SIAM J. Applied Dynamical Systems
Vol. 22, No. 3, pp. 2082–2122

Learning Theory for Dynamical Systems*

Tyrus Berry† and Suddhasattwa Das†

Abstract. The task of modeling and forecasting a dynamical system is one of the oldest problems, and it remains challenging. Broadly, this task has two subtasks: extracting the full dynamical information from a partial observation, and then explicitly learning the dynamics from this information. We present a mathematical framework in which the dynamical information is represented in the form of an embedding. The framework combines the two subtasks using the language of spaces, maps, and commutations. The framework also unifies two of the most common learning paradigms: delay-coordinates and reservoir computing. We use this framework as a platform for two other investigations of the reconstructed system, its dynamical stability and the growth of error under iterations. We show that these questions are deeply tied to more fundamental properties of the underlying system, i.e., the behavior of matrix cocycles over the base dynamics, its nonuniform hyperbolic behavior, and its decay of correlations. Thus, our framework bridges the gap between universally observed behavior of dynamics modeling and the spectral, differential, and ergodic properties intrinsic to the dynamics.

Key words. matrix cocycle, Lyapunov exponent, reservoir computing, delay-coordinates, mixing, direct forecast, iterative forecast

MSC codes. 37M99, 37N30, 37A20, 37D25

DOI. 10.1137/22M1516865

1. Introduction.

Many investigations of physical systems involve modeling and forecasting a dynamical system, in fields as diverse as climate sciences [71], traffic dynamics [22, 65], and epidemiology [62]. With the growth of computational power, many new techniques and paradigms of reconstructing a dynamical system have been proposed; we call this the *learning problem* for dynamics. Most of the common techniques seek to recreate a dynamical system by developing a conjugate or equivalent dynamical system, usually in a higher dimensional space. We present a theoretical framework which unifies these techniques. The framework, presented in the form of a commuting diagram (1.3) of maps and operators, helps to identify and distinguish between different conceptual components of this learning.

Figure 1 presents an outline of the paper. The primary requirement of all learning techniques is an embedding of the dynamics (Assumption 2), which may be explicit or implicit. Throughout the paper we shall use "embedding" in the topological/set-theoretic sense, rather than the differential topology sense. Thus by "embedding" we mean an injective map, and we do not require it to induce an injective map between tangent bundles. We show in section 2

*Received by the editors August 18, 2022; accepted for publication (in revised form) by J. Rubin February 23, 2023; published electronically August 8, 2023.
https://doi.org/10.1137/22M1516865
Funding: This research was supported by the NSF Sponsored Program fund 204839.
†Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030 USA (tyrus.berry@gmail.com, iamsuddhasattwa@gmail.com).

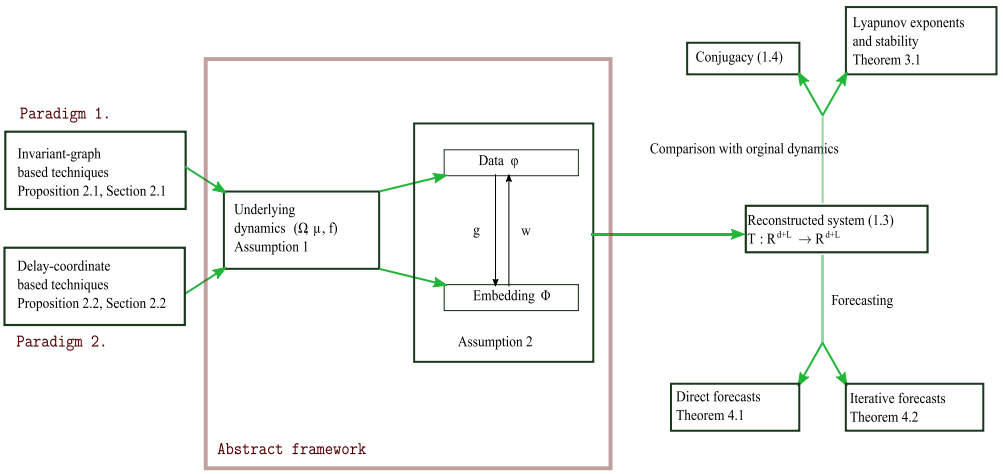


Figure 1. *Outline of results and theory.*

that there are two main paradigms of learning dynamics, based on invariant graphs and delay coordinates, in which the embedding is implicit and explicit, respectively. We unify both these paradigms in a common abstract mathematical framework (1.3) and show how the unknown dynamics can be reconstructed as a conjugate dynamical system (1.5) in the embedding space. We next investigate the stability of these reconstructions. We show that a proper quantitative assessment of the stability is related to the original dynamics, as well as the kind of interpolation done by the learning technique.

Another important consideration for us is the effectiveness of reconstruction models for the purpose of forecasting. The forecasting can be of two types: *direct* or *iterative*. We show in Theorem 4.1 that the direct method is limited by the rate of decay of correlations of the system. On the other hand, the iterative method is deeply connected to the embedding properties of the data, as well as the learning scheme employed. A key aspect of learning theory is the choice of a hypothesis space. This functional analytic consideration also fits seamlessly with our framework. We show how the rate at which the learned dynamics and the true dynamics diverge is a combination of the intrinsic dynamical properties as well as the effectiveness of the hypothesis space. We do so using the language of matrix cocycles. See Figure 2 for a comparative illustration of two computation techniques. We do an extensive comparison of various learning techniques in Tables 1 and 2.

The framework. We now build our general abstract framework by assigning mathematical objects and assumptions to various components of the entire prediction scheme. We begin with the dynamical system itself. A common practice is to assume an unknown dynamical system satisfying the following assumption.

Assumption 1. There is a C^1 dynamical system $f : \tilde{\Omega} \to \tilde{\Omega}$ on an m-dimensional C^1 manifold $\tilde{\Omega}$, with an ergodic measure μ with compact support Ω.

This minimal assumption on the underlying system allows it to be applied in many situations. The assumption of an ergodic invariant measure with compact support is fulfilled in any system with bounded trajectories. This assumption is, however, only for theoretical

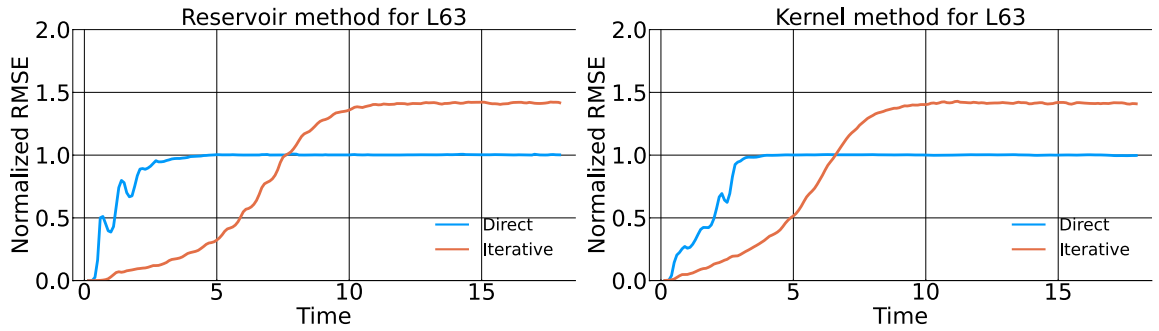


Figure 2. *Results of forecasting attempts on the Lorenz63 system, using the invariant-graph paradigm (section 2.1) on the left, and the delay-coordinate-based paradigm (section 2.2) on the right. For both these paradigms, we compare the performance of direct and iterative modes of forecasting (see (4.1), (4.2) for definitions). The horizontal axis shows the forecast time n, and the vertical axis is the root mean square (RMS) error of forecast as a function of n, for a signal of unit $L^2(\mu)$-norm. The RMS error is meant to approximate the $L^2(\mu)$-norm of the error as a function of the initial state of the underlying system. The iterative errors are seen to increase and eventually settle around $\sqrt{2}$, while the error from the direct mode settles at 1.0. We show that this is a universal behavior, based on a mathematical framework (1.3) that unifies both these paradigms, and both modes. Based on this framework, we develop Theorems 4.1 and 4.2, which provide expressions for the asymptotic behavior of these errors and are consistent with these graphs. Also see section 5 for an extended analysis.*

purposes, as the system in Assumption 1 is usually not presented explicitly. We next provide an abstract framework which describes how the system is manifested in a data-driven setting.

Assumption 2. There are maps $\phi : \Omega \to \mathbb{R}^d$ and $\Phi : \Omega \to \mathbb{R}^L$ such that

(i) Φ is an injective map;

(ii) there is a function $g : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^L$ such that $\Phi \circ f = g \circ (\phi \times \Phi)$.

The map ϕ is the measurement through which the dynamical system is observed. So the codomain of ϕ is often low-dimensional and may only be a partial observation the system. Since Φ is an injective map, it effectively serves as a representation of the dynamics space Ω in $\mathbb{R}^L$ space. The function g will be known and computable explicitly. Note that $\Phi \circ f$ is the evolution of Φ under one iteration of the dynamics of f. Thus g contains and encodes the evolution law, in terms of the current states of Φ and ϕ. This is summarized in the left half of the commutative diagram in (1.3). The function g is usually known explicitly, while Φ could be explicit (such as in delay-coordinate techniques; see section 2.2) or implicit (such as in invariant-graph-based techniques; see section 2.1). Let $\dim_\mu$ denote the box-dimension of the invariant set supported by the measure μ. We typically have

$$d \leq \dim_\mu \ll L.$$

The task now is to reconstruct the dynamics using the maps ϕ, Φ. This reconstruction will not be via the phase space $\tilde{\Omega}$ of the dynamics but rather through various function spaces which try to capture how these maps are transformed under the dynamics. For this reason, we shall use the operator theoretic language of dynamical systems. This is done using an operator-theoretic representation of the dynamics known as the *Koopman operator*.

Koopman operator. The Koopman operator [39, 19, 20] is a time-shift operator that acts on observables by composition with the map f. The space $L^2(\Omega,\mu)$ of square-integrable, complex-valued functions will be called our *space of observables*. The space $L^2(\Omega,\mu)$ can be unambiguously abbreviated as $L^2(\mu)$. For every $n\in\mathbb{N}$, the operator $U^n : L^2(\mu)\to L^2(\mu)$ is defined on every $h\in L^2(\mu)$ as

$$(U^n h) : \omega \mapsto (h\circ f^n)(\omega) \qquad \text{for } \mu\text{-a.e. } \omega\in\Omega.$$

The operator U and all its iterates U^n are unitary maps. The constant function 1_Ω is always an eigenfunction for U, with eigenvalue 1. A consequence of μ being ergodic is that 1 is a simple eigenvalue. Let $\mathcal{D}$ denote the closure in $L^2(\mu)$ of all eigenfunctions of U. Then one has the orthogonal decomposition

$$L^2(\mu) = \mathcal{D}\oplus\mathcal{D}^\perp. \tag{1.1}$$

The system is said to be *mixing* if the space $\mathcal{D}$ consists of only constant functions. Koopman operators allow the study of arbitrary nonlinear dynamics as linear dynamics on infinite-dimensional vector spaces. It has been used not only for forecasting tasks (see, e.g., [62]), but also in tasks such as harmonic analysis of dynamics generated data (see, e.g., [20]), control [50], and detection of coherent patterns (see, e.g., [19, 33]).

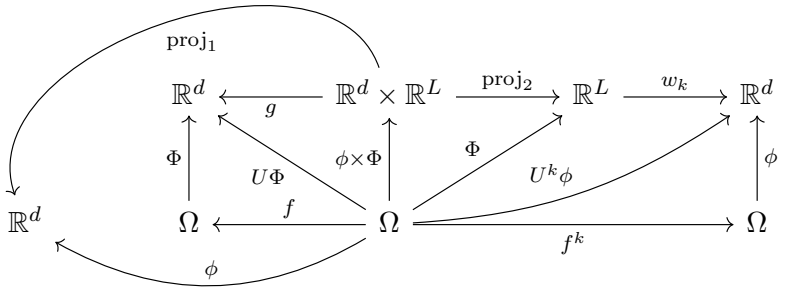
We are now prepared to present the reconstructed dynamics in terms of the Koopman operator.

Feedback function. Since Φ is an injective map (by Assumption 2), the current state of Φ determines the current and all future states in Ω and therefore of ϕ. Therefore for every $k\in\mathbb{N}$ there is a function w_k such that

$$w_k : \mathbb{R}^L\to\mathbb{R}^d, \quad w_k\circ\Phi = U^k\phi = \phi\circ f^k. \tag{1.2}$$

The learning task is to learn this map w_k. The following diagram connects all the maps and spaces we have discussed so far:

(1.3)



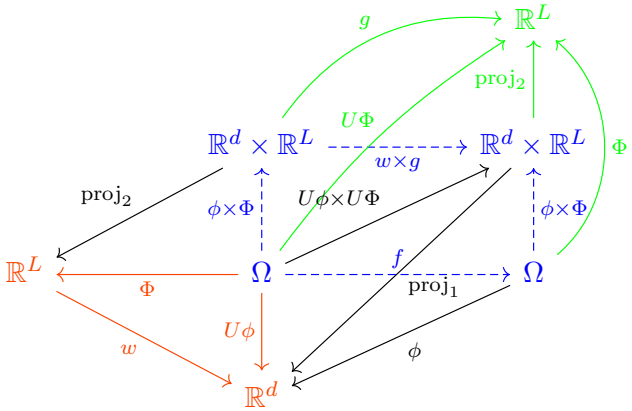
We next look at the specific case when $k=1$.

The reconstructed system. When $k=1$, w_k will be denoted as w. The following standalone or reconstructed dynamical system on $\mathbb{R}^d\times\mathbb{R}^L$ is conjugate to the dynamics on the attractor in Ω:

$$\mathcal{T} : \mathbb{R}^d\times\mathbb{R}^L\to\mathbb{R}^d\times\mathbb{R}^L, \quad \begin{bmatrix} u_{n+1}\\ y_{n+1}\end{bmatrix} = \mathcal{T}\begin{bmatrix} u_n\\ y_n\end{bmatrix} = \begin{bmatrix} w(y_n)\\ g(u_n,y_n)\end{bmatrix}. \tag{1.4}$$

We explore this conjugation further using the commutation diagram in (1.5). The blue loop reveals the conjugacy, the green loops are the result of Assumption 2, and the red loop is the result of (1.2).

(1.5)



The system (1.4) evolves in parallel to $f : \Omega \to \Omega$, while maintaining a conjugacy via the map $\phi \times \Phi$. The state vector $z_n := (u_n, y_n)$ is interpreted as follows: u_n represents the value $\phi(\omega_n)$, which is the unknown state ω_n observed through ϕ. $y_n \in \mathbb{R}^L$ is the embedded point $\Phi(\omega_n)$ in $\mathbb{R}^L$. With this interpretation in mind, (1.4) is to be initialized with the state vector

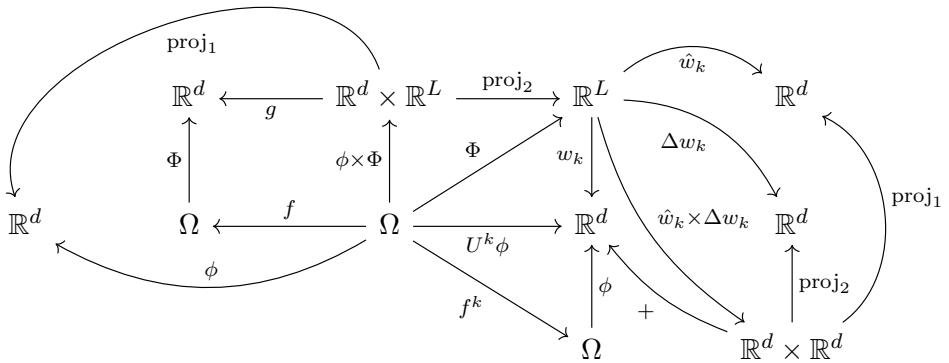
$$z_0 = z_0(\omega_0) := (\phi(\omega_0), \Phi(\omega_0)) \in \mathbb{R}^d \times \mathbb{R}^L \tag{1.6}$$

for some $\omega_0 \in \Omega$. The only component of (1.5) that needs to be learned is the feedback w. A crucial issue in the learning theory for dynamics is the learnability or injectivity of Φ. These have been proved theoretically or empirically in various situations. We have made the injectivity of Φ a part of our ground assumption and shifted the focus to the learning of w. One of our main focuses is the situation when w cannot be determined exactly but can only be approximated. In that case, one needs to consider a hypothesis space, and the error built into the model during learning.

Hypothesis space. In a practical situation, w_k is estimated from a search-set or hypothesis space $\mathcal{H}$ (see, e.g., [1]), which may be a linear subspace or nonlinear collection of functions. Thus the true function w_k can be expressed as

$$w_k = \hat{w}_k + \Delta w_k, \tag{1.7}$$

where $\hat{w}_k$ is the estimated function, and Δw_k is the error. Thus (1.3) can be rewritten as



Similarly to (1.4), the dynamics under the approximated feedback function becomes

$$\hat{\mathcal{T}}:\mathbb{R}^d\times\mathbb{R}^L\to\mathbb{R}^d\times\mathbb{R}^L,\quad \begin{bmatrix}u_{n+1}\\ y_{n+1}\end{bmatrix}=\hat{\mathcal{T}}\begin{bmatrix}u_n\\ y_n\end{bmatrix}=\begin{bmatrix}\hat{w}(y_n)\\ g(u_n,y_n)\end{bmatrix}. \tag{1.8}$$

The difference between (1.4) and (1.8) is the basis for evaluating the error of forecasts. We show in section 4 that the asymptotic rate at which these two systems diverge could depend on both the spectral properties of the dynamics and its Lyapunov exponents. This completes the description of our framework.

Outline. The rest of the paper is organized as follow. In section 2 we explore how the two main paradigms of learning dynamical systems fall under our framework. In section 3 we use (1.3) to derive stability results for the reconstructed system. In section 4, we again use (1.3) to obtain the rate of growth of errors when performing forecasts with the reconstructed system. We illustrate our results with some simple numerical examples in section 5. Sections 6, 7, 8, 9, 10, and 11 contain the proofs of our theorems.

2. Two paradigms of learning.

We now examine the two most important paradigms for realizing the scheme in (1.3): delay-coordinates and invariant-graph/echo-state network based techniques. We show that they follow the framework described in Assumptions 1 and 2, and how their implementations are special cases of (1.7), (1.4), and (1.6). Table 1 summarizes some features of these two techniques. Note that the questions of producing, constructing, or computing Φ and g are separate from the question of constructing w_k and $\hat{w}_k$. Table 2 gives an overview of various techniques used. We should note that connections between the delay-coordinates paradigm and invariant-graph paradigm have been explored recently from the viewpoint of *generalized synchronization* in [36, 44, 45]. Our investigation is motivated by the invariant-graph approach of Stark [73], and connections to generalized synchronization is a direction of future research.

