# Regularized system identification using orthonormal basis functions

Tianshi Chen and Lennart Ljung

*Abstract*— **Most of existing results on regularized system identification focus on regularized impulse response estimation. Since the impulse response model is a special case of orthonormal basis functions, it is interesting to consider if it is possible to tackle the regularized system identification using more compact orthonormal basis functions. In this paper, we explore two possibilities. First, we construct reproducing kernel Hilbert space of impulse responses by orthonormal basis functions and then use the induced reproducing kernel for the regularized impulse response estimation. Second, we extend the regularization method from impulse response estimation to the more general orthonormal basis functions estimation. For both cases, the poles of the basis functions are treated as hyper-parameters and estimated by empirical Bayes method. Then we further show that the former is a special case of the latter, and more specifically, the former is equivalent to ridge regression of the coefficients of the orthonormal basis functions.**

## I. INTRODUCTION

In this paper, we consider the system identification problem of linear discrete-time, time-invariant and causal systems, which is described as follows:

$$y(t) = g^0 * u(t) + v(t), \quad t = 1, \cdots, N \quad (1)$$

where $t = 1, \cdots, N$ are time indices at which the measured input $u(t)$ and output $y(t)$ are collected, and uniform sampling is used and the sampling interval $T_s = 1$, $v(t)$ is the disturbance and for convenience assumed to be a zero mean white Gaussian noise, $g^0(t)$ with $t = 1, 2, \cdots$, is the impulse response, $g^0 * u(t)$ is the convolution of $g^0(t)$ and $u(t)$ evaluated at the time $t$. The goal is to estimate $g^0(t)$ as well as possible based on the collected data $\{y(t), u(t)\}_{t=1}^N$.

The traditional method to this problem is the maximum likelihood/prediction error method (ML/PEM), see e.g., [1], [2]. Since $v(t)$ is white, PEM first postulates the so-called output error (OE) model structure $G(q, \theta)$ with $\theta \in \mathbb{R}^n$:

$$y(t) = G(q, \theta)u(t) + v(t), \quad (2)$$

where $q$ is the forward-shift operator and $qu(t) = u(t+1)$ and

$$G(q, \theta) = \frac{B(q)}{F(q)}, \quad \begin{matrix} B(q) = b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b} \\ F(q) = 1 + f_1 q^{-1} + \cdots + f_{n_f} q^{-n_f} \end{matrix} \quad (3)$$

with $\theta = [b_1, \cdots, b_{n_b} f_1, \cdots, f_{n_f}]^T$ and $n = n_b + n_f$. As long as a model structure $G(q, \theta)$ is chosen, ML/PEM minimizes the prediction error to get the model estimate

$$\hat{\theta} = \arg \min_\theta \sum_{t=1}^N (y(t) - G(q, \theta)u(t))^2. \quad (4)$$

Since the disturbance $v(t)$ in (1) is modeled as a stochastic process, the estimate $\hat{\theta}$ is a random variable. Let $\hat{g}$ denote the impulse response of $G(q, \hat{\theta})$. Then the mean square error (MSE) $\mathbb{E}(\|\hat{g} - g^0\|_2^2)$ tells the quality of the estimate $\hat{\theta}$. For the chosen model structure $G(q, \theta)$, a key issue to reduce the MSE is to find the "right" model complexity: it shall be parsimonious but capable to describe the data. Traditionally, it is suggested to use the model structure selection criterion, like AIC, BIC, to find a suitable $n$, the dimension of $\theta$. However, this way may not work well for short and noisy data records.

The model structure (3) has a very general form and includes many widely used model structures as special cases. One attractive class of model structures among many others is the linear-in-parameter model structures which can considerably simplify the optimization in (4). The most well-known instance is perhaps the finite impulse response (FIR) model structure

$$G(q, \theta) = \sum_{k=1}^n g_k q^{-k}, \quad \theta = [g_1, \cdots, g_n]^T. \quad (5)$$

However, the FIR model is often criticized for its large variance error when high order FIR models have to be used to describe "slow" systems with either slow dynamics or with high sampling rate. A more compact model structure is the linear combination of basis functions:

$$G(q, \theta) = \sum_{k=1}^m g_k \bar{F}_k(q), \theta = [g_1, \cdots, g_m, f_1, \cdots, f_{n_f}]^T \quad (6)$$

where $n = m + n_f$ and $\bar{F}_k(q) = q^{k-1}/F(q)$, $k = 1, \cdots, m$ are pre-specified basis functions. The model structure (6) has attracted a lot of interests in the last two decades, see e.g., [3] and the references therein. Two widely known special cases of (6) are the Laguerre model [4] and the Kautz model [5]. The Laguerre model takes the form

$$G(q, \theta) = \sum_{k=1}^m g_k \frac{\sqrt{(1-a^2)}}{q-a} \left(\frac{1-aq}{q-a}\right)^{k-1} \quad (7)$$

$$\theta = [g_1, \cdots, g_m, a]^T, \quad |a| < 1$$

where $a$ is pole of the Laguerre model and has to be pre-specified according to the *a priori* information on the

time constant of the underlying system [5]. Since the basis functions have infinite impulse responses, there is often no problem of describing "slow" systems with relatively small number of basis functions in (6). While the use of orthonormal basis functions (6) has been discussed a lot, still open problems are

