

# Control Occupation Kernel Regression for Nonlinear Control-Affine Systems

Moad Abudia, Tejasvi Channagiri, Joel A. Rosenfeld, and Rushikesh Kamalapurkar

**Abstract**—This manuscript presents an algorithm for obtaining an approximation of a nonlinear high order control affine dynamical system. Controlled trajectories of the system are leveraged as the central unit of information via embedding them in vector-valued reproducing kernel Hilbert space (vvRKHS). The trajectories are embedded as the so-called higher order control occupation kernels which represent an operator on the vvRKHS corresponding to iterated integration after multiplication by a given controller. The solution to the system identification problem is then the unique solution of an infinite dimensional regularized regression problem. The representer theorem is then used to express the solution as finite linear combination of these occupation kernels, which converts an infinite dimensional optimization problem to a finite dimensional optimization problem. The vector valued structure of the Hilbert space allows for simultaneous approximation of the drift and control effectiveness components of the control affine system. Several experiments are performed to demonstrate the effectiveness of the developed approach.

## I. INTRODUCTION

Consider a dynamical system of the form  $\dot{x} = f(x)$  with unknown dynamics,  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . Given an observed trajectory,  $\gamma : [0, T] \rightarrow \mathbb{R}^n$  satisfying the dynamics, system identification routines for dynamical systems have traditionally relied on taking numerical derivatives of observed trajectory data [1], [2], or numerical gradient of a cost function [3]. However, numerical derivatives are sensitive to signal noise, where an addition of white noise can cause unbounded disturbances for numerical differentiation. Methods such as the SINDy [2] and SINDYc algorithm [4] ameliorate this issue through a total variation regularization, but even these methods have limitations with the respect to noise.

This research was supported by the Air Force Office of Scientific Research (AFOSR) under contract numbers FA9550-20-1-0127 and FA9550-21-1-0134, and the National Science Foundation (NSF) under award numbers 2027976 and 2027999. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

Moad Abudia is with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: abudia@okstate.edu).

Tejasvi Channagiri is with the Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 USA (e-mail: tchannagiri@gmail.com, rosenfeldj@usf.edu).

Joel A. Rosenfeld is with the Department of Mathematics and Statistics, University of South Florida, Tampa, FL 33620 USA (e-mail: tchannagiri@gmail.com, rosenfeldj@usf.edu).

Rushikesh Kamalapurkar is with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: rkamalapurkar@ufl.edu).

Kernel methods developed by the machine learning community have been adopted for system identification purposes by the control community [5]–[7]. The complexity of kernel methods is further investigated in the system identification context in [8], where it is shown that the tuning of hyperparameters is a problem that still persists today.

Recently, a collection of results surrounding the concept of occupation kernels have appeared, where system identification problems are addressed not through numerical differentiation, but rather through integration [9]. While the analysis in this paper does not explicitly consider measurement noise, the developed algorithm relies on numerical integration, and as indicated by the simulation results, the integration approach is less sensitive to signal noise compared to state-of-the-art methods such as SINDYc. The integration approach can be incorporated in system identification routines naturally through reproducing kernel Hilbert spaces (RKHSs) and occupation kernels [9].

Occupation kernels are functions in a RKHS that represent trajectories of a dynamical system. Since the span of the occupation kernels can be shown to be dense in the RKHS, they can be leveraged as basis functions for approximation (see Proposition 3). Indeed, occupation kernels have been used in precisely that manner for motion tomography in [10] as well as for a regression approach to fractional order nonlinear system identification in [11]. Occupation kernels are also leveraged as basis functions for the construction of eigenfunctions for finite rank representations of Liouville operators in a continuous time dynamic mode decomposition routine in [12]. Occupation kernels generalize the idea of occupation measures, which have been leveraged extensively in optimal control, in a way analogous to how kernel functions generalize delta distributions [13]–[15]. Like kernel functions, occupation kernels provide a function theoretic analog of their respective measures in a RKHS.

The present manuscript generalizes two different approaches to system identification using occupation kernels, occupation kernels for higher order control affine systems are introduced in Section III, and are leveraged as basis functions for the resolution of a regularized regression problem for higher order control affine systems of the form  $\frac{d^s}{dt^s}x = f(x) + g(x)u$ . Specifically, this paper extends the regression approach for fractional order dynamical systems in [11] as well as giving a generalization of the control occupation kernels used in combination with control Liouville operators in [16].

System identification using control occupation kernel regression (COKR) is realized through a regularized regression

formulation, where the regularization forces the minimizer to be a linear combination of the occupation kernels, and in turn, leads to a finite dimensional optimization problem resolved by a matrix equation given in Section IV. In Section V, it is shown that COKR can be viewed as a generalization of kernel ridge regression (KRR) (see, for example, [17, Appendix A]) and that the convergence properties of KRR in terms of function evaluation can be extended to COKR as convergence properties in terms of inner product evaluation (see Proposition 4).

## II. PROBLEM STATEMENT

The objective of this manuscript is to learn an unknown higher order control affine system

$$\frac{d^s x}{dt^s} = f(x) + g(x)u \quad (1)$$

from observed piecewise continuous control signals  $\{u_j : [0, T_j] \rightarrow \mathbb{R}^m\}_{j=1}^M$  and corresponding Carathéodory solutions [18, Section 3.1]  $\{\gamma_{u_j} : [0, T_j] \rightarrow \mathbb{R}^n\}_{j=1}^M$  of (1), where  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is the drift function,  $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$  is the control effectiveness matrix, and  $s \in \mathbb{N}$  is the order of the system, which is assumed to be known.

For several selections of  $f$ ,  $g$ , and  $s$ , the formulation in (1) specializes to several different system identification problems. When  $g \equiv 0$ , the problem reduces to determining the unknown dynamics  $f$  for a higher order dynamical system. When  $s = 1$  the problem becomes a system identification problem for a first order control-affine system. If  $s = 1$  and  $g \equiv 0$ , the method developed in this manuscript agrees with that of [11] for integer-order systems with  $q = 1$ .

Systems of the form (1) encompass  $s$ -th order linear systems and Euler-Lagrange models with invertible inertia matrices, and hence, represent a wide class of physical plants, including but not limited to robotic manipulators and autonomous ground, aerial, and underwater vehicles.

In order to facilitate the description of the controlled dynamical system in terms of operators, a vector valued Reproducing Kernel Hilbert Space (vvRKHS) framework is utilized in this paper.

## III. HIGHER ORDER OCCUPATION KERNELS

Given a Hilbert space  $\mathcal{Y}$  and a set  $X$ , a vector valued RKHS,  $H$ , is a Hilbert space of functions from  $X$  to  $\mathcal{Y}$ , such that for each  $v \in \mathcal{Y}$  and  $x \in X$ , the functional that maps  $f \in H$  to  $\langle f(x), v \rangle_{\mathcal{Y}} \in \mathbb{R}$  is bounded. Hence for each  $x \in X$  and  $v \in \mathcal{Y}$ , there is a function  $K_{x,v} \in H$  such that  $\langle f(x), v \rangle_{\mathcal{Y}} = \langle f, K_{x,v} \rangle_H$ . The mapping  $v \mapsto K_{x,v}$  is linear over  $\mathcal{Y}$ ; hence,  $K_{x,v}$  can be expressed as an operator over  $\mathcal{Y}$  as  $K_x v := K_{x,v}$ , where  $K_x : \mathcal{Y} \rightarrow H$  is called the kernel operator of  $H$  centered at  $x$ . The operator  $K(x, y) := K_y^* K_x$  is called the reproducing kernel operator of  $H$ , where  $K_y^*$  is the adjoint of  $K_y$ .

