Physica D 418 (2021) 132829

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Kernel-based prediction of non-Markovian time series

Faheem Gilani [a], Dimitrios Giannakis [b], John Harlim [c,*]

[a] Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
[b] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
[c] Department of Mathematics, Department of Meteorology and Atmospheric Science, Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 8 July 2020
Received in revised form 8 December 2020
Accepted 14 December 2020
Available online 2 January 2021
Communicated by V.M. Perez-Garcia

Keywords:
Kernel analog forecast
Delay-embedding
Mori–Zwanzig formalism
Nyström method
Markovian kernel smoothing
Nonparametric smoother

ABSTRACT

A nonparametric method to predict non-Markovian time series of partially observed dynamics is developed. The prediction problem we consider is a supervised learning task of finding a regression function that takes a delay-embedded observable to the observable at a future time. When delay-embedding theory is applicable, the proposed regression function is a consistent estimator of the flow map induced by the delay-embedding. Furthermore, the corresponding Mori–Zwanzig equation governing the evolution of the observable simplifies to only a Markovian term, represented by the regression function. We realize this supervised learning task with a class of kernel-based linear estimators, the kernel analog forecast (KAF), which are consistent in the limit of large data. In a scenario with a high-dimensional covariate space, we employ a Markovian kernel smoothing method which is computationally cheaper than the Nyström projection method for realizing KAF. In addition to the guaranteed theoretical convergence, we numerically demonstrate the effectiveness of this approach on higher-dimensional problems where the relevant kernel features are difficult to capture with the Nyström method. Given noisy training data, we propose a nonparametric smoother as a de-noising method. Numerically, we show that the proposed smoother is more accurate than EnKF and 4Dvar in de-noising signals corrupted by independent (but not necessarily identically distributed) noise, even if the smoother is constructed using a data set corrupted by white noise. We show skillful prediction using the KAF constructed from the denoised data.

1. Introduction

A long-standing issue in the applications of dynamical systems is to predict time series of observables given partial observations. This problem has classically been studied from various angles under different names in the literature (i.e. reduced-order modeling, closure modeling, subgrid parameterization, etc.), but more recently it has also been viewed as a machine learning problem. In particular, at the core of this modeling problem is a supervised learning task to find a map that takes appropriate covariate data (an observable in the past and/or present times) to the desired response function (an observable at the future times). When the covariate data is a delay-embedded observable, the target map provides a non-Markovian prediction model. The realizations of this problem with state-of-art machine learning algorithms involving deep/recurrent neural networks have reported superb numerical performances even when the underlying dynamics are highly nonlinear and high-dimensional [1–4]. In the context of partially known dynamics, the recent work in [4] formulated the target function as a conditional expectation associated with an appropriate probability space and showed that the corresponding supervised learning framework (which is similar to the one proposed in [3,5]) produces an approximate closure model whose solutions converge (strongly) to those of the underlying dynamics for finite time when both models are initialized with the same initial conditions. Building on this positive result, one of the goals of this paper is to understand the regression problem corresponding to the supervised learning task from the viewpoint of dynamical systems theory and reduced-order modeling.

Due to the classical theory of dynamical systems, this modeling framework is closely related to the delay-embedding theorem [6] which has served as a foundation for attractor reconstruction from time series. We will argue that when the embedding theorem is satisfied, the regression (or target) function is theoretically consistent with the component of the flow induced by the delay-coordinate map. From the reduced-order modeling viewpoint, the same learning task can be formulated as a problem of deriving, from first principles, a set of effective equations that determines the evolution of the observable (e.g., [7–11]).

* Corresponding author.
E-mail addresses: fhg3@psu.edu (F. Gilani), dimitris@cims.nyu.edu (D. Giannakis), jharlim@psu.edu (J. Harlim).

https://doi.org/10.1016/j.physd.2020.132829
0167-2789/

The Mori–Zwanzig (MZ) formalism [12,13] has been proposed by this community as a natural framework for deriving such a set of effective equations for forecasting the time series of partially observed dynamical systems. The appeal of using the MZ formalism is that the resulting system is represented by an equation that involves projected linear evolution operators. In contrast to geometrical state-space approaches, the operator-theoretic approach focuses on the induced linear action of dynamical systems on appropriately chosen spaces of observables despite the nonlinearity of the flow map. In the context of the MZ formalism, this allows one to compartmentalize the contribution of the observable at the present time (the Markovian term), the observable in the past (the memory/non-Markovian term), and the orthogonal dynamics of the trajectory of the observables at the future times with a collection of linear operators.

While such a representation is attractive for understanding the modeling mechanism, it may not be easily translated into an efficient numerical method. This issue arises due to the fact that the MZ formula states the dependence of the observable at the future time on the entire history of observables and the initial condition. Besides, it is usually difficult to specify the memory kernel as it requires the solution of the high-dimensional orthogonal dynamics [7,14]. Ultimately, the desired computational objective is to have a finite memory approximation. This issue has given rise to many parametric approximations of the memory kernel, such as the delta function approximation [15], Krylov subspace approximation [16], series expansion [17,18], and rational approximation [19], just to name a few. While these approaches have shown positive results when addressing specific applications, they either require the knowledge of the full model and/or they are subjected to modeling error when the memory kernel is not adequately represented by the specified parametric model. We will argue that if the hypotheses of the delay-embedding theorems are satisfied, the representation of the MZ equation with the projection operator obtained through the corresponding regression framework can be simplified to a computationally tractable model. In particular, the MZ equation consists of only the "Markovian" term associated with the delay-embedded sequences, which is exactly the regression function given by the supervised learning framework. The connection between supervised learning, delay-embedding theory, and the MZ formalism suggests that the regression framework is indeed a natural approach for predicting time series of partially observed dynamics.

We should point out that this connection partially explains the empirical successes reported in [1–4] since they all adopted this regression modeling paradigm. One unexplained component of these empirical successes is the consistency of their estimators. In these papers, the authors approximated the target function using a neural network model (which is in the form of a composition of activation functions) which depends nonlinearly on possibly a very large number of parameters (depending on the depth and width of the neural network architecture). Thus, the training phase often involves a nonlinear, highly non-convex, optimization problem, and finding the global optimizer for such a problem can be a difficult task given that most solvers convergence is guaranteed locally. While this is an interesting direction, we will not explore it here. In this paper, we study a class of linear estimators that can be translated into computational algorithms with theoretical guarantees. In particular, we consider the kernel analog forecast (KAF) which has found applications in finance [20] and climate sciences [21–24]. KAF is a kernel regression method designed for the purpose of predicting time series generated by an observable of a dynamical system. The term "analog" refers to the fact that KAF is a generalization of the classical analog forecasting method proposed by Lorenz [25], for which the prediction is determined based on the affinity of the present states and the historical analog. In this context, the so-called "kernel trick" allows one to identify the analogs (feature space) with an appropriately chosen kernel. This, in turn, allows one to access an estimator that lies in a Reproducing Kernel Hilbert Space (RKHS) induced by the associated kernel features, with universal approximation properties. A key advantage of the RKHS formulation is that properties of the elements of the space are inherited by corresponding properties of the kernel. In particular, if the kernel is bounded, then functions in the RKHS are also bounded. Likewise, functions in an RKHS inherit the regularity of the kernel. This important property allows one to establish uniform convergence of the estimator, which justifies the use of KAF as an interpolator. In the context of dynamical systems forecasting, the natural function space (e.g., an L^2 space associated with an invariant measure) is usually not known explicitly, yet relationships between kernel integral operators and RKHSs allow one to empirically access the subspace of L^2 through a set of orthogonal basis functions corresponding to ordered eigenvalues. In this case, there is a natural mapping of the L^2 basis vectors corresponding to nonzero eigenvalues to orthogonal RKHS functions, and, under appropriate positivity conditions on the kernel, the latter span a dense subspace of the corresponding L^2 space. With orthogonality at hand, one can control the accuracy of the estimate by a finite eigenbasis representation and, simultaneously, avoid the large matrix inversion problem with the radial-type kernels. Finally, the RKHS structure allows one to evaluate the estimator on new data points using a classical interpolator, the Nyström projection method. It should be noted that this construction does not require that the covariate time series is Markovian, and is therefore well suited to forecasting under partial observations; e.g., see [26] for applications of KAF to prediction of slow components of multiscale systems exhibiting averaging or homogenization.

While KAF is theoretically sound [27], it may face practical limitations, especially when both the covariate space and the support of the pushforward of the invariant measure on the covariate space are high dimensional. This issue is mainly due to lack of guarantees that the leading eigenfunctions induced by a generic kernel on a high-dimensional covariate space adequately capture the response (predictand) variable of interest. To alleviate this limitation, while also reducing computational complexity, we propose to realize KAF with a *kernel smoothing* technique, whose basic idea is to apply a discrete convolution of a Markov operator on the response functions. We show that the proposed kernel smoothing method is a consistent estimator of the optimal regression function, i.e., the conditional expectation of the response given the covariate data. Using the variable-bandwidth kernels introduced in [28], we numerically demonstrate the effectiveness of kernel smoothing compared to the Nyström method in estimating the full discrete MZ equation in situations where the covariate space is relatively high-dimensional. On the other hand, when the covariate space is low dimensional, the Nyström method is generally a better choice since the response variable is more likely to be well represented by the leading empirical kernel features.

Another critical issue that often arises in practical applications is that the available observables are subjected to noises (of possibly unknown nature). This poses a question in the accuracy of the KAF estimators since the noises in the response and covariate data may yield an ill-posed regression problem. In this paper, we propose a *non-parametric smoother*, constructed using the Nyström projection method, to denoise observables corrupted by independent (but not necessarily identically distributed) noises. In our applications, we will show the effectiveness of the proposed smoother in denoising signals corrupted by various noise types,

including time varying noise, even if the smoother is constructed using a data set corrupted by independent and identically distributed (i.i.d.) Gaussian noise. From our numerical tests, we will find that the proposed smoother produces more accurate estimates than two popular data assimilation methods that are presently used in operational weather forecasts: the Ensemble Kalman filter [29] and the 4D-Variational approach [30], both of which require the true governing equations of the observed components. Using the smoothed data, we numerically verify that the kernel smoothing method is effective in predicting the response variable. We will show that this blended "projection-smoothing" approach is able to produce a reasonably accurate prediction from purely noisy observables.

This paper is organized as follows. In Section 2, we review the kernel-based regression framework for supervised learning tasks. In Section 2.1, we discuss the Nyström projection method. While the presentation follows closely that in [27], in the current discussion, we do not present the regression problem for time series generated by ergodic dynamical systems and only describe it on i.i.d. training data. We complete the discussion in Section 2.1 with a simple statistical error bound. In Section 2.2, we present the kernel smoothing method, and prove its consistency and associated error bounds using variable bandwidth kernels [28]. In Section 3, we discuss the problem of predicting observables of time series generated by dynamical systems. Since the only available training data is the time series of the relevant observables, we briefly review the discrete MZ formalism for reduced-order modeling in Section 3.1. In Section 3.2, we focus on estimating the solution operator of the projected discrete MZ equation with the KAF estimator. We demonstrate the performance of the estimator on a Hamiltonian system and the five-dimensional chaotic Lorenz-96 dynamical system. In Section 3.3, we discuss the connection of the proposed nonparametric regression framework with the delay-embedding and MZ formalism. In particular, we will show that if the hypothesis in the delay-embedding theory is satisfied, the regression function is indeed a component of the flow map. Furthermore, the MZ equation derived using the projection operator obtained by the regression framework consists of only the "Markovian" term and it is exactly represented by the corresponding regression function. Supporting numerical examples on the two same dynamical systems are given. In Section 4, we consider data corrupted by independently distributed noises. A non-parametric smoother based on the Nyström projection method is presented as a denoising method in Section 4.1. Subsequently, in Section 4.2, we numerically verify the prediction skill of the KAF estimator when it is trained using the smoothed data. In Section 5, we close this paper with a summary and outlook of open problems.

2. Nonparametric regression

Given spaces $\mathcal{X}$ and $\mathcal{Y}$, a basic problem of supervised learning is to construct a map $F : \mathcal{X} \to \mathcal{Y}$ from samples of labeled data, $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1,\dots,N}$, such that $F(x_i)$ optimally approximates y_i in a suitable sense. Here, we require that $\mathcal{Y}$ be a Hilbert space so that we can apply orthogonal projections, as well as compute expectations and other statistical functionals. On the other hand, we allow $\mathcal{X}$ to be nonlinear. In order for the target function F to be predictive, we relate x_i and y_i by assuming that they are realizations of random variables X and Y with common domain Ω. We assume that Ω is a probability space equipped with a σ-algebra $\mathcal{B}(\Omega)$ and probability measure μ. We call $\mathcal{X}$ the covariate space and $\mathcal{Y}$ the response space. The corresponding maps X and Y are called the covariate map and the response map, respectively.

Consider the Hilbert spaces $H = \left\{f : \Omega \to \mathcal{Y} \mid \int_\Omega f^2(\omega)\, d\mu(\omega) < \infty\right\}$, $V = \{g : \mathcal{X} \to \mathcal{Y} : g \circ X \in H\}$, and $H_X = \{f \in H : f = g \circ X \text{ for some } g \in V\}$. Note that $H = L^2(\mu)$ and $V = L^2(\nu)$ where $\nu = X_*\mu$ is the pushforward of μ via X. Moreover, H_X is the Hilbert subspace of $L^2(\mu)$ that contains equivalence classes of square-integrable functions which are measurable with respect to the σ-algebra generated by X. In general, there are many ways to construct a predictive map F. The least-squares approach is to construct an F that minimizes the mean square error. A standard result from statistics is that this estimator is given by the regression function, which is also known as the conditional expectation function. That is,

$$\mathbb{E}[Y|\cdot] = F \coloneqq \arg\min_{g \in V} \|Y - g \circ X\|_H^2, \tag{1}$$

where the conditional expectation $\mathbb{E}[\cdot|X]$ can be seen as an orthogonal projection of H onto H_X. In this paper, we will denote the orthogonal projection $P : H \to H_X$ as the conditional expectation $\mathbb{E}[\cdot|X]$. In appropriate context, we will also use $P : H \to S_X \subseteq H_X$, to denote an arbitrary orthogonal projection onto its range space, $S_X = \mathrm{ran}(P)$ such that $S_X^\perp = \mathrm{null}(P)$ and $H = S_X \oplus S_X^\perp$, where the orthogonality is defined with respect to the inner product of H.

When X is not injective, as in many applications, one cannot approximate the response $Y \in H$ to arbitrary precision by elements of H_X. However, one can still construct an optimal estimator of Y using the target function $F \in V$. In the remainder of this section, we discuss two methods for estimating $\mathbb{E}[Y|\cdot]$ from samples of labeled data $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \}_{i=1,\dots,N}$. The first one is the Nyström method which is an interpolation of an eigenbasis representation of the estimator. The second method is the kernel smoothing that employs a convolution operation associated with a Markov kernel. For the remainder of this section, we restrict our discussion to real-valued functions, so that $\mathcal{Y} = \mathbb{R}$ and $H = \{f : \Omega \to \mathbb{R} \mid \int_\Omega f^2(\omega)\, d\mu(\omega) < \infty\}$. Since our applications involve $\mathcal{Y} = \mathbb{R}^n$, a componentwise generalization to the finite-dimensional vector-valued case is immediate.

