Machine Learning and Dynamical Systems meet in Reproducing Kernel Hilbert Spaces

Boumediene Hamzi,
Department of Mathematics,
Imperial College London, London, UK.

joint work with Jake Bouvrie (MIT, USA), Peter Giesl (University of Sussex, UK), Christian Kuehn (TUM, Germany), Romit Maulik (Argonne National Lab., USA), Sameh Mohamed (SUTD, Singapore), Houman Owhadi (Caltech, USA), Martin Rasmussen & Kevin Webster (Imperial College London), Bernard Haasdonk and Dominik Wittwar (University of Stuttgart, Germany), Gabriele Santin (Bruno Kessler Foundation, Italy)

Research Supported by the European Commission through the Marie Curie Fellowships Scheme

Epistemological context

How to analyze complex systems: Among the current approaches to analyze complex systems

- Theory of Dynamical Systems allows to analyze complex systems when the model is known. It offers nontrivial ways to analyze dynamical systems. It has the status of Theory. Currently, it is limited to low-dimensional models.
- Machine Learning is concerned with algorithms designed to accomplish a certain task, whose performance improves with the input of more data. It allows the analysis of some very high-dimensional complex systems on the basis of data when the model is not even known.
  Current limitations: Mostly a set of techniques and algorithms. No Methodologies. Theory still underdeveloped. It is not clear why the algorithms work and what is their domain of applicability.

$\rightsquigarrow$ It makes sense to combine Dynamical Systems and Machine Learning.

Goal

Goal: Fill the gap between Machine Learning and Dynamical Systems in the following directions

- Machine Learning for Dynamical Systems: how to analyze dynamical systems on the basis of observed data rather than attempt to study them analytically (it allows to extend the boundaries of the classical theory of dynamical systems).
- Dynamical Systems for Machine Learning: how to analyze algorithms of Machine Learning using tools from the theory of dynamical systems (allows to give solid foundations to the existing methods and understand their true potential and limits- identify the domain of applicability of the algorithms in ML).

Context

As pointed out by Steve Smale, the interaction between Dynamical Systems and Learning Theory is an important problem[1]:

> *"Some years ago, Felipe (Cucker) and I were trying to find something about the brain science and artificial intelligence starting from literature on neural nets. It was in this setting that we encountered the beautiful ideas and fast algorithms of learning theory. Eventually we were motivated to write on the mathematical foundations of this new area of science.*
>
> *I have found this arena to, with its new challenges and growing number of applications, be exciting. For example, the unification of dynamical systems and learning theory is a major problem. Another problem is to develop a comparative study of useful algorithms currently available and to give unity to these algorithms."*

[1] Felipe Cucker and Ding Xuan Zhou (2007), Learning Theory: An Approximation Theory Viewpoint.

Context

"Personal computing has developed to the point where in many cases it ought to be easier to simulate a dynamical system and analyze the empirical data, rather than attempt to study the system analytically. Indeed, for large classes of nonlinear systems, numerical analysis may be the only viable option. Yet the mathematical theory necessary to analyze dynamical systems on the basis of observed data is still largely underdeveloped."

J. Bouvrie and BH (2012), Empirical Estimators for Stochastically Forced Nonlinear Systems: Observability, Controllability and the Invariant Measure,
`https://arxiv.org/pdf/1204.0563v1.pdf`

Research Context

Goal: Combining tools from the theories of Dynamical Systems and Learning in view of a Data-Based Qualitative Theory of Dynamical Systems for analysis, prediction of nonlinear systems and control.

Approach: View Reproducing Kernel Hilbert Spaces as "Linearizing Spaces". By linearization we mean the following: Nonlinear Systems will be embedded into an RKHS where Linear Systems Theory will be applied.

Motivation: Working in RKHSs allows to find a nonlinear version of algorithms that can be expressed in terms of inner products.

Outline

- Elements of Learning Theory and Function Approximation in RKHSs
- Probability Measures in RKHSs and the Maximum Mean Discrepancy
- Kernel Flows for Learning Chaotic Dynamical Systems
- Approximation of Center Manifolds in RKHSs
- Construction of Lyapunov Functions in RKHSs
- Detection of Critical Transitions for some Slow-Fast SDEs
- Review of Some Concepts of Linear Control Systems
- Approximation of Nonlinear Control Systems in RKHSs
- Review of Some Concepts of Linear SDEs
- Estimation of the Stationary Solution of the Fokker-Planck Equation of nonlinear SDEs

Summary of the Approach

• We assume that there is a $\phi : \mathbb{R}^n \to \mathcal{H}; x \mapsto z$ where $\mathcal{H}$ is an RKHS such that we can perform an analysis (in general, but not necessarily, a linear analysis) in $\mathcal{H}$ then come back to $\mathbb{R}^n$.

• The transformation ϕ is obtained from the kernel that defines the RKHS (in general, it is not necessary to explicitly find ϕ). In practice, we will use $\phi(x) = [\phi_1(x) \cdots \phi_N(x)]^T$ with

$$\phi_i(x) = K(x, x(t_i))$$

where K is a reproducing kernel and $x(t_i)$ are measurements at time t_i, $i = 1, \cdots, N$ and $N \gg n$.

• Measurements/Data are used to construct the Hilbert Space where computations become "simpler".

Reproducing Kernel Hilbert Spaces

- Historical Context: Appeared in the 1930s as an answer to the question: when is it possible to embed a metric space into a Hilbert space ? (Schoenberg, 1937)
- Answer: If the metric satisfies certain conditions, it is possible to embed a metric space into a special type of Hilbert spaces called RKHSs.
- Properties of RKHSs have been further studied in the 1950s and later (Aronszajn, 1950; Schwartz, 1964 etc.)

Reproducing Kernel Hilbert Spaces

• Definition: A Hilbert Space is an inner product space that is complete and separable with respect to the norm defined by the inner product.

• Definition: For a compact $\mathcal{X} \subseteq \mathbb{R}^d$, and a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, we say that $\mathcal{H}$ is a RKHS if there exists $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

i. k has the reproducing property, i.e. $\forall f \in \mathcal{H},\ f(x) = \langle f(\cdot), k(\cdot, x)\rangle$ (k is called a reproducing kernel).

ii. k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{\text{span}\{k(x, \cdot)|x \in \mathcal{X}\}}$.

• Definition: A Reproducing Kernel Hilbert Space (RKHS) is a Hilbert space H with a reproducing kernel whose span is dense in H. Equivalently, a RKHS is a Hilbert space of functions where all evaluation functionals are bounded and linear.

Reproducing Kernel Hilbert Spaces

The important properties of reproducing kernels are

- The RKHS is unique.
- $\forall x, y \in \mathcal{X}$, $K(x, y) = K(y, x)$ (symmetry).
- $\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ (positive definitness).
- $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$. Using this property, one can immediately get the canonical feature map (Aronszajn's feature map): $\Phi_c(x) = K(x, \cdot)$.

- A Mercer kernel is a continuous positive definite kernel.
- 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.
- Examples of kernels: $k(x, x') = \langle x, x' \rangle^d$, $k(x, x') = \exp\left(- \frac{||x - x'||_2^2}{2\sigma^2} \right)$, $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \theta)$.

Reproducing Kernel Hilbert Spaces

• Mercer Theorem: Let $(\mathcal{X}, \mu)$ be a finite-measure space, and suppose $k \in L_\infty(\mathcal{X}^2)$ is a symmetric real-valued function such that the integral operator

$$\begin{aligned} L_k : L_2(\mathcal{X}) &\rightarrow L_2(\mathcal{X}) \\ f &\mapsto (L_k f)(x) = \int_{\mathcal{X}} k(x, x') f(x') d\mu(x') \end{aligned}$$

is positive definite; that is, for all $f \in L_2(\mathcal{X})$, we have $\int_{\mathcal{X}^2} k(x, x') f(x) f(x') d\mu(x) d\mu(x') \geq 0$.
Let $\Psi_j \in L_2(\mathcal{X})$ be the normalized orthogonal eigenfunctions of L_k associated with the eigenvalues $\lambda_j > 0$, sorted in non-increasing order. Then

i. $(\lambda_j)_j \in \ell_1$,
ii. $k(x, x') = \sum_{j=1}^{N_\mathcal{X}} \lambda_j \Psi_j(x) \Psi_j(x')$ holds for almost all (x, x'). Either $N_\mathcal{X} \in \mathbb{N}$, or $N_\mathcal{X} = \infty$; in the latter case, the series converges absolutely and uniformly for almost all (x, x').

Reproducing Kernel Hilbert Spaces

• Proposition (Mercer Kernel Map): If k is a Mercer kernel, it is possible to construct a mapping Φ into a space where k acts as a dot product,

$$\langle \Phi(x), \Phi(x') \rangle = k(x, x'),$$

for almost all $x, x' \in \mathcal{X}$.

• From Mercer's theorem $\Phi : X \to \ell^2$ is

$$\Phi_i(x) = \sqrt{\lambda_i} \Psi_i(x).$$

• Φ is not unique and depends on the measure μ.

• Φ is difficult to compute in general.

Reproducing Kernel Hilbert Spaces

- It is unnecessary to invoke Mercer's theorem just for discussing feature maps/spaces.
- Example of non-Mercer feature maps using $\Phi(x) = K(x, \cdot)$
  - For a polynomial kernel $K(x,t) = \langle x, t \rangle^2$,

$$\Phi : (x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2} x_1 x_2) \in \mathbb{R}^3.$$

  - For a Gaussian kernel $K(x,t) = e^{-\frac{||x-t||^2}{\sigma^2}}$,

$$\Phi : x \to e^{-\frac{||x||^2}{\sigma^2}} \left(\sqrt{\frac{(2/\sigma^2)^k C_\alpha^k}{k!}} x^\alpha \right) \Big|_{|\alpha|=k, k=0}^{\infty} \in \ell^2.$$

- Mercer theorem is, however, fundamental to find error estimates and study the smoothing properties of kernels.

RKHS in Approximation Theory (aka Learning Theory)

- RKHS play an important role in learning theory whose objective is to find an unknown function $f : X \to Y$ from random samples $(x_i, y_i)|_{i=1}^m$.
- For instance, assume that the random probability measure that governs the random samples is ρ and is defined on $Z := X \times Y$. Let X be a compact subset of $\mathbb{R}^n$ and $Y = \mathbb{R}$. If we define the least square error of f as $\mathcal{E} = \int_{X\times Y}(f(x) - y)^2 d\rho$, then the function that minimizes the error is the regression function f_ρ defined as

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x), \quad x \in X,$$

where $\rho(y|x)$ is the conditional probability measure on $\mathbb{R}$.

RKHS in Approximation Theory (aka Learning Theory)

- Since ρ is unknown, neither f_ρ nor $\mathcal{E}$ is computable. We only have the samples $\mathbf{s} := (x_i, y_i)|_{i=1}^m$. The error $\mathcal{E}$ is approximated by the empirical error $\mathcal{E}_{\mathbf{s}}(f)$ by

$$\mathcal{E}_{\mathbf{s}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \lambda||f||_{\mathcal{H}}^2,$$

for $\lambda \geq 0$, λ plays the role of a regularization parameter.

RKHS in Approximation Theory (aka Learning Theory)

• 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 a kernel K, and the function $f_{\mathbf{s}}$ that minimizes the empirical error $\mathcal{E}_{\mathbf{s}}$ is

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

where the coefficients $(c_j)_{j=1}^m$ are obtained by solving the linear system

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

and $f_{\mathbf{s}}$ is taken as an approximation of the regression function f_ρ.

• We call *learning* the process of approximating the unknown function f from random samples on Z.

RKHS in Approximation Theory (aka Learning Theory)

- Now, suppose we are given a set of points $\mathbf{x} = (x_1, \cdots, x_m)$ sampled i.i.d. according to ρ. Many problems in Learning Theory deal with the empirical kernel matrix $\mathbb{K} \in \mathbb{R}^{m \times m}$ whose entries are

$$\mathbb{K}_{i,j} = \frac{1}{m} K(x_i, x_j).$$

- The *restriction operator* $\mathcal{R}_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ with a discrete subset $(\mathbf{x_i})|_{i=1}^m \in X$ is defined as

$$\mathcal{R}_{\mathbf{x}} f = (f(x_i))_{i=1}^m$$

The adjoint of the restriction operator, $\mathcal{R}_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}_K$ is given by

$$\mathcal{R}_{\mathbf{x}}^* c = \sum_{i=1}^m c_i K(x, x_i), \quad c \in \mathbb{R}^m$$

RKHS in Change Point Detection

- We will consider a sequence of samples $x_1, x_2, \cdots, x_n$ from a domain $\mathcal{X}$.
- We are interested in detecting a possible change-point τ, such that before τ, the samples $x_i \sim P$ i.i.d for $i \leq \tau$, where P is the so-called background distribution, and after the change-point, the samples $x_i \sim Q$ i.i.d for $i \geq \tau + 1$, where Q is a post-change distribution.
- We map the dataset in an RKHS $\mathcal{H}$ then compute a measure of discrepancy Δ_n.
- Δ_n is small if $P = Q$ and large if P and Q are far apart.
- We will use the maximum mean discrepancy (MMD)

$$\text{MMD}[\mathcal{H}, P, Q] := \sup_{f \in \mathcal{H}, ||f|| \leq 1} \{\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]\},$$

as a measure of heteregoneity.

