercer's theorem [28] allows us, to represent the kernel function K in terms of the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and the eigenfunctions $\{\vartheta_i\}_{i=1}^{\infty}$ as follows [15]:

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k\vartheta_k(x)\vartheta_k(x'),$$

where $x, x' \in \mathscr{X}$. Furthermore, $\{\sqrt{\lambda_k}\vartheta_k\}_{k=1}^{\infty}$ forms an orthonormal basis for $\mathscr{H}_K$, the associated RKHS of K. This means that any function $f \in \mathscr{H}_K$ can be represented as a linear combination of the orthonormal basis of the kernel K as follows

$$\mathscr{H}_K = \left\{ f:\mathscr{X} \to \mathbb{C} \mid f(x) = \sum_{i=1}^{\infty} d_i\vartheta_i(x), \sum_{i=1}^{\infty}\frac{|d_i|^2}{\lambda_i} < +\infty \right\}, \tag{4}$$

where $\{d_i\}_{i=1}^{\infty}$ is an absolute convergent sequence. Based on the inner product, $\|f\|_K^2 = \langle f, f\rangle_K = \sum_{i=1}^{\infty}|d_i|^2/\lambda_i$. Hence, the inner product $\langle f, g\rangle_K$ for any $f, g \in \mathscr{H}_K$ with $f = \sum_{i=1}^{\infty} a_i\vartheta_i$ and $g = \sum_{i=1}^{\infty} b_i\vartheta_i$ can be also represented as

$$\langle f, g\rangle_K = \sum_{i=1}^{\infty} a_i b_i^*/\lambda_i.$$

3.3 The spaces of stable discrete-time systems

Let us first introduce the space $\ell_2(\mathbb{N})$ as the space of squared summable sequences, i.e., $\sum_{k=1}^{\infty}|h(k)|^2 < \infty$, equipped with the well-defined inner product between any two elements $f, g \in \ell_2(\mathbb{N})$ as $\langle f, g\rangle_{\ell_2} = \sum_{k=1}^{\infty} f(k)g^*(k)$. An interesting subspace of $\ell_2(\mathbb{N})$ is the subspace $\mathcal{R}\ell_2(\mathbb{N})$ which contains only squared summable real sequences. Moreover, $\mathcal{R}\ell_1(\mathbb{N})$ is the subspace of absolutely summable real sequences, i.e., $\sum_{k=1}^{\infty}|h(k)| < \infty$, equipped with the norm $\|h\|_{\ell_1} = \sum_{k=1}^{\infty}|h(k)|$. Note that $\mathcal{R}\ell_1(\mathbb{N}) \subset \mathcal{R}\ell_2(\mathbb{N})$. The importance of the space $\mathcal{R}\ell_1(\mathbb{N})$ comes from the fact that impulse responses $h(k)$ of all finite-dimensional, discrete-time, stable and causal systems satisfy the necessary and sufficient condition $\sum_{k=1}^{\infty}|h(k)| < \infty$ [29], hence they belong to $\mathcal{R}\ell_1(\mathbb{N})$.

Definition 5 [The Hardy space over $\mathbb{E}$ [19]]. *Denote by $\mathscr{H}_2(\mathbb{E})$ the Hardy space of complex functions $F : \mathbb{C} \to \mathbb{C}$, which are analytic in $\mathbb{E}$ and squared integrable on $\mathbb{T}$. $\mathscr{H}_2(\mathbb{E})$ is equipped with an inner product that is defined as*

$$\langle F_1, F_2\rangle_{\mathscr{H}_2} = \frac{1}{2\pi \mathrm{j}}\oint_{\mathbb{T}} F_1(z)F_2^*(1/z^*)\frac{dz}{z}, \tag{5}$$

where $\mathrm{j} = \sqrt{-1}$ and $F_1, F_2 \in \mathscr{H}_2(\mathbb{E})$. $\square$

Due to the isomorphism between $\ell_2(\mathbb{N})$ and $\mathscr{H}_2(\mathbb{E})$, any sequence $f \in \ell_2(\mathbb{N})$ corresponds to one and only one function $F \in \mathscr{H}_2(\mathbb{E})$ and vice versa. The following z-transform defines this isomorphism

$$F(z) = \sum_{k=1}^{\infty} f(k)z^{-k}, \tag{6}$$

which holds for all $z \in \mathbb{C}$ in the corresponding region of convergence. $\mathscr{RH}_2(\mathbb{E})$ is the subspace of $\mathscr{H}_2(\mathbb{E})$ which contains all functions that have real-valued impulse responses. Note that $\mathcal{R}\ell_2(\mathbb{N})$ and $\mathscr{RH}_2(\mathbb{E})$ are also isomorphic.

It is worth to mention that functions in $\mathscr{RH}_2(\mathbb{E})$ are not necessarily rational. The subspace $\mathscr{RH}_{2-}(\mathbb{E})$ of $\mathscr{RH}_2(\mathbb{E})$ is defined to contain all strictly proper, finite-dimensional and real rational transfer functions which are analytic in $\mathbb{E}$ and are 0 for $z = \infty$.

Regarding these spaces the following canonical orthonormal basis can be given:

- $\mathcal{R}\ell_2(\mathbb{N})$: $\vartheta_i(k) = \delta_{ik}, \quad i \in \mathbb{N}$,
- $\mathscr{RH}_2(\mathbb{E})$: $\vartheta_i(z) = z^{-i}, \quad i \in \mathbb{N}$.

4 Rational Orthonormal Basis Functions

In this section, to develop our kernel construction, we introduce a general class of OBFs for $\mathscr{RH}_2(\mathbb{E})$, namely, the so-called *Takenaka-Malmquist functions*, and their special case, the *Generalized* OBFs (GOBFs).

4.1 Series expansion representation in terms of OBFs

In the sequel, we introduce OBFs which constitute a complete basis for $\mathscr{RH}_2(\mathbb{E})$. Let $G_0 \equiv 0$ and $\{G_i\}_{i=1}^{\infty}$ be a sequence of stable inner functions, also known as all-pass filters, which satisfy $G_i(z)G_i(1/z) = 1$. Let $\{\xi_1, \xi_2, \dots\} \subset \mathbb{D}$ denote the collection of all poles of the inner functions $G_1, G_2, \dots$ satisfying the completeness condition $\sum_{k=1}^{\infty}(1 - |\xi_k|) = \infty$. Based on $\{G_i\}_{i=1}^{\infty}$, the so-called Takenaka-Malmquist basis functions are defined as [19]

$$\psi_k(z) = \frac{\sqrt{1 - |\xi_k|^2}}{z - \xi_k} \prod_{i=1}^{k-1} \frac{1 - \xi_i^* z}{z - \xi_i}. \tag{7}$$

Note that, in the general case, in the sense that there are no further restrictions on the generating poles, such basis have complex-valued impulse responses, i.e., they span $\mathscr{H}_2(\mathbb{E})$. In order to guarantee that the associated impulse responses with the considered basis are real-valued, i.e., that they are restricted to span $\mathscr{RH}_2(\mathbb{E})$, the complex poles should appear in complex conjugate pairs.

The special case when all G_i are equal, i.e., $G_i(z) = G_{\mathrm{b}}, \forall i > 0$, where G_{b} has McMillan degree $n_{\mathrm{g}} > 0$, are know as GOBFs or *Hambo functions* for arbitrary $n_{\mathrm{g}} > 0$, *Laguerre functions* for $n_{\mathrm{g}} = 1$ and 2-parameters *Kautz functions* for $n_{\mathrm{g}} = 2$. Note that for these special cases, i.e., GOBFs, Laguerre and Kautz functions, the completeness condition is always fulfilled. In the sequel, we discuss GOBFs functions in more detail. Let $G_{\mathrm{b}} \in \mathscr{RH}_{2-}(\mathbb{E})$ be an inner function with McMillan degree $n_{\mathrm{g}} > 0$. Such a function is completely determined, modulo the sign, by its poles $\Lambda_{n_{\mathrm{g}}} = [\xi_1 \cdots \xi_{n_{\mathrm{g}}}] \in \mathbb{D}^{n_{\mathrm{g}}}$:

$$G_{\mathrm{b}}(z) = \pm \prod_{i=1}^{n_{\mathrm{g}}} \frac{1 - \xi_i^* z}{z - \xi_i}, \tag{8}$$

with $\Lambda_{n_{\mathrm{g}}}$ containing real poles and/or complex conjugate pole pairs. Let (A, B, C, D) be a minimal balanced *state-space* realization of $G_{\mathrm{b}}(z)$. The class of GOBFs is obtained by cascading copies of G_{b}, i.e., identical $n_{\mathrm{g}}^{\mathrm{th}}$ order all-pass filters, and can be written in a vector form as:

$$V_k(z) = V_1(z) G_{\mathrm{b}}^{k-1}(z), \quad \text{for } k > 1, \tag{9}$$

where $V_1(z) = (zI - A)^{-1}B$. Let $\varphi_i = [V_1]_i$ denote the i^{th} element of V_1. Then, the GOBFs consists of the functions

$$\Psi = \{\psi_k\}_{k=1}^{\infty} = \left\{\varphi_i G_{\mathrm{b}}^j\right\}_{i=1, j=0}^{n_{\mathrm{g}}, \infty}, \text{ with } k = j \cdot n_{\mathrm{g}} + i. \tag{10}$$

