Balanced Reduction of Nonlinear Control Systems in Reproducing Kernel Hilbert Space

Jake Bouvrie
Department of Mathematics
Duke University
Durham, NC 27708 USA
Email: jvb@math.duke.edu

Boumediene Hamzi
Department of Mathematics
Duke University
Durham, NC 27708 USA
Email: hamzi@math.duke.edu

Abstract—We introduce a novel data-driven order reduction method for nonlinear control systems, drawing on recent progress in machine learning and statistical dimensionality reduction. The method rests on the assumption that the nonlinear system behaves linearly when lifted into a high (or infinite) dimensional feature space where balanced truncation may be carried out implicitly. This leads to a nonlinear reduction map which can be combined with a representation of the system belonging to a reproducing kernel Hilbert space to give a closed, reduced order dynamical system which captures the essential input-output characteristics of the original model. Empirical simulations illustrating the approach are also provided.

I. INTRODUCTION

Model reduction of controlled dynamical systems has been a long standing, and as yet, unsettled challenge in control theory. The benefits are clear: a low dimensional approximation of a high dimensional system can be manipulated with a simpler controller, and can be simulated at lower computational cost. A complex, high dimensional system may even be replaced by a simpler model all together leading to significant cost savings, as in circuit design, while the "important variables" of a system might shed light on underlying physical or biological processes. Reduction of linear dynamical systems has been treated with some success to date. As we describe in more detail below, model reduction in the linear case proceeds by reducing the dimension of the system with an eye towards preserving its essential input-output behavior, a notion directly related to "balancing" observability and controllability of the system. The nonlinear picture, however, is considerably more involved.

In this paper we propose a scheme for balanced model-order reduction of general, nonlinear control systems. A key, and to our knowledge, novel point of departure from the literature on nonlinear model reduction is that our approach marries approximation and dimensionality reduction methods known to the machine learning and statistics communities with existing ideas in linear and nonlinear control. In particular, we apply a method similar to kernel PCA as well as function learning in Reproducing Kernel Hilbert Spaces (RKHS) to the problem of balanced model reduction. Working in RKHS provides a convenient, general functional-analytical framework for theoretical understanding as well as a ready source of existing results and error estimates. The approach presented here is also strongly empirical, in that observability and controllability, and in some cases the dynamics of the nonlinear system are estimated from simulated or measured trajectories. This emphasis on the empirical makes our approach broadly applicable, as the method can be applied without having to tailor anything to the particular form of the dynamics.

The approach we propose begins by constructing empirical estimates of the observability and controllability Gramians in a high (or possibly infinite) dimensional feature space. The Gramians are simultaneously diagonalized in order to identify directions which, in the feature space, are both the most observable and the most controllable. The assumption that a nonlinear system behaves linearly when lifted to a feature space is far more reasonable than assuming linearity in the original space, and then carrying out the linear theory hoping for the best. Working in the high dimensional feature space allows one to perform linear operations on a representation of the system's state and output which can capture strong nonlinearities. Therefore a system which is not model reducible using existing methods, may become reducible when mapped into such a nonlinear feature space. This situation closely parallels the problem of linear separability in data classification: A dataset which is not linearly separable might be easily separated when mapped into a nonlinear feature space. The decision boundary is linear in this feature space, but is nonlinear in the original data space.

Nonlinear reduction of the state space already opens the door to the design of simpler controllers, but is only half of the picture. One would also like to be able to write a closed, reduced dynamical system whose input-output behavior closely captures that of the original system. This problem is the focus of the second half of our paper, where we again exploit helpful properties of RKHS in order to provide such a closed system.

The paper is organized as follows. In the next section we provide the relevant background for model reduction and balancing. We then adapt and extend balancing techniques described in the background to the current RKHS setting in Section III. Section IV then proposes a method for determining a closed, reduced nonlinear control system in light of the reduction map described in Section III. Finally, Section V provides experiments illustrating an application of the proposed methods to a specific nonlinear system.

II. BACKGROUND

Several approaches have been proposed for the reduction of linear control systems in view of control, but few exist for finite or infinite-dimensional controlled nonlinear dynamical systems. For linear systems the pioneering "Input- Output balancing" approach proposed by B.C. Moore observes that the important states are the ones that are both easy to reach and that generate a lot of energy at the output. If a large amount of energy is required to reach a certain state but the same state yields a small output energy, the state is unimportant for the input-output behavior of the system. The goal is then to find the states that are *both* the most controllable and the most observable. One way to determine such states is to find a change of coordinates where the controllability and observability Gramians (which can be viewed as a measure of the controllability and the observability of the system) are equal and diagonal. States that are difficult to reach and that don't significantly affect the output are then ignored or truncated. A system expressed in the coordinates where each state is equally controllable and observable is called its *balanced realization.*

A proposal for generalizing this approach to nonlinear control systems was advanced by J. Scherpen [20], where suitably defined controllability and observability energy functions reduce to Gramians in the linear case. In general, to find the balanced realization of a system one needs to solve a set of Hamilton-Jacobi and Lyapunov equations (as we will discuss below). Moore [15] proposed an alternative, data-based approach for balancing in the linear case. This method uses samples of the impulse response of a linear system to construct empirical controllability and observability Gramians which are then balanced and truncated using Principal Components Analysis (PCA, or POD). This data-driven strategy was then extended to nonlinear control systems with a stable linear approximation by Lall et al. [11], by effectively applying Moore's method to a nonlinear system by way of the Galerkin projection. Despite the fact that the balancing theory under-pinning their approach assumes a linear system, Lall and colleagues were able to effectively reduce some nonlinear systems.

Phillips [18] et al. has also studied reduction of nonlinear circuit models in the case of linear but unbalanced coordinate transformations and found that approximation using a poly-nomial RKHS could afford computational advantages. Gray and Verriest mention in [6] that studying algebraically defined Gramian operators in RKHS may provide advantageous ap-proximation properties, though the idea is not further explored. Finally, Coifman et al. [3] discuss reduction of an uncontrolled stochastic Langevin system. There, eigenfunctions of a combi-natorial Laplacian, built from samples of trajectories, provide a set of reduction coordinates but does not provide a reduced system. This method is related to kernel principal components (KPCA) using a Gaussian kernel, however reduction in this study is carried out on a simplified linear system outside the context of control.