Probability Measures in RKHSes

- Let $\mathcal{H}$ be an RKHS on the separable metric space $\mathcal{X}$, with a continuous feature mapping $\phi : \mathcal{X} \to \mathcal{H}$. Assume that k is bounded, i.e. $\sup_{\mathcal{X}} k(x, x) < \infty$.
- Let $\mathcal{P}$ be the set of Borel probability measures on $\mathcal{X}$. We define the mapping to $\mathcal{H}$ of $P \in \mathcal{P}$ as the expectation of $\phi(x)$ with respect to P, i.e.

$$\begin{array}{rccl} \mu_P : \mathcal{P} & \to & \mathcal{H} & \\ P & \mapsto & \int_{\mathcal{X}} \phi(x) dP(x) =: \mu_k(P) & \text{(kernel mean embedding of P)} \end{array}$$

- The maximum mean discrepancy (MMD) between two probability measures P and Q is defined as the distance between two such mappings

$$MMD(P, Q) = ||\mu_k(P) - \mu_k(Q)||_{\mathcal{H}_k}$$

Probability Measures in RKHSes

- The maximum mean discrepancy (MMD) is defined as (Gretton et al., 2007)

$$\begin{aligned} \text{MMD}(P,Q) &:= ||\mu_P - \mu_Q||_{\mathcal{H}}, \\ &= \left(\mathbb{E}_{x,x'}(k(x,x')) + \mathbb{E}_{y,y'}(k(y,y')) - 2\mathbb{E}_{x,y}(k(x,y)\right)^{\frac{1}{2}} \end{aligned}$$

where x and x' are independent random variables drawn according to P, y and y' are independent random variables drawn according to Q, and x is independent of y.

- This quantity is a pseudo-metric on distributions, i.e. it satisfies all the qualities of a metric except $\text{MMD}(P,Q) = 0$ iff $P = Q$.
- For the MMD to be a metric, it is sufficient that the kernel is characteristic, i.e. the map $\mu_P : \mathcal{P} \to \mathcal{H}$ is injective. This is satisfied by the Gaussian kernel (both on compact domains and on $\mathbb{R}^d$) for example.

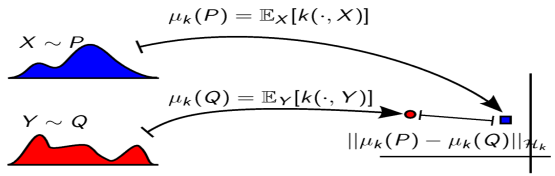
Probability Measures in RKHSes

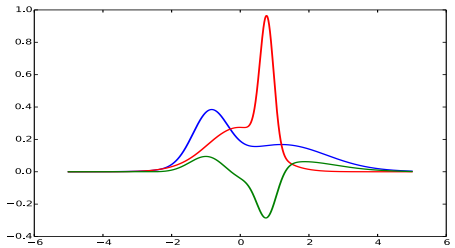
- **RKHS embedding**:

$$P \to \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$$

$$P \to [\mathbb{E}\varphi_1(X), \cdots, \mathbb{E}\varphi_s(X)] \in \mathbb{R}^s$$

- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:





$$\text{MMD}_k(P, Q) = ||\mu_k(P) - \mu_k(Q)||_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

Probability Measures in RKHSes

- For characteristic kernels, the MMD metrizes the weak- $\star$ topology on probability measures

$$\mathrm{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \rightsquigarrow P$$

- For characteristic kernels: convergence in distribution iff convergence in MMD.
- It is an Integral Probability Metric that can be computed directly from data without having to estimate the density as an intermediate step.
- Given two i.i.d samples $(x_1, \cdots, x_m)$ from P and $(y_1, \cdots, y_m)$ from Q, an unbiased estimate of the MMD is

$$\mathrm{MMD}_u^2 := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} [k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)]$$

Kernel Flows for Learning Chaotic Dynamical Systems

Kernel Flows for Learning Chaotic Dynamical Systems

- Problem **P** : Given input/output data $(x_1, y_1), \cdots, (x_N, y_N) \in \mathcal{X} \times \mathbb{R}$, recover an unknown function u^* mapping $\mathcal{X}$ to $\mathbb{R}$ such that $u^*(x_i) = y_i$ for $i \in \{1, ..., N\}$.
- In the setting of optimal recovery, Problem **P** can be turned into a well posed problem by restricting candidates for u to belong to a Banach space of functions $\mathcal{B}$ endowed with a norm defined as

$$||u||^2 = \sup_{\phi \in \mathcal{B}^*} \frac{(\int \phi(x) u(x) dx)^2}{(\int \phi(x) K(x, y) \phi(y) dx dy)}$$

and identifying the optimal recovery as the minimizer of the relative error

$$\min_v \max_u \frac{||u - v||^2}{||u||^2},$$

where the max is taken over $u \in \mathcal{B}$ and the min is taken over candidates in $v \in \mathcal{B}$ such that $v(x_i) = u(x_i) = y_i$.

Kernel Flows for Learning Chaotic Dynamical Systems

- The method of KFs is based on the premise that *a kernel is good if there is no significant loss in accuracy in the prediction error if the number of data points is halved*. This led to the introduction of

$$\rho = \frac{||v^* - v^s||^2}{||v^*||^2}$$

which is the relative error between v^*, the optimal recovery of u^* based on the full dataset $X = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, and v^s the optimal recovery of both u^* and v^* based on half of the dataset $X^s = \{(x_i, y_i) \mid i \in \mathcal{S}\}$ $(\mathrm{Card}(\mathcal{S}) = N/2)$ which admits the representation

$$v^s = (y^s)^T A^s K(x^s, \cdot)$$

with $y^s = \{y_i \mid i \in \mathcal{S}\}$, $x^s = \{x_i \mid i \in \mathcal{S}\}$, $A^s = (\Theta^s)^{-1}$, $\Theta^s_{i,j} = K(x^s_i, x^s_j)$.

Kernel Flows for Learning Chaotic Dynamical Systems

Given a family of kernels $K_\theta(x, x')$ parameterized by θ, the KF algorithm can then be described as follows :

1. Select random subvectors X^b and Y^b of X and Y (through uniform sampling without replacement in the index set $\{1, \ldots, N\}$)
2. Select random subvectors X^c and Y^c of X^b and Y^b (by selecting, at random, uniformly and without replacement, half of the indices defining X^b)
3. Let
$$\rho(\theta, X^b, Y^b, X^c, Y^c) := 1 - \frac{Y^{c,T} K_\theta(X^c, X^c)^{-1} Y_c}{Y^{f,T} K_\theta(X^b, X^b)^{-1} Y^b},$$
be the squared relative error (in the RKHS norm $\|\cdot\|_{K_\theta}$ defined by K_θ) between the interpolants u^b and u^c obtained from the two nested subsets of the dataset and the kernel K_θ
4. Evolve θ in the gradient descent direction of ρ, i.e. $\theta \leftarrow \theta - \delta \nabla_\theta \rho$
5. Repeat.

Kernel Flows for Learning Chaotic Dynamical Systems

- Let $x_1, \ldots, x_k, \ldots$ be a time series in $\mathbb{R}^d$. Our goal is to forecast x_{n+1} given the observation of $x_1, \ldots, x_n$.
- We work under the assumption that this time series can be approximated by a solution of a dynamical system of the form

$$z_{k+1} = f^\dagger(z_k, \ldots, z_{k-\tau^\dagger+1}),$$

where $\tau^\dagger \in \mathbb{N}^*$ and $f^\dagger$ may be unknown.
- Given $\tau \in \mathbb{N}^*$, the approximation of the dynamical can then be recast as that of interpolating $f^\dagger$ from pointwise measurements

$$f^\dagger(X_k) = Y_k \text{ for } k = 1, \ldots, N$$

with $X_k := (x_{k+\tau-1}, \ldots, x_k)$, $Y_k := x_{k+\tau}$ and $N = n - \tau$.

Kernel Flows for Learning Chaotic Dynamical Systems

- Given a reproducing kernel Hilbert space of candidates for $f^\dagger$, and using the relative error in the RKHS norm $\|\cdot\|$ as a loss, the regression of the data (X_k, Y_k) with the kernel K associated with provides a minimax optimal approximation of $f^\dagger$ in . This interpolant (in the absence of measurement noise) is

$$f(x) = K(x, X)(K(X, X))^{-1}Y$$

where $X = (X_1, \ldots, X_N)$, $Y = (Y_1, \ldots, Y_N)$, $k(X, X)$ for the $N \times N$ matrix with entries $k(X_i, X_i)$, and $k(x, X)$ is the N vector with entries $k(x, X_i)$.

- Use different variants of Kernel Flows (KF) to learn the kernel K from the data (X_k, Y_k).

Kernel Flows for Learning Chaotic Dynamical Systems

Assume the kernel K to be parameterized by θ. To update θ in K_θ, we minimize one of the following metrics (different variants of KFs)

- Metric associated to the RKHS norm

$$\rho(\theta, X^b, Y^b, X^c, Y^c) := 1 - \frac{Y^{c,T} K_\theta(X^c, X^c)^{-1} Y_c}{Y^{f,T} K_\theta(X^b, X^b)^{-1} Y^b}$$

- Metric associated to Lyapunov exponents and the premise that *a kernel is good if the estimate of the Lyapunov exponent obtained from the kernel approximation of the dynamics does not change if half of the data is used*:

$$\rho_L = |\lambda_{\mathsf{max},N} - \lambda_{\mathsf{max},N/2}|$$

- Metric associated to the Maximum Mean Discrepancy (MMD) and minimize

$$\rho_{\mathsf{MMD}} = \mathsf{MMD}(S_1, S_2)$$

between two different samples of the time series.

Kernel Flows for Learning Chaotic Dynamical Systems

- We use the kernel

$$
\begin{aligned}
k(x,y) &= \alpha_0 \max\{0, 1 - \frac{||x-y||_2^2|}{\sigma_0}\} + \alpha_1 e^{\frac{||x-y||_2^2}{\sigma_1^2}} + \alpha_2 e^{-\frac{||x-y||_2}{\sigma_2^2}} \\
&+ \alpha_3 e^{-\sigma_3 \sin^2(\sigma_4 \pi ||x-y||_2)} e^{-\frac{||x-y||_2^2}{\sigma_5^2}} + \alpha_4 ||x-y||_2^2
\end{aligned}
$$

Kernel Flows for Learning Chaotic Dynamical Systems

- Bernoulli map $x(k+1) = 2x(k) \mod 1$

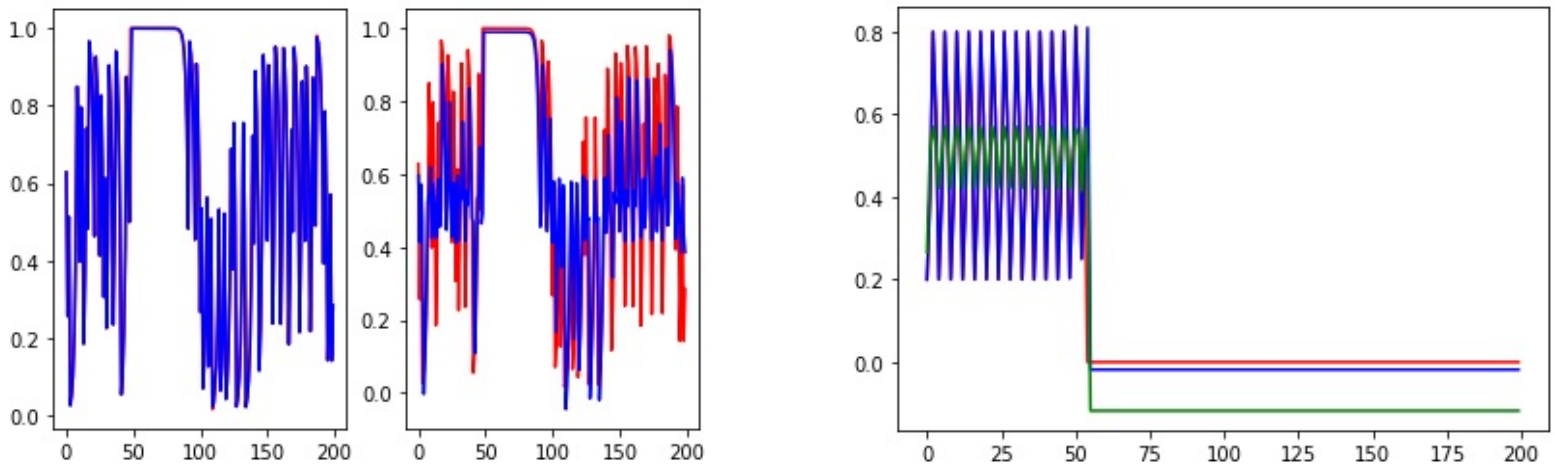


Figure: Time series generated by the true dynamics, approximation using the learned kernel and the kernel without learning for different initial conditions

Kernel Flows for Learning Chaotic Dynamical Systems

- Lorenz system

$$\begin{aligned} \frac{dx}{dt} &= s(y-x) \\ \frac{dy}{dt} &= rx - y - xz \\ \frac{dz}{dt} &= xy - bz \end{aligned}$$

with $s = 10$, $r = 28$, $b = 10/3$.

