
Forecasting Sequential Data Using Consistent Koopman Autoencoders

Omri Azencot^{*1} N. Benjamin Erichson^{*2} Vanessa Lin² Michael W. Mahoney²

Abstract

Recurrent neural networks are widely used on time series data, yet such models often ignore the underlying physical structures in such sequences. A new class of physics-based methods related to Koopman theory has been introduced, offering an alternative for processing nonlinear dynamical systems. In this work, we propose a novel Consistent Koopman Autoencoder model which, unlike the majority of existing work, leverages the forward and backward dynamics. Key to our approach is a new analysis which explores the interplay between consistent dynamics and their associated Koopman operators. Our network is directly related to the derived analysis, and its computational requirements are comparable to other baselines. We evaluate our method on a wide range of high-dimensional and short-term dependent problems, and it achieves accurate estimates for significant prediction horizons, while also being robust to noise.

1. Introduction

Sequential data processing and forecasting is a fundamental problem in the engineering and physics sciences. Recurrent Neural Networks (RNNs) provide a powerful class of models for these tasks, designed to learn long-term dependencies via their hidden state variables. However, training RNNs over long time horizons is notoriously hard (Pascanu et al., 2013) due to the problem of exploding and vanishing gradients (Bengio et al., 1994). Several approaches have been proposed to mitigate this issue using unitary hidden-to-hidden weight matrices (Arjovsky et al., 2016) or analyzing stability properties (Miller & Hardt, 2019), among other solutions. Still, attaining long-term memory remains a

challenge and moreover, modeling short-term dependencies might also be affected due to the limited expressivity of unitary RNNs (Kerg et al., 2019).

Another shortcoming of traditional RNNs is their difficulty to incorporate high-level constraints into the model. In this context, physics-based methods have been proposed, relating RNNs to dynamical systems (Sussillo & Barak, 2013) or differential equations (Chang et al., 2019). This point of view allows one to construct models which enjoy high-level properties such as time invertibility via Hamiltonian (Greydanus et al., 2019) or Symplectic (Chen et al., 2020; Zhong et al., 2020) networks. Other techniques suggested reversible RNNs (MacKay et al., 2018; Chen et al., 2018) to alleviate the large memory footprints RNNs induce during training. In this work, we advocate that modeling time-series data which exhibit strong short-term dependencies can benefit from *relaxing* the strict stability as well as time invertibility requirements.

An interesting physics motivated alternative for analyzing time series data has been introduced in Koopman-based models (Takeishi et al., 2017; Morton et al., 2018; 2019; Li et al., 2020). Koopman theory is based on the insight that a nonlinear dynamical system can be fully encoded using an operator that describes how scalar functions propagate in time. The Koopman operator is *linear*, and thus preferable to work with in practice, as tools from linear algebra and spectral theory can be directly applied. While Koopman’s theory (Koopman, 1931) was established almost a century ago, significant advances have been recently accomplished in the theory and methodology with applications in the fluid mechanics (Mezić, 2005) and geometry processing (Ovsjanikov et al., 2012) communities.

The Koopman operator maps between function spaces and thus it is infinite-dimensional and can not be represented on a computer. Nevertheless, most machine learning approaches hypothesize that there exists a data transformation under which an approximate finite-dimensional Koopman operator is available. Typically, this map is represented via an autoencoder network, embedding the input onto a low-dimensional latent space. In that space, the Koopman operator is approximated using a linear layer that encodes the dynamics (Takeishi et al., 2017). The main advantage of this framework is that the resulting models are easy to

^{*}Equal contribution ¹Department of Mathematics at UC Los Angeles, CA, USA. ²ICSI and Department of Statistics at UC Berkeley, CA, USA.. Correspondence to: Omri Azencot <azencot@math.ucla.edu>, N. Benjamin Erichson <erichson@berkeley.edu>.

analyze, and they allow for accurate prediction of short-term dependent data. Specifically, predicting forward or backward in time can be attained via subsequent matrix-vector products between the Koopman matrix and the latent observation. Similarly, stability features can be analyzed and constrained via the operator spectrum.