2.1. Paradigm I: Invariant graphs.

Let $g:\mathbb{R}^d\times\mathbb{R}^L$ be a smooth map for which there is constant $\lambda\in(0,1)$ such that

$$\|\nabla_y g(u,y)\|\le\lambda\quad\forall u\in\mathbb{R}^d,\, y\in\mathbb{R}^L. \tag{2.1}$$

Using this g, one can build a *reservoir* system, which is a skew-product system on $\Omega\times\mathbb{R}^L$ defined as

$$\begin{pmatrix}\omega_{n+1}\\ y_{n+1}\end{pmatrix}:=T_{\text{reservoir}}\begin{pmatrix}\omega_n\\ y_n\end{pmatrix}:=\begin{pmatrix}f(\omega_n)\\ g(\phi(\omega),y_n)\end{pmatrix}. \tag{2.2}$$

Table 1
The two learning paradigms satisfying Assumptions 1 and 2 and the scheme in (1.3).

Name	Φ	Basis for convergence of Φ	g
Invariant graphs	Implicitly obtained (2.3)	(2.4)	Explicit: (2.1)
Delay-coordinates	Explicitly obtained as basis functions	Ergodic convergence	Explicit: (2.7)

Table 2
Various learning techniques, applicable to learning w_k (1.2).

Technique	Hypothesis space	Advantages	Disadvantages
Linear reg.	Linear combination of fixed basis functions or coordinates	Availability of techniques for linear cases	Poor fit for nonlinear functions
Kernel reg. [9, 8]	$C^r(M)$ or $L^2(\mu)$ Spaces spanned by kernel sections or eigenvectors	Allows smooth interpolations and connections with underlying geometry	Localized nature of basis functions require large number of basis functions
RKHS [1, 20, 21]	Span of eigenfunctions of kernel integral operators	Completely data-driven, allows out of sample extension	Inexplicit, unspecified nature of basis functions
Nonlinear reg. [30]	Parameterized space of functions	Dependence on parameters allow application of manifold techniques such as gradient descent	Explicit knowledge of parameters as well as dependence on parameters required
Deep NNs [52]	Functions parameterized by network activation and coupling parameters	Simplicity of implementation; scalability; explicit dependence on parameters known	Little a priori knowledge known about dimension of layers or number of layers; little knowledge about convergence rate of learning; huge number of variables to optimize
LSTM [43, 59, 61]	Same as Deep NNs but with additional memory cells	Good for approximating functions which have sparse dependence over a long interval of time	Same as Deep NNs; more parameters to tune
Radial basis functions [72]	Similar to kernel techniques	Provides a global representation of the map	Lack of normalization lead to nonuniformity in predictability
Local approximation techniques, such as simplex methods [48], and local linear regression [51]	Nearest-neighbor based approximation of a neighborhood of the predictee point	Good approximation for low-curvature attractors, i.e., less oscillatory functions	Predictee point needs to be close to data cloud, feedback function unbounded.

The paradigm of invariant graphs was first investigated in [60, 47] and was studied eventually in greater detail as *echo-state networks* (see, e.g., [38, 36, 58, 34]) and *reservoir computers* [37, 57]. It is popular due to the simplicity of its construction and ease of use in learning problems. They are known for their robust performance in prediction [31, 12] but also for recovering other properties such as Lyapunov exponents [63]. A particular instance of g above introduced in [57] is

$$g(u, y) = \tanh\left(W_{in}u + W_Y y + v_{bias}\right),$$

where W_{in}, W_Y are random matrices of dimensions $L \times d$, $L \times L$, respectively, v_{bias} is a random vector of dimension L, and $\|W_Y\| \leq \lambda < 1$.

Although (2.2) involves the underlying dynamical map f, the actual knowledge of f is not needed. Note that the dynamics in the y-coordinate is linked to the ω-coordinate through the measurement ϕ. In the training phase, one provides as input the measurements $\{\phi(\omega_n)\}_{n=0}^{N}$. Thus (Ω, f) remains unknown but continues to drive the reservoir variable y. The variable y settles down into a representation of the attractor, in a manner which we make precise below.

Proposition 2.1. *Let Assumption 1 hold, let $\phi : \Omega \to \mathbb{R}^d$ be a continuous map, and let $g : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^L$ be a C^1 map satisfying (2.1). Then the following hold:*

(i) *There is a map $\Phi : \Omega \to \mathbb{R}^L$ such that Assumption 2 (ii) is satisfied.*

(ii) *The graph of Φ is invariant, i.e.,*

$$(U^n\Phi)(\omega) := \Phi(f^n(\omega)) = \mathrm{proj}_Y T^n_{\mathrm{reservoir}}(\omega, \Phi(\omega)) \quad \forall n \in \mathbb{N}, \quad \forall \omega \in \Omega. \tag{2.3}$$

(iii) *The graph of Φ is globally attracting, i.e.,*

$$\lim_{n\to\infty} (U^n\Phi)(\omega) = \lim_{n\to\infty} \mathrm{proj}_Y T^n(x, y) \quad \forall x \in \bar{\Omega}, y \in Y. \tag{2.4}$$

(iv) *If $\left\| \frac{\partial}{\partial_u} g \right\| \le 1$, then Assumption 3 is also satisfied.*

Assumption 3 is an additional assumption requiring that g be a nonexpansive map in each of its variables. It is described in section 3 and is used to establish stability properties. Proposition 2.1 is proved in section 8. Parts (i)–(iii) are immediate consequences of results by Stark [73] or by Grigoryeva, Hart, and Ortega [36, Thm. III.1]. We have put Stark's results along with the other paradigms in the common, general framework of (1.2).

The invariant graph property leads to a fulfillment of the identity in Assumption 2 (ii). However, the injectivity condition of Assumption 2 (i) remains to be proven rigorously. It has been generally observed that for L large enough, Φ is also injective. Ground Assumption 1 is assumed while running the system. Note that the embedding Φ is not obtained explicitly but implicitly through the state variables of the network. Given any arbitrary initialization to (2.2), by (2.4), the internal states of the reservoir converge to an invariant graph over Ω. The function Φ is precisely the function whose graph is invariant. Although it will remain indeterminate, its values over a dynamic trajectory, i.e., the values $\phi(f^n\omega_0)$, will be obtained for some unknown initial point ω_0.

Long short-term memory (LSTM). Long short-term memory (LSTM) networks [46, 61] are networks of units in which each unit is a skew-product system, usually much smaller in size than a reservoir network, and without the contraction requirement of a reservoir network. Each unit has internal states $y_n = (h_n, c_n)$, which is updated with the help of an additional input x_n which could originate from an external dynamical system. The functional equation is

$$y_n = (h_n, c_n) = G(x_n, h_{n-1}, c_{n-1}) = G(x_n, y_{n-1}), \quad n = 0, 1, 2, \ldots. \tag{2.5}$$

The variables h_n, c_n denote, respectively, a hidden state vector and a cell input activation vector to the LSTM unit. The units in the LSTM network can also be cascaded to each other Suppose there are Q LSTM units. Let us denote their internal states at time n as $y_n^{(1)}, \ldots, y_n^{(Q)}$. Due to the cascaded structure, we have

$$y_n^{(q)} = G\left(x_{n+q-1}, y_{n-1}^{(q-1)}\right), \quad n \in \mathbb{N}_0, q \in \{1, \ldots, Q\}. \tag{2.6}$$

Here $y_n^{(0)}$ is the constant sequence equal to zero. LSTMs implement delay-coordinates due to their full dependence on a delay-coordinate set $(x_n, \dots, x_{n+Q-1})$ for each $n \in \mathbb{N}_0$. In addition, this delay-coordinated input is fed into a skew-product system whose internal structure is block diagonal. Note that LSTM networks require tuning and do not automatically satisfy the contraction property in Assumption 3. It is an interesting question whether the tuning procedure with data from an ergodic trajectory leads to this criterion being met. Alternatively, optimization methods in learning dynamics could be modified to enforce additional constraints to satisfy Assumption 3. These are interesting directions of future work.

2.2. Paradigm II: Delay coordinates. An effective and numerically inexpensive means of obtaining an embedding of a dynamical system is using delay-coordinates [7, 9]. To relate to our framework, fix a number of delays $Q \in \mathbb{N}$ and set

$$L = Qd, \quad \Phi : \Omega \to \mathbb{R}^L, \Phi : \omega \mapsto \begin{bmatrix} \phi(\omega) \\ \vdots \\ \phi\left(f^{Q-1}\omega\right) \end{bmatrix}, \tag{2.7}$$

$$g : \mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^L, g : u \times \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(Q)} \end{bmatrix} \mapsto \begin{bmatrix} u \\ y^{(1)} \\ \vdots \\ y^{(Q-1)} \end{bmatrix}.$$

Using this paradigm, the reconstructed dynamics (1.4) becomes

$$\mathcal{T}_{\text{delay-coord}} : \mathbb{R}^{d \times dQ} \to \mathbb{R}^{d \times dQ} := \begin{bmatrix} u \\ y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(Q)} \end{bmatrix} \mapsto \begin{bmatrix} w\left(y^{(1)}, \dots, y^{(Q)}\right) \\ y^{(1)} \\ \vdots \\ y^{(Q-1)} \end{bmatrix}.$$

Note that in this case, g is a linear map. We have the following.

Proposition 2.2. *Let Assumption 1 hold. Then for a typical map $\phi : \Omega \to \mathbb{R}^d$, if $Q \in \mathbb{N}$ is large enough, then Φ defined through (2.7) is an injective, and thus Assumption 2 is satisfied. Moreover, Assumption 3 is also satisfied.*

The proof is a direct consequence of the delay-coordinate embedding theorem [68].

Thus the most common techniques for reconstruction and forecasting fall into the framework we introduced in (1.3). See Figures 2 and 3 for an illustration of application of these two techniques. In section 5.2, we also briefly review some techniques which do not fall under the schemes of Assumption 2 and (1.3). In the next two sections, we shall analyze two important features of our scheme, their stability, and the accuracy of their predictions.

3. Stability of reconstructed system. One could ask whether the reconstructed system could attain conjugacy or near-conjugacy with the original dynamics (see, e.g., [14, 70, 69]). In our case, the conjugacy map $\phi \times \Phi$ exists by virtue of Assumptions 1 and 2, as shown in (1.5). So instead of learning or discovering the conjugacy, our focus is on the stability

of the conjugate dynamics (1.4). The image of $h := \phi \times \Phi$ is a bijective image of Ω and is invariant under the dynamics of $\mathcal{T}$ (1.4). But $\mathcal{T}$ acts in the higher-dimensional ambient space $\mathbb{R}^{L+d}$, and one needs to calculate the rate of deviation perturbations from $X := h(\Omega)$. We track these using Lyapunov exponents. Let the distinct Lyapunov exponents of (Ω, f, μ) be $\lambda_1 > \lambda_2 \cdots > \lambda_r$, with corresponding Oseledets splitting $T\Omega = E_1 \oplus \cdots \oplus E_r$. Since the dimension of $\tilde{\Omega}$ is m, the multiplicities of the λ_i sum to m. The E_i's corresponding to negative valued λ_i constitute the stable directions, whereas the E_i corresponding to positive-valued λ_i constitute the unstable directions. Moreover,

$$\lim_{n\to\infty} \frac{1}{n} \ln \|Df^n(\omega)v_i\| = \lambda_i, \quad \mu\text{-a.e. } \omega \in \Omega \ \forall v_i \in E_i(\omega)\backslash\{0\}.$$

In general, the map $\mathcal{T}$ will have $L+d$ Lyapunov exponents (counting multiplicities), whereas f has m of them. We shall show in Theorem 3.1 (i) that m of the $d+L$ Lyapunov exponents of $\mathcal{T}$ coincide with the original d coefficients. We are interested in these other $d+L-m$ Lyapunov exponents of the reconstructed systems, and their positions relative to $\lambda_1(f), \ldots, \lambda_d(f)$. These additional Lyapunov exponents have been labeled as *spurious Lyapunov exponents* (see, e.g., [67, 25]). They pose significant challenges in data-driven identification of true Lyapunov exponents.

By the very definition of Lyapunov exponents, the $\lambda_i(\mathcal{T})$ depend not only on the invariant set X but also on its neighborhood. This leads to a problem of ambiguity. An essential part of $\mathcal{T}$ is the feedback function w. The function $w : \mathbb{R}^L \to \mathbb{R}^d$ is defined uniquely only on X. The conjugacy in (1.5) will be preserved on Ω irrespective of the nature of the extension of w to a neighborhood of X. We define a collection

$$\mathfrak{S} := \left\{ \hat{w} \in C^1\left(\mathbb{R}^L; \mathbb{R}^d\right) \,:\, \hat{w}|_X = w|_X \right\},$$

equipped with the C^1-topology. Every $\hat{w} \in \mathfrak{S}$ is a C^1 function satisfying $\hat{w} \circ \Phi(\omega) = (U\phi)(\omega)$ for every $\omega \in \Omega$. Any choice of $\hat{w} \in \mathfrak{S}$ leads to a different dynamics in $\mathbb{R}^{L+d}$ as in (1.4), with X as an invariant ergodic set. Therefore the top Lyapunov exponent λ_1 of (1.4) will be a function of $\hat{w}$. Thus we can define a function

$$\lambda_1 : \mathfrak{S} \to \mathbb{R}, \quad \lambda_1(\bar{w}) := \lambda_1(\mathcal{T}).$$

Our goal will be to study how close $\lambda_1(\mathcal{T})$ can be made to $\lambda_1(f, \mu)$.

Stability gap. As pointed out in [25, 24], the top Lyapunov exponent of the reconstructed system may exceed that of the original system. Moreover, some of the additional Lyapunov exponents may be positive. All these contribute to additional instabilities being introduced into the system. The *stability gap* in the reconstruction of a dynamical system (Ω, μ, f) is defined to be

$$\text{stability gap} := \inf_{\bar{w}\in\mathfrak{S}} \lambda_1(\bar{w}) - \lambda_1(f, \mu).$$

The stability gap is always nonnegative, as will be shown in Theorem 3.1. We shall study ways to obtain a bound on the stability gap in our next theorem. We shall need two additional assumptions.

Assumption 3. The function g from Assumption 2 further satisfies

$$\sup_{\omega\in\Omega} \|\partial_1 g\|_{(\phi(\omega),\Phi(\omega))} \leq 1, \quad \sup_{\omega\in\Omega} \|\partial_2 g\|_{(\phi(\omega),\Phi(\omega))} \leq 1.$$

Our next assumption requires a retraction to the range of Φ. Let $\mathcal{U}$ be a neighborhood of ran Φ. Recall that a continuous map $\text{ret} : \mathcal{U} \to \text{ran } \Phi$ is said to be a *retract* if $\text{ret}|_{\text{ran }\Phi} = \text{Id}_{\text{ran }\Phi}$.

Assumption 4. There is a continuous retraction $\text{ret} : \mathcal{U} \to \text{ran } \Phi$ for some open neighborhood $\mathcal{U}$ of ran Φ in $\mathbb{R}^L$.

For a retraction map such as ret, we shall be interested in the Lipschitz norm of the retraction

$$\kappa_{\text{ret}} := \sup_{y\in\text{ran }\Phi} \limsup_{y'\to y} \frac{d(\text{ret}(y), \text{ret}(y'))}{d(y, y')}. \tag{3.1}$$

Finally, we also need the following function $C_{\phi,\Phi}$ that depends on the (fixed) functions ϕ, Φ and a point $\omega \in \Omega$:

$$C_{\phi,\Phi} : \Omega \to \mathbb{R}^+, \quad C_{\phi,\Phi}(\omega) := \sup \left\{ \frac{\|D\phi(\omega)v\|}{\|D\Phi(\omega)v\|} \, : \, v \in T_\omega\Omega \setminus \{0\} \right\}.$$

We shall use $C_{\phi,\Phi}$ to bound the gap between $\lambda_1(\mathcal{T})$ and $\lambda(f)$.

Theorem 3.1 (stability of reconstruction). *Let Assumptions 1 and 2 hold. Then the following hold:*

(i) *The $d + L$ Lyapunov exponents of $\mathcal{T}$ contains as a subset the m Lyapunov exponents of f.*

(ii) *$\lambda_1(\bar{w})$ is upper semicontinuous with respect to $\bar{w}$. In other words, for every $\epsilon > 0$, there is a C^1 neighborhood $\mathcal{U}$ of $\bar{w}$ such that*

$$\lambda_1(\bar{w'}) < \lambda_1(\bar{w}) + \epsilon \quad \forall \bar{w'} \in \mathcal{U}.$$

(iii) *Suppose Assumptions 3 and 4 also hold. Then the stability gap is bounded by*

$$\inf_{\bar{w}\in\mathfrak{S}} \lambda_1(\bar{w}) - \lambda_1(f, \mu) \leq \int \ln\left[1 + (1 + C_{\phi,\Phi}(\omega))\,\kappa_{\text{ret}}\right] d\mu(\omega). \tag{3.2}$$

Claims (i) and (ii) of Theorem 3.1 are immediate consequences of results from [24, 11, 76]. We review these and prove claim (iii) in section 9.

Remark 3.2 *(instability of the reconstructed system).* Claims (ii) and (iii) imply that, at least in theory, given any bound ϵ, there is a robust (i.e., open) set of $\bar{w}$ for which the instability is no more than $C + \epsilon$ of the original dynamics, for some constant C depending on the dynamics, ϕ and Φ alone. In practice, $\hat{w}$ is obtained from some hypothesis space which is determined by the application domain. In such situations there is no guarantee of the stability being preserved up to an ϵ error.

Remark 3.3 *(continuity of Lyapunov exponents).* Theorem 3.1 is related to the important question of continuity of Lyapunov exponents. In our case, we show that the growth of

error in the system is related to a $GL(2L)$-valued matrix cocycle over the base dynamics (f, μ, ω), described in detail in (4.8). These cocycles are dependent (in a C^1-sense) on $\bar{w}$. Thus a relevant question for us is the continuity of λ_1 for these cocycles as a function of $\bar{w}$. There has been various results in this direction, such as for i.i.d. matrix cocycles [29, 28], in terms of *large deviation–type parameters* [26, Thm. 1.6], and in terms of dominated splittings [11, Thm. 5]. See [76] for a broad overview of this extensive field of investigation. However, none of these various sets of assumptions applies to our situation, in which the cocycle family is parameterized by a set of functions $\mathfrak{S}$.

Assumption 4 is of a topological nature and would depend on the topological or geometrical properties of X. The following corollary applies to the use of a large number of delay-coordinates.

Corollary 3.4. *Let $\Psi^t : \Omega \to \Omega$ be a smooth flow and f be the time-Δt map $f = \Psi^{\Delta t}$. Let all the conditions in Assumptions 1 and 2 be met and the delay-coordinate paradigm (2.7) be implemented. Suppose further that there is a retraction map as in Assumption 4 for which the Lipschitz constant $\kappa_{\text{ret}} = 1$. Then there is a constant C_2 depending only on the flow such that $C_{\phi,\Phi}(\omega) \le \frac{1}{Q} + 0.5 C_2 Q \Delta t$ for every $\omega \in \Omega$. In particular,*

$$0 \le \inf_{\bar{w} \in \mathfrak{S}} \lambda_1(\bar{w}) - \lambda_1(f, \mu) \le \ln \left[2 + \frac{2}{Q} + C_2 Q \Delta t \right].$$

The criterion that $\kappa_{\text{ret}} = 1$ is attained, for example, when $X = \operatorname{ran} \Phi$ is a manifold, and ret is a tubular neighborhood retract. Corollary 3.4 is proved in section 9.4.