Kernel Flows for Learning Chaotic Dynamical Systems

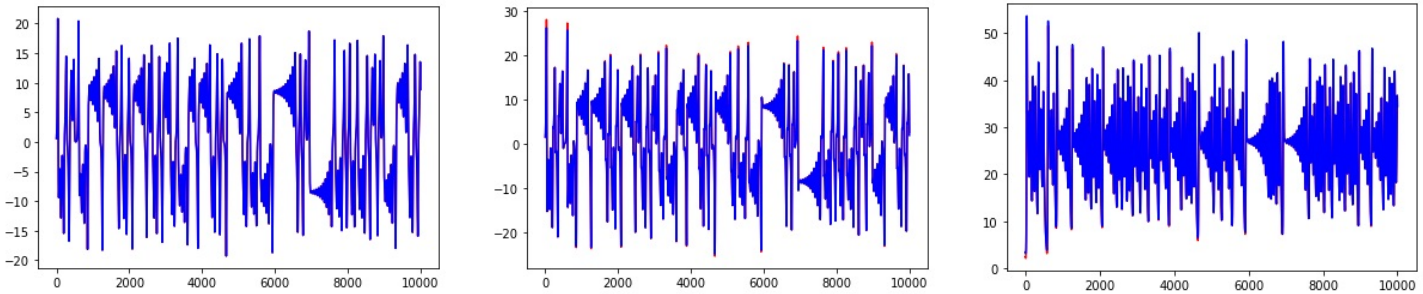


Figure: Time series generated by the true dynamics (red) and the approximation with the learned kernel (blue) - x component in the left figure, y component in the middle figure, z component in the right figure.

Kernel Flows for Learning Chaotic Dynamical Systems

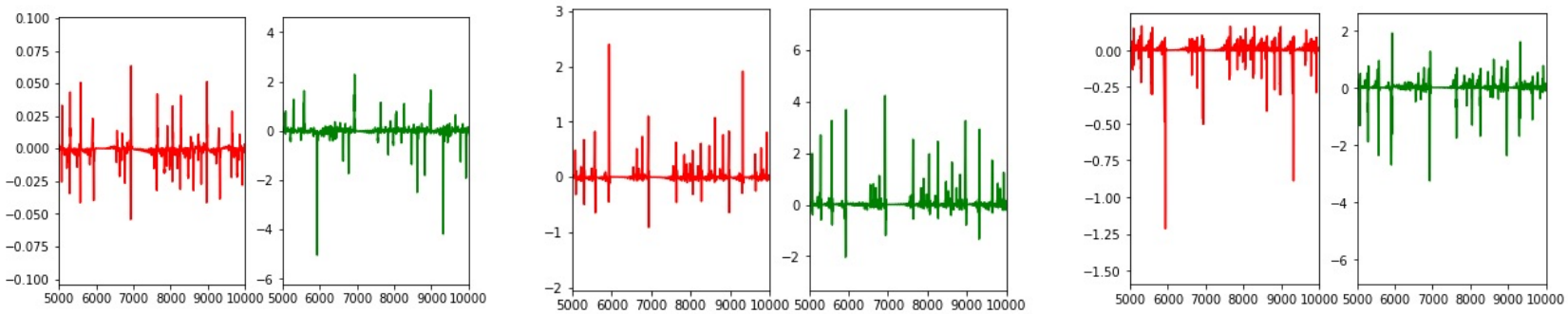


Figure: Difference between the true and the approximated dynamics with the learned kernel using ρ (red (first, third and fifth from the left)), with the initial kernel (green (second, fourth and sixth from the left)). x-component in the two figures at the left, y-component in the middle two figures, z-component in the right two figures.

Kernel Flows for Learning Chaotic Dynamical Systems

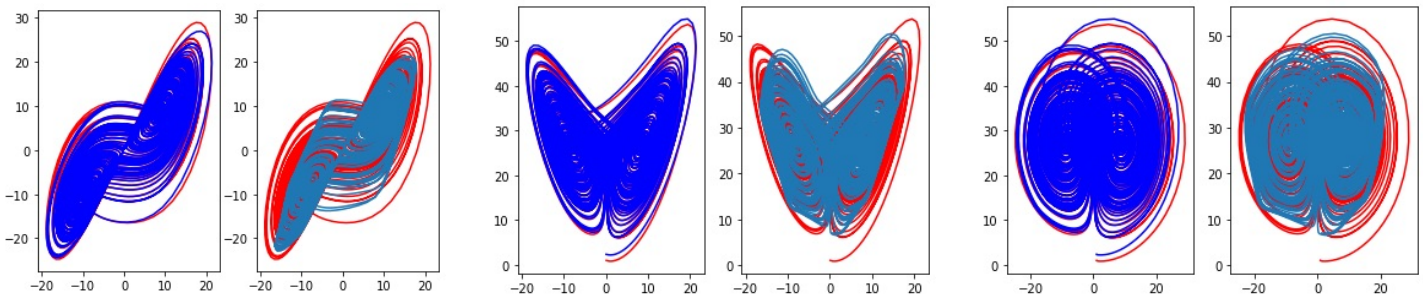


Figure: Projection of the true attractor and approximation of the attractor using a learned kernel on the XY,XZ and YZ axes (first, third and fifth from the left), Projection of the true attractor and approximation of the attractor using with initial kernel on the XY,XZ and YZ axes (second, fourth and sixth from the left)

Kernel Flows for Learning Chaotic Dynamical Systems

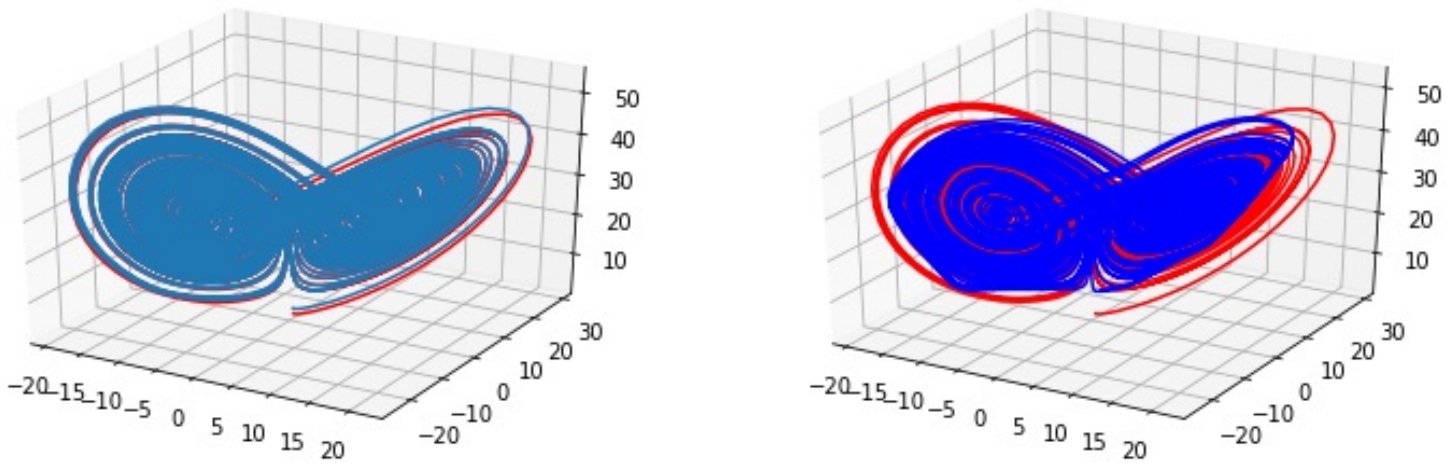
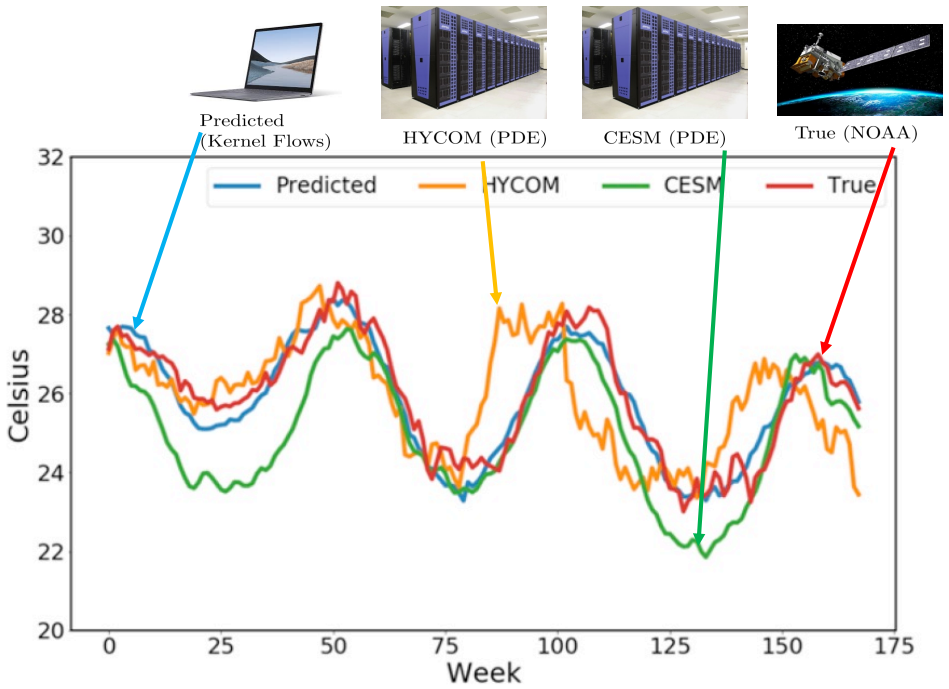


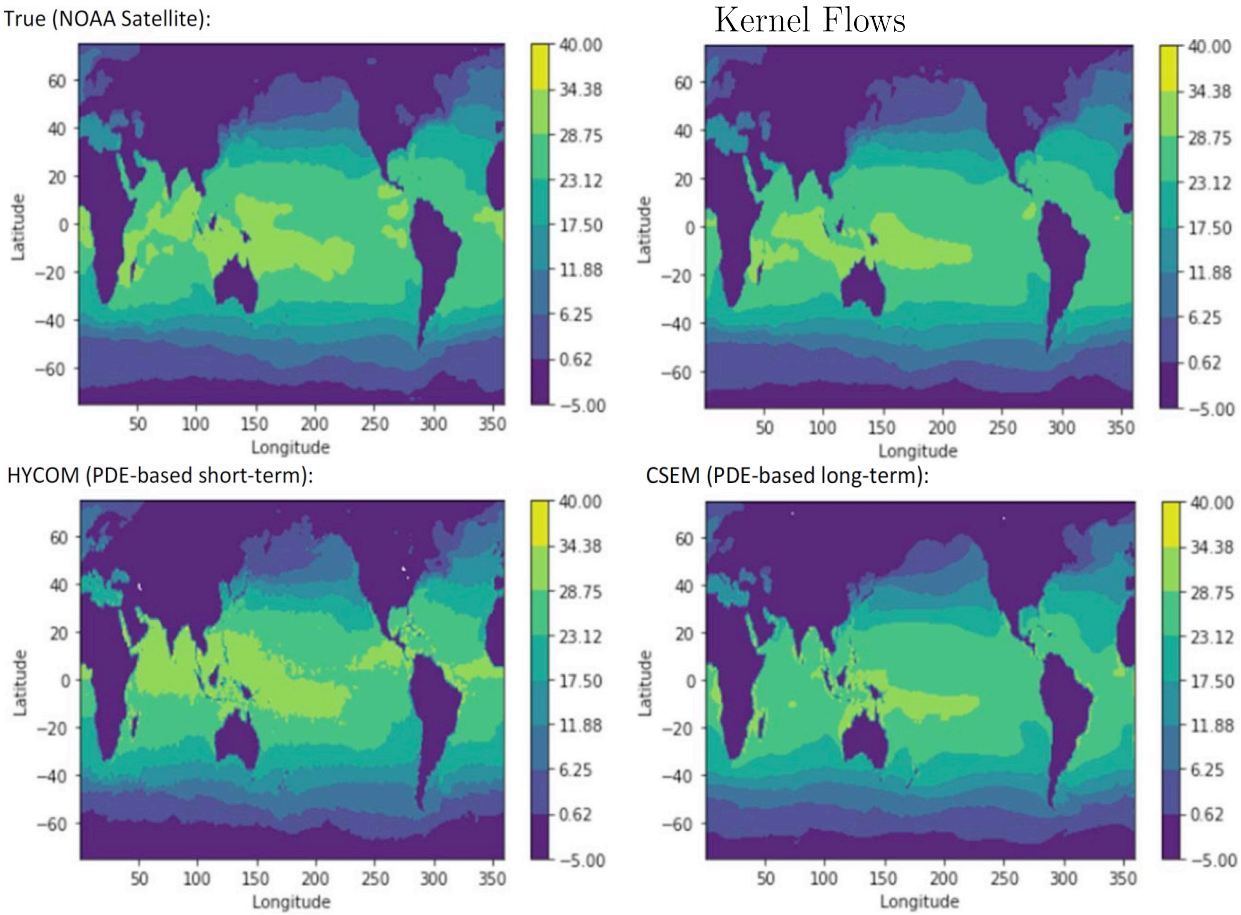
Figure: True attractor (blue) and approximation of the attractor using a learned kernel (red) [left], True attractor (blue) and approximation of the attractor using initial kernel (red) [right]