Based on the theory of Koopman, we propose a new model for forecasting high-dimensional time series data. In contrast to previous approaches, we assume that the *backward* map exists. That is, the system from a future to current time can be properly defined. While not all systems exhibit this feature (e.g., diffusive systems), there are many practical cases where this assumption holds. We investigate the interplay between the forward and backward maps and their consistency in the discrete and continuous space settings. Our work can be viewed as relaxing both reversibility and stability requirements, leading to higher expressivity and improved forecasts. Ideas close to ours have been applied to training Generative Adversarial Networks (Zhu et al., 2017; Hoffman et al., 2018). However, to the best of our knowledge, our work is first to establish the link between consistency of latent variables and dynamical systems.

1.1. Background and Problem Setup

In what follows, we focus on dynamical systems that can be described by a time-invariant model

$$z_{k+1} = \varphi(z_k), \quad z \in \mathcal{M} \subset \mathbb{R}^m, \quad (1)$$

where z_k denotes the state of the system at time $k \in \mathbb{N}$. The map $\varphi : \mathcal{M} \rightarrow \mathcal{M}$ is a (potentially non-linear) update rule on a finite dimensional manifold \mathcal{M} , pushing states from time k to time $k + 1$. The above model assumes that future states depend only on the current state z_k and not on information from a sequence of previous states.

In order to predict future states, one might be tempted to train a neural network which learns an approximation of the map φ . However, the resulting model ignores prior knowledge related to the problem, and it can be challenging to analyze. Alternatively, our approach is based on a data transformation of the states z_k for which the corresponding latent variables q_k evolve on a linear path, as illustrated in Figure 1. Our method allows one to integrate into the training process physics priors such as the relation between the forward and backward maps. Interestingly, Koopman theory suggests that indeed, there exists such a data transformation for any nonlinear dynamical system, so that the evolution of states can be fully represented by a linear map.

More formally, the dynamics φ induces an operator \mathcal{K}_φ that acts on scalar functions $f : \mathcal{M} \rightarrow \mathbb{R} \in \mathcal{F}$, with \mathcal{F} being some function space on \mathcal{M} (Koopman, 1931). The

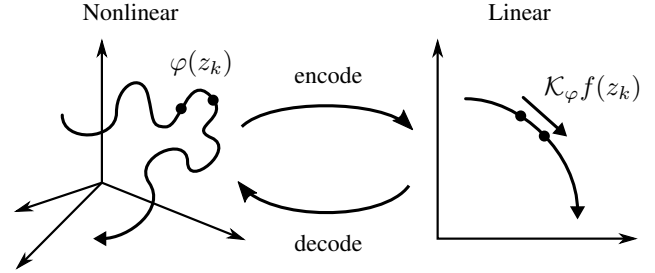


Figure 1. Illustration of the data transformation that maps the high-dimensional states z_k which evolve on a nonlinear trajectory via φ to a new space where the dynamics are linear and given by \mathcal{K}_φ .

Koopman operator is defined by

$$\mathcal{K}_\varphi f(z) = f \circ \varphi(z), \quad (2)$$

i.e., the function f is composed with the map φ . Essentially, the operator prescribes the evolution of a scalar function by pulling-back its values from a future time. In other words, $\mathcal{K}_\varphi f$ at z_k is the value of f evaluated at the future state z_{k+1} . Hence, the Koopman operator is also commonly known as the pull-back operator. Also, it is easy to show that \mathcal{K}_φ is *linear* for any $\alpha, \beta \in \mathbb{R}$

$$\begin{aligned} \mathcal{K}_\varphi(\alpha f + \beta g) &= (\alpha f + \beta g) \circ \varphi \\ &= \alpha f \circ \varphi + \beta g \circ \varphi = \alpha \mathcal{K}_\varphi f + \beta \mathcal{K}_\varphi g. \end{aligned}$$

Finally, we assume that the backward dynamics ψ exists, and we denote by \mathcal{U}_ψ the associated Koopman operator. In Fig. 2, we show an illustration of our setup.