These functions, i.e., (10), constitute a complete orthonormal basis for $\mathscr{RH}_2(\mathbb{E})$. As a result, any $G \in \mathscr{RH}_2(\mathbb{E})$ can be decomposed as

$$G(z) = \sum_{k=1}^{\infty} c_k \psi_k(z), \tag{11}$$

with $c_k \in \mathbb{R}$. Expansion (11) can be seen as the generalization of expansion with *pulse basis functions*, i.e., $\left\{z^{-k}\right\}_{k=1}^{\infty}$, used in the impulse response model structure (2). It can be shown that the rate of convergence of this series expansion is bounded by $\rho = \max_k |G_{\mathrm{b}}(\eta_k^{-1})|$, called the decay rate, where $\{\eta_k\}$ are the poles of $G(z)$ [30]. In practice, only a finite number of terms $\{\psi_k\}_{k=1}^{n_{\mathrm{b}}}$ is used, like in *Finite Impulse Response* (FIR) models, where $\left\{z^{-k}\right\}_{k=1}^{n}$ are used as basis functions. In contrast with FIR structures, the OBFs parameterization uses a broad class of basis functions with *Infinite Impulse Representation* (IIR). Therefore, OBFs parameterization can achieve an arbitrary low modeling error with a relatively small number of parameters due to the faster convergence of the series representation than in the FIR case, which in system identification results in decreased variance of the final model estimate [20,31].

Since we are interested in impulse response estimation based on time-domain data, it is more convenient to define the corresponding OBFs in the time-domain. Denote by $\phi_k(t) = \mathcal{Z}^{-1}\{\psi_k(z)\}$ the correspondent of ψ_k in the time domain, where $\mathcal{Z}^{-1}\{\cdot\}$ is the inverse z-transform on the appropriate region of convergence. Hence,

$$\Phi = \{\phi_k\}_{k=1}^{\infty}, \tag{12}$$

is a complete basis of $\mathcal{R}\ell_2(\mathbb{N})$ [32]. As a result, any impulse response $\mathrm{g} \in \mathcal{R}\ell_2(\mathbb{N})$ associated with a $G(z) \in$

$\mathscr{RH}_2(\mathbb{E})$ can be written as

$$\mathrm{g}(t) = \sum_{k=1}^{\infty} c_k \phi_k(t), \tag{13}$$

where $c_k \in \mathbb{R}$ and $\{c_k\}_{k=1}^{\infty}$ is equal to the expansion coefficients in (11). Next, we derive a magnitude bound for the Takenaka-Malmquist basis [19], which will be useful later.

Proposition 1 [Magnitude bound of OBFs]. *Consider the general Takenaka-Malmquist basis which is defined as in* (7) *with* $\{\xi_i\}_{i=1}^{\infty} \subset \mathbb{D}$*, which are assumed to appear as real or complex conjugate pairs, being the generating poles of* $\{\psi_k\}_{k=1}^{\infty} \in \mathscr{RH}_2(\mathbb{E})$ *and* $\{\phi_k\}_{k=1}^{\infty} \in \mathcal{R}\ell_2(\mathbb{N})$*. It holds that*

$$\|\phi_k\|_{\ell_1} \leq 2k\kappa, \tag{14}$$

where $\kappa \in \mathbb{R}$ *depends on the generating poles.*

Proof: See Appendix 9.1. □

4.2 RKHS associated with OBFs in the time-domain

A fundamental result on RKHS:

Proposition 2 [Unique kernel for an RKHS [27]]. *Let* $\mathscr{H}$ *be a separable*[1] *Hilbert space of functions over* $\mathscr{X}$ *with orthonormal basis* $\{\vartheta_k\}_{k=1}^{\infty}$*. Then,*

$$\mathscr{H} \text{ is an RKHS} \iff \sum_{k=1}^{\infty} |\vartheta_k(x)|^2 < \infty, \ \forall x \in \mathscr{X}.$$

The unique kernel K *that is associated with* $\mathscr{H}$ *is*

$$K(x,y) = \sum_{k=1}^{\infty} \vartheta_k(x)\vartheta_k(y). \qquad \square$$

Consider $\mathcal{R}\ell_2(\mathbb{N})$ and its standard orthonormal basis, see Section 3.3. Using the above result, it is immediate to conclude that $\mathcal{R}\ell_2(\mathbb{N})$ is an RKHS with a kernel given by the infinite-dimensional identity matrix, i.e. $K(i,j) = \delta_{ij}$. Now, the simplest kernel that can be built using GOBFs defined in (12) is

$$K_\Phi(i,j) = \sum_{k=1}^{\infty} \phi_k(i)\phi_k(j), \tag{15}$$

which represents the formulation of the OBFs-based kernel in time-domain and is a reproducing kernel for the RKHS space spanned by Φ, i.e., $\mathcal{R}\ell_2(\mathbb{N})$. The effectiveness of such kernel construction to identify impulse responses of LTI systems will be assessed in the sequel.

[1] A Hilbert space is said to be separable if it has a basis with countable number of elements.

5 Regularized Estimation of IIRs

In this section, we consider the problem of estimating the impulse response of the deterministic part of system (1) defined in (3), i.e., $\mathrm{g} = \{\mathrm{g}_k\}_{k=1}^{\infty}$, from a given set of observed data, i.e., $\mathcal{D}_\mathrm{N} = \{u(t), y(t)\}_{t=1}^{N}$. Following the classical approach, in the PEM setting and under the considered noise model, the following quadratic loss of the prediction error is minimized to get the model estimate (see [1])

$$\hat{\mathrm{g}} = \arg\min_{\mathrm{g}} \sum_{t=1}^{N} \big(y(t) - (\mathrm{g} \circledast u)(t)\big)^2. \tag{16}$$

Since the reconstruction of the impulse response function from a finite number of observations is a *deconvolution* problem, which is always ill-posed, hence, estimates with large variance are expected to be obtained, especially in case of short and noisy data set. An attractive way to cope with this problem is to introduce regularization into the estimation problem, which can be viewed from two equivalent perspectives: the first is functional approximation in RKHS and the second is a Bayesian interpretation.

5.1 Regularization in RKHS

In kernel based regularization, the estimation of the impulse response of a stable LTI system from noisy measurements is tackled by minimizing a regularized functional with respect to a "well-chosen" RKHS $\mathscr{H}_K$. The proposed estimator for impulse response estimation in [12] is based on solving the following Tikhonov-type variational problem [33]:

$$\hat{\mathrm{g}} = \operatorname*{argmin}_{\mathrm{g}\in\mathscr{H}_K} \ \sum_{t=1}^{N} \big(y(t) - (\mathrm{g} \circledast u)(t)\big)^2 + \mu\|\mathrm{g}\|_K^2, \tag{17}$$

where $\mu \geq 0$ is the regularization parameter. It is worth to mention that the cost function in (17) consists of two terms. The first term is the quadratic loss of the prediction error accounting for the data fit. The second term, i.e., the regularizer $\|\cdot\|_K^2$, controls the model complexity. The latter term also renders the problem well-posed by restricting the high degree of freedom offered by the nonparametric estimation. The restriction is introduced via expected properties of the impulse response function, e.g., smoothness and/or stability, expressed through K. The design of the structure of K involves choosing a parameterized form of K with some hyperparameters β which can express a wide variety of impulse responses, but at the same time restrict the high degree of freedom by encoding the expected dynamical properties like stability, oscillatory behaviour, etc. Moreover, β must be low dimensional such that its optimization by using an empirical Bayes approach in terms of marginal likelihood maximization [16] can be efficiently accomplished in the considered Bayesian setting. This provides automatic model structure selection whose efficiency depends on the choice of the structure of K [18]. Such a tuning approach has shown to better balance data fit and model complexity compared with classical tuning methods, e.g., CV.