In the following section we review balancing of linear and nonlinear systems as introduced in [15] and [20].

A. Balancing of Linear Systems

Consider a linear control system

$$\begin{aligned} \dot{x} &= Fx + Gu, \\ y &= Hx, \end{aligned} \tag{1}$$

where (F, G) is controllable, (F, H) is observable and F is Hurwitz. We define the controllability and the observability Gramians as, respectively,

$$\begin{aligned} W_c &= \int_0^\infty e^{Ft} GG^\top e^{F^\top t}\, dt, \\ W_o &= \int_0^\infty e^{F^\top t} H^\top H e^{Ft}\, dt. \end{aligned}$$

These two matrices can be viewed as a measure of the controllability and the observability of the system [15]. For instance, consider the past energy [20], $L_c(x_0)$, defined as the minimal energy required to reach x_0 from 0 in infinite time

$$L_c(x_0) = \inf_{\substack{u \in L_2(-\infty, 0), \\ x(-\infty)=0, x(0)=x_0}} \frac{1}{2} \int_{-\infty}^0 ||u(t)||^2\, dt, \tag{2}$$

and the future energy [20], $L_o(x_0)$, defined as the output energy generated by releasing the system from its initial state $x(t_0) = x_0$, and zero input $u(t) = 0$ for $t \geq 0$, i.e.

$$L_o(x_0) = \frac{1}{2} \int_0^\infty ||y(t)||^2\, dt, \tag{3}$$

for $x(t_0) = x_0$ and $u(t) = 0, t \geq 0$. In the linear case, it can be shown that $L_c(x_0) = \frac{1}{2} x_0^\top W_c^{-1} x_0$, and $L_o(x_0) = \frac{1}{2} x_0^\top W_o x_0$. The columns of W_c span the controllable subspace while the nullspace of W_o coincides with the unobservable subspace. As such, W_c and W_o (or their estimates) are the key ingredients in many model reduction techniques. It is also well known that W_c and W_o satisfy the Lyapunov equations [15]

$$\begin{aligned} FW_c + W_c F^\top &= -GG^\top, \\ F^\top W_o + W_o F &= -H^\top H. \end{aligned}$$

Several methods have been developed to solve these equations directly [12], [13].

The idea behind balancing is to find a representation where the system's observable and controllable subspaces are aligned so that reduction, if possible, consists of eliminating uncontrol-lable states which are also the least observable. More formally, we would like to find a new coordinate system such that

$$W_c = W_o = \Sigma = \text{diag}\{\sigma_1, \cdots, \sigma_n\},$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. If (F, G) is controllable and (F, H) is observable, then there exists a transformation such that the state space expressed in the transformed co-ordinates (TFT^{-1}, TG, HT^{-1}) is balanced and $TW_cT^\top = T^{-\top}W_oT^{-1} = \Sigma$. Typically one looks for a gap in the singular values $\{\sigma_i\}$ for guidance as to where truncation should occur. If we see that there is a k such that $\sigma_k \gg \sigma_{k+1}$, then the states most responsible for governing the input-output relationship of the system are $(x_1, \cdots, x_k)$ while $(x_{k+1}, \ldots, x_n)$ are assumed to make negligible contributions.

Although several methods exist for computing T [12], [13], the general idea is to compute the Cholesky decomposition of W_o so that $W_o = ZZ^\top$, and form the SVD $U\Sigma^2 U^\top$ of $Z^\top W_c Z$. Then T is given by $T = \Sigma^{\frac{1}{2}} U^\top Z^{-1}$. We also note that the problem of finding the coordinate change T can be seen as an optimization problem [1] of the form

$$\min_T \text{trace}[TW_c T^* + T^{-*}W_o T^{-1}].$$

B. Balancing of Nonlinear Systems

In the nonlinear case, the energy functions L_c and L_o in (2) and (3) are obtained by solving both a Lyapunov and a Hamilton-Jacobi equation. Here we follow the development of Scherpen [20]. Consider the nonlinear system

$$\begin{cases} \dot{x} &= f(x) + \sum_{i=1}^m g_i(x)u_i, \\ y &= h(x), \end{cases} \quad (4)$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0) = 0$, $g_i(0) = 0$ for $1 \le i \le m$, and $h(0) = 0$. Moreover, assume the following Hypothesis.

Hypothesis H: The linearization of (4) around the origin is controllable, observable and $F = \frac{\partial f}{\partial x}|_{x=0}$ is asymptotically stable.

Theorem 2.1: [20] If the origin is an asymptotically stable equilibrium of $f(x)$ on a neighborhood W of the origin, then for all $x \in W$, $L_o(x)$ is the unique smooth solution of

$$\frac{\partial L_o}{\partial x}(x)f(x) + \frac{1}{2}h^\top(x)h(x) = 0, \quad L_o(0) = 0 \quad (5)$$

under the assumption that (5) has a smooth solution on W. Furthermore for all $x \in W$, $L_c(x)$ is the unique smooth solution of

$$\frac{\partial L_c}{\partial x}(x)f(x) + \frac{1}{2}\frac{\partial L_c}{\partial x}(x)g(x)g^\top(x)\frac{\partial^\top L_c}{\partial x}(x) = 0, \quad L_c(0) = 0 \quad (6)$$

under the assumption that (6) has a smooth solution $\bar{L}_c$ on W and that the origin is an asymptotically stable equilibrium of $-(f(x) + g(x)g^\top(x)\frac{\partial \bar{L}_c}{\partial x}(x))$ on W.

With the controllability and the observability functions on hand, the input-normal/output-diagonal realization of system (4) can be computed by way of a coordinate transformation. More precisely,

Theorem 2.2: [20] Consider system (4) under Hypothesis H and the assumptions in Theorem 2.1. Then, there exists a neighborhood W of the origin and coordinate transformation $x = \varphi(z)$ on W converting the energy functions into the form

$$L_c(\varphi(z)) = \frac{1}{2}z^\top z,$$

$$L_o(\varphi(z)) = \frac{1}{2}\sum_{i=1}^n z_i^2 \sigma_i(z_i)^2,$$

where $\sigma_1(x) \ge \sigma_2(x) \ge \cdots \ge \sigma_n(x)$. The functions $\sigma_i(\cdot)$ are called *Hankel singular value functions*.

