KERNEL METHODS FOR THE
APPROXIMATION OF NONLINEAR SYSTEMS

JAKE BOUVRIE[†][¶] AND BOUMEDIENE HAMZI[‡][¶]

Abstract. We introduce a data-driven model approximation method for nonlinear control systems, drawing on recent progress in machine learning and statistical dimensionality reduction. The method is based on embedding the nonlinear system in a high (or infinite) dimensional reproducing kernel Hilbert space (RKHS) where linear 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 RKHS to give a closed, reduced order dynamical system which captures the essential input-output characteristics of the original model. Working in RKHS provides a convenient, general functional-analytical framework for theoretical understanding. Empirical simulations illustrating the approach are also provided.

1. Introduction. Data-based modelling of nonlinear dynamical systems has been addressed by many authors. For example, several methods have been developed in Time Series Analysis ([19] for example) and System Identification ([52], [51] for example). Coifman et al. discuss data-based modelling of a stochastic Langevin system [9]. Archambeau et al. [2] proposed methods to approximate SDEs from data. Smale and Zhou use kernel methods to approximate a hyperbolic dynamical system [45].

In this paper we propose a scheme for the approximation of nonlinear systems using balanced model-order reduction. 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 (Principal Component Analysis) as well as function approximation in Reproducing Kernel Hilbert Spaces (RKHSes) to the problem of balanced model reduction. Working in RKHSes 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.

The approach we propose begins by viewing the controllability and observability energies for nonlinear systems as Gramians in a high (or possibly infinite) dimensional RKHS. These Gramians are approximated empirically and then simultaneously diagonalized in order to identify directions which, in the RKHS, are both the most observable and the most controllable.

The assumption that it is possible to apply the method of linear balancing to a nonlinear system when lifted to a RKHS is far more reasonable than applying the linear theory in the original space hoping for the best. Working in the high dimensional RKHS allows one to perform linear operations on a representation of the system's state and output which can capture strong nonlinearities. Moreover,

---

[†]Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology, Cambridge, MA 02138, USA. (jvb@csail.mit.edu)

[‡]Department of Mathematics, Imperial College London, London SW7 2AZ, UK. (b.hamzi@imperial.ac.uk)

[¶]Parts of this work were done while both authors were at Department of Mathematics, Duke University, Durham, NC 27708, USA.

working in an RKHS readily include linear spaces[1] and, therefore, our approach not only covers existing Linear Theory but also extends the range of available spaces where it is reasonable to assume that one will obtain better results when dealing with a nonlinear problem. Therefore, a system for which existing model reduction methods fail, may be approximated by a lower dimensional system when mapped into a RKHS. 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. Essentially, we are proposing to apply linear methods to nonlinear systems once mapped into a high (possibly infinite-dimensional) Hilbert space[2].

Nonlinear reduction of the state space already allows 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 for linear and nonlinear control systems. We then adapt and extend balancing techniques described in the background section to the current RKHS setting in Section 3. Section 4 then proposes a method for determining a closed, reduced nonlinear control system in light of the reduction map described in Section 3. Finally, Section 5 provides experiments illustrating an application of the proposed methods to some nonlinear systems where the method of linear balancing does not apply in $\mathbb{R}^n$ since the systems we simulated are not linearly controllable and the origin is not asymptotically stable but the same method of linear balancing applies to the nonlinear system after being lifter to a RKHS. Appendix A contains a review of Learning Theory, Appendix B contains a description of kernel PCA.

Preliminary results of this work can be found in [5, 6].

2. 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 nonlinear control systems. For linear systems, the pioneering "Input- Output balancing" approach proposed by B.C. Moore [29] 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

---

[1]When using the embedding $\Phi : x \mapsto x$ and kernel $k(x,y) = \langle x, y \rangle$, cf. Appendix A for meaning of these objects.

[2]Let us note here that it is important to choose the right RKHS in order to perform this sort of computations. In our work, we used existing Universal Kernels such as the Gaussian or polynomial kernels (cf. Proposition A.4) but, in general, properly choosing the right RKHS is an open problem even in classical Learning Theory.

equally controllable and observable is called its *balanced realization*.

A proposal for generalizing this approach to nonlinear control systems was advanced by J. Scherpen [36], 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 [29] 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 Proper Orthogonal Decomposition (POD) [17]). This data-driven strategy was then extended to nonlinear control systems with a stable linear approximation by Lall et al. [25], by effectively applying Moore's method to a nonlinear system by way of the Galerkin projection. Despite the fact that the balancing theory underpinning their approach assumes a linear system, Lall and colleagues were able to effectively reduce some nonlinear systems.

Phillips et al. [32] has also studied reduction of nonlinear circuit models in the case of linear but unbalanced coordinate transformations and found that approximation using a polynomial RKHS could offer computational advantages. Gray and Verriest mention in [15] that studying algebraically defined Gramian operators in RKHS may provide advantageous approximation properties, though the idea is not further explored. Finally, Coifman et al. [9] discuss reduction of an uncontrolled stochastic Langevin system. There, eigenfunctions of a combinatorial 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 sections we review balancing of linear and nonlinear systems as introduced in [29] and [36]. See also [37] for a good survey on balancing for linear and nonlinear systems.

2.1. Balancing of Linear Systems. Consider a linear control system

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu, \\ y & = & Cx, \end{array} \qquad (2.1)$$

where (A, B) is controllable, (A, C) is observable and A is Hurwitz. We define the controllability and the observability Gramians as, respectively,

$$W_c = \int_0^\infty e^{At} BB^\top e^{A^\top t}\, dt, \quad \text{and} \quad W_o = \int_0^\infty e^{A^\top t} C^\top C e^{At}\, dt.$$

These two matrices can be viewed as a measure of the controllability and the observability of the system [29]. For instance, consider the past energy [36, 35], $L_c(x_0)$, defined as the minimal energy required to reach x_0 from 0 in infinite time

$$L_c(x_0) = \inf_{\substack{u \in L_2(-\infty, 0), \\ x(-\infty)=0, x(0)=x_0}} \frac{1}{2} \int_{-\infty}^0 ||u(t)||^2\, dt, \qquad (2.2)$$

and the future energy [36, 35], $L_o(x_0)$, 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.

$$L_o(x_0) = \frac{1}{2} \int_0^\infty ||y(t)||^2\, dt, \qquad (2.3)$$

for $x(t_0) = x_0$ and $u(t) = 0, t \geq 0$. In the linear case, it can be shown that

$$L_c(x_0) = \tfrac{1}{2} x_0^\top W_c^{-1} x_0, \quad \text{and} \quad L_o(x_0) = \tfrac{1}{2} x_0^\top W_o x_0. \qquad (2.4)$$

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 [29]

$$AW_c + W_c A^\top = -BB^\top, \quad A^\top W_o + W_o A = -C^\top C. \qquad (2.5)$$

Several methods have been developed to solve these equations (see [16, 26, 27] for example). As mentioned at the beginning of this section, it is also possible to estimate the Gramians from empirical data, cf. (3.4) and (3.6) below.

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 the states that are least controllable and which are also the least observable. More formally, we would like to find a new coordinate system such that $\Sigma := W_c = W_o = \mathrm{diag}\{\sigma_1, \cdots, \sigma_n\}$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$.

THEOREM 2.1. *[11] If (A, B) is controllable and (A, C) is observable, then the eigenvalues of $W_o W_c$ are similarity invariants, i.e. they are independent of the choice of the state-space representation of (2.1). Moreover, there exists a state-space representation where*

$$\Sigma := W_c = W_o = diag\{\sigma_1, \cdots, \sigma_n\} \qquad (2.6)$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ are the square roots of the eigenvalues of $W_o W_c$. Such representations are called balanced, and the system is in balanced form. The σ_i's, $i = 1, ..., n$ are called the Hankel singular values. The state space expressed in the transformed coordinates (QAQ^{-1}, QB, CQ^{-1}) is balanced and $QW_c Q^\top = Q^{-\top} W_o Q^{-1} = \Sigma$ where $Q \in \mathbb{R}^{n \times n}$.

The Hankel singular values, $\sigma_i|_{i=1}^n$, are the square roots of the eigenvalues of $W_o W_c$ and are the singular values of the Hankel operator

$$\mathbb{H} = \Psi_o \circ \Psi_c \qquad (2.7)$$

that characterizes the input-output behaviour of the system (2.1) with Ψ_c, the controllability operator, which maps $u \in L_2(-\infty, 0]$ to $x(0)$ and Ψ_o, the observability operator, which maps $x(0)$ to $y(t)$, $t \geq 0$ with no input applied for $t \geq 0$ [11]. More precisely,

$$\begin{aligned} \Psi_c : L_2(-\infty, 0] &\rightarrow \mathbb{C}^n \\ u &\mapsto \int_{-\infty}^0 e^{-A\tau} B u(\tau) d\tau \end{aligned} \qquad (2.8)$$

and

$$\begin{aligned} \Psi_o : \mathbb{C}^n &\rightarrow L_2[0, \infty) \\ x(0) = x_0 &\mapsto \begin{cases} Ce^{At} x_0, \text{ for } t \geq 0, \\ 0, \text{ otherwise.} \end{cases} \end{aligned} \qquad (2.9)$$

Clearly, $x_0 = \Psi_c u(t)$ for $u(t) \in L_2(-\infty, 0]$ is the system state at $t = 0$ due to the past input and $y(t) = \Psi_0 x_0$, $t \geq 0$ is the future output due to the initial state

x_0 with the input set to zero. In fact, $\mathbb{H}$ characterizes the system's future output $y(t) = \mathbb{H}u(t)$, $t \geq 0$ based on the past input $u(t), t \leq 0$. More precisely, if $x(-\infty) = 0$,

$$\mathbb{H}u(t) = \Psi_o \Psi_c u(t) = \int_{-\infty}^{0} Ce^{A(t-\tau)} Bu(\tau)d\tau, \quad \text{for } t \geq 0 \qquad (2.10)$$

When $\mathbb{H}$ is known to be a compact operator, then its adjoint operator $\mathbb{H}^*$ is also compact and the composition $\mathbb{H}^*\mathbb{H}$ is a self-adjoint compact operator with the spectral decomposition

$$\mathbb{H}^*\mathbb{H} = \sum_{i=1}^{\infty} \sigma_i^2 \langle \cdot, \Psi_i \rangle_{L_2} \Psi_i, \quad \sigma_i \geq 0, \qquad (2.11)$$

$$\langle \Psi_i, \Psi_j \rangle_{L_2} = \delta_{ij}, \quad \langle \Psi_i, (\mathbb{H}^*\mathbb{H})(\Psi_i) \rangle_{L_2} = \sigma_i^2 \qquad (2.12)$$

where σ_i^2 is an eigenvalue of $\mathbb{H}^*\mathbb{H}$ with the corresponding eigenvector Ψ_i, ordered as $\sigma_1 \geq \cdots \geq \sigma_n > 0$ and $\sigma_{i \geq n+1} = 0$ are the Hankel singular values of the input-output system Σ. For square linear systems, the nonzero eigenvalues of the Hankel operator associated to the system are the nonzero eigenvalues of the cross Gramian defined as the solution, W_x, of $AW_x + W_xA + BC = 0$ [37].

We also have, for every $x_0 \in \mathbb{C}^n$,

$$\Psi_c \Psi_c^* x_0 = W_c x_0, \quad \Psi_o^* \Psi_o x_0 = W_o x_0. \qquad (2.13)$$

Thus W_c and W_o are the matrix representations of the operators $\Psi_c \Psi_c^*$ and $\Psi_o^* \Psi_o$ [11].

For model reduction, typically one looks for a gap in the singular values $\{\sigma_i\}$ for guidance as to where truncation should occur. If 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.