Next, we analyze the divergence of the dynamics of $\hat{\mathcal{T}}$ (1.8) from that of the perfect reconstruction $\mathcal{T}$ (1.4).

4. Forecasts with reconstructed system.

We shall now analyze the effectiveness of forecasts made using the scheme in (1.3), and its approximation as (1.8). As suggested by Casdagli [15], given a reconstruction paradigm, there are two ways of estimating the value $\phi(f^n \omega)$ after n iterations of the base dynamics: We can iterate (1.8) n times, and the first coordinate of z_0 will serve as an approximation of $\phi(f^n \omega) = (U^n \phi)(\omega)$. We call this the iterative method, and its accuracy can be estimated via

$$\begin{aligned} \text{error}_{\text{iter}}(n, \omega) &:= \left\| U^n \phi(\omega) - \text{proj}_1 \circ \hat{\mathcal{T}}^n \circ (\phi, \Phi)(\omega) \right\|_{\mathbb{R}^d}, \\ \text{error}_{\text{iter}}(n) &:= \left[\int_\Omega \text{error}_{\text{iter}}(n, \omega)^2 d\mu(\omega) \right]^{1/2}. \end{aligned} \tag{4.1}$$

Or we can directly approximate w_n via (1.7) and obtain a *direct* estimate. The corresponding errors are

$$\begin{aligned} \text{error}_{\text{direct}}(n, \omega) &:= \| U^n \phi(\omega) - \hat{w}_n \circ \Phi(\omega) \|_{\mathbb{R}^L}, \\ \text{error}_{\text{direct}}(n) &:= \| U^n \phi - \hat{w}_n \circ \Phi \|_{L^2(\mu)} = \left[\int \text{error}^2_{\text{direct}}(n, \omega) d\mu(\omega) \right]^{1/2}. \end{aligned} \tag{4.2}$$

To aid the discussion, we will make further assumptions on the nature of the hypothesis space $\mathcal{H}$.

Linear hypothesis space. Usually the hypothesis space $\mathcal{H}$ will be a finite-dimensional space, spanned by a basis $h_1, \dots, h_m$. In that case

$$\mathcal{W} := \text{span } \{h_i \circ \Phi_l : 1 \le i \le m, \le l \le L\} \tag{4.3}$$

is a finite subspace of $L^2(\mu)$, and

$$\hat{w}_k \circ \Phi = \text{proj}_{\mathcal{W}} U^k \phi. \tag{4.4}$$

For example, if the hypothesis space is restricted to $\mathcal{L}(\mathbb{R}^L; \mathbb{R}^d)$, then $\mathcal{W} = \text{span } \Phi$. In the rest of this paper, we shall focus on this scenario where the hypothesis space is linear. We state this formally in the following assumption.

Assumption 5. The hypothesis space $\mathcal{W}$ is a finite-dimensional subspace of $L^2(\mu)$ and contains the constant function $1_{\mathbb{R}^L}$.

In most learning techniques, a bias or offset constant is calculated separately, thus satisfying the criterion that $\mathcal{W}$ contains constant functions.

Let π denote the projection $\text{proj}_{\mathcal{W}}$, and set $\Delta := \text{Id} - \pi$. For ease of notation, we will denote $\hat{w}_1$ simply by $\hat{w}$ in the rest of this section. Define the *projection error* to be the quantity

$$\delta = \delta(\mathcal{H}) := \|\Delta U \phi\|_{L^2(\mu)}. \tag{4.5}$$

This is the component of the measurement ϕ not recoverable using our choice of hypothesis space. Note that as the size of the hypothesis space increases, δ converges to 0.

We shall first examine the performance of the direct forecast method. For this purpose, we shall utilize a natural splitting of the space $L^2(\mu)$ induced by the Koopman operator. Let $\mathcal{D}$ be the closure of the span of the eigenfunctions of the Koopman operator U, and let $\mathcal{D}^\perp$ be its orthogonal complement. Thus we have the orthogonal splitting

$$L^2(\mu) = \mathcal{D} \oplus \mathcal{D}^\perp.$$

The space $\mathcal{D}$ always contain the constant functions. For mixing systems such as the Lorenz63 attractor, $\mathcal{D}$ contains only the constant functions. For quasiperiodic dynamics such as the dynamics on Hamiltonian tori, $\mathcal{D}^\perp = \{0\}$. These two components $\mathcal{D}, \mathcal{D}^\perp$ not only have different ergodic properties [39, 23], but also respond differently to data-analytic and harmonic analytic tools [19, 20]. This splitting is also natural in the sense that it is invariant under the action of the Koopman operator. We now provide an estimate on the rate of growth of the direct error.

Theorem 4.1 (error from direct forecast). *Let Assumptions 1 and 2 hold, and assume the notation in (1.7), (1.4), and (1.6). Let δ be as in (4.5). Then the error from direct iteration is given by*

$$error_{direct}(n) = \|(\text{Id} - \pi) U^n \phi\|_{L^2(\mu)}.$$

Now assume that Assumption 5 holds. Then there is a subset $\mathbb{N}' \subseteq \mathbb{N}$ with density 1 such that the following hold:

(i) *For every $\epsilon > 0$, if the hypothesis space $\mathcal{W}$ is chosen large enough, then*

$$\lim_{n\in\mathbb{N}',n\to\infty} error_{direct}(n) = \|\phi - \mathrm{proj}_{\mathcal{D}}\phi\|_{L^2(\mu)} + \epsilon.$$

(ii) *If f is weakly mixing, then for every choice of $\mathcal{W}$*

$$\lim_{n\in\mathbb{N}',n\to\infty} error_{direct}(n) = var_\mu := \|\phi - \mu(\phi)\|_{L^2(\mu)}.$$

(iii) *If f is strongly mixing, the set $\mathbb{N}'$ can be taken to be the entire set $\mathbb{N}$.*

(iv) *If f has purely a discrete spectrum, then for every $\epsilon > 0$, if the hypothesis space $\mathcal{W}$ is chosen large enough, then*

$$error_{direct}(n) < \epsilon \quad \forall n \in \mathbb{N}.$$

Theorem 4.1 is proved in section 10. An important basis for claims (i), (ii) is the decay of correlations seen in (weakly) mixing systems. See Remark 4.4 for further discussions on this topic. We next study the performance of the iterative method. It will be stated in terms of a construct called *matrix cocycles*.

Associated matrix cocycle. Matrix cocycles over the dynamics (Ω, μ, f) will be defined in more generality later in section 6.1. For the moment, we focus on the matrix-valued functions

$$\begin{aligned} W &: \Omega \to \mathbb{R}^{d\times L}, \quad W(\omega) := Dw \circ \Phi(\omega), \\ \hat{W} &: \Omega \to \mathbb{R}^{d\times L}, \quad \hat{W}(\omega) := D\hat{w}|_{\Phi(\omega)} = D\hat{w} \circ \Phi(\omega), \\ G^{(1)} &: \Omega \to \mathbb{R}^{L\times d}, \quad G^{(1)}(\omega) := \nabla_1 g|_{h(\omega)} = \nabla_1 g \circ h(\omega), \\ G^{(2)} &: \Omega \to \mathbb{R}^{L\times L}, \quad G^{(2)}(\omega) := \nabla_2 g|_{h(\omega)} \nabla_2 g \circ h(\omega) \end{aligned} \tag{4.6}$$

and their combination

$$M : \Omega \to \mathbb{R}^{(L+d)\times(L+d)}, \quad M(\omega) := \begin{bmatrix} 0^{d\times d} & W(\omega) \\ G^{(1)}(\omega) & G^{(2)}(\omega) \end{bmatrix}. \tag{4.7}$$

Next consider the vector-valued functions

$$c : \Omega \to \mathbb{R}^L,\ c(\omega) := G^{(1)}(\omega)\left(U^{-1}\Delta\phi\right)(\omega).$$

We shall use this to build a nonautonomous dynamical system on $\mathbb{R}^{d+L}$. Fix an $\omega \in \Omega$ and define

$$\begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} = M\left(f^n\omega\right)\begin{bmatrix} a_n \\ b_n \end{bmatrix} + \begin{bmatrix} 0 \\ c\left(f^n\omega\right) \end{bmatrix}, \quad a_1 = 0^d,\ b_0 = 0^L. \tag{4.8}$$

We call such a system a *perturbed matrix cocycle* (see section 7). Note that as the size of the hypothesis space is increased, the function c converges to 0 in the $L^2(\mu)$-norm, and the dynamics of (a_n, b_n) gets closer to that of the matrix cocycle generated by M. We shall examine this closely in Theorem 7.1. Note that (4.8) depends on the initial state ω. If ω is allowed to vary, then a_n, b_n become functions of ω. We shall overuse notation and also denote these functions as a_n, b_n.

Growth of the iterative error. In Theorem 4.2 below, we shall establish a rate at which the iterative error grows. Let (u_n, y_n) be iterates of the system (1.4). We are interested in the growth of the deviation quantities $\Delta u_n, \Delta y_n$ defined as

$$\begin{bmatrix} \Delta u_n \\ \Delta y_n \end{bmatrix} = \begin{bmatrix} U^{n-1}\pi U\phi \\ U^n \Phi \end{bmatrix} - \begin{bmatrix} u_n \\ y_n \end{bmatrix} \quad \forall n \in \mathbb{N}_0. \tag{4.9}$$

Note that when defining the deviation terms, we are using as reference the functions $U^{n-1}\pi U$ and $U^n\Phi$, both of which reflect the true state of the dynamics (Ω, f). See Remark 4.3 for further discussions on their significance. We now have the following theorem.

Theorem 4.2 (error from iterative forecast). *Let Assumptions 1 and 2 hold, and assume the notation in (1.7), (1.4), and (1.6). Fix an initial state $\omega \in \Omega$, and let (u_n, y_n) be successive iterations of the system (1.4), and let (a_n, b_n) be iterations of the dynamics in (4.8).*

(i) *Let δ be as in (4.5). The deviations (4.9) have the following relations with the states of the associated perturbed cocycle:*

$$\Delta u_n = a_n + O\left(a_{n-1}\right)^2, \quad \lim_{\delta \to 0} \frac{\|\Delta u_n\|}{\|a_n\|} = 1. \tag{4.10}$$

(ii) *Let $\lambda_1 = \lambda_1(\mathcal{M})$ the maximal Lyapunov exponent of the cocycle generated by $\hat{M}$. Then for every $\epsilon > 0$, there is a constant $C^{(1)}_{\omega,\epsilon}$ such that*

$$error_{iter}(n, \omega) = \|\Delta u_n(\omega)\|_{\mathbb{R}^L} = \delta C^{(1)}_{\omega,\epsilon} O\left(e^{n(\lambda_1+\epsilon)}\right) \quad as\ n \to \infty. \tag{4.11}$$

(iii) *If (Ω, μ, f) has the additional property of L^2 Pesin sets, then for every $\epsilon > 0$,*

$$error_{iter}(n) = \|\Delta u_n\|_{L^2(\mu)} = \delta C^{(2)}_{\epsilon} O\left(e^{n(\lambda_1+\epsilon)}\right) \quad as\ n \to \infty \tag{4.12}$$

for a constant $C^{(2)}_{\epsilon}$ that depends only on ϵ.

Pesin sets are subsets of Ω on which the nonuniformly hyperbolic map f has some degree of regularity. While Pesin sets always exist and cover the entire space Ω, the property of L^2 Pesin sets is an additional property, explained in more detail in section 6.2. Theorem 4.2 is proved in section 11.

Remark 4.3 ($U^{n-1}\pi U$ vs. $U^n\pi$). The explicit formulas for the direct and iterative schemes reveal a basic mathematical law that makes the direct method unsuitable for long-term prediction. It involves the operator πU^n, which projects the evolving measurement ϕ back into the space $\mathcal{W}$. For strongly mixing systems, the Koopman operator drives out any function from any finite-dimensional subspace, up to a constant function. On the other hand, the iterative method always involves the term $U^{n-1}\pi U$. The crucial difference is that the projection π is not made after the application of U^n, but always to the static operator U. The U^{n-1} in front of the π then merely acts as a rotation/unitary transform and thus does not change the $L^2(\mu)$ (i.e., RMS) magnitude of the error. Also see Remark 4.4 for a further discussion on decay of correlations.

Remark 4.4 (decay of correlations). Theorem 4.1 relates the growth of $\text{error}_{\text{direct}}$ with the rate of decay of correlation, while Theorem 4.2 relates the growth of $\text{error}_{\text{iter}}$ with the top

Lyapunov exponent. The former is a spectral/operator-theoretic property, while the latter is a combination of differential and ergodic properties such as Lyapunov exponents. The exact connections between mixing and positive Lyapunov exponents are still far from understood [2]. Connections have been established heuristically in some cases, such as [56, 17]. Rigorous proofs have been possible under additional assumptions such as the existence of finite Markov partitions [80, 79], or an expansive property of the map [3].

Remark 4.5 *(autocorrelations).* Given any nonzero function $\psi \in L^2(\mu)$, we define its normalized autocorrelation (with respect to the underlying dynamics) as

$$\mathrm{AutCorr}(n;\psi) := \|\psi\|^{-2} \langle U^n\psi, \psi\rangle .$$

Now, suppose that ψ lies in $\mathcal{W}$. Let $\{\mathfrak{w}_i\}_{i=1}^M$ be any orthonormal basis for the hypothesis space $\mathcal{W}$. Then

$$\begin{aligned}\mathrm{AutCorr}(n;\psi)^2 := \|\psi\|^{-4} |\langle U^n\psi,\psi\rangle|^2 &= \|\psi\|^{-4} \left| \left\langle U^n\psi, \sum_i \langle \mathfrak{w}_i,\psi\rangle\, \mathfrak{w}_i \right\rangle \right|^2 \\ &= \|\psi\|^{-4} \left| \sum_{i=1}^M \langle \mathfrak{w}_i,\psi\rangle \langle U^n\psi, \mathfrak{w}_i\rangle \right|^2 \\ &\le \|\psi\|^{-4} \sum_{i=1}^M |\langle \mathfrak{w}_i,\psi\rangle|^2 \sum_{i=1}^M |\langle U^n\psi,\mathfrak{w}_i\rangle|^2 = \|\psi\|^{-2} \sum_{i=1}^M |\langle U^n\psi,\mathfrak{w}_i\rangle|^2 .\end{aligned}$$

We show in section 10 that

$$\begin{aligned}\mathrm{error}_{\mathrm{direct}}(n)^2 = \|(\mathrm{Id}-\pi)\,U^n\phi\|^2_{L^2(\mu)} &= \|\phi\|^2 + \|\pi U^n\phi\|^2 - 2\langle U^n\phi, \pi U^n\phi\rangle \\ &= \|\phi\|^2 - \sum_{i=1}^M |\langle U^n\phi,\mathfrak{w}_i\rangle|^2 .\end{aligned}$$

Combining, we get

$$\phi \in \mathcal{W} \quad \Rightarrow \quad \mathrm{error}_{\mathrm{direct}}(n)^2 \le \|\phi\|^2 \left[1 - \mathrm{AutCorr}(n;\phi)^2\right]. \tag{4.13}$$

Thus, if the hypothesis space happens to include the initial observation map ϕ, then the growth of the direct error is directly related to the autocorrelation function of the observed signal ϕ. Autocorrelation is a statistical property of signals used frequently in classical time series analysis (see, e.g., [13]). Equation 4.13 thus combines concepts from learning theory, ergodic theory, and time series analysis.

Remark 4.6 *(overfitting error vs. projection error).* Equation (4.11) shows that the projection rate grows exponentially as expected from the presence of a Lyapunov exponent. The rate of growth is proportional to the smoothness of the learned function $\hat{w}_1$, while the multiplicative constant is proportional to the projection error δ. Thus this displays a trade-off between projection error and overfitting, and one can minimize the projection error by increasing the hypothesis space. But the resulting learned function may be too oscillatory, as a result increasing the instability of the feedback system (1.4). On the other hand, if one

approximates w_1 by a less oscillatory function, our base error δ itself will be large to begin with. To state this trade-off more precisely, define

$$\theta(\epsilon) := \inf \{\|D\hat{w}\| \,:\, \hat{w} \in \text{ some hypothesis space } \mathcal{H},\, \delta(\mathcal{H}) < \epsilon\}.$$

Then θ is a nondecreasing function of ϵ, satisfying

$$\theta(\|\phi\|) = 0, \quad \lim_{\epsilon \to 0^+} \theta(\epsilon) = \|Dw\|.$$

Thus (4.11) can be rewritten as

$$\text{error}_{\text{iter}}(k) = \epsilon O\left(k\theta(\epsilon)^k\right) \quad \text{as } k \to \infty.$$

Remark 4.7 *(cocycle structure).* Equations (4.11) and (4.8) together describe the evolution of the reconstructed dynamics as the normal and error parts, respectively. The format of (4.8) is that of a *matrix cocycle*, one of the major contributions of our paper. It also bears a resemblance to the *perturbed nonautonomous equations*, studied in the continuous-time case by Barreira and Valls [6, 5]. We look more closely at the growth or decay of these cocycles in section 7 and Theorem 7.1.

Remark 4.8 *(tightness of bounds).* The bound derived in (4.11) is not a tight bound. We obtain a better estimate in section 11 in terms of the full Lyapunov spectrum.

This completes the statement of our main results. In section 5, we discuss the consequences of our results and look at some numerical verification. In section 6 we review some concepts from random matrix cocycle theory. In the sections after that, we prove our theorems.

5. Examples and discussions. In this section we explore some of the consequences of Theorems 4.1 and 4.2.

1. According to Theorem 4.1 (iii), for a weakly mixing system, the direct prediction loses track of the signal and eventually only retains the mean of ϕ. The error from direct prediction thus converges to the variance of the observation ϕ.
2. For a mixed spectrum system, the direct method should recover a portion of $\text{proj}_{\mathcal{D}}\phi$, depending on the size of the hypothesis space $\mathcal{H}$, and lose track of the complementary component. Moreover, in (10.1), (10.5), which we derive later, the error $\text{error}_{\text{direct}}(n)$ does not converge to the variance, but fluctuates periodically.
3. The growth of the iterative error on the other hand does not depend on spectral properties of the dynamics. It depends on the top Lyapunov exponent $\lambda_1(\hat{w})$ of the reconstructed dynamics, which in turn depends on the top Lyapunov exponent of the original dynamics (Ω, f, μ) as well as the accuracy of the approximation $\hat{w}$. In a practical application, $\lambda_1(\hat{w})$ could be affected by the number of training data, and the smoothness of the true feedback function w.
4. Another feature of the iterative error is that unlike the direct error, it is not bounded by $\|\phi\|_{L^2(\mu)}$ as it is not the result of the applications of pure operators, but it is the deviation between the trajectories of two different dynamical systems. Thus the error could be of the order of $\sqrt{2}\|\phi\|_{L^2(\mu)}$.