1) how to choose suitable poles for the basis function?
2) how many basis functions shall be used?

There is another way to reduce the MSE, i.e, by using regularization. However, this way has not been investigated rigorously in system identification until the seminal work [6]. Instead of trimming the model complexity of $G(q, \theta)$ in terms of $n$, it was suggested to use a well-tuned regularization to regularize the impulse response to reduce the MSE [7]. Since then the followup results in [8], [7], [9], [10], [11] and the recent survey paper [12] show that the regularized high order FIR model (or high order ARX model) can lead to good model estimates in terms of accuracy and robustness. In this paper, we will make use of orthonormal basis functions for the regularized system identification and we will consider two cases. First, we construct reproducing kernel Hilbert space of impulse responses by orthonormal basis functions and then use the induced reproducing kernel for the regularized impulse response estimation. Second, we extend the regularization method from impulse response estimation to the more general orthonormal basis functions estimation. For both cases, the poles of the basis functions are treated as hyper-parameters and estimated by empirical Bayes method. Then we further show that the former is a special case of the latter, and more specifically, the former is equivalent to ridge regression of the coefficients of the orthonormal basis functions.

## II. REGULARIZED LEAST SQUARES METHOD

Consider a linear regression model

$$Y_N = \Phi_N \theta + V_N \qquad (8)$$

where $Y_N \in \mathbb{R}^N$ is the data, $\Phi_N \in \mathbb{R}^{N \times n}$ is the regression matrix, $\theta \in \mathbb{R}^n$ is the parameter to be estimated, and $V_N$ is the disturbance and assumed to be white Gaussian distributed as $\mathcal{N}(0, \sigma^2 I_N)$ with $I_N$ being the $N$-dimensional identity matrix. We estimate $\theta$ by minimizing the regularized least squares (RLS) criterion

$$\hat{\theta} = \arg\min_{\theta} \|Y_N - \Phi_N \theta\|_2^2 + \sigma^2 \theta^T \mathbf{K}(\alpha)^{-1} \theta \qquad (9a)$$

$$= K(\alpha) \Phi_N^T (\Phi_N \mathbf{K}(\alpha) \Phi_N^T + \sigma^2 I_N)^{-1} Y_N. \qquad (9b)$$

Here, $\mathbf{K}(\alpha) \succeq 0^1$ is called the *regularization* matrix (also often called the kernel matrix) and defined through the kernel function $K(k, j; \alpha)$ as $\mathbf{K}_{k,j}(\alpha) = K(k, j; \alpha)$, where $\alpha$ is a vector of tuning parameters and called hyper-parameter.

There are two key issues:

1) how to parameterize the kernel function $K(k, j; \alpha)$ which is often simply written as $K(\alpha)$ below?

---

[1] When $\mathbf{K}(\alpha)$ is singular, (9a) has to be interpreted in the way discussed in [9, Remark 2.1].

2) how to tune the hyper-parameter $\alpha$?

For 1), it is worth to note [7, Theorem 1] that the optimal regularization matrix in the sense of minimizing the MSE matrix of $\theta$ with respect to $\theta_0$ (the true value of $\theta$) exists and takes the form of $\mathbf{K}^{Opt} = \theta_0 \theta_0^T$. While it cannot be applied in practice, it gives a guideline to design the regularization matrix: let it mimic the behavior of $\mathbf{K}^{Opt}$. Apparently, if some prior information is known for $\theta_0$, it shall be used in the design of a suitable kernel function $K(\alpha)$.