Unfortunately, \mathcal{K}_φ is infinite-dimensional. Nevertheless, the key assumption in most of the practical approaches is that there exists a transformation χ whose conjugation with \mathcal{K}_φ leads to a *finite-dimensional* approximation which encodes “most” of the dynamics. Formally,

$$C = \chi \circ \mathcal{K}_\varphi \circ \chi^{-1}, \quad C \in \mathbb{R}^{\kappa \times \kappa}, \quad (3)$$

i.e., χ and its inverse extract the crucial structures from \mathcal{K}_φ , yielding an approximate Koopman matrix C . Similarly, we denote by $D = \chi \circ \mathcal{U}_\psi \circ \chi^{-1}$ the approximate backward system. The main focus in this work is to find the matrices C and D , and a nonlinear transformation χ such that the underlying dynamical system is recovered well.

We assume that we are given scalar observations of the dynamics $\{f_k : \mathcal{M} \rightarrow \mathbb{R}\}_{k=1}^n$ such that $f_{k+1} = f_k \circ \varphi + r_k$, where the function $r_k \in \mathcal{F}$ represents deviation from the true dynamics due to e.g., measurement errors or missing values. We focus on the case where φ and ψ are generally unknown, and our goal is to predict future observations from the given ones. Namely,

$$f_{k+l} = f_k \circ \varphi^l, \quad l = 1, 2, \dots, \quad (4)$$

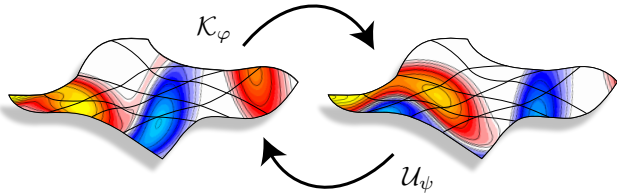


Figure 2. Our analysis and computational pipeline consider the forward and backward dynamics. Here, the operator \mathcal{K}_φ acts on a fluid flow field, detailing the evolution by pulling-back information from a future state. The operator \mathcal{U}_ψ acts in the reverse direction, i.e., prescribing the backward evolution by pushing-forward information from a past observation.

where φ^l means we repeatedly apply the dynamics. In practice, as C approximates the system, we exploit the relation $\chi^{-1} \circ C^l \circ \chi(f_k) \approx f_k \circ \varphi^l$ to produce further predictions. That is, the matrix C fully determines the forward evolution of the input observation f_k .

1.2. Main Contributions

Our main contributions are as follows.

- We developed a Physics Constrained Learning (PCL) framework based on Koopman theory and consistent dynamics for processing complex time series data.
- Our model is efficient and effective and its features include accurate predictions, time reversibility, and stable behavior even over long time horizons.
- We evaluate on high-dimensional clean and noisy systems including the pendulum, cylinder flow, vortex flow on a curved domain, and climate data, and we achieve exceptionally good results with our model.

2. Related Work

Modeling dynamical systems from Koopman’s point-of-view has gained increasing popularity in the last few years (Mezić, 2005). An approximation of the Koopman operator can be computed via the Dynamic Mode Decomposition (DMD) algorithm (Schmid, 2010). While many extensions of the original algorithm have been proposed, most related to our approach is the work of (Azencot et al., 2019) where the authors consider the forward and backward dynamics in a non-neural optimization setting. A network design similar to ours was proposed by (Lusch et al., 2018), but without our analysis, back prediction and consistency terms. Other techniques minimize the residual sum of squares (Takeishi et al., 2017; Morton et al., 2018), promote stability (Erichson et al., 2019; Pan & Duraisamy, 2020), use a variational approach (Mardt et al., 2018), or use graph convolutional networks (Li et al., 2020).