THEOREM 2.2. *[11] Consider a linear system (2.1) with its associated Hankel operator $\mathbb{H}$ (2.7). Let $\mathbb{H}_k$ be the Hankel operator of the reduced order linear system of order k. Then,*

$$||\mathbb{H} - \mathbb{H}_k|| = \sigma_{k+1} \qquad (2.14)$$

If $x(0) = 0$, the error between y, the output of the full order system, and y_r, the output of the reduced order system with k state variables, satisfies

$$||y - y_r||_2 \leq 2 \left(\sum_{j=k+1}^{n} \sigma_j \right) ||u||_2 \qquad (2.15)$$

If F is unstable then the controllability and observability quantities defined in (2.2) are undefined since the integrals will be unbounded. There may, however, still exist solutions to the Lyapunov equations (2.5) when F is unstable [48, 20]. Other approaches to balancing unstable linear systems exist (see [49, 50, 18, 54] for the method of LQG balancing for example).

Although several methods also exist for computing Q [26, 27], it is common to simply 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 Q is given by $Q = \Sigma^{\frac{1}{2}} U^\top Z^{-1}$. We also note that the problem of finding the coordinate change Q can be seen as an optimization problem [1] of the form $\min_Q \text{trace}[QW_c Q^* + Q^{-*} W_o Q^{-1}]$.

2.2. Balancing of Nonlinear Systems. In the nonlinear case, the energy functions L_c and L_o in (2.2) and (2.3) are obtained by solving both a Lyapunov and a Hamilton-Jacobi equation. Here we follow the development of Scherpen [36, 37]. Consider the nonlinear system

$$\Sigma : \begin{cases} \dot{x} & = & f(x) + \sum_{i=1}^{m} g_i(x)u_i, \\ y & = & h(x), \end{cases} \tag{2.16}$$

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

Assumption A: The linearization of (2.16) around the origin is controllable, observable and $A = \frac{\partial f}{\partial x}|_{x=0}$ is asymptotically stable.

The controllability operator $\Psi_c : \mathcal{U} \to X$ with $X = \mathbb{R}^n$ and $\mathcal{U} = L_2^m[0, \infty)$, and the observability operator $\Psi_o : X \to Y$ with $Y = L_2^p[0, \infty)$ for this system are defined by

$$\Psi_c : u \mapsto x^0 : \begin{cases} \dot{x} = -f(x) - g(x)u, & x(\infty) = 0, \\ x^0 = x(0) \end{cases} \tag{2.17}$$

$$\Psi_o : x^0 \mapsto y : \begin{cases} \dot{x} = f(x), & x^0 = x(0), \\ y = h(x) \end{cases} \tag{2.18}$$

As in the linear case, Ψ_c and Ψ_o represent the input-to-state behavior and the state-to-output behavior, respectively. The Hankel operator for the nonlinear system Σ in (2.16) is given by the composition of Ψ_c and Ψ_o

$$\mathbb{H} := \Psi_o \circ \Psi_c \tag{2.19}$$

Consider the norm-minimizing inverse $\Psi_c^{\dagger} : X \to \mathcal{U}$

$$\Psi_c^{\dagger} : x^0 \mapsto u := \operatorname{argmin}_{\Psi_c(u)=x^0} ||u||. \tag{2.20}$$

From this point of view L_c in (2.2) and L_o in (2.3) are

$$L_c(x^0) := \frac{1}{2}||\Psi_c^{\dagger}(x^0)||^2, \quad L_o(x^0) := \frac{1}{2}||\Psi_o(x^0)||^2 \tag{2.21}$$

THEOREM 2.3. *[36, 35] Consider the nonlinear system Σ defined in (2.16). 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*

$$\frac{\partial L_o}{\partial x}(x)f(x) + \frac{1}{2}h^{\top}(x)h(x) = 0, \quad L_o(0) = 0 \tag{2.22}$$

under the assumption that (2.22) has a smooth solution on W. Furthermore for all $x \in W$, $L_c(x)$ is the unique smooth solution of

$$\frac{\partial L_c}{\partial x}(x)f(x) + \frac{1}{2}\frac{\partial L_c}{\partial x}(x)g(x)g^{\top}(x)\frac{\partial^{\top} L_c}{\partial x}(x) = 0, \quad L_c(0) = 0 \tag{2.23}$$

under the assumption that (2.23) 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 (2.16) can be computed by way of a coordinate transformation. More precisely,

THEOREM 2.4. *[36, 35] Consider system* (2.16) *under Assumption A and the assumptions in Theorem 2.3. Then, there exists a neighborhood W of the origin and coordinate transformation $x = \varphi(z)$ on W converting the energy functions into the form $L_c(\varphi(z)) = \frac{1}{2} z^\top z$ and $L_o(\varphi(z)) = \frac{1}{2} \sum_{i=1}^{n} z_i^2 \sigma_i(z_i)^2$, where $\sigma_1(x) \geq \sigma_2(x) \geq \cdots \geq \sigma_n(x)$. 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 (2.22), (2.23) 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 [23, 22, 13]. Newman and Krishnaprasad [30] introduce 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 [25]. In this paper, we combine aspects of both data-driven approaches and analytic approaches by carrying out linear balancing of nonlinear control systems in a suitable RKHS.

3. Empirical Balancing of Nonlinear Systems in RKHS. We consider a general nonlinear system of the form

$$\begin{cases} \dot{x} &= f(x,u) \\ y &= h(x) \end{cases} \tag{3.1}$$

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 $x(0) = x_0$.

Hypothesis H:[3] The system (3.1) is zero-state observable, its linearization around the origin is controllable, and 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) [3], cf. Appendix for definition and key results on RKHSes. *Our approach hinges on the key modeling assumption that the nonlinear dynamical system can be embedded in an appropriate high (or possibly infinite) dimensional RKHS where the method of linear balancing can be applied.* More precisely, we will essentially assume that there is an RKHS $\mathcal{H}$ and maps $\Phi, \Psi : \mathbb{R}^n \to \mathcal{H}; x \mapsto \mathcal{H}$ such that controllability and observability energies of the nonlinear system (3.1) are "linearized", i.e. that in $\mathcal{H}$ they have an expression similar to the one in the linear case (2.4) and, therefore, can be written as

$$L_c(x) \cong \frac{1}{2} \Phi^T(x) \mathbb{W}_c^{-1} \Phi(x), \quad L_o(x) \cong \frac{1}{2} \Psi^T(x) \mathbb{W}_o \Psi(x), \tag{3.2}$$

with $\mathbb{W}_c, \mathbb{W}_o \in \mathbb{R}^{N \times N}$, $N \gg n$, are very large dimensional matrices[4].

---

[3]Let us note here that this assumption is similar to Assumption A above and that we made it mainly out of convenience but is not necessary as illustrated in the 2D and 7D examples below.

[4]If the system is affine in the input, we will use the PDEs (2.22) and (2.23) to find Φ, Ψ, $\mathbb{W}_c$ and $\mathbb{W}_o$. We leave such analysis for future work.

By "linearization" here, we mean *mapping the nonlinear system in a higher dimensional Hilbert space where linear theory can be applied.* To illustrate this point [46], consider a polynomial in $\mathbb{R}$, $p(x) = \alpha + \beta x + \gamma x^2$ where α, β, and γ are reals. If we consider the map $\Phi : \mathbb{R} \to \mathbb{R}^3$ defined as $\Phi(x) = [1 \ x \ x^2]^T$ then $p(x) = \boldsymbol{\alpha} \cdot [1 \ x \ x^2]^T = \boldsymbol{\alpha} \cdot \Phi(x)$ is an affine polynomial in the variable $\Phi(x)$.

Another example to illustrate our thought process is the one of Support Vector Machines (SVMs) that we referred to in the introduction. More precisely, consider the problem of classifying points in a data set $D = ((x_1, y_1), \cdots, (x_n, y_n))$ with $x_i \in X$ with X a set and $y_i = \pm 1$, i.e. trying to find $w \in \mathbb{R}^d$ with $||w||_2 = 1$ and $b \in \mathbb{R}$ such that $\langle w, x_i \rangle + b > 0$ for all i with $y_i = +1$, and $\langle w, x_i \rangle + b < 0$ for all i with $y_i = -1$, i.e. that the linear hyperplane characterized by (w, b) perfectly separates the set D into two groups of data points, the ones with $y_i = +1$ and the ones with $y_i = -1$. Sometimes, it is possible to find such an hyperplane but, in general, finding a linear hyperplane that perfectly separates a given data set D is not always possible and finding (w, b) will not be possible. To solve the classification problem, the SVM algorithm maps the input data $(x_1, \cdots, x_n)$ into a (possibly infinite-dimensional) Hilbert space $\mathcal{H}$, the so-called *feature space*, by a typically nonlinear map $\Phi : X \to \mathcal{H}$ called the *feature map*. Then one looks for a linear hyperplane that separates the data $((\Phi(x_1), y_1), \cdots, (\Phi(x_n), y_n))$, i.e. one looks for (w, b) in $\mathcal{H}$. When this is possible, the data in D will be classified in two categories, the ones with $y_i = +1$ and the ones with $y_i = -1$, but the separating curve for D will not be a linear hyperplane in the original space (even if it is a linear hyperplane in $\mathcal{H}$). An important property of SVMs is that for every dataset D without contradicting points, i.e. (x_i, y_i) and (x_j, y_j) with $x_i = x_j$ and $y_i \neq y_j$, there exists a feature map that allows the perfect separation by a hyperplane in the feature space [47].

Our aim is to generalize this way of thinking to nonlinear dynamical systems, i.e. given a problem for a nonlinear dynamical system, we map the state variables by a typically nonlinear map $\Phi : X \to \mathcal{H}$ where $\mathcal{H}$ is (possibly infinite-dimensional) Hilbert space in which the computations become simpler[5]. In our case, "simpler" means applying Linear Theory. In this paper, we will focus on the problem of approximation of nonlinear control systems by applying the method of linear balancing in RKHSes. We leave other applications for future work.

Covariance operators in RKHSes 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 RKHS, but capture nonlinearities in the original state space. The reduction approach we propose adapts ideas from kernel PCA (KPCA) [41] and is driven by a set of simulated or sampled system trajectories, extending and generalizing the work of Moore [29] and Lall et al. [25].

Our method works quite well since the controllability and the observability energies in the linear case can be expressed as inner products (2.4) and *working in RKHSes allows to find nonlinear versions of linear algorithms that can be expressed in terms of inner products* (this is the so-called *kernel trick* in Learning Theory, see Appendix A for more explanations). Hence the empirical Gramians defined below for nonlinear systems can be viewed as reasonable approximations of the controllability and observability energies for nonlinear systems.

In the development below we lift state vectors of the system (3.1) into a Hilbert

---

[5]One could also think of the methods in Quantum Mechanics where one constructs a Hilbert space from measurements in order to perform computations.

space $\mathcal{H}$, i.e. we consider a mapping $\Phi : \mathbb{R}^n \to \mathcal{H}$ and analyze the nonlinear system whose state is $\Phi(x)$.

3.1. Empirical Gramians in RKHS. Following [29], we estimate the controllability Gramian by exciting each coordinate of the input with impulses[6] while setting $x_0 = 0$. One can also further excite using rotations of impulses as suggested in [25], however for simplicity we consider only the original signals proposed in [29]. 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

$$X(t) = \begin{bmatrix} x^1(t) & \cdots & x^m(t) \end{bmatrix} \in \mathbb{R}^{n \times m}, \tag{3.3}$$

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

$$W_c = \frac{1}{m} \int_0^\infty X(t)X(t)^\top dt. \tag{3.4}$$

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

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^N X(t_i)X(t_i)^\top. \tag{3.5}$$

As described in [29], 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

$$W_o = \frac{1}{p} \int_0^\infty Y(t)^\top Y(t) dt \tag{3.6}$$

and

$$\widehat{W}_o = \frac{T}{pN} \sum_{i=1}^N \widetilde{Y}(t_i)\widetilde{Y}(t_i)^\top \tag{3.7}$$

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

$$d_j(t_i) = \left(y_j^1(t_i), \ldots, y_j^n(t_i) \right)^\top \in \mathbb{R}^n, \quad j = 1, \ldots, p, \ i = 1, \ldots, N \tag{3.8}$$

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 follow.