2.1. Nyström method

If V is equipped with an orthonormal basis $\{u_j\}_{j\in\mathbb{N}}$ and X is injective, then $\{\phi_j = u_j \circ X\}_{j\in\mathbb{N}}$ forms an orthonormal basis of H. In this case, any $Y \in H$ can be arbitrarily estimated, in H-norm, by

$$\mathbb{E}_L[Y|X] := \sum_{j=0}^{L} \langle Y, \phi_j\rangle_H \phi_j, \tag{2}$$

up to any desirable precision by taking $L \to \infty$. Due to the properties of orthogonal projection, the estimator

$$\mathbb{E}_L[Y|\cdot] = \sum_{j=0}^{L} \langle Y, \phi_j\rangle_H u_j, \tag{3}$$

is an optimal estimator from $\mathrm{span}\{\phi_0, \dots, \phi_L\} \subset H$. As mentioned above, when X is not injective, $\mathrm{span}\{\phi_j\}_{j\in\mathbb{N}} \subsetneq H$ so one cannot recover arbitrary target functions $Y \in H$. However, $\mathbb{E}_L[Y|\cdot]$ is a consistent estimator of $\mathbb{E}[Y|\cdot] \in V$ so that $\lim_{L\to\infty} \mathbb{E}_L[Y|\cdot] = \mathbb{E}[Y|\cdot]$ in V.

A practical issue in employing the estimator (3) is that orthonormal bases of H as well as V are not available. The whole point of nonparametric regression is to construct an estimator for $\{\phi_0, \phi_1, \dots\}$ from the random samples of observables $\{x_i : i = 1, \dots, N\}$, where $x_i = X(\omega_i)$ are realizations of the covariate map X. Kernel-based algorithms [28,31] are often used to obtain the function value $u_j(x_i) = u_j \circ X(\omega_i) = \phi_j(\omega_i)$, which can subsequently be used to estimate the inner product in (3). For our purposes, we

also need to evaluate the estimator in (3) on new covariate data that do not lie in the (finite) training data set. This evaluation can be done using an interpolation scheme such as the Nyström method that extends u_j on new covariate data disjoint from the finite sample of observations. To justify the validity of such an interpolation method, uniform convergence of the estimator is usually required rather than V-norm convergence.

One way to ensure uniform convergence is to construct an estimator in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of continuous functions such that $\mathcal{H}$ is dense in H_X. In particular, let $k : \Omega \times \Omega \to \mathbb{R}$ be the pullback of a kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on the covariate space. That is, k is symmetric positive definite and $k(\omega, \omega') = \kappa(X(\omega), X(\omega'))$. By the Moore–Aronszajn theorem, there exists a unique Hilbert space $\mathcal{H}$ (the RKHS), of real valued functions $f : \Omega \to \mathbb{R}$ with the reproducing property: $\mathcal{H} = \overline{\text{span}}\{k(\omega, \cdot), \forall \omega \in \Omega\}$ and every $f \in \mathcal{H}$ and $\omega \in \Omega$ satisfies $f(\omega) = \langle k(\omega, \cdot), f\rangle_{\mathcal{H}}$. Since the kernel k is a pullback kernel of κ, every function $f \in \mathcal{H}$ can be expressed as $f = g \circ X$ for some continuous function $g : \mathcal{X} \to \mathbb{R}$. If Ω is compact and k is continuous, one can show that $\mathcal{H}$-norm convergence implies uniform convergence so that $\mathcal{H} \subset C(\Omega)$. For non-compact domains, a bounded kernel ensures that $\mathcal{H} \subset C_b(\Omega)$ [32].

While it is convenient to represent functions in $\mathcal{H}$ as a linear superposition of kernel sections, namely, $f = \sum_{i=1}^{\infty} a_i k(\omega_i, \cdot)$ with $\omega_i \in \Omega$, empirical representations involve a partial summation of N terms, where N denotes the number of training samples. For large datasets, as in our applications, specification of the coefficients a_i involves an inversion of a large matrix and the repetitive function evaluation is numerically expensive. If a radial-type kernel is chosen, as in many applications, then we arrive at the so-called kernel ridge regression or radial basis function interpolation, depending on the literature. The estimator in (3) is proposed as an alternative to avoid this computational issue by leveraging the inner product structure of H. To that end, consider the reproducing kernel k from the perspective of an integral operator $K_\mu : H \to \mathcal{H}$ defined as

$$K_\mu f = \int_\Omega k(\cdot, \omega) f(\omega) d\mu(\omega), \tag{4}$$

where μ is assumed to be compactly supported on $M \subset \Omega$. This is a compact operator with adjoint $K_\mu^* : \mathcal{H} \to H$ that is also compact. By the spectral theorem, the compact, self-adjoint and positive-definite integral operator $G_\mu := K_\mu^* K_\mu : H \to H$ has eigenvalues $\lambda_0 \geq \lambda_1 \geq \cdots \searrow 0^+$ so that the corresponding eigenfunctions $\{\phi_0, \phi_1, \ldots\}$ form an orthonormal basis of H. In fact, defining, $\psi_j = K_\mu \phi_j / \lambda_j^{1/2}$ for $\lambda_j > 0$, we have,

$$\begin{aligned}\langle \psi_i, \psi_j\rangle_{\mathcal{H}} &= \frac{1}{\lambda_i^{1/2}\lambda_j^{1/2}} \langle K_\mu \phi_i, K_\mu \phi_j\rangle_{\mathcal{H}} = \frac{1}{\lambda_i^{1/2}\lambda_j^{1/2}} \langle K_\mu^* K_\mu \phi_i, \phi_j\rangle_H \\ &= \frac{\lambda_i^{1/2}}{\lambda_j^{1/2}} \langle \phi_i, \phi_j\rangle_H = \delta_{ij},\end{aligned}$$

which means that $\{\psi_0, \psi_1, \ldots\}$ is an orthonormal set in $\mathcal{H}$. By Mercer's theorem, we have an explicit representation $k(\omega, \omega') = \sum_{j=0}^{\infty} \lambda_j \varphi_j(\omega)\varphi_j(\omega') = \sum_{j=0}^{\infty} \psi_j(\omega)\psi_j(\omega')$, converging uniformly for $(\omega, \omega') \in M \times M$, where $\varphi_j = \lambda_j^{-1/2}\psi_j$ denotes the continuous representative of eigenfunction ϕ_j. The so-called "kernel trick" specifies an explicit choice of kernel k, such as the Gaussian kernel, to avoid computing the ℓ_2 inner-product between feature vectors $(\psi_0(\omega), \psi_1(\omega), \ldots)$ and $(\psi_0(\omega'), \psi_1(\omega'), \ldots)$. Our perspective is to rely on the orthogonality of the eigenbasis to approximate the target function of interest through the representation in (3) and use the RKHS theory to establish the convergence of the estimator as $L \to \infty$.

One of the most important aspects of the integral operator K_μ is that we can define an interpolation (Nyström) operator $\mathcal{N}_\mu : D(\mathcal{N}_\mu) \to \mathcal{H}$ as $\mathcal{N}_\mu \phi_j := \psi_j / \lambda_j^{1/2} = K_\mu \phi_j / \lambda_j := \varphi_j$, whose domain $D(\mathcal{N}_\mu) = \{f = \sum c_k \phi_k \in H | \sum_k c_k^2/\lambda_k < \infty\}$ contains functions of higher regularity than arbitrary elements of H. Note that if $D(\mathcal{N}_\mu)$ is equipped with the norm $\|f\|^2 = \sum_k c_k^2/\lambda_k$, then it is isometrically isomorphic to $\mathcal{H}(M)$, the restriction of $\mathcal{H}$ to the support M. Notice that the operator $\mathcal{N}_\mu$ maps the eigenfunction $\phi_j \in D(\mathcal{N}_\mu)$ to the continuous function φ_j. As a result, $f \in D(\mathcal{N}_\mu)$ has a continuous representation $\mathcal{N}_\mu f = \sum_j c_j \mathcal{N}_\mu \phi_j = \sum_j c_j \varphi_j$. Moreover, the map K_μ^* is a left inverse of $\mathcal{N}_\mu$ since

$$K_\mu^* \mathcal{N}_\mu f = \sum_j c_j K_\mu^* \varphi_j = \sum_j c_j K_\mu^* K_\mu \frac{\phi_j}{\lambda_j} = \sum_j c_j \phi_j = f. \tag{5}$$

This means that the map $K_\mu^* \mathcal{N}_\mu : D(\mathcal{N}_\mu) \to \overline{\text{ran} K_\mu}$ identifies functions in $D(\mathcal{N}_\mu)$ with their continuous representation in $\mathcal{H}$ through the Nyström operator, as a function in $\overline{\text{ran} K_\mu}$.

In our case, the target function is $\mathbb{E}[Y|\cdot] \in V$ or $\mathbb{E}[Y|X] \in H_X$. Thus we can consider the operator (4) but with domain H_X. In this case, an orthonormal set of continuous functions $\{\psi_0, \ldots, \psi_L\}$ in $\mathcal{H}$ satisfies $\psi_j = u_j \circ X$ for some continuous functions $\{u_0, \ldots, u_L\}$ that can be approximated from the covariate data. Using this basis, for each $\mathbb{E}_L[Y|X] \in D(\mathcal{N}_\mu)$, one can build an estimator for $\mathcal{N}_\mu \mathbb{E}_L[Y|X] \in \mathcal{H}$ which can be represented as

$$\mathcal{N}_\mu \mathbb{E}_L[Y|\cdot] = \sum_{j=0}^{L} \langle Y, \phi_j\rangle_H \frac{u_j}{\lambda_j^{1/2}}. \tag{6}$$

It is important to note that if the reproducing kernel k of the RKHS $\mathcal{H}$ is a pullback of a strictly positive definite kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then the domain $D(\mathcal{N}_\mu)$ is dense in H_X. To see this, take any function $f \in H_X$ and, since $\text{span}\{\phi_0, \phi_1, \ldots\}$ is dense in H_X, we have that $f = \sum_k c_k \phi_k$, where each eigenfunction is associated with a strictly positive eigenvalue. Furthermore,

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{{c_k}^2}{\lambda_k} &= \sum_{k=0}^{\infty} \frac{{c_k}^2}{\lambda_k} \langle \phi_k, \phi_k\rangle_{H_X} = \sum_{k=0}^{\infty} \frac{{c_k}^2}{\lambda_k} \langle K_\mu \phi_k, K_\mu \phi_k\rangle_{\mathcal{H}} \\ &= \sum_{k=0}^{\infty} {c_k}^2 \langle \psi_k, \psi_k\rangle_{\mathcal{H}} = \sum_{k=0}^{\infty} {c_k}^2 = \|f\|_{H_X}^2 < \infty,\end{aligned}$$

and we conclude that any function $f \in H_X$ can be approximated by a function in $D(\mathcal{N}_\mu)$ at arbitrary precision. From (5), one can see that the operator $K_\mu^* \mathcal{N}_\mu : D(\mathcal{N}_\mu) \to D(\mathcal{N}_\mu)$ is an identity map (a bounded operator). By the bounded linear transformation theorem, the closed extension of $K_\mu^* \mathcal{N}_\mu$ is the identity map on $\overline{D(\mathcal{N}_\mu)} = H_X$. This means that any function in H_X can be approximated to arbitrary precision in H-norm by a function in $K_\mu^* \mathcal{H} = D(\mathcal{N}_\mu)$. In particular, as $L \to \infty$, the estimator in (6) converges to the target function in H-norm, i.e., $\lim_{L\to\infty} K_\mu^* \mathcal{N}_\mu \mathbb{E}_L[Y|X] = \mathbb{E}[Y|X]$. If it now happens that $\mathbb{E}[Y|X]$ has a representative in $\mathcal{H}$, then the estimator converges to that representative in $\mathcal{H}$-norm, and thus uniformly, on the support of μ.

As mentioned above, in practice, we have no access to the basis functions $\{\phi_0, \ldots, \phi_L\}$ or $\{u_0, \ldots, u_L\}$. Given the pairs of labeled data points $\{(x_i, y_i)\}_{i=1,\ldots,N}$, where x_i are i.i.d. samples of X, we first describe an empirical estimate of $\phi_j(\omega_i) = u_j(x_i)$. Let $G_{\mu_N} := K_{\mu_N}^* K_{\mu_N}$, where $K_{\mu_N} : H_N \to \mathcal{H}$ and $K_{\mu_N}^* : \mathcal{H} \to H_N$ are defined as in (4) and the corresponding adjoint with $H = L^2(\mu_N)$ replaced by $H_N := L^2(\mu_N)$. Here, $\mu_N = \sum_{j=1}^{N} \delta_{\omega_j}/N$ is the discrete sampling measure, and $L^2(\mu_N)$ the corresponding finite-dimensional Hilbert space equipped with the inner product $\langle f, g\rangle_{H_N} = \frac{1}{N}\sum_{i=1}^{N} f(\omega_i)g(\omega_i)$. For simplicity of exposition, we will assume that all sampled states ω_i are distinct, so H_N is an N-dimensional Hilbert space, isomorphic to $\mathbb{R}^N$ equipped with a normalized dot product. In that case, the operator G_{μ_N} is represented by an $N \times N$ kernel matrix $\mathbf{G}_N = [\langle e_{i,N}, G_{\mu_N} e_{j,N}\rangle_{H_N}]$

$= [\kappa(x_i, x_j)]$, where $e_{j,N}$ are the standard orthonormal basis vectors of H_N with $e_{j,N}(\omega_i) = N^{1/2}\delta_{ij}$.

Let $\{\lambda_{j,N}, \phi_{j,N}\}$ be the jth eigenvalue and eigenvectors of $\mathbf{G}_N$, respectively. It is well known that the approximation of G by G_n is spectrally consistent [33]; that is the sequence of eigenvalues $\lambda_{j,N} \to \lambda_j$, as $N \to \infty$. Moreover, the continuous representative, $\mathcal{N}_{\mu_N}\phi_{j,N} = \psi_{j,N}/\lambda_{j,N}^{1/2}$ converges to $\mathcal{N}_\mu\phi_j = \psi_j/\lambda_j^{1/2}$ as $N \to \infty$ in $\mathcal{H}$. Denoting $\vec{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$, we have

$$\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)} = \frac{1}{N}\sum_{i=1}^{N} y_i\phi_{j,N}(\omega_i) = \int_\Omega Y(\omega)\mathcal{N}_{\mu_N}\phi_{j,N}(\omega)d\mu_N(\omega)$$

$$\longrightarrow \int_\Omega Y(\omega)\mathcal{N}_\mu\phi_j(\omega)d\mu(\omega) = \langle Y, \phi_j\rangle_H,$$

as $N \to \infty$, where we have used the law of large numbers for i.i.d. samples. For each j,

$$\begin{aligned}\mathbb{E}_\mu[\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)}] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\mu[Y\mathcal{N}_{\mu_N}\phi_{j,N}] \\ &= \mathbb{E}_\mu[Y\mathcal{N}_{\mu_N}\phi_{j,N}] = \mathbb{E}_\mu[Y\phi_j] + \mathcal{O}(\delta), \end{aligned} \tag{7}$$

where δ is an error bound of the eigenfunction estimation. In the proposition below, we will specify δ on a manifold without boundary based on the L^2 result from [34]. The standard Monte-Carlo error suggests that

$$\begin{aligned}&\mathbb{E}_\mu\left[\left(\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)} - \mathbb{E}_\mu[Y\mathcal{N}_{\mu_N}\phi_{j,N}]\right)^2\right] \\ &= \frac{1}{N}\mathbb{E}_\mu\left[(Y\mathcal{N}_{\mu_N}\phi_{j,N} - \mathbb{E}_\mu[Y\mathcal{N}_{\mu_N}\phi_{j,N}])^2\right] = \frac{\mathrm{Var}[Y\mathcal{N}_{\mu_N}\phi_{j,N}]}{N}.\end{aligned}$$

Without loss of generality, suppose that $\mathbb{E}_\mu[Y] = \mathbb{E}_\mu[\phi_j] = 0$. If Y is continuous on $M \subset \Omega$, the compact support of μ, then

$$\begin{aligned}\mathrm{Var}[Y\mathcal{N}_{\mu_N}\phi_{j,N}] &= \mathbb{E}_\mu[Y^2(\mathcal{N}_{\mu_N}\phi_{j,N})^2] \le \|Y^2\|_\infty\mathbb{E}_\mu[(\mathcal{N}_{\mu_N}\phi_{j,N})^2] \\ &\le \|Y^2\|_\infty(\mathbb{E}_\mu[\phi_j^2] + \mathcal{O}(\delta^2)) = \|Y^2\|_\infty(1 + \mathcal{O}(\delta^2)),\end{aligned}$$

where we have used the Hölder inequality and the orthonormality of ϕ_j. Together with (7), we have

$$\mathbb{E}_\mu\left[\left(\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)} - \mathbb{E}_\mu[Y\phi_j]\right)^2\right] \le C\left(\frac{1}{N} + \frac{\delta^2}{N} + \delta^2\right), \tag{8}$$

for some constant $C > 0$.