For 2), the current most effective method is to embed the regularization in the Bayesian framework and invoke the empirical Bayes method, i.e., the marginal likelihood maximization. Assume $\theta \sim \mathcal{N}(0, \mathbf{K}(\alpha))$. Then we estimate $\alpha$ by maximizing

$$\hat{\alpha} = \arg\max_{\alpha} p(Y_N | \alpha)$$
$$= \arg\min_{\alpha} Y_N^T (\Phi_N \mathbf{K}(\alpha) \Phi_N^T$$
$$+ \sigma^2 I_N)^{-1} Y_N + \log\det(\Phi_N \mathbf{K}(\alpha) \Phi_N^T + \sigma^2 I_N) \quad (10)$$

### A. Regularized impulse response estimation

For regularized impulse response estimation, we consider the model (5) with $n = \infty$. The system (2) can then be written as a linear regression model (8) with the $i$th row of $Y_N, V_N$ and $\Phi_N$ being $y(i), v(i)$ and $\varphi(i) = [u((i - 1)), \cdots, u((i - \infty))]^T$ where the unknown inputs $u(t)$ are set to zero, and $\theta = [g_1, g_2, \cdots, ]^T \in \mathbb{R}^{\infty}$. So we can use the RLS method to estimate the impulse response. The remaining issue is the design of a suitable kernel function $K(\alpha)$. Several choices have been suggested in [6], [8], [7]. For example, the diagonal-correlated (DC) kernel and its special case, the tuned-correlated (TC) kernel are defined as:

$$\text{DC} \quad K^{dc}(k, j; \alpha) = c\lambda^{(k+j)/2} \rho^{|k-j|}, \alpha = [c\ \lambda\ \rho]^T \quad (11)$$

$$\text{TC} \quad K^{tc}(k, j; \alpha) = c\min(\lambda^k, \lambda^j), \alpha = [c\ \lambda]^T \quad (12)$$

where the TC kernel has also been introduced as the first-order stable spline (SS) kernel, see [13], [14] for discussions. In practice, we however cannot handle infinite impulse response and we have to truncate the infinite impulse response to a finite one, i.e., the FIR model. In this case, we refer this method as the regularized FIR model method in [7].

## III. REGULARIZED IMPULSE RESPONSE ESTIMATION WITH KERNEL STRUCTURE CONSTRUCTED BY ORTHONORMAL BASIS FUNCTIONS

In the following, we consider a different kernel which is constructed by use of the orthonormal basis functions. Before proceeding to the details, recall that the RLS criterion (9a) for regularized impulse response estimation has a function estimation interpretation. The RLS (9a) is equivalent to

$$\hat{\vartheta} = \arg\min_{\vartheta \in \mathcal{H}_{K(\alpha)}} \sum_{t=1}^{N} |y(t) - \vartheta * u(t)|^2 + \sigma^2 \|\vartheta\|_{\mathcal{H}_{K(\alpha)}}^2 \qquad (13)$$

where $\vartheta(t) = \theta_t$ with $\theta_t$ being the $t$th element of $\theta \in \mathbb{R}^{\infty}$ is the impulse response, and $\mathcal{H}_{K(\alpha)}$ is the reproducing kernel

Hilbert space (RKHS) induced by the kernel $K(\alpha)$. Then the RLS estimate is also the function estimate that minimizes (13) within the RKHS $\mathcal{H}_{K(\alpha)}$. This implies that when we trim the kernel $K(\alpha)$, we equivalently trim the function space where we search for the impulse response.

The above observation gives us another idea to design the kernel structure: we can first construct a RKHS space of suitable impulse responses and this space then uniquely determines a reproducing kernel according to *Moore-Aronszajn* Theorem, see e.g., [15]. Note that looking for a RKHS space of impulse responses in time domain is equivalent to looking for a RKHS space of transfer functions in frequency domain. In system identification community, the idea of approximating or expressing the transfer function of the underlying system by expanding it in terms of orthogonal basis functions have been well studied, see e.g., [3], [16], [17], [18], [19] and the references therein. It is natural to ask if the space spanned by orthogonal basis functions could be a candidate for our use. To answer this question, we have to check if this space is a RKHS, and if it is, what its reproducing kernel is. Fortunately, there are standard answers to these questions.

### A. *Transfer function space spanned by the orthonormal basis functions on the unit circle [20], [18]*

Following [20], let $\{\alpha_k\}_{k=0}^{\infty}$ with $|\alpha_k| < 1$ be an arbitrary sequence of complex numbers which may appear as numbers of finite or even infinite multiplicity. Given $\{\alpha_k\}_{k=0}^{\infty}$, a system of functions $\{\phi_k(e^{i\omega})\}_{k=0}^{\infty}$ is defined as

$$\phi_0(e^{i\omega}) = \frac{\sqrt{1 - |\alpha_0|^2}}{1 - \overline{\alpha_0}e^{i\omega}},$$

$$\phi_j(e^{i\omega}) = \frac{\sqrt{1 - |\alpha_j|^2}}{1 - \overline{\alpha_j}e^{i\omega}} \prod_{k=0}^{j-1} \frac{\alpha_k - e^{i\omega}}{1 - \overline{\alpha_k}e^{i\omega}} \frac{|\alpha_k|}{\alpha_k}, \ j = 1, 2, \cdots,$$
$$(14)$$

where $\overline{\alpha_j}$ means the complex conjugate of $\alpha_j$, $\omega \in [-\pi \ \pi)$, and $\frac{|\alpha_j|}{\alpha_j} = \frac{\overline{\alpha_j}}{|\alpha_j|} = -1$ for $\alpha_j = 0$. Such a system is called the *Malmquist system*. It is well-known that the Malmquist system is orthonormal on the unit circle in the sense that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_k(e^{i\omega})\overline{\phi_j(e^{i\omega})}d\omega = \delta_{k,j} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases} \quad (15)$$