---

[6]This is not a limitation of our approach, other input signals can be used such as a white gaussian noise, cf. [6] for preliminary results.

3.2. 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. It is worth emphasizing however that we will be co-diagonalizing *two* Gramians in the feature space by way of a *non-orthogonal* transformation; the process bears a resemblance to (K)PCA, and yet is distinct. Indeed the favorable properties associated with an orthonormal basis are no longer available, the quantities we will in practice diagonalize are different, and the issue of data-centering must be considered with some additional care.

First note that the empirical controllability Gramian $\widehat{W}_c$ can be viewed as the sample covariance of a collection of $N \cdot m$ vectors, scaled by T

$$\widehat{W}_c = \frac{T}{mN} \sum_{i=1}^{N} X(t_i) X(t_i)^\top = \frac{T}{mN} \sum_{i=1}^{N} \sum_{j=1}^{m} x^j(t_i) x^j(t_i)^\top \tag{3.9}$$

where $X(t)$ is defined in (3.3) and the observability Gramian can be similarly viewed as the sample covariance of a collection of $N \cdot p$ vectors

$$\widehat{W}_o = \frac{T}{pN} \sum_{i=1}^{N} \sum_{j=1}^{p} d_j(t_i) d_j(t_i)^\top \tag{3.10}$$

where the d_j are defined in Equation (3.8).

We can thus consider three quantities of interest:
- The *controllability kernel matrix* $K_c \in \mathbb{R}^{Nm \times Nm}$ of kernel products

$$(K_c)_{\mu\nu} = K(x_\mu, x_\nu) = \langle \Phi(x_\mu), \Phi(x_\nu) \rangle_{\mathcal{H}} \tag{3.11}$$

  for $\mu, \nu = 1, \ldots, Nm$ where we have re-indexed the set of vectors $\{x^j(t_i)\}_{i,j} = \{x_\mu\}_\mu$ to use a single linear index.
- The *observability kernel matrix* $K_o \in \mathbb{R}^{Np \times Np}$,

$$(K_o)_{\mu\nu} = K(d_\mu, d_\nu) = \langle \Phi(d_\mu), \Phi(d_\nu) \rangle_{\mathcal{H}} \tag{3.12}$$

  for $\mu, \nu = 1, \ldots, Np$, where we have again re-indexed the set $\{d_j(t_i)\}_{i,j} = \{d_\mu\}_\mu$ for simplicity.
- The *Hankel kernel matrix* $K_{o,c} \in \mathbb{R}^{Np \times Nm}$,

$$(K_{o,c})_{\mu\nu} = K(d_\mu, x_\nu) = \langle \Phi(d_\mu), \Phi(x_\nu) \rangle_{\mathcal{H}} \tag{3.13}$$

  for $\mu = 1, \ldots, Np$, $\nu = 1, \ldots, Nm$.

We have chosen the suggestive terminology "Hankel kernel matrix" above because the square-roots of the nonzero eigenvalues of the matrix $K_{o,c} K_{o,c}^\top$ are the empirical Hankel singular values of the system mapped into feature space[7], where we assume

---

[7]The relation between the singular values and the eigenfunctions of the Hankel operator for the nonlinear system given by $\mathbb{H}$ in (2.19) and the extension of (2.11)-(2.12) to the nonlinear setting using the empirical Hankel singular values and eigenvectors of $K_{o,c} K_{o,c}^\top$ is an open problem that we leave for future work.

that the method of linear balancing can be applied. This assertion will be proved immediately below. Note that ordinarily, $Nm, Np \gg n$ and K_c, K_o will be rank deficient.

Before proceeding we consider the issue of data centering in feature space. PCA and kernel PCA assume that the data have been centered in order to make the problem translation invariant. In the setting considered here, we have two distinct sets of data: the observability samples and the controllability samples. A reasonable centering convention centers the data in each of these datasets separately. Let $\mathbf{\Psi}$ denote the matrix whose columns are the observability samples mapped into feature space by the feature map Φ, and let $\mathbf{\Phi}$ be the matrix similarly built from the feature space representation of the controllability samples. Then

$$K_o = \mathbf{\Psi}^\top \mathbf{\Psi}, \quad K_c = \mathbf{\Phi}^\top \mathbf{\Phi}, \quad \text{and} \quad K_{o,c} = \mathbf{\Psi}^\top \mathbf{\Phi}. \tag{3.14}$$

The above equation reduces to (3.5) and (3.7) in the linear case. In fact, when $\Phi(x) = x$, $(K_c)_{\mu\nu} = \langle x_\mu, x_\nu \rangle_{\mathbb{R}^n}$, $(K_o)_{\mu\nu} = \langle d_\mu, d_\nu \rangle_{\mathbb{R}^n}$, $(K_{o,c}) = \langle d_\mu, x_\nu \rangle_{\mathbb{R}^n}$.

Assume for the moment that there are M observability data samples and N controllability samples, and let $\mathbf{1}_N, \mathbf{1}_M$ denote the length N, M vectors of all ones, respectively. We can define centered versions of the feature space data matrices $\mathbf{\Phi}, \mathbf{\Psi}$ as

$$\widetilde{\mathbf{\Phi}} = \mathbf{\Phi} - \mu_c \mathbf{1}_N^\top, \qquad \widetilde{\mathbf{\Psi}} = \mathbf{\Psi} - \mu_o \mathbf{1}_M^\top \tag{3.15}$$

where $\mu_c := N^{-1}\mathbf{\Phi}\mathbf{1}_N$ and $\mu_o := M^{-1}\mathbf{\Psi}\mathbf{1}_M$. We will need two centered quantities in the development below. Let us note here that, in practice, we do not need to compute μ_c and μ_o as detailed above. Moreover, there is no need to compute the embedding Φ since all computations are done in terms of the kernels.

The first centered quantity we consider is the centered version of $K_{o,c}$, namely $\widetilde{K}_{o,c} = \widetilde{\mathbf{\Psi}}^\top \widetilde{\mathbf{\Phi}}$. Although one cannot compute μ_c, μ_o explicitly from the data, we can compute $\widetilde{K}_{o,c}$ by observing that

$$
\begin{aligned}
\widetilde{K}_{o,c} &= \left(\mathbf{\Psi} - \mu_o \mathbf{1}_M^\top\right)^\top \left(\mathbf{\Phi} - \mu_c \mathbf{1}_N^\top\right) \\
&= K_{o,c} - \tfrac{1}{N}K_{o,c}\mathbf{1}_N\mathbf{1}_N^\top - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top K_{o,c} + \tfrac{1}{NM}\mathbf{1}_M\mathbf{1}_M^\top K_{o,c}\mathbf{1}_N\mathbf{1}_N^\top.
\end{aligned}
\tag{3.16}
$$

The second quantity we'll need is a centered version of the *empirical observability feature map*

$$\mathbf{k}_o(x) := \mathbf{\Psi}^\top \Phi(x) = \left(K(x, d_1), \dots, K(x, d_M)\right)^\top \tag{3.17}$$

where $x \in \mathbb{R}^d$ is the state variable and the observability samples $\{d_j\}$ are again indexed by a single variable as in Equation (3.12). Centering follows reasoning similar to that of the Hankel kernel matrix immediately above:

$$
\begin{aligned}
\widetilde{\mathbf{k}}_o(x) &= \left(\mathbf{\Psi} - \mu_o \mathbf{1}_M^\top\right)^\top \left(\Phi(x) - \mu_c\right) \\
&= \mathbf{k}_o(x) - \tfrac{1}{N}K_{o,c}\mathbf{1}_N - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top \mathbf{k}_o(x) + \tfrac{1}{NM}\mathbf{1}_M\mathbf{1}_M^\top K_{o,c}\mathbf{1}_N.
\end{aligned}
\tag{3.18}
$$

Note: *Throughout the remainder of this paper we will drop the special notation $\widetilde{K}_{o,c}$, $\widetilde{\mathbf{k}}_o(x)$ and assume that $K_{o,c}, \mathbf{k}_o(x)$ are centered appropriately.*

With the quantities defined above, we can co-diagonalize the empirical Gramians (balancing) and reduce the dimensionality of the state variable (truncation) in feature

space by carrying out calculations in the original data space. As we are assuming that the method of linear balancing applies in the feature space, the order of the model can be reduced by discarding small Hankel values $\{\Sigma_{ii}\}_{i=q+1}^{n}$, and projecting onto the subspace associated with the first $q < n$ largest eigenvalues. The following key result describes this process:

THEOREM 3.1 (Balanced Reduction in Feature Space). *Consider the nonlinear control system (3.1) and its responses (3.3) and (3.8) to impulses from the input and initial condition, respectively. Let K be a Mercer kernel, $K_{o,c}$ be the Hankel kernel matrix defined in (3.13), and $K_{o,c}K_{o,c}^{\top} = V\Sigma^2 V^{\top}$ be its SVD decomposition with $\Sigma = diag\{\sigma_1, \cdots \sigma_{Np}\}$ and $\sigma_i \geq \sigma_{i+1}$ for $i = 1, \cdots Np$. Also consider the controllability kernel matrix K_c (3.11), the observability kernel matrix K_o (3.12).*

If there is a spectral gap in Σ, i.e. there is q such that $\sigma_q >> \sigma_{q+1}$ then balanced reduction in the RKHS can be accomplished by applying the state-space reduction map $\Pi : \mathbb{R}^n \to \mathbb{R}^q$ given by

$$\Pi(x) = T_q^{\top}\mathbf{k}_o(x), \quad x \in \mathbb{R}^n \tag{3.19}$$

where $T_q = V_q\Sigma_q^{-1/2}$, V_q are the eigenvectors that correspond to the largest q Hankel singular values. $\mathbf{k}_o(x)$ is the empirical observability feature map (3.18).

Proof. We assume the data have been centered in feature space. Let $\boldsymbol{\Phi}$ be a matrix with columns $\{\Phi(x^j(t_i))\}, i = 1, \ldots, N, j = 1, \ldots, m$, so that

$$X = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}, \tag{3.20}$$

is the feature space controllability Gramian counterpart to Equation (3.9). Similarly, let $\boldsymbol{\Psi}$ be a matrix with columns $\{\Phi(d_j(t_i))\}, i = 1, \ldots, N, j = 1, \ldots, p$, so that

$$Y = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}, \tag{3.21}$$

is the feature space observability Gramian counterpart to Equation (3.10). Since by definition $K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{F}}$, we also have that $K_c = \boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}$ and $K_o = \boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}$. In general the Gramians X, Y are infinite dimensional whereas the kernel matrices K_c, K_o are necessarily of finite dimension.

