Modern science is undergoing what might arguably be called a “data revolution,” manifested by a rapid growth of observed and simulated data from complex systems, as well as vigorous research on mathematical and computational frameworks for data analysis. In many scientific branches, these efforts have led to the creation of statistical models of complex systems that match or exceed the skill of first-principles models. Yet, despite these successes, statistical models are oftentimes treated as black boxes, providing limited guarantees about stability and convergence as the amount of training data increases. Black-box models also offer limited insights about the operating mechanisms (physics), the understanding of which is central to the advancement of science.

In this short review, we describe mathematical techniques for statistical analysis and prediction of time-evolving phenomena, ranging from simple examples such as an oscillator, to highly complex systems such as the turbulent motion of the Earth’s atmosphere, the folding of proteins, and the evolution of species populations in an ecosystem. Our main thesis is that combining ideas from the theory of dynamical systems with learning theory provides an effective route to data-driven models of complex systems, with refinable predictions as the amount of training data increases, and physical interpretability through discovery of coherent patterns around which the dynamics is organized. Our article thus serves as an invitation to explore ideas at the interface of the two fields.

This is a vast subject, and invariably a number of important developments in areas such as deep learning, reservoir computing, control, and nonautonomous/stochastic systems are not discussed here.¹ Our focus will be on topics drawn from the authors’ research and related work.

**Statistical Forecasting and Coherent Pattern Extraction**

Consider a dynamical system of the form \( \Phi^t : \Omega \to \Omega \), where \( \Omega \) is the state space and \( \Phi^t, t \in \mathbb{R} \), the flow map. For

¹See [https://arxiv.org/abs/2002.07928](https://arxiv.org/abs/2002.07928) for a version of this article with references to the literature on these topics.
example, $\Omega$ could be Euclidean space $\mathbb{R}^d$, or a more general manifold, and $\Phi^t$ the solution map for a system of ODEs defined on $\Omega$. Alternatively, in a PDE setting, $\Omega$ could be an infinite-dimensional function space and $\Phi^t$ an evolution group acting on it. We consider that $\Omega$ has the structure of a metric space equipped with its Borel $\sigma$-algebra, playing the role of an event space, with measurable functions on $\Omega$ acting as random variables, called observables.

In a statistical modeling scenario, we consider that available to us are time series of various such observables, sampled along a dynamical trajectory which we will treat as being unknown. Specifically, we assume that we have access to two observables, $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$, respectively referred to as covariate and response functions, together with corresponding time series $x_0, x_1, \ldots, x_{N-1}$ and $y_0, y_1, \ldots, y_{N-1}$, where $x_n = X(\omega_n)$, $y_n = Y(\omega_n)$, and $\omega_n = \Phi^t \Delta t(\omega_0)$. Here, $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces, $\Delta t$ is a positive sampling interval, and $\omega_0$ is an arbitrary point in $\Omega$ initializing the trajectory. We shall refer to the collection $\{(x_0, y_0), \ldots, (x_{N-1}, y_{N-1})\}$ as the training data. We require that $Y$ be a Banach space (so that one can talk about expectations and other functionals applied to $Y$), but allow the covariate space $\mathcal{X}$ to be nonlinear.

Many problems in statistical modeling of dynamical systems can be expressed in this framework. For instance, in a low-dimensional ODE setting, $X$ and $Y$ could both be the identity map on $\Omega = \mathbb{R}^d$, and the task could be to build a model for the evolution of the full system state. Weather forecasting is a classical high-dimensional application, where $\Omega$ is the abstract state space of the climate system, and $X$ a (highly noninvertible) map representing measurements from satellites, meteorological stations, and other sensors available to a forecaster. The response $Y$ could be temperature at a specific location, $Y = \mathbb{R}$, illustrating that the response space may be of considerably lower dimension than the covariate space. In other cases, e.g., forecasting the temperature field over a geographical region, $Y$ may be a function space. The two primary questions that will concern us here are:

**Problem 1** (Statistical forecasting). Given the training data, construct ("learn") a function $Z_t : \mathcal{X} \rightarrow \mathcal{Y}$ that predicts $Y$ at a lead time $t \geq 0$. That is, $Z_t$ should have the property that $Z_t \circ X$ is closest to $Y \circ \Phi^t$ among all functions in a suitable class.

**Problem 2** (Coherent pattern extraction). Given the training data, identify a collection of observables $z_j : \Omega \rightarrow \mathcal{Y}$ that have the property of evolving coherently under the dynamics. By that, we mean that $Z_t \circ \Phi^t$ should be relatable to $z_j$ in a natural way.

These problems have an extensive history of study from an interdisciplinary perspective spanning mathematics, statistics, physics, and many other fields. Here, our focus will be on nonparametric methods, which do not employ explicit parametric models for the dynamics. Instead, they use universal structural properties of dynamical systems to inform the design of data analysis techniques. From a learning standpoint, Problems 1 and 2 can be thought of as supervised and unsupervised learning, respectively. A mathematical requirement we will impose on methods addressing either problem is that they have a well-defined notion of convergence, i.e., they are refinable, as the number $N$ of training samples increases.

### Analog and POD Approaches

Among the earliest examples of nonparametric forecasting techniques is Lorenz’s analog method [Lor69]. This simple, elegant approach makes predictions by tracking the evolution of the response along a dynamical trajectory in the training data (the analogs). Good analogs are selected according to a measure of geometrical similarity between the covariate variable observed at forecast initialization and the covariate training data. This method posits that past behavior of the system is representative of its future behavior, so looking up states in a historical record that are closest to current observations is likely to yield a skillful forecast. Subsequent methodologies have also emphasized aspects of state space geometry, e.g., using the training data to approximate the evolution map through patched local linear models, often leveraging delay coordinates for state space reconstruction.

Early approaches to coherent pattern extraction include the proper orthogonal decomposition (POD), which is closely related to principal component analysis (PCA, introduced in the early twentieth century by Pearson), the Karhunen–Loève expansion, and empirical orthogonal function (EOF) analysis. Assuming that $Y$ is a Hilbert space, POD yields an expansion $Y \approx Y_L = \sum_{j=1}^L z_j \psi_j$. Arranging the data into a matrix $Y = (y_0, \ldots, y_{N-1})$, the $z_j$ are the singular values of $Y$ (in decreasing order), the $u_j$ are the corresponding left singular vectors, called EOFs, and the $\psi_j$ are given by projections of $Y$ onto the EOFs, $\psi_j(\omega) = \langle u_j, Y(\omega) \rangle_Y$. That is, the principal component $\psi_j : \Omega \rightarrow \mathbb{R}$ is a linear feature characterizing the unsupervised data $\{y_0, \ldots, y_{N-1}\}$. If the data is drawn from a probability measure $\mu$, as $N \to \infty$ the POD expansion is optimal in an $L^2(\mu)$ sense; that is, $Y_L$ has minimal $L^2(\mu)$ error $\|Y - Y_L\|_{L^2(\mu)}$ among all rank-$L$ approximations of $Y$. Effectively, from the perspective of POD, the important components of $Y$ are those capturing maximal variance.

Despite many successes in challenging applications (e.g., turbulence), it has been recognized that POD may not reveal dynamically significant observables, offering limited predictability and physical insight. In recent years, there has been significant interest in techniques that address this shortcoming by modifying the linear map $Y$ to...
have an explicit dependence on the dynamics [BK86], or replacing it by an evolution operator [DJ99, Mez05]. Either directly or indirectly, these methods make use of operator-theoretic ergodic theory, which we now discuss.

Operator-Theoretic Formulation

The operator-theoretic formulation of dynamical systems theory shifts attention from the state-space perspective, and instead characterizes the dynamics through its action on linear spaces of observables. Denoting the vector space of \( y \)-valued functions on \( \Omega \) by \( \mathcal{F} \), for every time \( t \) the dynamics has a natural induced action \( U^t : \mathcal{F} \to \mathcal{F} \) given by composition with the flow map, \( U^t f = f \circ \Phi^t \). It then follows by definition that \( U^t \) is a linear operator; i.e., \( U^t(\alpha f + g) = \alpha U^t f + U^t g \) for all observables \( f, g \in \mathcal{F} \) and every scalar \( \alpha \in \mathbb{C} \). The operator \( U^t \) is known as a composition operator, or Koopman operator after classical work of Bernard Koopman in the 1930s [Koo31], which established that a general (potentially nonlinear) dynamical system can be characterized through intrinsically linear operators acting on spaces of observables. A related notion is that of the transfer operator, \( P^t : \mathcal{M} \to \mathcal{M} \), which describes the action of the dynamics on a space of measures \( \mathcal{M} \) via the pushforward map, \( P^t m = \Phi^t_* m = m \circ \Phi^{-t} \). In a number of cases, \( \mathcal{F} \) and \( \mathcal{M} \) are dual spaces to one another (e.g., continuous functions and Radon measures), in which case \( U^t \) and \( P^t \) are dual operators.