In this paper, unless otherwise specified, we assume that  $\mathcal{Y} = \mathbb{R}^{m+1}$  (viewed as row vectors),  $X = \mathbb{R}^n$ , and  $H$  is a vvRKHS of  $\mathcal{Y}$ -valued continuous functions over  $X$ . Since  $\mathcal{Y}$  is the space of row-vectors, the operation on  $v \in \mathbb{R}^{m+1}$  by the kernel operator  $K_x$  will be expressed as  $v K_x$ . A trivial

extension of the proof of [19, Proposition 2] shows that for any vvRKHS  $H$  of continuous functions, given any continuous signal,  $\theta : [0, T] \rightarrow \mathbb{R}^n$  and any bounded measurable signal  $u : [0, T] \rightarrow \mathbb{R}^m$ , the functional  $\mathcal{T}_{\theta,u}^s : H \rightarrow \mathbb{R}$ , defined as  $\mathcal{T}_{\theta,u}^s h = \frac{1}{(s-1)!} \int_0^T (T-t)^{s-1} h(\theta(t)) \begin{pmatrix} 1 \\ u(t) \end{pmatrix} dt$  for all  $h \in H$  is bounded. Control occupation kernels were introduced in [16] as the unique functions in  $H$  that represent the functional  $\mathcal{T}_{\theta,u}^1$ , and can be generalized to higher order dynamical systems through the Cauchy iterated integral formula. Note that Definition 1 and Proposition 1 require weaker regularity assumptions than those made in the problem statement, which ensure existence and uniqueness of solutions of (1).

*Definition 1:* Given a continuous signal  $\theta : [0, T] \rightarrow \mathbb{R}^n$  and a bounded measurable signal  $u : [0, T] \rightarrow \mathbb{R}^m$ , the *control occupation kernel of order  $s \in \mathbb{N}$  corresponding to  $\theta$  in  $H$* , denoted by  $\Gamma_{\theta,u}^{(s)}$  is given as the unique function that represents the bounded functional  $\mathcal{T}_{\theta,u}^s$  as  $\langle h, \Gamma_{\theta,u}^{(s)} \rangle_H = \mathcal{T}_{\theta,u}^s h$ . This definition enables the treatment of particular higher order dynamical systems without state augmentation, i.e., without writing the system in terms of first order ordinary differential equations. State augmentation for systems studied in this note would result in needless approximation of already known parts of the dynamics.

*Proposition 1:* Given a continuous signal  $\theta : [0, T] \rightarrow \mathbb{R}^n$ , a bounded measurable signal  $u : [0, T] \rightarrow \mathbb{R}^m$ , and  $s \in \mathbb{N}$ ,

$$\Gamma_{\theta,u}^{(s)}(x) = \frac{1}{(s-1)!} \int_0^T (T-t)^{s-1} (1 - u^\top(t)) K_{\theta(t)}(x) dt, \quad (2)$$

where  $K_{\theta(t)}$  is the kernel operator of  $H$  centered at  $\theta(t)$ .

*Proof:* Consider  $\langle \Gamma_{\theta,u}^{(s)}(x), v \rangle_{\mathcal{Y}} = \langle \Gamma_{\theta,u}^{(s)}, v K_x \rangle_H$ , which follows from the definition of a vvRKHS. Symmetrically, since  $H$  is a real-valued vvRKHS, it follows that

$$\begin{aligned} \langle \Gamma_{\theta,u}^{(s)}(x), v \rangle_{\mathcal{Y}} &= \langle \Gamma_{\theta,u}^{(s)}, v K_x \rangle_H = \langle v K_x, \Gamma_{\theta,u}^{(s)} \rangle_H = \\ &= \frac{1}{(s-1)!} \int_0^T (T-t)^{s-1} [v K_x](\theta(t)) \begin{pmatrix} 1 \\ u(t) \end{pmatrix} dt \\ &= \left\langle \frac{1}{(s-1)!} \int_0^T (T-t)^{s-1} (1 - u(t)^\top) K_{\theta(t)}(x) dt, v \right\rangle_{\mathbb{R}^{m+1}}. \end{aligned}$$

Hence, the equality holds for all  $v \in \mathbb{R}^{m+1}$ , and the result follows. ■

Due to the presence of an integral in (2), a numerical integration routine, such as Simpson's rule, is needed to evaluate the occupation kernels and inner products involving the occupation kernels.

The following proposition shows that Definition 1 can be used to compute inner products in the vvRKHS. The notation  $(\cdot)_j$  is used to denote the  $j$ -th row of the matrix  $(\cdot)$ .

*Proposition 2:* If  $\gamma_u : [0, T] \rightarrow \mathbb{R}^n$  is a Carathéodory solution of (1) with  $((f)_j \ (g)_j) \in H$  for each  $j = 1, \dots, n$ , under a piecewise continuous controller  $u : [0, T] \rightarrow \mathbb{R}^m$ , starting from the initial conditions  $\left\{ \frac{d^\ell \gamma_u}{dt^\ell}(0) \right\}_{\ell=0}^{s-1}$ , then for all  $s \in \mathbb{N}$ ,  $\langle ((f)_j \ (g)_j), \Gamma_{\gamma_u, u}^{(s)} \rangle_H = \left( \gamma_u(T) - \sum_{\ell=0}^{s-1} \frac{T^\ell}{\ell!} \frac{d^\ell \gamma_u}{dt^\ell}(0) \right)_j$ .

*Proof:* Since  $((f(\gamma_u(t)))_j \quad (g(\gamma_u(t)))_j) \begin{pmatrix} 1 \\ u(t) \end{pmatrix} = (f(\gamma_u(t)))_j + (g(\gamma_u(t)))_j u(t) = \left( \frac{d^s \gamma_u}{dt^s}(t) \right)_j$  for almost all  $t \in [0, T]$ , and  $\Gamma_{\gamma_u, u}^{(s)}$  implements Cauchy's iterated integral formula through the inner product of the Hilbert space, the proposition follows through iterated application of the fundamental theorem of calculus. ■

Section IV explores a regularized regression approach for learning higher order control affine dynamical systems, where the dynamics can be expressed as a linear combination of higher order control occupation kernels. Therefore, it is necessary to have a cogent method for evaluating the higher order control occupation kernels, and this can be realized through inner products with the kernels  $e_j K_x$ , where  $e_j$  is the  $j$ -th cardinal basis function for  $\mathbb{R}^{m+1}$ . The following proposition provides an explicit formulation to compute the control occupation kernel in (2).

If variable length trajectories are admitted, then it can be shown that the span of the occupation kernels corresponding to trajectories  $\theta_u$  that result from the application of the control  $u$  to the system in (1) is dense in  $H$ .

*Proposition 3:* For any order  $s \in \mathbb{N}$  and any  $\{\gamma_\ell\}_{\ell=1}^{s-1} \subset \mathbb{R}^n$ , the span of the set

$$A_s := \left\{ \Gamma_{\gamma_u, \gamma_0, u}^{(s)} \mid \gamma_0 \in \mathbb{R}^n, T_u \in [0, T], u \in \mathcal{PC}^0([0, T_u]; \mathbb{R}^m) \right\}$$

is dense in  $H$ , where  $\gamma_{u, \gamma_0}$  is a Carathéodory solution of (1) under the control  $u$  and starting from the initial conditions  $\frac{d^j}{dt^j} \gamma_{u, \gamma_0}(0) = \gamma_j$  for  $j = 0, \dots, s-1$ ,  $\Gamma_{\gamma_u, \gamma_0, u}^{(s)}$  denotes the control occupation kernel of order  $s$  corresponding to the signal  $\gamma_{u, \gamma_0}$  in  $H$ , and  $\mathcal{PC}^0(A; B)$  denotes the space of piecewise continuous functions with domain  $A$  and co-domain  $B$ . ■

*Proof:* See the appendix. ■

Proposition 3 motivates the use of control occupation kernels for system identification. If  $H$  is a vvrKHS of a universal kernel [20, Definition 2], then any continuous function can be approximated, uniformly over any compact set, by a function in  $H$  [20, Theorem 3], and any function in  $H$  can be approximated using a linear combination of sufficiently many control occupation kernels.