We should point out that if the samples $\{\omega_i\}$ form a time series generated by an ergodic and stationary dynamical system, then the convergence can still be achieved via the Birkhoff ergodic theorem, but the convergence rate would depend on the mixing rate of the underlying processes [35,36]. Together with the convergence of the continuous representative, we can conclude that the discrete estimator

$$\mathbb{E}_{L,N}[Y|X] := \sum_{j=0}^{L}\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)}\phi_{j,N},$$

has a continuous representative

$$\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|X] = \sum_{j=0}^{L}\langle \vec{y}, \phi_{j,N}\rangle_{L^2(\mu_N)}\psi_{j,N}/\lambda_{j,N} \tag{9}$$

that converges in $\mathcal{H}$-norm to $\mathcal{N}_\mu\mathbb{E}_L[Y|X]$ as $N \to \infty$. Also, the left pseudo-inverse, $K_\mu^*\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|X] \to K_\mu^*\mathcal{N}_\mu\mathbb{E}_L[Y|X] = \mathbb{E}_L[Y|X]$ as $N \to \infty$ in H_X. Taking $L \to \infty$ after $N \to \infty$, we establish the consistency of the estimator with the target function, $\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|X] \to \mathbb{E}[Y|X] \in H_X$.

Let $\nu_N = \mu_N \circ X^{-1}$ be the pushforward of the sampling measure on covariate space $\mathcal{X}$. Computationally, we can estimate the discrete orthonormal basis $\{u_{0,N}, u_{1,N}, \ldots, u_{L,N}\}$ with respect to $L^2(\nu_N)$ by solving an eigenvalue problem associated with a Markov operator $G_{\mu_N,\epsilon}$ constructed using a decreasing kernel k_ϵ defined with bandwidth parameter ϵ (see also remark 3 of [34]). Note that the pullback is given as $\phi_{j,N,\epsilon} = u_{j,N,\epsilon} \circ X$. If $\mathcal{X}$ is a d-dimensional compact smooth manifold embedded in $\mathbb{R}^n$, then $\mathcal{N}_{\mu_N}u_{j,N,\epsilon}$ converges to the eigenfunctions u_j of the Laplace–Beltrami operator (positive definite with respect to V) as $N \to \infty$ and $\epsilon \to 0$. If ν_N has a smooth density with respect to the volume form, then the Laplace–Beltrami is defined with a conformally changed Riemannian metric inherited by $\mathcal{X}$ from the ambient space $\mathbb{R}^n$. In this case, we have:

Proposition 2.1. *Let $\mathcal{X}$ be a d-dimensional compact smooth Riemannian manifold with no boundary. Let $Y := F \circ X$ such that $F : \mathcal{X} \to \mathcal{Y}$ belongs to a Sobolev class, $H^\beta(\mathcal{X}) := \{F \in V | \hat{F}_j := \langle F, u_j\rangle_V, \sum_j \zeta_j^\beta\hat{F}_j^2 < \infty, \beta > 0\}$, where ζ_j is the eigenvalue of the Laplace–Beltrami operator associated with eigenfunction u_j, approximated with $u_{j,N,\epsilon}$ as discussed in the preceding paragraph. Assume also that $Y \in C(M)$, where $M \subset \Omega$ denotes the compact support of the invariant measure μ. Then, with $\mu_N = \sum_{j=1}^N \delta_{\omega_j}/N$ and $\mathcal{N}_\mu\mathbb{E}_{L,N}[Y|\cdot]$ defined as in (9), we have*

$$\mathbb{E}_\nu\left[(\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|\cdot] - \mathbb{E}[Y|\cdot])^2\right] = \mathcal{O}(LN^{-1}, \log(N)^{p_d}N^{-\frac{1}{d}}, L^{-\frac{2\beta}{d}}),$$

where $p_d = 3/4$ for $d = 2$ and $p_d = 1/d$ for $d \ge 3$.

Proof. To compute the error rate, we split the error into the variance error term that arises due to discrete data and the bias term that arises due to the truncation of eigenfunctions:

$$\begin{aligned}&\mathbb{E}_\nu\left[(\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|\cdot] - \mathbb{E}[Y|\cdot])^2\right] \\ &\le \mathbb{E}_\nu\left[(\mathcal{N}_{\mu_N}\mathbb{E}_{L,N}[Y|\cdot] - K_\mu^*\mathcal{N}_\mu\mathbb{E}_L[Y|\cdot])^2\right] + \mathbb{E}_\nu\left[(K_\mu^*\mathcal{N}_\mu\mathbb{E}_L[Y|\cdot] - \mathbb{E}[Y|\cdot])^2\right] \\ &\le \mathbb{E}_\nu\left[\left(\sum_{j=0}^{L}(\langle \vec{y}, \phi_{j,N,\epsilon}\rangle_{L^2(\mu_N)} - \langle Y, \phi_j\rangle_H)\mathcal{N}_{\mu_N}u_{j,N,\epsilon}\right)^2\right]\ldots \\ &+ \mathbb{E}_\nu\left[\left(\sum_{j=0}^{L}\langle Y, \phi_j\rangle_H(\mathcal{N}_{\mu_N}u_{j,N,\epsilon} - K_\mu^*\mathcal{N}_\mu u_j)\right)^2\right] + \mathbb{E}_\nu\left[\left(\sum_{j>L}\langle Y, \phi_j\rangle_H u_j\right)^2\right] \\ &\le \frac{L}{N}\mathbb{E}_\mu[Y^2] + C\left(\frac{\log(N)^{p_d}}{N^{\frac{1}{d}}}\right) + \sum_{j>L}\langle Y, \phi_j\rangle_H^2,\end{aligned}$$

for some constant C that is independent of ϵ, N, d but can depend on L. In the second equality above for the variance term, we isolate the errors due to Monte-Carlo approximation of the expansion coefficients (which is computed in (8)), where we suppressed the order δ^2/N term since it is dominated by the error of order-δ^2 in the discrete approximation of the eigenfunctions. Using the recent result in [34] for compact manifolds without boundary, the L^2-error bound for each eigenfunction (as $\epsilon \to 0$) is given by $\delta = \mathcal{O}\left(\frac{\log(N)^{p_d}}{N^{1/d}}\right)^{\frac{1}{2}}$, where d denotes the intrinsic dimension of $\mathcal{X}$ and $p_d = 3/4$ for $d = 2$ and $p_d = 1/d$ for $d \ge 3$.

For all $Y = F \circ X$, we have that $\langle Y, \phi_j\rangle_H = \langle F, u_j\rangle_V = \hat{F}_j$, and since $F \in H^\beta(\mathcal{X})$, we have

$$\sum_{j>L}\langle Y, \phi_j\rangle_H^2 = \sum_{j>L}\hat{F}_j \le \frac{1}{\zeta_{L+1}^\beta}\sum_{j=0}^{\infty}\zeta_j^\beta\hat{F}_j^2 \le C_2\zeta_{L+1}^{-\beta},$$

for some constant $C_2 > 0$. The proof follows by using the Weyl asymptotic estimate for the eigenvalue of the Laplace–Beltrami operator on compact Riemannian manifolds [37], $\zeta_{L+1} \sim L^{2/d}$. □

We should point out that balancing the first and last error rates yields the famous minimax optimal rate, $\mathcal{O}(N^{-\frac{2\beta}{2\beta+d}})$ for linear estimators [38]. Thus, unless the response function is highly smooth (e.g., $\beta = d$), such an estimator is subject to

the curse of dimension. In practice, the second error rate (corresponding to the estimation of eigenvectors) will dominate in high-dimensional problems even if the target function is smooth.

2.2. Kernel smoothing estimator

In the previous subsection, we approximated $\mathbb{E}[Y|X]$ with, $\mathbb{E}_{L,N}[Y|X]$, a superposition of eigenvectors of $\mathbf{G}_N$ and then used Nyström extension (9) to evaluate this representation on an out-of-sample point. In this subsection we show that the conditional expectation can also be approximated by an appropriate smoothing function in H.

The main idea is motivated by the fact that if $\mathcal{X}$ is a smooth manifold, any measurable function $g \in V$ can be represented as

$$g(x) = \mathbb{E}_{\delta_x}[g],$$

where δ_x denotes the Dirac mass centered at x. We can then attempt to regularize this integral operation by approximating δ_x with an appropriate family of Markov kernels that have a smooth density with respect to the pushforward measure ν.

To that end, we assume that $\mathcal{X} = \mathbb{R}^m$ and the support of ν is a smooth, compact d-dimensional submanifold $\mathcal{M} \subseteq \mathcal{X}$. We then start with a kernel $S_\epsilon : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\epsilon > 0$ is a bandwidth parameter, and perform a sequence of normalizations that yield, asymptotically, the kernel κ_ϵ so that

$$G_\epsilon g(x) := \int_{\mathcal{X}} \kappa_\epsilon(x, x') g(x') d\nu(x') = g(x) + \mathcal{O}(\epsilon), \tag{10}$$

holds for $g \in V$ and $x \in \mathcal{M}$. The integral operator G_ϵ can then be approximated by a matrix–vector multiplication. In this paper, we use the variable bandwidth construction of the kernel given in [28]. This expansion starts with a kernel S_ϵ on $\mathcal{X} \times \mathcal{X}$ of the form

$$S_\epsilon(x, x') = \epsilon^{-d/2} \exp\left(-\frac{\|x - x'\|^2}{\epsilon \rho(x)\rho(x')}\right),$$

where $\rho > 0$ is a bandwidth function that is chosen to be inversely proportional to a power of the sampling density as in [28].

For completeness, we describe the construction of the discrete approximation of the operator in (10). Let $x_1, \ldots, x_N$ be the observed m-dimensional data in $\mathcal{X}$. Then the following steps (which are the diffusion maps normalizations [28]) yield a discrete approximation $G_{N,\epsilon}$ of the integral operator G_ϵ, whose discrete representation is denoted by the matrix $\mathbf{G}_{N,\epsilon} = [\langle e_{i,N}, G_{N,\epsilon} e_{j,N} \rangle_{L^2(\mu_N)}] = [\kappa_\epsilon(x_i, x_j)]$,

$$\begin{aligned} q_\epsilon(x_i) &:= \sum_{j=1}^N \frac{S_\epsilon(x_i, x_j)}{\rho(x_i)^d}, & S_{\epsilon,\alpha}(x_i, x_j) &:= \frac{S_\epsilon(x_i, x_j)}{q_\epsilon(x_i)^\alpha q_\epsilon(x_j)^\alpha}, \\ q_{\epsilon,\alpha}(x_i, x_j) &:= \sum_{j=1}^N S_{\epsilon,\alpha}(x_i, x_j) & \mathbf{G}_{N,\epsilon}(x_i, x_j) &:= \frac{S_{\epsilon,\alpha}(x_i, x_j)}{q_{\epsilon,\alpha}(x_i)}. \end{aligned} \tag{11}$$

The two steps in the first row above are the "right-normalization" steps taken to de-bias the possibly non-uniform sampling distribution of the data with a parameter α. In our numerics, we set $\alpha = -d/4$ and $\rho = q_\epsilon^{-1/2}$ as in [28]. The two "left-normalization" steps in the second row of (11) turn $\mathbf{G}_{N,\epsilon}$ into a stochastic matrix. Note that the resulting kernel κ_ϵ from (11) is given in Appendix A5 of [28]. Based on the result in [28], for $\{x_1, \ldots, x_N\} \subset \mathcal{M}$ with sampling density $q = d\nu/d\mathrm{vol}$, where vol is the volume form on $\mathcal{M}$ through its embedding in $\mathcal{X}$, for fixed ϵ, we have the convergence rate

$$\left(\mathbf{G}_{N,\epsilon}\vec{g}\right)_i = G_\epsilon g(x_i) + \mathcal{O}\left(\frac{q(x_i)^{1/2+d/4}}{N^{1/2}\epsilon^{2+d/4}}, \frac{q(x_i)^{d(d/2-1/4)}}{N^{1/2}\epsilon^{1/2+d/4}}\right), \tag{12}$$

as $N \to \infty$. The second term in the error bound is due to the error in the discrete estimate and the first term is to ensure an order-ϵ^2estimate of q_ϵ.

Our choice of normalization is to ensure an asymptotically unbiased (up to order ϵ) estimate of g in (10). For many applications, it suffices to start with the standard Gaussian kernel with constant bandwidth ($\rho = 1$) and apply the steps in the second row in (11) to create a valid transition density. However, in this paper, we will always construct the integral operator in (10) using the variable bandwidth kernels due to their accurate estimation of densities in sparsely sampled regions. To tune the kernel bandwidth parameter ϵ, we use the auto-tuning algorithm in [28] which was found to be more effective for variable bandwidth kernels than the Gaussian kernel with a fixed bandwidth ($\rho = 1$). Furthermore, we have

Proposition 2.2. *Let P be the orthogonal projection of H onto H_X. Then for any $g \in V$ and $x_i = X(\omega_i) \in \mathcal{M}$, the relationship*

$$\left(\mathbf{G}_{N,\epsilon}\vec{g}\right)_i = Pg(x_i) + \mathcal{O}\left(\epsilon, \frac{q(x_i)^{1/2+d/4}}{N^{1/2}\epsilon^{2+d/4}}, \frac{q(x_i)^{d(d/2-1/4)}}{N^{1/2}\epsilon^{1/2+d/4}}\right), \tag{13}$$

holds in high probability.