If the space of observables under consideration is equipped with a Banach or Hilbert space structure, and the dynamics preserves that structure, the operator-theoretic formulation allows a broad range of tools from spectral theory and approximation theory for linear operators to be employed in the study of dynamical systems. For our purposes, a particularly advantageous aspect of this approach is that it is amenable to rigorous statistical approximation, which is one of our principal objectives. It should be kept in mind that the spaces of observables encountered in applications are generally infinite-dimensional, leading to behaviors with no counterparts in finite-dimensional linear algebra, such as unbounded operators and continuous spectrum. In fact, as we will see below, the presence of continuous spectrum is a hallmark of mixing (chaotic) dynamics.

In this review, we restrict attention to the operator-theoretic description of measure-preserving, ergodic dynamics. By that, we mean that there is a probability measure \( \mu \) on \( \Omega \) such that (i) \( \mu \) is invariant under the flow, i.e., \( \Phi^t_* \mu = \mu \); and (ii) every measurable, \( \Phi^t \)-invariant set has either zero or full \( \mu \)-measure. We also assume that \( \mu \) is a Borel measure with compact support \( A \subseteq \Omega \); this set is necessarily \( \Phi^t \)-invariant. An example known to rigorously satisfy these properties is the Lorenz 63 (L63) system on \( \Omega = \mathbb{R}^3 \), which has a compactly supported, ergodic invariant measure supported on the famous “butterfly” fractal attractor; see Figure 1. L63 exemplifies the fact that a smooth dynamical system may exhibit invariant measures with nonsmooth supports. This behavior is ubiquitous in models of physical phenomena, which are formulated in terms of smooth differential equations, but whose long-term dynamics concentrate on lower-dimensional subsets of state space due to the presence of dissipation. Our methods should therefore not rely on the existence of a smooth structure for \( A \).

In the setting of ergodic, measure-preserving dynamics on a metric space, two relevant structures that the dynamics may be required to preserve are continuity and \( \mu \)-measurability of observables. If the flow \( \Phi^t \) is continuous, then the Koopman operators act on the Banach space \( \mathcal{F} = C(A, \mathcal{Y}) \) of continuous, \( y \)-valued functions on \( A \), equipped with the uniform norm, by isometries, i.e., \( ||U^t f||_\mathcal{F} = ||f||_\mathcal{F} \). If \( \Phi^t \) is \( \mu \)-measurable, then \( U^t \) lifts to an operator on equivalence classes of \( y \)-valued functions in \( L^p(\mu, \mathcal{Y}) \). \( 1 \leq p \leq \infty \), and acts again by isometries. If \( \mathcal{Y} \) is a Hilbert space (with inner product \( \langle \cdot, \cdot \rangle_y \)), the case \( p = 2 \) is special, since \( L^2(\mu, \mathcal{Y}) \) is a Hilbert space with inner product \( \langle f, g \rangle_{L^2(\mu, \mathcal{Y})} = \int_{\Omega} \langle f(\omega), g(\omega) \rangle_y \, d\mu(\omega) \), on which \( U^t \) acts as a unitary map, \( U^{it} = U^{-t} \).

Clearly, the properties of approximation techniques for observables and evolution operators depend on the underlying space. For instance, \( C(A, \mathcal{Y}) \) has a well-defined notion of pointwise evaluation at every \( \omega \in \Omega \) by a continuous linear map \( \delta_\omega : C(A, \mathcal{Y}) \to \mathcal{Y} \), \( \delta_\omega f = f(\omega) \), which is useful for interpolation and forecasting, but lacks an inner-product structure and associated orthogonal projections. On the other hand, \( L^2(\mu, \mathcal{Y}) \) has inner-product structure, which is very useful theoretically as well as for numerical algorithms, but lacks the notion of pointwise evaluation.

Letting \( \mathcal{F} \) stand for any of the \( C(A, \mathcal{Y}) \) or \( L^p(\mu, \mathcal{Y}) \) spaces, the set \( U = \{U^t : \mathcal{F} \to \mathcal{F}\}_{t \in \mathbb{R}} \) forms a strongly continuous group under composition of operators. That is, \( U^t \circ U^s = U^{t+s} \), \( U^{-1} = U^{-t} \), and \( U^0 = \text{Id} \), so that \( U \) is a group, and for every \( f \in \mathcal{F} \), \( U^t f \) converges to \( f \) in the norm of \( \mathcal{F} \) as \( t \to 0 \). A central notion in such evolution groups is that of the generator, defined by the \( \mathcal{F} \)-norm limit \( V f = \lim_{t \to 0} (U^t f - f)/t \) for all \( f \in \mathcal{F} \) for which the limit exists. It can be shown that the domain \( D(V) \) of all such \( f \) is a dense subspace of \( \mathcal{F} \), and \( V : D(V) \to \mathcal{F} \) is a closed, unbounded operator. Intuitively, \( V \) can be thought of as a directional derivative of observables along the dynamics. For example, if \( \mathcal{Y} = \mathcal{C} \), \( A \) is a \( \mathcal{C}^1 \) manifold, and the flow \( \Phi^t : A \to A \) is generated by a continuous vector field \( \vec{V} : A \to TA \), then the generator of the Koopman group on \( C(A) \) has as its domain the space \( C^1(A) \subseteq C(A) \) of continuously differentiable, complex-valued functions, and \( V f = \vec{V} \cdot \nabla f \) for \( f \in C^1(A) \). A strongly continuous
evolution group is completely characterized by its generator, as any two such groups with the same generator are identical.

The generator acquires additional properties in the setting of unitary evolution groups on $H = L^2(\mu, y)$, where $y$ is skew-adjoint, $V^* = -V$. Note that the skew-adjointness of $V$ holds for more general measure-preserving dynamics than Hamiltonian systems, whose generator is skew-adjoint with respect to Lebesgue measure. By the spectral theorem for skew-adjoint operators, there exists a unique projection-valued measure $E : \mathcal{B}(\mathbb{R}) \rightarrow B(H)$, giving the generator and Koopman operator as the spectral integrals

$$V = \int_{\mathbb{R}} i\alpha dE(\alpha), \quad U^t = e^{itV} = \int_{\mathbb{R}} e^{it\alpha} dE(\alpha).$$

Here, $\mathcal{B}(\mathbb{R})$ is the Borel $\sigma$-algebra on the real line, and $B(H)$ the space of bounded operators on $H$. Intuitively, $E$ can be thought of as an operator analog of a complex-valued representation $H$.

Here, $\mathcal{B}(\mathbb{R})$ is the Borel $\sigma$-algebra on the real line, and $B(H)$ the space of bounded operators on $H$. Intuitively, $E$ can be thought of as an operator analog of a complex-valued representation $H$.

**Data-Driven Forecasting**

Based on the concepts introduced above, one can formulate statistical forecasting in Problem 1 as the task of constructing a function $Z_t : X \rightarrow Y$ on covariate space $X$, such that $Z_t \circ X$ optimally approximates $U^t Y$ among all functions in a suitable class. We set $Y = C$, so the response variable is scalar-valued, and consider the Koopman operator on $L^2(\mu)$, so we have access to orthogonal projections. We also assume for now that the covariate function $X$ is injective, so $\hat{Y}_t := Z_t \circ X$ should be able to approximate $U^t Y_0$ to arbitrarily high precision in $L^2(\mu)$ norm. Indeed, let $\{u_0, u_1, \ldots\}$ be an orthonormal basis of $L^2(\gamma)$, where $\gamma = X_0\mu$ is the pushforward of the invariant measure onto $X$. Then, $\{\phi_0, \phi_1, \ldots\}$ with $\phi_j = u_j \circ X$ is an orthonormal basis of $L^2(\mu)$.