It is worth to mention that, by considering an FIR truncation of (3) of order n with $\theta = [\mathrm{g}_1 \cdots \mathrm{g}_n]^\top \in \mathbb{R}^n$, the truncation of (3) can be written in the following matrical form

$$Y \cong U_n\theta + V, \tag{18}$$

where $Y{=}[y(1) \cdots y(N)]^\top$, $U_n{=}[\bar{U}_n^\top(1) \cdots \bar{U}_n^\top(N)]^\top$with $\bar{U}_n(i){=}[u(i-1) \cdots u(i-n)]$ and $V{=}[v(1) \cdots v(N)]^\top$.

Hence, (17) becomes equivalent to the following *Regularized Least-Squares* (ReLS) problem [25, Section 11.3]

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \ \|Y - U_n\theta\|_2^2 + \mu\theta^\top \mathbf{K}^{-1}(\beta)\theta, \tag{19a}$$

$$= \mathbf{K}(\beta)U_n^\top \left(U_n\mathbf{K}(\beta)U_n^\top + \mu I_N\right)^{-1} Y, \tag{19b}$$

where $\mathbf{K}(\beta)$ is an $n \times n$ kernel matrix, which is defined as $[\mathbf{K}]_{ij} = K(i,j)$ and parameterized in terms of β that contains the hyperparameters requiring tuning.

5.2 Bayesian perspective on regularization

Under the Gaussian regression framework [15], the estimator (17) admits a Bayesian interpretation. To show that, consider the linear regression model (18), where θ is modeled as a zero-mean Gaussian process [15], independent of the noise V, with a covariance (kernel) matrix $\mathbf{K}(\beta)$, i.e., $\theta \sim \mathscr{N}(0, \mathbf{K}(\beta))$. Accordingly, $Y \sim \mathscr{N}(0, \Sigma)$, where Σ is the covariance of Y given by $\Sigma := U_n\mathbf{K}(\beta)U_n^\top + \sigma^2 I_N$. As a result, θ and Y will be jointly Gaussian distributed. Hence, in terms of this Bayesian interpretation, the minimum variance estimator of θ for known Y, U_n and β is $E\{\theta \mid Y, U_n, \beta\}$ given by (19b), where E is the expectation operator. This interpretation provides an efficient way to estimate the unknown hyperparameters from data following an empirical Bayes approach [16,17]. Denote $p(Y|\beta)$ the likelihood function of the observations Y given β. Then, the maximum likelihood estimate of β is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \ \ p(Y|\beta) = \underset{\beta}{\operatorname{argmin}} \ \ Y^\top\Sigma^{-1}Y + \log|\Sigma|. \tag{20}$$

By replacing β in (19b) by its maximum marginal likelihood estimate $\hat{\beta}$, we obtain the so-called empirical Bayes estimate [2] $\hat{\theta}$.

[2] It is known from [25] that the optimal value of μ is the noise variance σ^2 in the sense that it minimizes the *Mean Squared Error* (MSE) of the estimator $\hat{\theta}$, which is typically not known a priori. One possible way is to treat σ^2 as an additional hyperparameter. Alternatively, a low-bias high-order ARX [12] or FIR model [13] can be estimated with least squares and then use the sample variance of the residuals as an estimate of σ^2 [1], [34]. In this paper, we follow the approach of estimating a low-bias high-order FIR model as in [13].

5.3 Kernel structures for impulse response estimation

For impulse response estimation, the kernel function K should reflect what is reasonable to assume about the impulse response, e.g., if the system is exponentially stable, the impulse response coefficients g_k should decay exponentially, and if the impulse response is smooth, neighboring values should have a positive correlation [25]. For this purpose, it is useful to recall that the optimal kernel [13, Theorem 1] for the estimation problem (17) is given by:

$$K(i,j) = \mathrm{g}_i\mathrm{g}_j, \tag{21}$$

where $i,j \in \mathbb{N}$ and $\mathrm{g} = \{\mathrm{g}_i\}_{i=1}^\infty$ is the true impulse response. Even if (21) is impossible to be used in practice since the true impulse is unknown, it provides a guideline to design a suitable kernel function for regularized impulse response estimation.

In the literature, many kernel structures have been introduced, e.g., SS kernel [12], DI kernel, DC kernel [13], TC kernel [13,14]:

$$K^{\mathrm{DI}}_{\mathrm{i,j}}(\beta) = \begin{cases} \beta_1\beta_2^i, \ i \geq j \\ 0, \ \text{otherwise} \end{cases}, \ \beta = [\beta_1 \ \beta_2]^\top \tag{22a}$$

$$K^{\mathrm{DC}}_{\mathrm{i,j}}(\beta) = \beta_1\beta_2^{|i-j|}\beta_3^{\frac{i+j}{2}}, \ \beta = [\beta_1 \ \beta_2 \ \beta_3]^\top \tag{22b}$$

$$K^{\mathrm{TC}}_{\mathrm{i,j}}(\beta) = \beta_1 \min(\beta_2^i, \beta_2^j), \ \beta = [\beta_1 \ \beta_2]^\top \tag{22c}$$

$$K^{\mathrm{SS}}_{\mathrm{i,j}}(\beta) = \begin{cases} \beta_1\frac{\beta_2^{2\mathrm{i}}}{2}(\beta_2^{\mathrm{j}} - \frac{\beta_2^{\mathrm{i}}}{3}), \ i \geq j \\ \beta_2\frac{\beta_2^{2\mathrm{j}}}{2}(\beta_2^{\mathrm{i}} - \frac{\beta_2^{\mathrm{j}}}{3}), \ \text{otherwise} \end{cases}, \ \beta{=}[\beta_1 \ \beta_2]^\top. \tag{22d}$$

These kernels have a well-known spectral decomposition [35, Theorem 3.1], [12, Section 5] and their orthonormal basis decay to zero guaranteeing the stability of the impulse responses in the associated RKHS. However, there are other dynamical properties that could be included, e.g., resonance behaviour, damping, etc. In the next section, we propose an advanced kernel structure that is capable of expressing these dynamical aspects, which can be seen as a refined form of the above kernels to step closer to (21) with a flexible parameterization.

6 OBFs kernels based IIR estimation

In this section, we give a systematic way to construct kernel functions for impulse response estimation based on OBFs that are capable of describing a wide range of dynamical properties. This results in a tailored made RKHS improving the accuracy of the model estimates by exploiting a better bias/variance trade-off compared to existing kernels.

6.1 System theory perspective

Starting from (13) and remembering (21), it follows that the optimal kernel in terms of the OBFs sequence $\Phi = \{\phi_k\}_{k=1}^{\infty}$ is given by:

$$\begin{aligned} K(i,j) = \mathrm{g}(i)\mathrm{g}(j) &= \sum_{k=1}^{\infty} c_k\phi_k(i)\sum_{l=1}^{\infty} c_l\phi_l(j) \\ &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} c_k c_l \phi_k(i)\phi_l(j). \end{aligned}$$

In the Bayesian setting, g is assumed to be a particular realization of a Gaussian random process. This corresponds to the assumption that the expansion coefficients, i.e., $\{c_k\}_{k=1}^{\infty}$ is a sequence of independent random variables with zero-mean and variance ς_k^2, i.e., $c_k \sim \mathscr{N}(0, \varsigma_k^2)$, then, by taking the expectation, we have

$$K(i,j) = E\left\{\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} c_k c_l \phi_k(i)\phi_l(j)\right\} = \sum_{k=1}^{\infty} \varsigma_k^2 \phi_k(i)\phi_k(j).$$

It is well-known that the expansion coefficients $\{c_k\}_{k=1}^{\infty}$ satisfy $\sum_{k=1}^{\infty} |c_k|^2 < \infty$, i.e., $\{c_k\}_{k=1}^{\infty} \in \mathcal{R}\ell_2(\mathbb{N})$. $\mathcal{R}\ell_2(\mathbb{N})$ is a rich space that contains the impulse responses of possibly infinite-dimensional and time-varying systems. However, we are interested only in finite-dimensional LTI systems with impulse responses that belong to $\mathcal{R}\ell_1(\mathbb{N})$, hence the expansion coefficients must satisfy a more restrictive condition $\sum_{k=1}^{\infty} |c_k| < \infty$. One possible way to impose such behavior in the kernel definition is to enforce $\{\varsigma_k^2\}$ to decay exponentially to zero. This will become more clear in the sequel.