Proof. Note that for any $g \in V$, where $f = g \circ X \in H_X$, $\nu = X_*\mu$, and $x_i = X(\omega_i) \in \mathcal{M}$, a change of variables shows that

$$\begin{aligned} G_\epsilon g(x_i) &= \int_{\mathcal{X}} \kappa_\epsilon(x_i, x') g(x') d\nu(x') \\ &= \int_\Omega \kappa_\epsilon(X(\omega_i), X(\omega'))(g \circ X)(\omega') d\mu(\omega') := J_\epsilon f(\omega_i). \end{aligned} \tag{14}$$

Let $k_{\epsilon,\omega_i} := k_\epsilon(\omega_i, \cdot) = \kappa_\epsilon(X(\omega_i), X(\cdot))$. Since $f \in H_X$, we see that $J_\epsilon f(\omega_i) = J_\epsilon P f(\omega_i)$. Thus, for each $\omega_i \in \Omega$,

$$J_\epsilon P f(\omega_i) = (G_\epsilon P g) \circ X(\omega_i) = Pg \circ X(\omega_i) + \mathcal{O}(\epsilon) = Pf(\omega_i) + \mathcal{O}(\epsilon),$$

as $\epsilon \to 0$, due to the asymptotic expansion in (10). Together with (12) and (14), we have

$$\begin{aligned} \left(\mathbf{G}_{N,\epsilon}\vec{g}\right)_i - Pg(x_i) &= \left(\left(\mathbf{G}_{N,\epsilon}\vec{g}\right)_i - G_\epsilon g(x_i)\right) + (G_\epsilon g(x_i) - Pg(x_i)) \\ &= \left(\left(\mathbf{G}_{N,\epsilon}\vec{g}\right)_i - G_\epsilon g(x_i)\right) + (J_\epsilon f(\omega_i) - Pf(\omega_i)) \\ &= \mathcal{O}\left(\epsilon, \frac{q(x_i)^{1/2+d/4}}{N^{1/2}\epsilon^{2+d/4}}, \frac{q(x_i)^{d(d/2-1/4)}}{N^{1/2}\epsilon^{1/2+d/4}}\right). \end{aligned}$$

For a function $\vec{g}$ whose components are function values at the training data $\{x_1, \ldots, x_N\}$, the kernel smoothing estimate of $g(x_{out})$ on a new point x_{out} is given by $\mathbf{G}_{N_{out},\epsilon}\vec{g}$, where $\mathbf{G}_{N_{out},\epsilon}$ is a row vector consisting of $\kappa_\epsilon(x_{out}, x_i)$ for $i = 1, \ldots, N$. The definition of G_ϵ can be extended to $g \in \mathbb{R}^n$, where $n > 1$ componentwise. That is, if $g(x) = (g_1(x), \ldots, g_n(x))$, where $g_i(x) \in \mathbb{R}$ for $1 \le i \le n$ then $G_\epsilon g(x) := (G_\epsilon g_1(x), \ldots, G_\epsilon g_n(x))$. Then, the componentwise convergence in probability holds due to the preceding proposition. The discrete estimator then becomes a matrix–matrix multiplication, $\mathbf{G}_{N,\epsilon}\mathbf{g}$, where the ijth component of the matrix $\mathbf{g}$ is given by $g_i(x_j)$.

The kernel smoothing estimate is conceptually simple and computationally fast to construct. While naive, we will show in the next section that the kernel smoothing estimate of the conditional expectation performs well when accurate estimation of the eigenbasis of V is not available, especially when the covariate space is high-dimensional.

3. Predicting the dynamics of observables

In this section, we discuss the problem of predicting observables (e.g.,partial components) of a measure preserving discrete time dynamical system. We start by reviewing the MZ formalism [39], which is a classical reduced-order modeling framework

often used for this task. The MZ formalism expresses the evolution of the desired reduced order dynamics in terms involving Markovian, non-Markovian, and orthogonal dynamics through the use of an orthogonal projection operator. The total contributions of the Markovian and non-Markovian terms in this decomposition will coincide with the optimal solution to the regression problem with observables at initial and future times as covariate and response data, respectively. For a low-dimensional covariate space, we show that the regression estimator can be accurately constructed using the Nyström method in Section 3.2. In Section 3.3, we argue that if the hypotheses of delay-embedding theorems are satisfied [6,40], the MZ-equations can be simplified to only a "Markovian" term, whose representation is precisely the regression function that maps the delay-embedded observable to the observable at a future time. This regression function, which can be estimated by KAF, is nothing but the component of the flow map induced by the lag embedding. In such high-dimensional covariate space regression problems, we numerically demonstrate that the kernel smoothing estimate is a more accurate estimator than the Nyström method.

3.1. Mori–Zwanzig formalism for reduced order modeling

Let (Ω, Φ) be a discrete-time deterministic dynamical system, generated by an invertible map $\Phi : \Omega \to \Omega$. Furthermore, we assume that there is a Φ-invariant probability measure $\mu : \mathcal{B}(\Omega) \to [0, 1]$, where $\mathcal{B}(\Omega)$ is the Borel σ-algebra on Ω. That is, for all $B \in \mathcal{B}(\Omega)$, $\mu(\Phi^{-1}(B)) = \mu(B)$. As in Section 2, we assume that μ is supported on a compact set $M \subseteq \Omega$. For a given $\omega_0 \sim \mu$, we let $\omega_i := \Phi^i(\omega_0)$.

In what follows, we assume that only partial observations $x_i \in \mathbb{R}^n$ of ω_i are available and they are defined through a measurable function $X : \Omega \to \mathcal{X} = \mathbb{R}^n$ such that $x_i := X(\omega_i) = X \circ \Phi^i(\omega_0)$. Since our goal is to use the observed time series of $\{x_i\}$ to estimate $x_{i+t} \in \mathcal{X}$ for some $t \in \mathbb{N}$, we set the response space equals to the covariate space, $\mathcal{Y} = \mathcal{X}$, and consider the response map $Y = X_t : \Omega \to \mathcal{X}$, where $X_t = X \circ \Phi^t$; that is, $X_t(\omega_0) = X(\omega_t)$. Note that since $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$, we have $H = \{f : \Omega \to \mathcal{X} : \int_\Omega \|f^2(\omega)\|^2 d\mu(\omega) < \infty\}$, $V = \{g : \mathcal{X} \to \mathcal{X} : g \circ X \in H\}$, and $H_X = \{f \in H : f = g \circ X \text{ for some } g \in V\}$.

Let us define an orthogonal projection operator $P : H \to S_X \subseteq H_X \subseteq H$, where $S_X = \text{ran}(P)$ is a closed subspace of H_X, and let $Q = I - P$ be the orthogonal projection onto the orthogonal space, $S_X^\perp = \text{null}(P)$. With these projection operators, we have $H = S_X \oplus S_X^\perp$. Let $U : H \to H$ be the Koopman operator defined as $Uf = f \circ \Phi$, for all $f \in H$. Then by "the Dyson's formula" [41,42], the map $U^i : H \to H$ given by $U^i f = f \circ \Phi^i$ can be written as,

$$U^{i+1} = \sum_{k=0}^{i} U^{i-k} PU(QU)^k + (QU)^{i+1}. \tag{15}$$

Applying (15) on $X \in H$, we obtain the discrete MZ equation that describes the evolution of $x_i = X(\omega_i)$. In detail, letting $\Xi_i := (QU)^i X$ and noting that

$$\begin{aligned} U^{i+1}X &= X_{i+1} \\ U^{i-k}PU(QU)^k X &= PU(QU)^k X \circ \Phi^{i-k} = P(\Xi_k \circ \Phi)(X \circ \Phi^{i-k}) \\ (QU)^{i+1}X &= \Xi_{i+1}, \end{aligned}$$

along with the fact that $PQ = 0$, yields

$$X_{i+1}(\omega_0) = (PUX \circ \Phi^i)(\omega_0) + \sum_{k=1}^{i} P(\Xi_k \circ \Phi) \circ (X \circ \Phi^{i-k})(\omega_0) + \Xi_{i+1}(\omega_0). \tag{16}$$

Since $PUX \in S_X$, there exists an $M_0 \in V$ such that $PUX = M_0 \circ X$, by the definition of S_X. Similarly, since $P(\Xi_k \circ \Phi) \circ X \in S_X$, there exists $M_k \in V$ such that $P(\Xi_k \circ \Phi) \circ X = M_k \circ X$. Therefore, Eq. (16) can be written in terms of the observable values x_i as

$$x_{i+1} = M_0(x_i) + \sum_{k=1}^{i} M_k(x_{i-k}) + \Xi_{i+1}(\omega_0). \tag{17}$$

Note that (17) decomposes x_{i+1} into the Markovian term M_0, the memory terms M_i and a term Ξ_i that is orthogonal to S_X.

3.2. Approximation of the projected Mori–Zwanzig equation

If we consider the specific choice $S_X = H_X$, and thus the projection operator $P := \mathbb{E}[\cdot|X]$, we obtain the projected MZ-equation,

$$\mathbb{E}[X_{i+1}|x_0] = M_0(x_i) + \sum_{k=1}^{i} M_k(x_{i-k}), \tag{18}$$

since $M_k \circ X \in H_X$, and the orthogonal term $P\Xi_{i+1} = 0$ since $\Xi_{i+1} \in H_X^\perp$. In this case, notice that $\mathbb{E}[X_{i+1} \mid \cdot]$ is precisely the minimizer in (1) with X_{i+1} in place of Y. The main takeaway here is that the solutions of the projected MZ-equation in (18) is the regression function, $\mathbb{E}[X_{i+1}|\cdot]$, of the dynamical map $X_0 \mapsto X_{i+1}$. The importance of this observation is that one can approximate $\mathbb{E}[X_{i+1}|X_0]$ from the historical data $\{x_i\}$. In the next two examples, we will numerically verify this claim with the two nonparametric estimators discussed in Section 2, the Nyström method and kernel smoothing.

Hamiltonian system: First, consider the 16-dimensional dynamical system given by the Hamiltonian

$$H(\omega) = \frac{1}{2}\left(\sum_{i=1}^{16} \omega_{(i)}^2 + \sum_{i=1}^{7} \omega_{(2i-1)}^2 \omega_{(2i+1)}^2\right), \tag{19}$$

where $(\omega_{(2i-1)}, \omega_{(2i)})$ for $i = 1, \ldots, 8$ are the canonical conjugate variables and $\omega = (\omega_{(1)}, \ldots, \omega_{(16)}) \in \mathbb{R}^{16}$. Thus the full system is derived through the relations

$$\frac{d\omega_{(2i-1)}}{dt} = \frac{\partial H(\omega)}{\partial \omega_{(2i)}}, \qquad \frac{d\omega_{(2i)}}{dt} = \frac{\partial H(\omega)}{\partial \omega_{(2i-1)}}. \tag{20}$$

In addition to the subscript-(i) used to denote the ith component of $\omega \in \mathbb{R}^{16}$, we will use the notation ω_j to denote the jth sample of Ω with sampling measure $\frac{d\mu}{d\omega} \propto e^{-H(\omega)}$, where $\omega_j = (\omega_{j,(1)}, \ldots, \omega_{j,(16)})$. We should point out that this example is a high-dimensional version of the main example in [7].

Suppose we are interested in the conditional expectation $\mathbb{E}[\omega_{(1)}(t), \omega_{(2)}(t) \mid \omega_{(1)}(0), \omega_{(2)}(0)]$, where the expectation is drawn from the canonical invariant density μ corresponding to H with fixed $\omega_{(1)}(0)$ and $\omega_{(2)}(0)$. We can view the problem of estimating the conditional density in the regression framework as follows. Let us define the covariate map $X : \Omega \to \mathcal{X}$ by $X(\omega(0)) = (\omega_{(1)}(0), \omega_{(2)}(0)) := x_0 \in \mathcal{X}$ and the response variable $X_t(\omega) = X(\omega(i\Delta t)) = (\omega_{(1)}(i\Delta t), \omega_{(2)}(i\Delta t))$, where $\Delta t > 0$ is a fixed time step.

In this example, we will consider estimates based on the Nyström method with $L = 100$ and the kernel smoothing estimator. For this application, let $\vec{x}_0 = (x_{0,1}, \ldots, x_{0,N})$ and $\vec{x}_0^{out} = (x_{0,1}^{out}, \ldots, x_{0,N_{out}}^{out})$ be two vectors that will be used for training and verification, respectively. Each component of these vectors is an i.i.d sample of X_0, that is, $x_{0,j} = X_0(\omega_j)$, where ω_j is drawn independently from μ. Let $\vec{x}_t := (x_{t,1}, \ldots, x_{t,N})$ be a vector of the training time series with components given by $x_{t,j} = U^t \circ X_0(\omega_j)$.

In the following numerical experiments, the time series $\{x_{t,1}, \ldots, x_{t,N}\}_{t=0,1,\ldots}$ was observed at the sampling interval $\Delta t = .1$ time units, and the initial conditions $\{x_{0,1}, \ldots, x_{0,N}\}$ are samples of the invariant density $\nu = X_* \mu$. We verify the quality of the estimators on $N_{out} = 1000$ out-of-sample initial conditions, also

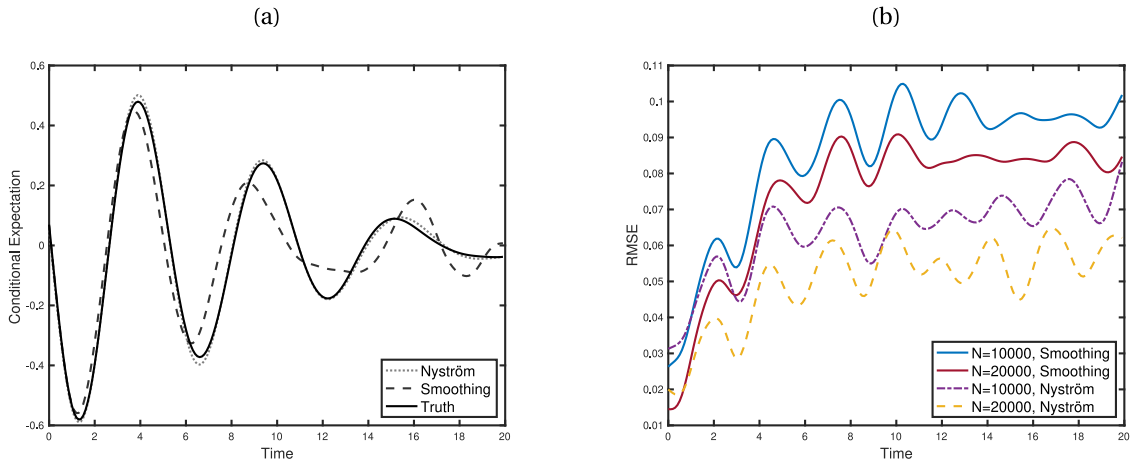

Fig. 1. Hamiltonian Example: (a) Comparison of the kernel smoothing estimate (Smoothing), the Nyström method estimate, and the MC empirical estimate (which is considered as the truth) of the conditional expectation of the first component of a particular out-of-sample trajectory, trained using $N = 20{,}000$ samples. (b) The RMSEs (based on $N_{out} = 1000$ samples) of both estimators as functions of lead time forecast, constructed using $N = 10{,}000$ and $N = 20{,}000$ data points.

sampled from ν. To verify the performance of the two estimators, we compare them to the empirical conditional expectation obtained from a Monte-Carlo simulation. This calculation requires samples of the conditional distribution of $(\omega_{(3)}(0), \ldots, \omega_{(16)}(0))$ given each out-of-sample initial condition, $(\omega_{(1)}(0), \omega_{(2)}(0)) = x_{0,j}$, where $j = 1, \ldots, N_{out}$. Numerically, we obtain these samples, denoted by $(\omega^{(k)}_{(3),j}(0), \ldots, \omega^{(k)}_{(16),j}(0))$, using the Hamiltonian Monte Carlo method [43] on the reduced Hamiltonian in (19) with fixed $(\omega_{(1)}(0), \omega_{(2)}(0)) = x_{0,j}$. Concatenating these samples and the fixed $x_{0,j}$, we define $\omega^{(k)}_j = (x_{0,j}, \omega^{(k)}_{(3),j}(0), \ldots, \omega^{(k)}_{(16),j}(0))$, for $j = 1, \ldots, N_{out}$, $k = 1, \ldots, N_{MC}$. In the numerical result below, we use $N_{MC} = 20{,}000$ samples for each initial condition $x_{0,j}$. Given these samples, the Monte-Carlo approximation of $\mathbb{E}[X_t \mid \cdot]$ is given by

$$\mathbb{E}[X_t \mid x_{0,j}] \approx \frac{1}{N_{MC}} \sum_{k=1}^{N_{MC}} U^t \circ X(\omega^{(k)}_j) = \frac{1}{N} \sum_{k=1}^{N_{MC}} X \circ \Phi^t(\omega^{(k)}_j), \qquad (21)$$

where each realization $\Phi^t(\omega^{(k)}_j)$ is the solution of the full dynamics with solution map denoted by Φ. Note that the solution map of (19) is the result of a temporal discretization of the Hamiltonian dynamics in (20); in our numerics, we use the Runge–Kutta-4 (RK4) method.