We are interested in the space spanned by a subset of the Malmquist system (14). It can be shown see e.g., [18] that the space spanned by $\{\phi_0(e^{i\omega}), \phi_1(e^{i\omega}), \cdots, \phi_m(e^{i\omega})\}$ with the inner product defined as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\omega})\overline{g(e^{i\omega})}d\omega \quad (16)$$

is a RKHS space with the reproducing kernel

$$K_{freq}^{ob}(e^{i\omega}, e^{i\omega'}) = \sum_{k=0}^{m} \phi_k(e^{i\omega})\overline{\phi_k(e^{i\omega'})} \quad (17)$$

which we will refer below as the $(m+1)$th order orthonormal basis (OB) kernel in frequency domain.

Setting in the following

$$B_{m+1}(e^{i\omega}) = \prod_{k=0}^{m} \frac{\alpha_k - e^{i\omega}}{1 - \overline{\alpha_k}e^{i\omega}} \frac{|\alpha_k|}{\alpha_k} \quad (18)$$

where same as before, $\frac{|\alpha_k|}{\alpha_k} = \frac{\overline{\alpha_k}}{|\alpha_k|} = -1$ for $\alpha_k = 0$. Then the kernel (17) has a simplified expression [20, Lemma 5]

$$K_{freq}^{ob}(e^{i\omega}, e^{i\omega'}) = \frac{1 - B_{m+1}(e^{i\omega})\overline{B_{m+1}(e^{i\omega'})}}{1 - e^{i(\omega-\omega')}} \quad (19)$$

which is also known as the Christoffel-Darboux (C-D) formula, see e.g., [18, Theorem 3.1]. The C-D formula is useful to simplify the construction of the kernel matrix (i.e., the regularization matrix).

### B. *The Laguerre kernel*

The simplest case of OB kernels (19) is perhaps the case where $\alpha_i = a$ for $i = 0, 1, \cdots, m$ with $a \in \mathbb{R}$ and $|a| < 1$. In this case, the OB kernel (19) becomes

$$K_{freq}^{lag}(e^{i\omega}, e^{i\omega'}) = \begin{cases} \frac{1 - B_{m+1}(e^{i\omega})\overline{B_{m+1}(e^{i\omega'})}}{1 - e^{i(\omega-\omega')}} & \omega \neq \omega' \\ (m+1)\frac{1-a^2}{|1-ae^{i\omega}|^2} & \omega = \omega' \end{cases}$$
$$(20)$$

which we will refer below as the $(m+1)$th order Laguerre kernel. This is because it is the reproducing kernel of the RKHS space spanned by the first $m+1$ Laguerre rational basis function of (7) in the frequency domain. For the Laguerre kernel (20), there is only one hyper-parameter $a$, the real pole of the Laguerre basis functions, which is convenient to estimate for the hyper-parameter estimation.

### C. *Regularized frequency response estimation*

Since OB kernels are defined in frequency domain, one may wonder if it is possible to work in frequency domain directly without going back to the time domain. The answer is affirmative. Recently, we have derived the dual of the regularized impulse response estimation in frequency domain, i.e., the regularized frequency response estimation, see [21] for details. By using the implementation in [21], we can derive the regularized frequency response with OB kernels.

## IV. REGULARIZED ORTHONORMAL BASIS FUNCTIONS ESTIMATION

It is worth to note that the hyper-parameters of OB kernels (19) are $\{\alpha_k\}_{k=0}^{m}$ which are the poles of the basis functions. So tuning OB kernels is equivalent to tuning the location of the poles of its underlying basis functions. This finding motivates another way of using orthonormal basis functions for regularized system identification:

1) formulate the orthonormal basis functions based model as a linear regression model;
2) treat the poles of the orthonormal basis functions as hyper-parameters and design a suitable kernel for the coefficients of the orthonormal basis functions;
3) estimate the hyper-parameter by empirical Bayes method and then obtain the regularized orthonormal basis functions by using RLS method.