We next describe some numerical experiments conducted to verify and illustrate these universal behaviors.

5.1. Numerical experiments. We now compare the reconstruction technique using the two paradigms of invariant graphs and delay-coordinate embedding. The former was implemented using reservoir systems, and the latter using kernel regression. We applied both of these techniques to three systems:

(i) A quasiperiodic rotation on a two-dimensional torus (Figure 3, top panel):

$$\left(\theta^1(n+1), \theta^{(2)}(n+1)\right) = \left(\theta^1(n+1), \theta^{(2)}(n+1)\right) + (\rho_1, \rho_2) \bmod 2\pi. \tag{5.1}$$

Here $\theta^{(1)}$ and $\theta^{(2)}$ are angular coordinates on the torus, and (ρ_1, ρ_2) is the rotation vector.

(ii) The Lorenz63 (L63) system (Figure 2). Let Φ^t_{L63} denote the flow under the Lorenz63 system. Fix a sampling interval Δt. This leads to the discrete time system

$$(x_{n+1}, y_{n+1}, z_{n+1}) = \Psi^{\Delta t}_{\mathrm{L63}}(x_n, y_n, z_n). \tag{5.2}$$

Φ^t_{L63} has a unique physical measure which has been proved to be nonuniformly hyperbolic and mixing.

(iii) A dynamical system formed by taking the Cartesian product of L63 with a simple harmonic oscillator (Figure 3, bottom panel). Such a system will have a mixed spectrum, with the space $\mathcal{D}$ generated by a single base eigenfunction. We shall refer to this system as L63Rot.

$$\begin{aligned} \theta_{n+1} &= \theta_n + \rho \bmod 2\pi, \\ (x_{n+1}, y_{n+1}, z_{n+1}) &= \Psi^{\Delta t}_{\mathrm{L63}}(x_n, y_n, z_n). \end{aligned} \tag{5.3}$$

This system is analyzed in the bottom panel of Figure 3.

The results of our computations in Figures 2, 3, and 4 illustrate the consequences of Theorems 4.1 and 4.2. Figure 4 highlights the two most important conclusions from our results. First, as seen in the top row, the iterative errors grow at an exponential rate comparable to the top Lyapunov exponent λ_1. Second, if the hypothesis space is large enough, then the direct error is bounded above by a formula (4.13) involving the autocorrelation function of the direct observation map ϕ. There are three things to note concerning Figure 4:

(i) Theorem 4.2 gives an upper bound for the long-term behavior of the iterative error. The theoretical bound for exponential rate of growth is indicated by the slope of the black dashed line and is $\approx 0.9056\Delta t$. So although the initial exponential rate of errors seem to be larger than this, by choosing a multiplicative constant large enough, the error graph still remains underneath the theoretical curve. The offset of the straight dashed line equals the logarithm of this multiplicative constant. Thus as long as the long-term averaged error growth rate is less than $\approx 0.9056\Delta t$, there will always be a multiplicative constant large enough to satisfy the bounds in (4.11) and (4.12).

(ii) The errors from the direct error occasionally cross the theoretical bound indicated by the black dashed line. This is because the bound in (4.13) assumes that the learning error for w is zero, i.e., w lies in the hypothesis space $\mathcal{H}$. In most situations such as in our experiments, there is always a small learning error. An extended analysis for this situation is an interesting and open task.

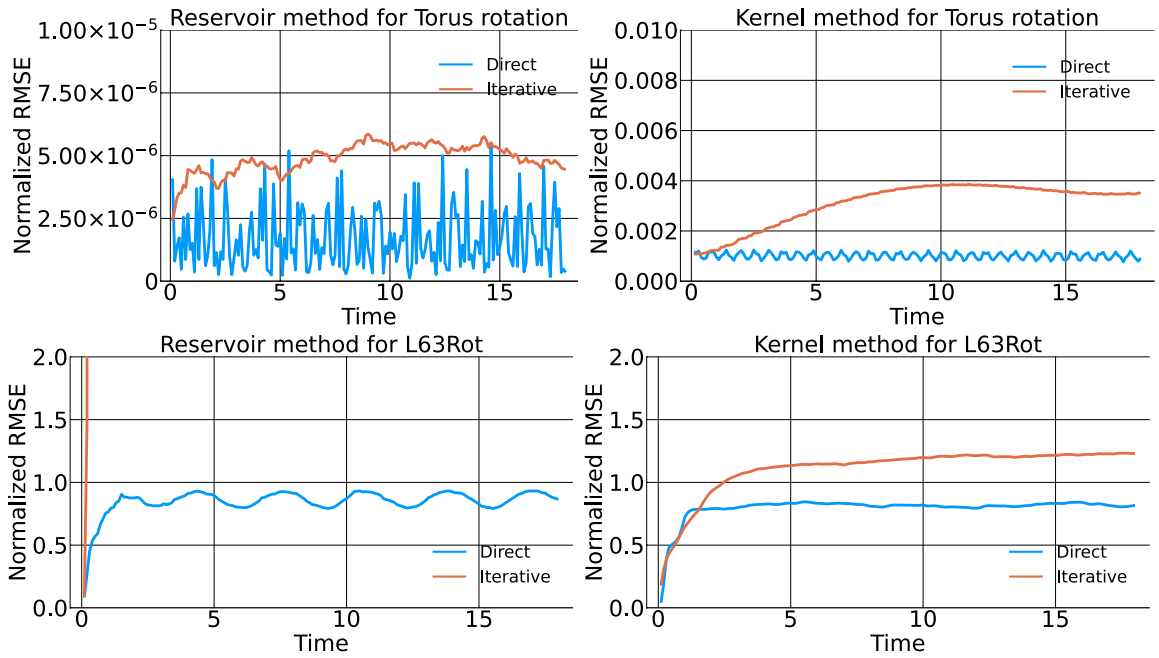


Figure 3. *Performance of the two reconstruction techniques for (i) a quasiperiodic rotation on the torus $\mathbb{T}^2$ (bottom panels), and (ii) a Cartesian product of Lorenz63 with a simple harmonic oscillator (top panels). By Theorem 4.1 (iv), if one has a proper embedding and a good approximation $\hat{w}$ of w, one can achieve arbitrarily small errors for the torus rotation for all forecast times. This is supported by the fact that the direct methods for both the paradigms show errors of the order of 10^{-6}. Since the torus rotation has all Lyapunov exponents zero, by Theorem 4.2 (ii) and Theorem 3.1 (ii), the error from the iterative techniques should grow subexponentially, as supported by the figures. The system (5.3) is a mixed spectrum system, i.e., the splitting in (1.1) is nontrivial. The plots conform to the expected behavior discussed in points (1)–(4) of section 5.*

(iii) The errors from the iterative forecasts made using the reservoir blow up. This is because the standard reservoir computers are not guaranteed to be stable. This drawback is a important subject for further study.

This completes the presentation of our main theoretical and numerical results. The framework that we have built provides many new directions of research into the field of learning of dynamical systems. We now present some other directions of work.

5.2. Methods based on Koopman approximation. There are many techniques of forecasting which do not attempt to reconstruct the dynamics using some form of embedding. Instead, they directly try to approximate the Koopman operator by tracking its action on a limited subspace of functions. In this section we review some of these methods and relate them loosely to our main mathematical constructions.

Kernel analogue forecasting. This technique [81] is a direct method for pointwise forecast, using locally decaying kernels. Suppose that μ is a smooth volume measure, and p_ϵ is a C^2, strictly positive definite, locally decaying kernel, which is Markov with respect to μ, i.e.,

$$\int p_\epsilon(\cdot, y) d\mu(y) \equiv 1_\Omega.$$

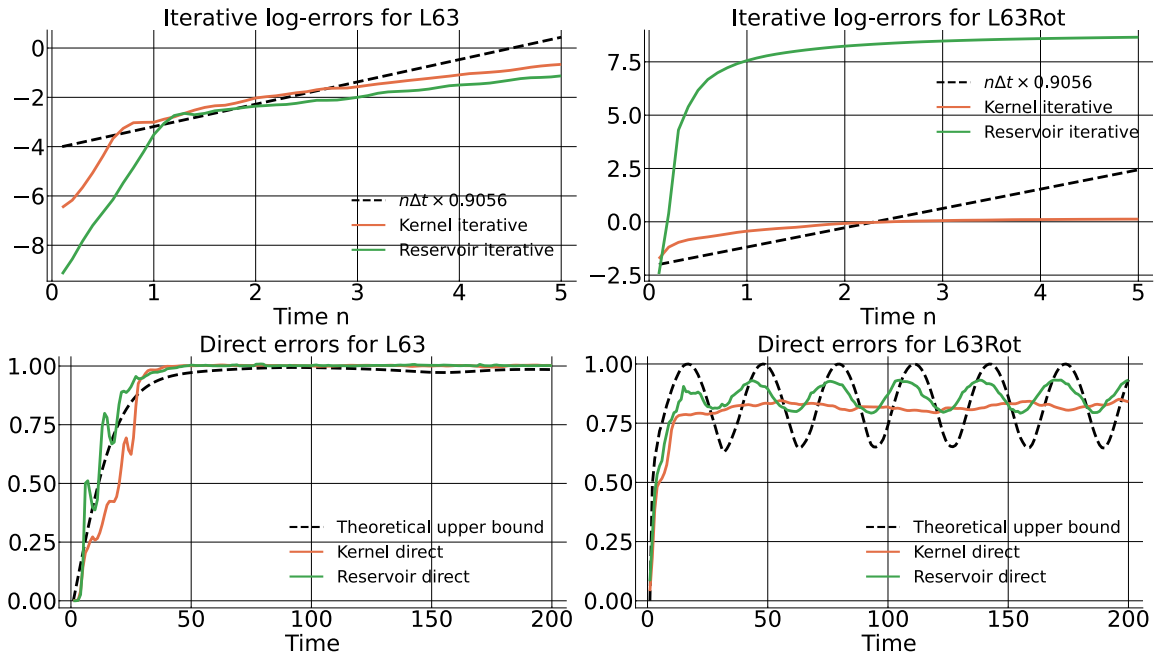


Figure 4. *Error analysis using theoretical results. The top row shows how the iterative errors of the L63 and L63Rot systems compare with the theoretical bounds of Theorem 4.2. The dashed lines each have slopes $\lambda_1 \Delta t$, with Δt being the sampling interval, and $\lambda_1 \approx 0.9056$ for both the systems. The bottom row compares the errors from direct forecast with the autocorrelation bound of (4.13). The assumption there that w lies in the hypothesis space is not met, and thus we see some fluctuations above the theoretical upper bound. Together, these plots indicate conformity with our theoretical predictions.*

Let P_ϵ be the kernel integral operator corresponding to p_ϵ and μ. The core idea of this technique is the following pointwise estimate (see, e.g., [16, 75]):

$$|\phi(\omega) - (P_\epsilon \phi)(\omega)| \leq \|D\phi(\omega)\| \, \epsilon + O\left(\epsilon^2\right) \quad \forall \phi \in C^1(\Omega), \quad \forall \omega \in \Omega. \tag{5.4}$$

Using a change-of-variables formula, one can write

$$\begin{aligned}(P_\epsilon U^t \phi)(\omega) &= \int p_\epsilon(\omega, \omega')(U^t \phi)(\omega') d\mu(\omega') = \int p_\epsilon(\omega, \omega') \phi(\Phi^t \omega') d\mu(\omega'), \quad \text{take } \omega'' := \Phi^t \omega' \\ &= \int p_\epsilon(\omega, \Phi^{-t} \omega'') \phi(\omega'') d(\Phi_*^t \mu)(\omega'').\end{aligned}$$

Therefore

$$(U^t \phi)(\omega) = \int p_\epsilon(\omega, \Phi^{-t} \omega'') \phi(\omega'') d(\Phi_*^t \mu)(\omega'') + \left\| D(U^t \phi)(\omega) \right\| O(\epsilon). \tag{5.5}$$

In the above inequality the integral is approximated as

$$\int p_\epsilon(\omega, \Phi^{-t} \omega'') \phi(\omega'') d(\Phi_*^t \mu)(\omega'') \approx \frac{1}{N} \sum_{n=0}^{N-1} p_\epsilon(\omega, \omega_n) \phi(\omega_{n+t}).$$

One of the major drawbacks of this method is that the pointwise approximation deteriorates as t increases, since the function $U^t \phi$ becomes increasingly oscillatory.

Diffusion forecast. This involves choosing an orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ for some choice of a Hilbert space H, choosing the size of a truncation L, setting $H_L := \text{span}\,\{\phi_j : j = 1, \ldots, L\}$, and setting

$$U^{(L)} := \pi_L U \pi_L, \quad U_n^{(L)} := \pi_L U^n \pi_L,$$

where $\pi_L : H \to H_L$ is the orthogonal projection. $U^{(L)}$ and $U_n^{(L)}$ are the L-dimensional approximations of the Koopman operator. Typical choices of H are $L^2(\mu)$ or Sobolev spaces, and the ϕ_j are typically Laplacian eigenfunctions, or eigenfunctions of symmetric kernel integral operators. The choice between $U^{(L)}$ and $U_n^{(L)}$ is similar to the choice between the iterative and direct methods (4.1) and (4.2). However, since these methods are not dependent on an actual embedding of the dynamics, the error of both of these forecasts grows at the same rate as the rate of decay of correlations.

Spectral techniques. The diffusion forecast is one among many techniques of approximating the Koopman operator. A more robust approach is a spectral approximation technique developed in [32], in which the goal is to approximate the spectral measure of the generator V associated with a continuous-time dynamical system. This technique is convergent and works for any kind of ergodic dynamical system. The Koopman group $\{U^t : t \in \mathbb{R}\}$ is then approximated by the 1-parameter unitary group generated by a compact, spectral approximation $\tilde{V}$ of V. In this technique, U^t is not approximated by its action on a fixed subspace of functions, but on a subspace spanned by *approximate eigenfunctions*. This also leads to a discovery of nearly periodic structures present within the possibly chaotic system.

5.3. Future work. There are several promising directions of research that can be built upon our framework.

1. Multimodal forecasting: One of the main ideas verified theoretically and via numerical experiments is that the error of direct forecasts increase at the rate of mixing of the system, which is usually larger than the top Lyapunov exponent. However, if there are quasiperiodic components, the direct method is effective in retaining that component. On the other hand, the error from the iterative method increases at a slower rate, but does not preserve the quasiperiodic components of the signal. The iteration model eventually behaves effectively in an uncorrelated fashion with the true dynamics. A *multimodal* forecasting technique would be a combination of these two modes, which combines their best features.
2. The direct method is essentially $\pi U^n \phi$, and the iterative is $U^{n-1}\pi U\phi$. Another possibility is a k-time step iterative forecast which would be $U^{n-k}\pi U^k\phi$. As k increases the leading term $\pi U^k\phi$ will have an error that decays according to the decay of correlations. To make amends, we could incorporate several k's in a window $[1, K]$ as
$$\sum_{k=1}^{K} \alpha_k U^{n-k}\pi U^k\phi.$$
3. Yet another idea we are pursuing is *ensemble forecasting*, which has long been suggested as a prediction technique for chaotic systems [42, 18].

4. Numerical approximation of optimal feedback function $\bar{w}$: A key insight of Theorem 3.1 is that the top Lyapunov exponent of the reconstructed model depends on the behavior of the feedback function in a neighborhood of the image of the attractor. Theorem 4.2 then shows that this Lyapunov exponent describes the exponential rate at which the reconstructed model diverges from the true system. In most learning techniques, one tries to find a feedback function that simultaneously minimizes a fitting error and an oscillation penalty term. Theorems 3.1 and 4.2 suggest that instead of measuring the overfitting error via the usual oscillation bond, a good candidate would be to take into account the behavior of $\bar{w}$ in a neighborhood of the dataset. The precise manner in which this ambient space behavior is to be translated into a penalty function is a promising field of research. A related and inseparable question concerns that of an appropriate choice of hypothesis space.
5. Effect of noise: Our techniques have not addressed the challenges posed by noise, either in measurement or dynamic. It is well known (see, e.g., [74]) that measurement noise could be hard to distinguish from chaos and can severely restrict the accuracy of even short-term predictions. Numerical averages rely on ergodic convergences, and the stability of ergodic averages to noise is a complicated and broad question of its own. Stability results have been shown in systems with Sinai–Ruelle–Bowen (SRB) measures [77, 10]. In such settings, the use of Kalman filtering in a model-free approach [40, 41] may yield promising results.

6. Review of nonuniform hyperbolicity.

This section provides an overview of the topics of matrix cocycles and Lyapunov exponent theory.

6.1. Matrix cocycles.

Let Assumption 1 hold, and let $G : \Omega \to GL(\mathbb{R}, m)$ be a measurable map. Then it generates a *matrix cocycle* (see [27], [4, sect. 3.4]), which is the map

$$\mathcal{G} : \Omega \times \mathbb{N}_0 \to GL(\mathbb{R}; m), \quad \mathcal{G}(n,\omega) := \begin{cases} \mathrm{Id}_d & \text{if } n = 0, \\ G(f^{n-1}\omega)\cdots G(\omega) & \text{if } n > 0, \\ G(f^{-|n|}\omega)^{-1}\cdots G(f^{-1}\omega)^{-1} & \text{if } n < 0. \end{cases} \tag{6.1}$$

$\mathcal{G}$ is called a $GL(m;\mathbb{R})$-valued cocycle over the dynamics (Ω, μ, f) generated by f. It has the property

$$\mathcal{G}(m+n,\omega) = \mathcal{G}(n, f^m\omega)\cdot\mathcal{G}(m,\omega) \quad \forall \omega \in \Omega, \quad \forall m, n \in \mathbb{Z}. \tag{6.2}$$

Here the $\cdot$ notation denotes the matrix multiplication. Equation (6.2) is the defining equation of a matrix cocycle. Conversely, given any map $\mathcal{G} : \Omega \times \mathbb{N}_0 \to GL(m;\mathbb{R})$ satisfying (6.2), one has a generator $G : \Omega \to GL(m;\mathbb{R})$ so that $\mathcal{G}$ is related to G via (6.1). One of the immediate consequences of (6.2) is that

$$\mathcal{G}(0,\omega) = \mathrm{Id}_m, \quad \mathcal{G}(-n,\omega) = \mathcal{G}(n, f^{-n}\omega)^{-1} \quad \forall \omega \in \Omega.$$

When the initial point $\omega_0 \in \Omega$ is fixed, we will drop it from the notation and define

$$\mathcal{G}(n-1, j) := \mathcal{G}(n-j, f^j\omega_0) = G(f^{n-1}\omega_0)\cdots G(f^j\omega_0).$$

Matrix-valued cocycles arise naturally in multiple ways in dynamical systems. For example, if Ω is an m-dimensional manifold and f a differentiable map, then $\mathcal{G}(\omega, n) := Df^n(\omega)$ is a $GL(m;\mathbb{R})$ cocycle.