In Fig. 1, we show a comparison of the Nyström method and the kernel smoothing estimate of the conditional expectation, constructed using $N = 20{,}000$ training samples and the empirical Monte-Carlo estimate in (21) for a particular out-of-sample data (which is considered as the truth). Notice that the Nyström method is significantly more accurate than the kernel smoothing estimate. In Fig. 1(b), we also show the Root-Mean-Square-Errors (RMSEs) between the two estimators and the empirical Monte-Carlo estimate of the conditional expectation, averaged over $N_{out} = 1000$ out-of-sample initial conditions for training data sizes, $N = 10{,}000$ and $20{,}000$. Besides the clear advantage of the Nyström method over kernel smoothing, notice that both estimators are improved as the size of training data, N, increases.

Although the Nyström method performs better than the kernel smoothing estimate, the former is computationally more expensive. For both methods we construct $\mathbf{G}_{N,\epsilon}$ using the steps in (11). To alleviate memory and computation costs associated with full $N \times N$ kernel matrices for $N \gg 1$, in practice a k-nearest neighbor algorithm is employed so the resulting matrix $\mathbf{G}_{N,\epsilon}$ is sparse with approximately k nonzero entries on each row. To predict on a new data point using the kernel smoothing method, one only needs to extend $\mathbf{G}_{N,\epsilon}$ to the new point and multiply the extended row vector by the column of the response training data. On the other hand, the Nyström method requires the eigenvectors of the matrix $\mathbf{G}_{N,\epsilon}$ and then employs the Nyström extension method to approximate the eigenfunctions evaluated on the new out of sample data points. Thus, after constructing $\mathbf{G}_{N,\epsilon}$, the computational cost of the kernel smoothing method is $\mathcal{O}(k)$ where k is the number of nearest neighbors employed, while the cost for the Nyström method is $\mathcal{O}(kL)+\mathcal{O}(\text{E.D.})$, where L is the number of eigenfunctions used and $\mathcal{O}(\text{E.D.})$ is the cost of acquiring the L eigenvectors.

The Lorenz-96 model: Next, we consider the Lorenz-96 model [44] given by

$$\frac{d\omega_{(i)}}{dt} = (\omega_{(i+1)} - \omega_{(i-2)})\omega_{(i-1)} - \omega_{(i)} + F \qquad (22)$$

for $i = 1, \ldots, 5$, forcing parameter $F = 8$ and, with periodic boundary condition, $\omega_{(-1)} = \omega_{(4)}$, $\omega_{(0)} = \omega_{(5)}$, and $\omega_{(6)} = \omega_{(1)}$. In this regime, the dynamics is chaotic with attractor dimension 2.9 and two positive Lyapunov exponents as reported in [45]. We estimate the conditional expectation $\mathbb{E}[X_t \mid X_0]$, where the covariate function is $X(\omega) = \omega_{(1)}(0)$, the response function is $X_t(\omega) = X(\Phi^t(\omega)) = \omega_{(1)}(t)$, and the initial conditions are drawn from the standard Gaussian distribution. Note that this distribution is not invariant under the dynamics of the system. Here Φ is given by the RK4 discretization of (22) with time step $1/64$.

Numerically, we generate $N = 20{,}000$ and $N_{out} = 1000$ initial conditions for training and verification, respectively, from the standard five-dimensional multivariate Gaussian and integrate the training data forward 2.5 time units to generate training time series observations. Subsequently, we used only the first component, $\omega_{(1)}$, of the initial conditions and the training time series to construct, both, the Nyström and kernel smoothing estimates of the conditional expectation. For the Nyström method, we use $L = 300$ eigenfunctions. Both estimators are compared to an empirical estimator which is obtained by averaging (21) over $\omega^{(k)}_j = (x_{0,j}, \omega^{(k)}_{(2),j}(0), \ldots, \omega^{(k)}_{(5),j}(0))$, for $j = 1, \ldots, N_{out}$, $k = 1, \ldots, N_{MC} = 20{,}000$ samples of initial conditions. Here, the first component of each initial condition, $x_{0,j} = X(\omega^{(k)}_j)$, is one of the $N_{out} = 1000$ verification samples and the other components are drawn from the four-dimensional standard Gaussian. In Fig. 2(a), we show the evolution of one of the 1000 verification samples. Comparing the three estimates, notice the closer agreement between the

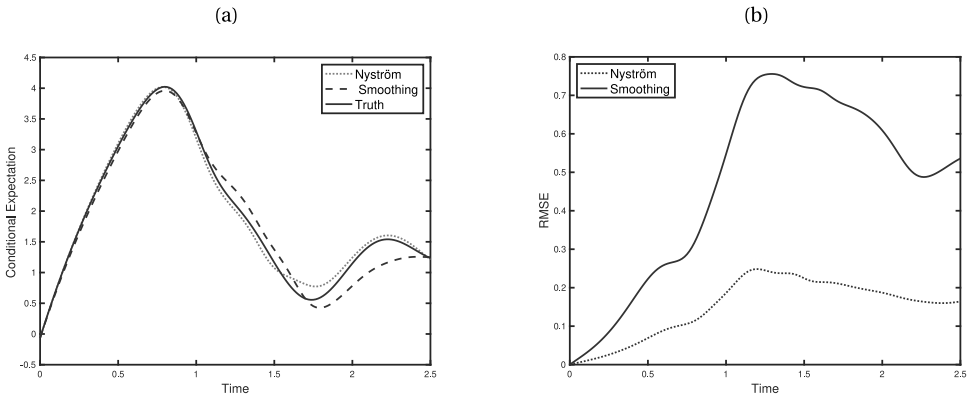

Fig. 2. The Lorenz-96 example: (a) Comparison of the kernel smoothing estimate (Smoothing) and the Nyström method of the conditional expectation of the first component of a particular out-of-sample trajectory, trained using $N = 20{,}000$ training data. (b) The RMSEs (based on $N_{out} = 2000$) of both estimators as functions of lead time forecast, constructed using $N = 20{,}000$ data points.

Nyström and the empirical estimates. In Fig. 2(b) one can see that the RMSE (based on averaging over $N_{out} = 1000$ out-of-sample points) of the Nyström-based estimate is more accurate than the kernel smoothing estimate. This result is consistent with the previous example.

Next, we will show that the full MZ equation (16) can be constructed by an optimal least squares estimator of a regression framework with appropriate choice of covariate space.

3.3. Mori–Zwanzig projection and delay-coordinate maps

In the preceding subsection, we considered estimating the conditional expectation $\mathbb{E}[X_t \mid X_0]$ and showed that it can be numerically approximated using the time series data. In this subsection, we are interested in predicting the realization of x_t in (17). While the MZ representation suggests that the solution depends on the entire historical data, for practical computation, finite-memory models to collectively represent these terms as a finitely supported function is desirable. Since the memory terms depend on the orthogonal dynamics (see Eq. (16)), such an approximation can be achieved, e.g., by delta function approximation [15], Krylov subspace approximation [16], rational approximation [19], or a series representation of the orthogonal dynamics [17,18]. Note that while a finite-dimensional (matrix) representation is the computational object of interest, a series representation may not converge since it involves expansion of semigroups generated by unbounded operators.

On the other hand, we should point out that depending on the choice of the projection operator P, the explicit representation of the terms in the MZ equation, $(M_j)_{j=0}^{t-1}$ as well as the orthogonal dynamics Ξ_t, may or may not be easily translated into an efficient algorithm that yields a consistent approximation. As we showed in the preceding subsection, choosing $P = \mathbb{E}[\cdot|X]$ as an estimator will not yield an accurate approximation to x_t since this estimator truncates the orthogonal dynamics. Other common choice of projection operators can be found in [7,39]. For example, while the popular Mori projection, defined as $P = \langle X, X\rangle_H^{-1}\langle \cdot, X\rangle_H X$, yields a linear model for $(M_j)_{j=0}^{t-1}$, the representation of the orthogonal dynamics in such a basis expansion may not be computationally tractable [10].

Recently, it was shown in [42] that by choosing P to be the Wiener projection, one can simplify the MZ equation so that only the Markovian term M_0 and orthogonal terms Ξ_t remain, where M_0 is now a function that takes a delay coordinate of the observable. Building on this result, our intuition is to construct a non-decreasing sequence of projection operators $\{P_m : m \in \mathbb{N}\}$ which allows one to access the entire function space H with a finite m and a simple representation of the MZ equation. In what follows, we argue that delay-embedding theorem [6] provides a natural candidate for achieving this goal.

To that end, we define the delay coordinate map $\mathbb{X}_m : \Omega \to \mathcal{X}^m$ by $\mathbb{X}_m(\omega) = (X_{-m+1}(\omega), \ldots, X_{-1}(\omega), X_0(\omega))$, $X_i = U^i \circ X$, which we will consider as the covariate function. Simultaneously, we consider the response function $X_t : \Omega \to \mathcal{X}$, where $\mathcal{X}$ is the response space as in the preceding sections. Note that the optimal estimator for the map $\mathbb{X}_m \mapsto X_t$ is given by the conditional expectation $P_m X_t := \mathbb{E}[X_t \mid \mathbb{X}_m]$. Under mild assumptions on the covariate X, the dynamical flow Φ, and the sampling interval Δt, the theory of delay-coordinate maps [6,40] states that $\mathbb{X}_m$ is a homeomorphism between the support, M, of the invariant measure and $\mathcal{X}^m$ for sufficiently large m. Consequently, the Borel sigma algebra on M is identical to the sigma algebra generated by $\mathbb{X}_m$, $\sigma(\mathbb{X}_m)$. Thus, X_t is measurable with respect to the sigma algebra generated by $\mathbb{X}_m$, which means that $P_m X_t := \mathbb{E}[X_t \mid \mathbb{X}_m] = X_t$ is the identity map for sufficiently large m.

Let m be such that the embedding result stated above holds. Then, letting $P = P_m$ in (16), the memory and the orthogonal terms vanish since they involve $Q_m UX = (I - P_m)UX = 0$. While P_m is an identity operator, the MZ equation reduces to a contribution of a Markovian term,

$$x_{i+1} = (P_m UX \circ \Phi^i)(\omega_0) = \mathbb{E}[UX|\mathbb{X}_m(\omega_i)] = M_0(x_{i-m}, \ldots, x_i),$$

for some $M_0 \in V_m := \{f : \mathcal{X}^m \to \mathcal{X} : f \circ \mathbb{X}_m \in H\}$. If we define the flow map T on $\mathcal{X}^m$ induced by Φ as $T \circ \mathbb{X}_m(\omega_i) := \mathbb{X}_m \circ \Phi(\omega_i)$, then M_0 is the mth component of the flow map T, which is also the regression function of the supervised learning task $\mathbb{X}_m \mapsto X_t$.

In light of this connection, we will employ the nonparametric estimators discussed in Section 2 to approximate the regression function M_0 and numerically show that true trajectory of the observables can be recovered with adequate accuracy for sufficiently large m.

Hamiltonian system: As an example, consider again the Hamiltonian system in (19)–(20). Here, we are interested in approximating $\mathbb{E}[\omega_{(1)}(t) \mid \omega_{(1)}(-m+1), \ldots, \omega_{(1)}(-1), \omega_{(1)}(0)]$ for $t \in \mathbb{Z}_+$. Letting the response function be $X_t(\omega) = \omega_{(1)}(t) = x_t$ and the covariate function be $\mathbb{X}_m := (X_{-m+1}, X_{-1}, X_0)$, we can rewrite the conditional expectation of interest as, $\mathbb{E}[X_t|\mathbb{X}_m(\omega)]$. As before, we approximate this conditional expectation using the Nyström method with $L = 300$ eigenfunctions and the kernel smoothing estimator. The training data was generated by evolving $N = 20{,}000$ initial conditions $\{\omega_0^{(k)}\}_{k=1,\ldots,N}$, drawn from the invariant measure μ, for m units in time, using RK4 with the same specification as in the previous example. Fig. 3(a) shows a

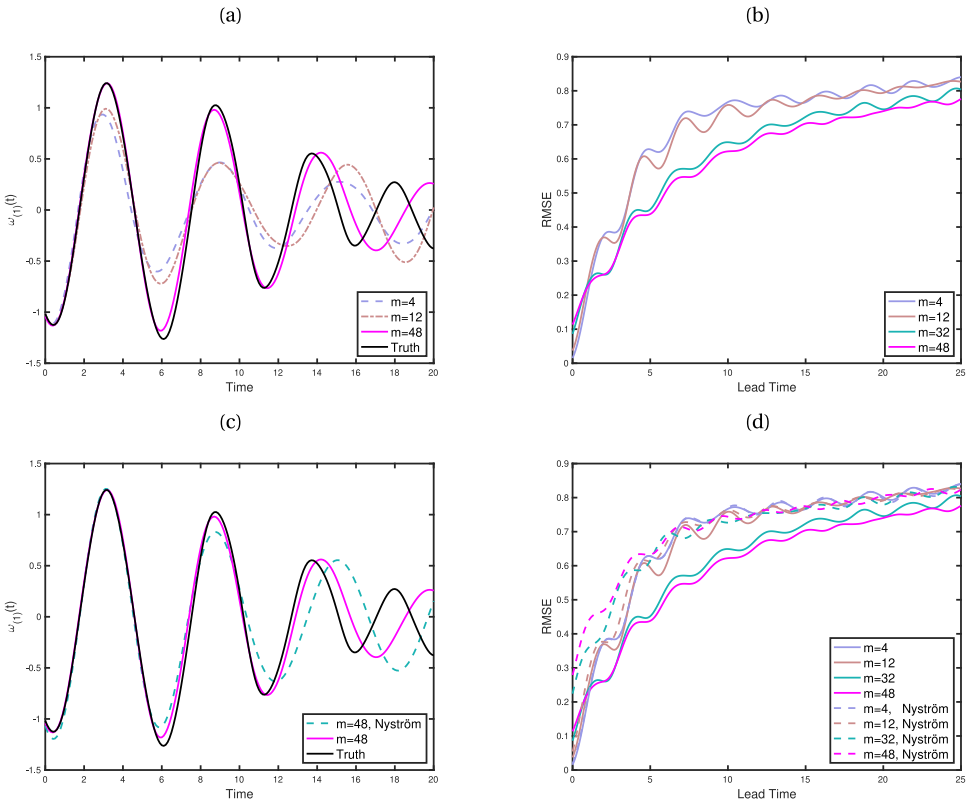

Fig. 3. Hamiltonian Example: (a) The trajectory of the first component, $\omega_{(1)}(t)$ for a particular out-of-sample initial condition along with the kernel smoothing estimates using $m = 4$, 12, and 48 past observations. (b) The RMSE between the true trajectory and the kernel smoothing estimates of the conditional expectation, calculated over 10,000 out-of-sample points. (c) A comparison of the kernel smoothing estimate and the Nyström estimate of the trajectory using $m = 48$ past data points. (d) The RMSEs of the Nyström and the kernel smoothing estimates of the trajectory for $m = 4$, 12, 32, and 48. Note that the RMSE plots show the RMSE for the lead time and omits the respective training windows for each m.

particular out-of-sample trajectory along with the kernel smoothing estimates of the trajectory for various choices of m. Notice that as m increases, the kernel smoothing estimator $\mathbb{E}[X_t|\mathbb{X}_m]$ approaches the true trajectory. In Fig. 3(b), we show the RMSE, averaged over $N_{out} = 10\,000$ out-of-sample verification points. Notice that the RMSEs are smaller as m increases except at initial time. The worse performance at initial time is not so surprising since the kernel smoothing is not an interpolation method, and thus will not be consistent with the given initial conditions. In panel (c), we show the quality of the prediction for $m = 48$ for a particular trajectory. Notice that while the trajectory is well estimated by both methods up to about 6 time units, the kernel smoothing method is more accurate compared to the Nyström method. The improved prediction of the kernel smoothing method compared to the Nyström method at longer times is consistent for different length of memory, m, as shown by the RMSE metric in panel (d), computed over $N_{out} = 10{,}000$ out-of-sample verification points.