We now carry out linear balancing on (X, Y) in the feature space (RKHS). First, take the SVD of $X^{1/2}\boldsymbol{\Psi}$ so that

$$U\Sigma V^{\top} = X^{1/2}\boldsymbol{\Psi} \tag{3.22}$$

$$U\Sigma^2 U^{\top} = (X^{1/2}\boldsymbol{\Psi})(X^{1/2}\boldsymbol{\Psi})^{\top} = X^{1/2}YX^{1/2} \tag{3.23}$$

$$V\Sigma^2 V^{\top} = (X^{1/2}\boldsymbol{\Psi})^{\top}(X^{1/2}\boldsymbol{\Psi}) = \boldsymbol{\Psi}^{\top}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}\boldsymbol{\Psi} = K_{o,c}K_{o,c}^{\top}. \tag{3.24}$$

The last equality in Equation (3.23) follows since X is symmetric and therefore $X^{1/2}$ is too. The linear balancing transformation is then given by

$$\mathcal{T} = \Sigma^{1/2}U^{\top}X^{-1/2}, \tag{3.25}$$

and one can readily verify that $\mathcal{T}X\mathcal{T}^{\top} = \mathcal{T}^{-\top}Y\mathcal{T}^{-1} = \Sigma$. Here, inverses should be interpreted as peudo-inverses when appropriate[8]. From Equations (3.22)-(3.24), we

---

[8]Such as in the case of $X^{-1/2}$ when the number of data points is less than the dimension of the RKHS.

see that $U^\top = \Sigma^{-1}V^\top\mathbf{\Psi}^\top X^{1/2}$ and thus $\mathcal{T} = \Sigma^{-1/2}V^\top\mathbf{\Psi}^\top$. We can project an arbitrary mapped data point $\Phi(x)$ onto the (balanced) "principal" subspace of dimension q spanned by the first q rows of $\mathcal{T}$ by computing

$$\mathcal{T}_q\Phi(x) = \Sigma_q^{-1/2}V_q^\top\mathbf{\Psi}^\top\Phi(x) = \Sigma_q^{-1/2}V_q^\top\mathbf{k}_o(x) \tag{3.26}$$

where $\mathbf{k}_o(x) := \mathbf{\Psi}^\top\Phi(x)$ is the empirical observability feature map, recalling that V_q is the matrix formed by taking the top q eigenvectors of $K_{o,c}K_{o,c}^\top$ by Equation (3.24).
☐

We note that square roots of the non-zero eigenvalues of $K_{o,c}K_{o,c}^\top$ are exactly the Hankel singular values of the system mapped into the feature space, under the assumption of linearity in the feature space. This can be seen by noting that $\lambda_+(YX) = \lambda_+(X^{1/2}YX^{1/2}) = \lambda_+(K_{o,c}K_{o,c}^\top)$, where $\lambda_+(\cdot)$ refers to the non-zero eigenvalues of its argument. In practice, we compute the largest eigenvalues of $K_{o,c}K_{o,c}^\top$ instead of performing its SVD.

In Section 4 below we show how to use the nonlinear reduction map (3.19) to realize a closed, reduced order system which can approximate the original system to a high degree of accuracy.

4. 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,

$$\dot{x}_r = \left(J_\Pi(x)f(x,u)\right)\big|_{x=\Pi^\dagger(x_r)} \tag{4.1}$$

where $\Pi^\dagger$ refers to an appropriate notion (to be defined) of the inverse of Π. However we are faced with the difficulty that the map Π is not in general surjective (even if $q = n$), and moreover one cannot guarantee that an arbitrary point in the RKHS has a non-empty preimage under Φ [28]. 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 $\Pi^\dagger$ 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 $J_\Pi(\Pi^\dagger(x_r))$ which takes into account further properties of the reproducing kernel as well as the fact that the Jacobian is to be evaluated at $x = \Pi^\dagger(x_r)$ in particular. In both cases, the important ability of the map Π to capture strong nonlinearities will not be significantly diminished.

4.1. 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 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 using the representer theorem in Appendix A, with the reduction map Π. It is reasonable to expect that this approximation will be better than directly computing $f(\Pi^\dagger(x_r), u)$ using, for instance, a Taylor expansion approximation for $\Pi^\dagger$ which may ignore important nonlinearities at a stage where crude approximations must be avoided. Let us note here that, for this approximation part, we are using the representer theorem as described in Appendix

13

A and that the RKHS we use is not necessarily the same as the one where we have been performing balancing as in the previous section.

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 (cf. [53, 12, 10, 34]). As mentioned in section A, equation (A.12), $\hat{f}_i$ takes the form

$$\hat{f}_i(z) = \hat{f}_i(\Pi(x), u) = \sum_{j=1}^{\ell} c_j^i K^f(z, z_j), i = 1, \dots, n, \tag{4.2}$$

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 $\{c_j^i\}|_{(i,j)=(1,1)}^{(i,j)=(n,\ell)}$ comprise a set of coefficients found using the regularized least squares (RLS) algorithm and satisfying the algebraic equation (A.13) with $\mathbf{s} = \{(f_i(x_j, u_j), z_j)\}_{j=1}^{\ell}$ as training examples. 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 [36] 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 can estimate the functions f and g_i from examples $\{(\Pi(x_j), f(x_j))\}_j$ and $\{(\Pi(x_j), g(x_j))\}_j$.

4.2. Approximation of the Jacobian Contribution. We turn to approximating the component $J_\Pi\big(\Pi^\dagger(x_r)\big)$ appearing in Equation (4.1).

4.2.1. 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 $\Pi^\dagger(x_r)$ (in the first-order, least-norm sense) as

$$\widehat{\Pi}^\dagger(x_r) := \big(J_\Pi(a)\big)^\dagger(x_r - \Pi(a)) + a. \tag{4.3}$$

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 [24]. If $\widetilde{\mathbf{k}}_o(x)$ is the centered version of the length M vector $\mathbf{k}_o(x)$ defined by (3.17), and since $\Pi(x) = T_q^T \widetilde{\mathbf{k}}_o(x)$, then

$$J_\Pi(x) = \frac{\partial \Pi(x)}{\partial x} = T_q^\top \big(I - \tfrac{1}{M}\mathbf{1}_M \mathbf{1}_M^\top\big) \frac{\partial \mathbf{k}_o(x)}{\partial x} \tag{4.4}$$

where $\mathbf{1}_M$ is the length M vector of all ones. $J_\Pi(x)$ is a matrix in $\mathbb{R}^{q \times n}$ since $T_q = V_q \Sigma_q^{-\frac{1}{2}} \in \mathbb{R}^{M \times q}$ with $V_q \in \mathbb{R}^{M \times q}$, $K_{oc} \in \mathbb{R}^{Np \times Nm}$, $\Sigma_q \in \mathbb{R}^{q \times q}$, $(I - \frac{1}{M} \mathbf{1_M} \mathbf{1_M}^T) \in \mathbb{R}^{M \times M}$, $k_o(x) \in \mathbb{R}^{M \times 1}$, $\frac{\partial k_o(x)}{\partial x} \in \mathbb{R}^{M \times n}$. An example calculation of $\left(\partial_x \mathbf{k}_o(x) \right)_i = \partial_x K(x, d_i)$ in the case of a polynomial kernel is given in the section immediately below.

4.2.2. 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 $J_\Pi(\Pi^\dagger(x_r))$. 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 $x_r = \Pi(x)$ given by (3.19), if Π were 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* $(\Phi \circ \Pi^\dagger)(x_r)$ as the solution to the convex optimization problem

$$\begin{aligned} \min_{z \in \mathcal{H}} \quad & \|z\|_{\mathcal{H}} \\ \text{subj. to} \quad & \|M_q z - x_r\|_{\mathbb{R}^k} = 0 \end{aligned} \tag{4.5}$$

where $M_q : \mathcal{H} \to \mathbb{R}^k$ is defined as in Equation (3.26). 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 $\Pi \circ \Phi^{-1} = M_q(\Phi \circ \Phi^{-1}) = M_q z$. 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 $\Pi^\dagger$ is used in our formulation of the reduction map and corresponding reduced dynamical system.

One will recognize that the solution z^* to (4.5) is just the Moore-Penrose pseudoinverse $z^* = M_q^\dagger x_r$. Inserting this solution into the feature map representation of a kernel K gives the following definition for $K(\Pi^\dagger(x_r), y)$:

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

where the final equality follows applying Equations (3.22)-(3.24) and T_q is defined as in Theorem 3.1. Substituting into the derivative for a polynomial kernel $K = K_d$

15

gives

$$\left.\frac{\partial K_d(x,y)}{\partial x}\right|_{x=\Pi^\dagger(x_r)} = d\big\langle x_r, (T_q^\top K_o T_q)^{-1}\Pi(y)\big\rangle^{\frac{d-1}{d}} y^\top$$

which immediately gives an expression for $J_\Pi(\Pi^\dagger(x_r))$ using (4.4):

$$J_\Pi(x)|_{x=\Pi^\dagger(x_r)} = T_q^\top\big(I - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top\big)d\big\langle x_r, (T_q^\top K_o T_q)^{-1}\Pi(d_i)\big\rangle^{\frac{d-1}{d}} d_i^\top \qquad (4.6)$$

Note that this approximation is global in the sense that the $q \times q$ matrix inverse $(T_q^\top K_o T_q)^{-1}$ need only be computed once[9]; no updating is required during simulation of the closed system.

4.3. Reduced System Dynamics. Given an estimate $\hat{f}\big(\Pi(x),u\big)$ of $f(x,u)$ in the RKHS $\mathcal{H}_f$ as in (4.2) and a notion of $J_\Pi\big(\Pi^\dagger(x_r)\big)$ from (4.4) and the section above, we can write down a closed dynamical system on the reduced statespace. We have

$$\dot{x}_r \approx \big(J_\Pi(x)\hat{f}(\Pi(x),u)\big)\big|_{x=\Pi^\dagger(x_r)}$$
$$\approx \big(J_\Pi(x)\big)\big|_{x=\Pi^\dagger(x_r)}\mathbf{C}^\top\mathbf{k}^f(x_r,u)$$
$$= T_q^\top\big(I - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top\big)\frac{\partial\mathbf{k}_o(x)}{\partial x}\Big|_{x=\Pi^\dagger(x_r)}\mathbf{C}^\top\mathbf{k}^f(x_r,u)$$
$$:= T_q^\top\big(I - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top\big)J_\mathbf{k}\big(\Pi^\dagger(x_r)\big)\mathbf{C}^\top\mathbf{k}^f(x_r,u) \qquad (4.7)$$

where $\mathbf{C}$ is a matrix with the vectors $\mathbf{c}_i$ as its rows, and $J_\mathbf{k}\big(\Pi^\dagger(x_r)\big) := \frac{\partial\mathbf{k}_o(x)}{\partial x}\big|_{x=\Pi^\dagger(x_r)}$ is the Jacobian of the empirical feature map defined in Equation (3.17). Here the expression $J_\mathbf{k}\big(\Pi^\dagger(x_r)\big)$ should be interpreted as notation for either of the Jacobian approximations suggested in Section 4.2.

Equation (4.7) is seen to give a closed nonlinear control system expressed solely in terms of the reduced variable $x_r \in \mathbb{R}^q$:

$$\begin{cases} \dot{x}_r = J_\Pi(x)|_{x=\Pi^\dagger(x_r)}\mathbf{C}^\top\mathbf{k}^f(x_r,u) \\ \hat{y} = \hat{h}(x_r) \end{cases} \qquad (4.8)$$

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-surjectivity 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 (4.7) to future work[10].

---

[9]We use the word "inverse" loosely. In practice one would use a numerically stable method, such as an LU-factorization, which can be used to rapidly compute $A^{-1}b$ for fixed A but many different b.

[10]There are different sources of error in this process of approximation. A first one due to truncation in the RKHS which could be characterized by extending (2.15) using the eigenvalues of the Hankel kernel matrix (3.13). There is also the error in finding the pseudo inverse of Φ in (4.5), i.e. that it is tempting to write $||y - \hat{y}|| \leq 2\Big(\sum_{j=q+1}^{Np}\sigma_j\Big)||u||$ modulo the fact that Φ may not be surjective and that $\hat{y}$ in (4.8) is not what we will end up observing as output for the reduced order system in $\mathbb{R}^q$ and the fact that there is also the error in approximating $f(x,u)|_{x=\Pi^\dagger(x_r)}$ by $\mathbf{C}^\top\mathbf{k}^f(x_r,u)$ that might be computed using results from [42] for a given u as illustrated in section 4.5 below.

4.4. Outputs of the Reduced System. Analogous to the case of the dynamics f, we are faced with two possibilities for approximating

$$y = h\big(\Pi^\dagger(x_r)\big). \qquad (4.9)$$

We can apply a crude Taylor series approximation to estimate $\Pi^\dagger$ and therefore $h\big(\Pi^\dagger(x_r)\big)$, or as in Section 4.1, we can estimate a map

$$(\hat h \circ \Pi): \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R}^p, \\ x_r & \mapsto & y \end{array} \qquad (4.10)$$

from the reduced state space to the output space directly, using RKHS methods. Given samples $\mathbf{s} = \{\Pi(x_j), y_j\}_{j=1}^\ell$, each coordinate function $\big(\hat h_i\big)_{i=1}^p$ is given in the familiar form (A.12)

$$\hat h_i(\Pi(x)) = \sum_{j=1}^\ell b_j^i K^h\big(\Pi(x), \Pi(x_j)\big), \qquad (4.11)$$

where K^h is the kernel chosen to define the RKHS, and may be different for each coordinate and the b_j satisfy the algebraic set of equations (A.13).