Proposition 6.1 (multiplicative ergodic theorem [66], [27 Thm. 4.1, p. 10]). *Let Assumption 1 hold. Then there exists a forward invariant set Ω' of full μ-measure such that the limit*

$$\Lambda(\omega) := \lim_{n\to\infty} [\mathcal{G}(n,\omega)^* \mathcal{G}(n,\omega)]^{1/2n}$$

exists for every $\omega \in \Omega'$. Moreover, there is a splitting $\mathbb{R}^M = \oplus_{i=1}^{l} E_i(\omega)$ and constants $\lambda_1 \geq \cdots \geq \lambda_l \geq -\infty$ such that

$$v \in E_i \quad \Rightarrow \quad \lim_{n\to\infty} \frac{1}{n} \ln \|\mathcal{G}(n,\omega)v\| = \lambda_i.$$

The numbers l, $\lambda_1, \ldots, \lambda_l$ are constant on Ω'. The subspaces E_i depend measurably on $\omega \in \Omega'$ and

$$\mathcal{G}(n,\omega)v \in E_i\,(f^n\omega) \quad \forall v \in E_i(\omega), \forall n \in \mathbb{Z}. \tag{6.3}$$

The vector spaces E_i are called the Lyapunov subspaces and λ_i the Lyapunov exponents. These measure the asymptotic rate of expansion or contraction along the Lyapunov directions.

6.2. Pesin sets. Lyapunov exponents describe the asymptotic behavior of orbits and not the local differential properties of the map. By Proposition 6.1, at almost every ω, the limits λ_i are attained along the various Oseledet subspaces E_i. However, the rate at which the limits are attained are in general not uniform or even continuous as a function of ω. The Oseledet subspace of $T_\omega\tilde{\Omega}$ corresponding to λ_i itself is usually a measurable but noncontinuous function of ω.

Pesin sets [64, 66] were introduced to capture the regularity and boundedness in the highly nonuniform nature of the Oseledet splitting. Fix a constant $\epsilon_l > 0$, such that $\epsilon_l \ll \min_i |\lambda_i|$. ϵ_l is called the *leakage rate*. Then there is a nested sequence of compact sets $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_3 \subseteq \cdots$ whose union has μ-measure 1, such that for every $k \in \mathbb{N}$

$$e^{-k\epsilon_l} e^{(\lambda_i - \epsilon_l)n} \leq \|Df^n(\omega)|E_i\| \leq e^{k\epsilon_l} e^{(\lambda_i+\epsilon_l)n} \quad \forall \omega \in \Omega_k, \forall i \in \{1,\ldots,r\}, \forall n \in \mathbb{Z}.$$

Moreover, the subspaces E_i vary smoothly on the sets Ω_k.

Although the norm of the Jacobian $Df^n(\omega)$ when restricted to E_i grows asymptotically at the rate $e^{\lambda_i n}$, this exact exponential growth need not be attained for finite n. There is a constant $C_{\mathrm{NUH}}(\omega) = C_{\mathrm{NUH}}(\omega;\epsilon_l)$ depending on ω such that

$$\frac{1}{C_{\mathrm{NUH}}(\omega)} e^{(\lambda_i-\epsilon_l)n} \leq \|Df^n(\omega)|E_i\| \leq C_{\mathrm{NUH}}(\omega) e^{(\lambda_i+\epsilon_l)n} \quad \forall n \in \mathbb{Z}, \forall i \in 1,\ldots,r. \tag{6.4}$$

$C_{\mathrm{NUH}}(\omega)$ plays the role of a multiplicative constant, and ϵ_l behaves as the extent of fluctuation around the limiting rate λ_i. The decomposition into Pesin sets imply that if ω is restricted to Ω_k, then $C_{\mathrm{NUH}}(\omega)$ can be uniformly bounded by $e^{k\epsilon_l}$.

Thus the Pesin sets Ω_k have uniformly hyperbolic behavior. However, they need not be uniformly hyperbolic sets, as they are not necessarily invariant sets. In general $f(\Omega_k) \subseteq \Omega_{k+1}$. Note that if Ω_k is invariant, then it is a uniformly hyperbolic set. Despite not being invariant sets, Pesin sets are useful for obtaining concrete bounds on the rate of hyperbolicity. The Poincaré recurrence theorem guarantees that a typical trajectory returns to a Pesin set infinitely many times. These two properties of recurrence and uniform hyperbolic rates have been used effectively to establish strong global properties of the system, such as approximation by periodic points [49], shadowing [78], and metric properties of local stable and unstable manifolds [53, 54]. These techniques will play an important role in our proofs.

L^p Pesin sets. For every choice of the leakage rate ϵ_l, the Pesin sets Ω_k grow to form an invariant set of full measure. The rate at which $\mu(\Omega_k)$ approaches 1 is an important consideration. It is important for obtaining estimates on various statistical properties of the nonuniformly hyperbolic system. However, there are not many estimates on how quickly the Pesin sets grow, except under additional conditions, such as the existence of reasonably good Markov approximations (see, e.g., [35]). For our purpose, we say that a nonuniformly hyperbolic system has *L^p Pesin sets* if the function $\omega \mapsto C_{\mathrm{NUH}}(\omega; \epsilon_l)$ is L^p-integrable with respect to ω. This property will be used later to obtain global bounds from local behavior in Theorem 7.1.

6.3. Lyapunov exponents in Euclidean space. Given a dynamical system on $F : \mathbb{R}^M \to \mathbb{R}^M$, one has the following alternative definition of Lyapunov exponents:

$$\lambda(z, v) := \limsup_{n\to\infty} \frac{1}{n} \ln \lim_{\delta\to 0^+} \frac{1}{\delta} \left\| F^n(z + \delta v) - F^n(z) \right\|, \quad z \in \mathbb{R}^M, v \in \mathbb{R}^M.$$

Lemma 6.2. *Let $\mathcal{T} : \mathbb{R}^M \to \mathbb{R}^M$ be a C^1 map, with an invariant ergodic measure $\bar{\mu}$ with compact support X. Let $T_X\mathbb{R}^M = E_1 \oplus E_2 \oplus \cdots \oplus E_k$ be a splitting of $T\mathbb{R}^M$ restricted to X. Let λ_1 be the maximal Lyapunov exponent with respect to the measure $\bar{\mu}$. Then for μ-a.e. $z \in X$,*

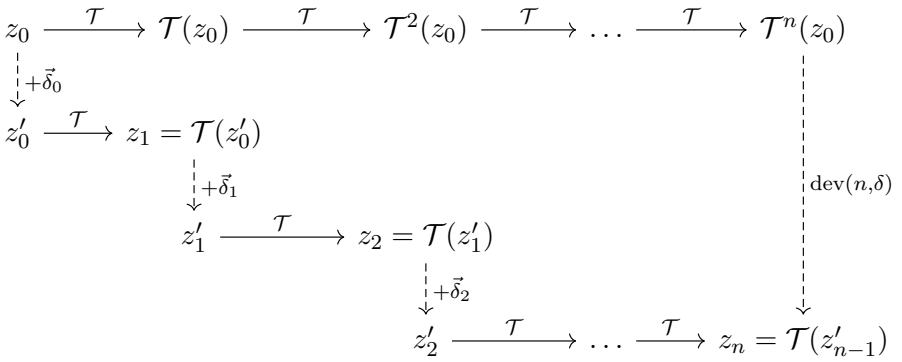
$$\lambda_1 = \max_{1\le i\le k} \sup_{v\in E_i(z)\setminus\{0\}} \lambda(z, v) = \sup_{v\in T_z\mathbb{R}^M\setminus\{0\}} \lambda(z, v).$$

The proof is a direct consequence of the definition of Lyapunov exponents and will be omitted. We next consider a special type of perturbed sequences of points.

Pseudotrajectories. Let $\mathcal{T} : \mathbb{R}^M \to \mathbb{R}^M$ be a C^1 map on a manifold M, with an ergodic invariant measure $\bar{\mu}$ with compact support X. Fix a sequence of positive numbers $(c_j)_{j=0}^{\infty}$ and initial point $z_0 \in X$. Now define

(6.5)
$$\mathcal{S}\left(z_0, (c_j)_{j=0}^{\infty}\right) := \left\{ (z_j)_{j=0}^{\infty} \in \mathbb{R}^M \,:\, z_{n+1} = \mathcal{T}(z_n'),\, d(z_{n+1}', z_{n+1}) \le c_n d(z_n', z_n)\, \forall n \in \mathbb{N} \right\},$$
$$\mathcal{S}\left(\delta; z_0, (c_j)_{j=0}^{\infty}\right) := \left\{ (z_j)_{j=0}^{\infty} \in \mathcal{S}\left(z_0, (c_j)_{j=0}^{\infty}\right) \,:\, d(z_0', z_0) \le \delta \right\}.$$

Thus $\mathcal{S}(z_0, (c_j)_{j=0}^{\infty})$ is the set of all pseudotrajectories $z_n \in \mathbb{R}^m$ such that at each stage n, z_{n+1} is the image of a perturbation z_n' of z_n. Moreover, the perturbation to z_{n+1} is at most c_n times the perturbation to z_n. The set $\mathcal{S}(\delta; z_0, (c_j)_{j=0}^{\infty})$ is the subset of these sequences such that the initial perturbation is no more than δ. Thus $S(z_0, c) = \cup_{\delta>0} S(\delta; z_0, c)$. The figure below illustrates such a sequence in $\mathcal{S}(\delta; z_0, (c_j)_{j=0}^{\infty})$:



The top row shows the reference trajectory $\{\mathcal{T}^n z_0 : n \in \mathbb{N}_0\}$, starting at an initial point z_0. At each time step $n = 1, 2, \dots$, z_n is the image of a point z'_{n-1}, which is a $\vec{\delta}_{n-1}$ perturbation of the earlier point z_{n-1}. The magnitude of $\vec{\delta}_n$ is bounded by $c_{n-1}\delta_{n-1}$. Thus at every stage, the error accumulates and is scaled by the factor of at most c_n. The magnitude of the initial perturbation is $\|\vec{\delta}_0\| \leq \delta$. Such perturbed sequences arise in our proof of Theorem 3.1. We study the rate of growth of the divergence between the two trajectories, as a ratio of the initial error magnitude δ. The maximum possible deviation after n steps can be written as

$$\mathrm{dev}(z_0, n, \delta) := \sup\left\{ d\left(z_n, \tau^n(z_0)\right) \,:\, (z_j)_{j=0}^{\infty} \in \mathcal{S}\left(\delta; z_0, (c_j)_{j=0}^{\infty}\right)\right\}.$$

We are more interested in the growth of this deviation as a multiplier of the initial error margin δ, namely,

$$\mathrm{dev}(z_0, n) := \limsup_{\delta \to 0^+} \frac{1}{\delta} \mathrm{dev}(z_0, n, \delta).$$

We next derive the asymptotic rate at which these rates of divergence $\mathrm{dev}(z_0, n)$ grow.

Proposition 6.3 (δ-pseudotrajectory). *Let $\mathcal{T} : M \to M$ be a C^1 map on a manifold M, with an ergodic, nonuniformly hyperbolic invariant measure $\bar{\mu}$ with compact support X. Assume the notation in (6.5). Let $(c_j)_{j=0}^{\infty}$ be a sequence of positive numbers for which the limit $C = \lim_{N\to\infty} \frac{1}{N} \sum_{j=0}^{N} \ln c_j$ exists. Then*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathrm{dev}(z_0, n) \leq \lambda_1(\tau) + C, \quad \bar{\mu}\text{-a.e. } z_0 \in X.$$

The proof is a direct consequence of the local stable/unstable manifold theorem [66, sect. 6] and will be omitted. Let $c : X \to \mathbb{R}^+$ be a continuous function. Now define, similarly to (6.5),

$$\mathcal{S}(z_0, c) := \left\{ (z_j)_{j=0}^{\infty} \,:\, z_{n+1} = \mathcal{T}(z'_n),\, d(z'_{n+1}, z_{n+1}) \leq c(z_n) d(z'_n, z_n),\, \forall n \in \mathbb{N} \right\}, \tag{6.6}$$

$$\mathcal{S}(\delta; z_0, c) := \left\{ (z_j)_{j=0}^{\infty} \in \mathcal{S}(z_0, c) \,:\, d(z'_0, z_0) \leq \delta \right\},$$

$$\mathrm{dev}(n, \delta; z_0, c) := \sup\left\{ d\left(z_n, \mathcal{T}^n(z_0)\right) \,:\, (z_j)_{j=0}^{\infty} \in \mathcal{S}(\delta; z_0, c) \right\}.$$

Proposition 6.4 (δ-pseudotrajectory II). *Let $\mathcal{T} : M \to M$ be a C^1 map on a manifold M, with an ergodic, nonuniformly hyperbolic invariant measure $\bar{\mu}$ with compact support X, and $c : M \to \mathbb{R}^+$ is a continuous map. Assume the notation in (6.6). Then*

$$\limsup_{n\to\infty} \frac{1}{n} \limsup_{\delta\to 0^+} \ln \frac{1}{\delta} \mathrm{dev}(n, \delta; z_0, c) \le \lambda_1(f) + \int \ln c d\bar{\mu}.$$

Proof. Set $c_j := c(z_j)$. Then as $\delta \to 0$, the sum $\frac{1}{n}\sum_{j=0}^{n-1} \ln c_j$ converge by the uniform continuity of c and the ergodic theorem. Thus Proposition 6.3 applies. ∎

7. Cocycles with random perturbations. This section investigates a type of dynamics that we shall call a *perturbed random cocycle* in (4.8). The results in this section are of independent interest, but also directly apply to our study of the growth of error under iterations of the reconstructed system (1.4). We shall later use them in the proof of Theorem 4.2. A version of perturbed random cocycles was investigated by Barreira and Valls [6, 5] in the context of nonautonomous differential equations of the form

$$\frac{d}{dt} v(t) = A(t)v(t) + f_{\text{perturb}}(v(t), t)$$

in Euclidean space. The nonautonomous behavior is due to the dependence on the time parameter t. In our case, for a fixed initial state ω_0 of the underlying dynamical system (Ω, f), the time dependence is via the orbit of ω_0. The authors assumed that the function f_{perturb} decays at a polynomial rate with respect to the norm of v, a property not applicable to our case. We shall study the problem in more generality.

Consider dynamics of the form

$$z_{n+1} = G_n z_n + d_{n+1}, \tag{7.1}$$

where the z_n's and d_n's are m-vectors, and the G_n's are invertible $m \times m$ matrices. z_n represents a state vector. The G_n, d_n form a sequence of random matrices and perturbation vectors. If we assume that the source of this randomness is the dynamical system (Ω, μ, f), then the iterates of (7.1) can be realized as iterates of the skew-product map

$$F : \Omega \times \mathbb{R}^M \to \Omega \times \mathbb{R}^M, \quad F : \begin{pmatrix} \omega \\ z \end{pmatrix} \mapsto \begin{pmatrix} f\omega \\ G(\omega)z + d(f\omega) \end{pmatrix}, \tag{7.2}$$

where $d \in L^2(\Omega, \mu; \mathbb{R}^M)$ can be interpreted as a random perturbation vector. We call the system (7.2) a *perturbed random matrix cocycle*. If we fix an initial state (ω_0, z_0) and set

$$z_n := \mathrm{proj}_2 F^n(\omega_0, z_0), \quad G_n := G(f^n\omega_0), \quad d_n := d(f^n\omega_0), \quad n \in \mathbb{N}_0,$$

then we get (7.1). Iterating (7.1) n times gives

$$z_n = [G_{n-1} \cdots G_0]\, z_0 + d_n + \sum_{j=1}^{n-1} [G_{n-1} \cdots G_j]\, d_j = \mathcal{G}(n-1, 0)z_0 + \sum_{j=1}^{n} \mathcal{G}(n-1, j)d_j. \tag{7.3}$$

Equation (7.3) can be rewritten in terms of (7.2) as

$$F^n(\omega, z) = \left(f^n\omega, \mathcal{G}(n,\omega)z + (\Psi^n d)(\omega)\right), \tag{7.4}$$

where

$$(\Psi^n d)(\omega) := \sum_{j=1}^{n} \mathcal{G}\left(n-j, f^j\omega\right) d\left(f^j\omega\right) \in \mathbb{R}^d \tag{7.5}$$

for every $n \in \mathbb{N}$. We shall call this map Ψ^n the *graph-transform* operator. If one views the map d as an $\mathbb{R}^d$-valued graph over Ω, then $\Psi^n d$ is a new graph over Ω. Also note that the transformation is linear in d, justifying the name of "operator." Moreover,

$$\Psi^0 u \equiv 0, \quad \Psi^1 u \equiv u \quad \forall u : \Omega \to \mathbb{R}^d. \tag{7.6}$$

Moreover, if the initial value $z_0 = 0$, then

$$z_n(\omega) = \mathrm{proj}_2 F^n(\omega, z_0 \equiv 0) = (\Psi^n d)(\omega) \quad \forall n \in \mathbb{N}.$$

By (7.4), the growth of z_n depends on the initial value z_0 only through the action of $\mathcal{G}(n,\omega)$, which is well tractable by the multiplicative ergodic theorem. We are mainly interested in the remaining part, i.e., the behavior of the operator Ψ^n. In the following theorem, we shall use λ_i^+ to denote $\max(\lambda_i, 0)$.

Theorem 7.1. *Let Assumption 1 hold and suppose $\mathcal{G}$ is a $GL(m;\mathbb{R})$-valued cocycle as in (6.1), (6.2), along with a perturbed matrix cocycle as in (7.2). Assume the notation of the associated Oseledet splitting as in Proposition 6.1, so that the vector-valued function d has the splitting*

$$d = \oplus_{i=1}^{r} d^{(i)}, \quad d^{(i)} \in E^{(i)}.$$

Finally let Ψ be as in (7.5). Then the following hold:

(i) *Suppose d is essentially bounded. Then for μ-a.e. $\omega \in \Omega$,*

$$\left\|\Psi^n d^{(i)}(\omega)\right\| = \left\|d^{(i)}\right\|_{L^\infty} C_{NUH}(\omega) O\left(e^{(\lambda_i^+ + \epsilon_l)n}\right) \text{ as } n \to \infty \quad \forall 1 \le i \le r, \tag{7.7}$$

where $\lambda_i^+ := \max(\lambda_i, 0)$.