6.2 Machine learning perspective

Given a sequence of OBFs $\Phi = \{\phi_k\}_{k=1}^{\infty}$, it is shown that these basis span an RKHS with the reproducing kernel given in (15). Since Φ is an orthonormal basis in $\mathcal{R}\ell_2(\mathbb{N})$, from Proposition 2, it comes that $K_\Phi(i,j) = \delta_{ij}$. If the system to be identified is stable, this kernel will perform poorly (this coincides with the conclusion in [21, Section V]): in fact, the optimal structure (21) suggests that the diagonal elements of the kernel should decay to zero, instead of being constant. In addition, the off-diagonal elements should be different from zero. Also the Bayesian interpretation of regularization, as described, e.g., in [25, subsection 4.3], supports the same conclusions from a Bayesian perspective. The estimator (17) can in fact be seen as the minimum variance estimator of the impulse response when the latter is a zero-mean Gaussian process, independent of the noise, with covariance proportional to K. When (15) is adopted, g becomes proportional to a stationary white noise. But the variability of a stable impulse response is expected to decay to zero as time progresses. Hence, (15) defines a kernel which is not stable according to the following definition (which extends to the discrete-time case the one contained in [25, Section 13]):

Definition 6 [Stable RKHS]. *Let $\mathscr{H}_K$ be the RKHS of real-valued functions with domain $\mathbb{N}$ induced by a kernel K. Then, $\mathscr{H}_K$ is said to be a stable RKHS, and the associated K is called stable, if $\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N})$.* □

Based on the isomorphism explained in Section 4, the kernel defined by the OBFs, i.e., Φ, leads to an RKHS as a hypothesis space given by $\mathcal{R}\ell_2(\mathbb{N})$. However, $\mathcal{R}\ell_2(\mathbb{N}) \not\subset \mathcal{R}\ell_1(\mathbb{N})$, hence the kernel is not stable. The following proposition provides a sufficient condition for a kernel to be stable. The proof is omitted since it is derived from the results contained in [36] following the same arguments as in [25, Section 13]. In particular, one can first think of the function domain as $\mathbb{N}$ equipped with a counting measure. Then, the rationale in [25, Section 13] holds by replacing integrals with infinite sums.

Proposition 3 [Sufficient condition for kernel stability]. *Let $\mathscr{H}_K$ be the RKHS on $\mathbb{N}$ induced by a real reproducing kernel K. Then,*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |K(i,j)| < \infty \implies \mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N}). \quad \square$$

In view of the above results from both machine learning and system theory perspectives, to include the stability constraint, the approach proposed in this paper is to consider the following kernel construction

$$K_\Phi^{\mathrm{s}}(i,j) = \beta_{\mathrm{s}} \sum_{k=1}^{\infty} r_k(\beta_{\mathrm{d}})\phi_k(i)\phi_k(j), \tag{23}$$

where $r_k(\beta_{\mathrm{d}}) \geq 0$ converges to zero as $k \to \infty$, β_{d} is considered to be a hyperparameter that determines the decay rate of the expansion (23). The decay term, i.e., $r_k(\beta_{\mathrm{d}})$, with β_{d} tuned by marginal likelihood optimization acts as an automatic way to select the number of significant basis functions that are needed to construct the kernel. In absence of more sophisticated prior information, in many cases, monotonically decreasing weights, e.g.,

$$r_k(\beta_{\mathrm{d}}) = k^{-\beta_{\mathrm{d}}}, \quad \beta_{\mathrm{d}} > 0, \tag{24}$$

or

$$r_k(\beta_{\mathrm{d}}) = \beta_{\mathrm{d}}^{-k}, \quad \beta_{\mathrm{d}} > 1, \tag{25}$$

are effective choices. This is also supported by system theory, where it is known that the decay rate of the expansion coefficients can be always upper bounded by an exponential term [19], [37, Equation 5.17]. However, depending on the available knowledge, other parameters can be introduced in the decay term that describe more complicated shapes for the weights. Similarly, when prior information is available, the choice of the basis functions can be further restricted. This fits in the framework developed in this paper, e.g., if the number of resonance peaks is known, we can use such information to decide the number of complex pairs/real poles that should be considered for the GOBFs.

The following proposition provides information on the stability of the kernel constructed using general Takenaka-Malmquist [3] OBFs.

Proposition 4 [Stability of the OBFs based kernels]. *Consider the kernel (23) built using the general OBFs basis. Then, such a kernel is stable if* $r_k(\beta_\mathrm{d}) = k^{-\beta_\mathrm{d}}$ *and* $\beta_\mathrm{d} > 3$ *or* $r_k(\beta_\mathrm{d}) = \beta_\mathrm{d}^{-k}$ *and* $\beta_\mathrm{d} > 1$.

Proof: See Appendix 9.2. □

Next to β_d, the other hyperparameters that characterizes an OBFs kernel are: the scale factor β_s and the poles used to generate the sequence $\phi_k(\cdot)$, i.e., the poles Λ_{n_g} of the associated inner function G_b, collected in a vector β_p. All the hyperparameters β_s, β_d, β_p are collected into β. This allows to introduce an identification scheme for regularized impulse response estimation with the OBFs-based kernel (23) as summarized in Algorithm 1. For a frequency domain formulation of this kernel and the associated results see [38].

Algorithm 1 Regularized impulse response estimation with OBFs-based kernel (23)

Require: A data record $\mathcal{D}_\mathrm{N} = \{u(k), y(k)\}_{k=1}^N$

1: Estimate the noise variance σ^2 with $\hat{\sigma}^2$ using a low-bias high-order ARX or FIR model estimated with the least-squares approach.
2: Hyperparameters estimation: Solve (20) to get the empirical Bayes estimate $\hat{\beta}$ for $\beta = \left[\beta_\mathrm{s}\ \beta_\mathrm{d}\ \beta_\mathrm{p}^\top\right]^\top$.
3: Impulse response estimation: With $\beta = \hat{\beta}$ and $\mu = \hat{\sigma}^2$, the estimate of the impulse response is computed via (19b).
4: **return** Estimated impulse response $\hat{\theta}$.

[3] We prove the stability under a general class of OBFs and hence the results hold for the special cases, e.g., GOBFs, Laguerre and Kautz basis.

6.3 Regularized OBFs expansion estimation

In [21], a regularization-based estimation of OBFs expansions (ROBFs) has been investigated. More specifically, an $n_\mathrm{b}-$truncated version of the OBFs model structure (13) has been considered:

$$y(t) = \sum_{k=1}^{n_\mathrm{b}} c_k(\phi_k \circledast u)(t) + v(t). \tag{26}$$

As an extension of the Bayesian impulse response identification, the expansion coefficients $c = [c_1 \cdots c_{n_\mathrm{b}}]^\top$ are estimated by minimizing the following modification of the ReLS criterion (19a)

$$\begin{aligned}\hat{c} &= \operatorname*{argmin}_{c\in\mathbb{R}^{n_\mathrm{b}}} \|Y - \Gamma(\beta_\mathrm{p})c\|_2^2 + \sigma^2 c^\top \mathbf{K}_\mathrm{c}^{-1}(\beta_\mathrm{c})c,\\ &= \operatorname*{argmin}_{c\in\mathbb{R}^{n_\mathrm{b}}} \sum_{t=1}^{N}\left(y(t) - \sum_{k=1}^{n_\mathrm{b}} c_k(\phi_k \circledast u)(t)\right)^2 + \sigma^2 c^\top \mathbf{K}_\mathrm{c}^{-1}(\beta_\mathrm{c})c,\end{aligned} \tag{27}$$

where β_p is the vector that contains the generating poles for the OBFs, $\mathbf{K}_\mathrm{c}(\beta_\mathrm{c})$ is the regularization matrix on the expansion coefficients $\{c_k\}_{k=1}^{n_\mathrm{b}}$, β_c is the vector that contains the hyperparameters associated with $\mathbf{K}_\mathrm{c}$ that describe the behaviour of the expansion coefficients, and

$$\Gamma(\beta_\mathrm{p}) = \left[\gamma^\top(1)\ \cdots\ \gamma^\top(N)\right]^\top$$

is the regression matrix with

$$\gamma(k) = \left[(\phi_1 \circledast u)(k)\ \cdots\ (\phi_{n_\mathrm{b}} \circledast u)(k)\right].$$

The expansion coefficients are assumed to be absolutely summable, according to the stability assumption on the data-generating system, hence kernels such as the TC, DC and SS can be used to regularize the estimation of these coefficients, i.e., to construct $\mathbf{K}_\mathrm{c}$. Moreover, the generating poles β_p of the OBFs are considered as additional hyperparameters and can be estimated by the empirical Bayes method besides of other hyperparameters, like β_c, that parameterize the kernel, used to describe the distribution of the expansion coefficients.