#### IV. A REGRESSION METHOD USING CONTROL OCCUPATION KERNELS FOR CONTROL AFFINE SYSTEMS

The COKR method developed in this manuscript solves a regularized regression problem posed over the vvrKHS,  $H$ . Through the representer theorem, it is shown that that higher order control occupation kernels arise naturally as a collection of basis functions for approximating the control affine dynamics.

$$\begin{aligned} \langle ((f)_i \quad (g)_i), \Gamma_{\gamma_u, u}^{(s)} \rangle_H &= \left( \gamma_u(T) - \sum_{\ell=0}^{s-1} \frac{T^\ell}{\ell!} \frac{d^\ell \gamma_u}{dt^\ell}(0) \right)_i \\ &= \frac{1}{(s-1)!} \int_0^T (T-t)^{s-1} ((f(\gamma_u(t)))_i + (g(\gamma_u(t)))_i u(t)) dt, \end{aligned}$$

given  $\lambda > 0$  and the controllers and controlled trajectories from Section II, the regularized regression problem to determine an approximation of the  $i$ -th row of  $f$  and  $g$  within the vector valued RKHS is given as

$$\begin{aligned} \min_{((\hat{f})_i \quad (\hat{g})_i) \in H} & \sum_{j=1}^M \left[ \left\langle ((\hat{f})_i \quad (\hat{g})_i), \Gamma_{\gamma_{u_j}, u_j}^{(s)} \right\rangle_H \right. \\ & \left. - \left( \gamma_{u_j}(T_j) - \sum_{\ell=1}^{s-1} \frac{T_j^\ell}{\ell!} \frac{d^\ell \gamma_{u_j}}{dt^\ell}(0) \right)_i \right]^2 + \lambda \| ((\hat{f})_i \quad (\hat{g})_i) \|_H^2, \end{aligned} \quad (3)$$

where  $\lambda > 0$  is a regularization parameter, and  $\frac{1}{(s-1)!} \int_0^{T_j} (T_j - t)^{s-1} ((\hat{f})_i(\gamma_{u_j}(t)) + (\hat{g})_i(\gamma_{u_j}(t)) u_j(t)) dt$  is equal to the inner product  $\langle ((\hat{f})_i \quad (\hat{g})_i), \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H$ . Using the Representer Theorem for occupation kernels (cf. [21] and [11, Proposition 1]), the minimizer of (3) can be expressed as a linear combination of occupation kernels  $((\hat{f})_i \quad (\hat{g})_i) = w_{i,1} \Gamma_{\gamma_{u_1}, u_1}^{(s)} + \dots + w_{i,M} \Gamma_{\gamma_{u_M}, u_M}^{(s)}$ . Using this representation of the minimizer, the inner products in the optimization problem can be computed as  $\langle ((\hat{f})_i \quad (\hat{g})_i), \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H = \langle \sum_{k=1}^M w_{i,k} \Gamma_{\gamma_{u_k}, u_k}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H = (w)_i(G)^j$

where  $(w)_i := (w_{i,1} \quad \dots \quad w_{i,M})$  denotes the  $i$ -the row of the weight matrix  $w \in \mathbb{R}^{n \times M}$ ,  $(G)^j$  denotes the  $j$ -th column of the Gram matrix  $G \in \mathbb{R}^{M \times M}$ , whose  $i, j$ -th element, denoted by  $G_{i,j}$ , is given by  $G_{i,j} = \langle \Gamma_{\gamma_{u_i}, u_i}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H$ , where

$$\begin{aligned} \langle \Gamma_{\gamma_{u_i}, u_i}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H &= \left( \frac{1}{(s-1)!} \right)^2 \int_0^{T_j} \int_0^{T_i} (T_i - t)^{s-1} \\ &\quad (T_j - t)^{s-1} [1 \quad u_i^\top(\tau)] K(\gamma_{u_j}(t), \gamma_{u_i}(\tau)) \begin{bmatrix} 1 \\ u_j(t) \end{bmatrix} d\tau dt. \end{aligned} \quad (4)$$

Furthermore, the norm in the optimization problem can be computed as  $\| ((\hat{f})_i \quad (\hat{g})_i) \|_H^2 = (w)_i G(w)_i^\top$ . Hence, the resolution (3) reduces to the finite dimensional convex optimization problem

$$\min_{(w)_i \in \mathbb{R}^{1 \times M}} \sum_{j=1}^M \left( \left[ (w)_i(G)^j - (D)_i^j \right]^2 + \lambda (w)_i G(w)_i^\top \right). \quad (5)$$

where  $(D)_i^j$  is the  $i, j$ -th element of the end point difference matrix  $D \in \mathbb{R}^{n \times M}$ , whose  $j$ -th column is  $(D)^j = \gamma_{u_j}(T_j) - \sum_{\ell=1}^{s-1} \frac{T_j^\ell}{\ell!} \frac{d^\ell \gamma_{u_j}}{dt^\ell}(0)$ .

The optimization problem in (5) is written with respect to a single dimension in the state space. Once all the state space dimensions are concatenated, the combined problem can be solved via the resolution of the linear system  $(G + \lambda I_M) w^\top = D^\top$ .

The resultant approximation is given as  $(\hat{f}(x) \quad \hat{g}(x)) = \sum_{j=1}^M (w)_j \Gamma_{\gamma_{u_j}, u_j}^{(s)}(x)$  where  $(w)_j$  denotes the  $j$ -th column of  $w$ . The COKR technique is summarized in Algorithm 1.

**Algorithm 1** The COKR algorithm

**Input:** Trajectories  $\{\gamma_{u_i}\}_{i=1}^M$ , control signals  $\{u_i\}_{i=1}^M$ , a regularization parameter  $\lambda$ , the order of the system  $s$ , a numerical integration procedure, and a reproducing kernels  $K$  of  $H$ .

**Output:**  $(\hat{f}(x) \quad \hat{g}(x))$

- 1:  $G \leftarrow \left( \langle \Gamma_{\gamma_{u_i}, u_i}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H \right)_{i,j=1}^M$ ; using (4)
- 2:  $D \leftarrow \left( (\gamma_{u_j}(T_j))_i - \left( \sum_{\ell=1}^{s-1} \frac{T_j^\ell}{\ell!} \frac{d^\ell \gamma_{u_j}}{dt^\ell}(0) \right)_i \right)_{i,j=1}^{n,M}$
- 3:  $w \leftarrow ((G + \lambda I_M)^{-1} D^\top)^\top$
- 4:  $(\hat{f}(x) \quad \hat{g}(x)) \leftarrow \sum_{j=1}^M (w)^j \Gamma_{\gamma_{u_j}, u_j}^{(s)}(x)$ ; using (2)
- 5: **return**  $(\hat{f}(x) \quad \hat{g}(x))$

**V. FUNCTIONAL RIDGE REGRESSION**

In the developed COKR method, we are given a set of functions  $\{\Gamma_{\gamma_{u_i}, u_i}^{(s)}\}_{i=1}^M$  and a set of constants  $(D)_i^j$  that satisfy  $\langle (h)_i, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H = (D)_i^j$  with  $(h)_i = ((f)_i \quad (g)_i)$ . We then proceed to solve  $\min_{(h)_i \in H} \sum_{j=1}^M \left( \left[ \langle (h)_i, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H - (D)_i^j \right]^2 + \lambda \| (h)_i \|_H^2 \right)$ .