(ii) *If the system has L^2 Pesin sets, then*

$$\left\|\Psi^n d^{(i)}(\omega)\right\|_{L^2(\mu)} = \left\|d^{(i)}\right\|_{L^2(\mu)} \|C_{NUH}\|^2_{L^2(\mu)} O\left(e^{(\lambda_i^+ + \epsilon_l)n}\right) \text{ as } n \to \infty \quad \forall 1 \le i \le r. \tag{7.8}$$

Proof. To gain more insight into the growth of z_n, we use (6.2) to get

$$\begin{aligned}(\Psi^n d)(\omega) &= \sum_{j=1}^{n} \mathcal{G}\left(n-j, f^j\omega\right) d\left(f^j\omega\right) = \sum_{j=1}^{n} \mathcal{G}\left(n-j, f^j\omega\right) \mathcal{G}(j,\omega)\, \mathcal{G}(j,\omega)^{-1} d\left(f^j\omega\right) \\ &= \mathcal{G}(n,\omega) \sum_{j=1}^{n} \mathcal{G}(j,\omega)^{-1} d\left(f^j\omega\right).\end{aligned}$$

We have thus related the error in the prediction to the growth of the vector $\mathcal{G}(n,\omega)$. The growth of the matrix $\mathcal{G}(n,\omega)$ with n can be estimated using Proposition 6.1. Broadly speaking, the different Oseledet subspaces grow approximately at rate $e^{\lambda_i n}$ under the action of $\mathcal{G}(n,\omega)$. We now estimate the growth of the components of the summand along the Oseledet splitting. Define

$$e_j(\omega) := \mathcal{G}(j,\omega)^{-1} d\left(f^j\omega\right), \quad e_j^{(i)}(\omega) := \mathcal{G}(j,\omega)^{-1} d^{(i)}\left(f^j\omega\right).$$

Then we have

$$\mathcal{G}\left(n-j, f^j\omega\right) d\left(f^j\omega\right) = \mathcal{G}(n,\omega)\, e_j(\omega), \quad \mathcal{G}\left(n-j, f^j\omega\right) d^{(i)}\left(f^j\omega\right) = \mathcal{G}(n,\omega)\, e_j^{(i)}(\omega). \tag{7.9}$$

The analysis of the growth of this term will depend on the sign of λ_i.

Case: $\lambda_i > 0$. Then

$$\begin{aligned}\left\|e_j^{(i)}(\omega)\right\| &= \left\|\mathcal{G}(j,\omega)^{-1} d^{(i)}(f^j\omega)\right\| \le \left\|\mathcal{G}(j,\omega)^{-1}|_{E^{(i)}(f^j\omega)}\right\| \left\|d^{(i)}(f^j\omega)\right\| \\ &\le C_{\mathrm{NUH}}(\omega) e^{-j(\lambda_i-\epsilon)} \left\|d^{(i)}(f^j\omega)\right\|.\end{aligned}$$

Therefore by (6.4) and (7.9),

$$\begin{aligned}\left\|\mathcal{G}\left(n-j, f^j\omega\right) d^{(i)}\left(f^j\omega\right)\right\| &= \left\|\mathcal{G}(n,\omega)\, e_j^{(i)}(\omega)\right\| \le \left\|\mathcal{G}(n,\omega)\,|_{E^{(i)}(\omega)}\right\| \left\|e_j^{(i)}(\omega)\right\| \\ &\le C_{\mathrm{NUH}}^2(\omega) e^{n(\lambda_i+\epsilon)} e^{-j(\lambda_i-\epsilon)} \left\|d^{(i)}\left(f^j\omega\right)\right\|.\end{aligned}$$

Thus $\lambda_i > 0$ implies

$$\begin{aligned}\left\|\left(\Psi^n d^{(i)}\right)(\omega)\right\| &\le \sum_{j=1}^{n} \left\|\mathcal{G}\left(n-j, f^j\omega\right) d^{(i)}\left(f^j\omega\right)\right\| \\ &\le C_{\mathrm{NUH}}(\omega)^2 e^{n(\lambda_i+\epsilon)} \sum_{j=1}^{n} e^{-j(\lambda_i-\epsilon)} \left\|d^{(i)}\left(f^j\omega\right)\right\|.\end{aligned} \tag{7.10}$$

At this point the following identity is relevant to us:

$$\begin{aligned}e^{n(\lambda_i+\epsilon)} \sum_{j=1}^{n} e^{-j(\lambda_i-\epsilon)} &= e^{-(\lambda_i-\epsilon)} e^{n(\lambda_i+\epsilon)} \frac{1-e^{-n(\lambda_i-\epsilon)}}{1-e^{-(\lambda_i-\epsilon)}} = \frac{e^{-(\lambda_i-\epsilon)}}{1-e^{-(\lambda_i-\epsilon)}} \left[e^{n(\lambda_i+\epsilon)} - e^{2\epsilon}\right] \\ &= c_i O\left(e^{n(\lambda_i+\epsilon)}\right)\end{aligned} \tag{7.11}$$

for some constant $c_i > 0$. Suppose that d is essentially bounded. Then (7.10) and (7.11) give

$$\lambda_i > 0 \quad \Rightarrow \quad \left\|\left(\Psi^n d^{(i)}\right)(\omega)\right\| \le c_i \left\|d^{(i)}\right\|_{L^\infty} O\left(e^{n(\lambda_i+\epsilon)}\right).$$

This proves claim (i) for the case $\lambda_i > 0$. Now suppose that the map has L^2 Pesin sets. Integrating both sides of (7.10) with respect to ω and then summing according to (7.11) gives

$$\left\|\Psi^n d^{(i)}\right\|_{L^1(\mu)} = \int \left\|\left(\Psi^n d^{(i)}\right)(\omega)\right\| d\mu(\omega) \le C^2 \left\|d^{(i)}\right\|_{L^1(\mu)} O\left(e^{n(\lambda_i+\epsilon)}\right). \tag{7.12}$$

Case: $\lambda_i < 0$. Suppose that $\omega \in \Omega_k$. In this case note that for each $1 \le j \le n$,

$$\begin{aligned}\left\| \mathcal{G}\left(n-j, f^j\omega\right) d^{(i)}\left(f^j\omega\right)\right\| &\le \left\| \mathcal{G}\left(n-j, f^j\omega\right)|_{E^{(i)}(f^j\omega)}\right\| \left\| d^{(i)}\left(f^j\omega\right)\right\| \\ &\le e^{(n-j)(\lambda_i+\epsilon_l)} e^{\epsilon_l(k+j)} \left\| d^{(i)}\left(f^j\omega\right)\right\|.\end{aligned}$$

Summing over j gives

$$\lambda_i < 0 \quad \Rightarrow \quad \left\| \left(\Psi^n d^{(i)}\right)(\omega)\right\| \le e^{k\epsilon_l} e^{n(\lambda_i+\epsilon_l)} \sum_{j=1}^{n} e^{j\lambda_1} \left\| d^{(i)}\left(f^j\omega\right)\right\|. \tag{7.13}$$

The rest of the analysis is similar to the previous analysis, now made simpler by the fact that $\lambda_i < 0$ and the right-hand side above is bounded uniformly with respect to n. This completes the proof of theorem. ∎

8. Proof of Proposition 2.1. Proposition 2.1 is a direct consequence of a result of J. Stark, which we state below.

Skew-product systems. Let $\tilde{\Omega}, Y$ be smooth manifolds, and let $T : (x, y) \mapsto (fx, g(x, y))$ be a skew-product map on $\tilde{\Omega} \times Y$. For every $n \in \mathbb{N}$, let $g^{(n)} : \tilde{\Omega} \times Y \to Y$ be the map such that

$$T^n(x, y) = \left(f^n x, g^{(n)}(x, y)\right) \quad \forall (x, y) \in \tilde{\Omega} \times Y.$$

Lemma 8.1 (invariant graphs for skew-product systems [73, Thm. 1.3]). *Assume the notation above, and suppose that the following hold:*

(i) *f is a $C^{1+\alpha}$ diffeomorphism and there are constants $\mu \ge 0, C_2 > 0$ such that $\|Df^{-n}\| \le C_2 e^{\mu n}$.*

(ii) *There is a closed and f-invariant subset $\Omega \subseteq \tilde{\Omega}$.*

(iii) *There exist constants $\lambda, C_3 > 0$ such that*

$$\text{Lip}\left(g^{(n)}(x, \cdot)\right) \le C_3 \exp(-\lambda n) \quad \forall x \in \tilde{\Omega}. \tag{8.1}$$

(iv) *g is uniformly $C^{1+\alpha}$ on compact sets.*

Then there is a continuous map $\Phi : \Omega \to Y$ such that the graph of Φ is invariant and globally attracting under T. Moreover, for every $\gamma \in (0, \alpha]$ such that $\mu(1+\gamma) < \lambda$, Φ is $C^{1+\gamma}$ in the Whitney sense.

Note that for every $n \in \mathbb{N}$, $x \in \tilde{\Omega}, y \in Y$,

$$g^{(1)} = g, \quad g^{(n+1)}(x, y) = g\left(f^n(\omega), g^{(n)}(x, y)\right).$$

For our purposes, set $\tilde{\Omega} = \Omega$ and $Y = \mathbb{R}^L$, and let f, g be smooth maps satisfying Assumptions 1 and 2 and (2.1). Then clearly all the conditions of Lemma 8.1 are satisfied. Thus there is a smooth map $\Phi : \Phi \to \mathbb{R}^L$ whose graph is invariant under T. This completes the proof of the proposition. ∎

9. Proof of Theorem 3.1.

9.1. Proof of claims (i), (ii). Claim (i) was proved by Dechert and Gençay. We restate their result using our terminology. Although they prove their result in the context of delay-coordinate maps, their proof is based on a commutation identity [24, eqn. (3.4)] which also holds in our more general case.

Lemma 9.1 (see [24, Thm. 3.1]). *Let M, N be C^1 manifolds of dimension m, n, respectively. Let $f : M \to M$ and $g : N \to N$ be two C^1 diffeomorphisms, conjugate via a C^1 map $J : M \to N$ as $g \circ J = J \circ f$. Let μ be an invariant ergodic measure μ of f. Let $\lambda_1(f,\mu) > \cdots > \lambda_r(f,\mu)$ be the distinct Lyapunov exponents of the ergodic system (f,μ), and let $E_1 \oplus \cdots \oplus E_r$ be the corresponding Oseledet splitting.*

(i) *For every $1 \le j \le r$, $\lambda_j = \lambda_j(f,\mu)$ is also a Lyapunov exponent of the ergodic system $(g, J_*\mu)$. The Oseledet subspace of TN corresponding to λ_j contains the subspace $DJ(E_j)$.*

(ii) *In particular, the Lyapunov exponents of g contains as a subset the Lyapunov exponents of f.*

In our case, the conjugation is via the map

$$h := (\phi, \Phi) : \Omega \to \mathbb{R}^{d+L}.$$

We next prove claim (ii). Under our assumption of ergodicity of μ, the Lyapunov exponents are constant μ-a.e. and coincide with their averages. The semicontinuity of averaged Lyapunov exponents is well known, either as functions of the map [11, Prop. 2.2.] or as a function of a cocycle over a fixed base dynamics [76, Rem. 1.4].

9.2. Proof of claim (iii). Fix a generic point $\omega_0 \in \operatorname{supp}(\mu)$ and set

$$z_0 := (\phi(\omega_0), \Phi(\omega_0)) = h(\omega_0).$$

z_0 is a point in X. To determine the maximal Lyapunov exponent of $\mathcal{T}$, we have to determine the maximum rate of deviation of orbits under perturbations. By Lemma 6.2, it is sufficient to consider the perturbation to occur either only in the first d coordinates or in the last L coordinates in the space $\mathbb{R}^{d+L}$. We call these ϕ-perturbations and Φ-perturbations, respectively.

ϕ-perturbations. First perturb z_0 to $z_0' = (\phi(\omega_0) + \vec{\delta}, \Phi(\omega_0))$ for some $\vec{\delta} \in \mathbb{R}^d$. Then

$$\mathcal{T}(z_0') = \mathcal{T}\left(\phi(\omega_0) + \delta, \Phi(\omega_0)\right) = \left(w \circ \Phi(\omega_0), g\left(\phi(\omega_0) + \delta, \Phi(\omega_0)\right)\right).$$

Therefore setting $\delta' = g(\phi(\omega_0) + \delta, \Phi(\omega_0)) - g(\phi(\omega_0), \Phi(\omega_0))$, we get

$$\mathcal{T}\left(z_0 + \begin{pmatrix} \delta \\ 0 \end{pmatrix}\right) = \mathcal{T}(z_0) + \begin{pmatrix} 0 \\ \delta' \end{pmatrix}, \quad \|\delta'\| \le \|\partial_1 g\|_{\sup} \delta. \tag{9.1}$$

Thus by Assumption 3, the map is contractive under ϕ-perturbations. In light of this observation, it is sufficient to bound the rate of growth of Φ-perturbations by $\lambda_1(f)$.

Φ-perturbations. Next perturb z_0 to $z_0' = (\phi(\omega_0), \Phi(\omega_0) + \vec{\delta}_0)$ for some $\vec{\delta}_0 \in \mathbb{R}^L$. Then

$$\mathcal{T}(z_0') = \mathcal{T}\begin{pmatrix} \phi(\omega_0) \\ \Phi(\omega_0) + \vec{\delta}_0 \end{pmatrix} = \begin{pmatrix} \hat{w}\left(\Phi(\omega_0) + \vec{\delta}_0\right) \\ g\left(\phi(\omega_0), \Phi(\omega_0) + \vec{\delta}_0\right) \end{pmatrix}.$$

We wish to show that if one starts with a Φ-perturbation of z_0, then after one iteration of $\mathcal{T}$, one still ends up with a Φ-perturbation with the image of a perturbed point. More precisely, we have the picture

(9.2)
$$\begin{array}{ccccc} z_0 & \xrightarrow{+\vec{\epsilon}_0} & z_0'' & \xrightarrow{\mathcal{T}} & z_1 \\ \downarrow{\scriptstyle +(0,\vec{\delta}_0)} & & & & \downarrow{\scriptstyle +(0,\vec{\delta}_1)} \\ z_0' & & \xrightarrow{\mathcal{T}} & & \mathcal{T}z_0' \end{array}$$

as described below.

Lemma 9.2. *Let Assumptions 1, 2, 3, and 4 hold. Let $z_0 = h(\omega_0) \in X$, and let z_0' be a Φ-perturbation of z_0 by a vector $\vec{\delta} \in \mathbb{R}^L$. Then there is a point $z_1 \in X$ such that the following hold:*

(i) *$z_1 = \mathcal{T}(z_0'')$ for some point $z_0'' \in X$ such that the perturbation $\vec{\epsilon}_0 := z_0'' - z_0$ has length at most $C_{\phi,\Phi,\mathrm{ret}}(\omega_0)\|\vec{\delta}_0\|$ from z_0, where*

$$C_{\phi,\Phi,\mathrm{ret}}(\omega) := (1 + C_{\phi,\Phi}(\omega))\,\kappa_{\mathrm{ret}} \quad \forall \omega \in \Omega.$$

(ii) *$\mathcal{T}(z_0')$ is a Φ-perturbation of $z_1 = \mathcal{T}(z_0'')$ with perturbation magnitude at most $[1 + C_{\phi,\Phi,\mathrm{ret}}(\omega_0)]\|\vec{\delta}_0\|$.*

Before proving Lemma 9.2 we show how its repeated application leads to a proof of Theorem 3.1 (iii). Repeated applications of (9.2) gives

(9.3)
$$\begin{array}{ccccccccccccc} z_0 & \xrightarrow{+\vec{\epsilon}_0} & z_0'' & \xrightarrow{\mathcal{T}} & z_1 & \xrightarrow{+\vec{\epsilon}_1} & z_1'' & \xrightarrow{\mathcal{T}} & z_2 & \cdots & z_n & \xrightarrow{+\vec{\epsilon}_n} & z_n'' & \xrightarrow{\mathcal{T}} & z_{n+1} & \cdots & \cdots \\ \downarrow{\scriptstyle +(0,\vec{\delta}_0)} & & & & \downarrow{\scriptstyle +(0,\vec{\delta}_1)} & & & & \downarrow{\scriptstyle +(0,\vec{\delta}_2)} & & \downarrow{\scriptstyle +(0,\vec{\delta}_n)} & & & & \downarrow{\scriptstyle +(0,\vec{\delta}_{n+1})} & & \\ z_0' & & \xrightarrow{\mathcal{T}} & & \mathcal{T}z_0' & & \xrightarrow{\mathcal{T}} & & \mathcal{T}^2 z_0' & \cdots & \mathcal{T}^n z_0' & & \xrightarrow{\mathcal{T}} & & \mathcal{T}^{n+1} z_0' & \cdots & \cdots \end{array}$$

Thus we have the following.

Lemma 9.3. *Let Assumptions 1, 2, 3, and 4 hold. Then for every $\omega_0 \in \Omega$ and any Φ-perturbation z_0' of $z_0 = h(\omega_0)$, there is a sequence of points $z_0'', z_1'', z_2'' \ldots \in X$ such that for every $n \in \mathbb{N}$, the following hold:*

(i) *$\mathcal{T}^n(z_0')$ is a Φ-perturbation of the point $z_n := \mathcal{T}(z_n'')$.*

(ii) *Since the points z_n'' and z_n lie on X, there is a sequence of points $\omega_n := h^{-1}(z_n)$ on Ω.*

(iii) *The perturbation $\vec{\delta}_n := \mathcal{T}^n(z_0') - z_n$ satisfies $\|\vec{\delta}_n\| \le [1 + C_{\phi,\Phi,\mathrm{ret}}(\omega_{n-1})]\|\vec{\delta}_{n-1}\|$.*

(iv) *The perturbation $\vec{\epsilon}_n := z_n'' - z_n$ satisfies $\|\vec{\epsilon}_n\| \le C_{\phi,\Phi,\mathrm{ret}}(\omega_n)\|\vec{\delta}_n\|$.*

Combining Lemma 9.3 (ii) and (iii) gives

$$\|\vec{\epsilon}_n\| = d\left(z_n'', z_n\right) \le \left\|\vec{\delta}_0\right\| C_{\phi,\Phi,\mathrm{ret}}(\omega_n) \prod_{i=0}^{n} \left[1 + C_{\phi,\Phi,\mathrm{ret}}(\omega_i)\right].$$

Thus the sequence z_n is a pseudotrajectory, and similarly to Proposition 6.3 it follows that

$$\inf_{\bar{w}\in\mathfrak{S}} \lambda_1(\bar{w}) - \lambda_1(f,\mu) \le \int \ln\left[1 + (1 + C_{\phi,\Phi}(\omega))\,\kappa_{\text{ret}}\right] d\mu(\omega).$$

This completes the proof of claim (iii) and of the theorem. ∎

9.3. Proof of Lemma 9.2. Lemma 9.2 is where Assumption 4 is needed. We first show how a neighborhood retraction of the attractor leads to an extension $\bar{w}$ of w.

$\bar{w}$ from retraction. Since ret is a retraction, we have $\text{ret}|X \equiv \text{Id}_X$. Let $\text{proj}_2 : \mathbb{R}^{d+L} \to \mathbb{R}^L$ be the projection onto the last L coordinates. Note that $\mathcal{U}_X := \text{proj}_2^{-1}(\mathcal{U})$ is a neighborhood of X in $\mathbb{R}^{d+L}$. Now define

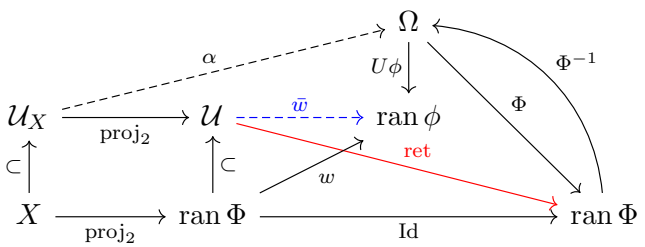
$$\bar{w} = (U\phi) \circ \Phi^{-1} \circ \text{ret} : \mathcal{U} \to \text{ran}\ \phi$$

and

$$\alpha := \Phi^{-1} \circ \text{ret} \circ \text{proj}_2 : \mathcal{U}_X \to \Omega.$$

Then we have the following commutations.