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.

4.5. On Error Estimates. The problems of approximating the functions f and h from time series can be viewed as learning problem as described in Appendix A and a theoretical justification of our algorithm is guaranteed by the error estimates in Theorem A.12. In fact, for the dynamical system $x(k+1) = f(x(k))$ (resp. the control system $x(k+1) = f(x(k), u(k))$), we have that f^* in (A.5) is the map $f^*(x) = f_i(x)$ (resp. $f^*(z) = f_i(x, u)$ with $z = (x, u)'$) and the samples $\mathbf{s}$ in (A.7) are $(x(k), x_i(k+1) + \eta_i)$ (resp. $\left(\begin{pmatrix} x(k) \\ u(k) \end{pmatrix}, x_i(k+1) + \eta_i\right)$).

Here f_i is the unknown map $x(k) \to x_i(k+1)$ (resp. $\begin{pmatrix} x(k) \\ u(k) \end{pmatrix} \to x_i(k+1)$) and it plays the role of the unknown function (A.5).

The initial condition $x(0)$ (resp. the input) is known and η_i are distributed according to a probability measure ρ_x that satisfies the following condition (this is the *Special Assumption* in [43]).

Assumption The measure ρ_x is the marginal on $X = \mathbb{R}^n$ of a Borel measure ρ on $X \times \mathbb{R}$ with zero mean supported on $[-M_x, M_x], M_x > 0$.

Let's note that in [43], the authors do not consider time series and that we apply their results to time series. In the case of learning the map (4.9) $\hat h \circ \Pi : \mathbb{R}^n \to \mathbb{R}^p$, we are looking at learning the map $x_r \mapsto y$ from the samples $\mathbf{s} = \{\Pi(x_j), y_j\}_{j=1}^\ell$ while in the case of learning the vector-valued map $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, this corresponds to learning the i-th coordinate of the map $(\tilde x, u) \mapsto f(x, u)$ for $i = 1, \cdots, n$.

In the case of the dynamics $x(k+1) = f^*(x(k))$ (resp. the control system $x(k+1) = f^*(x(k), u(k))$), the error estimate (A.20) has the form

$$||\hat x_i(k+1) - x_i(k+1)||^2 \le 2C_{\bar x}\mathcal{E}_{\text{samp}} + 2||x(k+1)||_K^2(\gamma + 8C_{\bar x}\Delta), \qquad (4.12)$$

where $||x_i(k+1)||_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{c_{i,j}^2}{\lambda_j}$.

In the case of the unknown map (4.9) $y = h\big(\Pi^{\dagger}(x_r)\big) = \tilde{h}(x_r)$, the error estimate (A.20) has the form

$$||\hat{y} - y||^2 \leq 2C_{\bar{x}}\mathcal{E}_{\text{samp}} + 2||\tilde{h}||_K^2(\gamma + 8C_{\bar{x}}\Delta), \tag{4.13}$$

where $||\tilde{h}||_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \frac{\tilde{c}_{i,j}^2}{\tilde{\lambda}_j}$.

The first term in the right hand side of inequalities (4.12)-(4.13) represents the error due to the noise (sampling error) and the second term represents the error due to regularization (regularization error) and the finite-number of samples (integration error).

4.6. Algorithm Summary. To summarize, the approach we have proposed proceeds as follows

1. Given a nonlinear control system (2.16), let $u^i(t) = \delta(t)e_i$ be the i-th excitation signal for $i = 1, \ldots, m$, and let $x^i(t) : t \in [0, \infty) \mapsto x^i(t) \in \mathbb{R}^n$ be the corresponding response of the system. Run the system and sample the trajectories at times $\{t_j\}_{j=1}^N$ to generate a collection of $N \cdot m$ vectors $\{x^i(t_j) \in \mathbb{R}^n\}$.

2. Fixing $u(t) = 0$ and setting $x_0 = e_i$ for $i = 1, \ldots, n$ (separately), measure the corresponding system output responses $y^i(t) : t \in [0, \infty) \mapsto y^i(t) \in \mathbb{R}^p$. As before, sample the responses at times $\{t_j\}_{j=1}^N$ and save the collection of $N \cdot p$ vectors $\{d_k(t_i)\}$ defined as

$$d_k(t_j) = \big(y_k^1(t_j), \ldots, y_k^n(t_j)\big)^{\top} \in \mathbb{R}^n, \quad k = 1, \ldots, p, \; j = 1, \ldots, N \tag{4.14}$$

3. Choose a kernel K defining a RKHS $\mathcal{H}$, and form the Hankel kernel matrix $K_{o,c} \in \mathbb{R}^{Np \times Nm}$,

$$(K_{o,c})_{\mu\nu} = K(d_{\mu}, x_{\nu}) \quad \mu = 1, \ldots, Np, \; \nu = 1, \ldots, Nm \tag{4.15}$$

where we have re-indexed the sets $\{d_k(t_i)\} = \{d_{\mu}\}, \{x^i(t_j)\} = \{x_{\nu}\}$ to use single indices.

4. Compute the eigendecomposition $K_{o,c}K_{o,c}^{\top} = V\Sigma^2 V^{\top}$ assuming $K_{o,c}$ has been centered according to Equation (3.16).

5. 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

$$\begin{aligned} \Pi : \mathbb{R}^n &\to \mathbb{R}^q \\ x &\mapsto x_r = \Pi(x) = T_q^{\top}\mathbf{k}_o(x) \end{aligned} \tag{4.16}$$

where $T_q = V_q\Sigma_q^{-1/2}$ and $\mathbf{k}_o(x)$ is the centered empirical observability feature map given by Equation (3.18).

$$\begin{aligned} \widetilde{\mathbf{k}}_o(x) &= \big(\boldsymbol{\Psi} - \mu_o\mathbf{1}_M^{\top}\big)^{\top}\big(\Phi(x) - \mu_c\big) \\ &= \mathbf{k}_o(x) - \tfrac{1}{N}K_{o,c}\mathbf{1}_N - \tfrac{1}{M}\mathbf{1}_M\mathbf{1}_M^{\top}\mathbf{k}_o(x) + \tfrac{1}{NM}\mathbf{1}_M\mathbf{1}_M^{\top}K_{o,c}\mathbf{1}_N. \end{aligned} \tag{4.17}$$

and

$$k_o(x) = \Psi^T\Phi(x) = (K(x, d_1), \cdots, K(x, d_M))^T \tag{4.18}$$

6. From input/output pairs or simulated/measured trajectories, learn approximations of the dynamics and output function defined on the reduced state space using, for instance, the representation theorem in section §A (equations (A.12)-(A.13). The RKHS used to approximate these functions need not be the same as the RKHS in which balanced truncation was carried out.
For the approximation of the dynamics, the representer theorem will be applied to learn the $i-th$ coordinate of the map $(\Pi(x), u) \mapsto f(x, u)$ using the samples $\mathbf{s} = (f_i(x_j, u_j), (\Pi(x_j), u_j))|_{j=1}^{\ell}$.
For the approximation of the output map, the representer theorem will be applied to learn the map $x_r = \Pi(x) \mapsto y$ using the samples $\mathbf{s} = (\Pi(x_j), y_j)|_{j=1}^{\ell}$.
7. Approximate the Jacobian contribution as described in section §4.2. If the chosen kernels are polynomials, directly use (4.6).
8. Combine the approximations to determine an expression for a closed, reduced, nonlinear dynamical system (4.8) as described in sections §. 4.3 and §. 4.4.

5. Experiments. We demonstrate an application of our method on two examples appearing in [31] (Examples 3.1. and 3.2, pgs. 52-54).

5.1. Two-Dimensional Exactly Reducible System. Consider the nonlinear system

$$
\begin{aligned}
\dot{x}_1 &= -3x_1^3 + x_1^2 x_2 + 2x_1 x_2^2 - x_2^3 \\
\dot{x}_2 &= 2x_1^3 - 10x_1^2 x_2 + 10x_1 x_2^2 - 3x_2^3 - u \\
y &= 2x_1 - x_2 .
\end{aligned}
\tag{5.1}
$$

It can be shown that this system has the same input-output relationship as the system $\dot{y} = -y^3 + u$ by rearranging terms so that

$$
\begin{aligned}
\dot{x}_1 &= -(2x_1 - x_2)^2 x_1 + (x_1 - x_2)^3 \\
\dot{x}_2 &= -(2x_1 - x_2)^2 x_2 + 2(x_1 - x_2)^3 - u \\
y &= 2x_1 - x_2 .
\end{aligned}
$$

Defining the new variables $z_1 = 2x_1 - x_2$ and $z_2 = x_1 - x_2$, the system can then be re-written

$$
\begin{aligned}
\dot{z}_1 &= -z_1^3 + u \\
\dot{z}_2 &= -z_1^2 z_2 - z_2^3 + u \\
y &= z_1 .
\end{aligned}
\tag{5.2}
$$

It can be seen that the variable z_2 may be truncated because it doesn't appear in the expression of the output and thus doesn't affect z_1. Let's note here that the change of variables $x \mapsto z$ is a linear one and, importantly, that the method of nonlinear balancing as developed by Scherpen doesn't apply for this example since the system is not linearly controllable and the jacobian of the linearization has zero eigenvalues.

5.2. Seven-Dimensional System. We will also consider a 7-dimensional nonlinear system with one dimensional input and output:

$$
\begin{aligned}
\dot{x}_1 &= -x_1^3 + u \\
\dot{x}_2 &= -x_2^3 - x_1^2 x_2 + 3x_1 x_2^2 - u \\
\dot{x}_3 &= -x_3^3 + x_5 + u \\
\dot{x}_4 &= -x_4^3 + x_1 - x_2 + x_3 + 2u \\
\dot{x}_5 &= x_1 x_2 x_3 - x_5^3 + u \\
\dot{x}_6 &= x_5 - x_6^3 - x_5^3 + 2u \\
\dot{x}_7 &= -2x_6^3 + 2x_5 - x_7 - x_5^3 + 4u
\end{aligned}
\tag{5.3}
$$

with $y = x_1 - x_2^2 + x_3 + x_4 x_3 + x_5 - 2x_6 + 2x_7$.

Here also the method of nonlinear balancing as developed by Scherpen doesn't apply for this example since the system is not linearly controllable and the jacobian of the linearization has zero eigenvalues.

5.3. Experimental Setup. For both systems 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 Hankel kernel matrix $K_{o,c}$ given by the third degree polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^3$. For the 2-D system we retained one component, and for the 7-D system we retained two for the sake of variety. Thus the reduction map Π was defined by taking the top one or two eigenvectors (scaled columns of T) corresponding to the largest Hankel singular values, giving a reduced state space of dimension one or two for the 2-D and 7-D systems, respectively.

Next, a map from the reduced variable x_r to $\dot{x}$ was estimated following Section 4.1. The same procedure was followed in both experiments. The control input was chosen to be a 10hz square wave with peaks at ± 1 at 50% duty cycle, and 1000 samples from the simulated system in the interval $[0, 5s]$ were mapped down using Π and then used to solve the 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 function 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 an offset, and used a fast leave-one-out cross-validation (LOOCV) computation [33] to select the optimal regularization parameter.

We followed a similar process to learn the output function $y = \hat{h}(x_r)$ for both systems. Here we used a 10Hz square wave control input (peaks at ± 2, 50% duty cycle), 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 reciprocal of the average squared-distance between the training examples. We again used LOOCV to select the RLS regularization parameter.

Finally, the comparisons between the exact and approximate reduced order systems were done using $x_0 = 0$ and a control input different from those used to learn the dynamics and output functions: $u(t) = \frac{1}{4}\big(\sin(2\pi 3t) + \text{sq}(2\pi 5t - \pi/2)\big)$ where $\text{sq}(\cdot)$ denotes the square wave function.