Using the representer theorem, we can also show that given a new function  $\Gamma_{\gamma_{u}, u}^s$ , the COKR predictor  $(\hat{h})_i$  satisfies  $(\delta)_i := \langle (\hat{h})_i, \Gamma_{\gamma_{u}, u}^s \rangle_H = (D)_i(G + \lambda I_M)^{-1} \left( \langle \Gamma_{\gamma_{u_1}, u_1}^{(s)}, \Gamma_{\gamma_{u}, u}^{(s)} \rangle_H \dots \langle \Gamma_{\gamma_{u_M}, u_M}^{(s)}, \Gamma_{\gamma_{u}, u}^{(s)} \rangle_H \right)^\top$ . The COKR method can therefore be seen as a generalization of kernel ridge regression (KRR) (see, for example, [17, Appendix A]), where the control occupation kernels replace the reproducing kernels.

The developed generalization is useful in problems where direct samples of functions to be approximated are not available for training. If inner products of the unknown functions against a set of basis functions can be computed, and if that set of basis functions satisfies the representer theorem, then the generalized KRR method can be used to generate useful estimates of the unknown functions. The affine system identification problem studied in this paper is one example of approximation problems that have this property. The price paid for the generalization is that instead of performance guarantees in terms of function evaluation, we get performance guarantees in terms of inner product evaluation against the set of basis used for the representation. Performance in terms of function evaluation then depends on the selected set of basis, and needs to be analyzed separately.

Performance guarantees developed for KRR in terms of function evaluation can be interpreted as performance guarantees in terms of inner product evaluation via the reproducing property. Such re-interpretation allows generalization of these performance guarantees to COKR. For instance, for a set of  $j$  trajectories, let  $G_j \in \mathbb{R}^{j \times j}$  denote the Gram matrix of the control occupation kernels,  $D_j \in \mathbb{R}^{n \times j}$  denote the end point difference matrix, and  $\Gamma_j \in \mathbb{R}^{j \times j}$  denote the column vector  $\left( \langle \Gamma_{\gamma_{u_1}, u_1}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H \dots \langle \Gamma_{\gamma_{u_{j-1}}, u_{j-1}}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H \right)^\top$ . The first  $j-1$  trajectories can then be used to predict the

$i$ -th component of the end point difference for the  $j$ -th trajectory, denoted by  $(D_j)_i^j$ , using COKR. The predicted end point difference is given by  $(\delta_j)_i := \langle \hat{h}_i, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H = (D_{j-1})_i (G_{j-1} + \lambda I_{j-1})^{-1} \Gamma_j$ . A straightforward adaptation of the proof of [22, Theorem 3] then establishes the following proposition

*Proposition 4:* The cumulative end point difference prediction loss satisfies

$$\begin{aligned} \sum_{j=1}^M \frac{((\delta_j)_i - (D_j)_i^j)^2}{1 + \nu_j} &= \\ \min_{(h)_i \in H} \left( \sum_{j=1}^M \left[ \langle (h)_i, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H - (D_j)_i^j \right]^2 + \lambda \| (h)_i \|_H^2 \right), \\ \text{where } \nu_j &= \frac{\langle \Gamma_{\gamma_{u_j}, u_j}^{(s)}, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle_H - \Gamma_j^\top (G_{j-1} + \lambda I_{j-1})^{-1} \Gamma_j}{\lambda}. \end{aligned}$$

*Proof:* Follows from Proposition 7 below.  $\blacksquare$

Furthermore, similar to KRR, it can be shown that in the limit as  $M$  goes to infinity, the CORK predictor is at least as good as any continuous function for prediction of end point differences.

*Proposition 5:* If  $H$  is a vvRKHS of a universal kernel and  $X$  is compact then for any continuous function  $h : X \rightarrow \mathbb{R}^{1 \times m+1}$ ,

$$\limsup_{M \rightarrow \infty} \frac{1}{M} \left\{ \sum_{j=1}^M \left( (\delta_j)_i - (D_j)_i^j \right)^2 - \sum_{j=1}^M \left[ \langle h, \Gamma_{\gamma_{u_j}, u_j}^{(s)} \rangle - (D_j)_i^j \right]^2 \right\} \leq 0$$

*Proof:* Follows from Proposition 8 below.  $\blacksquare$

**A. A generalization of kernel ridge regression**

It is well-known that the system identification problem can be generalized to other applications where the objective is to approximate an unknown function  $g$  using the data  $\{(F_j, y_j)\}_{j=1}^M$  where  $\{F_j\}_{j=1}^M \subset H^*$ , and  $H^*$  is the set of bounded linear functionals on  $H$ , such that  $[F_j](g) = y_j$  (see, for example, [23, Theorem 1.3.1]). In the functional ridge regression (FRR) approach, the aim is to find a function  $\hat{g} \in H$  that solves

$$\min_{g \in H} \sum_{i=1}^M ([F_i](g) - y_i)^2 + \lambda \| g \|_H^2. \quad (6)$$

The following proposition is a special case of [23, Theorem 1.3.1]. For completeness, a proof is included in the appendix.

*Proposition 6:* If  $f_i$  is the representative of  $F_i$  for all  $i$  then there exist constants  $\{c_i\}_{i=1}^M \subseteq \mathbb{R}$  such that the minimizer of (6) is given by  $\hat{g} = \sum_{i=1}^M c_i f_i$ .

*Proof:* See the appendix.  $\blacksquare$

The vector of constants  $c = (c_1, \dots, c_M)^\top$  in the FRR minimizer are then given by the resolution of the linear system  $(G_f + \lambda I_M)c = y$ , where  $y = (y_1, \dots, y_M)^\top$  denotes the vector of measurements and  $G_f \in \mathbb{R}^{M \times M}$  is the Gram matrix of the representing functions  $\{f_i\}_{i=1}^M$ , whose  $i, j$ -th element is  $\langle f_i, f_j \rangle_H$ . Given a new  $F \in H^*$ , with

representative  $f \in H$ , the action of  $F$  on the unknown function  $g$  can then be approximated as  $[F](g) \approx [F](\hat{g}) = c^\top (\langle f, f_1 \rangle_H, \dots, \langle f, f_M \rangle_H)^\top$ . Note that for implementation of FRR, all that is required is the ability to compute inner products between the representing functions.

The COKR problem studied in this paper is an example of FRR where the functionals  $F_j$  are the integration functionals  $[f \ g] \mapsto \left\langle [f \ g], \Gamma_{\gamma_{u_j}, u_j} \right\rangle_H$  and the representatives are the control occupation kernels  $\Gamma_{\gamma_{u_j}, u_j}$ . The KRR problem is another example of FRR, where the functionals  $F_j$  are the evaluation functionals  $f \mapsto [E_{x_j}](f) = f(x_j)$  and the representatives are data-centered kernels  $K(\cdot, x_j)$ . In this paper, we show that cumulative error guarantees similar to those derived for KRR in results such as [24, Theorem 11.1] also hold for FRR.

### B. Cumulative error bounds for FRR

Let  $H$  be a RKHS of  $\mathbb{R}^n$ -valued continuous functions defined on a compact set  $X$ . Let  $\{F_j\}_{j=1}^\infty$  be a family of bounded linear functionals defined on  $C^0(X, \mathbb{R}^n)$  that satisfies the following assumption.