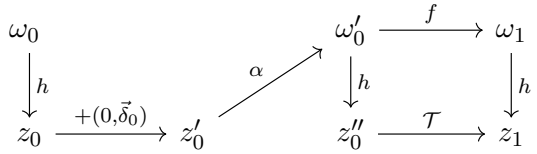
(9.4)



Note that by definition, α is a continuous map which coincides with $\Phi^{-1} \circ \text{proj}_2$ when restricted to ran Φ. Moreover,

$$U\phi \circ \alpha = \bar{w} \circ \text{proj}_2. \tag{9.5}$$

The construction. Set $\omega_0' = \alpha(z_0')$ and $z_0'' := h(\omega_0')$ and $\omega_1 := f(\omega_0')$, as shown below:



Proof of claim (i). We first obtain a bound for $\Phi(\omega_0') - \Phi(\omega_0)$. Note that

$$\begin{aligned}\Phi(\omega_0') &= \Phi \circ \alpha\left(z_0 + (0,\vec{\delta}_0)\right) = \Phi \circ \Phi^{-1} \circ \text{ret} \circ \text{proj}_2\left(z_0 + (0,\vec{\delta}_0)\right) \\ &= \text{ret} \circ \text{proj}_2\left(z_0 + (0,\vec{\delta}_0)\right) = \text{ret}\left(\Phi(\omega_0) + \vec{\delta}_0\right).\end{aligned}$$

Therefore

$$\left\|\Phi(\omega_0') - \Phi\left(\omega_0\right)\right\| = \left\|\text{ret}\left(\Phi(\omega_0) + \vec{\delta}_0\right) - \text{ret}\left(\Phi(\omega_0)\right)\right\| \le \kappa_{\text{ret}} \left\|\vec{\delta}_0\right\|. \tag{9.6}$$

We next estimate the gap $\phi(\omega_0') - \phi(\omega_0)$. By the definition of the constant $C_{\phi,\Phi}(\omega_0)$ and by (9.6),

$$\left\|\phi(\omega_0') - \phi(\omega_0)\right\| \leq C_{\phi,\Phi}(\omega_0) \left\|\Phi(\omega_0') - \Phi(\omega_0)\right\| \leq C_{\phi,\Phi}(\omega_0)\kappa_{\text{ret}} \left\|\vec{\delta}_0\right\|. \tag{9.7}$$

Equip the space $\mathbb{R}^{d+L}$ with the norm $\|(x,y)\|_{\mathbb{R}^{d+L}} := \|x\|_{\mathbb{R}^d} + \|y\|_{\mathbb{R}^L}$. Then we have

$$\begin{aligned}
\vec{\epsilon}_0 = \left\|z_0'' - z_0'\right\| &= \left\| \begin{bmatrix} \phi(\omega_0') \\ \Phi(\omega_0') \end{bmatrix} - \begin{bmatrix} \phi(\omega_0) \\ \Phi(\omega_0) \end{bmatrix} \right\| = \left\|\phi(\omega_0') - \phi(\omega_0)\right\| + \left\|\Phi(\omega_0') - \Phi(\omega_0) - \delta_0\right\| \\
&\leq \left(1 + C_{\phi,\Phi}(\omega_0)\right) \kappa_{\text{ret}} \left\|\vec{\delta}_0\right\| \quad [(9.6),(9.7)], \\
&= C_{\phi,\Phi,\text{ret}}(\omega_0) \left\|\vec{\delta}_0\right\|.
\end{aligned}$$

Similarly, we have

$$\left\|z_0'' - z_0'\right\| \leq \left\|z_0'' - z_0\right\| + \left\|z_0 - z_0'\right\| = \left[1 + C_{\phi,\Phi,\text{ret}}(\omega_0)\right] \left\|\vec{\delta}_0\right\|. \tag{9.8}$$

This completes the proof of claim (i).

Proof of claim (ii). Next, by the contractiveness of g from Assumption 3,

$$\left\|g\left(z_0''\right) - g\left(z_0'\right)\right\| \leq \left\|z_0'' - z_0'\right\| \overset{\text{by (9.8)}}{\leq} \left[1 + C_{\phi,\Phi,\text{ret}}(\omega_0)\right] \left\|\vec{\delta}_0\right\|. \tag{9.9}$$

We have from definition that

$$\text{proj}_2(z_1) = \text{proj}_2 \circ h(\omega_1) = \text{proj}_2 \circ h \circ f(\omega_0') = \text{proj}_2 \circ \mathcal{T} \circ h(\omega_0') = g(z_0'').$$

Thus

$$\left\|\text{proj}_2(z_1) - \text{proj}_2 \circ \mathcal{T}(z_0')\right\| = \left\|g(z_0'')) - g(z_0')\right\| \overset{\text{by (9.9)}}{\leq} \left[1 + C_{\phi,\Phi,\text{ret}}(\omega_0)\right] \left\|\vec{\delta}_0\right\|. \tag{9.10}$$

Finally set $y' := \Phi(\omega_0) + \vec{\delta}$. Then note that

$$\begin{aligned}
\text{proj}_1(z_1) &= \text{proj}_1 \circ h(\omega_1) = \text{proj}_1 \circ h \circ f(\omega_0') = \text{proj}_1 \circ \mathcal{T} \circ h(\omega_0') = w \circ \text{proj}_2 \circ h(\omega_0') \\
&= w \circ \Phi(\omega_0') = w \circ \Phi \circ \Phi^{-1} \circ \text{ret}(y') = \bar{w}(y') \\
&= \text{proj}_1 \circ \mathcal{T}(z_0').
\end{aligned}$$

This completes the proof of claim (ii) and thus of the lemma. ∎

9.4. Proof of Corollary 3.4. Corollary 3.4 satisfies the conditions of Theorem 3.1. The claim will be proved if it can be shown that $C_{\phi,\Phi} \leq \frac{1}{Q} + QO(\Delta t)$.

$$D\Phi(\omega) = \left(D\Phi\left(\Psi^{0\Delta t}\omega\right), D\Phi\left(\Psi^{-\Delta t}\omega\right), \ldots, D\Phi\left(\Psi^{-(Q-1)\Delta t}\omega\right)\right).$$

Since ϕ is a C^2 function,

$$\left\|D\phi\left(\Psi^{q\Delta t}\omega\right) - D\Phi(\omega)\right\| = O\left(|q|\Delta t\right) \text{ as } \Delta t \to 0^+.$$

Thus

$$\|D\Phi(\omega)\| = Q\,\|D\phi\,(\omega)\| + \sum_{q=0}^{Q} O\,(q\Delta t) = Q\,\|D\phi\,(\omega)\| + Q^2 O\,(\Delta t) \;\text{ as } \Delta t \to 0^+.$$

Therefore

$$C_{\phi,\Phi}(\omega) = \frac{\|D\phi(\omega)\|}{\|D\Phi(\omega)\|} = \frac{\|D\phi(\omega)\|}{Q\,\|D\phi(\omega)\| + Q^2 O\,(\Delta t)} = \frac{1}{Q} + QO\,(\Delta t) \;\text{ as } \Delta t \to 0^+.$$

This proves the claim. ∎

10. Proof of Theorem 4.1. By (4.4), the direct forecast error can be expressed in terms of the Koopman operator as

$$\text{error}_{\text{direct}}(n) := \|U^n\phi - \text{proj}_{\mathcal{W}} U^n\phi\|_{L^2(\mu)} = \|(\text{Id} - \text{proj}_{\mathcal{W}})\,U^n\phi\|_{L^2(\mu)}\,,$$

as claimed. To proceed further, we have to separately examine the components of ϕ along $\mathcal{D}$ and its complement. For that purpose, define

$$\phi^{(d)} := \text{proj}_{\mathcal{D}}\phi, \quad \phi^{(c)} := \phi - \phi^{(d)}.$$

This decomposition is possible due to the linearity of U^n and the invariance of the subspaces $\mathcal{D}, \mathcal{D}^\perp$. Therefore

$$\begin{aligned} \text{error}_{\text{direct}}(n)^2 &= \|(\text{Id} - \text{proj}_{\mathcal{W}})\,U^n\phi\|^2_{L^2(\mu)} \\ &= \left\|(\text{Id} - \text{proj}_{\mathcal{W}})\,U^n\phi^{(c)}\right\|^2_{L^2(\mu)} + \left\|(\text{Id} - \text{proj}_{\mathcal{W}})\,U^n\phi^{(d)}\right\|^2_{L^2(\mu)}. \end{aligned} \tag{10.1}$$

We call them the discrete and continuous components, respectively, and analyze them separately.

Continuous component. We begin with a review of some concepts from ergodic theory related to mixing.

Lemma 10.1 (weak mixing). *Let (Ω, μ, f) be a measure preserving system, with the splitting as in (1.1). Then for every $\phi_1, \in \mathcal{D}^\perp$ and every $\phi_2 \in L^2(\mu)$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left|\langle \phi_1, U^n\phi_2\rangle_{L^2(\mu)} - \mu\,(\phi_1)\,\mu\,(\phi_2)\right| = 0,$$
$$\lim_{N\in\mathbb{N}', N\to\infty} \langle \phi_1, U^n\phi_2\rangle_{L^2(\mu)} = \mu\,(\phi_1)\,\mu\,(\phi_2)\,,$$

where $\mathbb{N}'$ is a subset $\mathbb{N}$ with density 1.

Proof. The first identity follows from [39, Mixing Theorem, p. 45]. The second identity follows from [55, sect. 2.1]. ∎

If $\mathbb{N}'$ above can be taken to be $\mathbb{N}$ and $\mathcal{D} = \{\text{constant}\}$, then the system (Ω, f, μ) will be called *strongly mixing*. In other words the following holds:

$$\lim_{N\to\infty} \left\langle \phi_1, U^N\phi_2\right\rangle_{L^2(\mu)} = \mu\,(\phi_1)\,\mu\,(\phi_2) \quad \forall \phi_1, \phi_2 \in L^2(\mu). \tag{10.2}$$

We are now ready to prove the following:

$$\lim_{n\in\mathbb{N}',n\to\infty} \left\| (\mathrm{Id} - \mathrm{proj}_{\mathcal{W}})\, U^n \phi^{(c)} \right\|_{L^2(\mu)} = \left\| \phi^{(c)} \right\|_{L^2(\mu)}. \tag{10.3}$$

We will in fact prove the stronger result

$$\lim_{n\in\mathbb{N}',n\to\infty} \mathrm{proj}_{\mathcal{W}} U^n \phi^{(c)} = 0. \tag{10.4}$$

Proof of (10.4). By (4.3), $\mathcal{W}$ is spanned by a finite orthonormal basis $\{\mathfrak{w}_i : i = 1, \dots, M\}$. Then note that

$$\mathrm{proj}_{\mathcal{W}} \psi = \sum_{i=1}^{M} \langle \mathfrak{w}_i, \psi \rangle_{L^2(\mu)} \mathfrak{w}_i \quad \forall \psi \in L^2(\mu).$$

Therefore,

$$\begin{aligned}\lim_{n\in\mathbb{N}',n\to\infty} \mathrm{proj}_{\mathcal{W}} U^n \phi^{(c)} &= \lim_{n\in\mathbb{N}',n\to\infty} \sum_{i=1}^{M} \langle \mathfrak{w}_i, U^n \phi^{(c)} \rangle_{L^2(\mu)} \mathfrak{w}_i \\ &= \sum_{i=1}^{M} \lim_{n\in\mathbb{N}',n\to\infty} \langle \mathfrak{w}_i, U^n \phi^{(c)} \rangle_{L^2(\mu)} \mathfrak{w}_i = \sum_{i=1}^{M} \mu(\mathfrak{w}_i) \mu(\phi^{(c)}) \mathfrak{w}_i = 0.\end{aligned}$$

The identity in (10.4) now follows.

Discrete component. Next, let $z_1, z_2, \dots$ be an orthonormal basis for $\mathcal{D}$ in terms of the Koopman eigenfunctions. Then one has

$$\mathrm{proj}_{\mathcal{D}} \mathfrak{w}_l = \sum_j a_{l,j} z_j, \quad 1 \le l \le M, \quad a_{l,j} := \langle z_j, \mathfrak{w}_l \rangle_{L^2(\mu)}.$$

Let Π be the $\mathbb{N} \times \mathbb{N}$ matrix defined as $\Pi_{j,k} := \langle \pi z_j, \pi z_k \rangle_{L^2\mu)}$. Then

$$\begin{aligned}\sum_{l=1}^{M} a_{l,k}^* a_{l,j} &= \sum_{l=1}^{M} \langle \mathfrak{w}_l, z_k \rangle_{L^2(\mu)} \langle z_j, \mathfrak{w}_l \rangle_{L^2(\mu)} = \left\langle \sum_{l=1}^{M} \langle z_k, \mathfrak{w}_l \rangle_{L^2(\mu)} \mathfrak{w}_l, \sum_{l=1}^{M} \langle z_j, \mathfrak{w}_l \rangle_{L^2(\mu)} \mathfrak{w}_l \right\rangle_{L^2(\mu)} \\ &= \langle \pi z_j, \pi z_k \rangle_{L^2(\mu)} = \Pi_{j,k}.\end{aligned}$$

Now let $\phi^{(d)} = \sum_j \phi_j z_j$. Then $U^n \phi^{(d)} = \sum_j \phi_j e^{\iota \omega_j n} z_j$. Therefore,

$$\left\langle \mathfrak{w}_l, U^n \phi^{(d)} \right\rangle_{L^2(\mu)} = \left\langle \mathrm{proj}_{\mathcal{D}} \mathfrak{w}_l, U^n \phi^{(d)} \right\rangle_{L^2(\mu)} = \sum_j \phi_j e^{\iota \omega_j n} \langle \mathfrak{w}_l, z_j \rangle_{L^2(\mu)} = \sum_j \phi_j e^{\iota \omega_j n} a_{l,j}^*.$$

Therefore,

$$\pi U^n \phi^{(d)} = \sum_{l=1}^{M} \langle \mathfrak{w}_l, U^n \phi \rangle_{L^2(\mu)} \mathfrak{w}_l = \sum_{l=1}^{M} \sum_j \phi_j e^{\iota \omega_j n} a_{l,j}^* \mathfrak{w}_l.$$

Note that the infinite sequence $\vec{\phi} := (\phi_j)_{j\in\mathbb{N}}$ is an ℓ^2 sequence. Define the operator

$$\mathcal{F} : \ell^2 \to \ell^2, \quad \left(\mathcal{F}\vec{\phi}\right)_j := e^{\iota\omega_j}\phi_j \quad \forall j \in \mathbb{N}.$$

Then

$$\left\|\pi U^n \phi^{(d)}\right\|^2_{L^2(\mu)} = \sum_{k,j} \phi_k^* \phi_j e^{\iota\omega_j n} e^{-\iota\omega_k n} \sum_{l=1}^{M} a_{l,k}^* a_{l,j} = \left(\mathcal{F}^n\vec{\phi}\right)^* \Pi \left(\mathcal{F}^n\vec{\phi}\right).$$

Therefore

$$\left\|(\mathrm{Id} - \mathrm{proj}_{\mathcal{W}})\, U^n \phi^{(d)}\right\|^2_{L^2(\mu)} = \vec{\phi}^* \mathcal{F}^{n*} \left[\mathrm{Id} - \Pi\right] \mathcal{F}^n \vec{\phi}. \tag{10.5}$$

The operator $\mathcal{F}$ is a unitary operator which is diagonal with respect to the usual basis of ℓ^2. Thus

$$\lim_{\Pi \to \mathrm{Id}} \left\|(\mathrm{Id} - \mathrm{proj}_{\mathcal{W}})\, U^n \phi^{(d)}\right\|^2_{L^2(\mu)} = 0. \tag{10.6}$$

If the hypothesis space is increased, then Π converges strongly to Id and the above limit is approached. Thus for any $\epsilon > 0$, if $\mathcal{W}$ is large enough, then $\|(\mathrm{Id} - \mathrm{proj}_{\mathcal{W}})U^n\phi^{(d)}\|^2_{L^2(\mu)} < \epsilon$.

Proof of Theorem 4.1. Claim (i) follows from (10.1), (10.4), and (10.6). In claim (ii), if (f, μ) is weakly mixing, then $\mathcal{D} = \{\text{constant}\}$, and thus $\phi^{(d)}$ is just the average $\mu(\phi)$. The claim now follows from (10.1) and (10.4). Claim (iii) follows from the definition of strong mixing and (10.2). In claim (iv), $\mathcal{D}^\perp = \{0\}$, and the claim follows from (10.6). ∎

11. Proof of Theorem 4.2. We next look at the iterates of the map $\mathcal{T}$ in (1.4), with initial conditions in (1.6). Let $\hat{w} = \hat{w}_1$ as in (4.4). The following identity will be used repeatedly:

$$\hat{w}\left(U^n\Phi\right) = \hat{w} \circ \Phi \circ f^n \overset{\text{by (4.4)}}{=} (\mathrm{proj}_{\mathcal{W}} U\phi) \circ f^n = U^n \mathrm{proj}_{\mathcal{W}} U\phi = U^n \pi U \phi \quad \forall n \in \mathbb{N}. \tag{11.1}$$

The proof of (4.10) will be by induction on n. For the base case, note that

$$z_1 = z_1(\omega_0) = \mathcal{T}(z_0) = \left[\hat{w}_1\left(\Phi\right), g \circ (\phi, \Phi)\right] \overset{(4.4)}{=} \left[\pi U\phi, U\Phi\right],$$

and thus $\Delta u_1 = a_1 = 0^d$ and $\Delta y_1 = b_1 = 0^L$. Next suppose that the statement is true up to some $n \in \mathbb{N}$. Using the notation in (4.9) we have

$$\begin{aligned} u_{n+1} = \hat{w}\left(y_n\right) &= \hat{w}\left(U^n\Phi - \Delta y_n\right) = \hat{w}\left(U^n\Phi\right) - D\hat{w}|_{U^n\Phi}\Delta y_n + O\left(\|\Delta y_n\|^2\right) \\ &= U^n\pi U\phi - \hat{W}\left(f^n(\cdot)\right)\Delta y_n + O\left(\|\Delta y_n\|^2\right) \quad \text{by (4.6), (11.1).} \end{aligned}$$

So

$$\Delta u_{n+1} := U^n\pi U\phi - u_{n+1} = \hat{W}\left(f^n(\cdot)\right)\Delta y_n + O\left(\|\Delta y_n\|^2\right).$$

Similarly,

$$\begin{aligned} y_{n+1} &= g(u_n, y_n) = g\left(U^{n-1}\pi U\phi - \Delta u_n, U^n\Phi - \Delta y_n\right) \\ &= g\left(U^n\phi - \Delta u_n - U^{n-1}\Delta U\phi, U^n\Phi - \Delta y_n\right) \\ &= g\left(U^n\phi, U^n\Phi\right) - \nabla_1 g|_{h\circ f^n}\left(\Delta u_n + U^{n-1}\Delta U\phi\right) \\ &\quad - \nabla_2 g|_{h\circ f^n}\Delta y_n + O\left(\|\Delta u_n\|^2\right) + O\left(\|\Delta y_n\|^2\right) \\ &= U^{n+1}\Phi - G^{(1)}\left(f^n(\cdot)\right)\Delta u_n - G^{(2)}\left(f^n(\cdot)\right)\Delta y_n + c\left(f^n(\cdot)\right) \\ &\quad + O\left(\|\Delta u_n\|^2\right) + O\left(\|\Delta y_n\|^2\right). \end{aligned}$$

So

$$\begin{aligned} \Delta y_{n+1} &:= U^{n+1}\Phi - y_{n+1} \\ &= G^{(1)}\left(f^n(\cdot)\right)\Delta u_n + G^{(2)}\left(f^n(\cdot)\right)\Delta y_n + c\left(f^n(\cdot)\right) + O\left(\|\Delta u_n\|^2\right) + O\left(\|\Delta y_n\|^2\right). \end{aligned}$$

Combining we get

$$\begin{bmatrix} u_{n+1} \\ y_{n+1} \end{bmatrix} = \hat{M}\left(f^n(\cdot)\right)\begin{bmatrix} u_n \\ y_n \end{bmatrix} + \begin{bmatrix} 1 \\ c\left(f^n(\cdot)\right) \end{bmatrix} + \begin{bmatrix} O\left(\|\Delta y_n\|^2\right) \\ O\left(\|\Delta u_n\|^2\right) + O\left(\|\Delta y_n\|^2\right) \end{bmatrix}. \tag{11.2}$$

The evolution equation (11.2) for (u_n, y_n) is thus the addition of the Taylor series error terms to the evolution equation (4.8) for (a_n, b_n). Claim (i) and (4.10) immediately follow. Since the evolution of (a_n, b_n) is that of a perturbed random cocycle, Theorem 7.1 applies and Claims (ii) and (iii) follow. This completes the proof of Theorem 4.2. ∎

REFERENCES

[1] R. Alexander and D. Giannakis, *Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques*, Phys. D, 409 (2020), 132520, https://doi.org/10.1016/j.physd.2020.132520.