The Taylor series approximation for Π was done once, about x_0, and was not updated further.

5.4. Results. For the 2D example, there is a clear gap in the singular values of $K_{oc}^T K_{oc}$, $\sigma_1(K_{oc}^T K_{oc}) = 89.3419$, $\sigma_2(K_{oc}^T K_{oc}) = 0.2574$ while the other ones are negligible. For the 7D example, there is also a clear gap in the singular values of $K_{oc}^T K_{oc}$, $\sigma_1(K_{oc}^T K_{oc}) = 197.7821$, $\sigma_2(K_{oc}^T K_{oc}) = 46.9314$, $\sigma_3(K_{oc}^T K_{oc}) = 2.7132$, $\sigma_4(K_{oc}^T K_{oc}) = 0.3293$, $\sigma_5(K_{oc}^T K_{oc}) = 0.0718$, $\sigma_6(K_{oc}^T K_{oc}) = 0.0028$, $\sigma_7(K_{oc}^T K_{oc}) = 0.0010$, $\sigma_6(K_{oc}^T K_{oc}) = 0.0001$ while the other ones are negligible. We plot the first 100 singular values of the Hankel kernel matrix for both examples in logarithmic scale in figure 5.3.

The simulated outputs $\hat{y}(t)$ of the closed reduced systems as well as the output $y(t)$ of the original system together with a comparison with Lall et al. are plotted in Figures 5.1 and 5.2, respectively. One can see that, even for a significantly different
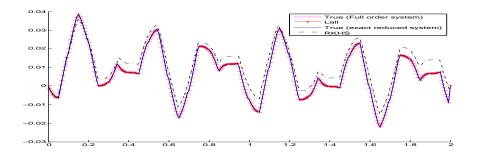
FIG. 5.1. Simulated output trajectories for the original and reduced 2-dimensional system.

input, the reduced systems closely capture the original systems. 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 $\Pi, \hat{f}$ and $\hat{h}$. Indeed, we have had some success driving example systems with random uniform input in some cases. Finally, we note that our method is as good as Lall et al.'s method especially that we assumed that the dynamics (3.1) is unknown and the simulation results are obtained using (4.8) while in Lall et al. the dynamics is assumed to be known.

6. Conclusion. We have introduced a new, empirical model reduction method for nonlinear control systems. The method assumes that the method of linear balancing applies to nonlinear systems in a high dimensional feature 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 pair of nonlinear systems 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 computing error estimates, reduction of unstable systems, structure preserving systems, stochastic differential equations (SDEs), and finding easily verifiable conditions of model reducibility of nonlinear systems. For instance, we conjecture that a nonlinear system is model reducible if there exists an RKHS where there is a spectral gap in the gramians. Finally, our results allows us to argue that working in RKHSes allows to develop methods for
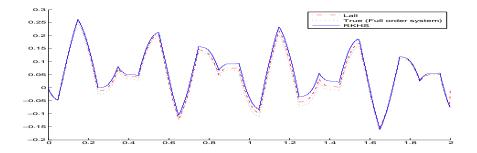
Fig. 5.2. *Simulated output trajectories for the original and reduced 7-dimensional system.*

a Data-Based Theory for (Nonlinear) Dynamical Systems.

Appendix A. Elements of Learning Theory.

In this section, we give a brief overview of reproducing kernel Hilbert spaces as used in statistical learning theory. The discussion here borrows heavily from [10, 53, 46]. Early work developing the theory of RKHS was undertaken by I.J. Schoenberg [38, 39, 40] and then N. Aronszajn [3]. Historically, RKHSes came from the question: when is it possible to embed a metric space into a Hilbert space ?[11]

DEFINITION A.1. *Let $\mathcal{H}$ be a Hilbert space of functions on a set $\mathcal{X}$. Denote by $\langle f, g \rangle$ the inner product on $\mathcal{H}$ and let $||f|| = \langle f, f \rangle^{1/2}$ be the norm in $\mathcal{H}$, for f and $g \in \mathcal{H}$. We say that $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

    i. K has the reproducing property, i.e. $\forall f \in \mathcal{H}$, $f(x) = \langle f(\cdot), K(\cdot, x) \rangle$.
    ii. K spans $\mathcal{H}$, i.e. $\mathcal{H} = span\{K(x, \cdot) | x \in \mathcal{X}\}$.

K will be called a reproducing kernel of $\mathcal{H}$. $\mathcal{H}_K(X)$ will denote the RKHS $\mathcal{H}$ with reproducing kernel K.

DEFINITION A.2. *Given a kernel K and inputs $x_1, \cdots, x_n \in \mathcal{X}$, the $n \times n$ matrix*

$$k := (K(x_i, x_j))_{ij}, \tag{A.1}$$

---

[11]A quasi-metric space (X, d) is embeddable in a Hilbert space $\mathcal{H}$ if there exists a mapping $\Phi : X \to \mathcal{H}$ such that $d(x, y) = ||\Phi(x) - \Phi(y)||$ for all x and y in X. Schoenberg [38, 39, 40] proved that a finite metric space (X, d) whose points are $x_0 \cdots, x_n$ can be embedded into a Hilbert space if and only if the matrix A, whose entries are $A_{ij} = d(x_i, x_0)^2 + d(x_j, x_0)^2 - d(x_i, x_j)^2$, is nonnegative definite, i.e. A is a Gram matrix [8].
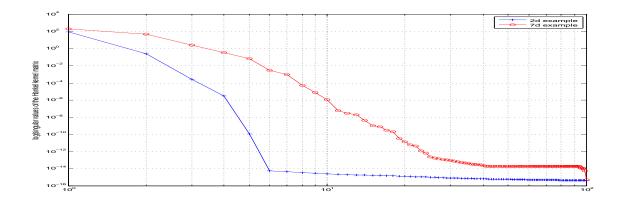
FIG. 5.3. *First hundred singular values of the Hankel kernel matrix, in logarithmic scale, for the 2D and 7D examples.*

is called the Gram Matrix *of k with respect to $x_1, \cdots, x_n$. The kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which for all $n \in \mathbb{N}$ and distinct $x_i \in \mathcal{X}$ gives rise to a strictly positive definite Gram matrix is called a strictly positive definite kernel.*

DEFINITION A.3. *(Mercer kernel map) A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a Mercer kernel if it is continuous, symmetric and positive definite.*

The important properties of reproducing kernels are summarized in the following proposition

PROPOSITION A.4. *If K is a reproducing kernel of a Hilbert space $\mathcal{H}$, then*

 i. $K(x, y)$ is unique.
 ii. $\forall x, y \in \mathcal{X}$, $K(x, y) = K(y, x)$ (symmetry).
 iii. $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ (positive definitness).
 iv. $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$.
 v. Let $c \neq 0$. The following kernels, defined on a compact domain $\mathcal{X} \subset \mathbb{R}^n$, are Mercer kernels: $K(x, y) = x \cdot y'$ (Linear), $K(x, y) = (1 + x \cdot y')^d$, $d \in \mathbb{N}$ (Polynomial), $K(x, y) = e^{-\frac{||x - y||^2}{\sigma^2}}$, $\sigma > 0$ (Gaussian) .

THEOREM A.5. *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric and positive definite function. Then, there exists a Hilbert space of functions $\mathcal{H}$ defined on $\mathcal{X}$ admitting K as a reproducing Kernel. Moreover, there exists a function $\Phi : X \to \mathcal{H}$ such that*[12]

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \quad for \quad x, y \in \mathcal{X}.$$

Φ *is called a feature map*[13].

---

[12]This decomposition shows that kernels can be viewed as generalized dot products.

[13]The dimension of the RKHS can be infinite and corresponds to the dimension of the eigenspace of the integral operator $L_K : \mathcal{L}_\nu^2(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ defined as $(L_K f)(x) = \int K(x, t) f(t) d\nu(t)$ if K is a Mercer kernel, for $f \in \mathcal{L}_\nu^2(\mathcal{X})$ and ν is a Borel measure on $\mathcal{X}$.

Conversely, let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X}$ compact, satisfying $\forall x \in \mathcal{X}, \exists \kappa_x > 0,$ such that $|f(x)| \leq \kappa_x \|f\|_{\mathcal{H}}$. Then, $\mathcal{H}$ has a reproducing kernel K.

THEOREM A.6. *Every sequence of functions $(f_n)_{n \geq 1}$ which converges strongly to a function f in $\mathcal{H}_K(X)$, converges also in the pointwise sense, that is, $\lim_{n \to \infty} f_n(x) = f(x)$, for any point $x \in X$. Further, this convergence is uniform on every subset of X on which $x \mapsto K(x, x)$ is bounded.*

Remarks.

   i. In theorem A.5, and using property [iv.] in Proposition A.4, we can take $\Phi(x) := K_x := K(x, \cdot)$ in which case $\mathcal{F} = \mathcal{H}$ – the "feature space" is the RKHS. This is called the *canonical feature map*.

   ii. The fact that Mercer kernels are positive definite and symmetric reminds us of similar properties of Gramians and covariance matrices. This is an essential fact that we are going to use in the following.

   iii. In practice, we choose a Mercer kernel, such as the ones in [v.] in Proposition A.4, and theorem A.5 guarantees the existence of a Hilbert space admitting such a function as a reproducing kernel.

   iv. Working in RKHSes allows to find nonlinear version of algorithms expressed in terms of inner products [46]. In fact, if an algorithm contains the quantity $\langle x, x' \rangle$ then a nonlinear version of it, i.e. when x is replaced by $\phi(x)$ with $\phi : \mathbb{R}^n \to \mathcal{H}$, would contain the quantity $\langle \phi(x), \phi(x') \rangle$ where ϕ represents the nonlinearity. If we are working in an RKHS then $\langle \phi(x), \phi(x') \rangle := K(x, x')$ and therefore we can replace all the quantities involving $\langle x, x' \rangle$ in the original algorithm by $K(x, x')$ in its nonlinear version.

$\triangleleft$

Example. The following example is taken from [4]. Let V be the collection of functions f with $f'' \in L^2[0, 1]$ and consider the subspace

$$W_2^0 = \{f(x) \in V : f, f' \text{ absolutely continuous and } f(0) = f'(0) = 0\}.$$

Define an inner product on W_2^0 as

$$\langle f, g \rangle = \int_0^1 f''(t) g''(t) dt. \tag{A.2}$$

Using integration by parts and the Fundamental Theorem of Calculus, it can be shown that for $f \in W_2^0$ and any $s \in [0, 1]$, $f(s)$ can be written as

$$f(s) = \int_0^1 (s - u)_+ f''(u) du, \tag{A.3}$$

Since the reproducing kernel of the space W_2^0 must satisfy $f(s) = \langle f(\cdot), R(\cdot, s) \rangle$. From (A.2) and (A.3), we deduce that $K(\cdot, s)$ is a function such that $\frac{d^2 K(u,s)}{d^2 u} = (s - u)_+$. Moreover, since $K(\cdot, s) \in W_2^0$ and using the property $K(s, t) = \langle K(\cdot, t), K(\cdot, s) \rangle$, we deduce that

$$K(s, t) = \langle K(\cdot, t), K(\cdot, s) \rangle = \int_0^1 (t - u)_+ (s - u)_+ \, du = \frac{\max(s, t) \min^2(s, t)}{2} - \frac{\min^3(s, t)}{6} \tag{A.4}$$

$\triangle$

RKHS play an important role in learning theory whose objective is to find an unknown function

$$f^* : X \to Y \tag{A.5}$$

from random samples

$$\mathbf{s} = (x_i, y_i)|_{i=1}^m, \tag{A.6}$$

In the following we review results from [43] (for a more general setting, cf. [10]) about the special case when the data samples $\mathbf{s}$ are such that

Assumption 1: The samples in (A.6) have the special form

$$\mathcal{S} : \quad \mathbf{s} = (x, y_x)|_{x \in \bar{x}}, \tag{A.7}$$

where $\bar{x} = \{x_i\}|_{i=1}^{d+1}$ and y_x is drawn at random from $f^*(x) + \eta_x$, where η_x is drawn from a probability measure ρ_x.