Furthermore, it has been shown in [21, Section V] that regularized impulse response estimation with the OBFs-based kernel K_Φ (15), i.e., the unstable formulation, is a special ill-defined case of the ROBFs estimation problem (27) pointing out that using OBFs in defining kernel functions in the time-domain in a naive way is not advisable. More specifically, when adopting (15), the regularized impulse response estimation with (15) becomes equivalent to the Ridge regression of c, i.e., $\mathbf{K}_\mathrm{c} = I_{n_\mathrm{b}}$ in (27), which cannot guarantee the absolute convergence of c. As a further extension of this result, in the sequel, the connection of the impulse response estimation with the stable OBFs-based kernel K_Φ^{s} defined in (23) and the ROBFs approach is shown. By considering an $n_{\mathrm{b}}-$truncated kernel representation[4] of K_Φ^{s}, i.e., $K_\Phi^{\mathrm{s}}(i,j) = \beta_{\mathrm{s}}\sum_{k=1}^{n_{\mathrm{b}}} r_k(\beta_{\mathrm{d}})\phi_k(i)\phi_k(j)$, the associated RKHS $\mathscr{H}_{K_\Phi^{\mathrm{s}}}$ can be written as

$$\begin{aligned}\mathscr{H}_{K_\Phi^{\mathrm{s}}} &= \mathrm{Span}\{\phi_1,\phi_2,\dots,\phi_{n_{\mathrm{b}}}\}\\ &= \left\{ \mathrm{g} \mid \mathrm{g}(t) = \sum_{k=1}^{n_{\mathrm{b}}} c_k\phi_k(t), c_k\in\mathbb{R}\right\},\end{aligned} \tag{28}$$

with

$$c_k = \langle \mathrm{g},\phi_k\rangle_{K_\Phi^{\mathrm{s}}},$$

and

$$\|\mathrm{g}\|^2_{K_\Phi^{\mathrm{s}}} = \sum_{k=1}^{n_{\mathrm{b}}} c_k^2/r_k(\beta_{\mathrm{d}}) = c^\top \mathbf{K}_{\mathrm{c}}^{-1}(\beta_{\mathrm{c}})c,$$

where $c = [c_1 \dots c_{n_{\mathrm{b}}}]^\top$ and $[\mathbf{K}_{\mathrm{c}}]_{ij} = r_j(\beta_{\mathrm{d}})\delta_{ij}$.

As $\mathrm{g}(\cdot) = \sum_{k=1}^{n_{\mathrm{b}}} c_k\phi_k(\cdot)$ and $\|\mathrm{g}\|^2_{K_\Phi^{\mathrm{s}}} = c^\top\mathbf{K}_{\mathrm{c}}^{-1}(\beta_{\mathrm{c}})c$, (17) can be written as

$$\hat{c}=\underset{c\in\mathbb{R}^{n_{\mathrm{b}}}}{\mathrm{argmin}} \sum_{t=1}^{N}\left(y(t)-\sum_{k=1}^{n_{\mathrm{b}}} c_k(\phi_k\circledast u)(t)\right)^2+\sigma^2 c^\top\mathbf{K}_{\mathrm{c}}^{-1}(\beta_{\mathrm{c}})c. \tag{29}$$

This gives that (29) is identical with (27). Although they are identical from the optimization point of view with different parameterization of the solution, conceptually they are significantly different. The approach in [21] utilizes the Bayesian approach to regularize the estimation of the expansion coefficients. On the other hand, the approach presented in this paper uses the OBFs to construct a kernel function that results in a stable RKHS directly in the time-domain, which can be used for impulse response estimation and provides a better understanding of that space. Both approaches consider the generating poles as hyperparameters and tune them with marginal likelihood maximization. To clarify the connection between RFIR with OBFs based kernels (RFIR-OBFs) and ROBFs, see Table 1.

6.4 *OBFs-based kernels with Laguerre and Kautz basis*

As an illustration of the kernel construction mechanism, two special cases of GOBFs defined in (9) are considered, namely, 2-*parameter Kautz functions* with $n_{\mathrm{g}} = 2$, and *Laguerre functions* with $n_{\mathrm{g}} = 1$ [19]. Laguerre basis are defined as

$$\psi_k(z) = \frac{\sqrt{1-\xi^2}}{z-\xi}\left(\frac{1-\xi z}{z-\xi}\right)^{k-1}, \qquad \xi\in(-1,1), \tag{30}$$

where the parameter ξ is known as the Laguerre parameter or generating pole. The impulse response of Laguerre basis functions exhibit an exponential decay as shown in Fig. 1 for $\xi = 0.5$. However, Laguerre functions construction do not allow the use of complex poles, hence, they are less suitable to capture oscillatory response. In that case, two-parameter Kautz basis functions result in a more appropriate structure. The two-parameter Kautz basis are the set of orthonormal functions

$$\begin{aligned}\psi_{2k-1} &= \frac{\sqrt{1-c^2}(z-b)}{z^2+b(c-1)z-c}\left(\frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}\right)^{k-1}\\ \psi_{2k} &= \frac{\sqrt{(1-c^2)(1-b^2)}}{z^2+b(c-1)z-c}\left(\frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}\right)^{k-1},\end{aligned} \tag{31}$$

where $b, c\in(-1,1)$. Note that (31) corresponds to a repeated complex pair $\xi,\xi^*\in\mathbb{D}$, where $b = (\xi+\xi^*)/(1+\xi\xi^*)$ and $c = -\xi\xi^*$.

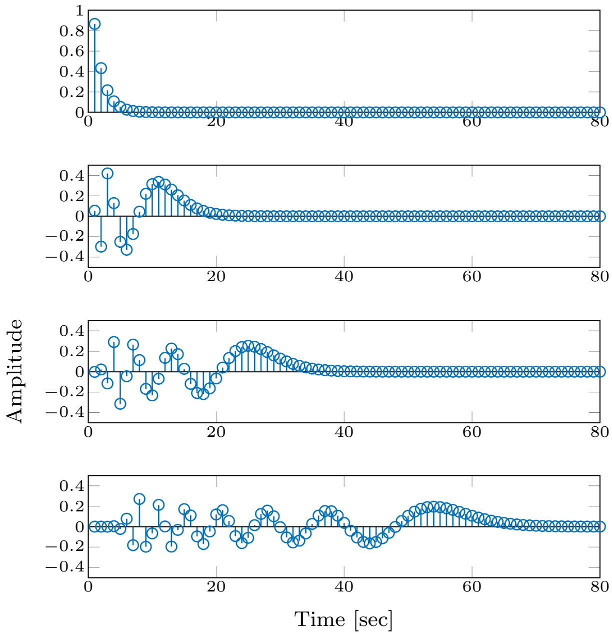


Fig. 1. Laguerre basis functions, ψ_1, ψ_5, ψ_{10}, ψ_{20} for $\xi = 0.5$.

6.5 *Hyperparameter estimation and computational complexity*

In case of the OBFs-based kernel defined in (23), the hyperparameters that are need to be estimated from data are the scaling parameter β_{s}, decay parameter β_{d} and the generating poles collected in β_{p}. Note that in case of a Laguerre-based kernel, only one real pole, i.e., ξ in (30), is needed to generate the full sequence of basis and for

[4] If we consider that $\hat{\mathrm{g}}$ to be the estimate with the infinite kernel representation, i.e., $n_{\mathrm{b}} = \infty$ and $\hat{\mathrm{g}}_{n_{\mathrm{b}}}$ is the estimate with the $n_{\mathrm{b}}-$truncated representation of the kernel, then, the following result holds $\lim_{n_{\mathrm{b}}\to\infty}\|\hat{\mathrm{g}}-\hat{\mathrm{g}}_{n_{\mathrm{b}}}\|_{K_\Phi^{\mathrm{s}}} = 0$, see [39, Theorem 7].