We first formulate (2) with the linear combination of orthonormal basis functions (6) as a linear regression model:

$$y(t) = \sum_{k=1}^{m} g_k \varphi_k * u(t) + v(t) \qquad (21)$$

where $g = [g_1, \cdots, g_m]^T$, $\varphi_k(t)$ is the impulse response of $\bar{F}_k(q)$ in (6). Let $p$ be the vector consisting of all poles of $\bar{F}_k(q) = 0$, $k = 1, \cdots, m$. Then the impulse response $\varphi_k(t)$, $k = 1, \cdots, m$ depend on $p$.

The feature of OB kernels that their hyper-parameters are poles of the basis functions motivates to treat poles of the basis functions as hyper-parameters and estimate them by the empirical Bayes method. It should be noted that this idea has also been figured out independently by Darwish, Tóth, and Van den Hof in [22]. For now we assume that $p$ is known and then we can estimate $g$ by minimizing the RLS criterion

$$\hat{g} = \arg\min_{g} \sum_{t=1}^{N} |y(t) - \sum_{k=1}^{m} g_k \varphi_k * u(t)|^2 + \sigma^2 g^T \mathbf{K}(\alpha)^{-1} g \qquad (22)$$

$$= \arg\min_{g} \|Y_N - \Phi_N(p)g\|_2^2 + \sigma^2 g^T \mathbf{K}(\alpha)^{-1} g \qquad (23)$$

where $\mathbf{K}(\alpha)$ is the regularization matrix on the coefficients $\{g_k\}_{k=1}^{m}$ of the orthonormal basis functions, and $\Phi_N(p)$ is the regression matrix that can be formed in a natural way.

As discussed in Section II, it is a key issue to design a suitable kernel structure, which relies on the prior knowledge that we know about the coefficients of the orthonormal basis functions. Apparently, this issue depends on what orthonormal basis functions we use. For illustration, we consider the Laguerre model (7) as an example below.

The assumptions on the Laguerre coefficients $\{g_k\}_{k=1}^{\infty}$ and the convergence property of Laguerre model, i.e, how fast (7) converges as $m \to \infty$ has been discussed, see e.g., [4]. It is suggested in [4] to assume the absolute convergence of the sum of the Laguerre coefficients $\{g_k\}_{k=1}^{\infty}$, i.e.,

$$\sum_{k=1}^{\infty} |g_k| < \infty \qquad (24)$$

If we treat $\{g_k\}_{k=1}^{\infty}$ as the impulse response of a linear system, then the above assumption (24) says nothing but the linear system is stable. This observation implies that the kernels introduced for regularized impulse response estimation, the SS, TC and DC kernels can be candidates to regularize the Laguerre coefficients $\{g_k\}_{k=1}^{\infty}$.

*Remark 4.1:* As pointed out in [4], the convergence rate of the Laguerre model (7) can be slow, e.g., if the system has poles close to the unit circle or has high resonant poles. In this case, one can try the adapted DC kernel as follows:

$$K^{adc}(k,j) = c\lambda(k+j)\rho^{|k-j|}, \qquad (25)$$

where $\lambda(\cdot)$ is a nonnegative function such that $\lambda(\cdot)$ decays slower than the exponential function and $K^{adc}$ is a valid kernel. Or one can choose to use the other regularized orthonormal basis functions, such as the Kautz model in [5] to handle the case where the system has high resonant poles.

For more general orthonormal basis functions, we can always first try the SS, TC and DC kernels if (24) is assumed. If they do not work so well, we shall spend more efforts on investigating the prior knowledge or assumption on the coefficients of orthonormal basis functions and design a suitable kernel structure accordingly.

Now it remains to estimate the hyper-parameters: the pole $p$ of the orthonormal basis functions and the hyper-parameter $\alpha$ used to parameterize the kernel structure. Assume $\theta \sim \mathcal{N}(0, \mathbf{K}(\alpha))$. Then from (23) we have

$$\hat{p}, \hat{\alpha} = \arg\max_{p,\alpha} p(Y_N|p,\alpha)$$
$$= \arg\min_{p,\alpha} Y_N^T (\Phi_N(p)\mathbf{K}(\alpha)\Phi_N(p)^T$$
$$+ \sigma^2 I_N)^{-1} Y_N + \log\det(\Phi_N(p)\mathbf{K}(\alpha)\Phi_N(p)^T + \sigma^2 I_N)$$

Finally, solving (22) or (23) by replacing $p, \alpha$ with $\hat{p}, \hat{\alpha}$ yields the regularized orthonormal basis function estimate.