[2] J. Alves, S. Luzzatto, and V. Pinheiro, *Markov structures and decay of correlations for non-uniformly expanding dynamical systems*, Ann. Inst. H. Poincaré C Non Linéaire, 22 (2005), pp. 817–839, https://doi.org/10.1016/j.anihpc.2004.12.002.

[3] J. Alves, S. Luzzatto, and V. Pinheiro, *Lyapunov exponents and rates of mixing for one-dimensional maps*, Ergodic Theory Dynam. Systems, 24 (2004), pp. 637–657, https://doi.org/10.1017/S0143385703000579.

[4] L. Arnold, *Random Dynamical Systems*, Springer, 1991, https://doi.org/10.1007/978-3-662-12878-7.

[5] L. Barreira and C. Valls, *Smoothness of invariant manifolds for nonautonomous equations*, Comm. Math. Phys., 259 (2005), pp. 639–677, https://doi.org/10.1007/s00220-005-1380-z.

[6] L. Barreira and C. Valls, *Stable manifolds for nonautonomous equations without exponential dichotomy*, J. Differential Equations, 221 (2006), pp. 58–90, https://doi.org/10.1016/j.jde.2005.04.005.

[7] T. Berry, J. R. Cressman, Z. Gregurić-Ferenček, and T. Sauer, *Time-scale separation from diffusion-mapped delay coordinates*, SIAM J. Appl. Dyn. Syst., 12 (2013), pp. 618–649, https://doi.org/10.1137/12088183x.

[8] T. Berry, D. Giannakis, and J. Harlim, *Nonparametric forecasting of low-dimensional dynamical systems*, Phys. Rev. E, 91 (2015), 032915. https://doi.org/10.1103/PhysRevE.91.032915.

[9] T. Berry and J. Harlim, *Correcting biased observation model error in data assimilation*, Monthly Weather Rev., 145 (2017), pp. 2833–2853, https://doi.org/10.1175/MWR-D-16-0428.1.

[10] A. Blumenthal and L. Young, *Equivalence of physical and SRB measures in random dynamical systems*, Nonlinearity, 32 (2019), pp. 1494–1524, https://doi.org/10.1088/1361-6544/aafaa8.
[11] J. Bochi and M. Viana, *The Lyapunov exponents of generic volume-preserving and symplectic maps*, Ann. of Math. (2), 161 (2005), pp. 1423–1485, https://doi.org/10.4007/annals.2005.161.1423.
[12] E. Bollt, *On explaining the surprising success of reservoir computing forecaster of chaos? The universal machine learning dynamical system with contrast to VAR and DMD*, Chaos, 31 (2021), 013108, https://doi.org/10.1063/5.0024890.
[13] G. Box, G. Jenkins, G. Reinsel, and G. Ljung, *Time Series Analysis: Forecasting and Control*, John Wiley & Sons, 2015.
[14] J. Bramburger, S. Brunton, and J. Kutz, *Deep learning of conjugate mappings*, Phys. D, 427 (2021), 133008, https://doi.org/10.1016/j.physd.2021.133008.
[15] M. Casdagli, *Nonlinear prediction of chaotic time series*, Phys. D, 35 (1989), pp. 335–356, https://doi.org/10.1016/0167-2789(89)90074-2.
[16] R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30, https://doi.org/10.1016/j.acha.2006.04.006.
[17] P. Collet and J.-P. Eckmann, *Liapunov multipliers and decay of correlations in dynamical systems*, J. Stat. Phys., 115 (2004), pp. 217–254, https://doi.org/10.1023/B:JOSS.0000019817.71073.61.
[18] C. Danforth and J. Yorke, *Making forecasts for chaotic physical processes*, Phys. Rev. Lett., 96 (2006), 144102, https://doi.org/10.1103/PhysRevLett.96.144102.
[19] S. Das and D. Giannakis, *Delay-coordinate maps and the spectra of Koopman operators*, J. Stat. Phys., 175 (2019), pp. 1107–1145, https://doi.org/10.1007/s10955-019-02272-w.
[20] S. Das and D. Giannakis, *Koopman spectra in reproducing kernel Hilbert spaces*, Appl. Comput. Harmon. Anal., 49 (2020), pp. 573–607, https://doi.org/10.1016/j.acha.2020.05.008.
[21] S. Das and D. Giannakis, *Reproducing kernel Hilbert algebras on compact Lie groups*, J. Funct. Anal. Appl., 29 (2023), https://doi.org/10.1007/s00041-023-09992-4.
[22] S. Das, S. Mustavee, and S. Agarwal, *Uncovering Quasi-Periodic Nature of Physical Systems: A Case Study of Signalized Intersections*, https://arxiv.org/abs/2109.08623, 2021.
[23] S. Das and J. Yorke, *Super convergence of ergodic averages for quasiperiodic orbits*, Nonlinearity, 31 (2018), pp. 491–501, https://doi.org/10.1088/1361-6544/aa99a0.
[24] W. Dechert and R. Gençay, *The topological invariance of Lyapunov exponents in embedded dynamics*, Phys. D, 90 (1996), pp. 40–55, https://doi.org/10.1016/0167-2789(95)00225-1.
[25] W. Dechert and R. Gençay, *Is the largest Lyapunov exponent preserved in embedded dynamics?*, Phys. Lett. A, 276 (2000), pp. 59–64, https://doi.org/10.1016/S0375-9601(00)00657-5.
[26] P. Duarte and S. Klein, *Lyapunov Exponents of Linear Cocycles*, Atlantis Stud. Dynam. Sys. 3, Atlantis Press, 2016, https://doi.org/10.2991/978-94-6239-124-6.
[27] G. Froyland, S. Lloyd, and A. Quas, *Coherent structures and isolated spectrum for Perron-Frobenius cocycles*, Ergodic Theory Dynam. Systems, 30 (2010), pp. 729–756, https://doi.org/10.1017/S0143385709000339.
[28] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Stat., 31 (1960), pp. 457–469, https://doi.org/10.1214/aoms/1177705909.
[29] H. Furstenberg and Y. Kifer, *Random matrix products and measures on projective spaces*, Israel J. Math., 46 (1983), pp. 12–32, https://doi.org/10.1007/BF02760620.
[30] Gallant, *Nonlinear regression*, Amer. Statist., 29 (1975), pp. 73–81, https://doi.org/10.1080/00031305.1975.10477374.
[31] D. Gauthier, E. Bollt, A. Griffith, and W. Barbosa, *Next generation reservoir computing*, Nature Comm., 12 (2021), pp. 1–8, https://doi.org/10.1038/s41467-021-25801-2.
[32] D. Giannakis, S. Das, and J. Slawinska, *Reproducing kernel Hilbert space compactification of unitary evolution groups*, Appl. Comput. Harmon. Anal., 54 (2021), pp. 75–136, https://doi.org/10.1016/j.acha.2021.02.004.
[33] D. Giannakis, J. Slawinska, and Z. Zhao, *Spatiotemporal feature extraction with data-driven Koopman operators*, J. Mach. Learn. Res. Proc., 44 (2015), pp. 103–115.
[34] L. Gonon and J. Ortega, *Fading memory echo state networks are universal*, Neural Networks, 138 (2021), pp. 10–13, https://doi.org/10.1016/j.neunet.2021.01.025.

[35] S. Gouëzel and L. Stoyanov, *Quantitative Pesin theory for Anosov diffeomorphisms and flows*, Ergodic Theory Dynam. Systems, 39 (2019), pp. 159–200, https://doi.org/10.1017/etds.2017.25.
[36] L. Grigoryeva, A. Hart, and J. Ortega, *Chaos on compact manifolds: Differentiable synchronizations beyond the Takens theorem*, Phys. Rev. E, 103 (2021), 062204, https://doi.org/10.1103/PhysRevE.103.062204.
[37] L. Grigoryeva, A. Hart, and J. Ortega, *Learning Strange Attractors with Reservoir Systems*, https://arxiv.org/abs/2108.05024, 2021.
[38] L. Grigoryeva, J. Henriques, L. Larger, and J. Ortega, *Stochastic nonlinear time series forecasting using time-delay reservoir computers: Performance and universality*, Neural Networks, 55 (2014), pp. 59–71, https://doi.org/10.1016/j.neunet.2014.03.004.
[39] P. Halmos, *Lectures on Ergodic Theory*, American Mathematical Society, 1956.
[40] F. Hamilton, T. Berry, and T. Sauer, *Ensemble Kalman filtering without a model*, Phys. Rev. X, 6 (2016), 011021, https://doi.org/10.1103/PhysRevX.6.011021.
[41] F. Hamilton, T. Berry, and T. Sauer, *Kalman-Takens filtering in the presence of dynamical noise*, Eur. Phys. J. Special Topics, 226 (2017), pp. 3239–3250, https://doi.org/10.1140/epjst/e2016-60363-2.
[42] S. Hammel, J. Yorke, and C. Grebogi, *Numerical orbits of chaotic processes represent true orbits*, Bull. Amer. Math. Soc., 19 (1988), pp. 465–469, https://doi.org/10.1090/S0273-0979-1988-15701-1.
[43] J. Harlim, S. Jiang, S. Liang, and H. Yang, *Machine learning for prediction with missing dynamics*, J. Comput. Phys., 428 (2021), 109922, https://doi.org/10.1016/j.jcp.2020.109922.
[44] A. Hart, J. Hook, and J. Dawes, *Embedding and approximation theorems for echo state networks*, Neural Networks, 128 (2020), pp. 234–247, https://doi.org/10.1016/j.neunet.2020.05.013.
[45] A. Hart, J. Hook, and J. Dawes, *Echo state networks trained by Tikhonov least squares are $L^2(\mu)$ approximators of ergodic dynamical systems*, Phys. D, 421 (2021), 132882, https://doi.org/10.1016/j.physd.2021.132882.
[46] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput., 9 (1997), pp. 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[47] H. Jaeger and H. Haas, *Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication*, Science, 304 (2004), pp. 78–80, https://doi.org/10.1126/science.1091277.
[48] K. Jimenez, J. Moreno, and G. Ruggeri, *Forecasting on chaotic time series: A local optimal linear-reconstruction method*, Phys. Rev. A, 45 (1992), pp. 3553–3558, https://doi.org/10.1103/PhysRevA.45.3553.
[49] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publ. Math. Inst. Hautes Études Sci., 51 (1980), pp. 137–173, https://doi.org/10.1007/BF02684777.
[50] M. Korda and I. Mezić, *Optimal construction of Koopman eigenfunctions for prediction and control*, IEEE Trans. Automat. Control, 65 (2020), pp. 5114–5129, https://doi.org/10.1109/TAC.2020.2978039.
[51] D. Kugiumtzis, O. C. Lingjaerde, and N. Christophersen, *Regularized local linear prediction of chaotic time series*, Phys. D, 112 (1998), pp. 344–360, https://doi.org/10.1016/s0167-2789(97)00171-1.
[52] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444, https://doi.org/10.1038/nature14539.
[53] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms: Part* I: *Characterization of measures satisfying Pesin's entropy formula*, Ann. of Math. (2), (1985), pp. 509–539, https://doi.org/10.2307/1971328.
[54] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms: Part* II: *Relations between entropy, exponents and dimension*, Ann. of Math. (2), (1985), pp. 540–574, https://doi.org/10.2307/1971329.
[55] G. Liao, W. Sun, E. Vargas and S. Wang, *Approximation of Bernoulli measures for non-uniformly hyperbolic systems*, Ergodic Theory Dynam. Systems, 40 (2022), pp. 233–247, https://doi.org/10.1017/etds.2018.33.
[56] M. Logunov and O. Butkovskii, *Mixing and Lyapunov exponents of chaotic systems*, Tech. Phys., 53 (2008), pp. 959–965, https://doi.org/10.1134/S106378420808001X.

[57] Z. Lu, J. Pathak, B. Hunt, M. Girvan, R. Brockett, and E. Ott, *Reservoir observers: Model-free inference of unmeasured variables in chaotic systems*, Chaos, 27 (2017), p. 041102, https://doi.org/10.1063/1.4979665.

[58] M. Lukoševičius and H. Jaeger, *Reservoir computing approaches to recurrent neural network training*, Comput. Sci. Rev., 3 (2009), pp. 127–149, https://doi.org/10.1016/j.cosrev.2009.03.005.

[59] C. Ma, J. Wang, and W. E, *Model reduction with memory and the machine learning of dynamical systems*, Commun. Comput. Phys., 25 (2018), pp. 947–962, https://doi.org/10.4208/CICP.OA-2018-0269.

[60] M. Matthews, *On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*, Ph.D. thesis, ETH Zurich, 1992.

[61] R. Maulik, A. Mohan, B. Lusch, S. Madireddy, P. Balaprakash, and D. Livescu, *Time-series learning of latent-space dynamics for reduced-order model closure*, Phys. D, 405 (2020), 132368, https://doi.org/10.1016/j.physd.2020.132368.

[62] S. Mustavee, S. Agarwal, C. Enyioha, and S. Das, *A linear dynamical perspective on epidemiology: Interplay between early Covid-19 outbreak and human activity*, Nonlinear Dyn., 109 (2022), pp. 1233–1252, https://doi.org/10.1007/s11071-022-07469-5.

[63] J. Pathak, Z. Lu, R. B. Hunt, M. Girvan, and E. Ott, *Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data*, Chaos, 27 (2017), 121102, https://doi.org/10.1063/1.5010300.

[64] Y. Pesin, *Families of invariant manifolds corresponding to nonzero characteristic exponents*, Math. USSR-Izvestiya, 10 (1976), p. 1261, https://doi.org/10.1070/IM1976v010n06ABEH001835.

[65] R. Rahman and S. Hasan, *Real-time signal queue length prediction using long short-term memory neural network*, Neural Comp. Appl., 33 (2021), pp. 3311–3324, https://doi.org/10.1007/s00521-020-05196-9.

[66] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Publ. Math. Inst. Hautes Études Sci., 50 (1979), pp. 27–58, https://doi.org/10.1007/BF02684768.

[67] T. Sauer, J. Tempkin, and J. Yorke, *Spurious Lyapunov exponents in attractor reconstruction*, Phys. Rev. Lett, 81 (1998), pp. 4341–4344, https://doi.org/10.1103/PhysRevLett.81.4341.

[68] T. Sauer, J. A. Yorke, and M. Casdagli, *Embedology*, J. Stat. Phys., 65 (1991), pp. 579–616, https://doi.org/10.1007/bf01053745.

[69] J. Skufca and E. Bollt, *Relaxing conjugacy to fit modeling in dynamical systems*, Phys. Rev. E, 76 (2007), 026220, https://doi.org/10.1103/PhysRevE.76.026220.

[70] J. Skufca and E. Bollt, *A concept of homeomorphic defect for defining mostly conjugate dynamical systems*, Chaos, 18 (2008), 013118, https://doi.org/10.1063/1.2837397.

[71] J. Slawinska and D. Giannakis, *Indo-Pacific variability on seasonal to multidecadal time scales. Part* I: *Intrinsic SST modes in models and observations*, J. Climate, 30 (2017), pp. 5265–5294, https://doi.org/10.1175/JCLI-D-16-0176.1.

[72] L. Smith, *Identification and prediction of low dimensional dynamics*, Phys. D, 58 (1992), pp. 50–76, https://doi.org/10.1016/0167-2789(92)90101-R.

[73] J. Stark, *Regularity of invariant graphs for forced systems*, Ergodic Theory Dynam. Systems, 19 (1999), pp. 155–199, https://doi.org/10.1017/S0143385799126555.

[74] G. Sugihara and R. M. May, *Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series*, Nature, 344 (1990), pp. 734–741, https://doi.org/10.1038/344734a0.

[75] N. Trillos and D. Slepčev, *A variational approach to the consistency of spectral clustering*, Appl. Comput. Harmon. Anal., 45 (2018), pp. 239–281, https://doi.org/10.1016/j.acha.2016.09.003.

[76] M. Viana, *(Dis)continuity of lyapunov exponents*, Ergodic Theory Dynam. Systems, 40 (2020), pp. 577–611, https://doi.org/10.1017/etds.2018.50.

[77] L. Y. W. Cowieson, *SRB measures as zero-noise limits*, Ergodic Theory Dynam. Systems, 25 (2005), pp. 1115–1138, https://doi.org/10.1017/S0143385704000604.

[78] Z. Wang and W. Sun, *Lyapunov exponents of hyperbolic measures and hyperbolic periodic orbits*, Trans. Amer. Math. Soc., 362 (2010), pp. 4267–4282, https://www.jstor.org/stable/25733367.

[79] L. Young, *Statistical properties of dynamical systems with some hyperbolicity*, Ann. of Math. (2), 147 (1998), pp. 585–650, https://doi.org/10.2307/120960.

[80] L. Young, *Recurrence times and rates of mixing*, Israel J. Math., 110 (1999), pp. 153–188, https://doi.org/10.1007/BF02808180.

[81] Z. Zhao and D. Giannakis, *Analog forecasting with dynamics-adapted kernels*, Nonlinearity, 29 (2016), pp. 2888–2939, https://doi.org/10.1088/0951-7715/29/9/2888.