Kernel Flows for Learning Chaotic Dynamical Systems

- HYCOM: 800 core-hours per day of forecast on a Cray XC40 system
- CESM: 17 million core-hours on Yellowstone, NCAR's high-performance computing resource
- Architecture optimized LSTM: 3 hours of wall time on 128 compute nodes of the Theta supercomputer.
- Our method: 40 seconds to train on a single node machine (laptop) without acceleration



Kernel Flows for Learning Chaotic Dynamical Systems



Center Manifold Approximation

Center Manifold Analysis

Consider a dynamical system

$$\dot{x} = f(x) = Fx + \bar{f}(x)$$

of large dimension n, and $F = \frac{\partial f}{\partial x}(x)|_{x=0}$.
Suppose $x = 0$ is an equilibrium, i.e. $f(0) = 0$.

- Goal: Analyze the stability of this equilibrium.
- If F has all its eigenvalues with negative real parts $\Rightarrow$ The origin is asymptotically stable.
- If F has some eigenvalues with positive real parts $\Rightarrow$ The origin is unstable.

Center Manifold Analysis

- If $\sigma(F) \leq 0$ (some eigenvalues of F are with zero real parts with the rest of the eigenvalues having negative real parts): The linearization fails to determine the stability properties of the origin.
- After a linear change of coordinates, we have

$$\begin{aligned} \dot{x}_1 &= F_1 x_1 + \bar{f}_1(x_1, x_2) \\ \dot{x}_2 &= F_2 x_2 + \bar{f}_2(x_1, x_2) \end{aligned}$$

where $\sigma(F_1) = 0$ and $\sigma(F_2) < 0$.

- Intuitively, we expect the stability of the equilibrium to only depend on the nonlinear terms $\bar{f}_1(x_1, x_2)$. The center manifold theorem correctly formalizes this intuition.

Center Manifold Analysis

- A center manifold is an invariant manifold, $x_2 = \theta(x_1)$, tangent to the x_1 directions at $x = 0$.
- Since

$$\begin{aligned} \dot{x}_1 &= F_1 x_1 + \bar{f}_1(x_1, x_2) \\ \dot{x}_2 &= F_2 x_2 + \bar{f}_2(x_1, x_2) \end{aligned}$$

and $x_2 = \theta(x_1)$, we deduce that θ satisfies the PDE

$$F_2\theta(x_1)+\bar{f}_2(x_1,\theta(x_1))=\frac{\partial\theta}{\partial x_1}(x_1)\Big(F_1x_1+\bar{f}_1(x_1,\theta(x_1))\Big).$$

- The Center Manifold Theorem ensures that there are smooth solutions to this PDE.

Center Manifold Analysis

- The center dynamics is the dynamics on the center manifold,

$$\dot{x}_1 = F_1 x_1 + \bar{f}_1(x_1, \theta(x_1)).$$

- Center Manifold Theorem: The equilibria $x_1 = 0, x_2 = 0$ of the original dynamics is locally asymptotically stable iff the equilibria $x_1 = 0$ of the center dynamics is locally asymptotically stable.
- After solving the PDE, this reduces the problem to analyzing the nonlinear stability of a lower dimensional system.
- Our Contributions: kernel methods to approximate the center manifold, a data-based version of the center manifold theorem.

Center Manifold Analysis: Main results

- Let $\hat{\theta}$ be an approximant of the center manifold θ. Given the constraints $\theta(0) = 0$ and $D_x\theta(0) = 0$, we use a generalized version of the representer theorem and write

$$\hat{\theta}(x) = \sum_{i=1}^{N+1} k(x, x_i)\alpha_i + \sum_{i=1}^{m} \partial_i^{(2)} k(x, 0)\beta_i,$$

- ($\Rightarrow$) Under certain conditions, we prove that if the equilibrium $x_1 = 0$ of

$$\dot{x}_1 = F_1 x_1 + \bar{f}_1(x_1, \hat{\theta}(x_1)).$$

is asymptotically stable then the equilibrium $x_1 = 0$, $x_2 = 0$ of the full order dynamics is asymptotically stable ((asymptotic) stability-preserving property- in one direction at least, second direction is still missing).
- We also prove that $||x_{1,\theta}(t) - x_{1,\hat{\theta}}(t)||$ is bounded.

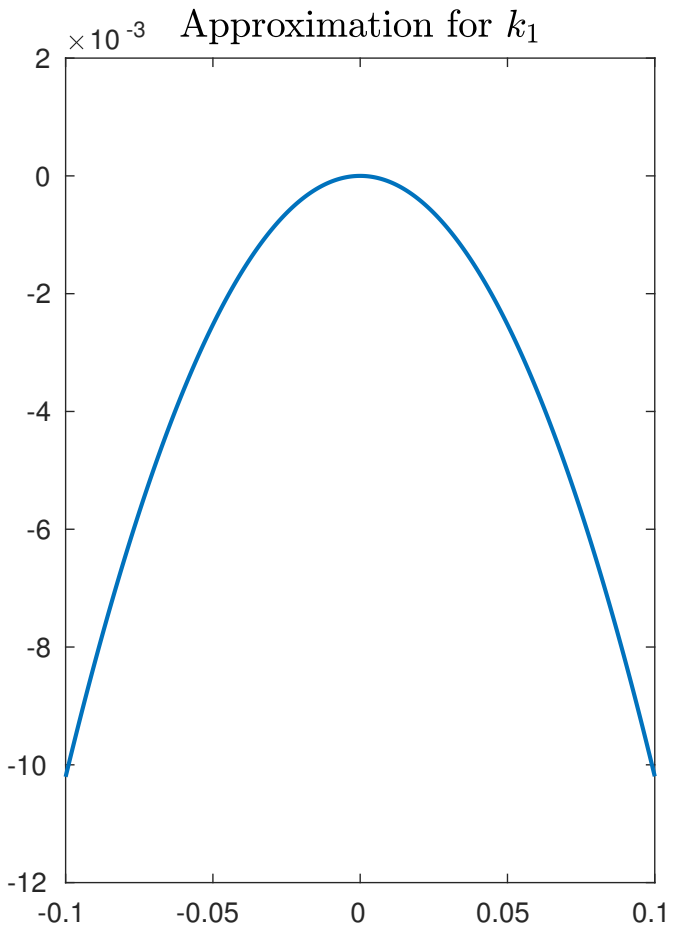
Numerical Experiments: Example 1

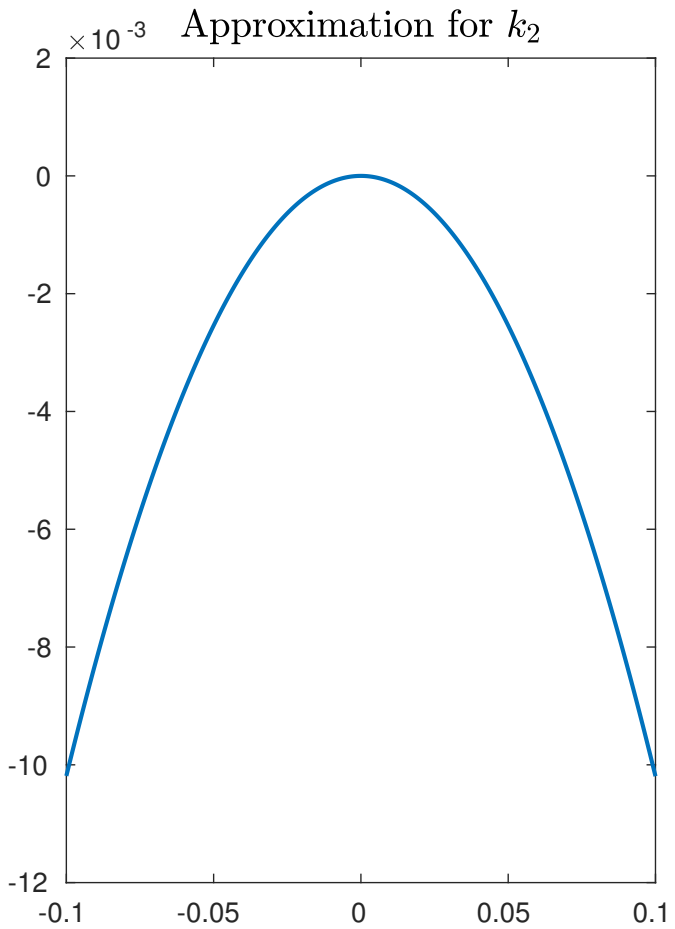
- We consider the 2-dimensional system

$$\begin{aligned} \dot{x} &= f_1(x,y) = xy \\ \dot{y} &= f_2(x,y) = -y - x^2 \end{aligned} \tag{1}$$

- Analytically, the center manifold is $y = -x^2 + O(x^3)$.
- We generate the training data by solving the system with an implicit Euler scheme for initial time $t_0 = 0$, final time $T = 1000$ and with the timestep $\Delta t = 0.1$. We initiate the numerical procedure with initial values $(x_0, y_0) \in \{\pm 0.8\} \times \{\pm 0.8\}$ and store the resulting data pairs in X and Y after discarding all data whose x-values are not contained in the neighborhood $[-0.1, 0.1]$ which results in $N = 38248$ data pairs. We use the kernels $k_1(x,y) := (1 + xy/2)^4$ and $k_2(x,y) = e^{-(x-y)^2/2}$.

Numerical Experiments: Example 1





Numerical Experiments: Example 2

- Consider the $(2+1)$-dimensional system

$$\dot{x} = L_1 x + N_1(x,y) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + y \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$
$$\dot{y} = L_2 y + N_2(x,y) = -y - x_1^2 - x_2^2 + y^2.$$

Numerical Experiments: Example 2

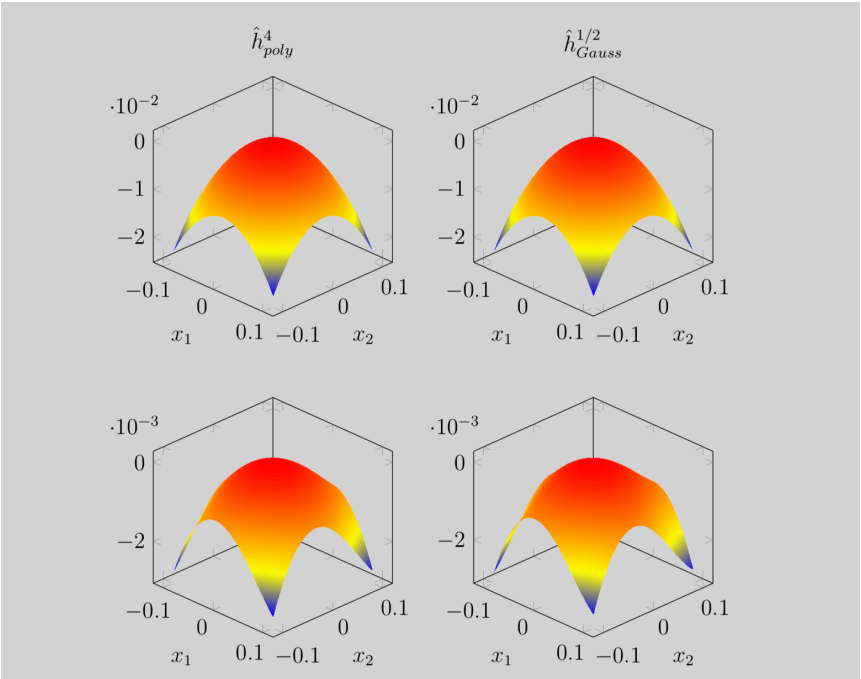


Figure: Approximations $\hat{h}^4_{poly}$ and $\hat{h}^{1/2}_{Gauss}$ of the center manifold (first row), and corresponding residuals r^4_{poly} and $r^{1/2}_{Gauss}$ (second row)

Construction of Lyapunov Functions from Data

Summary of the Approach

- We will consider a nonlinear ODE $\dot{x} = f(x)$, $x \in \mathbb{R}^n$ and assume that f is not known but $x(t_i)$, $i = 1, \cdots, N$, are known.
- We approximate f from $x(t_i)$, $i = 1, \cdots, N$.
- We find a Lyapunov function $\hat{V}$ for $\hat{f}$.
- We prove that $\hat{V}$ is also a Lyapunov function for f.