**Sampling measures and ergodicity.** The dynamical trajectory $\{\omega_0, \omega_1, \ldots, \omega_{N-1}\}$ in state space underlying the training data is the support of a discrete sampling measure $\mu_N := \sum_{n=0}^{N-1} \delta_{\omega_n}/N$. A key consequence of ergodicity is that for Lebesgue-a.e. sampling interval $\Delta t$ and $\mu$-a.e. starting point $\omega_0 \in \Omega$, as $N \rightarrow \infty$, the sampling measures $\mu_N$ weak-converge to the invariant measure $\mu$; that is,

$$\lim_{N \rightarrow \infty} \int_{\Omega} f d\mu_N = \int_{\Omega} f d\mu \quad \forall f \in C(\Omega).$$ (1)
Since integrals against $\mu_N$ are time averages on dynamical trajectories, i.e., $\int_\Omega f \, d\mu_N = \sum_{n=0}^{N-1} f(\omega_n)/N$, ergodicity provides an empirical means of accessing the statistics of the invariant measure. In fact, many systems encountered in applications possess so-called physical measures, where (1) holds for $\omega_0$ in a “larger” set of positive measure with respect to an ambient measure (e.g., Lebesgue measure) from which experimental initial conditions are drawn. Hereafter, we will let $M$ be a compact subset of $\Omega$, which is forward-invariant under the dynamics (i.e., $\Phi^t(M) \subseteq M$ for all $t \geq 0$), and contains $A$. For example, in dissipative dynamical systems such as L63, $M$ can be chosen as a compact absorbing ball.

Shift operators. Ergodicity suggests that appropriate data-driven analogs are the $L^2(\mu_N)$ spaces induced by the sampling measures $\mu_N$. For a given $N$, $L^2(\mu_N)$ consists of equivalence classes of measurable functions $f : \Omega \to \mathbb{C}$ having common values at the sampled states $\omega_n$, and the inner product of two elements $f, g \in L^2(\mu_N)$ is given by an empirical time-correlation, $(f, g)_{\mu_N} = \int_\Omega f(\omega) \, g(\omega) \, d\mu_N = \sum_{n=0}^{N-1} f(\omega_n)g(\omega_n)/N$. Moreover, if the $\omega_n$ are distinct (as we will assume for simplicity of exposition), $L^2(\mu_N)$ has dimension $N$, and is isomorphic as a Hilbert space to $\mathbb{C}^N$ equipped with a normalized dot product. Given that, we can represent every function $f \in L^2(\mu_N)$ by a column vector $\mathbf{f} = (f(\omega_0), \ldots, f(\omega_{N-1}))^T \in \mathbb{C}^N$, and every linear map $A : L^2(\mu_N) \to L^2(\mu_N)$ by an $N \times N$ matrix $A$, so that $A(\mathbf{f}) = \mathbf{g} = A\mathbf{f}$ is the column vector representing $g = Af$. The elements of $\mathbf{f}$ can also be understood as expansion coefficients in the standard basis $\{e_0, e_1, \ldots, e_{N-1}\}$ of $L^2(\mu_N)$, where $e_{i,N}(\omega_n) = N^{1/2}\delta_{nj}$; that is, $f(\omega_n) = (e_n^T, f)_{L^2(\mu_N)}$. Similarly, the elements of $A$ correspond to the operator matrix elements $A_{ij} = (e_i^T, A e_j)_{L^2(\mu_N)}$.

Next, we would like to define a Koopman operator on $L^2(\mu_N)$, but this space does not admit such an operator as a composition map induced by the dynamical flow $\Phi^t$ on $\Omega$. This is because $\Phi^t$ does not preserve null sets with respect to $\mu_N$, and thus does not lead to a well-defined composition map on equivalence classes of functions in $L^2(\mu_N)$. Nevertheless, on $L^2(\mu_N)$ there is an analogous construct to the Koopman operator on $L(\mu)$, namely, the shift operator $U^q_N : L^2(\mu_N) \to L^2(\mu_N)$, $q \in \mathbb{Z}$, defined as

$$U^q_N f(\omega_n) = \begin{cases} f(\omega_{n+q}), & 0 \leq n + q \leq N - 1, \\ 0, & \text{otherwise}. \end{cases}$$

Even though $U^q_N$ is not a composition map, intuitively it should have a connection with the Koopman operator $U^q \Delta t$. One could consider, for instance, the matrix representation $U^q_N(q) = [e_{i,N} U^q e_{j,N}]_{L^2(\mu_N)}$ in the standard basis, and attempt to connect it with a matrix representation of $U^q \Delta t$ in an orthonormal basis of $L^2(\mu)$. However, the issue with this approach is that the $e_{j,N}$ do not have $N \to \infty$ limits in $L^2(\mu)$, meaning that there is no suitable notion of $N \to \infty$ convergence of the matrix elements of $U^q_N$ in the standard basis. In response, we will construct a representation of the shift operator in a different orthonormal basis with a well-defined $N \to \infty$ limit. The main tools that we will use are kernel integral operators, which we now describe.

Kernel integral operators. In the present context, a kernel function will be a real-valued, continuous function $k : \Omega \times \Omega \to \mathbb{R}$ with the property that there exists a strictly positive, continuous function $d : \Omega \to \mathbb{R}$ such that

$$d(\omega)k(\omega, \omega') = d(\omega')k(\omega', \omega) \quad \forall \omega, \omega' \in \Omega. \quad (2)$$

Notice the similarity between (2) and the detailed balance relation in reversible Markov chains. Now let $\rho$ be any Borel probability measure with compact support $S \subseteq M$ included in the forward-invariant set $M$. It follows by continuity of $k$ and compactness of $S$ that the integral operator $K_\rho : L^2(\rho) \to C(M)$,

$$K_\rho f = \int_\Omega k(\cdot, \omega) f(\omega) \, d\rho(\omega), \quad (3)$$

is well-defined as a bounded operator mapping elements of $L^2(\rho)$ into continuous functions on $M$. Using $\rho_\| : C(M) \to L^2(\rho)$ to denote the canonical inclusion map, we consider two additional integral operators, $G_\rho : L^2(\rho) \to L^2(\rho)$ and $\tilde{G}_\rho : C(M) \to C(M)$, with $G_\rho = \iota_\rho K_\rho$ and $G_\rho = K_\rho \iota_\rho$, respectively.

The operators $G_\rho$ and $G_\rho$ are compact operators acting with the same integral formula as $K_\rho$ in (3), but their codomains and domains, respectively, are different. Nevertheless, their nonzero eigenvalues coincide, and $\phi \in L^2(\rho)$ is an eigenfunction of $G_\rho$ corresponding to a nonzero eigenvalue $\lambda$ if and only if $\phi \in C(M)$ with $\phi = K_\rho \phi/\lambda$ is an eigenfunction of $G_\rho$ at the same eigenvalue. In effect, $\phi \mapsto \phi$ “interpolates” the $L^2(\rho)$ element $\phi$ (defined only up to $\rho$-null sets) to the continuous, everywhere-defined function $\phi$. It can be verified that if (2) holds, $G_\rho$ is a trace-class operator with real eigenvalues, $|\lambda_0| \geq |\lambda_1| \geq \cdots \to 0^+$. Moreover, there exists a Riesz basis $\{\phi_0, \phi_1, \ldots\}$ of $L^2(\rho)$ and a corresponding dual basis $\{\phi_0^*, \phi_1^*, \ldots\}$ with $\langle \phi_i, \phi_j \rangle_{L^2(\rho)} = \delta_{ij}$, such that $G_\rho \phi_j = \lambda_j \phi_j$ and $G_\rho^* \phi_j = \lambda_j^* \phi_j$. We say that the kernel $k$ is $L^2(\rho)$-universal if $G_\rho$ has no zero eigenvalues; this is equivalent to $\rho_\|$ being dense in $L^2(\rho)$. Moreover, $k$ is said to be $L^2(\rho)$-Markov if $G_\rho$ is a Markov operator, i.e., $G_\rho \geq 0, G_\rho f \geq 0$ if $f \geq 0$, and $G_1 = 1$.