Table 1
RFIR-OBFs vs. ROBFs estimation approaches.

	RFIR-OBFs	ROBFs
Model structure	$G_0(q) = \sum_{k=1}^{\infty} \mathrm{g}_k q^{-k}$	$\mathrm{g}(t) = \sum_{k=1}^{\infty} c_k \phi_k(t)$
Expansion basis	$\{q^{-k}\}_{k=1}^{\infty}$	$\{\phi_k\}_{k=1}^{\infty}$
Model parameters	$\{\mathrm{g}_k\}_{k=1}^{\infty}$	$\{\mathrm{c}_k\}_{k=1}^{\infty}$
Optimization problem	(19a)	(27)
Utilized kernel	g is regularized using K_{Φ}^{s} defined in (23)	c is regularized by K_{c}, which can be any of the kernels used for impulse response estimation, e.g., (22)
Hyperparameters	β_{s}: scaling parameter of the kernel β_{d}: decay rate parameter β_{p}: generating poles of the OBFs	β_{c}: hyperparameters of K_{c} β_{p}: generating poles of the OBFs
Role of OBFs	Construct the kernel	Filter the input to construct the regression matrix Γ

a Kautz-based kernel only two conjugate complex poles, defined by $\mathscr{b}$ and $\mathscr{c}$ in (31), are needed to generate that sequence. Estimation of these hyperparameters following the empirical Bayes approach can be accomplished by solving the optimization (20).

The algorithm of regularized impulse response estimation consists of two main steps [40]:

(1) Hyperparameters estimation: This step involves the minimization of the cost function in (20), which is a nonlinear optimization problem for which a single evaluation of the cost function is $O(N^3)$.
(2) Impulse response estimation: The computational complexity of this step is $O(N^3)$.

In [41], a new computational strategy has been proposed which may reduce significantly the computational load and extend the practical applicability of this methodology to large-scale scenarios. The proposed algorithm [41, Algorithm 2] is developed for SS kernels and exploits the spectral decomposition of these kernels [12]. With this approach, the computational complexity now scales as $O(l^3)$, where l is the number of the used eigenfunctions. Moreover, it can effectively compute the marginal likelihood with $O(N^2 l)$ for a single evaluation of the cost, see [41, Table 1]. This algorithm is directly applicable for kernels that exhibit a spectral decomposition, like our OBFs-based kernel. Moreover, the effectiveness of this algorithm depends on if the to-be-estimated impulse response can be approximated with a few number of eigenfunctions [41, Page 5]. This motivates also the use of OBFs as eigenfunctions, offering a wide range of basis, which if properly chosen, can achieve a high approximation accuracy with only a few active basis in the expansion.

7 Numerical Simulation

In this section, the performance of the proposed OBFs based kernels in the considered Bayesian identification setting is assessed via Monte Carlo based simulation studies using randomly generated discrete-time LTI systems.

7.1 Simulation studies

By using the setting of (1) as the data-generating system, five simulation studies have been accomplished for the following scenarios:

(1) `S1D1`: fast systems, $\mathcal{D}_{\mathrm{N}}$ with $N = 500$, *Signal-to-Noise Ratio* (SNR) = 10dB.
(2) `S1D2`: fast systems, $\mathcal{D}_{\mathrm{N}}$ with $N = 375$, SNR = 1dB.
(3) `S2D1`: slow systems, $\mathcal{D}_{\mathrm{N}}$ with $N = 500$, SNR = 10dB.
(4) `S2D2`: slow systems, $\mathcal{D}_{\mathrm{N}}$ with $N = 375$, SNR = 1dB.
(5) `S3`: oscillatory systems, $\mathcal{D}_{\mathrm{N}}$ with $N = 400$, SNR = 10dB.

Each Scenario 1) to 4) corresponds to 100 randomly generated (by the `drss` Matlab function) 30^{th} order discrete-time SISO LTI systems as G_0. The fast systems have all poles inside $0.95\mathbb{D}$ and the slow systems have at least one pole in the ring $\mathbb{D} - 0.95\mathbb{D}$, i.e., slow dominant poles. These systems are used to generate data sets for a white u, with $u(t) \sim \mathcal{N}(0, 1)$ and v being additive independent white Gaussian noise. The variance of v is set such that the SNR, i.e.,

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_{k=1}^{N} \tilde{y}^2(k)}{\sum_{k=1}^{N} v^2(k)} \right)$$

where $\tilde{y}(k)$ denotes the noise-free system output, i.e., $\tilde{y}(k) = G_0(q)u(k)$, is SNR = 1dB or 10dB for various Monte Carlo experiments. Whereas, Scenario 5) has been generated as reported in [42], but with only one dominant complex conjugate pole pair.

7.2 *Identification setting*

In all of the five scenarios, we estimate FIR models, i.e., the $n-$truncated impulse responses of (3) or equivalently θ in (18), with $n = 125$ using the following approaches:

(1) RFIR-TC: regularized impulse response estimation where the impulse response coefficients are estimated by solving (17) and regularized with the TC kernel (22c).
(2) RFIR-OBF-L, -K or -G: regularized impulse response estimation where the impulse response coefficients are estimated by solving (17) and regularized with the OBFs based kernel (23) with three different basis functions, i.e., Laguerre with one real pole, Kautz with one complex conjugate pair or GOBFs with two real poles where the generating inner function (8) is a 2^{nd} order one.
(3) ROBF-L, -K or -G: regularized OBFs expansion estimation, where the expansion coefficients are estimated by solving (27) and regularized with the diagonal kernel (22a), i.e., DI kernel, with three different basis functions, i.e., Laguerre with one real pole, Kautz with one complex conjugate pair or GOBFs with two real poles.

Note that, two different scenarios are considered for the number of basis functions to construct the OBFs based kernel and the OBFs model structure, i.e., $n_{\mathrm{b}} = 40$ and $n_{\mathrm{b}} = 100$. This is provided to show the effectiveness of the presented approach to control the flexibility offered by even a large number of basis. The performance index that is used to measure the quality of the impulse response estimation is the *Best Fit Rate* (BFR) of the estimated impulse response $\hat{\mathrm{g}}_k$

$$\text{BFR}{=}100\%. \left(1-\sqrt{\frac{\sum_{k=1}^{125}|\mathrm{g}_k-\hat{\mathrm{g}}_k|^2}{\sum_{k=1}^{125}|\mathrm{g}_k-\bar{\mathrm{g}}|^2}}\right), \bar{\mathrm{g}} = \frac{1}{125}\sum_{k=1}^{125}\mathrm{g}_k,$$

where g_k are the true coefficient values. The hyperparameters have been estimated by the discussed marginal likelihood maximization, i.e., (20). Note that, in this work the approach proposed in [40], i.e., QR factorization, is employed to solve the optimization problem to tune the unknown hyperparameters and to estimate the unknown impulse response.

7.3 *Identification results*

The average model fits over the considered five data sets estimated with TC kernel are reported in Table 2, whereas the average model fits in case of (RFIR-OBF)/ROBF-L, (RFIR-OBF)/ROBF-K and (RFIR-OBF)/ROBF-G are reported in Table 3, 4 and 5, respectively. Moreover, in each case, the considered scenarios of different number of basis functions, i.e., 40 basis and 100 basis, are given [5] . The highest average model fit over the RFIR-OBF alternatives or the ROBF alternatives is highlighted in bold. While both weightings, (24) and (25), are implemented for RFIR-OBF estimators, their performance was roughly the same . Hence, they are not distinguished in the provided results.

Table 2
Average of the BFR of the estimated FIR models with TC kernel.

RFIR-TC	S1D1	S1D2	S2D1	S2D2	S3
	90.82	77.25	84.08	63.61	86.01

Table 3
Average of the BFR of the estimated FIR models with Laguerre basis.

RFIR-OBF-L	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.85	78.97	85.71	67.88	85.31
100 basis	91.90	79.20	87.86	68.88	87.67
ROBF-L	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.90	78.92	85.70	69.28	83.41
100 basis	91.90	79.18	88.13	69.73	88.93

Table 4
Average of the BFR of the estimated FIR models with Kautz basis.