## V. REGULARIZED IMPULSE RESPONSE ESTIMATION WITH THE OB KERNEL IS A SPECIAL CASE OF REGULARIZED ORTHONORMAL BASIS FUNCTIONS ESTIMATION

In this section, we show that the regularized impulse response estimation with the OB kernel (17) is a special case of the regularized orthonormal basis functions estimation. More specifically, it is equivalent to ridge regression of the coefficients of the orthonormal basis functions, see e.g., [23].

To show this, it is more convenient to go back to time domain. For the orthonormal basis functions $\{\phi_k(e^{i\omega})\}_{k=0}^{\infty}$ in frequency domain, we can define their correspondents $\{\varphi_k(t)\}_{k=0}^{\infty}$ in time domain. Here, $\varphi_k(t)$ is the impulse response of $\phi_k(e^{i\omega})$ and moreover, we have

$$\phi_k(e^{i\omega}) = \mathcal{F}\{\varphi_k(t)\}, \quad \varphi_k(t) = \mathcal{F}^{-1}\{\phi_k(e^{i\omega})\} \qquad (26)$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the discrete time Fourier transform and its inverse transform, respectively.

Then it is straightforward to verify by using (15) and Cauchy's integral formula that $\{\varphi_k(t)\}_{k=0}^{\infty}$ are orthonormal in the sense that

$$\sum_{t=0}^{\infty} \varphi_k(t)\varphi_j(t) = \delta_{k,j} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases} \qquad (27)$$

Moreover, the space spanned by $\{\varphi_0(t), \varphi_1(t), \cdots, \varphi_m(t)\}$ with the inner product

$$\langle f, h \rangle = \sum_{t=0}^{\infty} f(t)h(t) \qquad (28)$$

is a RKHS space with the reproducing kernel

$$K_{time}^{ob}(t,t') = \sum_{k=0}^{m} \varphi_k(t)\varphi_k(t') \qquad (29)$$

which we will refer below as the $(m+1)$th order OB kernel in time domain. Apparently, the OB kernel (29) in time domain and the OB kernel (17) in frequency domain are related through Fourier transform, e.g., $K_{freq}^{ob}(e^{i\omega}, e^{i\omega'}) = \mathcal{F}\{\mathcal{F}\{K_{time}^{ob}(t,t')\}\}$.

Now consider (13) with the kernel $K(\alpha)$ replaced by the OB kernel (29). The RKHS $\mathcal{H}_{K(\alpha)}$ becomes

$$\mathcal{H}_{K(\alpha)} = \text{ span of } \varphi_0(t), \varphi_1(t), \cdots, \varphi_m(t)$$
$$= \{\vartheta(t)|\vartheta(t) = \sum_{k=1}^{m} g_k \varphi_k(t), g_k \in \mathbb{R}\} \quad (30)$$

and moreover,

$$\|\vartheta\|_{\mathcal{H}_{K(\alpha)}}^2 = \sum_{k=1}^{m} g_k^2 \quad (31)$$

Therefore, (13) is equivalent to

$$\hat{g} = \arg\min_g \sum_{t=1}^{N} |y(t) - \sum_{k=1}^{m} g_k \varphi_k * u(t)|^2 + \sigma^2 \|g\|_2^2 \quad (32)$$

where the regularization $\|g\|_2^2$ is a ridge regression of $g$.

We have the following interesting observations:

1) The regularized impulse response estimation with the OB kernel (29) (equivalently, (17)) is equivalent to a ridge regression of the coefficients of the orthonormal basis functions (32), which is a special case of the regularized orthonormal basis functions estimation (22) with the regularization matrix $\mathbf{K}(\alpha) = I_m$.

2) For the Laguerre kernel (20), the ridge regression $\|g\|_2^2$, i.e., the kernel $K(k, j; \alpha) = \alpha\delta_{k,j}$ cannot guarantee the absolute convergence of the sum of Laguerre model coefficients, i.e., (24). Since the kernel $K(k, j; \alpha) = \alpha\delta_{k,j}$ does not reflect our prior knowledge, it is not a good kernel and the regularized impulse response estimation with the OB kernel (29) will not work well for high order OB kernel. This claim will be verified by numerical simulations shortly.

## VI. Numerical simulation

### A. Data-bank

For this preliminary work, we use a portion of the data-bank in [7, Section 2], which consists of 4 data collections:

- S1D1: fast systems, data sets with $N = 500$, SNR=10
- S2D1: slow systems, data sets with $N = 500$, SNR=10
- S1D2: fast systems, data sets with $N = 375$, SNR=1
- S2D2: slow systems, data sets with $N = 375$, SNR=1

Each collection contains 250 randomly generated 30th order discrete-time systems and data sets. The fast systems have all poles inside the circle with center at the origin and radius 0.95 and the slow systems have at least one pole outside this circle. The signal to noise ratio (SNR) is defined as the ratio of the variance of the noise-free output over the variance of the white Gaussian noise. In all cases the input is Gaussian random signal with unit variance. For more details regarding the data bank, see [7, Section 2].