Observe now that the operators $G_\mu_N$ associated with the sampling measures $\mu_N$, henceforth abbreviated by $G_N$, are represented by $N \times N$ kernel matrices $G_N = [e_{i,N} \langle e_{j,N}, \mu_N \rangle]_{L^2(\mu_N)}$ in the standard basis of $L^2(\mu_N)$. Further, if $k$ is a pullback kernel from covariate space, i.e., $k(\omega, \omega') = k_0(\omega, \omega')\Phi(\omega, \omega')$ for $\Phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then $G_N = [k_0(\omega, \omega')]$ is empirically accessible from the
training data. Popular kernels in applications include the covariance kernel \(k(x, x') = \langle x, x' \rangle_X\) on an inner-product space and the radial Gaussian kernel \(k(x, x') = e^{-\|x-x'\|^2/\varepsilon}\). It is also common to employ Markov kernels constructed by normalization of symmetric kernels [CL06, BH16]. We will use \(k_N\) to denote kernels with data-dependent normalizations.

A widely used strategy for learning with integral operators [vLB08] is to construct families of kernels \(k_N\) converging in \((CM \times M)\) norm to \(k\). This implies that for every nonzero eigenvalue \(\lambda_j \in G \equiv G_\mu\) the sequence of eigenvalues \(\lambda_{j,N}\) of \(G_N\) satisfies \(\lim_{N \to \infty} \lambda_{j,N} = \lambda_j\). Moreover, there exists a sequence of eigenfunctions \(\phi_{j,N} \in L^2(\mu_N)\) corresponding to \(\lambda_{j,N}\), whose continuous representatives, \(\varphi_{j,N} = K_N \phi_{j,N}/\lambda_{j,N}\), converge in \(C(M)\) to \(\varphi_j = K \phi_j/\lambda_j\), where \(\phi_j \in L^2(\mu)\) is any eigenfunction of \(G\) at eigenvalue \(\lambda_j\). In effect, we use \(C(M)\) as a “bridge” to establish spectral convergence of the operators \(G_N\), which act on different spaces. Note that \((\lambda_{j,N}, \varphi_{j,N})\) does not converge uniformly with respect to \(j\), and for a fixed \(N\), eigenvalues/eigenfunctions at larger \(j\) exhibit larger deviations from their \(N \to \infty\) limits. Under measure-preserving, ergodic dynamics, convergence occurs for \(\mu\)-a.e. starting state \(\omega_n \in M\), in a set of positive ambient measure if \(\mu\) is physical. In particular, the training states \(\omega_n\) need not lie on \(A\). See Figure 1 for eigenfunctions of \(G_N\) computed from data sampled near the L63 attractor.

Diffusion forecasting. We now have the ingredients to build a constant statistical forecasting scheme based on data-driven approximations of the Koopman operator. In particular, note that if \(\phi_{i,N}, \phi_{j,N}\) are biothogonal eigenfunctions of \(G_N\) and \(G_N\), respectively, at nonzero eigenvalues, we can evaluate the matrix element \(U_{i,j,N}(q)\) of the shift operator using the continuous representatives \(\varphi_{i,N}, \varphi_{j,N}\),

\[
U_{i,j,N}(q) = \frac{1}{N} \sum_{n=0}^{N-1-q} \varphi_{i,N}(\omega_n) \varphi_{j,N}(\omega_{n+q})
\]

\[
= \frac{N-q}{N} \int_{\Omega} \varphi_{i,N}(q U^{q \Delta t} q_{j,N}) d\mu_{N-q},
\]

where \(U^{q \Delta t}\) is the Koopman operator on \(C(M)\). Therefore, if the corresponding eigenvalues \(\lambda_i, \lambda_j\) of \(G\) are nonzero, by the weak convergence of the sampling measures in (1) and \(C(M)\) convergence of the eigenfunctions, as \(N \to \infty\), \(U_{i,j,N}(q)\) converges to the matrix element \(U_{i,j}(q \Delta t) = \langle \phi_i, U^{q \Delta t} \phi_j \rangle_{L^2(\mu)}\) of the Koopman operator on \(L^2(\mu)\). This convergence is not uniform with respect to \(i, j\), but if we fix a parameter \(L \in \mathbb{N}\) (which can be thought of as spectral resolution) such that \(\lambda_{L-1} \neq 0\), we can obtain a statistically consistent approximation of \(L^2(\mu)\) Koopman operator matrices, \(U(q \Delta t) = [U_{ij}(q \Delta t)]\), by shift operator matrices, \(U_N(q) = [U_{i,j,N}(q)]\), with \(i, j \in \{0, \ldots, L-1\}\). Checkerboard plots of \(U_N(q)\) for the L63 system are displayed in Figure 1.

This method for approximating matrix elements of Koopman operators was proposed in a technique called diffusion forecasting (named after the diffusion kernels employed) [BGH15]. Assuming that the response \(Y\) is continuous and by spectral convergence of \(G_N\), for every \(j \in \mathbb{N}\) such that \(\lambda_j > 0\), the inner products \(\tilde{Y}_{j,N} = \langle \phi_j, \Psi \rangle_{L^2(\mu)}\) converge, as \(N \to \infty\), to \(\tilde{Y}_j = \langle \phi_j, \Psi \rangle_{L^2(\mu)}\). This implies that for any \(L \in \mathbb{N}\) such that \(\lambda_{L-1} > 0, \sum_{j=0}^{L-1} \tilde{Y}_{j,N} \varphi_{j,N}\) converges in \(C(M)\) to a continuous representative of \(\Pi_L Y\), where \(\Pi_L\) is the orthogonal projection on \(L^2(\mu)\) mapping into span\{\(\varphi_0, \ldots, \varphi_{L-1}\)\}. Suppose now that \(\varphi_N\) is a sequence of continuous functions converging uniformly to \(\varphi \in C(M)\), such that \(\varphi_N\) are probability densities with respect to \(\mu\) (i.e., \(\varphi_N \geq 0\) and \(\int \varphi_N d\mu_N = 1\)). By similar arguments as for \(Y\), as \(N \to \infty\), the continuous function \(\sum_{j=0}^{L-1} \tilde{Y}_{j,N} \varphi_{j,N}\) with \(\tilde{Y}_{j,N} = \langle \varphi_j, \varphi_N \rangle_{L^2(\mu)}\) converges to \(\Pi_L \varphi\) in \(L^2(\mu)\). Putting these facts together, and setting \(\hat{\varphi}_N = (\hat{\varphi}_0, \ldots, \hat{\varphi}_{L-1})^T\) and \(\hat{Y}_N = (\hat{Y}_0, \ldots, \hat{Y}_{L-1})^T\), we conclude that

\[
\hat{\varphi}_N^T U_N(q) \hat{Y}_N \xrightarrow{N \to \infty} \langle \Pi_L \hat{\varphi}, \Pi_L U^{q \Delta t} \Psi \rangle_{L^2(\mu)},
\]

Here, the left-hand side is given by matrix–vector products obtained from the data, and the right-hand side is equal to the expectation of \(\Pi_L U^{q \Delta t} \varphi\) with respect to the probability measure \(\rho\) with density \(d\rho/d\mu = \varphi\); i.e.,

\[
\langle \Pi_L \hat{\varphi}, \Pi_L U^{q \Delta t} \Psi \rangle_{L^2(\mu)} = \mathbb{E}_\rho(\Pi_L U^{q \Delta t} \Psi),
\]

where \(\mathbb{E}_\rho(\cdot) := \int_\Omega (\cdot)\, d\rho\).

What about the dependence of the forecast on \(L\)? As \(L\) increases, \(\Pi_L\) converges strongly to the orthogonal projection \(\Pi_G : L^2(\mu) \to L^2(\mu)\) onto the closure of the range of \(G\). Thus, if the kernel \(k\) is \(L^2(\mu)\)-universal (i.e., \(\text{ran} G = L^2(\mu)\)), \(\Pi_G = \text{Id}\), and under the iterated limit of \(L \to \infty\) after \(N \to \infty\) the left-hand side of (4) converges to \(\mathbb{E}_\rho U^{q \Delta t} \Psi\).

In summary, implemented with an \(L^2(\mu)\)-universal kernel, diffusion forecasting consistently approximates the expected value of the time-evolution of any continuous observable with respect to any probability measure with continuous density relative to \(\mu\). An example of an \(L^2(\mu)\)-universal kernel is the pullback of a radial Gaussian kernel on \(X = \mathbb{R}^m\). In contrast, the covariance kernel is not \(L^2(\mu)\)-universal, as in this case the rank of \(G\) is bounded by \(m\). This illustrates that forecasting in the POD basis may be subject to intrinsic limitations, even with full observations.