RFIR-OBF-K	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.91	79.08	87.29	70.35	93.40
100 basis	91.92	78.99	88.44	70.83	**93.70**
ROBF-K	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.89	78.64	87.50	70.52	**93.72**
100 basis	91.98	78.93	88.74	71.33	93.64

For illustration, the distributions of the model fits over the five data sets, with TC and the estimates with RFIR-OBF, ROBF, which are highlighted in bold, are shown by boxplots in Fig. 2 to 4.

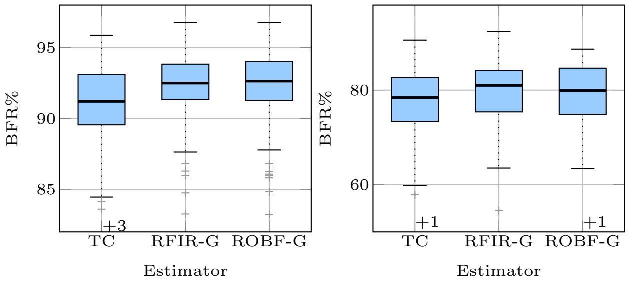
Discussion

Next, we give some insights to the obtained results.

[5] Note that, when K_{Φ}^{s} is constructed based on n_{b} basis functions, it can be easily seen that $\text{rank}(\mathbf{K}) \le n_{\mathrm{b}}$, where $\mathbf{K}$ is the resulting kernel matrix. This means that with $n_{\mathrm{b}} = 40$ or 100, $\mathbf{K}$ will be positive semidefinite. However, full rankness of K is not necessary regarding the solution $\hat{\theta}$ in terms of (19b) or the Bayesian interpretation of the problem, but it makes the analogy with regularization more complicated than (19a).

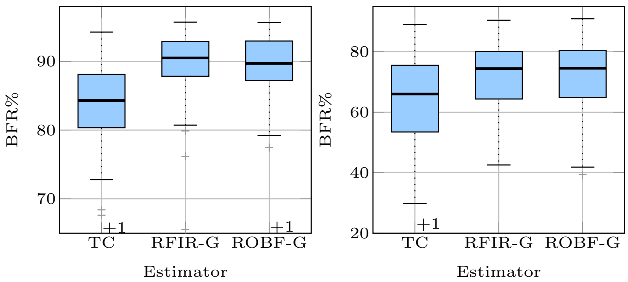
Table 5
Average of the BFR of the estimated FIR models with GOBFs basis.

RFIR-OBF-G	S1D1	S1D2	S2D1	S2D2	S3
40 basis	92.07	**79.54**	87.74	69.82	83.53
100 basis	**92.21**	79.52	**89.26**	**71.10**	88.76
ROBF-G	S1D1	S1D2	S2D1	S2D2	S3
40 basis	92.18	**79.27**	87.37	71.12	83.40
100 basis	**92.31**	79.18	**89.42**	**72.38**	89.20



(a) S1D1 (b) S1D2

Fig. 2. Boxplot for model fits over S1D1, S1D2. The shown estimators are those highlighted in bold in the tables. Note that RFIR-TC and RFIR-OBF-G are denoted as TC and RFIR-G, respectively.



(a) S2D1 (b) S2D2

Fig. 3. Boxplot for model fits over S2D1, S2D2. The shown estimators are those highlighted in bold in the tables. Note that RFIR-TC and RFIR-OBF-G are denoted as TC and RFIR-G, respectively.

(1) In general, RFIR-OBF with all its alternatives performs better than RFIR-TC, because RFIR-OBF estimators employ kernels that are capable to capture dynamical properties, e.g., resonance behaviour, damping, etc., via the generating poles of the OBFs, rather than only focusing on smoothness and stability.
(2) For resonating systems, i.e., S3: RFIR-OBF-L/ROBF-L have difficulties. It is well-known that for a system with resonance behaviour, a long Laguerre expansion is needed to get good accuracy. This can be easily seen from the poor performance in case of 40 basis compared to the TC kernel. However, when increasing the number of basis to

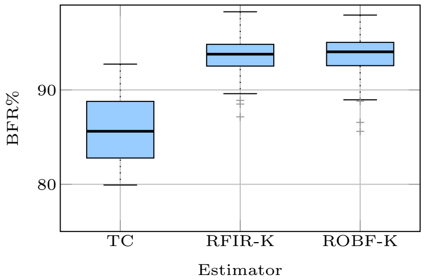


Fig. 4. Boxplot for model fits over S3. The shown estimators are those highlighted in bold in the tables. Note that RFIR-TC and RFIR-OBF- K are denoted as TC and RFIR-K, respectively

100, the results improve a lot due to the employed long expansion and the regularization that keeps the variance low.
(3) In case of S3, Kautz basis perform significantly better compared to other estimators. This is due to the fact that Kautz basis is generated by two repeated conjugate complex poles that are tuned by marginal likelihood optimization and hence can capture the true dominant poles.
(4) Due to the regularization acting on the estimation problem, in most of the cases we gain from increasing the number of basis functions, which is not the case in classical identification due to the increased variance resulting from increasing the number of the expansion coefficients.
(5) For slow systems, i.e., S2D1 and S2D2: RFIR-OBF estimators show a significant improvement over the TC kernel, especially RFIR-OBF-G, which gives the best performance for the first four data sets, i.e., S1D1, . . ., S2D2. Since, it is known that for slow systems, if the basis functions are properly chosen, the OBFs offer a more compact model structure which results in a better RKHS as a hypothesis space.
(6) The results of RFIR-OBF and ROBF are very close to each other due to the equivalence provided in Section 6.3. Note that the difference in results is due to numerical issues, i.e., in case of RFIR-OBF we directly estimate 125 length impulse response, whereas in case of ROBF we only estimate as many expansion coefficients as the number of basis functions and then compute a 125 length impulse response of the estimated OBFs model.

8 Conclusion

In this paper, a systematic construction mechanism of stable time-domain kernels based on OBFs for impulse response estimation of LTI systems has been proposed. Two proposed weightings of the OBFs are introduced as decay terms, guaranteeing stability of impulse responses in the associated hypothesis space. The resulting OBFs based kernel is parameterized in terms of hyperparameters as the scale factor, the decay parameter, and the poles used to generate the OBFs sequence. Tuning of these hyperparameters is performed by empirical Bayes, maximizing the marginal likelihood. The flexible parameterization and marginal likelihood maximization of the estimated IIR enable to overcome both the difficulty of selecting the number of basis functions to be introduced in the kernel and the difficult task of choosing the poles of the OBFs. Three special cases have been illustrated, namely, Laguerre, Kautz and GOBFs based kernel structures. Their performance has been evaluated and compared with the TC kernel by Monte-Carlo simulations. Results show that the novel kernels perform well compared with the TC kernel especially for slow systems. Moreover, OBFs based kernels with Kautz basis perform significantly better on resonating systems, pointing out the capability of the proposed approach to describe a wide variety of dynamical systems properties, progressing the achievable quality of the model estimates via the empirical Bayes methods. Interesting topics to be investigated in the future work is to extend the applicability of the presented kernel formulation for model classes capable to represent varying dynamical phenomena, i.e., *Linear Parameter-Varying* (LPV) and *Linear Time-Varying* (LTV) systems.