Analogous to the linear case, the system's states can be sorted in order of importance by sorting the singular value functions, and reduction proceeds by removing the least important states.

In the above framework for balancing of nonlinear systems, one needs to solve (or numerically evaluate) the PDEs (5), (6) and compute the coordinate change $x = \varphi(z)$, however there are no systematic methods or tools for solving these problems. Various approximate solutions based on Taylor series expansions have been proposed [9], [8], [5]. Newman [16] introduces a statistical approximation based on exciting the system with white Gaussian noise and then computing the balancing transformation using an algorithm from differential topology. As mentioned earlier, an essentially linear empirical approach was proposed in [11]. In this paper, we combine aspects of both data-driven approaches and analytic approaches by carrying out balancing in a suitable RKHS.

III. EMPIRICAL BALANCING OF NONLINEAR SYSTEMS IN RKHS

We consider a general nonlinear system of the form

$$\begin{cases} \dot{x} &= f(x,u) \\ y &= h(x) \end{cases} \quad (7)$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0,0) = 0$, and $h(0) = 0$. Let $\mathcal{R}(x_0) = \{x' \in \mathbb{R}^n : \exists u \in L_\infty(\mathbb{R}, \mathbb{R}^m)$ and $\exists T \in [0, \infty)$ such that $x(0) = x_0$ and $x(T) = x'\}$ be the reachable set from the initial condition $x(0) = x_0$. We assume that the system is zero-state observable, and that the linearization of (7) around the origin is controllable. We also assume that the origin of $\dot{x} = f(x, 0)$ is asymptotically stable.

We treat the problem of estimating the observability and controllability Gramians as one of estimating an integral operator from data in a reproducing kernel Hilbert space (RKHS) [2]. *Our approach hinges on the key modeling assumption that the nonlinear dynamical system is linear in an appropriate high (or possibly infinite) dimensional lifted feature space.* Covariance operators in this feature space and their empirical estimates are the objects of primary importance and contain the information needed to perform model reduction. In particular, the (linear) observability and controllability Gramians are estimated and diagonalized in the feature space, but capture nonlinearities in the original state space. The reduction approach we propose adapts ideas from kernel PCA (KPCA) [21] and is driven by a set of simulated or sampled system trajectories, extending and generalizing the work of Moore [15] and Lall et al. [11].

A. Definitions

In the development below we lift state vectors of the system into a reproducing kernel Hilbert space [2], $\mathcal{H}$, endowed with a symmetric positive definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which we assume here to be continuous and bounded by $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)} < \infty$. In particular, we make use of the following important properties: For all $f \in \mathcal{H}$, $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$, where $K_x := K(x, \cdot)$. This is the reproducing property. Second, to any RKHS we can associate a feature map $\Phi : \mathcal{X} \to \mathcal{F}$ satisfying $\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = K(x, x')$. For example, we can take $\Phi(x) := K_x$ in which case $\mathcal{F} = \mathcal{H}$ – the "feature space" is the RKHS. We will further assume that $\mathcal{H}$ is always separable.

B. Empirical Gramians

Following [15], we estimate the controllability Gramian by exciting each coordinate of the input with impulses while setting $x_0 = 0$. One can also further excite using rotations of impulses as suggested in [11], however for simplicity we consider only the original signals proposed in [15]. Let $u^i(t) = \delta(t)e_i$ be the i-th excitation signal, and let $x^i(t)$ be the corresponding response of the system. Form the matrix $X(t) = \begin{bmatrix} x^1(t) & \cdots & x^m(t) \end{bmatrix} \in \mathbb{R}^{n \times m}$, so that $X(t)$ is seen as a data matrix with column observations given by the respective responses $x^i(t)$. Then $W_c \in \mathbb{R}^{n \times n}$ is given by

$$W_c = \frac{1}{m} \int_0^\infty X(t)X(t)^\top.$$

We can approximate this integral by sampling the matrix function $X(t)$ within a finite time interval $[0, T]$ assuming the regular partition $\{t_i\}_{i=1}^N, t_i = (T/N)i$. This leads to the empirical controllability Gramian

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^N X(t_i)X(t_i)^\top.$$

As described in [15], the observability Gramian is estimated by fixing $u(t) = 0$, setting $x_0 = e_i$ for $i = 1, \ldots, n$, and measuring the corresponding system output responses $y^i(t)$. As before, assemble the responses into a matrix $Y(t) = [y^1(t) \cdots y^n(t)] \in \mathbb{R}^{p \times n}$. The observability Gramian $W_o \in \mathbb{R}^{n \times n}$ and its empirical counterpart $\widehat{W}_o$ are given by

$$W_o = \frac{1}{p} \int_0^\infty Y(t)^\top Y(t), \quad \widehat{W}_o = \frac{T}{pN} \sum_{i=1}^N \widetilde{Y}(t_i)\widetilde{Y}(t_i)^\top$$

where $\widetilde{Y}(t) = Y(t)^\top$. The matrix $\widetilde{Y}(t_i) \in \mathbb{R}^{n \times p}$ can be thought of as a data matrix with column observations

$$d_j(t_i) = \left(y_j^1(t_i), \ldots, y_j^n(t_i)\right)^\top \in \mathbb{R}^n, \quad j = 1, \ldots, p, \quad (8)$$

so that $d_j(t_i)$ corresponds to the response at time t_i of the single output coordinate j to each of the (separate) initial conditions $x_0 = e_k, k = 1, \ldots, n$. This convention will lead to greater clarity in the steps that follows.