Sequential data are commonly processed using RNNs (Elman, 1990; Graves, 2012). The main difference between standard neural networks and RNNs is that the latter networks maintain a hidden state which uses the current input and previous inner states. Variants of RNNs such as Long Short Term Memory (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (Cho et al., 2014) have achieved groundbreaking results on various tasks including language modeling and machine translation, among others. Still, training RNNs involves many challenges and a recent trend in machine learning focuses on finding new interpretations of RNNs based on dynamical systems theory (Laurent & von Brecht, 2016; Miller & Hardt, 2019).

Several physics-based models have been recently proposed. Based on Lagrangian mechanics theory, Lutter et al. (2019) encoded the Euler–Lagrange equations into their network to attain physics plausibility and to alleviate poor generalization of deep models. Other methods attempt to learn conservation laws from data and their associated Hamiltonian representation, leading to exact preservation of energy (Greydanus et al., 2019) and better handling of stiff problems (Chen et al., 2020). To deal with the limited expressivity of unitary RNNs, Kerg et al. (2019) suggested to employ the Schur decomposition to their connectivity matrices. By considering the normal and nonnormal components, their network allows for transient expansion and compression, leading to improved results on tasks which require continued computations across timescales.

3. Method

In what follows, we describe our PCL model which we use to handle time series data. Similar to other methods, we model the transformation χ above via an encoder χ_e and a decoder χ_d . Our approach differs from other Koopman-based techniques (Lusch et al., 2018; Takeishi et al., 2017; Morton et al., 2018) in two key components. First, in addition to modeling the forward dynamics, our network also takes into account the *backward* system. Second, we require that the resulting forward and backward Koopman operators are *consistent*.

3.1. Autoencoding Observations

Given a set of observations $F = \{f_k\}_{k=1}^n$ as defined in Sec. 1.1, we design an autoencoder (AE) to embed our inputs in a low-dimensional latent space using a nonlinear map χ_e . The decoder χ_d map allows to reconstruct latent variables in the spatial domain. To train the AE, we define

$$\tilde{f} = \chi_d \circ \chi_e(f), \quad (5)$$

$$\mathcal{E}_{\text{id}} = \frac{1}{2n} \sum_{k=1}^n \|f_k - \tilde{f}_k\|_2^2, \quad (6)$$

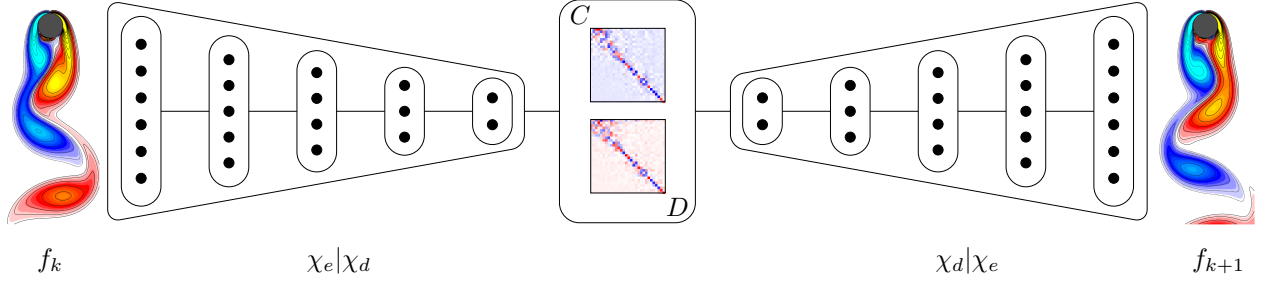


Figure 3. Our network takes observations f_k , and it learns a latent representation of them via an autoencoder architecture χ_e and χ_d . Then, the latent variables are propagated forward and backward in time using the linear layers C and D , respectively. We emphasize that all the above connections are bi-directional and so information can flow freely from left to right or right to left.

i.e., \tilde{f} is the reconstructed version of f , and \mathcal{E}_{id} derives the optimization to obtain an AE such that $\chi_d \circ \chi_e \approx \text{id}$. We note that the specific requirements from χ_e and χ_d are problem dependent, and we detail the particular design we use in the supplementary material (SM).