**Assumption 1:** There exists a positive constant  $\bar{F} < \infty$  such that for all  $j \in \{1, \dots, \infty\}$ ,  $f \in C^0(X, \mathbb{R}^n)$ , and  $g \in H$ ,  $|[F_j](f)| \leq \bar{F} \|f\|_\infty$  and  $|[F_j](g)| \leq \bar{F} \|g\|_H$ . Furthermore, there exists a positive constant  $\bar{Y} < \infty$  such that for all  $j \in \{1, \dots, \infty\}$ ,  $|y_j| \leq \bar{Y}$ .

If the measurements  $y_j$  correspond to the action of the functionals on a continuous scalar field  $g$  defined on a compact set, then the existence of  $\bar{Y}$  follows from uniform boundedness of the functionals. To facilitate the convergence analysis, let  $\hat{g}_j = \arg \min_{g \in H} \sum_{i=1}^{j-1} ([F_i](g) - y_i)^2 + \lambda \|g\|_H^2$  be the FRR minimizer and let  $[F_j](\hat{g}_j)$  be the FRR prediction of the action of the  $j$ -th functional on the unknown function  $g$ , both computed using the first  $j-1$  functionals and the corresponding  $j-1$  measurements. Note that  $[F_j](\hat{g}_j)$  is a prediction since the measurement  $y_j$  is not used to compute  $[F_j](\hat{g}_j)$ . Under assumption 1, Proposition 4 can be generalized to obtain the following result.

**Proposition 7:** The cumulative prediction loss satisfies  $\sum_{j=1}^M \frac{(y_j - [F_j](\hat{g}_j))^2}{1 + \nu_j} = \min_{g \in H} \sum_{j=1}^M ([F_j](g) - y_j)^2 + \lambda \|g\|_H^2$ , where  $\nu_j = \frac{\langle f_j, f_j \rangle_H - \bar{f}_j^\top (G_{f,j-1} + \lambda I)^{-1} \bar{f}_j}{\lambda}$ ,  $\bar{f}_j^\top = (\langle f_1, f_j \rangle_H \dots \langle f_{j-1}, f_j \rangle_H)$ , and  $G_{f,j-1}$  is the Gram matrix of the first  $j-1$  representing functions.

*Proof:* The proof follows from a straightforward adaptation of the proof of [22, Theorem 3]. ■

Additionally, if  $H$  is a vvRKHS of a universal kernel, then the FRR predictor is asymptotically at least as good as any continuous function in the cumulative prediction error metric.

**Proposition 8:** If  $H$  is a vvRKHS of a universal kernel,  $X$  is compact, and assumption 1 is satisfied then for any continuous function  $g : X \rightarrow \mathbb{R}^{1 \times m+1}$ ,

$$\limsup_{M \rightarrow \infty} \frac{1}{M} \sum_{j=1}^M \left( ([F_j](\hat{g}_j) - y_j)^2 - ([F_j](g) - y_j)^2 \right) \leq 0.$$

*Proof:* See the appendix. ■

Proposition 8 shows that the FRR solution converges to a function that performs better than any continuous function, with respect to the cumulative average prediction error. However, Proposition 8 does not imply that the functions  $\hat{g}_j$  converge to  $g$  in the supremum norm, a limitation FRR shares with KRR. The following section nevertheless demonstrates COKR as an effective tool for approximation of nonlinear dynamical systems.

## VI. NUMERICAL EXAMPLES

### A. Higher order system - an academic example

This example utilizes the second order one dimensional nonlinear model of the Duffing oscillator given by  $\ddot{x} = (x - x^3) + (2 + \sin(x))u$ , where  $f(x) = (x - x^3)$  is the drift function,  $g(x) = 2 + \sin(x)$  is the control effectiveness function, and  $u$  is the controller. To approximate the system dynamics, 200 trajectories of the system are recorded, along with the corresponding control signals, starting from a grid of initial conditions  $[-3, 3] \times [-3, 3]$ , under a control signal that is composed of the sum of three sinusoidal signals with randomly generated frequencies ranging from  $1 \text{ rad/s}$  and  $3 \text{ rad/s}$ , with coefficients sampled randomly from  $[-1, 1]$ . Each trajectory is corrupted by Gaussian measurement noise where the signal-to-noise ratio is about  $3 \text{ dB}$ . The initial velocities are obtained by numerically differentiating the measured noisy trajectories. The recorded trajectories and control signals are then utilized to approximate  $f$  and  $g$ . The kernel used in this example is  $K(x, y) = \exp\left(\frac{x^\top y}{\mu}\right)$  with  $\mu = 5$ .

**Experiment 1:** The first experiment is an ablation study done to probe the effect of the regularization parameter  $\lambda$ . To examine the effect of  $\lambda$ , 22 COKR models of the system are generated from the same dataset of trajectories, each using a different value of  $\lambda$ . The values of  $\lambda$  are selected to be between  $10^{-13}$  and  $10^7$ . Figure 1 illustrates the results of the ablation study. The ablation study indicates that there is a wide range of values of  $\lambda$  that result in similar model performance. While guidelines to select  $\lambda$  directly based on the signal-to-noise ratio are difficult to provide, standard techniques such as cross-validation can be used to numerically identify a value of  $\lambda$  that yields acceptable model performance.

Based on this ablation study, the regularization parameter is selected to be  $\lambda = 0.001$  in the following Monte-Carlo trials.

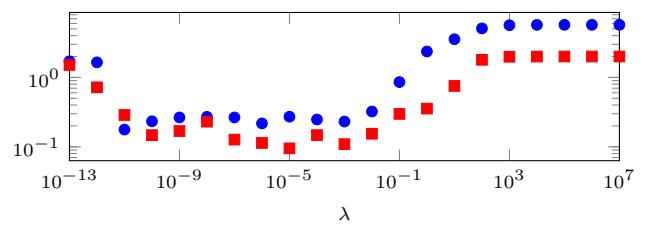
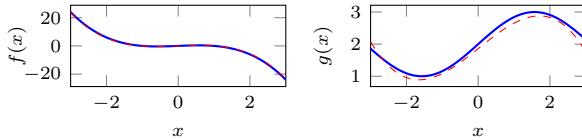
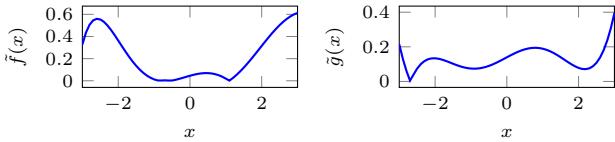


Fig. 1. The blue circle marks are the mean of  $|\tilde{f}(x)|$  and the red square marks are the mean of  $|\tilde{g}(x)|$  over  $x \in [-3, 3]$  for different values of  $\lambda$  using COKR trained with noisy trajectories in Experiment A1.

**Experiment 2:** A Monte-Carlo simulation with 1000 trials is conducted to evaluate robustness of the developed method



**Fig. 2.** The evaluation of  $f$  and  $\hat{f}$  (left) and  $g$  and  $\hat{g}$  (right) for the Duffing oscillator with measurement noise. The solid blue lines represent the true values of the functions and the dotted red lines represent the COKR estimates in Experiment A2.



**Fig. 3.** The evaluation of  $|\tilde{f}|$  (left) and  $|\tilde{g}|$  (right) for the Duffing oscillator with measurement noise in Experiment A2.