The Lorenz-96 model: In this example, we consider predicting the first component $\omega_{(1)}(t)$ of the five-dimensional Lorenz-96 model given by (22), again with $F = 8$. As in the previous example, we will compare the Nyström method with $L = 300$ eigenfunctions and the kernel smoothing method in approximating $\mathbb{E}[X_t|\mathbb{X}_m(\omega)]$, where $X_t(\omega) = \omega_{(1)}(t)$. In this example, the delay-embedded data, $\mathbb{X}_m(\omega)$, are sampled from the invariant distribution of the system by running initial conditions sufficiently forward in time. In particular, we take $N = 20{,}000$ samples from the invariant distribution and construct the conditional expectation using observations of the first component of the samples. Here, the time series $U^t \circ X(\omega_1)$ used for constructing the estimator were observed at time steps of $\Delta t = 1/64$. In Fig. 4(a), we show a prediction of a particular out-of-sample realization of $\omega_{(1)}(t)$. Notice, as with the previous example, that the quality of the kernel smoothing estimator increases with m, except at the initial time as seen in Fig. 4(b). The RMSE in 4(b) was calculated over $N_{out} = 10{,}000$ out-of-sample initial conditions, also sampled from the invariant measure. As one can see, the kernel smoothing estimator is consistently more accurate than the Nyström method for similar m.

In principle, we should point out that the Nyström method can be improved with a larger number of eigenfunctions L. However, there is a practical issue in realizing improved accuracies. In our numerical tests, we do not find any meaningful improvement using any larger L compared to the present results with $L = 300$. We suspect that as the covariate space dimension increases (here, controlled by the number of delays m), the Nyström method requires an increasingly higher number of eigenfunctions L to reconstruct the response at a given level of accuracy, and for the available number of training samples N, these eigenfunctions cannot be accurately estimated. Thus, unless a mechanism is in place to ensure that the response is well approximated by the leading kernel eigenfunctions, or the eigenfunctions corresponding to large L can be robustly estimated with modest amounts of data (both of which are highly nontrivial problems), there may be practical limitations to improving the performance of the Nyström method simply by increasing the number of eigenfunctions employed. From these numerical results, we conclude that the kernel smoothing method (which requires much less

computational effort) is an effective alternative when an accurate estimation of the eigenfunctions is not available.

4. Smoothing and predicting with noisy data

To apply the prediction framework discussed in preceding section to real applications, one has to take into account that the available data set is most likely corrupted by noise. To overcome this issue, we design a nonparametric state estimation method, which we will call a smoother, to denoise the data. We should point out that we adopted the terminology smoother since the object of interest is the conditional expectation of the classical Bayesian smoothing problem [46], which is different than the kernel smoothing method described in Section 2. In Section 4.1, we describe a nonparametric smoother, formulated using the nonparametric regression framework discussed in Section 2. Subsequently, in Section 4.2, we numerically verify the prediction skill of the framework in Section 2 where the estimator is trained using the smoothed data obtained from the method in Section 4.1.

4.1. A nonparametric smoother

We consider smoothing noisy time series observations of the form

$$z_t = x_t + \theta_t, \qquad t = 1, \ldots, N,$$

where $\theta_t := \Theta_t(\alpha)$ are realizations of a centered random variable $\Theta_t : A \to \mathcal{X}$ that is independent of x_t. As in Section 3, $x_t = X_t(\omega)$ where $X_t : \Omega \to \mathcal{X}$, $\Omega \subset \mathbb{R}^n$, $\mathcal{X} = \mathbb{R}^p$ and $p \leq n$. The goal here is to construct the smoother $\mathbb{E}[X_k \mid z_1, \ldots, z_m]$, where $0 \leq k \leq m-1$, and use it as an estimator for x_k. In this application, we assume that the full vector ω_0 of initial conditions is drawn from an invariant density. To pose this smoothing problem in the regression framework presented in Section 2, we take the covariate space $\mathcal{Z}_m = (\mathbb{R}^p)^m$ to be the range of the covariate mapping $\mathbb{Z}_m : A \times \Omega \to \mathcal{Z}_m$ given by $\mathbb{Z}_m(\alpha, \omega) := \{Z_1(\alpha, \omega), Z_2(\alpha, \omega), \ldots, Z_m(\alpha, \omega)\}$, where $Z_k(\alpha, \omega) = X_k(\omega) + \Theta_k(\alpha)$. We also let the response space $\mathcal{X}_k$ to be the range of $X_k = \mathbb{R}^p$. Then the least squares estimator is given by $\mathbb{E}[X_k \mid \mathbb{Z}_m]$. If the distribution of x_k is invariant, then the covariate space $\mathcal{Z}_m$ and the response space $\mathcal{X}_k$ do not depend on time so the optimal estimator can be trained once using training data drawn from the invariant density.

In general, the noise-free time series x_t is not available for training. Given this constraint, we consider estimating $\mathbb{E}[Z_k \mid \mathbb{Z}_m]$ instead. Denote the available noisy data by $\vec{z} = \{z_1, \ldots, z_N\}$ and let $\vec{x} = \{x_1, \ldots, x_N\}$, and $\vec{\theta} = \{\theta_1, \ldots, \theta_N\}$ be the uncorrupted data, and the noise, respectively. We employ the VBDM algorithm to obtain the basis functions $\hat{u}_{j,N}(z_{i+1}, \ldots, z_{i+m}) = \hat{\phi}_{j,N}(\alpha_i, \omega_i)$, for $i = 1, \ldots, N$ and then represent $E[Z_k \mid \mathbb{Z}_m]$ as a superposition of these basis functions. We motivate the construction of this conditional expectation by noting that the diffusion maps algorithm is robust to low noise perturbations (see Criterion 5 of [31]); more specifically, the error in the spectrum of the graph Laplacian can be controlled as long as the size of perturbation $|\theta_t| < \sqrt{\epsilon}$. Assuming that this argument holds for the estimation of the eigenvectors, we can reasonably expect that $\hat{u}_{j,N}(z_{i+1}, \ldots, z_{i+m}) = \hat{\phi}_{j,N}(\alpha_i, \omega_i) \approx \phi_{j,N}(\omega_i) = u_{j,N}(x_{i+1}, \ldots, x_{i+m})$ when the noise size is smaller than $\sqrt{\epsilon}$.

For a particular class of dynamical systems, the noise robustness of the graph-theoretic techniques can be considerably strengthened by performing delays. In [47], it was shown that if the Koopman operator of the dynamical system on Ω has a pure point spectrum, and the noise is i.i.d. with finite first four moments, the pointwise estimator for the graph Laplacian determined from the noisy data can be made to agree with the noise-free estimator at any desired tolerance by increasing the embedding window length m. Here, we do not assume that the dynamics has pure point spectrum, so the estimates in [47] do not necessarily apply, but we can heuristically deduce that,

$$\begin{aligned}
\langle \vec{z}, \hat{\phi}_{j,N} \rangle_{L^2(\hat{\mu}_N)} &:= \frac{1}{N} \sum_{i=1}^{N} z_i \hat{u}_{j,N}(z_{i+1}, \ldots, z_{i+m}) \\
&\approx \frac{1}{N} \sum_{i=1}^{N} z_i u_{j,N}(x_{i+1}, \ldots, x_{i+m}) \\
&= \frac{1}{N} \sum_{i=1}^{N} x_i u_{j,N}(x_{i+1}, \ldots, x_{i+m}) \\
&\quad + \frac{1}{N} \sum_{i=1}^{N} \theta_i u_{j,N}(x_{i+1}, \ldots, x_{i+m}) \\
&= \langle \vec{x}, \phi_{j,N} \rangle_{L^2(\mu_N)},
\end{aligned}$$

due to the fact that θ_i is independent of x_i. Here, $\hat{\mu}_N = \sum_{i=1}^{N} \delta_{\alpha_i, \omega_i}/N$ is the discrete sampling measure on noisy data, whereas μ_N is the discrete sampling measure on uncorrupted data. This suggests that we can approximate the smoother $\mathbb{E}_{L,N}[X_k \mid \mathbb{Z}_m]$ as,

$$\begin{aligned}
\mathbb{E}_{L,N}[X_k \mid \mathbb{Z}_m] &:= \sum_{j=0}^{L} \langle \vec{x}, \phi_{j,N} \rangle_{L^2(\mu_N)} \phi_{j,N} \approx \sum_{j=0}^{L} \langle \vec{z}, \hat{\phi}_{j,N} \rangle_{L^2(\hat{\mu}_N)} \phi_{j,N} \\
&= \mathbb{E}_{L,N}[Z_k \mid \mathbb{Z}_m].
\end{aligned}$$

Here, the approximation is due to the fact that the construction of the estimator is based solely on the noisy data. A more detailed error analysis is an open problem that is beyond the scope of this paper. Note that the kernel smoothing estimator discussed in Section 2.2 does not possess the robustness-to-noise property that the VBDM basis does and furthermore the kernel smoothing estimate of z_k is not an approximation of the kernel smoothing estimate trained using x_k. Thus, an application of the kernel smoothing estimator discussed in Section 2.2 trained solely on noisy data would not yield a good approximation to the desired smoother, $\mathbb{E}[X_k \mid \mathbb{Z}_m]$.

In the remainder of this subsection, we will demonstrate the effectiveness of this smoother in recovering x_k from an out-of-sample sequence $(z_1, \ldots, z_m)$ of noisy observations. We will show numerically that choosing $m > 1$ reduces the RMSE when estimating x_k. We will also demonstrate the sensitivity of the RMSE as the parameter k is varied. To verify the accuracy of the proposed smoother, we compare the RMSE of this method to the RMSEs of the Ensemble Kalman Filter [29] and 4D-Var [30], both of which are very popular data assimilation methods that are operationally used in weather forecasting.

Smoothing noisy observations of the Lorenz-63 system: Consider the Lorenz-63 system given by

$$\begin{aligned}
\frac{d\omega_{(1)}}{dt} &= \sigma(\omega_{(2)} - \omega_{(1)}) \\
\frac{d\omega_{(2)}}{dt} &= \omega_{(1)}(\rho - \omega_{(3)}) - \omega_{(1)} \\
\frac{d\omega_{(3)}}{dt} &= \omega_{(1)}\omega_{(2)} - \beta\omega_{(3)}
\end{aligned} \tag{23}$$

with the standard parameters $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ that give rise to the famous "butterfly"-like attractor [48]. In this example, we are interested in estimating $\omega_{(1)}$, from observations of the form $Z_t(\omega(0)) = \omega_{(1)}(t) + \theta_t$, and we will use the VBDM algorithm to construct the smoother $\mathbb{E}_{L,N}[Z_k \mid \mathbb{Z}_m]$. Practically, given an out-of-sample sequence $\mathbf{z}_{t,m}^{out} = (z_{t+1}^{out}, \ldots, z_{t+m}^{out})$ of consecutive noisy observations, we use the Nyström method to evaluate

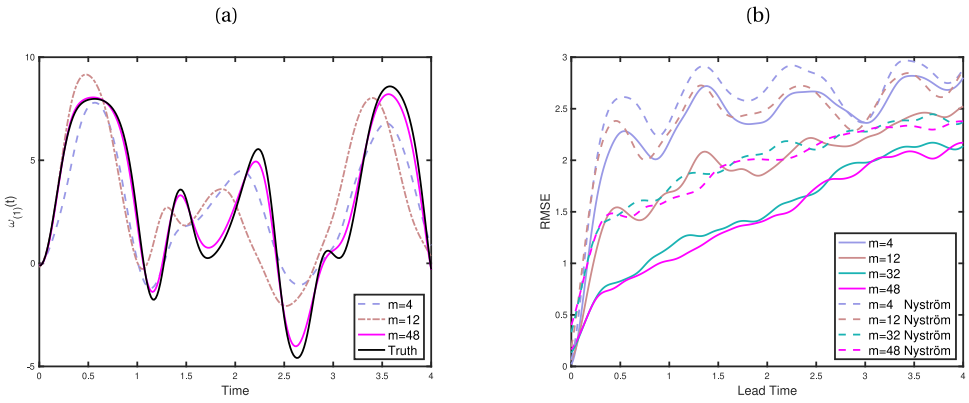

Fig. 4. 5D Lorenz-96 example: (a) Predicting a particular out-of-sample trajectory of the 5-d Lorenz-96 system with $F = 8$ via the kernel smoothing estimate of the conditional expectation using $m = 4$, 12, and 48 past observations. The training and testing data were both drawn from the invariant distribution. (b) The empirical RMSE of the kernel smoothing estimates with m past observations, calculated over 10,000 out-of-sample points. (c) The Empirical RMSE of both the Nyström and the kernel smoothing estimate for $m = 4$, 12, 32 and 48.

$\mathbb{E}_{L,N}[X_{t+k}|\mathbf{z}^{out}_{t,m}]$, where $k = 1, \ldots, m$ and use this as an estimator for x^{out}_{t+k}. To smooth a long sequence of noisy observations, we independently apply the constructed conditional expectation on each t and the corresponding $\mathbf{z}^{out}_{t,m}$ sequence of the trajectory. In all of the following numerical experiments, the training data $\mathbb{Z}_m$ is constructed by taking sequential m observations of $\omega_{(1)}(t)$ and adding θ_t to each of these elements.

In the first experiment, we use $N = 12{,}000$ observations of z_t with $\theta_t \sim \mathcal{N}(0, 4)$ and construct the conditional expectation estimators, $\mathbb{E}_{L,N}[Z_k|\mathbb{Z}_m]$, with $m = 5$ and $k = 1, \ldots, 5$ using $L = 120$ eigenfunctions. The observation time step for z_t is $\Delta t = 0.1$. We evaluate this smoothing operator on an out-of-sample trajectory of length $N_{out} = 10{,}000$, corrupted by four noise types with variance approximately 4 : (1) Gaussian noises $\mathcal{N}(0, 4)$, (2) Student's-t noises with 8/3 degrees of freedom, (3) Uniformly distributed noises over $(-\sqrt{48}/2, \sqrt{48}/2)$, and (4) Time varying noises of the type $2\sin(tU)$, where U is uniformly distributed over $[-1/2, 1/2]$. In each of the following experiments, the basis is constructed using data corrupted by $\mathcal{N}(0, 4)$ noise. This choice of standard deviation, 2, is roughly 25% of the climatological standard deviation, 7.9246. Table 1 shows the RMSEs of the smoothers $\mathbb{E}[Z_k|\mathbb{Z}_5]$ when the observed component is corrupted by $\mathcal{N}(0, 4)$ noise for $1 \le k \le 5$. From Table 1, one can see that the smallest error is obtained for $k = 2$. Table 2 shows the RMSEs of the smoother $\mathbb{E}[Z_2 \mid \mathbb{Z}_5]$ in the cases when the observed component is corrupted by the four noise types mentioned previously. In Fig. 5, we show the smoothed trajectories compared to the truth and noisy observations. We should point out that the $k = 2$ smoother forces us to discard the first and the last $N_{out} - (m - 2)$ observations in the trajectory. Note that the RMSEs shown in each of these tables are the errors in recovering $\omega_{(1)}$, computed by averaging the errors of time indices k to $N_{out} - (m - k)$.