C. Kernel PCA

Kernel PCA [21] generalizes linear PCA by carrying out PCA in a high dimensional feature space defined by a feature map $\Phi : \mathbb{R}^n \to \mathcal{F}$. Taking the feature map $\Phi(x) = K_x$ and given the set of data $\mathbf{x} := \{x_i\}_{i=1}^N \in \mathbb{R}^n$, we can consider PCA in the feature space by simply working with the covariance of the mapped vectors,

$$C_\mathbf{x} = \frac{1}{N} \sum_{i=1}^N \Phi(x_i) \otimes \Phi(x_i), \quad (9)$$

where $\Phi(x_i) \otimes \Phi(x_i) = \langle \Phi(x_i), \cdot \rangle \Phi(x_i)$ denotes the tensor product between two vectors in $\mathcal{H}$. We will assume the data are centered in the feature space so that $\sum_i \Phi(x_i) = 0$. If not, data may be centered according to the prescription in [21]. The principal subspaces are computed by diagonalizing $C_\mathbf{x}$,

however as is shown in [21], one can equivalently form the matrix $K \in \mathbb{R}^{N \times N}$ of kernel products $(K)_{ij} = K(x_i, x_j)$ for $i, j = 1, \ldots, N$, and solve the eigenproblem $K\boldsymbol{\alpha} = N\lambda\boldsymbol{\alpha}$. If $C_\mathbf{x} v_i = \lambda_i v_i$, then we have that $v_i = \Psi \boldsymbol{\alpha}_i$ where $\Psi := \left(\Phi(x_1) \cdots \Phi(x_N)\right)$, and the non-zero eigenvalues of K and $C_\mathbf{x}$ coincide. The eigenvectors $\boldsymbol{\alpha}_i$ of K are then normalized so that the eigenvectors v_i of $C_\mathbf{x}$ have unit norm in the feature space, leading to the condition $\|\boldsymbol{\alpha}_i\|^2 = \lambda_i^{-1}$. Assuming this normalization convention, sort the eigenvectors according to the magnitudes of the corresponding eigenvalues in descending order, and form the matrix $A_q = \begin{bmatrix} \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_q \end{bmatrix}, 1 \leq q \leq \min(n, N)$. Similarly, form the matrix $V_q = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix}, 1 \leq q \leq n$ of sorted eigenvectors of $C_\mathbf{x}$. The first q principal components of a vector $x = \Phi(\tilde{x})$ in the feature space are then given by $V_q^\top x$. It can be shown however (see [21]) that principal components in the feature space can be computed in the original space with kernels using the map $\Pi(x) := A_q^\top \mathbf{k}(x)$, where $\mathbf{k}(x) = \left(K(x, x_1), \ldots, K(x, x_N)\right)^\top$.

D. Model Order Reduction Map

The method we propose consists, in essence, of collecting samples and then performing a process similar to "simultaneous principal components analysis" on the controllability and observability Gramian estimates in the (same) RKHS. As mentioned above, given a choice of the kernel K defining a RKHS $\mathcal{H}$, principal components in the feature space can be computed implicitly in the original input space using K. Because we will find *non-orthogonal* coordinates in the feature space in which the Gramians become simultaneously diagonal, the process is not strictly speaking PCA, and the favorable properties associated with an orthonormal basis are no longer available. We will, however, continue to refer to the process of diagonalizing a covariance matrix as (K)PCA.

Turning to the controllability Gramian (the case of the observability Gramian is analogous), first note that $\widehat{W}_c$ can be viewed as the sample covariance of a collection of $N \cdot m$ vectors, scaled by T:

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^N X(t_i)X(t_i)^\top = \frac{T}{mN} \sum_{i=1}^N \sum_{j=1}^m x^j(t_i)x^j(t_i)^\top.$$

Thus we can form the controllability kernel matrix $K_c \in \mathbb{R}^{Nm \times Nm}$ of kernel products $(K_c)_{ij} = K(x_i, x_j)$ for $i, j = 1, \ldots, Nm$ in order to carry out PCA in the feature space, where we have re-indexed the set of vectors $\{x^k(t_\ell)\}$ to use a single linear index. Similarly, we can compute the observability kernel matrix $K_o \in \mathbb{R}^{Np \times Np}$ consisting of the pairwise kernel products of the collection of data vectors described in (8). Ordinarily, $Nm, Np \gg n$ and K_c, K_o will be rank deficient.

We assume here for simplicity that the number of input excitation signals m is equal to the dimension of the output p so that the number of samples N taken from the output and state trajectories can be the same, leading to kernel matrices K_c and K_o of the same size. If one adopts the set of input excitations $\{u^i(t)\}$ as above, then an alternative although more

restrictive assumption can be that the number of inputs to the system is equal to the number of outputs. Then the pair K_c, K_o is simultaneously diagonalized by taking the (reduced) SVD of $K_c^{1/2} K_o K_c^{1/2}$ so that $K_c^{1/2} K_o K_c^{1/2} = U\Sigma^2 U^\top$. Conjugation by $T = \Sigma^{1/2} U^\top \sqrt{K_c^\dagger}$ diagonalizes K_c and conjugation by $T^{-\top} = \sqrt{\Sigma^\dagger} U^\top K_c^{1/2}$ diagonalizes K_o, where $X^\dagger$ denotes the pseudoinverse of X. Finally, the order of the model is reduced by discarding small eigenvalues $\{\Sigma_{ii}\}_{i=q+1}^n$, and projecting onto the subspace associated with the first $q < n$ largest eigenvalues. This leads to the state-space reduction map $\Pi : \mathbb{R}^n \to \mathbb{R}^q$ given by

$$\Pi(x) = T_q^\top \mathbf{k}_c(x), \quad x \in \mathbb{R}^n \tag{10}$$

where

$$\mathbf{k}_c(x) := \big(K(x, x^1(t_1)), \dots, K(x, x^m(t_N))\big)^\top. \tag{11}$$

IV. CLOSED DYNAMICS OF THE REDUCED SYSTEM

Given the nonlinear state space reduction map $\Pi : \mathbb{R}^n \to \mathbb{R}^q$, a remaining challenge is to construct a corresponding (reduced) dynamical system on the reduced state space which well approximates the input-output behavior of the original system on the original state space. Setting $x_r = \Pi(x)$ and applying the chain rule,

$$\dot{x}_r = \big(J_\Pi(x) f(x, u)\big)\big|_{x = \Pi^{-1}(x_r)}. \tag{12}$$

However we are faced with the difficulty that the map Π is not in general injective (even if $q = n$), and moreover one cannot guarantee that an arbitrary point in the RKHS has a non-empty preimage under Φ [14]. We propose an approximation scheme to get around this difficulty: The dynamics f will be approximated by an element of an RKHS *defined on the reduced state space*. When f is assumed to be known explicitly it can be approximated to a high degree of accuracy. An approximate, least-squares notion of "Π^{-1}" will be given to first or second order via a Taylor series expansion, but only where it is strictly needed – and at the last possible moment – so that a first or second order approximation will not be as crude as one might suppose. We will also consider, as an alternative, a direct approximation of $J_\Pi(\Pi^{-1}(x_r))$ which takes into account further properties of the reproducing kernel as well as the fact that the Jacobian is to be evaluated at $x = \Pi^{-1}(x_r)$ in particular. In both cases, the important ability of the map Π to capture strong nonlinearities will not be significantly diminished.