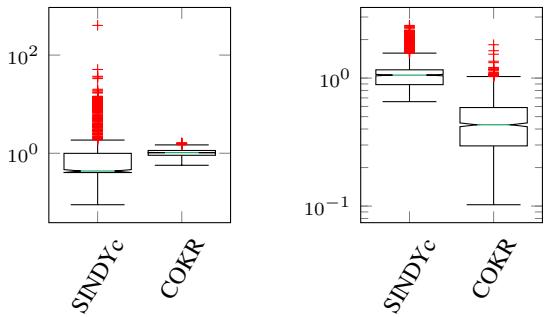
to excitation signals and sensor noise. Figure 2 shows the estimated and actual values of the drift function and the estimated and actual values of the control effectiveness function in the first Monte-Carlo trial. Figure 3 shows the drift approximation error,  $\tilde{f}$ , as a function of  $x$  and the control effectiveness approximation error,  $\tilde{g}$ , as a function of  $x$  in the first Monte-Carlo trial. For each trial, the maximum values of  $|\tilde{f}|$  over the interval  $[-3, 3]$  are collected using COKR and SINDYc for comparison. Box plots for the Monte-Carlo trials are presented in Figure 4 for the maximum errors  $\max_{x \in [-3, 3]} |\tilde{f}|$  and  $\max_{x \in [-3, 3]} |\tilde{g}|$ , respectively.

### B. Control-affine model of a two-link robot manipulator

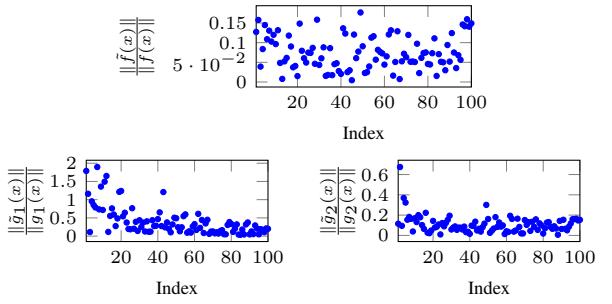
This example utilizes a first-order four-dimensional model of a two-link robot manipulator given by

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ -M^{-1}(x_1)C(x_1, x_2) \end{pmatrix} + \begin{pmatrix} 0_{2 \times 2} \\ M^{-1}(x_1) \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}, \quad (7)$$

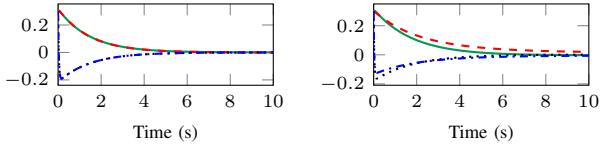
where  $x_1 := \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$ ,  $x_2 := \begin{pmatrix} \dot{q}_1 \\ \dot{q}_2 \end{pmatrix}$ ,  $q_1$ , and  $q_2$  are the angular positions of the two links, respectively,  $M(x_1) = \begin{pmatrix} p_1 + 2p_3 \cos(q_2) & p_2 + p_3 \cos(q_2) \\ p_2 + p_3 \cos(q_2) & p_2 \end{pmatrix}$  is the inertia matrix,  $C(x_1, x_2) = (V(x_1, x_2) + F_d)x_2 + F_s(x_2)$ , where  $F_d = \begin{pmatrix} f_{d_1} & 0 \\ 0 & f_{d_2} \end{pmatrix}$  denotes the viscous friction at the two joints,



**Fig. 4.** Monte-Carlo results of the max of  $|\tilde{f}|$  (left) and  $|\tilde{g}|$  (right) over 1000 trials.



**Fig. 5.** The relative errors  $\frac{\|\tilde{f}(x)\|}{\|f(x)\|}$  (top),  $\frac{\|\tilde{g}_1(x)\|}{\|g_1(x)\|}$  (bottom left), and  $\frac{\|\tilde{g}_2(x)\|}{\|g_2(x)\|}$  (bottom right) evaluated at 100 points, indexed by decreasing distance from the origin in Experiment B1.



**Fig. 6.** Regulation result using the true model (left) and estimated model (right) for computed torque control. The green line is the the angular position  $q_1(\text{rad})$ , the dashed red line is the the angular position  $q_2(\text{rad})$ , the dashed black line is the the angular velocity  $\dot{q}_1(\text{rad/s})$ , and the dashed blue line is the the angular velocity  $\dot{q}_2(\text{rad/s})$  in Experiment B2.

$F_s(x_2) = \begin{pmatrix} f_{s_1} \tanh(\dot{q}_1) \\ f_{s_2} \tanh(\dot{q}_2) \end{pmatrix}$  denotes the static friction at the two joints,  $V(x_1, x_2) = \begin{pmatrix} -p_3 \sin(q_2) \dot{q}_2 & -p_3 \sin(q_2)(\dot{q}_1 + \dot{q}_2) \\ p_3 \sin(q_2) \cdot \dot{q}_1 & 0 \end{pmatrix}$  is the centrifugal-Coriolis matrix and  $\tau_1$  and  $\tau_2$  are the control torques applied by the two motors. The parameters are given by  $p_1 = 3.473$ ,  $p_2 = 0.196$ ,  $p_3 = 0.242$ ,  $f_{d_1} = 5.3$ ,  $f_{d_2} = 1.1$ ,  $f_{s_1} = 8.45$ , and  $f_{s_2} = 2.35$ .

**Experiment 1:** Trajectories of the system are collected, starting from 1000 different initial conditions, sampled using pseudorandom Halton sampling over a cube of side 6 centered at the origin of  $\mathbb{R}^4$ . Each trajectory is recorded using control signals  $\tau_1$  and  $\tau_2$ , comprised of a sum of three sinusoidal signals with randomly generated frequencies and coefficients. For each run, the control signals and the trajectories of the system are recorded to generate the database that is used to approximate  $f$  and  $g$ . Since this system has two controllers  $\tau_1$  and  $\tau_2$ , the control effectiveness matrix  $g$  is decomposed into  $g_1$  and  $g_2$ , where  $g_1$  is the first row of  $g$  and  $g_2$  is the second row of  $g$ .

The approximation error between the identified system and the actual system is computed through evaluation at 100 sample points sampled using pseudorandom Halton sampling over a cube of side 4 centered at the origin of  $\mathbb{R}^4$ . The functions  $f$ ,  $\hat{f}$ ,  $g$ , and  $\hat{g}$  are evaluated at each vertex of the grid to yield the approximation errors  $\tilde{f}$ ,  $\tilde{g}_1$ , and  $\tilde{g}_2$ .

Figure 5 shows the relative errors  $\frac{\|\tilde{f}(x)\|}{\|f(x)\|}$ ,  $\frac{\|\tilde{g}_1(x)\|}{\|g_1(x)\|}$ , and  $\frac{\|\tilde{g}_2(x)\|}{\|g_2(x)\|}$ , respectively as a function of the node index, all using a regularization parameter of  $\lambda = 0.001$ .

**Experiment 2:** To demonstrate the utility of the model developed via the novel system identification method, we simulate the actual two-link robot manipulator using a fourth

	COKR		SIDNYc	
	Mean	SD	Mean	SD
$\max_{x \in [-3,3]} ( \tilde{f}(x) )$	1.0322	0.1765	2.5224	13.3437
$\max_{x \in [-3,3]} ( \tilde{g}(x) )$	0.4718	0.2346	1.1095	0.3417

TABLE I

COMPARISON RESULT OF THE MONTE-CARLO SIMULATIONS OF THE TWO METHODS.

order Runge-Kutta method with a computed torque controller designed to regulate the angular positions and velocities to zero. The computed torque controller is calculated using the estimated model, and is given by

$$\tau = \hat{M}(q)[-K_v \dot{q} - K_p q] + \hat{C}(q, \dot{q}) \quad (8)$$

where  $\hat{M}(q)$  and  $\hat{C}(q, \dot{q})$  are recovered from the estimated data-driven drift and control-effectiveness functions  $\hat{f}$  and  $\hat{g}$ , and  $K_p = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$ ,  $K_v = \begin{pmatrix} 30 & 0 \\ 0 & 30 \end{pmatrix}$  are feedback gains. Figure 6 shows the performance of the computed torque regulator used on the true two-link robot manipulator model.