Lyapunov Functions

- Consider the system of ODEs $\Sigma : \left\{ \begin{array}{rcl} \dot{x} & = & f(x), \\ x(0) & = & \xi \end{array} \right.$ with $x \in \mathbb{R}^n$, $f \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$ where $\sigma \geq 1$, $n \in \mathbb{N}$.
Flow $S_t\xi := x(t)$, solution of Σ.
- **Assumptions**
  - 0 is an equilibrium $(f(0) = 0)$
  - 0 is exponentially asymptotically stable (real parts of all eigenvalues of $Df(0)$ are negative)
- Definition (Basin of Attraction) The basin of attraction of 0 is

$$\mathcal{A} := \{\xi \in \mathbb{R}^n | S_t\xi \to_{t\to\infty} 0\}$$

- The basin of attraction $\mathcal{A}$ can be determined using Lyapunov functions.

Lyapunov Functions

Theorem (Lyapunov 1893)
Let $V : \mathbb{R}^n \to \mathbb{R}^+$, $K \subset \mathbb{R}^n$ a compact set.

- V decreases along solutions, i.e. (if V is smooth)

$$V'(x) = \frac{d}{dt} V(x(t))|_{t=0} = \nabla V(x) \cdot f(x) < 0$$

  for all $x \in K \setminus \{0\}$ (V' is the orbital derivative $=$ derivative along the solution)
- K is sublevel set of V, i.e. $K = \{x \in \mathbb{R}^n | V(x) \leq R\}$.

Then $K \subset \mathcal{A}$.

Existence of Lyapunov Functions

- "Converse Theorems" (Massera 1949) etc. - but not constructive !
- Theorem (Existence of V, Bhatia) Let $f \in C^\sigma$, $\sigma \geq 1$, 0 exponentially stable equilibrium. Then there exists $V \in C^\sigma(\mathcal{A}, \mathbb{R})$ with

$$V'(x) := \nabla V(x) \cdot f(x) = -||x||^2 \quad \text{for all} \quad x \in \mathcal{A}$$

The Lyapunov function V is uniquely defined up to a constant.

- Idea: $V(x) = \int_0^\infty ||S_t x||^2 dt$.

Computation of Lyapunov Functions

- Giesl proposed an algorithm to approximate Lyapunov functions using radial basis functions.
- Error estimates for this approach have been proved by Giesl and Wendland.
- The method is based on finding an approximate solution of a first-order linear PDE:

$$LV(x) = -||x||^2 \quad (LV(x) = -p(x) \quad \text{with} \quad p(x) > 0)$$

with $LV := V'(x) := \nabla V(x) \cdot f(x)$.

Computation of Lyapunov Functions (Giesl, 2007)

- Theorem (Giesl, 2007)

Consider $\dot{x} = f(x)$ with $f \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$ and let x_0 be an equilibrium such that all eigenvalues of $Df(x_0)$ have a negative real part. Let $p(x) \in C^\sigma(\mathbb{R}^n, \mathbb{R})$ satisfy the following conditions: a.) $p(x) > 0$ for $x \neq x_0$, b.) $p(x) = O(||x - x_0||_2^\eta)$ with $\eta > 0$ for $x \to x_0$, c.) For all $\epsilon > 0$, p has a lower positive bound on $\mathbb{R}^n \setminus B(x_0, \epsilon)$ where $B(x_0, \epsilon)$ is a the ball centered at x_0 of radius ϵ.
Then there exists a Lyapunov function $V_1 \in C^\sigma(A(x_0), \mathbb{R})$ such that $V_1(x_0) = 0$ and

$$LV_1(x) = f_1(x) := -p(x), \quad \text{for all } \ x \in A(x_0),$$

where $A(x_0)$ is the basin of attraction of x_0.

Computation of Lyapunov Functions (Giesl, 2007)

Algorithm: Let $\Phi(x) = \psi_k(||x||)$ be a radial function where ψ_k is a Wendland function (compact support). Consider the grid points $X_N = \{x_1, \cdots, x_N\} \subset \mathbb{R}^n$. Consider the following ansatz

$$V_1(x) = \sum_{k=1}^{N} \beta_k (\delta_{x_k} \circ L)^y \Phi(x - y),$$

where $(\delta_{x_k} \circ L)^y$ denotes differentiation with respect to y then evaluation at $y = x_k$.

Computation of Lyapunov Functions (Giesl, 2007)

By considering the interpolation conditions

$$LV_1(x_j) = LV(x_j) = f_1(x_j),$$

and by plugin in the ansatz

$$\sum_{i=1}^{N} \beta_k \underbrace{(\delta_{x_j} \circ L)^x (\delta_{x_k} \circ L)^y \Phi(x-y)}_{=a_{jk}} = LV(x_j) = f_1(x_j) =: \gamma_j,$$

one gets a system of linear algebraic equations for the β in βs:

$$A\beta = \gamma,$$

where the matrix A is symmetric and positive definite.

Estimates on Lyapunov Functions (Giesl and Wendland, 2007)

- Theorem(Giesl & Wendland, 2007)

Let ψ_k, $k \in \mathbb{N}$, be a Wendland function and let $\Phi(x) = \psi_k(||x||) \in C^{2k}(\mathbb{R}^n, \mathbb{R})$ be a radial basis function. Let $f \in C^\sigma(\mathbb{R}^n, \mathbb{R})$ where $\sigma \geq \frac{n+1}{2} + k$. Then, for each compact set $K_0 \subset A(x_0)$ there is C^* such that

$$|V'(x) - V_1'(x)| \leq C^* h^\theta \text{ for all } x \in K_0,$$

where $h := \max_{y \in K_0} \min_{x \in X_n} ||x - y||$ is the fill distance and $\lambda = 1/2$ for $k = 1$ and $\lambda = 1$ for $k \geq 2$ (or $\lambda = k - 1/2$).

Computation of Lyapunov Functions from Data

- Giesl's approach assumes that the right hand side of (ODE) is known, and sampled values of f are used at chosen grid points.
- We assume the underlying system Σ where f is unknown but, instead, we have sampled data values $(x_i; y_i)|_{i=1}^m$ with $y_i = f(x_i) + \eta$, $i = 1, \cdots, m$ with each $x_i \in A(\bar{x})$, and $\eta \in \mathbb{R}^d$ is an independent random variable drawn from a probability distribution with zero mean and variance $\sigma^2 \in \mathbb{R}^d$.
- Our approximation algorithm looks for suitable functions in an RKHS.
- Error estimates are derived for some RKHSes that are also Sobolev spaces.

Numerical Experiment

Consider the nonlinear system

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_1 x_2^2 \\ \dot{x}_2 &= -x_2 - x_2 x_1^2 \end{aligned} \tag{2}$$

It can be checked that $V(x) = x_1^2 + x_2^2$ is a Lyapunov function for the system. First, we used Algorithm 1 to approximate the right hand side of (2) with $m = 400$ points and $z := (x_i, y_i)_{i=1}^m$ are such that the points x_i are equidistantly distributed over $[-0.95, 0.95]$.

Numerical Experiment

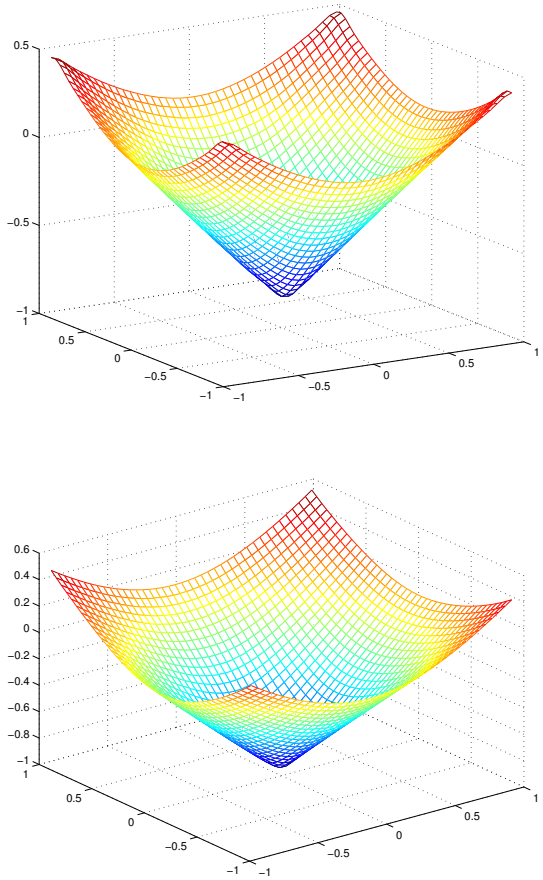


Figure: Lyapunov function using Algorithm 2 with 360 points(top), 1520 points (bottom)

Numerical Experiment

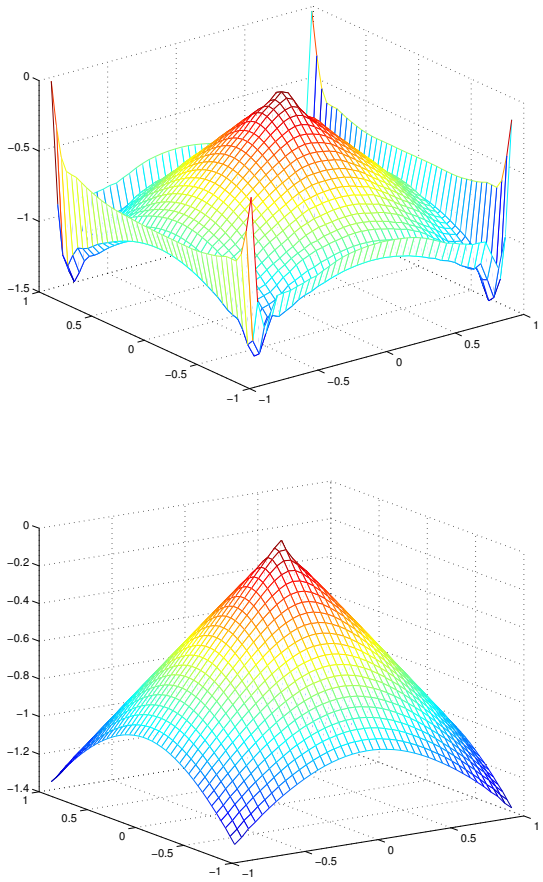


Figure: Orbital derivative of the Lyapunov function with respect to the original system using Algorithm 2 with 360 points(top), 1520 points (bottom).

Detection of Critical Transitions for MultiScale Systems

Detection of Critical Transitions for MultiScale Systems

- Consider the fast-slow SDE

$$\begin{aligned} \dot{x}_1 &= \frac{1}{\epsilon} f_1(x_1, x_2) + \frac{\sigma_1}{\sqrt{\epsilon}} \eta_1(\tau), \\ \dot{x}_2 &= f_2(x_1, x_2) + \sigma_2 \eta_2(\tau) \end{aligned}$$

where $f_1 \in \mathcal{C}(\mathbb{R}^2; \mathbb{R})$ and $f_2 \in \mathcal{C}(\mathbb{R}^2; \mathbb{R})$ are Lipschitz and η_1, η_2 are independent white Gaussian noises.

- x_1 is a fast variable in comparison to the slow variable x_2.
- The set $C_0 = \{(x_1, x_2) \in \mathbb{R}^2 : f_1(x_1, x_2) = 0\}$ is called the critical manifold.