### B. Examined methods

We examine three methods:

1) **RLAG-TC,RLAG-DI**: the regularized Laguerre basis functions estimation. The Laguerre model with orders $m = 10, 20, 30, 40$ are considered. The TC

kernel (12) and the diagonal (DI) kernel $K(k, j; \alpha) = \text{diag}(\alpha, \alpha^2, \cdots, \alpha^m)$ are used to regularize the Laguerre coefficients. The results are represented as RLAG-TC and RLAG-DI, respectively.

2) **LS-LAG**: the Laguerre basis function estimation with least squares method. The estimate of the pole of the Laguerre model is obtained from RLAG-TC and then the least squares method is used to estimate the Laguerre coefficients without regularization.

3) **RFIR-TC,RFIR-LAG**: the regularized impulse response estimation. The order of the FIR model (5) is chosen to be 125 and the unknown input are set to zero when forming the regression matrix. The TC kernel (12) and the Laguerre kernel (20) are used to regularize the impulse response coefficients. The results are represented as RFIR-TC and RFIR-LAG, respectively. As shown in Section V, RFIR-LAG is equivalent to regularized Laguerre basis functions estimation with the scaled identity kernel $K(k, j; \alpha) = \alpha\delta_{k,j}$.

### C. Model fit

To measure the performance of the examined methods, we compare the impulse response of the estimated model with that of the true system: we let $\hat{g}_k$ and $g_k^0$ to denote the $k$th coefficient of the former and the latter impulse response, respectively. Then the model fit is defined as

$$W = 100 \left(1 - \left[\frac{\sum_{k=1}^{125} |g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^{125} |g_k^0 - \bar{g}^0|^2}\right]^{1/2}\right), \quad \bar{g}^0 = \frac{1}{125}\sum_{k=1}^{125} g_k^0 \quad (33)$$

### D. Simulation result

The average model fit over the corresponding data collections are shown in the table below.

| LS-LAG | S1D1 | S1D2 | S2D1 | S2D2 |
|---|---|---|---|---|
| $m = 10$ | 80.2 | 70.2 | 73.6 | 59.5 |
| $m = 20$ | 88.8 | 68.6 | 82.1 | 56.3 |
| $m = 30$ | 90.0 | 62.8 | 84.0 | 38.1 |
| $m = 40$ | 88.7 | 56.6 | 84.1 | -4.9 |

| RFIR-TC | S1D1 | S1D2 | S2D1 | S2D2 |
|---|---|---|---|---|
| $n = 125$ | 91.4 | 76.1 | 81.2 | 66.1 |

| RFIR-LAG | S1D1 | S1D2 | S2D1 | S2D2 |
|---|---|---|---|---|
| $m = 10$ | 80.1 | 69.6 | 72.0 | 60.6 |
| $m = 20$ | 88.3 | 68.5 | 80.4 | 61.3 |
| $m = 30$ | 89.1 | 64.7 | 82.5 | 59.9 |
| $m = 40$ | 88.1 | 62.6 | 83.1 | 58.4 |

| RLAG-TC | S1D1 | S1D2 | S2D1 | S2D2 |
|---|---|---|---|---|
| $m = 10$ | 80.2 | 71.3 | 72.9 | 63.0 |
| $m = 20$ | 89.2 | 75.3 | 81.9 | 67.8 |
| $m = 30$ | 91.3 | 76.1 | 85.2 | 69.2 |
| $m = 40$ | **91.8** | **76.3** | **86.8** | **70.1** |

| RLAG-DI | S1D1 | S1D2 | S2D1 | S2D2 |
|---|---|---|---|---|
| $m = 10$ | 80.4 | 71.8 | 73.2 | 64.0 |
| $m = 20$ | 89.2 | 75.7 | 82.3 | 68.6 |
| $m = 30$ | 91.2 | 76.0 | 85.9 | 69.7 |
| $m = 40$ | 91.6 | 76.1 | **86.8** | 70.0 |

### E. Findings

First, RLAG can achieve comparable performance as RFIR but with more compact model stucture in terms of the number of basis functions. In particular, for slow systems S2D1 and S2D2, RLAG has clearly better performance (about 5%) than RFIR.