3.2. Backward Dynamics

In general, a dynamical system φ prescribes a rule to move forward in time. There are numerous practical scenarios where it makes sense to consider the backward system, i.e., $\psi : z_k \rightarrow z_{k-1}$. For instance, the Euler equation, which describes the motion of an inviscid fluid, is invariant to sign changes in its time parameter (see Sec. 5 for an example). Previous approaches incorporated the backward dynamics into their model as in bi-directional RNN (Schuster & Paliwal, 1997). However, the inherent nonlinearities of a typical neural network make it difficult to constrain the forward and backward models. To this end, a few approaches were recently proposed (Greydanus et al., 2019; Chen et al., 2020) where the obtained dynamics are reversible by construction due to the leapfrog integration. In comparison, most existing Koopman-based techniques do not consider the backward system in their modeling or training.

To account for the forward as well as backward dynamics, we incorporate two linear layers with no biases into our network to represent the approximate Koopman operators. As we assume that χ_e transforms our data into a latent space where the dynamics are linear, we can directly evolve the dynamics in that space. We introduce the following notation

$$\hat{f}_{k+1} = \chi_d \circ C \circ \chi_e(f_k), \quad (7)$$

$$\check{f}_{k-1} = \chi_d \circ D \circ \chi_e(f_k), \quad (8)$$

for every admissible k . Namely, the Koopman operators $C, D \in \mathbb{R}^{\kappa \times \kappa}$ allow to obtain forward estimates \hat{f}_{k+1} , and backward forecasts \check{f}_{k-1} .

In practice, we noticed that our models predict as well as generalize better if we employ a *multistep* forecasting, instead of computing one step forward and backward in time.

Given a choice of $\lambda_s \in \mathbb{N}$, the total number of prediction steps, we define the following loss terms

$$\mathcal{E}_{\text{fwd}} = \frac{1}{2\lambda_s n} \sum_{l=1}^{\lambda_s} \sum_{k=1}^n \|f_{k+l} - \hat{f}_{k+l}\|_2^2, \quad (9)$$

$$\mathcal{E}_{\text{bwd}} = \frac{1}{2\lambda_s n} \sum_{l=1}^{\lambda_s} \sum_{k=1}^n \|f_{k-l} - \check{f}_{k-l}\|_2^2, \quad (10)$$

where we assume that f_{k+l} and f_{k-l} are provided during training for any $l \leq \lambda_s$, see Eq. (4). Also, \hat{f}_{k+1} and \check{f}_{k-1} are obtained by taking powers l of C , respectively D , in Eqs. (7) and (8). We show in Fig. 3 a schematic illustration of our network design including the encoder and decoder components as well as the Koopman matrices C and D . Notice that all the connections are bi-directional, that is, data can flow from left to right *and* right to left.

Backward prediction. Koopman operators and their approximating matrices are linear objects that allow for greater flexibility when compared to other models for time series processing. One consequence of this linearity is that while existing Koopman-based nets (Lusch et al., 2018; Takeishi et al., 2017; Morton et al., 2018) are geared towards forward prediction, their evolution matrix C can be exploited for backward prediction as well. This computation is obtained via the inverse of C , i.e.,

$$\bar{f}_{k-1} = \chi_d \circ C^{-1} \circ \chi_e(f_k). \quad (11)$$

However, models that were trained for forward prediction typically produce poor backward predictions as we show in the SM. In contrast, we note that our model allows for the direct back prediction using the D operator and Eq. (8). Thus, while other techniques can technically produce backward predictions, our model supports it by construction.