9 Appendix

9.1 Appendix A: Proof of Proposition 1

By following the same line of reasoning as in [23]: the $\mathscr{L}_\infty(\mathbb{T})-$norm of the k^{th} Takenaka-Malmquist basis ψ_k is uniformly bounded

$$\sup_\omega \left| \frac{\sqrt{1-|\xi_k|^2}}{e^{j\omega}-\xi_k} \prod_{i=1}^{k-1} \frac{1-\xi_i^* e^{j\omega}}{e^{j\omega}-\xi_i} \right| \leq \frac{\sqrt{1-|\xi_k|^2}}{1-|\xi_k|}.$$

Based on the fact that the $\ell_1(\mathbb{N})$-norm of the impulse response of a k^{th}-order stable system is less than twice the nuclear norm of the associated Hankel operator [43, Section 2], and that nuclear norm is less than k times the $\mathscr{L}_\infty(\mathbb{T})$-norm [43, Theorem 2.1]:

$$\|\phi_k\|_{\ell_1} \leq k \cdot 2 \frac{\sqrt{1-|\xi_k|^2}}{1-|\xi_k|}.$$

Let

$$\kappa = \sup_{\xi \in \{\xi_k\}_{k=1}^\infty} \frac{\sqrt{1-|\xi|^2}}{1-|\xi|}.$$

Accordingly,

$$\|\phi_k\|_{\ell_1} \leq k \cdot 2\kappa.$$

9.2 Appendix B: Proof of Proposition 4

The proof can be accomplished by the application of Proposition 3 and Proposition 1. From Proposition 3, one have to check that:

$$\sum_{i=1}^\infty \sum_{j=1}^\infty |K_\Phi^{\text{s}}(i,j)| = \sum_{i=1}^\infty \sum_{j=1}^\infty \left| \sum_{k=1}^\infty r_k(\beta_{\text{d}}) \phi_k(i) \phi_k(j) \right| < \infty.$$

To this end:

$$\begin{aligned}
&\sum_{i=1}^\infty \sum_{j=1}^\infty \left| \sum_{k=1}^\infty r_k(\beta_{\text{d}}) \phi_k(i) \phi_k(j) \right| \\
&\leq \sum_{i=1}^\infty \sum_{j=1}^\infty \sum_{k=1}^\infty r_k(\beta_{\text{d}}) |\phi_k(i)| |\phi_k(j)| \\
&= \sum_{k=1}^\infty r_k(\beta_{\text{d}}) \sum_{i=1}^\infty \sum_{j=1}^\infty |\phi_k(i)| |\phi_k(j)| \\
&= \sum_{k=1}^\infty r_k(\beta_{\text{d}}) \underbrace{\sum_{i=1}^\infty |\phi_k(i)|}_{\|\phi_k\|_{\ell_1}} \underbrace{\sum_{j=1}^\infty |\phi_k(j)|}_{\|\phi_k\|_{\ell_1}} \leq (2\kappa)^2 \sum_{k=1}^\infty k^2 r_k(\beta_{\text{d}}),
\end{aligned}$$

where the last equation is obtained by the bound provided in Proposition 1. Hence, $\sum_{k=1}^\infty k^2 r_k(\beta_{\text{d}}) < \infty$ should be satisfied to guarantee the stability of the kernel. In case of (24): $\sum_{k=1}^\infty k^2 r_k(\beta_{\text{d}}) = \sum_{k=1}^\infty k^{(2-\beta_{\text{d}})}$, and with $\beta_{\text{d}} > 3$, the series will have a convergent sum, which guarantees the stability of the kernel. In case of (25): $\sum_{k=1}^\infty k^2 r_k(\beta_{\text{d}}) = \sum_{k=1}^\infty k^2 \beta_{\text{d}}^{-k}$, and with $\beta_{\text{d}} > 1$, the series will have a convergent sum, which guarantees the stability of the kernel.

References

[1] L. Ljung, *System Identification, Theory for the User.* Upper Saddle River: Prentince Hall, 1999.

[2] T. Söderström and P. Stoica, *Instrumental Variable Methods for System Identification.* Springer-Verlag, New York, 1983.

[3] R. Pintelon and J. Schoukens, *System Identification, A Frequency Domain Approach.* IEEE Press, Wiley, 2012.

[4] A. Chiuso, "The role of vector autoregressive modeling in predictor-based subspace identification," *Automatica*, vol. 43, no. 6, pp. 1034–1048, 2007.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[6] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[7] 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.

[8] R. Tibshirani, "Regression shrinkage and selection with the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[9] H. Hjalmarsson, J. S. Welsh, and C. R. Rojas, "Identification of Box-Jenkins models using structured ARX models and nuclear norm relaxation," in *Proc. of the 16th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012, pp. 322–327.

[10] L. Breiman, "Better subset regression using the nonnegative garotte," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[11] C. R. Rojas, R. Tóth, and H. Hjalmarsson, "Sparse estimation of polynomial and rational dynamic models," *Special issue, IEEE Trans. on Automatic Control*, vol. 59, pp. 2962–9286, 2014.

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

[13] 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.

[14] G. Pillonetto and G. De Nicolao, "Kernel selection in linear system identification part I: A Gaussian process perspective," in *Proc. of the 50th IEEE Conf. on Decision and Control and the European Control Conf.*, Orlando, FL, USA, Dec. 2011, pp. 4318–4325.

[15] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.* The MIT Press, 2006.

[16] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[17] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis.* CRC Press, 2000.

[18] G. Pillonetto and A. Chiuso, "Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator," *Automatica*, vol. 58, pp. 106–117, 2015.

[19] P. S. C. Heuberger, P. M. J. Van den Hof, and B. Wahlberg, *Modeling and Identification with Rational Orthonormal Basis Functions.* Springer-Verlag, 2005.

[20] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra, "A generalized orthonormal basis for linear dynamical systems," *IEEE Trans. on Automatic Control*, vol. 40, no. 3, pp. 451–465, 1995.

[21] T. Chen and L. Ljung, "Regularized system identification using orthonormal basis functions," in *Proc. of the European Control Conf.*, Linz, Austria, July 2015, pp. 1291 – 1296.

[22] M. A. H. Darwish, G. Pillonetto, and R. Tóth, "Perspectives of orthonormal basis functions based kernels in Bayesian system identification," in *Proc. of the 54th IEEE Conf. on Decision and Control*, Osaka, Japan, Dec. 2015, pp. 2713–2718.

[23] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. on Automatic Control*, vol. 36, no. 5, pp. 551–562, 1991.

[24] ——, "System identification using Kautz models," *IEEE Trans. on Automatic Control*, vol. 39, no. 6, pp. 1276–1282, 1994.

[25] 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.

[26] G. Wahba, *Spline models for observational data.* Philadelphia: Siam, 1990.

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

[28] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Trans. of the Royal Society of London, Series A*, vol. 209, pp. 415–446, 1909.

[29] R. K. Pearson, *Discrete-time dynamic models.* Oxford University Press on Demand, 1999.

[30] T. Oliveira e Silva, "A n-width result for the generalized orthonormal basis function model," in *Proc. of the 13th IFAC World Congress*, Sydney, Australia, July 1996, pp. 375–380.

[31] R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof, "Asymptotically optimal orthonormal basis functions for LPV system identification," *Automatica*, vol. 45, no. 6, pp. 1359–1370, 2009.

[32] T. Oliveira e Silva, "Rational orthonormal functions on the unit circle and on the imaginary axis, with applications in system identification," Automatic Control Group, Royal Institute of Technology, Stockholm, Sweden, Tech. Rep. ver. 1.0 a, 1995.

[33] A. Tikhonov and V. Arsenin, *Solutions of ill-posed problems.* Washington DC: Winston, Wiley, 1977.

[34] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. on Automatic Control*, vol. 37, no. 7, pp. 913–928, 1992.

[35] T. Chen, G. Pillonetto, A. Chiuso, and L. Ljung, "Spectral analysis of the DC kernel for regularized system identification," in *Proc. of the 54th IEEE Conf. on Decision and Control*, Osaka, Japan, Dec. 2015, pp. 4017 – 4022.

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

[37] R. Tóth, *Modeling and Identification of Linear Parameter-Varying Systems*, ser. Lecture Notes in Control and Information Sciences, Vol. 403. Heidelberg: Springer, 2010.

[38] M. A. H. Darwish, J. Lataire, and R. Tóth, "Bayesian frequency domain identification of LTI systems with OBFs kernels," in *Proc. of the 20th IFAC World Congress*, Toulouse, France, July 2017, pp. 6412–6417.

[39] G. Pillonetto and B. Bell, "Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance," *Automatica*, vol. 43, no. 10, pp. 1698–1712, 2007.

[40] T. Chen and L. Ljung, "Implementation of algorithms for tuning parameters in regularized least squares problems in system identification," *Automatica*, vol. 49, no. 7, pp. 2213–2220, 2013.

[41] F. P. Carli, A. Chiuso, and G. Pillonetto, "Efficient algorithms for large scale linear system identification using stable spline estimators," in *Proc. of the 16th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012, pp. 119–124.

[42] A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto, "On the design of multiple kernels for nonparametric linear system identification," in *Proc. of the 53rd IEEE Conf. on Decision and Control*, Los Angeles, USA, Dec. 2014, pp. 3346–3351.

[43] K. Glover, R. F. Curtain, and J. R. Partington, "Realisation and approximation of linear infinite-dimensional systems with error bounds," *SIAM Journal on Control and Optimization*, vol. 26, no. 4, pp. 863–898, 1988.