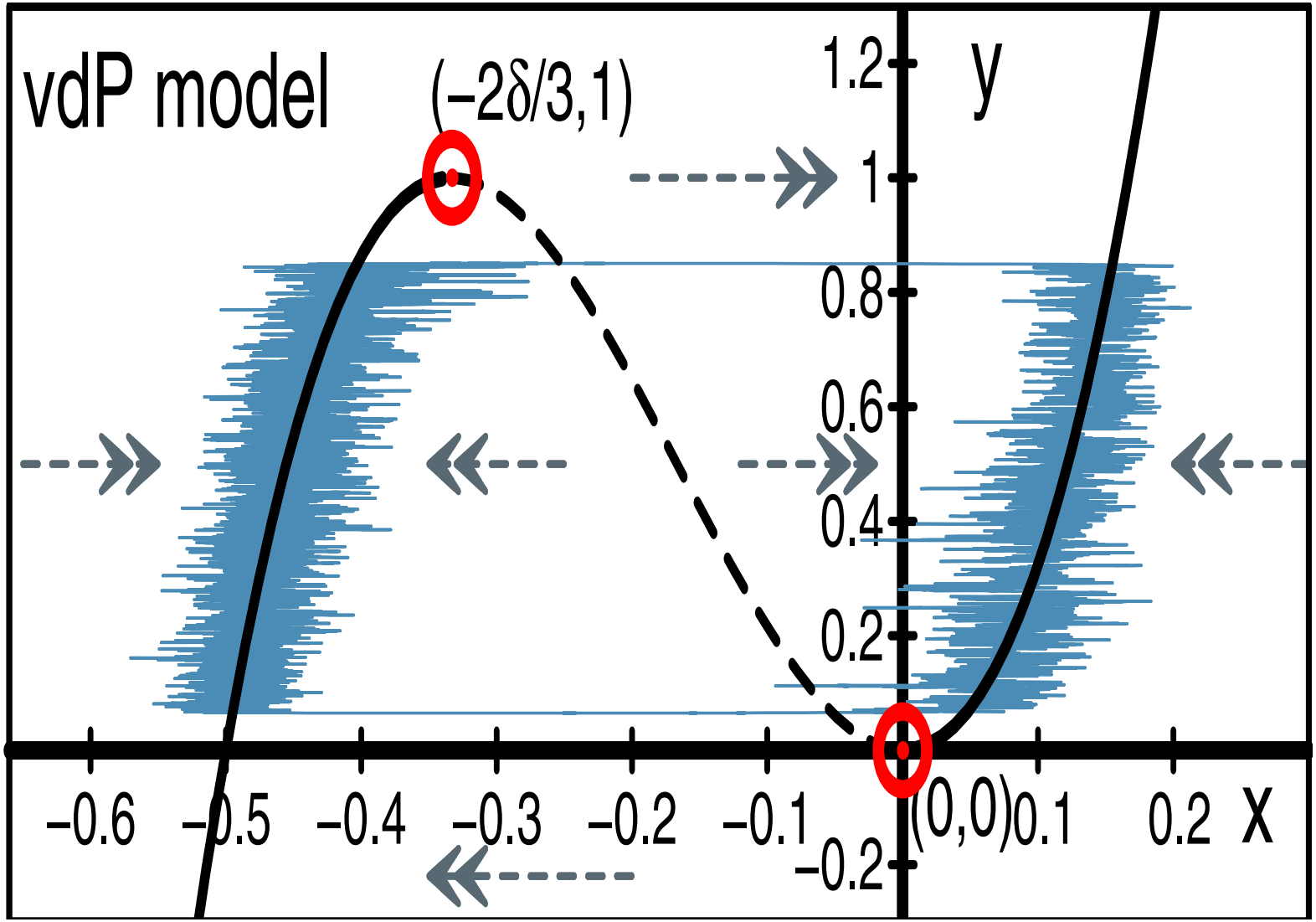
MultiScale Systems

- The van der Pol model.
- The equations of the model are

$$\begin{aligned} \dot{x}_1 &= \frac{1}{\epsilon}(x_2 - \frac{27}{4\,\delta^3}x_1^2(x_1+\delta)) + \frac{\sigma_1}{\sqrt{\epsilon}}\eta_1(t) \\ \dot{x}_2 &= -\frac{\delta}{2} - x_1 + \sigma_2\eta_2(t) \end{aligned}$$

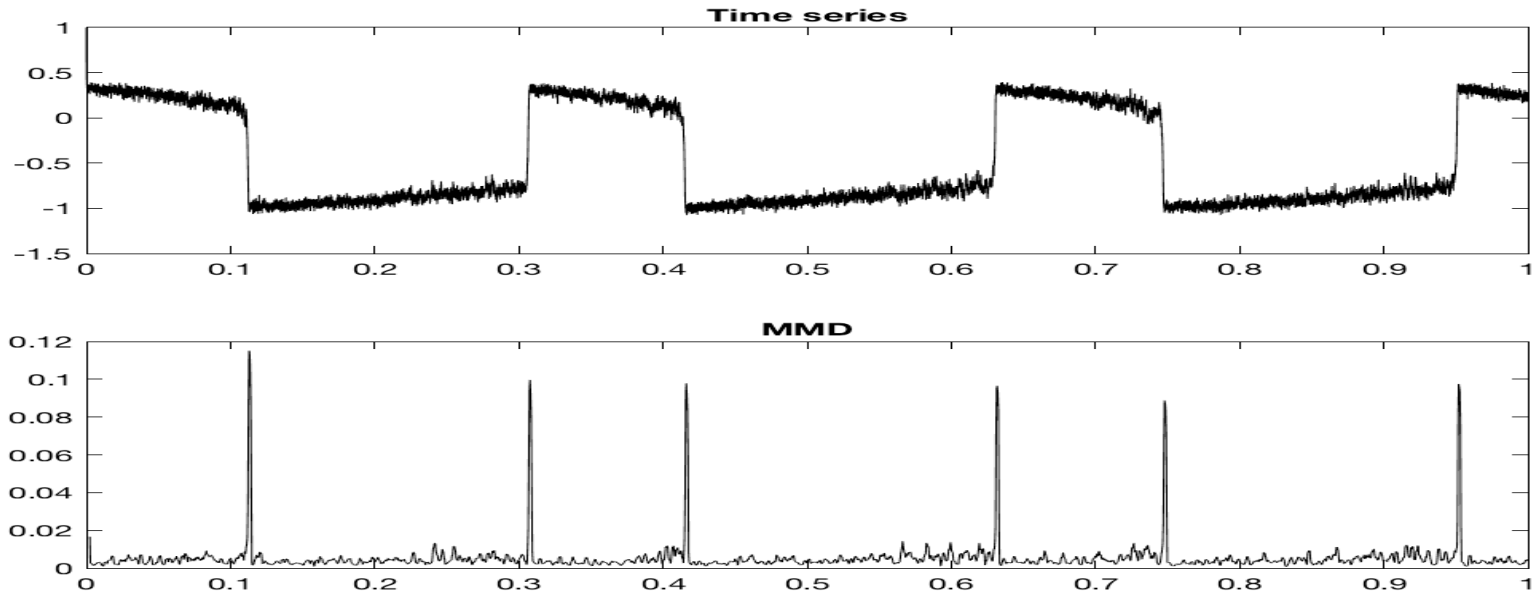
$\delta = 1, \sigma_1 = 0.1, \sigma_2 = 0.1, \varepsilon = 0.01.$

MultiScale Systems



MultiScale Systems

- Numerical Simulation



Approximation of Control Systems in Reproducing Kernel Hilbert Spaces

Review of Some Concepts from Linear Control Theory

- Consider a linear control system

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

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^q$, $y \in \mathbb{R}^p$, (A, B) is controllable, (A, C) is observable and A is Hurwitz.

- We define the controllability and the observability Gramians as, respectively, $W_c = \int_0^\infty e^{At} B B^\top e^{A^\top t}\, dt$, $W_o = \int_0^\infty e^{A^\top t} C^\top C e^{At}\, dt$.
- These two matrices can be viewed as a measure of the controllability and the observability of the system.

Review of Some Concepts from Linear Control Theory

- Consider the past energy, $L_c(x_0)$, defined as the minimal energy required to reach x_0 from 0 in infinite time

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

- Consider the future energy, $L_o(x_0)$, 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.

$$L_o(x_0) = \frac{1}{2} \int_{0}^{\infty} \|y(t)\|^2 \, dt,$$

for $x(t_0) = x_0$ and $u(t) = 0, t \geq 0$.

Review of Some Concepts from Linear Control Theory

- In the linear case, it can be shown that

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

- Moreover, W_c and W_o satisfy the following Lyapunov equations

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

Controllability and Observability Energies in Model Reduction of Linear Control Systems

- Gramians have several uses in Linear Control Theory. For example, for the purpose of model reduction.
- Balancing: find a representation where the system's observable and controllable subspaces are aligned so that reduction, if possible, consists of eliminating uncontrollable states which are also the least observable.
- More formally, we would like to find a new coordinate system such that

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

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. If (F, G) is controllable and (F, H) is observable, then there exists a transformation such that the state space expressed in the transformed coordinates (TFT^{-1}, TG, HT^{-1}) is balanced and $TW_cT^\top = T^{-\top}W_oT^{-1} = \Sigma$.

Balancing of Linear Control Systems

- Typically one looks for a gap in the singular values $\{\sigma_i\}$ for guidance as to where truncation should occur. If we see that there is a k such that $\sigma_k \gg \sigma_{k+1}$, then the states most responsible for governing the input-output relationship of the system are $(x_1, \cdots, x_k)$ while $(x_{k+1}, \ldots, x_n)$ are assumed to make negligible contributions.
- Although several methods exist for computing T, the general idea is to compute the Cholesky decomposition of W_o so that $W_o = ZZ^\top$, and form the SVD $U\Sigma^2 U^\top$ of $Z^\top W_c Z$. Then T is given by

$$T = \Sigma^{\frac{1}{2}} U^\top Z^{-1}.$$

Controllability and Observability Energies for Nonlinear Systems

- Consider the nonlinear system Σ

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

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 $h(0) = 0$.

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

Controllability and Observability Energies for Nonlinear Systems

- Theorem (Scherpen, 1993) 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

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

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

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

under the assumption that this equation 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.

Balancing of Nonlinear Systems

• Theorem (Scherpen) Consider system Σ under Hypothesis H and the assumptions in the preceding theorem. Then, there exists a neighborhood W of the origin and coordinate transformation $x = \varphi(z)$ on W converting the energy functions into the form

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

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

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

Balancing of Nonlinear Systems

- In the above framework for balancing of nonlinear systems, one needs to solve (or numerically evaluate) the PDEs and compute the coordinate change $x = \varphi(z)$.
- However there are no systematic methods or tools for solving these equations.
- Various approximate solutions based on Taylor series expansions have been proposed Krener (2007, 2008), Fujimoto and Tsubakino (2008).
- Newman and Krishnaprasad (2000) 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.
- An essentially linear empirical approach, similar to Moore's empirical approach, was proposed by Lall, Marsden and Glavaski (2002).

Computing the Controllability and Observability Energies: Linear Case

- Analytic Approach: The Gramians W_c and W_o satisfy the Lyapunov equations

$$FW_c + W_cF^\top = -GG^\top,$$
$$F^\top W_o + W_oF = -H^\top H.$$

- Data-Based Approach: Moore showed that W_c and W_o can be obtained from the impulse responses of Σ_L. For instance,

$$W_c = \int_0^\infty X(t)X(t)^T dt, \quad W_o = \int_0^\infty Y^T(t)Y(t)dt$$

where $X(t)$ is the response to $u^i(t) = e_i$ with $x(0) = 0$, and $Y(t)$ is the output response to $u(t) = 0$ and $x(0) = e_i$.
Given $X(t)$ and $Y(t)$, one can perform PCA to obtain W_c and W_o respectively.

Empirical Estimates of the Gramians

The observability and controllability Gramians may be estimated statistically from typical system trajectories:

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

where $t_i \in [0,T], i = 1,\ldots,N$, $X(t) = \begin{bmatrix} x^1(t) & \cdots & x^m(t) \end{bmatrix}$, and $Y(t) = [y^1(t) \ \cdots \ y^n(t)]^\top$ if $\{x^j(t)\}_{j=1}^m, \{y^j(t)\}_{j=1}^n$ are measured (vector-valued) responses and outputs of the system.

Computing the Controllability and Observability Energies for Nonlinear Systems

Questions

- How to compute the controllability and observability energies from data ?
- How to extend Moore's empirical approach to Nonlinear Control Systems ?
- Are there "Gramians" for Nonlinear Systems ? and in the affirmative, how to compute them from data ?
- Idea ! Use of kernel methods. A kernel based procedure may be interpreted as mapping the data, through "feature maps", from the original input space into a potentially higher dimensional Reproducing Kernel Hilbert Space where linear methods may then be used.

Controllability and Observability Energies of Nonlinear Systems in RKHSes

- We consider a general nonlinear system of the form

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

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0,0) = 0$, and $h(0) = 0$.