3.3. Consistent Dynamics

The backward prediction penalty \mathcal{E}_{bwd} in itself only affects D , and it is completely independent of C . That is, C will not change due to backpropagating the error in Eq. (10). To

link between the forward and backward evolution matrices, we need to introduce an additional penalty that promotes consistent dynamics. Formally, we say that the maps φ and ψ are *consistent* if $\psi \circ \varphi(f) = f$ for any $f \in \mathcal{F}$. In the Koopman setting, we will show below that this condition is related to requiring that $DC = I_\kappa$, where I_κ is the identity matrix of size κ . However, our analysis shows that in fact the continuous space and discrete space settings differ, yielding related but different penalties. In this work, we will incorporate the following loss to promote consistency

$$\mathcal{E}_{\text{con}} = \sum_{k=1}^{\kappa} \frac{1}{2k} \|D_{k*} C_{*k} - I_k\|_F^2 + \frac{1}{2k} \|C_{k*} D_{*k} - I_k\|_F^2, \quad (12)$$

where D_{k*} and C_{*k} are the upper k rows of D and leftmost k columns of the matrix C , and $\|\cdot\|_F$ is the Frobenius norm.

Stability. Recently, stability has emerged as an important component for analyzing neural nets (Miller & Hardt, 2019). Intuitively, a dynamical system is stable if nearby points stay close under the dynamics. Mathematically, the eigenvalues of a linear system fully determine its behavior, providing a powerful tool for stability analysis. Indeed, the challenging problem of vanishing and exploding gradients can be elegantly explained by bounding the modulus of the weight matrices' eigenvalues (Arjovsky et al., 2016). To overcome these challenges, one can design networks that are stable by construction; see, e.g., Chang et al. (2019), among others.

We, on the other hand, relax the stability constraint and allow for quasi-stable models. In practice, our loss term (12) regularizes the nonconvex minimization by promoting the eigenvalues to get closer to the unit circle. The comparison in Fig. 4 highlights the stability features our model attains, whereas a non regularized network obtains unstable modes. From an empirical viewpoint, unstable behavior leads to rapidly diverging forecasts, as we show in Sec. 5.

3.4. A Consistent Dynamic Koopman Autoencoder

Combining all the pieces together, we obtain our PCL model for processing time series data. Our model is trained by minimizing a loss function whose minimizers guarantee that we achieve a good AE, that predictions through time are accurate, and that the forward and backward dynamics are consistent. We define our loss

$$\mathcal{E} = \lambda_{\text{id}} \mathcal{E}_{\text{id}} + \lambda_{\text{fwd}} \mathcal{E}_{\text{fwd}} + \lambda_{\text{bwd}} \mathcal{E}_{\text{bwd}} + \lambda_{\text{con}} \mathcal{E}_{\text{con}}, \quad (13)$$

where $\lambda_{\text{id}}, \lambda_{\text{fwd}}, \lambda_{\text{bwd}}, \lambda_{\text{con}} \in \mathbb{R}^+$ are user-defined positive parameters that balance between reconstruction, prediction and consistency. Finally, $\mathcal{E}_{\text{id}}, \mathcal{E}_{\text{fwd}}, \mathcal{E}_{\text{bwd}}, \mathcal{E}_{\text{con}}$ are defined in Eqs. (6), (9), (10), and (12), respectively.