Here for each $x \in X$, ρ_x is a probability measure with zero mean, and its variance σ_x^2 satisfies $\sigma^2 := \sum_{x \in \bar{x}} \sigma_x^2 < \infty$. Let X be a closed subset of $\mathbb{R}^n$ and $\bar{t} \subset X$ is a discrete subset. Now, consider a kernel $K : X \times X \to \mathbb{R}$ and define a matrix (possibly infinite) $K_{\bar{t}, \bar{t}} : \ell^2(\bar{t}) \to \ell^2(\bar{t})$ as

$$(K_{\bar{t}, \bar{t}} a)_s = \sum_{t \in \bar{t}} K(s, t) a_t, \quad s \in \bar{t}, a \in \ell^2(\bar{t}), \tag{A.8}$$

where $\ell^2(\bar{t})$ is the set of sequences $a = (a_t)_{t \in \bar{t}} : \bar{t} \to \mathbb{R}$ with $\langle a, b \rangle = \sum_{t \in \bar{t}} a_t b_t$ defining an inner product. For example, we can take $X = \mathbb{R}$ and $\bar{t} = \{0, 1, \cdots, d\}$.

In the case of dynamical systems such as the ones we are studying in this paper, we are interested in learning the map $\mathcal{F} : x(k) \mapsto x(k+1)$ that characterises a dynamical system $x(k+1) = \mathcal{F}(x(k))$ or the map $\mathcal{F} : (x(k), u(k)) \mapsto x(k+1)$ that characterises a controlled dynamical system $x(k+1) = \mathcal{F}(x(k), u(k))$. In order to get error estimates, we could easily apply the following results to the case of the dynamical systems we are interested in.

The case of approximating a function $f^* \in \mathcal{H}_K$ from samples of the form (A.6) has been studied in [43, 44]. The problem we are interested in is the one of reconstructing f^* from $\mathbf{s}$ which can be expressed as the minimisation problem

$$\bar{f}_{\mathbf{s}, \gamma} := \operatorname{argmin}_{f \in \mathcal{H}_{K, \bar{t}}} \left\{ \sum_{x \in \bar{x}} (f(x) - y_x)^2 + \gamma ||f||_K^2 \right\}, \tag{A.9}$$

where $\gamma \geq 0$. Moreover, in order to consider the case where $\bar{x}$ is not defined by a uniform grid on X, the authors of [43] introduced a weighting $w := \{w_x\}_{x \in \bar{x}}$ on $\bar{x}$ with $w_x > 0$[14]. Let D_w be the diagonal matrix with ,aim diagonal entries $\{w_x\}_{x \in \bar{x}}$. Then, $||D_w|| \leq ||w||_\infty$.

In this case, the regularisation scheme (A.9) becomes

$$\bar{f}_{\mathbf{s}, \gamma} := \operatorname{argmin}_{f \in \mathcal{H}_{K, \bar{t}}} \left\{ \sum_{x \in \bar{x}} w_x (f(x) - y_x)^2 + \gamma ||f||_K^2 \right\}, \tag{A.10}$$

---

[14]A suggestion proposed in [43] is to consider the ρ_X-volume of the Voronoi associated with $\bar{x}$. Another example is $w = 1$ or if $|\bar{x}| = m < \infty$, $w = \frac{1}{m}$.

In learning theory, the minimization is taken over functions from a hypothesis space often taken to be a ball of a RKHS $\mathcal{H}_K$ associated to Mercer kernel K, and the function $f_{\mathbf{s}}$ that minimizes the empirical error $\mathcal{E}_{\mathbf{s}}$ is given by *the representer theorem* (cf. [53, 12, 10] for derivation and more general forms of this theorem)

THEOREM A.7. *Assume $f^* \in \mathcal{H}_{K,\bar{t}}$ and the standing hypotheses with X, K, $\bar{t}$, ρ as above, y as in (A.7). Suppose $K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}}$ is invertible. Define $\mathcal{L}$ to be the linear operator $\mathcal{L} = (K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1} K_{\bar{t},\bar{x}} D_w$. Then the problem (A.10) has a unique solution*

$$f_{\mathbf{s},\gamma} = \sum_{t \in \bar{t}} (\mathcal{L}y)_t K_t \tag{A.11}$$

The above theorem can be reformulated as follows

THEOREM A.8. *Let $\mathbf{s} \in \mathbb{Z}^m$ and $\lambda \in \mathbb{R}$, $\lambda > 0$. The empirical target, i.e. the function $f_{\lambda,\mathbf{s}} = f_{\mathbf{s}}$ minimizing the regularized empirical error (A.9) over $f \in \mathcal{H}_K$, may be expressed as*

$$f_{\mathbf{s}}(x) = \sum_{j=1}^{m} c_j K(x, x_j), \tag{A.12}$$

where $c = (c_1, \cdots, c_m)$ is the unique solution of the well-posed linear system in $\mathbb{R}^m$

$$\lambda m \, c_i + \sum_{j=1}^{m} K(x_i, x_j) c_j = y_i, \quad i = 1, \cdots m, \tag{A.13}$$

Assumption 2: For each $x \in X$, ρ_x is a probability measure with zero mean supported on $[-M_x, M_x]$ with $\mathcal{B}_w := (\sum_{x \in \bar{x}} w_x M_x^2)^{\frac{1}{2}} < \infty$.

DEFINITION A.9. *We say that $\bar{x}$ is $\Delta-$dense in X if for each $y \in X$ there is some $x \in \bar{x}$ satisfying $||x - y||_{\ell^\infty(\mathbb{R}^n)} \le \Delta$.*

THEOREM A.10. *(Sample Error) [Theorem 4, Propositions 2 and 3 in[43]]*

Suppose $K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}}$ is invertible. Under the assumption (A.7), let $f_{\mathbf{s},\gamma} = \sum_{t \in \bar{t}} c_t K_t$ be the solution of (A.10) given in Theorem A.7 by $c = \mathcal{L}y$. Let $\mathcal{L}_w$ and κ be

$$\mathcal{L}_w = (K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1} K_{\bar{t},\bar{x}} D_w^{1/2} \tag{A.14}$$

$$\kappa := ||K_{\bar{t},\bar{t}}|| \, ||(K_{\bar{t},\bar{x}} D_w K_{\bar{x},\bar{t}} + \gamma K_{\bar{t},\bar{t}})^{-1}||^2 \tag{A.15}$$

Then for every $0 < \delta < 1$, with confidence $1 - \delta$ we have the sample error estimate

$$||f_{\mathbf{s},\gamma} - f_{\bar{x},\gamma}||_K^2 \le \mathcal{E}_{samp} := \kappa \sigma_w^2 \alpha^{-1} \left(\frac{2||K_{\bar{t},\bar{t}} \mathcal{L}_w|| \, ||\mathcal{L}_w|| \, \mathcal{B}_w^2}{\kappa \sigma_w^2} \, \log \frac{1}{\delta} \right), \tag{A.16}$$

where α is the increasing function defined for $u > 1$ as $\alpha(u) = (u - 1) \log u$. In particular, $\mathcal{E}_{samp} \to 0$ when $\gamma \to \infty$ or $\sigma_w^2 \to 0$.

THEOREM A.11. *(Regularization Error and Integration Error, Proposition 4 and Theorem 5 in[43])*

Under Assumptions 1 and 2. Let $\bar{X} = (X_x)_{x \in \bar{x}}$ be the Voronoi of X associated with $\bar{x}$ and $w_x = \rho_X(X_x)$. Define the Lipschitz norm on a subset $X' \subset X$ as

$||f||_{Lip(X')} := ||f||_{L^\infty(X')} + \sup_{s,u \in X} \frac{|f(s)-f(u)|}{||s-u||_{\ell^\infty(\mathbb{R}^n)}}$ *and assume that the inclusion map of $\mathcal{H}_{K,\bar{t}}$ into the Lipschitz space satisfies*[15]

$$C_{\bar{x}} := \sup_{f \in \mathcal{H}_{K,\bar{t}}} \frac{\sum_{x \in \bar{x}} w_x ||f||^2_{Lip(X_x)}}{||f||^2_K} < \infty. \tag{A.17}$$

i. If $f^* \in \mathcal{H}_{K,\bar{t}}$ and $\lambda_{\bar{x},w} > 0$, then

$$||f_{\bar{x},\gamma} - f^*||^2_K \le \frac{\gamma ||K_{\bar{t},\bar{t}}|| \, ||f^*||^2_K}{\lambda^2_{\bar{x},w}} \tag{A.18}$$

ii. If $\bar{x}$ is $\Delta-$dense, $C_{\bar{x}} < \infty$, and $f^* \in \mathcal{H}_{K,\bar{t}}$, then

$$||f_{\bar{x},\gamma} - f^*||^2 \le ||f^*||^2_K (\gamma + 8C_{\bar{x}}\Delta) \tag{A.19}$$

THEOREM A.12. *(Sample, Regularization and Integration Errors) (Corollary 5 in [43])*

Under Assumptions 1 and 2. Let $\bar{X} = (X_x)_{x \in \bar{x}}$ be the Voronoi of X associated with $\bar{x}$ and $w_x = \rho_X(X_x)$. If $\bar{x}$ is $\Delta-$dense, $C_{\bar{x}} < \infty$, and $f^* \in \mathcal{H}_{K,\bar{t}}$, then, for every $0 < \delta < 1$, with probability at least $1 - \delta$ there holds

$$||f_{\mathbf{s},\gamma} - f^*||^2 \le 2C_{\bar{x}}\mathcal{E}_{samp} + 2||f^*||^2_K(\gamma + 8C_{\bar{x}}\Delta), \tag{A.20}$$

where $\mathcal{E}_{samp}$ is given in (A.16).

If f^* is not an element of $\mathcal{H}_{K,\bar{t}}$ then one also needs an estimate for the approximation error [42],[10].

THEOREM A.13. *Define $f_{\mathbf{s},\gamma}$ by (A.11). If $L_K^{-r} f_\rho \in L^2_{\rho_X}$, then*

$$||f_{\mathbf{s},\gamma} - f_\rho||_{L^2_{\rho_X}} \le \lambda^r ||L_K^{-r} f_\rho||_{L^2_{\rho_X}}, \quad if \quad 0 < r \le 1. \tag{A.21}$$

When $\frac{1}{2} < r \le 1$, we have

$$||f_{\mathbf{s},\gamma} - f_\rho||_K \le \lambda^{r - \frac{1}{2}} ||L_K^{-r} f_\rho||_{L^2_{\rho_X}} \tag{A.22}$$

Here $f_{\mathbf{s},\gamma}$ is taken as an approximation of the regression function f_ρ. Hence, minimizing over the (possibly infinite dimensional) Hilbert space, reduces to minimizing over $\mathbb{R}^m$. The series (A.11) converges absolutely and uniformly to f. We call *learning* the process of approximating the unknown function f from random samples on Z.

In the following, we assume that the kernels K are continuous and bounded by $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)} < \infty$.

Appendix B. Kernel PCA. Kernel PCA [41] will be a helpful starting point for understanding the approach to balanced reduction introduced in this paper. We briefly review the relevant background here.

Kernel PCA (KPCA) is a generalization of linear PCA that allows to take into account nonlinear versions of observations. This is done by carrying out PCA in a high dimensional RKHS through an injective, not necessarily surjective, map

$$\Phi : \mathbb{R}^n \to \mathcal{H}, \; x \mapsto \Phi(x). \tag{B.1}$$

---

[15]This assumption is true if X is compact and the inclusion map of $\mathcal{H}_{K,\bar{t}}$ into the space of Lipschitz functions on X is bounded which is the case when K is a C^2 Mercer kernel [55]. In fact, if $||f||_{\text{Lip}(X)} \le C_0 ||f||_K$ for each $f \in \mathcal{H}_{K,\bar{t}}$, then $C_{\bar{x}} \le C_0^2 \rho_X(X)$.