Kernel analog forecasting. While providing a flexible framework for approximating expectation values of observables under measure-preserving, ergodic dynamics, diffusion forecasting does not directly address the problem of constructing a concrete forecast function, i.e., a function \(Z_t : X \to \mathbb{C}\) approximating \(U^q Y\) as stated in
Figure 1. Panel (a) shows eigenfunctions $\phi_{1,N}$ of $G_N$ for a dataset sampled near the L63 attractor. Panel (b) shows the action of the shift operator $U^q_N$ on the $\phi_{1,N}$ from (a) for $q = 50$ steps, approximating the Koopman operator $U^q \Delta t$. Panels (c, d) show the matrix elements $\langle \phi_{i,N}, U^q_N \phi_{j,N} \rangle$ of the shift operator for $q = 5$ and 50. The mixing dynamics is evident in the larger far-from-diagonal components in $q = 50$ vs. $q = 5$. Panel (e) shows the matrix representation of a finite-difference approximation of the generator $V$, which is skew-symmetric. Panels (f, g) summarize the diffusion forecast (DF) and kernel analog forecast (KAF) for lead time $t = q \Delta t$. In each diagram, the data-driven finite-dimensional approximation (top row) converges to the true forecast (middle row). DF maps an initial state $\omega \in M \subseteq \Omega$ to the future expectation of an observable $E_{\Psi(\omega)} U^t Y = E_{U^t \Psi(\omega)} Y$, and KAF maps a response function $Y \in C(M)$ to the conditional expectation $E(U^t Y | X)$. 
Problem 1. One way of defining such a function is to let \( \kappa_N \) be an \( L^2(X_N) \)-Markov kernel on \( X \) for \( X_N = X \mu_N \), and to consider the “feature map” \( \Psi_N : X \to C(M) \) mapping each point \( x \in X \) in covariate space to the kernel section \( \Psi_N(x) = \kappa_N(x, X(\cdot)) \). Then, \( \Psi_N(x) \) is a continuous probability density with respect to \( \mu_N \), and we can use diffusion forecasting to define \( Z_q(Y) = \Psi_N(T_x U_N q) \) for covariate \( x \in X \) with notation as in (4).

While this approach has a well-defined \( N \to \infty \) limit, it does not provide optimality guarantees, particularly in situations where \( X \) is noninjective. Indeed, the \( L^2(\mu) \)-optimal approximation to \( U^T Y \) of the form \( Z_t \) is given by the conditional expectation \( E(U^T Y \mid X) \). In the present \( L^2 \) setting we have \( E(U^T Y \mid X = x) = \Pi_X U^T Y \), where \( \Pi_X \) is the orthogonal projection into \( L^2(X) \) along all pullbacks \( f \in \tilde{L}^2(\mu) \) from covariate space. Even though \( E(U^T Y \mid X = x) \) can be expressed as an expectation with respect to a conditional probability measure \( \mu(\cdot \mid x) \) on \( \Omega \), that measure will generally not have an \( L^2(\mu) \) density, and there is no map \( \Psi : X \to C(M) \) such that \( \langle \Psi(x), U^T Y \rangle_{L^2(\mu)} \) equals \( E(U^T Y \mid X = x) \).

To construct a consistent estimator of the conditional expectation, we require that \( k \) be a pullback of a kernel \( k : X \times X \to \mathbb{R} \) on covariate space which is (i) symmetric, \( \kappa(x, x') = \kappa(x', x) \) for all \( x, x' \in X \) (so (2) holds); (ii) strictly positive; and (iii) strictly positive-definite. The latter means that for any sequence \( x_0, \ldots, x_{n-1} \) of distinct points in \( X \) the matrix \( [\kappa(x_i, x_j)] \) is strictly positive. These properties imply that there exists a Hilbert space \( \mathcal{H} \) of complex-valued functions on \( \Omega \), such that (i) for every \( \omega \in \Omega \), the kernel sections \( k_\omega = k(\omega, \cdot) \) lie in \( \mathcal{H} \); (ii) the evaluation functional \( \delta_\omega : \mathcal{H} \to \mathbb{C} \) is bounded and satisfies \( \delta_\omega(f) = \langle k_\omega, f \rangle_{\mathcal{L}^2(\mu)} \); (iii) every \( f \in \mathcal{H} \) has the form \( f = g \circ X \) for a continuous function \( g : X \to \mathbb{C} \); and (iv) \( \mu_{\mathcal{L}^2(\mu)} \) lies dense in \( \mathcal{H}_X(\mu) \).

A Hilbert space of functions satisfying (i) and (ii) above is known as a reproducing kernel Hilbert space (RKHS). RKHSs have many useful properties for statistical learning [CS02], not least because they combine the Hilbert space structure of \( L^2 \) spaces with pointwise evaluation in spaces of continuous functions. The density of \( \mathcal{H} \) in \( \tilde{L}^2(\mu) \) is a consequence of the strict positive-definiteness of \( \kappa \). In particular, because the conditional expectation \( E(U^T Y \mid X) \) lies in \( \tilde{L}^2(\mu) \), it can be approximated by elements of \( \mathcal{H} \) to arbitrarily high precision in \( \tilde{L}^2(\mu) \) norm, and every such approximation will be a pullback \( \tilde{Y}_t = Z_t \circ X \) of a continuous function \( Z_t \) that can be evaluated at arbitrary covariate values.

We now describe a data-driven technique for constructing such a prediction function, which we refer to as kernel analog forecasting (KAF) [AG20]. Mathematically, KAF is closely related to kernel principal component regression. To build the KAF estimator, we work again with integral operators as in (3), with the difference that now \( K_\rho : L^2(\rho) \to \mathcal{H}(M) \) takes values in the restriction of \( \mathcal{H} \) to the forward-invariant set \( M \), denoted \( \mathcal{H}(M) \). One can show that the adjoint \( K_\rho^* : \mathcal{H}(M) \to L^2(\rho) \) coincides with the inclusion map \( i_\rho \) on continuous function, so that \( K_\rho^* \) maps \( f \in \mathcal{H}(M) \subset C(M) \) to its corresponding \( L^2(\rho) \) equivalence class. As a result, the integral operator \( G_\rho : L^2(\rho) \to L^2(\rho) \) takes the form \( G_\rho = K_\rho^* K_\rho \), becoming a self-adjoint, positive-definite, compact operator with eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0^+ \), and a corresponding orthonormal eigenbasis \( \{ \phi_0, \phi_1, \ldots \} \) of \( L^2(\rho) \). Moreover, \( \{ \psi_0, \psi_1, \ldots \} \) with \( \phi_j = K_\rho \phi_j / \sqrt{\lambda_j} \) is an orthonormal set in \( \mathcal{H}(M) \). In fact, Mercer’s theorem provides an explicit representation \( k(\omega, \omega') = \sum_{j=0} \psi_j(\omega) \overline{\psi_j(\omega')} \), where direct evaluation of the kernel in the left-hand side (known as “kernel trick”) avoids the complexity of inner-product computations between feature vectors \( \psi_j \). Here, our perspective is to rely on the orthogonality of the eigenbasis to approximate observables of interest at fixed \( L \), and establish convergence of the estimator as \( L \to \infty \). A similar approach was adopted for density estimation on noncompact domains, with Mercer-type kernels based on orthogonal polynomials [ZHL19].

Now a key operation that the RKHS enables is the Nyström extension, which interpolates \( L^2(\rho) \) elements of appropriate regularity to RKHS functions. The Nyström operator \( N_\rho : D(N_\rho) \to \mathcal{H}(M) \) is defined on the domain \( D(N_\rho) = \{ \sum_j c_j \phi_j : \sum_j |c_j|^2 / \lambda_j < \infty \} \) by linear extension of \( N_\rho \phi_j = \psi_j / \sqrt{\lambda_j} \). Note that \( N_\rho \phi_j = K_\rho \phi_j / \sqrt{\lambda_j} \), so \( N_\rho \) maps \( \phi_j \) to its continuous representative, and \( K_\rho^* N_\rho \phi_j = f \), meaning that \( N_\rho f = f, \rho \text{-a.e.} \) While \( D(N_\rho) \) may be a strict \( L^2(\rho) \) subspace, for any \( L \) with \( \lambda_{L+1} > 0 \) we define a spectrally truncated operator \( N_{L\rho} : L^2(\rho) \to \mathcal{H}(M), N_{L\rho} \sum_j \psi_j = \sum_{j=0} L \psi_j / \sqrt{\lambda_j} \). Then, as \( L \) increases, \( K_\rho^* N_{L\rho} f \) converges to \( \Pi_{G_\rho} f \) in \( L^2(\rho) \).