As seen in Table 2, the smoother, $\mathbb{E}_{L,N}[Z_2|\mathbb{Z}_5]$ performs better than an ensemble Kalman filter with 64 ensemble members constrained to observing the same one-dimensional noisy component $\omega_{(1)}$ used for training the smoother (denoted by 1 observation in the table). We also found that the proposed smoother is more accurate than 4D-Var constrained to using only one observation ($\omega_{(1)}$ only) or two observations (both $\omega_{(1)}(t)$ and $\omega_{(2)}(t)$). In this table, for diagnostic purpose, we also report the RMSEs of EnKF and 4D-Var when noisy observations of all three components are available. One can see that, in this case, 4D-Var is superior; however, when only one component is observed, the proposed smoother (which requires no knowledge of the dynamics) is more accurate than both EnKF and 4D-Var. We

Table 1
RMSEs of the smoother for the $\omega_{(1)}$ component of the Lorenz-63 subjected to i.i.d $\mathcal{N}(0, 4)$ noise. The smoother $\mathbb{E}_{L,N}[Z_k|\mathbb{Z}_m]$ was constructed using $N = 12{,}000$ training data points, consisting of $m = 5$ sequential observations and $L = 120$ eigenfunctions. Each RMSE is averaged over out-of-sample data points with indices k to $N_{out-(m-k)}$, where $N_{out} = 10{,}000$ data points.

k	1	2	3	4	5
RMSE	1.3591	1.0663	1.1243	1.2394	1.4928

Table 2
RMSEs of the smoother for the $\omega_{(1)}$ component of the Lorenz-63 model subjected to: (1) Gaussian noises $\mathcal{N}(0, 4)$, (2) Student's-t noises with 8/3 degrees of freedom, (3) Uniformly distributed noises over $(-\sqrt{48}/2, \sqrt{48}/2)$, and (4) Time varying noise of the type $2\sin(tU)$, where U is uniformly distributed over $[-1/2, 1/2]$. The smoothing operator was constructed, as in Table 1, by using $N = 12{,}000$ training data points consisting of $m = 5$ sequential observations, corrupted by Gaussian noise and $L = 120$ eigenfunctions. The RMSEs were also calculated the same way as in Table 1. The same underlying trajectory was used for all of the RMSE calculations. The last six rows report benchmark results of applying the EnKF with 64 ensemble members and 4D-Var with $m = 5$, observing one to three noisy components, respectively.

	Gaussian	Student's-t	Uniform	Time varying
Smoother m = 5, k = 2	1.0663	1.1339	1.1406	1.1455
EnKF (1 obs)	1.3439	1.6589	1.4055	1.3704
EnKF(2 obs)	.7435	0.7871	.7282	.7538
EnKF(3 obs)	.6343	.7806	.6242	.6674
4D-VAR (1 obs)	2.0935	2.2131	2.3187	2.3145
4D-VAR (2 obs)	1.4971	1.4099	1.6088	1.7032
4D-VAR (3 obs)	.5436	.6437	.5198	.7785

should point out that in these numerical experiments the 4D-Var is implemented with $m = 5$ so that the configuration is similar to that of the non-parametric smoother.

Finally, we note here that if we instead use the kernel smoothing estimator described in Section 2.2 to approximate the smoother $\mathbb{E}[Z_k \mid \mathbf{Z}_m]$, we find that the RMSE of smoothing the same observations with Gaussian noise (as in Table 2) is 1.5605. Therefore, while the smoothed time series are less noisy than the observed data, the kernel smoothing based smoother performs worse than the EnKF with only 1 observation.

Smoothing noisy observations of the Lorenz-96 system: We consider the proposed smoother for $k = 2$ to estimate the first component $\omega_{(1)}(t)$ of the 40-dimensional Lorenz-96 system (22) with $F = 8$. In this numerical experiment, the training data consist of observations of $\omega_{(1)}(t)$ perturbed by Gaussian noise,

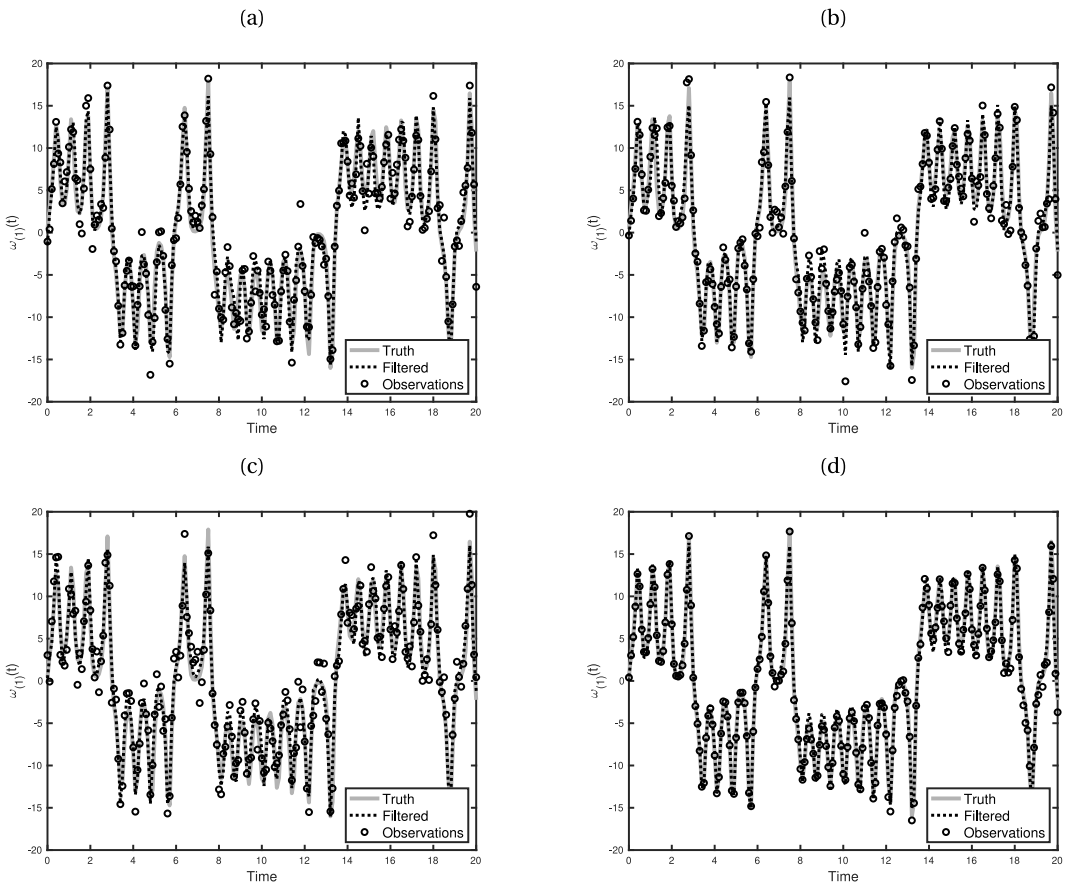

Fig. 5. Smoothed trajectories of $\omega_{(1)}$, compared to the truth and noisy observations. Here, the same smoother, constructed using data corrupted by Gaussian noise $\mathcal{N}(0, 4)$, was used to recover out-of-sample data corrupted by: (a) Gaussian noises $\mathcal{N}(0, 4)$; (b) Student's-t noises with 8/3 degrees of freedom; (c) Uniformly distributed noises over $(-\sqrt{48}/2, \sqrt{48}/2)$; and (d) Time varying noises of the type $2\sin(tU)$, where U is uniformly distributed over $[-1/2, 1/2]$. See Tables 1 and 2 for the RMSEs.

$\mathcal{N}(0, 1)$ but, as in Lorenz-63 example, we apply the filter to out-of-sample trajectories with various noise types with variance close to 1. This choice of noise standard deviation, 1, is roughly 25% of the climatological standard deviation, 3.5868.

For brevity, we only report the results for $m = 6$, $k = 3$, $L = 200$, and $N = 12{,}000$ training data, which we verified on three out-of-sample trajectories, each of length $N_{out} = 10{,}000$ sequential observations. These three trajectories were respectively generated by perturbing a single out-of-sample trajectory of length N_{out} with $\mathcal{N}(0, 1)$ noise, Student's-t noise with 10 degrees of freedom, and noise drawn uniformly from $[-1.8, 1.8]$. The training and testing time series were generated using RK4 with an observation time step of 0.05. See Table 3 for the RMSEs as well as benchmark results of applying EnKF and 4D-VAR in the same noise regimes. Note that the proposed smoother, $\mathbb{E}_{L,N}[Z_3 \mid \mathbb{Z}_6]$, performs better than both EnKF and 4D-Var even when the these two schemes were allowed to observe 30 noisy components.

Finally, we consider a more chaotic regime with $F = 16$. We construct an estimator to smooth observations of $\omega_1(t)$ corrupted with Gaussian noise, $\mathcal{N}(0, 1)$. The smoother is trained with parameters $L = 300$, $N = 20{,}000$ and $m = 6$ using a set of training data corrupted by Gaussian noise, $\mathcal{N}(0, 1)$. The RMSE is computed over $N_{out} = 10{,}000$ verification data which was generated by perturbing the true $\omega_{(1)}$ component of a trajectory by Gaussian noise, $\mathcal{N}(0, 1)$.

In Table 4, we report the RMSEs of the smoothing estimates using various choices of $1 \le k \le 6$. Fig. 6 shows the results of the smoother with $k = 3$. As a benchmark, we report the results of applying EnKF with 64 ensemble members and 4D-Var with $m = 6$, observing 10, 30 and 40 components of the system in Table 5. Note that the estimator, which is solely constructed using noisy data $\omega_{(1)}$, with $k = 3$ performs better than both the EnKF and 4D-Var estimates observing 30 noisy components.

Table 3
RMSEs of the smoother for the $\omega_{(1)}$ component of the Lorenz-96, with $F = 8$, subjected to: (1) Gaussian noise $\mathcal{N}(0, 1)$; (2) Student's-t noise with 10 degrees of freedom; (3) uniform noise drawn from $[-1.8, 1.8]$. The smoothing operator for $k = 3$ was constructed using $N = 12{,}000$ training data points consisting of $m = 6$ sequential observations constructed by Gaussian noise and $L = 200$ eigenfunctions. The RMSEs are calculated using the same underlying trajectory of length $N_{out} = 10{,}000$. The last six rows report benchmark results of applying the EnKF with 64 ensemble members and 4D-Var with $m = 6$, observing $10-40$ noisy components, respectively.

	Gaussian	Student's-t	Uniform
Smoother $m = 6, k = 3$	.5958	.6100	.5968
EnKF (10 obs)	.7234	0.7387	.7151
EnKF (30 obs)	.6229	.6785	.6416
EnKF (40 obs)	.2492	.2428	.2211
4D-VAR (10 obs)	2.4722	2.5051	2.4691
4D-VAR (30 obs)	2.0373	2.0764	2.0141
4D-VAR (40 obs)	.2933	.2983	.2822

4.2. Prediction with smoothed training data

In the preceding subsection, we introduced a smoother for denoising an out-of-sample sequence of noisy observations. In

Table 4
RMSEs for smoothing the $\omega_{(1)}$ component of Lorenz-96 with $F = 16$ subject to i.i.d $\mathcal{N}(0, 1)$ noise. The smoother is constructed using $L = 300$ eigenfunctions, estimated from $N = 20{,}000$ training data points consisting of $m = 6$ sequential observations. The RMSEs are calculated over a noisy trajectory consisting of 10,000 data points. The same out-of-sample trajectory was used for each of the RMSE calculations.

k	1	2	3	4	5	6
RMSE	.8768	.6688	.6543	.7321	.7946	.8935

Table 5
RMSEs for smoothing $\omega_{(1)}$ component of Lorenz-96 with $F = 16$ subject to i.i.d $\mathcal{N}(0, 1)$ noise. The configuration here is similar to that in Table 4.

No. Observation	10	30	40
EnKF	.90558	.8802	.35061
4D-Var $m = 6$	4.1632	4.052	.36414

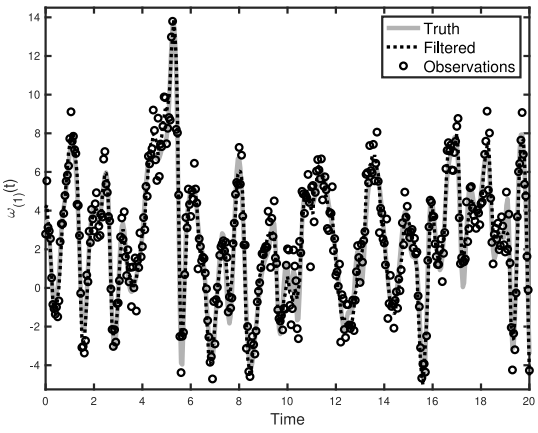

Fig. 6. Smoothed trajectories of $\omega_{(1)}$ compared to the truth and noisy observations. The smoother is constructed as described in Table 4 for $m = 6, k = 3$. The RMSE is .6543.

this subsection, we first denoise the noisy time series data using the smoother and then use the smoothed data as a surrogate for the true data to construct a kernel smoothing estimator as discussed in Section 2.2. In the following numerical result, we verify the prediction skill of employing these two steps given noisy time series observations of the five-dimensional Lorenz-96 example from Section 3.3 and compare it with the true prediction model.

The training data consists of $N = 20{,}000$ sequences, each consisting of sequential observations of the $\omega_{(1)}$ component of the 5D Lorenz-96 model observed at time steps of $\Delta t = 1/64$. However, unlike the example in Section 3.3, these observations are generated by perturbing the true $\omega_{(1)}(t)$ component by $\mathcal{N}(0, 1)$ noise. We also generate an additional $N_{out} = 10{,}000$ sequences as above for verification. We denoise both the noisy training and verification data using the smoother $\mathbb{E}_{L,N}[Z_k|\mathbb{Z}_{m_s}]$ with $k = 3$, $m_s = 6$ and $L = 120$. Here, we defined m_s in place of m to avoid the conflict of notation in the later discussion which refers to m as the memory length in the kernel smoothing estimates. The smoothed training data is then used to construct the kernel smoothing estimator described in Section 2.2. Subsequently, the kernel smoothing estimator is evaluated at each of the smoothed verification data. We should point out that while each smoothed verification data point required $m_s - k = 3$ future observations, the numerical experiments below, which evaluate the prediction skill beyond $3\Delta t = 3/64$ time units, are still valid prediction tests.