Φ is possibly nonlinear and $\mathcal{H}$ is a high-dimensional, possibly infinite-dimensional, RKHS. Given the set of data $\mathbf{x} := \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, the covariance matrix $C \in \mathbb{R}^n \times \mathbb{R}^n$ defined as

$$C = \frac{1}{N} \sum_{i=1}^N x_i x_i^T \tag{B.2}$$

becomes a covariance matrix in $\mathcal{H}$

$$\mathbf{C}_{\mathcal{H}} = \frac{1}{N} \sum_{i=1}^N \Phi(x_i) \Phi(x_i)^T \tag{B.3}$$

If $\mathcal{H}$ is infinite-dimensional, we think of $\Phi(x_i)\Phi(x_i)^T$ as a linear operator on $\mathcal{H}$, mapping $x \mapsto \Phi(x_j)\langle \Phi(x_j), x \rangle$. 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 [41].

Taking the feature map[16] $\Phi : \mathbb{R}^n \to \mathbb{R}^N$ with $\Phi_i(x) = K(x, x_i)$ and given the set of data $\mathbf{x} := \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, we can consider PCA in $\mathbb{R}^N$ by simply working with the covariance of the mapped vectors (B.3).

The principal subspaces are computed by diagonalizing $C_{\mathcal{H}}$ through solving the eigenvalue problem

$$\lambda v = \mathbf{C}_{\mathcal{H}} v, \tag{B.4}$$

for eigenvalues $\lambda \geq 0$ and nonzero eigenvectors $v \in \mathbb{R}^N$. This problem can be expressed in term of an inner-product by plugging (B.3) into (B.4) and getting

$$\lambda v = \mathbf{C}_{\mathcal{H}} v = \frac{1}{N} \sum_{i=1}^N \langle \Phi(x_i), v \rangle_{\mathcal{H}} \Phi(x_i), \tag{B.5}$$

However as is shown in [41], we can perform the computations directly in terms of the kernel without explicitly knowing Φ. One can equivalently form the matrix $\mathbb{K}$ of kernel products whose entries are $(\mathbb{K})_{ij} = K(x_i, x_j)$ for $i, j = 1, \ldots, N$, and solve the eigenproblem

$$\mathbb{K}\boldsymbol{\alpha} = N\lambda\boldsymbol{\alpha}, \tag{B.6}$$

in $\mathbb{R}^N$. Moreover the eigenvectors v in (B.4) and the eigenvectors $\boldsymbol{\alpha}$ in (B.6) are related through

$$v_i = \Psi\boldsymbol{\alpha}_i, \tag{B.7}$$

where $\Psi := \begin{pmatrix} \Phi(x_1) & \cdots & \Phi(x_M) \end{pmatrix}$, and the non-zero eigenvalues of $\mathbb{K}$ and $\mathbf{C}_{\mathcal{H}}$ coincide.

The eigenvectors $\boldsymbol{\alpha}_i$ of $\mathbb{K}$ are subsequently normalized so that the eigenvectors v_i of $\mathbf{C}_{\mathcal{H}}$ have unit norm in the RKHS, 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

$$A_q = \begin{bmatrix} \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_q \end{bmatrix}, 1 \leq q \leq \min(n, M). \tag{B.8}$$

---

[16]This feature map is valid given property iv. in Proposition A.3.

Similarly, form the matrix $V_q = \begin{bmatrix} v_1 & \cdots & v_q \end{bmatrix}, 1 \leq q \leq \min(n, M)$ of sorted eigenvectors of $C_{\mathcal{H}}$. 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 [41]) that principal components in the feature space can be computed in the original space with kernels using the map $\Pi : \mathbb{R}^M \to \mathbb{R}^q$

$$\Pi(x) := A_q^\top \mathbf{k}(x), \tag{B.9}$$

where $\mathbf{k}(x) = \big(K(x, x_1), \ldots, K(x, x_M)\big)^\top$.

Kernel methods can be used to develop nonlinear generalizations of any algorithm that can be expressed in terms of inner products [41] and KPCA is an illustration of this approach. KPCA is viewed as a nonlinear version of PCA since PCA in $\mathbb{R}^n$ can be reformulated as an eigenvalue problem in terms of inner products as in (B.5) but with $C_{\mathcal{H}}$ replaced by C and $\Phi(x) = x$, i.e.

$$\lambda v = C v = \frac{1}{M} \sum_{i=1}^{M} \langle x_i, v \rangle x_i, \tag{B.10}$$

given the expression of C in (B.2). If one wants to perform PCA on a nonlinear version of the data $(x_i)|_{i=1}^{M}$ through a nonlinear map Φ, it is enough to replace x by $\Phi(x)$ in the eigenvalue problem (B.10) to get (B.5).

Our goal in this paper is to extend this method to linear balancing in view of applying it to nonlinear control systems.

REFERENCES

[1] Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*, SIAM Publications.
[2] C Archambeau, D Cornford, M Opper, J Shawe-Taylor, Gaussian process approximations of stochastic differential equations, Journal of machine learning research, pp. 1-16, 64, 2007.
[3] Aronszajn, N. (1950). Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.*, 68:337-404.
[4] Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Norwell, MA.
[5] Bouvrie, J. and B. Hamzi (2010). Balanced Reduction of Nonlinear Control Systems in Reproducing Kernel Hilbert Space, in *Proc. 48th Annual Allerton Conference on Communication, Control, and Computing*, pp. 294-301. `http://arxiv.org/abs/1011.2952`.
[6] Bouvrie, J. and B. Hamzi (2012), Empirical Estimators for the Controllability Energy and Invariant Measure of Stochastically Forced Nonlinear Systems,in *Proc. of the 2012 American Control Conference* (long version at `http://arxiv.org/abs/1204.0563`).
[7] Bouvrie, J. and B. Hamzi (2014), Embedology for Control and Random Dynamical Systems in Reproducing Kernel Hilbert Spaces, in preparation.
[8] Cheney, W. and W. Light (2009). A Course in Approximation Theory, Graduate Studies in Mathematics, vol. 101, AMS.
[9] 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.
[10] Cucker, F. and S. Smale (2001). On the mathematical foundations of learning. *Bulletin of AMS*, 39:1-49.
[11] Dullerud, G. E., and F. Paganini (2000). *A Course in Robust Control Theory: a Convex Approach*, Springer.
[12] Evgeniou, T., M. Pontil, and T. Poggio (2000). Regularization networks and support vector machines, Advances in Computational Mathematics, vol. 13, no. 1, pp. 1-50.
[13] 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.
[14] Fujimoto, K. and J. Scherpen (2010). Balanced realization and model order reduction for nonlinear systems based on singular value analysis, SIAM Journal on Control and Optimization, Vol. 48, 7, pp. 180-194.

[15] Gray, W. S. and E. I. Verriest (2006). Algebraically Defined Gramians for Nonlinear Systems, *Proc. of the 45th IEEE CDC.*

[16] Hammarling, S. J. (1982). Numerical Solution of the Stable, Non-negative Definite Lyapunov Equation, IMA Journal of Numerical Analysis, vol. 2, pp. 303-323.

[17] Jolliffe, I.T. (2002). Principal Component Analysis, Springer.

[18] Jonckheere, E.A., and L. M. Silverman (1983). *A New Set of Invariants for Linear Systems - Application to Reduced Order Compensator Design.* IEEE Transactions on Automatic Control, **AC-28**, 10, 953-964.

[19] Kantz, H. and T. Schreiber. Nonlinear Time Series Analysis, Cambridge University Press, 2004.

[20] Kenney, C. and G. Hewer (1987). *Necessary and Sufficient Conditions for Balancing Unstable Systems*, IEEE Transactions on Automatic Control, **32**, 2, pp. 157-160.

[21] Krener, A. (2006). Model reduction for linear and nonlinear control systems. Bode Lecture, 45th IEEE Conference on Decision and Control.

[22] 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, pp. 161-170. Springer.

[23] 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, pp. 41-62. Springer.

[24] 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).*

[25] 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.

[26] Laub, A.J. (1980). On Computing "balancing" transformations, *Proc. of the 1980 Joint Automatic Control Conference (ACC).*

[27] Li, J.-R. (2000). *Model Reduction of Large Linear Systems via Low Rank System Grammians.* Ph.D. thesis, Massachusetts Institute of Technology.

[28] 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.

[29] Moore, B. (1981). Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction, *IEEE Tran. Automat. Control*, **26**, 1, pp. 17-32.

[30] Newman, A.J., and P. S. Krishnaprasad (2000). Computing balanced realizations for nonlinear systems, *Proc. of the Math. Theory of Networks and Systems (MTNS).*

[31] Nilsson, O. (2009). *On Modeling and Nonlinear Model Reduction in Automotive Systems*, Ph.D. thesis, Lund University.

[32] 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.*

[33] 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.

[34] Schaback, R. and H. Wendland. Kernel techniques: From machine learning to meshless methods, Acta Numerica, vol.15, pp. 543-639, 2006.

[35] Scherpen, J. (1993). Balancing for nonlinear systems. Syst. & Contr. Let., 21, pp 143-153.

[36] Scherpen, J. (1994). *Balancing for Nonlinear Systems*, Ph.D. thesis, University of Twente, `http://www.dcsc.tudelft.nl/~jscherpen/thesis.html`.

[37] Scherpen, J. (2011). Balanced Realizations, Model Order Reduction, and the Hankel Operator, the Control Systems Handbook, 2nd ed., Advanced Methods, Eds. William Levine, CRC Press, Taylor & Francis Group, Chap. 4, pp. 1-24.

[38] Schoenberg, I. J. (1935). Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert", Annals of Mathematics 36 (1935), 724–732.

[39] Schoenberg, I. J. (1937) On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in Hilbert space, Annals of Mathematics, (2), vol. 38 (1937), pp. 787-793.

[40] Schoenberg, I. J. (1938). Metric spaces and positive definite functions, Trans. Amer. Math. Soc. 44 (1938), pp. 522-536.

[41] Schölkopf, B., Smola, A., and Müller, K (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

[42] Smale, S. and D.,-X. Zhou (2003). Estimating the Approximation Error in Learning Theory, Analysis and Applications, Volume 01, Issue 01.

[43] Smale, S. and D.,-X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41, 2004, pp. 279-305.

[44] Smale, S. and D.,-X. Zhou, Shannon sampling II: Connections to learning theory, Applied and Computational Harmonic Analysis, Volume 19, Issue 3, November 2005, pp. 285-302.

[45] S. Smale and D.-X. Zhou, *Online Learning with Markov Sampling*, Anal. Appl., 7, pp. 87-113, 2009.

[46] Schölkopf, B., and A. J. Smola (2002). Learning with Kernels, The MIT Press.

[47] Steinwart, I. and A. Christmann (2008). Support Vector Machines, Springer.

[48] Therapos, C. P. (1989). *Balancing Transformations for Unstable Nonminimal Linear Systems*, IEEE Transactions on Automatic Control, **34**, 4, pp. 455-457.

[49] Verriest, E. (1981). Suboptimal LQG-Design via Balanced Realizations. Proc. of the 20th IEEE CDC, pp. 686-687.

[50] Verriest, E. (1984). Approximation and Order Reduction in Nonlinear models using an RKHS Approach, Proc. of the 18th Annual Conference on Information Sciences and Systems, pp. 197-201.

[51] Box, G. E. P. and G. M. Jenkins, Time Series Analysis: Forecasting and Control, Wiley, 2008.

[52] L. Ljung, System Identification: Theory for the User, Prentice Hall, 1999.

[53] Wahba, G. (1990). Spline Models for Observational Data, *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics* 59.

[54] Weiland, S. (1991). Theory of approximation and disturbance attenuation of linear systems, Doctoral dissertation, University of Groningen.

[55] Zhou, D.-X. (2003). Capacity of reproducing kernel spaces in learning theory, IEEE Transactions on Information Theory, Volume:49 , Issue: 7, pp. 1743 - 1752.