To make empirical forecasts, we set \( \rho = \mu_N \), compute the expansion coefficients \( c_j(\rho)(t) \) of \( U^T Y \) in the \( \{ \phi_j \} \) basis of \( L^2(\mu_N) \), and construct \( Y_{L,\rho} = N_{L\rho} U^T Y \). Because \( \phi_j \) are pullbacks of known functions \( u_j \in C(X) \), we have \( Y_{L,\rho} = Z_{L,\rho} \circ X \), where \( Z_{L,\rho} = \sum_{j=0}^{L-1} c_j(\rho)(t) u_j \), can be evaluated at any \( x \in X \).

The function \( Y_{L,\rho} \) is our estimator of the conditional expectation \( E(U^T Y \mid X) \). By spectral convergence of kernel integral operators, as \( N \to \infty \), \( Y_{L,\rho} \) converges to \( Y_{L} := N_{L} U^T Y \) in \( C(M) \) norm, where \( N_{L} \equiv N_{L\rho} \). Then, as \( L \to \infty \), \( K^* Y_{L,\rho} \) converges in \( L^2(\mu) \) norm to \( \Pi_{G} U^T Y \). Because \( \kappa \) is strictly positive-definite, \( G \) has dense range in \( L^2(\mu) \), and thus \( \Pi_{G} U^T Y \equiv \Pi_{X} U^T X = E(U^T Y \mid X) \). We therefore conclude that \( Y_{L,\rho} \) converges to the conditional
Coherent Pattern Extraction

We now turn to the task of coherent pattern extraction in Problem 2. This is a fundamentally unsupervised learning problem, as we seek to discover observables of a dynamical system that exhibit a natural time evolution (by some suitable criterion), rather than approximate a given observable as in the context of forecasting. We have mentioned POD as a technique for identifying coherent observables through eigenfunctions of covariance operators. Kernel PCA [SSM98] is a generalization of this approach utilizing integral operators with potentially nonlinear kernels. For data lying on Riemannian manifolds, it is popular to employ kernels approximating geometrical operators, such as heat operators and their associated Laplacians. Examples include Laplacian eigenmaps [BN03], diffusion maps [CL06], and variable-bandwidth kernels [BH16]. Meanwhile, coherent pattern extraction techniques based on evolution operators have also gained popularity in recent years. These methods include spectral analysis of transfer operators for detection of invariant sets [DJ99, DFS00], harmonic averaging [Mez05], and dynamic mode decomposition (DMD) [RMB+09, Sch10, WKR15, KNK+18] techniques for approximating Koopman eigenfunctions, and Darboux kernels for approximating spectral projectors [KPM20]. While natural from a theoretical standpoint, evolution operators tend to have more complicated spectral properties than kernel integral operators, including nonisolated eigenvalues and continuous spectrum. The following examples illustrate distinct behaviors associated with the point \((H_\sigma)\) and continuous \((H_c)\) spectrum subspaces of \(L^2(\mu)\).

Example 1 (Torus rotation). A quasiperiodic rotation on the 2-torus, \(\Omega = T^2\), is governed by the system of ODEs

\[
\dot{\omega} = \vec{V}(\omega),
\]

where \(\omega = (\omega_1, \omega_2) \in [0, 2\pi)^2\), \(\vec{V} = (\nu_1, \nu_2)\), and \(\nu_1, \nu_2 \in \mathbb{R}\) are rationally independent frequency parameters. The resulting flow, \(\Phi(t)(\omega) = (\omega_1 + t\nu_1, \omega_2 + t\nu_2)\mod 2\pi\), has a unique Borel ergodic invariant probability measure \(\mu\) given by a normalized Lebesgue measure. Moreover, there exists an orthonormal basis of \(L^2(\mu)\) consisting of Koopman eigenfunctions \(z_{jk}(\omega) = e^{i(j\omega_1 + k\omega_2)}\), \(j, k \in \mathbb{Z}\), with eigenfrequencies \(\alpha_{jk} = j\nu_1 + k\nu_2\). Thus, \(H_\sigma = L^2(\mu)\), and \(H_c\) is the zero subspace. Such a system is said to have a pure point spectrum.

Example 2 (Lorenz 63 system). The L63 system on \(\Omega = \mathbb{R}^3\) is governed by a system of smooth ODEs with quadratic nonlinearities. This system is known to exhibit a physical ergodic invariant probability measure \(\mu\) supported on a compact set (the L63 attractor), with mixing dynamics. This means that \(H_\sigma\) is the one-dimensional subspace of \(L^2(\mu)\) consisting of constant functions, and \(H_c\) consists of all \(L^2(\mu)\) functions orthogonal to the constants (i.e., with zero expectation value with respect to \(\mu\)).

Delay-coordinate approaches. For the point spectrum subspace \(H_\sigma\), a natural class of coherent observables is provided by the Koopman eigenfunctions. Every Koopman eigenfunction \(z_j \in H_\sigma\) evolves as a harmonic oscillator at the corresponding eigenfrequency, \(U^t z_j = e^{iz_j^*/z_j}\), and the associated autocorrelation function, \(C_{z_j}(t) = e^{|z_j|^2t}\), also has a harmonic evolution. Short of temporal invariance (which only occurs for constant eigenfunctions under measure-preserving ergodic dynamics), it is natural to think of a harmonic evolution as being “maximally” coherent. In particular, if \(z_j\) is continuous, then for any \(\omega \in A\), the real and imaginary parts of the time series \(t \mapsto U^t z_j(\omega)\) are pure sinusoids, even if the flow \(\Phi^t\) is aperiodic. Further attractive properties of Koopman eigenfunctions include the facts that they are intrinsic to the dynamical system generating the data, and they are closed under pointwise multiplication, \(z_j z_k = z_{jk}\), allowing one to generate every eigenfunction from a potentially finite generating set.
Yet, consistently approximating Koopman eigenfunctions from data is a nontrivial task, even for simple systems. For instance, the torus rotation in Example 1 has a dense set of eigenfrequencies by rational independence of the basic frequencies $v_1$ and $v_2$. Thus, any open interval in $\mathbb{R}$ contains infinitely many eigenfrequencies $\sigma_{jk}$, necessitating some form of regularization. Arguably, the term “pure point spectrum” is somewhat of a misnomer for such systems since a nonempty continuous spectrum is present. Indeed, since the spectrum of an operator on a Banach space includes the closure of the set of eigenvalues, $i\mathbb{R} \setminus \{i\alpha_{jk}\}$ lies in the continuous spectrum.

As a way of addressing these challenges, observe that if $G$ is a self-adjoint, compact operator commuting with the Koopman group (i.e., $U^t G = GU^t$), then any eigenspace $W_\lambda$ of $G$ corresponding to a nonzero eigenvalue $\lambda$ is invariant under $U^t$, and thus under the generator $V$. Moreover, by compactness of $G$, $W_\lambda$ has finite dimension. Thus, for any orthonormal basis $\{\phi_0, \ldots, \phi_{l-1}\}$ of $W_\lambda$, the generator $V$ on $W_\lambda$ is represented by a skew-symmetric, and thus unitarily diagonalizable, $l \times l$ matrix $V = [(\phi_i, V\phi_j)]_{i,j=0}^{l-1}$. The eigenvectors $u_i = (u_{i0}, \ldots, u_{i,l-1})^T \in \mathbb{C}^l$ of $V$ then contain expansion coefficients of Koopman eigenfunctions $z = \sum_{j=0}^{l-1} u_j \phi_j$ in $W_\lambda$, and the eigenvalues corresponding to $u_i$ are eigenvalues of $V$.

On the basis of the above, since any integral operator $G$ on $L^2(\mu)$ associated with a symmetric kernel $k \in L^2(\mu \times \mu)$ is Hilbert–Schmidt (and thus compact), and we have a wide variety of data-driven tools for approximating integral operators, we can reduce the problem of consistently approximating the point spectrum of the Koopman group on $L^2(\mu)$ to the problem of constructing a commuting integral operator. As we now argue, the success of a number of techniques, including singular spectrum analysis (SSA) [BK86], diffusion-mapped delay coordinates (DMDC) [BCGFS13], nonlinear Laplacian spectral analysis (NLSA) [GM12], and Hankel DMD [BBP17], in identifying coherent patterns can at least be partly attributed to the fact that they employ integral operators that approximately commute with the Koopman operator.