A. Representation of the dynamics in RKHS

The vector-valued map $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ can be approximated by a composing a set of n regression functions (one for each coordinate) $\hat{f}_i : \mathbb{R}^{q \times m} \to \mathbb{R}$ in an RKHS, with the reduction map Π. It is reasonable to expect that this approximation will be better than directly computing $f(\widehat{\Pi}^{-1}(x_r), u)$ using, for instance, a Taylor expansion notion of "Π^{-1}", which may ignore important nonlinearities at a stage where crude approximations must be avoided.

Let $\tilde{x} = \Pi(x)$ denote a reduced state variable, and concatenate the input examples $\tilde{x}_j = \Pi(x_j) \in \mathbb{R}^q, u_j \in \mathbb{R}^m$ so that $z_j = (\tilde{x}_j, u_j) \in \mathbb{R}^{q \times m}$, and $\{(f_i(x_j, u_j), z_j)\}_{j=1}^\ell$ is a set of input-output training pairs describing the i-th coordinate of the map $(\tilde{x}, u) \mapsto f(x, u)$. The training examples should characterize "typical" behaviors of the system, and can even re-use those trajectories simulated in response to impulses for estimating the Gramians above. We will seek the function $\hat{f}_i \in \mathcal{H}$ which minimizes

$$\sum_{j=1}^\ell \big(\hat{f}_i(z_j) - f_i(x_j, u_j)\big)^2 + \lambda_i \|\hat{f}_i\|_{\mathcal{H}}^2$$

where λ_i here is a regularization parameter. We have chosen the square loss, however other suitable loss functions may be used. It can be shown [22] that in this case $\hat{f}_i$ takes the form $\hat{f}_i(z) = \sum_{j=1}^\ell c_j^i K^f(z, z_j), i = 1, \dots, n$, where K^f defines the RKHS $\mathcal{H}_f$ (and is unrelated to K used to estimate the Gramians). Note that although our notation takes the RKHS for each coordinate function to be the same, in general this need not be true: different kernels may be chosen for each function. Here the $\{c_j^i\}$ comprise a set of coefficients learned using the regularized least squares (RLS) algorithm. The kernel family and any hyper-parameters can be chosen by cross-validation. For notational convenience we will further define the vector-valued empirical feature map

$$\big(\mathbf{k}^f(\tilde{x}, u)\big)_i := K^f\big((\tilde{x}, u), z_i\big)$$

for $i = 1, \dots, \ell$. In this notation $\hat{f}_i\big(\Pi(x), u\big) = \mathbf{c}_i^\top \mathbf{k}^f(\tilde{x}, u)$ where $(\mathbf{c}_i)_j = c_j^i$.

A broad class of systems seen in the literature [20] are also characterized by separable dynamics of the form $\dot{x} = f(x) + \sum_{i=1}^m g_i(x) u_i$. In this case one need only estimate the functions f and g_i from examples $\{(\Pi(x_j), f(x_j))\}_j$ and $\{(\Pi(x_j), g(x_j))\}_j$.

B. Approximation of the Jacobian Contribution

We turn to approximating the component $J_\Pi\big(\Pi^{-1}(x_r)\big)$ appearing in Equation (12).

1) Inverse-Taylor Expansion: A simple solution is to compute a low-order Taylor expansion of Π and then invert it using the Moore-Penrose pseudoinverse to obtain the approximation. For example, consider the first order expansion $\Pi(x) \approx \Pi(a) + J_\Pi(a)(x - a)$. Then we can approximate $\Pi^{-1}(x_r)$ (in the first-order, least-norm sense) as

$$\widehat{\Pi}^{-1}(x_r) := \big(J_\Pi(a)\big)^\dagger (x_r - \Pi(a)) + a. \tag{13}$$

We may start with $a = x_0$, but periodically update the expansion in different regions of the dynamics if desired. A good expansion point could be the estimated preimage of $x_r(t)$ returned by the algorithm proposed in [10].

2) Exploiting Kernel Properties: For certain choices of the kernel K defining the Gramian feature space $\mathcal{H}$, one can exploit the fact that K_x and its derivative bear a special relationship, and potentially improve the estimate for $J_\Pi(\Pi^{-1}(x_r))$. Perhaps the most commonly used off-the-shelf kernel families

are the polynomial and Gaussian families. For any two kernels with hyperparameters p and q (respectively) in one of these classes, we have that $K_p = (K_q)^{p/q}$. We'll consider the polynomial kernel of degree d, $K_d(x,y) := (1 + \langle x, y \rangle)^d$ in particular; the Gaussian case can be derived using similar reasoning. For a polynomial kernel we have that

$$\frac{\partial K_d(x,y)}{\partial x} = dK_{d-1}(x,y)y^\top = d\big(K_d(x,y)\big)^{\frac{d-1}{d}} y^\top.$$

Recalling that $K_d(x,y) = \langle \Phi(x), \Phi(y) \rangle_\mathcal{H}$ and $x_r = \Pi(x) = V_q^\top \Phi(x)$, if Π was invertible then we would have

$$\frac{\partial K_d(x,y)}{\partial x}\bigg|_{x=\Pi^{-1}(x_r)} = d\big\langle (\Phi \circ \Pi^{-1})(x_r), \Phi(y) \big\rangle^{\frac{d-1}{d}} y^\top.$$