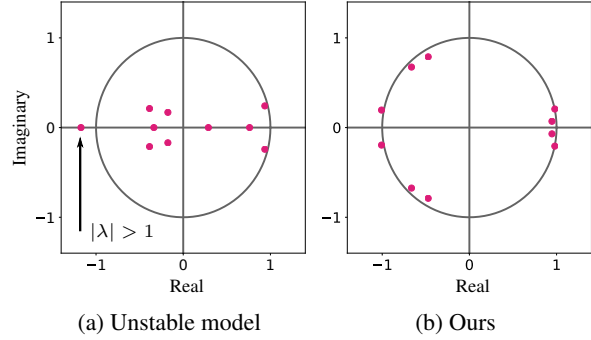


Figure 4. Koopman operators are linear and thus their spectrum can be investigated. We visualize the eigenvalues of an unregularized model (a) and ours (b) in the complex plane. Indeed, our PCL framework promotes stability, whereas the other model exhibits an eigenvalue with modulus greater than one.

4. Consistent Dynamics via Koopman

We now turn to prove a necessary and sufficient condition for a dynamical system to be consistent from a Koopman viewpoint in the continuous space setting. We then show a similar result in the spatial discrete case, yielding a more elaborate condition which we use in practice.

We recall that Koopman operators take inputs and return outputs from a function space \mathcal{F} . Thus, many properties of the underlying dynamics can be related to the action of \mathcal{K}_φ on every function in \mathcal{F} . A natural approach in this case is to consider a *spectral* representation of the associated objects. Specifically, we choose an orthogonal basis for \mathcal{F} which we denote by $\{\xi_k\}_{k=0}^\infty$ where for any i, j we have

$$\langle \xi_i, \xi_j \rangle_{\mathcal{M}} = \int_{\mathcal{M}} \xi_i(z) \xi_j(z) dz = \delta_{ij}, \quad (14)$$

with δ_{ij} being the Kronecker delta function. Under this choice of basis, any function $f \in \mathcal{F}$ can be represented by $f = \sum_k \langle f, \xi_k \rangle_{\mathcal{M}} \xi_k$. Moreover, due to the linearity of \mathcal{K}_φ we also have that $\mathcal{K}_{i,j} = \langle \xi_i, \mathcal{K}_\varphi \xi_j \rangle_{\mathcal{M}}$.

In the following proposition we characterize consistency (invertibility) in the continuous-space case, which is a known result in Ergodic theory (Eisner et al., 2015).

Proposition 1. *Given a manifold \mathcal{M} , the dynamical system φ is invertible if and only if for every i and j the Koopman operators \mathcal{K}_φ and \mathcal{U}_ψ satisfy $\langle \xi_i, \mathcal{U}_\psi \mathcal{K}_\varphi \xi_j \rangle_{\mathcal{M}} = \delta_{ij}$.*

Proof. If φ is invertible, then the composition $\psi \circ \varphi = \text{id}$ for every $z \in \mathcal{M}$. Thus,

$$\langle \xi_i, \mathcal{U} \mathcal{K} \xi_j \rangle_{\mathcal{M}} = \int_{\mathcal{M}} \xi_i(z) \xi_j(\psi \circ \varphi(z)) dz = \langle \xi_i, \xi_j \rangle_{\mathcal{M}}.$$

Conversely, we assume that $\langle \xi_i, \mathcal{U} \mathcal{K} \xi_j \rangle_{\mathcal{M}} = \delta_{ij}$ for all i, j . It follows that for every k we have $\mathcal{U} \mathcal{K} \xi_k = \xi_k$ since ξ_k is

orthogonal to every $\xi_l, l \neq k$. Let f be some scalar function, then $f(\psi \circ \varphi(z)) = \mathcal{U}\mathcal{K}f(z) = f(z)$ and thus $\psi \circ \varphi = \text{id}$.

The main advantage of Prop. 1 is that it can be used directly in a computational pipeline. In particular, if we denote by C and D the $\kappa \times \kappa$ matrices that approximate \mathcal{K} and \mathcal{U} , respectively, then the above condition takes the form

$$\frac{1}{2} \|DC - I_\kappa\|_F^2 = 0, \quad (15)$$

where I_κ is an identity matrix of size κ . This loss term was recently used in (Azencot et al., 2019) to construct a robust scheme for computing Dynamic Mode Decomposition operators. However, we prove below that in the *discrete-space* setting, a more elaborate condition is required. In addition, we refer to the work (Melzi et al., 2019) which deals with a closely related setup where isometric maps are being sought.