- Assume that the method of linear balancing can be applied to the nonlinear system when lifted into an RKHS.
- In the linear case, $L_c(x_0) = \frac{1}{2} x_0^T W_c^{-1} x_0$ and $L_o(x_0) = \frac{1}{2} x_0^T W_o x_0$ can be rewritten as $L_c(x_0) = \frac{1}{2} \left\langle W_c^{\dagger} x_0, x_0 \right\rangle$ and $L_o(x_0) = \frac{1}{2} \langle W_o x_0, x_0 \rangle$.
- In the nonlinear case, it may be tempting to write, in $\mathcal{H}$, $L_c(x) = \frac{1}{2} \left\langle W_c^{\dagger} h, h \right\rangle$ and $L_o(x) = \frac{1}{2} \langle W_o h, h \rangle$ where $h = \Phi(x) = K(x, \cdot)$ and $\Phi : \mathbb{R}^n \to \mathcal{H}$. However, there are some complications...

Controllability and Observability Energies of Nonlinear Systems in RKHSes

- We can show that

$$\begin{aligned}\hat{L}_c(x) &= \tfrac{1}{2}\left\langle (\tfrac{1}{m}\mathcal{R}_{\mathbf{x}}^*\mathcal{R}_{\mathbf{x}} + \lambda I)^{-2}\tfrac{1}{m}\mathcal{R}_{\mathbf{x}}^*\mathcal{R}_{\mathbf{x}}K_x, K_x\right\rangle \\ &= \tfrac{1}{2m}\left\langle \mathcal{R}_{\mathbf{x}}^*(\tfrac{1}{m}\mathcal{R}_{\mathbf{x}}\mathcal{R}_{\mathbf{x}}^* + \lambda I)^{-2}\mathcal{R}_{\mathbf{x}}K_x, K_x\right\rangle \\ &= \tfrac{1}{2m}\mathbf{k_c}(x)^\top(\tfrac{1}{m}K_c + \lambda I)^{-2}\mathbf{k_c}(x),\end{aligned}$$

where $\mathbf{k_c}(x) := \mathcal{R}_{\mathbf{x}}K_x = \big(K(x, x_\mu)\big)_{\mu=1}^{Nq}$ is the Nq-dimensional column vector containing the kernel products between x and the controllability samples.

Controllability and Observability Energies of Nonlinear Systems in RKHSes

- Similarly, letting $\mathbf{x}$ now denote the collection of $m = Np$ observability samples, we can approximate the future output energy by

$$\begin{aligned} \hat{L}_o(x) &= \tfrac{1}{2}\left\langle \widehat{W}_o K_x, K_x \right\rangle \\ &= \tfrac{1}{2m}\left\langle \mathcal{R}_{\mathbf{x}}^* \mathcal{R}_{\mathbf{x}} K_x, K_x \right\rangle \\ &= \tfrac{1}{2m}\mathbf{k_o}(x)^\top \mathbf{k_o}(x) = \tfrac{1}{2m}\left\|\mathbf{k_o}(x)\right\|_2^2 \end{aligned} \tag{3}$$

where $\mathbf{k_o}(x) := \big(K(x, d_\mu)\big)_{\mu=1}^{Np}$ is the Np-dimensional column vector containing the kernel products between x and the observability samples.

Balanced Reduction of Nonlinear Control Systems in RKHS

- We consider a general nonlinear system of the form

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

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $f(0,0) = 0$, and $h(0) = 0$. We assume that the origin of $\dot{x} = f(x,0)$ is asymptotically stable.

Proposed Data-Driven Approach:

- Assume that we can apply the method of linear balancing *when the system is lifted to a high (possibly infinite) dimensional feature space*.
- Carry out balancing and truncation (linear techniques) implicitly in the feature space (discard unimportant states).
- Construct a nonlinear reduced-order model by learning approximations to f, h defined directly on the reduced state space.

Balancing in RKHS

Idea: We can perform balancing/truncation in feature space by lifting the data into $\mathcal{H}$ via Φ, and simultaneously diagonalizing the corresponding covariance operators.

The standard empirical controllability Gramian (in $\mathbb{R}^n$)

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

becomes

$$C_c = \frac{T}{mN} \sum_{i=1}^{N} \sum_{j=1}^{m} \left\langle \Phi\left(x^j(t_i)\right), \cdot \right\rangle_{\mathcal{H}} \Phi\left(x^j(t_i)\right)$$

for example.

Balancing in RKHS

- "Balancing" is carried out implicitly in $\mathcal{H}$ by simultaneous diagonalization of K_c and K_o.
- If $K_c^{1/2} K_o K_c^{1/2} = U \Sigma^2 U^\top$, we can define the aligning transformation

$$T = \Sigma^{1/2} U^\top \sqrt{K_c^\dagger}.$$

- The dimension of the state space is reduced by discarding small eigenvalues $\{\Sigma_{ii}\}_{i=q+1}^{n}$, and projecting onto the subspace in $\mathcal{H}$ associated with the first $q < n$ largest eigenvalues.
- This leads to the *nonlinear* state-space dimensionality reduction map $\Pi : \mathbb{R}^n \to \mathbb{R}^q$ given by

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

where

$$\mathbf{k}_c(x) := \left(K(x, x^1(t_1)), \ldots, K(x, x^m(t_N))\right)^\top.$$

An Experiment

Consider the $7-D$ system (Nilsson, 2009)

$$\dot{x}_1 = -x_1^3 + u \qquad \dot{x}_2 = -x_2^3 - x_1^2 x_2 + 3x_1 x_2^2 - u$$

$$\dot{x}_3 = -x_3^3 + x_5 + u \qquad \dot{x}_4 = -x_4^3 + x_1 - x_2 + x_3 + 2u$$

$$\dot{x}_5 = x_1 x_2 x_3 - x_5^3 + u \qquad \dot{x}_6 = x_5 - x_6^3 - x_5^3 + 2u$$

$$\dot{x}_7 = -2x_6^3 + 2x_5 - x_7 - x_5^3 + 4u$$

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

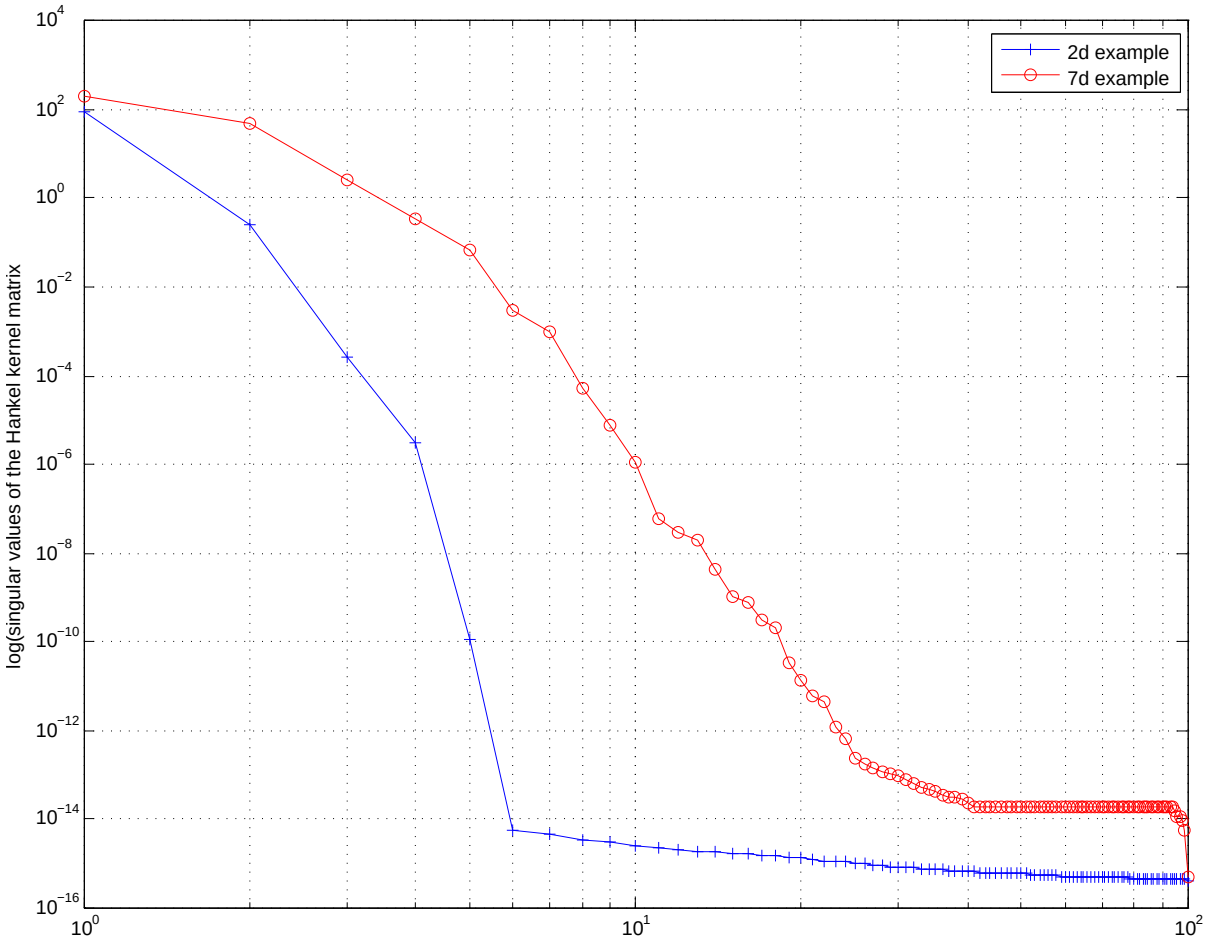
Experiment: Inputs

- Excite with impulses: inputs (K_c) and initial conditions $(K_o, u = 0)$.
- Learn $\hat{f}, \hat{h}$ using a 10Hz square wave input signal u.
- Reduce to a second-order system.
- Simulate the reduced system with a different input,

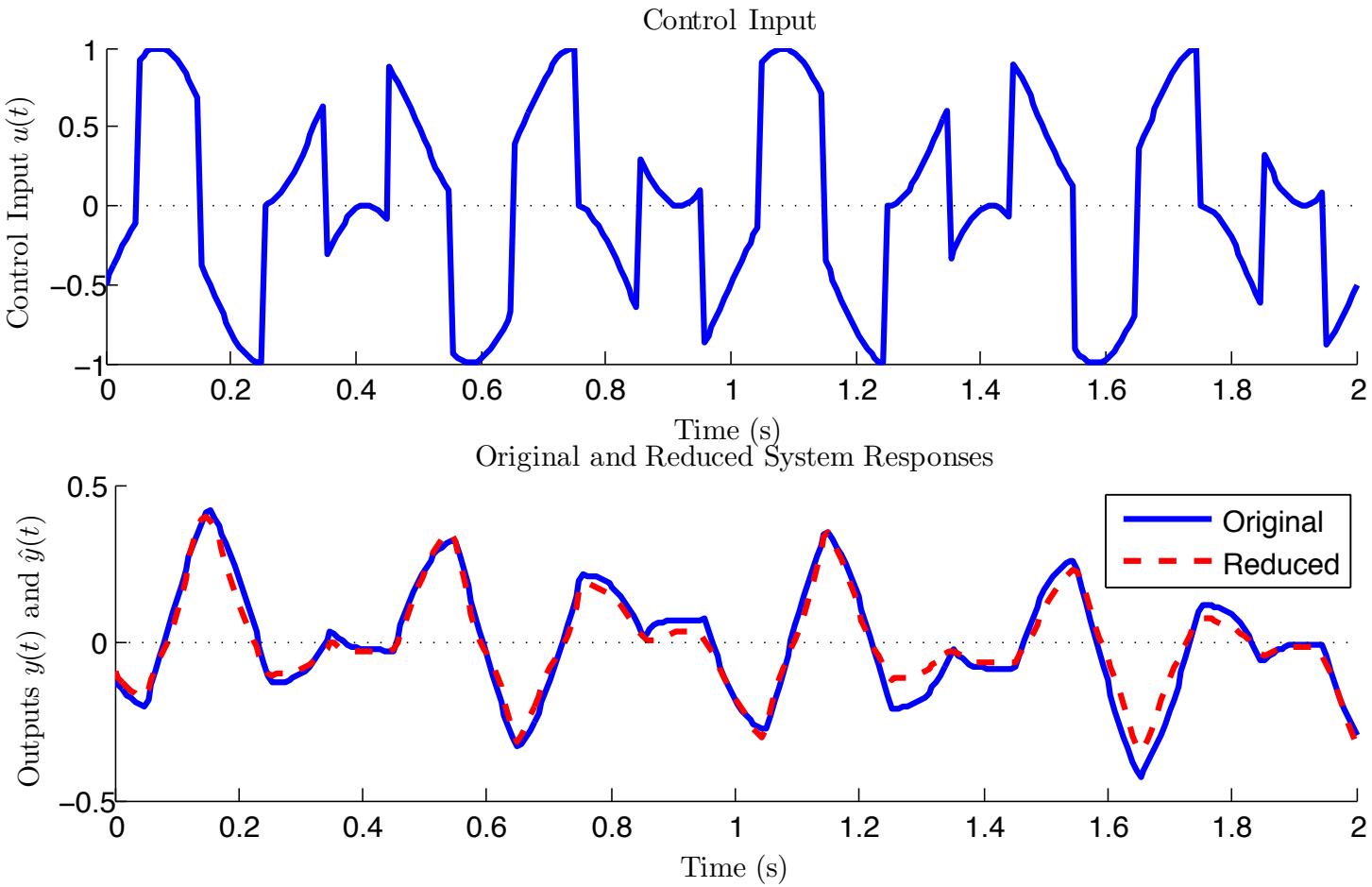
$$u(t) = \tfrac{1}{2}\big(\sin(2\pi 3t) + \mathsf{sq}(2\pi 5t - \pi/2)\big)$$

and compare the output to that of the original system.