Second, for RLAG, RLAG-DI has very close performance as RLAG-TC, which is different from RFIR studied in [7] where RFIR-DI is much worse than RFIR-TC. For RFIR, TC kernel is clearly a better kernel than the DI kernel because on the one hand, the impulse response is often smooth and on the other hand, the latter does not assume smoothness. However, for RLAG, no prior knowledge regarding the Laguerre coefficients is available except the absolute convergence of the sum of the Laguerre coefficients (24). Both TC kernel and DI kernel can guarantee (24). The simulation results indicate that to assume independence between neighboring Laguerre coefficients is not a bad choice for the tested data bank.

Third, RFIR-LAG has worse performance than RLAG. This coincides with our observation in Section V that the ridge regression is not a suitable regularization for Laguerre basis functions. The influence of the unsuitable regularization is enlarged for high order Laguerre kernels and cause larger difference in the performance.

Fourth, RLAG has better performance than LS-LAG shows the importance of the regularization on the Laguerre coefficients.

## VII. Conclusion and future works

In this preliminary work, we have explored the possibilities to tackle regularized system identification problems using orthonormal basis functions.

Interestingly, the idea of constructing kernels using orthonormal basis functions for regularized impulse response estimation turns out to be a special case of the regularized orthonormal basis functions estimation, and moreover, it is equivalent to ridge regression of the coefficients of the orthonormal basis functions.

The idea of regularizing the orthonormal basis functions works fine but still requires more careful investigation. Due to the space limitation we have mainly studied the regularized Laguerre basis functions as an instance, but the proposed idea applies to the more general orthonormal basis functions, e.g., the Kautz model. Such extensions are necessary and will be examined in our future works because it is known that the convergence rate of the Laguerre model is slow when the system has poles close to the unit circle. Another interesting topic is regarding how to design a suitable kernel for the coefficients of the orthonormal basis functions.

## VIII. Acknowledgement

## References

[1] L. Ljung, *System Identification - Theory for the User*, 2nd ed. Upper Saddle River, N.J.: Prentice-Hall, 1999.

[2] T. Söderström and P. Stoica, *System Identification*. London: Prentice-Hall Int., 1989.

[3] P. S. Heuberger, P. M. Van den Hof, and B. Wahlberg, *Modelling and identification with rational orthogonal basis functions*. London: Springer, 2005.

[4] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. Automatic Control*, vol. AC-36, pp. 551–562, 1991.

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

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

[7] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes - Revisited," *Automatica*, vol. 48, pp. 1525–1535, 2012.

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

[9] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, "System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques," *IEEE Transactions on Automatic Control*, no. 11, pp. 2933–2945, 2014.

[10] G. Pillonetto and A. Chiuso, "Tuning complexity in kernel-based linear system identification: The robustness of the marginal likelihood estimator," in *Control Conference (ECC), 2014 European*, June 2014, pp. 2386–2391.

[11] A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto, "On the design of multiple kernels for nonparametric linear system identification," in *Proceedings of the IEEE Conference on Decision and Control*, Los Angeles, CA., 2014.

[12] 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.

[13] G. Pillonetto and G. D. Nicolao, "Kernel selection in linear system identification. Part I: A Gaussian process perspective," in *Proc. 50th IEEE Conference on Decision and Control*, Orlando, Florida, 2011, pp. 4318–4325.

[14] T. Chen, H. Ohlsson, G. C. Goodwin, and L. Ljung, "Kernel selection in linear system identification. Part II: A classical perspective," in *Proc. 50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, Florida, 2011, pp. 4326–4331.

[15] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, pp. 337–404, 1950.

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

[17] P. M. Van den Hof, P. S. Heuberger, and J. Bokor, "System identification with generalized orthonormal basis functions," *Automatica*, vol. 31, no. 12, pp. 1821–1834, 1995.

[18] B. Ninness, H. Hjalmarsson, and F. Gustafsson, "Generalized Fourier and Toeplitz results for rational orthonormal bases," *SIAM Journal on Control and Optimization*, vol. 37, no. 2, pp. 429–460, 1999.

[19] ——, "The fundamental role of general orthonormal bases in system identification," *IEEE Transactions Autom. Control*, vol. AC-44, no. 7, pp. 1384–1406, July 1999.

[20] M. Djrbashian, "Orthonormal sets of rational functions on the unit circle," *Izvestiya Akademii Nauk Armyanskoi SSR, ser. Mathematika*, no. 1,2, pp. 3–24,106–125, 1966. (Translated by K. Müller and A. Bultheel, TW Reports vol:TW253, Department of Computer Science, K.U. Leuven, 1997).

[21] J. Latarie and T. Chen, "Transfer function and transient estimation by Gaussian process regression in frequency domain," *Automatica*, under review, submitted on 08/01/2014.

[22] M. Darwish, R. Tóth, and P. V. den Hof, "Bayesian system identification based on generalized orthonormal basis functions," in *ERNSI Workshop*, 2014.

[23] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.