## VII. DISCUSSION

In the first numerical example, where the Duffing oscillator system is identified, the approximation  $\hat{f}$  is nearly identical to the true  $f$  as seen in Figure 2 and the maximum error is 0.61, which is a 2.6% error relative to  $\|f\|_\infty$  as seen in Figure 3. The approximation  $\hat{g}$  captures the underlying structure of  $g$  where  $g(x) = 2 + \sin(x)$ , but  $\hat{g}$  deviates slightly from the true value as seen in Figure 2 with a maximum error of 0.4 which is a 23% error relative to  $\|g\|_\infty$  as seen in Figure 3. The maximum errors are a function of kernel type, parameters, number and spatial coverage of the recorded data, and measurement noise, and can be reduced through segmentation of the recorded trajectories to generate more data points and increase the resolution of the approximations.

The Monte-Carlo trials summarized in Figure 4 demonstrate the robustness of the developed system identification method to excitation signals and measurement noise. Figure 2 show the full error plots of a representative sample trial.

Figure 4 indicates that COKR is less sensitive to measurement noise than SIDNYc. In Table I the mean and standard deviation of  $\max(|\tilde{f}(x)|)$  and  $\max(|\tilde{g}(x)|)$  for COKR and SIDNYc are listed, which shows that COKR is more precise than SIDNYc. This result is expected in a scenario where measurement noise is present, since COKR uses numerical integration whereas SIDNYc uses numerical differentiation. Although the median of the error measurement of SINDYc for  $f$  in Figure 4 is lower than that of COKR, SINDYc results in a significant number of outliers, where the worst performing run produces an error measurement that is two orders of magnitude higher than COKR.

In the second numerical example, where the model of a two-link robot manipulator is identified, Figure 5 indicates that the estimation errors get worse as one approaches the boundary of the domain covered by the trajectories, and better as one approaches the origin. While the errors are hard to

visualize as a function of distance from the training data in four dimensions, the trends observed in Figure 5 could heuristically be attributed to limited coverage of the corners of the domain.

The identified system  $(\hat{f}, \hat{g})$  is used to compute the necessary torque to regulate the actual two-link robot manipulator, which results in a performance similar to a computed torque controller implemented using exact model knowledge, as seen in Figure 6. The main downside of regulating the system using  $(\hat{f}, \hat{g})$  is the long computation time necessary to evaluate  $(\hat{f}, \hat{g})$  at any given  $x$ , which highlights the need for optimization of the evaluation function to improve computational performance.

## VIII. CONCLUSION

A data-driven control occupation kernel regression method is developed in this manuscript for identification of nonlinear affine control systems, where identification of higher order systems is possible without numerical differentiation for higher order systems. As a result, as indicated by the numerical experiments, the developed method is robust to measurement noise. While the model developed using this method can be used to compute the feedforward component of a controller, depending on the number of trajectories used for modeling, evaluation of the model at a given state can be computationally expensive. Further research is required to develop a more efficient evaluation method to render it useful for real-time feedback.

A distinct advantage of the use of an occupation kernel basis to approximate the dynamics of the system is that under the assumptions of the framework developed in this paper, the solution of the minimization problem in (3) is guaranteed to be a linear combination of the occupation kernels by the representer theorem. This means that the occupation kernels are natural basis functions arising from a dynamics context, and in principle they should perform better than less structured bases. In particular, through the representer theorem, it is clear that the occupation kernel basis will, in general, result in models that yield a lower modeling error in the Hilbert space norm than any other generic basis, resulting in better generalization of the estimates away from the training data. Data-richness in the developed algorithm is related to orthogonality of the control occupation kernels in the Hilbert space. Examination of the relationship between excitation in the input signal in terms of independent frequencies, and orthogonality of the resulting control occupation kernels, is a part of future research.

## APPENDIX

*Proposition 3:* For any order  $s \in \mathbb{N}$  and any  $\{\gamma_\ell\}_{\ell=1}^{s-1} \subset \mathbb{R}^n$ , the span of the set

$$A_s := \left\{ \Gamma_{\gamma_u, \gamma_0, u}^{(s)} \mid \gamma_0 \in \mathbb{R}^n, T_u \in [0, T], u \in \mathcal{PC}^0([0, T_u]; \mathbb{R}^m) \right\}$$

is dense in  $H$ , where  $\gamma_{u, \gamma_0}$  is a Carathéodory solution of (1) under the control  $u$  and starting from the initial conditions  $\frac{d^j}{dt^j} \gamma_{u, \gamma_0}(0) = \gamma_j$  for  $j = 0, \dots, s-1$ ,  $\Gamma_{\gamma_u, \gamma_0, u}^{(s)}$  denotes the control occupation kernel of order  $s$  corresponding to the signal  $\gamma_{u, \gamma_0}$  in  $H$ , and  $\mathcal{PC}^0(A; B)$  denotes the space of

piecewise continuous functions with domain  $A$  and co-domain  $B$ .

*Proof:* Select  $\gamma_0$  such that  $h(\gamma_0) \neq 0$ . Continuity of  $h$  and  $\gamma_{u,\gamma_0}$  can then be invoked to conclude that for any constant control signal  $u(t) = b$  such that  $h(\gamma_0) \begin{pmatrix} 1 \\ b \end{pmatrix} \neq 0$ , there exists a  $T_u > 0$  for which  $\int_0^{T_u} h(\gamma_{u,\gamma_0}(t)) \begin{pmatrix} 1 \\ b \end{pmatrix} dt \neq 0$ . For example, one can select  $T_u = \inf \left\{ t \in [0, T] \mid h(\gamma_{u,\gamma_0}(t)) \begin{pmatrix} 1 \\ b \end{pmatrix} = 0 \right\}$ . A straightforward extension of the above argument to the higher order case allows us to conclude that for any order  $s$  and any nonzero  $h$ , there exist  $\gamma_0$ ,  $u$ , and  $T_u$  such that  $\langle h, \Gamma_{\gamma_0,u}^{(s)} \rangle_H \neq 0$ . That is,  $H \cap A_s^\perp = \{0\}$ , and as a result,  $(A_s^\perp)^\perp = H$ . Since  $(A_s^\perp)^\perp = \overline{\text{span } A_s}$ , we conclude that  $H = \text{span } A_s$ . ■

*Proposition 6:* If  $f_i$  is the representative of  $F_i$  for all  $i$  then there exist constants  $\{c_i\}_{i=1}^M \subseteq \mathbb{R}$  such that the minimizer of (6) is given by  $\hat{g} = \sum_{i=1}^M c_i f_i$ .