In Fig. 7(a), we show the kernel smoothing estimates for different memory length, m, and the true trajectory, all starting from a particular out-of-sample initial condition. While the prediction is less accurate than the one obtained from the kernel smoothing estimator constructed using noise-free data (see panel (c) for $m = 48$), one can still see that the prediction is somewhat improved as the memory length m is increased. This is also confirmed by the RMSEs plot in panel (b). In panel (d), we overlay the RMSEs in panel (b) with those corresponding to the prediction model constructed from the noise-free data (as in Fig. 4). While the model constructed from the noise-free data is more accurate, the discrepancy between the noise-free model and the noisy model decreases as m increases.

5. Summary

In this paper, we have studied aspects of forecasting and denoising of non-Markovian time series generated by partially observed dynamical systems. Within the context of kernel methods, where there is a mature theory on the consistency of empirical estimators, we have explored two distinct approaches, namely projection-based methods using the Nyström out-of-sample extension approach and smoothing methods using Markov integral operators. We refer collectively to these approaches as kernel analog forecasting (KAF) [21], since in many ways they can be viewed as kernel-based generalizations of the classical analog forecasting approach proposed by Lorenz in 1969 [25].

Previously, the consistency of KAF in the large-data limit was studied from the perspective of the Nyström approach [27], where it was shown that, under suitable ergodicity assumptions, the KAF estimator approximates the conditional expectation of observables of partially observed systems acted upon by the Koopman operator, thus yielding statistically optimal forecasts in the L^2 (root mean square error) sense. Here, we have shown that an analogous consistency property also holds if KAF is implemented using a one-parameter family of Markovian kernels in a limit of vanishing kernel bandwidth. The advantages of this smoothing approach over the Nyström estimator are that it is positivity-preserving, and avoids the need for a kernel eigendecomposition. The latter carries the risk that the number of eigenfunctions needed to approximate the conditional expectation is larger than what can be feasibly computed, both in terms of computational cost and statistical robustness. This practical limitation of the Nyström method is particularly prone to occur when the covariate space is high-dimensional, as we have demonstrated with numerical experiments involving a Hamiltonian system and the L96 system. In problems were the regression function for the response (predictand) projects well onto the leading kernel eigenfunctions, the Nyström method is still the method of choice, however.

Next, we studied the connection between KAF and the Mori–Zwanzig (MZ) framework for reduced dynamical modeling with memory. In particular, a major challenge in MZ approaches is to construct appropriate projection operators onto the covariate space (resolved dynamics), simplifying the structure of the memory kernel and rendering it amenable to approximation. We have argued that kernel methods provide a natural way of constructing improved projections by embedding the covariate data into a higher-dimensional space using delay-coordinate maps. This leads to a family of projections whose ranges are nested subspaces of the L^2 space associated with the invariant measure, increasing with the number of delays, and recovering the whole of L^2 using finitely-many delays. Correspondingly, the MZ equation in this limit involves only a Markovian term, which was estimated here nonparametrically.

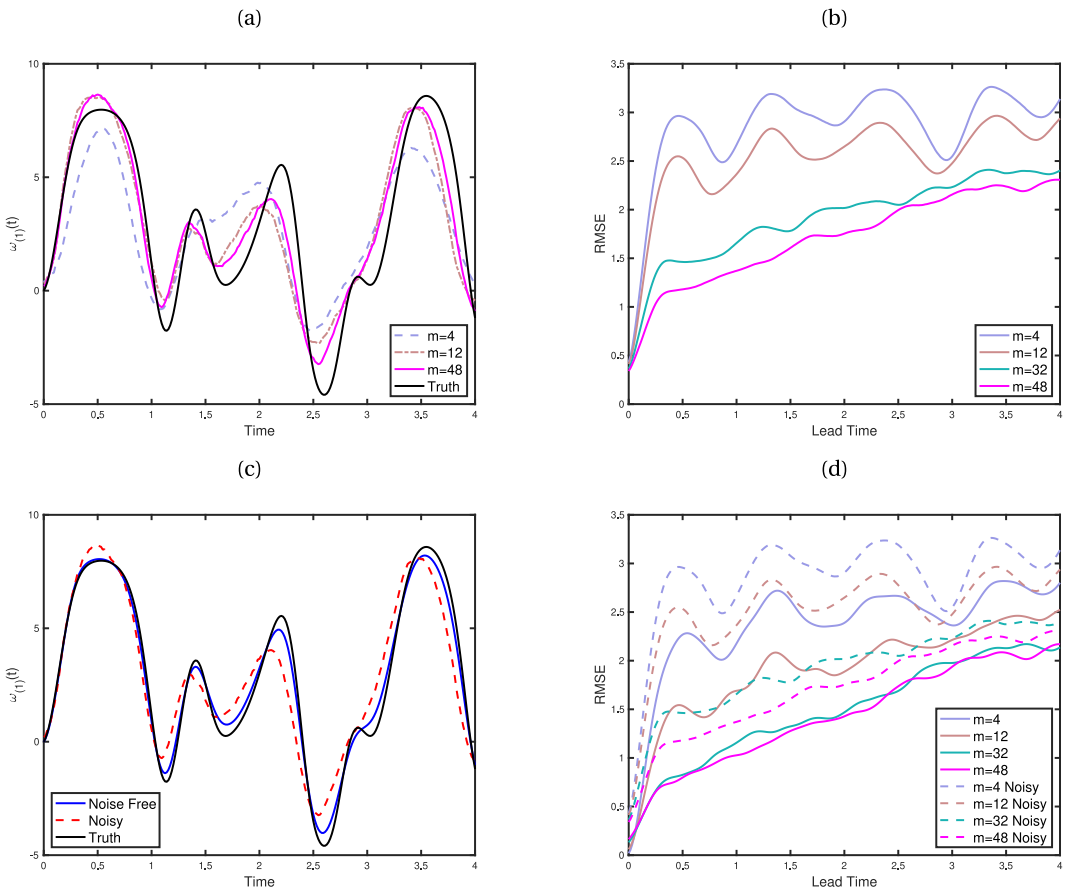

Fig. 7. Result of applying the kernel smoothing estimate on noisy observations of the first component of the 5D Lorenz-96 model with $F = 8$. The observations were first denoised using the nonparametric smoother and subsequently used for training the kernel smoothing estimator discussed in Section 2.2. (a) Trajectory of a particular out-of-sample prediction using the smoothed data for various memory lengths; (b) The corresponding RMSEs as functions of lead time for the kernel smoothing estimator constructed using smoothed observations; (c) A comparison of the predicted trajectory against that constructed from the corresponding noise-free data set for $m = 48$; (d) The RMSEs in panel (b) overlaid with the corresponding RMSEs obtained from noise-free data experiment (exactly the RMSEs in Fig. 4(b)).

Our third topic of study was smoothing and predicting with noisy data. We proposed a scheme whereby delay-coordinate maps are used to obtain high-quality kernel eigenfunctions from noisy data, which are then used for denoising via subspace projection. Using again the L96 model as a testbed, we demonstrated that this nonparametric denoising scheme oftentimes outperforms classical state-estimation methods such as the ensemble Kalman filter and the 4D-VAR approach. Once denoised, the data can be used to train skillful forecast models via the Nyström or smoothing formulations of KAF.

Possible directions for future research include data-informed methods for kernel design that bias the leading eigenspaces of the corresponding integral operators such that they capture the response variable with minimal loss, thus improving the performance of Nyström-based forecasting in high-dimensional covariate spaces. In addition, the efficacy of delay-coordinate maps in improving prediction skill of non-Markovian time series motivates the development of non-parametric kernel-based methodologies to estimate individual terms in the MZ equation.

CRediT authorship contribution statement

Faheem Gilani: Conceptualization, Methodology, Investigation, Visualization, Software, Validation, Formal analysis, Writing - original draft. **Dimitrios Giannakis:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **John Harlim:** Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research of JH was partially supported under the NSF, USA grant DMS-1854299. DG acknowledges support from NSF, USA grant 1842538, NSF, USA grant DMS-1854383, and ONR, USA grant N00014-16-1-2649. The authors thank Jonathan Poterjoy for sharing the EnKF and 4D-Var codes which he is developing as part of his NSF CAREER award.

References

[1] Pantelis R.V.lachas, Wonmin Byeon, Zhong Y.W.an, Themistoklis P. Sapsis, Petros Koumoutsakos, Data-driven forecasting of high-dimensional chaotic systems with Long Short-Term Memory networks, Proc. R. Soc. A 474 (2213) (2018) 20170844.

[2] Chao Ma, Jianchun Wang, Weinan E, Model reduction with memory and the machine learning of dynamical systems, Commun. Comput. Phys. 25 (4) (2019) 947–962.

[3] Shaowu Pan, Karthik Duraisamy, Data-driven discovery of closure models, SIAM J. Appl. Dyn. Syst. 17 (4) (2018) 2381–2413.

[4] John Harlim, Shixiao W. Jiang, Senwei Liang, Haizhao Yang, Machine learning for prediction with missing dynamics, J. Comput. Phys. 428 (2021) 109922.
[5] Shixiao W. Jiang, John Harlim, Modeling of missing dynamical systems: Deriving parametric models using a nonparametric framework, Res. Math. Sci. 7 (3) (2020) 1–25.
[6] Floris Takens, Reconstruction theory and nonlinear time series analysis, in: H.W. Broer, B. Hasselblatt, F. Takens (Eds.), Handbook of Dynamical Systems, Vol. 3, Elsevier Science, 2010, pp. 345–377 (Chapter 7).
[7] A.J. Chorin, O.H. Hald, R. Kupferman, Optimal prediction with memory, Physica D 166 (3) (2002) 239–257.
[8] Heyrim Cho, Daniele Venturi, George E. Karniadakis, Numerical methods for high-dimensional probability density function equations, J. Comput. Phys. 305 (2016) 817–837.
[9] Eric J. Parish, Karthik Duraisamy, A dynamic subgrid scale model for large eddy simulations based on the Mori–Zwanzig formalism, J. Comput. Phys. 349 (2017) 154–175.
[10] Weiqi Chu, Xiantao Li, The Mori–Zwanzig formalism for the derivation of a fluctuating heat conduction model from molecular dynamics, Commun. Math. Sci. 217 (2) (2019) 539–563.
[11] Jacob Price, Panos Stinis, Renormalized reduced order models with memory for long time prediction, Multiscale Model. Simul. 17 (1) (2019) 68–91.
[12] R. Zwanzig, Statistical mechanics of irreversiblity, Lectures Theor. Phys. 3 (1961) 106–141.
[13] H. Mori, Transport, collective motion, and Brownian motion, Progr. Theoret. Phys. 33 (1965) 423–450.
[14] Ayoub Gouasmi, Eric J. Parish, Karthik Duraisamy, A priori estimation of memory effects in reduced-order models of nonlinear systems using the Mori–Zwanzig formalism, Proc. R. Soc. A 473 (2205) (2017) 20170385.
[15] Carmen Hijón, Pep Español, Eric Vanden-Eijnden, Rafael Delgado-Buscalioni, Mori–Zwanzig formalism as a practical computational tool, Faraday Discuss 144 (2010) 301–322.
[16] Minxin Chen, Xiantao Li, Chun Liu, Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules, J. Chem. Phys. 141 (2014) 064112.
[17] Jing Li, Panos Stinis, Mori–Zwanzig reduced models for uncertainty quantification, J. Comput. Dyn. 6 (1) (2019) 39–68.
[18] Yuanran Zhu, Daniele Venturi, Faber approximation of the Mori–Zwanzig equation, J. Comput. Phys. 372 (2018) 694–718.
[19] J. Harlim, X. Li, Parametric reduced models for the nonlinear Schrödinger equation, Phys. Rev. E. 91 (2015) 053306.
[20] Francis E.H. Tay, Lijuan Cao, Application of support vector machines in financial time series forecasting, Omega 29 (4) (2001) 309–317.
[21] Zhizhen Zhao, Dimitrios Giannakis, Analog forecasting with dynamics-adapted kernels, Nonlinearity 29 (9) (2016) 2888.
[22] Romeo Alexander, Zhizhen Zhao, Eniko Székely, Dimitrios Giannakis, Kernel analog forecasting of tropical intraseasonal oscillations, J. Atmos. Sci. 74 (4) (2017) 1321–1342.
[23] Darin Comeau, Zhizhen Zhao, Dimitrios Giannakis, Andrew J. Majda, Data-driven prediction strategies for low-frequency patterns of north pacific climate variability, Clim. Dynam. 48 (5–6) (2017) 1855–1872.
[24] Darin Comeau, Dimitrios Giannakis, Zhizhen Zhao, Andrew J. Majda, Predicting regional and pan-Arctic sea ice anomalies with kernel analog forecasting, Clim. Dynam. 52 (9–10) (2019) 5507–5525.
[25] Edward N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, J. Atmos. Sci. 26 (4) (1969) 636–646.
[26] Dmitry Burov, Dimitrios Giannakis, Krithika Manohar, Andrew Stuart, Kernel analog forecasting: Multiscale test problems, 2020, arXiv preprint arXiv:2005.06623.
[27] Romeo Alexander, Dimitrios Giannakis, Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques, 2019.
[28] T. Berry, J. Harlim, Variable bandwidth diffusion kernels, Appl. Comput. Harmon. Anal. 40 (2016) 68–96.
[29] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. 99 (1994) 10143–10162.
[30] A.C. Lorenc, Analysis methods for numerical weather prediction, Q. J. R. Meteorol. Soc. 112 (1986) 1177–1194.
[31] Ronald R. Coifman, Stéphane Lafon, Diffusion maps, Applied and computational harmonic analysis 21 (1) (2006) 5–30.
[32] Andreas Christmann, Ingo Steinwart, Support Vector Machines, Springer, 2008.
[33] U. von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, Ann. Statist. 26 (2) (2008) 555–586.
[34] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, Dejan Slepčev, Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator, Found. Comput. Math. (2019).
[35] Yu A. Davydov, Convergence of distributions generated by stationary stochastic processes, Theory Probab. Appl. 13 (4) (1968) 691–696.
[36] Hanyuan Hang, Ingo Steinwart, Fast learning from α-mixing observations, J. Multivariate Anal. 127 (2014) 184–199.
[37] Bruno Colbois, Ahmad El Soufi, Alessandro Savo, Eigenvalues of the Laplacian on a compact manifold with density, Comm. Anal. Geom. 23 (3) (2015) 639–670.
[38] Charles J. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Stat. (1982) 1040–1053.
[39] R. Zwanzig, Nonequilibrium Statistical Mechanics, Oxford University Press, 2001.
[40] Tim Sauer, James A. Yorke, Martin Casdagli, Embedology, J. Stat. Phys. 65 (3–4) (1991) 579–616.
[41] Eric Darve, Jose Solomon, Amirali Kia, Computing generalized Langevin equations and generalized Fokker–Planck equations, Proc. Natl. Acad. Sci. 106 (27) (2009) 10884–10889.
[42] Kevin K. Lin, Fei Lu, Data-driven model reduction, Wiener projections, and the Koopman–Mori–Zwanzig formalism, J. Comput. Phys. 424 (2021) 109864.
[43] Michael Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, 2017, arXiv preprint arXiv:1701.02434.
[44] E.N. Lorenz, Predictability - a problem partly solved, in: Proceedings on predictability, held at ECMWF on 4–8 September 1995, 1996, pp. 1–18.
[45] Georg A. Gottwald, A. Majda, A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks, Nonlinear Process. Geophys. 20 (5) (2013).
[46] Simo Särkkä, Bayesian Filtering and Smoothing, Vol. 3, Cambridge University Press, 2013.
[47] D. Giannakis, Data-driven spectral decomposition and forecasting of ergodic dynamical systems, Appl. Comput. Harmon. Anal. 62 (2) (2019) 338–396.
[48] E.N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (1963) 130–141.