The map Π is not injective however, and in addition the fibers of Φ may be potentially empty, so we must settle for an approximation. It is reasonable then to *define* $(\Phi \circ \Pi^{-1})(x_r)$ as the solution to the convex optimization problem

$$\begin{aligned} \min_{z \in \mathcal{H}} \quad & \|z\|_\mathcal{H} \\ \text{subj. to} \quad & \big\|V_q^\top z - x_r\big\|_{\mathbb{R}^k} = 0. \end{aligned} \tag{14}$$

If a point $z \in \mathcal{H}$ has a pre-image in $\mathbb{R}^n$ this definition is consistent with composing Φ with the formal definition $\Phi^{-1}(z) = \{x \in \mathbb{R}^n \mid \Phi(x) = z\}$ and noting that in this case $\Pi \circ \Phi^{-1} = V_q^\top(\Phi \circ \Phi^{-1}) = V_q^\top z$. Furthermore, a trajectory $x_r(t)$ of the closed dynamical system on the reduced statespace need not (and may not) have a counterpart in the original statespace by virtue of the way in which "Π^{-1}" is used in our formulation of the reduction map and corresponding reduced dynamical system.

One will recognize that the solution z^* to (14) is just the Moore-Penrose pseudoinverse $z^* = (V_q^\top)^\dagger x_r$. Inserting this solution into the feature map representation of a kernel K gives the following definition for $K(\Pi^{-1}(x_r), y)$:

$$\begin{aligned} K(\Pi^{-1}(x_r), y) &= \big\langle (\Phi \circ \Pi^{-1})(x_r), \Phi(y) \big\rangle_\mathcal{H} \\ &= \big\langle (V_q^\top)^\dagger x_r, \Phi(y) \big\rangle_\mathcal{H} = \big\langle x_r, V_q^\dagger \Phi(y) \big\rangle_{\mathbb{R}^k} \\ &= \big\langle x_r, (V_q^\top V_q)^{-1} V_q^\top \Phi(y) \big\rangle \\ &= \big\langle x_r, (V_q^\top V_q)^{-1} \Pi(y) \big\rangle \\ &= \big\langle x_r, (T_q^\top K_c T_q)^{-1} \Pi(y) \big\rangle \\ &= \big\langle x_r, (T_q^\top T_q \Sigma_q)^{-1} \Pi(y) \big\rangle. \end{aligned}$$

Substituting into the derivative for a polynomial kernel $K = K_d$ gives

$$\frac{\partial K_d(x,y)}{\partial x}\bigg|_{x=\Pi^{-1}(x_r)} = d\big\langle x_r, (T_q^\top T_q \Sigma_q)^{-1} \Pi(y) \big\rangle^{\frac{d-1}{d}} y^\top$$

which immediately gives an expression for $J_\Pi(\Pi^{-1}(x_r))$. Note that this approximation is global in the sense that the $q \times q$ matrix inverse $(T_q^\top T_q \Sigma_q)^{-1}$ need only be computed once; no updating is required during simulation of the closed system.

C. Reduced System Dynamics

Given an estimate $\hat{f}\big(\Pi(x), u\big)$ of $f(x,u)$ in the RKHS $\mathcal{H}_f$ and a notion of $J_\Pi\big(\Pi^{-1}(x_r)\big)$ from above, we can write down a closed dynamical system on the reduced statespace. We have

$$\begin{aligned} \dot{x}_r &\approx \big(J_\Pi(x)\hat{f}(\Pi(x), u)\big)\big|_{x=\Pi^{-1}(x_r)} \\ &\approx \big(J_\Pi(x)\big)\big|_{x=\Pi^{-1}(x_r)} \mathbf{C}^\top \mathbf{k}^f(x_r, u) \\ &\approx T_q^\top J_\mathbf{k}\big(\Pi^{-1}(x_r)\big)\mathbf{C}^\top \mathbf{k}^f(x_r, u) \end{aligned} \tag{15}$$

where $\mathbf{C}$ is a matrix with the vectors $\mathbf{c}_i$ as its rows, and $J_\mathbf{k}$ is the Jacobian of the empirical feature map defined in Equation (11). Here the expression $J_\mathbf{k}\big(\Pi^{-1}(x_r)\big)$ should be interpreted as notation for either of the Jacobian approximations suggested in Section IV-B.

Equation (15) is seen to give a closed nonlinear control system expressed solely in terms of the reduced variable $x_r \in \mathbb{R}^q$:

$$\begin{cases} \dot{x}_r = T_q^\top J_\mathbf{k}\big(\widehat{\Pi}^{-1}(x_r)\big)\mathbf{C}^\top \mathbf{k}^f(x_r, u) \\ y = \hat{h}(x_r) \end{cases}$$

where the map $\hat{h} \circ \Pi$ modeling the output function $h : \mathbb{R}^n \to \mathbb{R}^p$ is estimated as described immediately below. Although the "true" reduced system does not actually exist due to non-injectivity of the feature map Φ, in many situations one can expect that the above system will capture the essential input-output behavior of the original system. We leave a precise analysis of the error in the approximations appearing in (15) to future work.

D. Outputs of the Reduced System

Analogous to the case of the dynamics f, we are faced with two possibilities for approximating $y = h\big(\Pi^{-1}(x_r)\big)$. We can apply the Taylor approximation $\widehat{\Pi}^{-1}$, or as in Section IV-A we can estimate a map $(\hat{h} \circ \Pi) : \mathbb{R}^n \to \mathbb{R}^p$, $x_r \mapsto y$ from the reduced state space to the output space directly, using RKHS methods. Given samples $\{\Pi(x_j), y_j\}_{j=1}^\ell$, each coordinate function $(\hat{h}_i)_{i=1}^p$ is given in the familiar form $\hat{h}_i(\Pi(x)) = \sum_{j=1}^\ell b_j^i K^h\big(\Pi(x), \Pi(x_j)\big)$, where K^h is the kernel chosen to define the RKHS, and may be different for each coordinate. It should be noted that just given the state space reduction map Π, one can immediately compare the output of the system defined by $\hat{h}(x_r)$ to the original system without defining a closed dynamics as above. In fact with Π and $\hat{h}$ one can design a simpler controller which takes as input the reduced state variable x_r, but controls the original system.