*Proof:* write  $g = g_1 + g_2$ , with  $g_1 \in \text{span}\{f_i\}_{i=1}^M := S$  and  $g_2 \in S^\perp$ . Then, the first term in the cost,  $\sum_{i=1}^M ([F_i](g) - y_i)^2 = \sum_{i=1}^M (\langle g_1, f_i \rangle + \langle g_2, f_i \rangle - y_i)^2 = \sum_{i=1}^M (\langle g_1, f_i \rangle - y_i)^2$ , is independent of  $g_2$ . The second term is given by  $\|g\|_H^2 = \langle g_1 + g_2, g_1 + g_2 \rangle_H = \langle g_1, g_1 \rangle_H + 2\langle g_1, g_2 \rangle_H + \langle g_2, g_2 \rangle_H = \|g_1\|_H^2 + \|g_2\|_H^2$ . Since the first term is independent of  $g_2$  and the second term is monotonic in  $\|g_2\|_H^2$ , for any  $\hat{g}$  that minimizes the cost,  $\|g_2\|_H^2 = 0$ . ■

*Proposition 8:* If  $H$  is a vvRKHS of a universal kernel,  $X$  is compact, and assumption 1 is satisfied then for any continuous function  $g : X \rightarrow \mathbb{R}^{1 \times m+1}$ ,

$$\limsup_{M \rightarrow \infty} \frac{1}{M} \sum_{j=1}^M \left( ([F_j](\hat{g}_j) - y_j)^2 - ([F_j](g) - y_j)^2 \right) \leq 0.$$

*Proof:* Given  $\epsilon > 0$  and a continuous function  $g$ , let  $\bar{g}_\epsilon \in H$  be a function that satisfies  $\sup_{x \in X} \|g(x) - \bar{g}_\epsilon(x)\| \leq \epsilon$ . The probabilistic interpretation in section 5 of [22] implies that  $\langle f_j, f_j \rangle_H - \bar{f}_j^\top (G_{f,j-1} + \lambda I)^{-1} \bar{f}_j > 0$ , and as a result,  $1 + \frac{\langle f_j, f_j \rangle_H - \bar{f}_j^\top (G_{f,j-1} + \lambda I)^{-1} \bar{f}_j}{\langle f_j, f_j \rangle_H} > 1$ . Proposition 7 thus implies that  $\sum_{j=1}^M (y_j - [F_j](\hat{g}_j))^2 < \min_{h \in H} \sum_{j=1}^M ([F_j](h) - y_j)^2 + \lambda \|h\|_H^2$ . Letting  $h = \bar{g}_\epsilon$  and adding and subtracting  $\sum_{j=1}^M ([F_j](g) - y_j)^2$  we conclude that  $\sum_{j=1}^M (y_j - [F_j](\hat{g}_j))^2 - \sum_{j=1}^M ([F_j](g) - y_j)^2 + \sum_{j=1}^M \left( ([F_j](g) - y_j)^2 - ([F_j](\bar{g}_\epsilon) - y_j)^2 \right) < \lambda \|\bar{g}_\epsilon\|_H^2$ . Using linearity of  $F_j$ , the identity  $([F_j](g))^2 - ([F_j](\bar{g}_\epsilon))^2 = ([F_j](g + \bar{g}_\epsilon))([F_j](g - \bar{g}_\epsilon))$ , and Assumption 1, we have  $\sum_{j=1}^M (y_j - [F_j](\hat{g}_j))^2 - \sum_{j=1}^M ([F_j](g) - y_j)^2 < \lambda \|\bar{g}_\epsilon\|_H^2 + \sum_{j=1}^M \left( \bar{F}^2 \|g + \bar{g}_\epsilon\|_\infty \|g - \bar{g}_\epsilon\|_\infty + 2\bar{F} \|\bar{g}_\epsilon - g\|_\infty \bar{Y} \right)$ . Using the fact that  $g + \bar{g}_\epsilon = 2g + \bar{g}_\epsilon - g$ , we get the bound  $\sum_{j=1}^M (y_j - [F_j](\hat{g}_j))^2 - \sum_{j=1}^M ([F_j](g) - y_j)^2 < \lambda \|\bar{g}_\epsilon\|_H^2 + \sum_{j=1}^M \left( (\bar{F}^2 2\|g\|_\infty + 2\bar{F}\bar{Y}) \|\bar{g}_\epsilon - g\|_\infty + \bar{F}^2 \|\bar{g}_\epsilon - g\|_\infty^2 \right)$ . Since  $\sup_{x \in X} \|g(x) - \bar{g}_\epsilon(x)\| \leq \epsilon$ ,

$$\frac{1}{M} \left( \sum_{j=1}^M (y_j - [F_j](\hat{g}_j))^2 - \sum_{j=1}^M ([F_j](g) - y_j)^2 \right) < \frac{\lambda}{M} \|\bar{g}_\epsilon\|_H^2 + \left( \bar{F}^2 \|2g\|_\infty + 2\bar{F}\bar{Y} \right) \epsilon + \bar{F}^2 \epsilon^2.$$

Since  $\|\bar{g}_\epsilon\|_H^2$  is independent of  $M$ , the proposition is established by taking the limit superior as  $M \rightarrow \infty$ , and then taking the limit as  $\epsilon \rightarrow 0$ . ■

## REFERENCES

- [1] G. V. Chowdhary and E. N. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *J. Guid. Control Dynam.*, vol. 34, no. 2, pp. 592–607, Mar. 2011.
- [2] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [3] L. Ljung, "System identification," in *Signal Analysis and Prediction*, ser. Applied and Numerical Harmonic Analysis, A. Procházka, J. Uhlíř, P. W. J. Rayner, and N. G. Kingsbury, Eds. Birkhäuser Boston, 1998, pp. 163–173.
- [4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Sparse identification of nonlinear dynamics with control (SINDYC)," *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 710–715, 2016.
- [5] 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.
- [6] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.
- [7] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinear system identification," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2825–2840, 2011.
- [8] 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.
- [9] J. A. Rosenfeld, B. Russo, R. Kamalapurkar, and T. Johnson, "The occupation kernel method for nonlinear system identification," *SIAM J. Control Optim.*, to appear, see arXiv:1909.11792.
- [10] B. P. Russo, R. Kamalapurkar, D. Chang, and J. A. Rosenfeld, "Motion tomography via occupation kernels," *J. Comput. Dyn.*, vol. 9, no. 1, pp. 27–45, 2022.
- [11] X. Li and J. A. Rosenfeld, "Fractional order system identification with occupation kernel regression," *IEEE Control Syst. Lett.*, to appear.
- [12] J. A. Rosenfeld, R. Kamalapurkar, L. F. Gruss, and T. T. Johnson, "Dynamic mode decomposition for continuous time systems with the Liouville operator," *J. Nonlinear Sci.*, vol. 32, no. 1, pp. 1–30, Feb. 2022.
- [13] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat, "Nonlinear optimal control via occupation measures and LMI-relaxations," *SIAM J. Control Optim.*, vol. 47, no. 4, pp. 1643–1666, 2008.
- [14] A. Majumdar, R. Vasudevan, M. M. Tobenkin, and R. Tedrake, "Convex optimization of nonlinear feedback controllers via occupation measures," *Int. J. Robot. Res.*, vol. 33, no. 9, pp. 1209–1230, 2014.
- [15] M. Korda, D. Henrion, and C. N. Jones, "Controller design and value function approximation for nonlinear dynamical systems," *Automatica*, vol. 67, pp. 54 – 66, 2016.
- [16] J. A. Rosenfeld and R. Kamalapurkar, "Dynamic mode decomposition with control Liouville operators," in *IFAC-PapersOnLine*, vol. 54, no. 9, Jul. 2021, pp. 707–712.
- [17] F. Zhdanov and Y. Kalnishkan, "An identity for kernel ridge regression," *Theoretical Computer Science*, vol. 473, pp. 157–178, 2013.
- [18] H. K. Khalil and J. W. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.
- [19] J. A. Rosenfeld and R. Kamalapurkar, "Dynamic mode decomposition with control liouville operators," *IEEE Transactions on Automatic Control*, 2024.
- [20] C. Carmeli, E. De Vito, A. Toigo, and V. Umaná, "Vector valued reproducing kernel hilbert spaces and universality," *Analysis and Applications*, vol. 8, no. 01, pp. 19–61, 2010.
- [21] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *Ann. Math. Statist.*, vol. 41, pp. 495–502, 1970.
- [22] F. Zhdanov and V. Vovk, "Competing with gaussian linear experts," arXiv:0910.4683, 2009.
- [23] G. Wahba, *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.
- [24] V. Vovk, "Kernel ridge regression," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Springer, 2013, pp. 105–116.