A common characteristic of these methods is that they employ, in some form, delay-coordinate maps [SYC91]. With our notation for the covariate function $X : \Omega \to \mathcal{X}$ and sampling interval $\Delta t$, the $Q$-step delay-coordinate map is defined as $X_Q : \Omega \to \mathcal{X}^Q$ with $X_Q(\omega) = (X(\omega_0), X(\omega_{-1}), \ldots, X(\omega_{-Q+1}))$ and $\omega_q = \Phi^q(\omega)$. That is, $X_Q$ can be thought of as a lift of $X$, which produces “snapshots,” to a map taking values in the space $\mathcal{X}^Q$ containing “videos.” Intuitively, by virtue of its higher-dimensional codomain and dependence on the dynamical flow, a delay-coordinate map such as $X_Q$ should provide additional information about the underlying dynamics on $\Omega$ over the raw covariate map $X$. This intuition has been made precise in a number of “embedding” theorems [SYC91], which state that under mild assumptions, for any compact subset $S \subseteq \Omega$ (including, for our purposes, the invariant set $A$), the delay-coordinate map $X_Q$ is injective on $S$ for sufficiently large $Q$. As a result, delay-coordinate maps provide a powerful tool for state space reconstruction, as well as for constructing informative predictor functions in the context of forecasting.

Aside from considerations associated with topological reconstruction, however, observe that a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on covariate space pulls back to a distance-like function $\tilde{d}_Q : \Omega \times \Omega \to \mathbb{R}$ such that

$$\tilde{d}_Q^2(\omega, \omega') = \frac{1}{Q} \sum_{q=0}^{Q-1} d^2(X(\omega_{-q}), X(\omega'_{-q})).$$

In particular, $\tilde{d}_Q^2$ has the structure of an ergodic average of a continuous function under the product dynamical flow $\Phi^t \times \Phi^t$ on $\Omega \times \Omega$. By the von Neumann ergodic theorem, as $Q \to \infty$, $\tilde{d}_Q$ converges to $d_\infty$ a.e. constant, as $\Phi^t \times \Phi^t$ need not be ergodic, and aside from special cases it will not be continuous on $A \times A$. Nevertheless, based on the $L^2(\mu \times \mu)$ convergence of $\tilde{d}_Q$ to $d_\infty$, it can be shown [DG19] that for any continuous function $h : \mathbb{R} \to \mathbb{R}$, the integral operator $G_\infty$ on $L^2(\mu)$ associated with the kernel $k_\infty = h \circ d_\infty$ commutes with $U^t$ for any $t \in \mathbb{R}$. Moreover, as $Q \to \infty$, the operators $G_{Q,N}$ associated with $k_Q = h \circ d_Q$ converge to $G_\infty$ in $L^2(\mu)$ operator norm, and thus in spectrum.

Many of the operators employed in SSA, DMDC, NLSA, and Hankel DMD can be modeled after $G_Q$ described above. In particular, because $G_Q$ is induced by a continuous kernel, its spectrum can be consistently approximated by data-driven operators $G_{Q,N}$ on $L^2(\mu_N)$, as described in the context of forecasting. The eigenfunctions of these operators at nonzero eigenvalues approximate eigenfunctions of $G_\infty$, which approximate in turn eigenfunctions of $G_\infty$ lying in finite unions of Koopman eigenspaces. Thus, for sufficiently large $N$ and $Q$, the eigenfunctions of $G_{Q,N}$ at nonzero eigenvalues capture distinct timescales associated with the point spectrum of the dynamical system, providing physically interpretable features. These kernel eigenfunctions can also be employed in Galerkin schemes to approximate individual Koopman eigenfunctions.

Besides the spectral considerations described above, in [BCGFS13] a geometrical characterization of the eigenspaces of $G_Q$ was given based on Lyapunov metrics of dynamical systems. In particular, it follows by Oseledets’s multiplicative ergodic theorem that for $\mu$-a.e. $\omega \in M$ there exists a decomposition $T_\omega M = F_{1,\omega} \oplus \cdots \oplus F_r,\omega$, where $T_\omega M$
is the tangent space at $\omega \in M$, and $F_{\omega,\omega}$ are subspaces satisfying the equivariance condition $D\Phi F_{\omega,\omega} = F_{\omega,\Phi(\omega)}$. Moreover, there exist $\Lambda_1 > \cdots > \Lambda_r$, such that for every $v \in F_{\omega,\omega}$, 

$$\Lambda_j = \lim_{t \to \infty} \frac{1}{t} \int_0^t \log \|D\Phi^t v\| \, ds/t,$$

where $\|\cdot\|$ is the norm on $\mathcal{T}_\omega M$ induced by a Riemannian metric. The numbers $\Lambda_j$ are called Lyapunov exponents, and are metric-independent. Note that the dynamical vector field $\mathcal{V}(\omega)$ lies in a subspace $F_{0,\omega}$ with a corresponding zero Lyapunov exponent.

If $F_{0,\omega}$ is one-dimensional, and the norms $||D\Phi^t v||$ obey appropriate uniform growth/decay bounds with respect to $\omega \in M$, the dynamical flow is said to be uniformly hyperbolic. If, in addition, the support $A$ of $\mu$ is a differentiable manifold, then there exists a class of Riemannian metrics, called Lyapunov metrics, for which the $F_{\omega,\omega}$ are mutually orthogonal at every $\omega \in A$. In [BCGFS13], it was shown that using a modification of the delay-distance in (5) with exponentially decaying weights, as $Q \to \infty$, the top eigenfunctions $\phi_j(Q)$ of $G_Q$ vary predominantly along the subspace $F_{\omega,\omega}$ associated with the most stable Lyapunov exponent. That is, for every $\omega \in \Omega$ and tangent vector $v \in T_\omega M$ orthogonal to $F_{\omega,\omega}$ with respect to a Lyapunov metric, 

$$\lim_{Q \to \infty} \langle v, \nabla \phi_j(Q) \rangle = 0.$$
In [DGS18], a constructive procedure was proposed for obtaining the kernel family $k_\tau$ through a Markov semigroup on $L^2(\mu)$. This method has a data-driven implementation, with analogous spectral convergence results for the associated integral operators $G_{\tau,N}$ on $L^2(\mu_N)$ to those described in the setting of forecasting. Given these operators, we approximate $V_\tau$ by $\hat{V}_{\tau,N} = G_{\tau,N}^{1/2}V_N G_{\tau,N}^{1/2}$, where $V_N$ is a skew-adjoint, finite-difference approximation of the generator. For example, $V_N = (U_N^1 - U_N^0)/(2\Delta t)$ is a second-order finite-difference approximation based on the 1-step shift operator $U^1_N$. See Figure 1 for a graphical representation of a generator matrix for L63. As with our data-driven approximations of $U^1$, we work with a rank-$L$ operator $\hat{V}_\tau := \Pi_{\tau,N,L,\tau,N,L}\Pi_{\tau,N,L}$, where $\Pi_{\tau,N,L}$ is the orthogonal projection to the subspace spanned by the first $L$ eigenfunctions of $G_{\tau,N}$. This family of operators converges spectrally to $V_\tau$ in a limit of $N \to 0$, followed by $\Delta t \to 0$ and $L \to \infty$, where we note that $C^1$ regularity of $k_\tau$ is important for the finite-difference approximations to converge.

At any given $\tau$, an a posteriori criterion for identifying candidate eigenfunctions $\hat{z}_{j,\tau}$ satisfying (6) for small $\varepsilon$ is to compute a Dirichlet energy functional, $D(\hat{z}_{j,\tau}) = \|N_{\tau,N} \hat{z}_{j,\tau}\|_{L^2(\mu)}^2/\|\hat{z}_{j,\tau}\|_{L^2(\mu_N)}^2$. Intuitively, $D$ assigns a measure of roughness to every nonzero element in the domain of the Nyström operator (analogously to the Dirichlet energy in Sobolev spaces on differentiable manifolds), and the smaller $D(\hat{z}_{j,\tau})$ is, the more coherent $\hat{z}_{j,\tau}$ is expected to be. Indeed, as shown in Figure 3, the $\hat{z}_{j,\tau}$ corresponding to low Dirichlet energy identify observables of the L63 system with a coherent dynamical evolution, even though this system is mixing and has no nonconstant Koopman eigenfunctions. Sampled along dynamical trajectories, the approximate Koopman eigenfunctions resemble amplitude-modulated wavepackets, exhibiting a low-frequency modulating envelope while maintaining phase coherence and a precise carrier frequency. This behavior can be thought of as a “relaxation” of Koopman eigenfunctions, which generate pure sinusoids with no amplitude modulation.