V. Experiments

We demonstrate an application of our method to a 7-dimensional nonlinear system with one dimensional input and
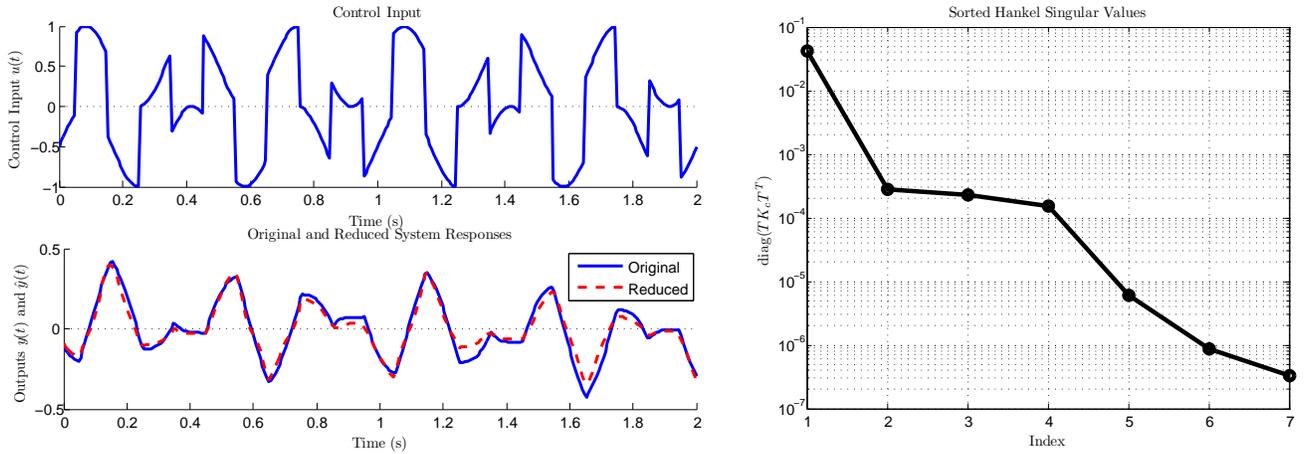
Fig. 1. *(Left) Simulated output trajectories for the original and reduced (2-dimensional) system. (Right) Top Hankel singular values (zeros omitted).*

output appearing in [17] (Example 3.2, pg. 54):

$$\dot{x}_1 = -x_1^3 + u \qquad \dot{x}_2 = -x_2^3 - x_1^2 x_2 + 3x_1 x_2^2 - u$$
$$\dot{x}_3 = -x_3^3 + x_5 + u \qquad \dot{x}_4 = -x_4^3 + x_1 - x_2 + x_3 + 2u$$
$$\dot{x}_5 = x_1 x_2 x_3 - x_5^3 + u \quad \dot{x}_6 = x_5 - x_6^3 - x_5^3 + 2u$$
$$\dot{x}_7 = -2x_6^3 + 2x_5 - x_7 - x_5^3 + 4u$$
$$y = x_1 - x_2^2 + x_3 + x_4 x_3 + x_5 - 2x_6 + 2x_7$$

Impulse and initial-condition responses of the system were simulated as described above, and 800 samples equally spaced in the time interval $[0, 5s]$ were sampled to build the kernel matrices K_c and K_o using the third degree polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^3$. Recall that these kernel matrices, are the inner product counterparts to the empirical Gramians. Examples of K_c and K_o for this system are shown in Figure 2. We imposed a small amount of regularization when computing the balancing transformation T, taking the Cholesky decomposition of $K_c + 0.001 \cdot I$ instead of K_c. Figure 1 (right pane) shows the Hankel singular values $\Sigma = TK_cT^\top$ for this problem on a log scale. One can see that perhaps the first two components ought to capture most of the system's behavior. Thus the reduction map Π was defined by taking only the eigenvectors (scaled columns of T) corresponding to the largest two Hankel singular values, giving a reduced state space of dimension two.

Next, a map from the reduced variable x_r to $\dot{x}$ was estimated following Section IV-A. The control input was chosen to be a 10hz square wave, and 1000 samples from the simulated system in the interval $[0, 5s]$ were mapped down using Π and then used to solve the (n) RLS regression problems, one for each state variable, again using a third degree polynomial kernel. All initial conditions were set to zero. The desired outputs (dependent variable examples) used to learn $\hat{f}$ were taken to be the true f evaluated at the samples from the simulated state trajectory. We also added a bias dimension of 1's to the data to account for any offset, and used a fast leave-one-out cross-validation (LOOCV) computation [19] to select the optimal regularization parameter. Two remarks are in order. The above dynamics can in fact be represented explicitly and

exactly in a 3rd degree polynomial RKHS; only monomials up to degree 3 appear in the dynamics. Second, the control input is decoupled from the state. Both of these facts can be used to obtain an improved reduced model, however we did not make use of these special properties and instead applied the simplest version of the techniques described above which assume no special structure.

We followed a similar process to learn the output function $y = \hat{h}(x_r)$. Here we used a 10Hz square wave control input, zero initial conditions and 700 samples in the interval $[0, 5s]$. For this function the Gaussian kernel $K(x, y) = \exp(-\gamma \|x - y\|_2^2)$ was used to demonstrate that our method does not rely on any particular match between the form of the dynamics and the type of kernel. The scale hyperparameter γ was chosen to be the average distance between the training examples. We again used LOOCV to select the RLS regularization parameter.

Finally, the closed system was simulated as described above using $x_0 = 0$ and a control input different from those used to learn the dynamics and output functions: $u(t) = \frac{1}{2}\big(\sin(2\pi 3t) + \mathrm{sq}(2\pi 5t - \pi/2)\big)$ where $\mathrm{sq}(\cdot)$ denotes the square wave function. This input is shown at the top of Figure 1 (left panel). The Taylor series approximation for Π was done once about x_0 and was not updated further. The simulated outputs $\hat{y}(t)$ of the closed reduced system as well as the output $y(t)$ of the original system are plotted at the bottom in Figure 1 (left panel). One can see that, even for a significantly different input, the two dimensional reduced system closely captures the original system. The main source of error is seen to be over- and under-shoot near the square wave transients. This error can be further reduced by simulating the system for different sorts of inputs (and/or frequencies) and including the collected samples in the training sets used to learn Π, $\hat{f}$ and $\hat{h}$. Indeed, we have had some success driving example systems with random uniform input in some cases.