Experiment


2d example
7d example
log(singular values of the Hankel kernel matrix
10^{4}
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
10^{-16}
10^{0}
10^{1}
10^{2}

Experiment


Control Input
Control Input $u(t)$
1
0.5
0
−0.5
−1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Time (s)
Original and Reduced System Responses
Outputs $y(t)$ and $\hat{y}(t)$
0.5
0
−0.5
Original
Reduced
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Time (s)

SDEs in Reproducing Kernel Hilbert Spaces

Review of Some Concepts for Linear Stochastic Differential Equations

- Consider the stochastically excited stable dynamical control systems affine in the input $u \in \mathbb{R}^q$

$$\dot{x} = f(x) + G(x)u \ ,$$

where $G : \mathbb{R}^n \to \mathbb{R}^{n \times q}$ is a smooth matrix-valued function. We replace the control inputs by sample paths of white Gaussian noise processes, giving the corresponding stochastic differential equation (SDE)

$$dX_t = f(X_t)dt + G(X_t)dW_t^{(q)}$$

with $W_t^{(q)}$ a $q-$dimensional Brownian motion. The solution X_t to this SDE is a Markov stochastic process with transition probability density $\rho(t, x)$ that satisfies the *Fokker-Planck (or Forward Kolmogorov) equation*

$$\frac{\partial \rho}{\partial t} = -\langle \frac{\partial}{\partial x}, f\rho \rangle + \frac{1}{2} \sum_{j,k=1}^{n} \frac{\partial^2}{\partial x_j \partial x_k} [(GG^T)_{jk}\rho] =: L\rho \ .$$

Review of Some Concepts for Linear Stochastic Differential Equations

- In the context of linear Gaussian theory where we are given an $n-$dimensional system of the form $dX_t = AX_t dt + BdW_t^{(q)}$, with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times q}$, the transition density is Gaussian.
- It is therefore sufficient to find the mean and covariance of the solution $X(t)$ in order to uniquely determine the transition probability density.

Review of Some Concepts for Linear Stochastic Differential Equations

- The mean satisfies $\frac{d}{dt}\mathbb{E}[x] = A\mathbb{E}[x]$ and thus $\mathbb{E}[x(t)] = e^{At}\mathbb{E}[x(0)]$. If A is Hurwitz, $\lim_{t\to\infty}\mathbb{E}[x(t)] = 0$.
- The covariance satisfies $\frac{d}{dt}\mathbb{E}[xx^T] = A\mathbb{E}[xx^T] + \mathbb{E}[xx^T]A + BB^T$.
- Hence, $\mathcal{Q} = \lim_{t\to\infty}\mathbb{E}[xx^\top]$ satisfies the Lyapunov system $A\mathcal{Q} + \mathcal{Q}A^\top = -BB^\top$. So, $\mathcal{Q} = W_c = \int_0^\infty e^{At}BB^\top e^{A^\top t}\,dt$, where W_c is the controllability Gramian, which is positive iff the pair (A, B) is controllable.

Review of Some Concepts for Linear Stochastic Differential Equations

- Combining the above facts, the steady-state probability density is given by

$$\rho_\infty(x) = Z^{-1} e^{-\frac{1}{2} x^\top W_c^{-1} x} = Z^{-1} e^{-L_c(x)}$$

with $Z = \sqrt{(2\pi)^n \det(W_c)}$.

Extension to the Nonlinear Case

• The preceding suggests the following key observations in the linear setting: Given an approximation $\hat{L}_c$ of L_c we obtain an approximation for ρ_∞ of the form

$$\hat{\rho}_\infty(x) \propto e^{-\hat{L}_c(x)}$$

• Although the above relationship between ρ_∞ and L_c holds for only a small class of systems (e.g. linear and some Hamiltonian systems), by mapping a nonlinear system into a suitable reproducing kernel Hilbert space we may reasonably extend this connection to a broad class of nonlinear systems.

Nonlinear SDEs in RKHSes

• Assumption1: Given a suitable choice of kernel K, if the $\mathbb{R}^d$-valued stochastic process $x(t)$ is a solution to the (ergodic) stochastically excited nonlinear system

$$dX_t = f(X_t)dt + G(X_t) \circ dW_t^{(q)}$$

the $\mathcal{H}$-valued stochastic process $(\Phi \circ x)(t) =: X(t)$ can be reasonably modelled as an Ornstein-Uhlenbeck process

$$dX(t) = AX(t)dt + \sqrt{C}dW(t), \quad X(0) = 0 \in \mathcal{H}$$

where A is linear, negative and is the infinitesimal generator of a strongly continuous semigroup e^{tA}, C is linear, continuous, positive and self-adjoint, and $W(t)$ is the cylindrical Wiener process.

Nonlinear SDEs in RKHSes

- Assumption2: The measure P_∞ is the invariant measure of the OU process and P_∞ is the pushforward along Φ of the unknown invariant measure μ_∞ on the statespace $\mathcal{X}$ we would like to approximate.
- Assumption3: The measure μ_∞ is absolutely continuous with respect to Lebesgue measure, and so admits a density.

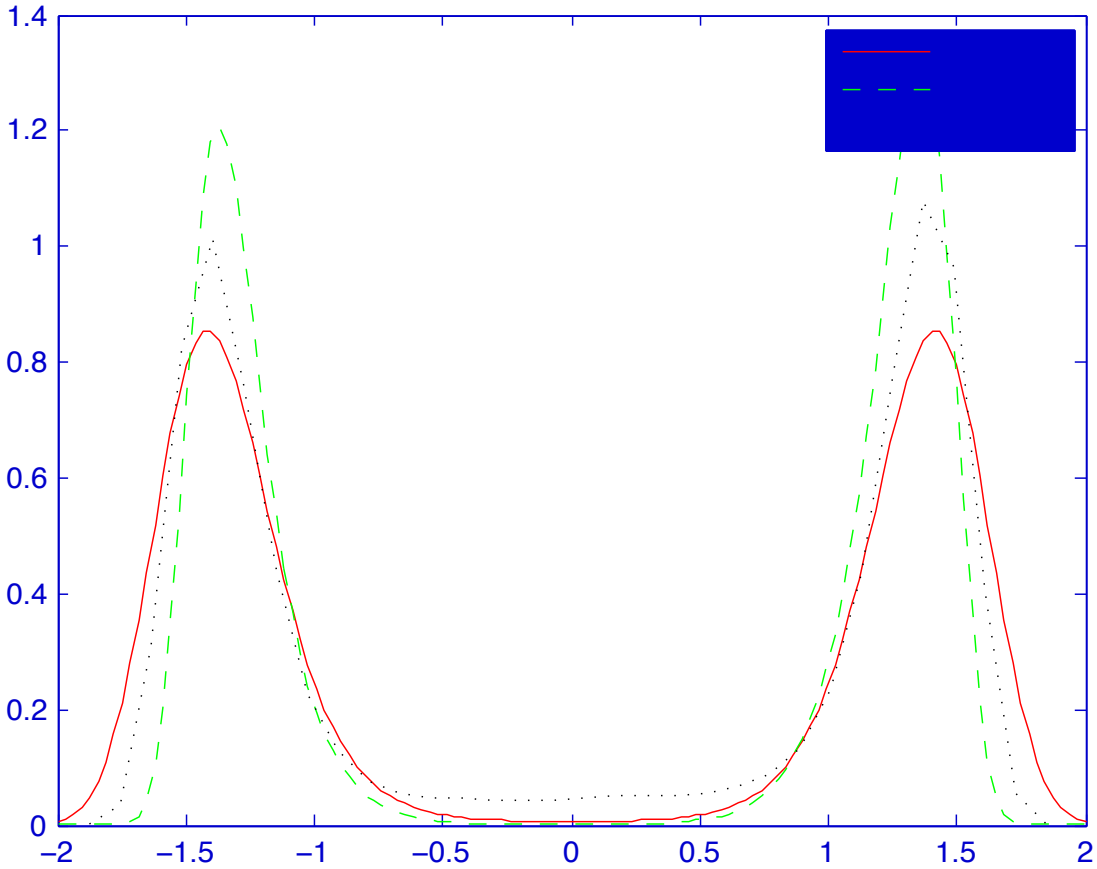
Nonlinear SDEs in RKHSes

- The stationary measure μ_∞ is defined on a finite dimensional space, so together with part (iii) of Assumption A, we may consider the corresponding density

$$\rho_\infty(x) \propto \exp\bigl(-\hat{L}_c(x)\bigr)$$

Experiment

Consider the SDE $dX = -5X^5 + 10X^3 + \sqrt{2}dW$.



Conclusions

- We used kernel flows to approximate chaotic dynamical systems.
- We used the maximum mean discrepancy to detect critical transitions.
- We introduced estimators for the controllability/observability energies of nonlinear control systems. We used these energies to perform model approximation of nonlinear control systems using a linear technique.
- We showed that the controllability energy estimator may be used to estimate the stationary solution of the Fokker-Planck equation governing nonlinear SDEs using a linear estimate.
- The estimators we derived were based on applying linear methods for control and random dynamical systems to nonlinear control systems and SDEs, once mapped into an infinite-dimensional RKHS acting as a "linearizing space".
- We introduced a data-based approach for the construction of Lyapunov functions, Center Manifold Approximation and Center Manifold Theorem.
- These results collectively argue that working in reproducing kernel Hilbert spaces offers tools for a data-based theory of nonlinear dynamical systems.

References

1. B. Hamzi, R. Maulik, H. Owhadi (2021), Data-driven geophysical forecasting: Simple, low-cost, and accurate baselines with kernel methods, https://arxiv.org/abs/2103.10935.
2. B. Hamzi and H. Owhadi (2020), Learning dynamical systems from data: a simple cross-validation perspective, https://arxiv.org/abs/2007.05074.
3. B. Haasdonk, B. Hamzi, G. Santin and D. Wittwar (2020), Kernel methods for center manifold approximation and a data-based version of the Center Manifold Theorem , https://arxiv.org/pdf/2012.00338.pdf.
4. Stefan Klus, Feliks Nüske, Boumediene Hamzi (2020), Kernel-based approximation of the Koopman generator and the Schrödinger operator, https://arxiv.org/abs/2005.13231.
5. Andreas Bittracher, Stefan Klus, Boumediene Hamzi, Peter Koltai and Christof Schütte (2020), Dimensionality Reduction of Complex Metastable Systems via Kernel Embeddings of Transition Manifolds https://arxiv.org/pdf/1904.08622.pdf.
6. B. Hamzi, C. Kuehn, S. Mohamed (2019), A Note on Kernel Methods for Multiscale Systems with Critical Transitions, Mathematical Methods in the Applied Sciences, Vol. 42, No. 3, pp. 907-917, https://arxiv.org/abs/1804.09415.
7. B. Haasdonk, B. Hamzi, G. Santin and D. Wittwar (2018), Greedy Kernel Methods for Center Manifold Approximation, Proc. of ICOSAHOM 2018, https://arxiv.org/abs/1810.11329.
8. J. Bouvrie and B. Hamzi (2017), Kernel Methods for the Approximation of Some Key Quantities of Nonlinear Systems, Journal of Computational Dynamics, vol. 4, no. 1, http://arxiv.org/abs/1204.0563.
9. J. Bouvrie and B. Hamzi (2017), Kernel Methods for the Approximation of Nonlinear Systems, SIAM J. Control & Optimization, vol. 55, no. 4, http://arxiv.org/abs/1108.2903.
10. P. Giesl, B. Hamzi, M. Rasmussen, K. Webster (2016), Approximation of Lyapunov Functions from Noisy Data, Journal of Computational Dynamics, http://arxiv.org/abs/1601.01568.
11. J. Bouvrie 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).
12. J. Bouvrie and B. Hamzi (2010), Balanced Reduction of Nonlinear Control Systems in Reproducing Kernel Hilbert Spaces, Proc.48th Annual Allerton Conference on Communication, Control, and Computing, pp. 294-301. http://arxiv.org/abs/1011.2952.