Conclusions and Outlook
We have presented mathematical techniques at the interface of dynamical systems theory and data science for statistical analysis and modeling of dynamical systems. One of our primary goals has been to highlight a fruitful interplay of ideas from ergodic theory, functional analysis, and differential geometry, which, coupled with learning theory, provide an effective route for data-driven prediction and pattern extraction, well-adapted to handle nonlinear dynamics and complex geometries.

There are several open questions and future research directions stemming from these topics. First, it should be possible to combine pointwise estimators derived from methods such as diffusion forecasting and KAF with the

Figure 3. A representative eigenfunction $\hat{z}_{j,\tau}$ of the compactified generator $\hat{V}_\tau$ for the L63 system, with low corresponding Dirichlet energy. Top: Scatterplot of $\text{Re}\,\hat{z}_{j,\tau}$ sampled along a dynamical trajectory. Bottom: Time series of $\text{Re}\,\hat{z}_{j,\tau}$.
ACKNOWLEDGMENTS. Research of the authors described in this review was supported by DARPA grant HR0011-16-C-0116; NSF grants 1842538, DMS-1317919, DMS-1521775, DMS-1619661, DMS-1723171, DMS-1854383; and ONR grants N00014-12-1-0912, N00014-14-1-0150, N00014-13-1-0797, N00014-16-1-2649, N00014-16-1-2888.

References

[AG20] R. Alexander and D. Giannakis, Operator-theoretic forecasting for nonlinear time series with kernel analog techniques, Phys. D 409 (2020), 132520, 24, DOI 10.1016/j.physd.2020.132520. MR4093838

[BBP+17] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, Chaos as an intermittently forced linear system, Nat. Commun. 8 (2017), no. 19, DOI 10.1038/s41467-017-00303-8.

[BCGFS13] 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), no. 2, 618–649, DOI 10.1137/12088183X. MR3047433

[BGH15] T. Berry, D. Giannakis, and J. Harlim, Nonparametric forecasting of low-dimensional dynamical systems, Phys. Rev. E 91 (2015), 032915, DOI 10.1103/PhysRevE.91.032915.

[BH16] T. Berry and J. Harlim, Variable bandwidth diffusion kernels, Appl. Comput. Harmon. Anal. 40 (2016), no. 1, 68–96, DOI 10.1016/j.acha.2015.01.001. MR3431485

[BK86] D. S. Broomhead and G. P. King, Extracting qualitative dynamics from experimental data, Phys. D 20 (1986), no. 2-3, 217–236, DOI 10.1016/0167-2789(86)90031-X. MR859354

[BN03] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003), 1373–1396.

[CLO06] R. R. Coifman and S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006), no. 1, 5–30, DOI 10.1016/j.acha.2006.04.006. MR2238665

[CS02] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. (N.S.) 39 (2002), no. 1, 1–49, DOI 10.1090/S0273-0979-01-00923-5. MR1864085

[DEF00] M. Dellnitz, G. Froyland, and S. Sertl, On the isolated spectrum of the Perron-Frobenius operator, Nonlinearity 13 (2000), no. 4, 1171–1188, DOI 10.1088/0951-7715/13/4/310. MR1767953

[DG19] S. Das and D. Giannakis, Delay-coordinate maps and the spectra of Koopman operators, J. Stat. Phys. 175 (2019), no. 6, 1107–1145, DOI 10.1007/s10955-019-02272-w. MR3962976

[DGS18] S. Das, D. Giannakis, and J. Slawinska, Reproducing kernel Hilbert space compactification of unitary evolution groups, 2018, https://arxiv.org/abs/1808.01515.

[DJ99] M. Dellnitz and O. Junge, On the approximation of complicated dynamical behavior, SIAM J. Numer. Anal. 36 (1999), no. 2, 491–515, DOI 10.1137/S0036142996313002. MR1668207

[GM12] D. Giannakis and A. J. Majda, Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability, Proc. Natl. Acad. Sci. USA 109 (2012), no. 7, 2222–2227, DOI 10.1073/pnas.1118984109. MR2898568

[KNK+18] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, Data-driven model reduction and transfer operator approximation, J. Nonlinear Sci. 28 (2018), no. 3, 985–1010, DOI 10.1007/s00332-017-9437-7. MR3800253

[Koo31] B. O. Koopman, Hamiltonian systems and transformation in Hilbert space, Proc. Natl. Acad. Sci. 17 (1931), no. 5, 315–318, DOI 10.1073/pnas.17.5.315.

[KPM20] M. Korda, M. Putinar, and I. Mezić, Data-driven spectral analysis of the Koopman operator, Appl. Comput. Harmon. Anal. 48 (2020), no. 2, 599–629, DOI 10.1016/j.acha.2018.08.002. MR4047538

[Lor69] E. N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, J. Atmos. Sci. 26 (1969), 636–646, DOI 10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2.

[Mey05] I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, Nonlinear Dyn. 41 (2005), no. 1-3, 309–325, DOI 10.1007/s11071-005-2824-8. MR2157184

[RMB09] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, Spectral analysis of nonlinear flows, J. Fluid Mech. 641 (2009), 115–127, DOI 10.1017/S0022112009992059. MR2577895

[Sch10] P. J. Schmid, Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech. 656 (2010), 5–28, DOI 10.1017/S0022112010001217. MR2669948

[SMM98] B. Schölkopf, A. Smola, and K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998), 1299–1319.

[SYC91] T. Sauer, J. A. Yorke, and M. Casdagli, Embedology, J. Statist. Phys. 65 (1991), no. 3-4, 579–616, DOI 10.1007/BF01053745. MR1137425

[vLBB08] U. von Luxburg, M. Belkin, and O. Bousquet, Consistency of spectral clustering, Ann. Statist. 36 (2008), no. 2, 555–586, DOI 10.1214/009053607000000640. MR2396807

[WRK15] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, A data-driven approximation of the Koopman operator: extending dynamic mode decomposition, J. Nonlinear Sci. 25 (2015), no. 6, 1307–1346, DOI 10.1007/s00332-015-9258-5. MR3315049

[ZHL19] H. Zhang, J. Harlim, and X. Li, Computing linear response statistics using orthogonal polynomial based estimators: an RKHS formulation, 2019, https://arxiv.org/abs/1912.11110.
The School of Basic Sciences (Physics, Chemistry and Mathematics) at EPFL seeks to appoint a Tenure-Track Assistant Professor of Mathematics. We seek outstanding candidates with research interests in the theory of dynamical systems, broadly construed. Indicative areas of interest include, but are not limited to: algebraic/analytical dynamics, dynamics and geometry, ergodic theory, locally homogeneous spaces.

We expect candidates to establish leadership and strengthen the EPFL’s profile in the field. Priority will be given to the overall originality and promise of the candidate’s work over any particular specialization area.

Candidates should hold a PhD and have an excellent record of scientific accomplishments in the field. In addition, commitment to teaching at the undergraduate, master and doctoral levels is expected. Proficiency in French teaching is not required, but willingness to learn the language expected.

EPFL, with its main campus located in Lausanne, Switzerland, on the shores of lake Geneva, is a dynamically growing and well-funded institution fostering excellence and diversity. It has a highly international campus with first-class infrastructure, including high performance computing.

As a technical university covering essentially the entire palette of engineering and science, EPFL offers a fertile environment for research cooperation between different disciplines. The EPFL environment is multi-lingual and multi-cultural, with English often serving as a common interface.

Applications should include a cover letter, a CV with a list of publications, a concise statement of research (maximum 3 pages) and teaching interests (one page), and the names and addresses (including e-mail) of at least three references.

Applications should be uploaded (as PDFs) by September 30, 2020 to:

https://facultyrecruiting.epfl.ch/positiondetails/23691270

Enquiries may be addressed to:
Prof. Victor Panaretos
Chair of the Search Committee
E-mail: direction.math@epfl.ch

For additional information, please consult www.epfl.ch, sb.epfl.ch, math.epfl.ch

EPFL is an equal opportunity employer and family friendly university. It is committed to increasing the diversity of its faculty. It strongly encourages women to apply.