VI. CONCLUSION

In this paper we introduced a new model reduction method for nonlinear control systems. The method assumes that the nonlinear system is approximately linear in a high dimensional
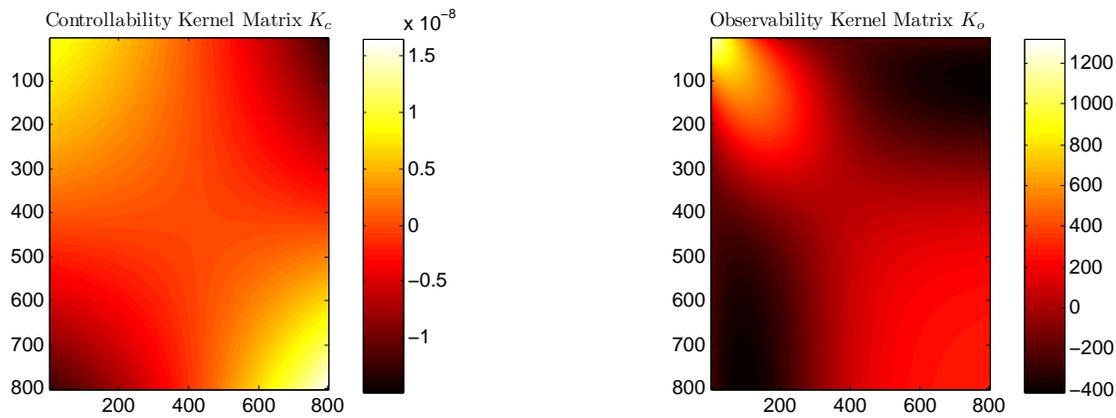
Fig. 2. *Plots of the kernel matrices encoding controllability properties (left) and observability properties (right) of the system.*

feature space, and carries out linear balanced truncation in that space. This leads to a nonlinear reduction map, which we suggest can be combined with representations of the dynamics and output functions by elements of an RKHS to give a closed reduced order dynamical system which captures the input-output characteristics of the original system. We then demonstrated an application of our technique to a 7-dimensional system and simulated the original and reduced models for comparison, showing that the approach proposed here can yield good low-order nonlinear reductions of strongly nonlinear control systems. We believe that techniques well known to the machine learning and statistics communities can offer much to control and dynamical systems research, and many further directions remain, including reduction of unstable systems and structure preserving systems.

REFERENCES

[1] Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*, SIAM Publications.
[2] Aronszajn, N. (1950). Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.*, 68:337-404.
[3] Coifman, R. R., I. G. Kevrekidis, S. Lafon, M. Maggioni and B. Nadler (2008). Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems, *Multiscale Model. Simul.*, 7(2):842-864.
[4] Dullerud, G. E., and F. Paganini (2000). *A Course in Robust Control Theory: a Convex Approach*, Springer.
[5] Fujimoto, K. and D. Tsubakino (2008). Computation of nonlinear balanced realization and model reduction based on Taylor series expansion, *Systems and Control Letters*, **57**, 4, pp. 283-289.
[6] Gray, W. S. and E. I. Verriest (2006). Algebraically Defined Gramians for Nonlinear Systems, *Proc. of the 45th IEEE CDC*.
[7] Krener, A. (2006). Model reduction for linear and nonlinear control systems. The Bode Lecture in the 45th IEEE CDC.
[8] Krener, A. J. (2007). The Important State Coordinates of a Nonlinear System. In *"Advances in control theory and applications"*, C. Bonivento, A. Isidori, L. Marconi, C. Rossi, editors, pages 161-170. Springer.
[9] Krener, A. J. (2008). Reduced order modeling of nonlinear control systems. In *"Analysis and Design of Nonlinear Control Systems"*, A. Astolfi and L. Marconi, editors, pages 41-62. Springer.
[10] Kwok, J. T. and I.W. Tsang (2003). "The Pre-Image Problem in Kernel Methods". In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*.
[11] Lall, S., J. Marsden, and S. Glavaski (2002). A subspace approach to balanced truncation for model reduction of nonlinear control systems, *International Journal on Robust and Nonlinear Control*, **12**, 5, pp. 519-535.
[12] Laub, A.J. (1980). On Computing "balancing" transformations, *Proc. of the 1980 Joint Automatic Control Conference (ACC)*.
[13] Li, J.-R. (2000). *Model Reduction of Large Linear Systems via Low Rank System Grammians*. Ph.D. thesis, Massachusetts Institute of Technology.
[14] Mika, S., B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch (1998). Kernel PCA and de-noising in feature spaces, In *Proc. Advances in Neural Information Processing Systems (NIPS) 11*, pp. 536–542, MIT Press.
[15] Moore, B. (1981). Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, *IEEE Tran. Automat. Control*, **26**, 1, pp. 17-32.
[16] Newman, A.J., and P. S. Krishnaprasad (2000). Computing balanced realizations for nonlinear systems, *Proc. of the Math. Theory of Networks and Systems (MTNS)*.
[17] Nilsson, O. (2009). *On Modeling and Nonlinear Model Reduction in Automotive Systems*, Ph.D. thesis, Lund University.
[18] Phillips, J., J. Afonso, A. Oliveira and L. M. Silveira (2003). Analog Macromodeling using Kernel Methods. In *Proceedings of the IEEE/ACM International Conference on Computer-aided Design*.
[19] Rifkin, R., and R.A. Lippert. *Notes on Regularized Least-Squares*, CBCL Paper 268/AI Technical Report 2007-019, Massachusetts Institute of Technology, Cambridge, MA, May, 2007.
[20] Scherpen, J. (1994). *Balancing for Nonlinear Systems*, Ph.D. thesis, University of Twente.
[21] Schölkopf, B., Smola, A., and Müller, K (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5): 1299–1319.
[22] Wahba, G. (1990). Spline Models for Observational Data, *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics* 59.