To discretize the above objects, we assume that our manifold \mathcal{M} is represented by the domain $M \subset \mathbb{R}^d$ which is sampled using m vertices. In this setup, scalar functions $f : M \rightarrow \mathbb{R}$ are vectors $f \in \mathbb{R}^m$ storing values on vertices with in-between values obtained via interpolation. A map $\varphi : M \rightarrow M$ can be encoded using a matrix $P_\varphi \in \mathbb{R}^{m \times m}$ defined by

$$P_\varphi \delta_z = h_{\varphi(z)}, \quad (16)$$

where h_x is a function that stores the vertices' coefficients such that $h_x^T X = x^T$, with $X \in \mathbb{R}^{m \times d}$ being the spatial coordinates of M . Similarly, we denote by Q_ψ the matrix associated with ψ , i.e., $Q_\psi \delta_z = h_{\psi(z)}$. Note that P and Q are in fact discrete Koopman operators represented in the canonical basis. We show in Fig. 5 an illustration of our spatial discrete setup including some of the notations.

Similar to the continuous setting, we can choose a basis for the function space on M . We denote $B \in \mathbb{R}^{m \times m}$ as the matrix that contains the orthogonal basis elements in its columns, i.e., $\langle b_i, b_j \rangle_M = b_i^T b_j = \delta_{ij}$ for every i, j . We use this basis to define the matrices C and D by

$$C = B^T P_\varphi B, \quad D = B^T Q_\psi B. \quad (17)$$

Finally, instead of invertible maps, we consider *consistent* maps. That is, a discrete map φ is consistent if for every z ,

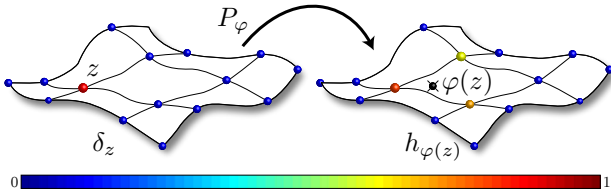


Figure 5. The Kronecker delta δ_z centered at z is mapped via P_φ to the function $h_{\varphi(z)}$ which holds the coefficients of the vertices of M whose combination generates $\varphi(z)$.

we have that $\psi \circ \varphi(z) = z$. Using the above constructions and notations, we are ready to state our main result.

Proposition 2. *Given a domain M , the map φ is consistent if and only if for every i and j the matrices C and D satisfy $\sum_k \frac{1}{2k} \|D_{k*} C_{*k} - I_k\|_F^2 = 0$, where D_{k*} and C_{*k} are the upper k rows of D and leftmost k columns of the matrix C .*

Proof. If φ is a consistent map, then $Q P \delta_z = \delta_z$ for every z and thus $Q P = I$. In addition, for every k we have that

$$D_{k*} C_{*k} = B_k^T Q B B^T P B_k = B_k^T Q P B_k = B_k^T B_k = I_k,$$

where B_k are the first k basis elements of B and $B B^T = I$. Conversely, we assume that C, D are related to some maps φ and ψ and are constructed via Eq (17). In addition, the condition $\sum_k \frac{1}{2k} \|D_{k*} C_{*k} - I_k\|_F^2 = 0$ holds. Then $B_k^T Q P B_k = I_k$ for every k . By induction on k , it follows that $Q P b_k = b_k$ for every k , where b_k is the k th column of B , and thus $Q P = I$, as B spans the space of scalar functions on M .

Discussion. Importantly, the indices of the sums in the penalty terms of Eq. (12) take values from $[1, \kappa]$, whereas the indices in Prop. 2 reach up to m (the sampling space of M). Thus, Eq. (12) may be viewed as an approximation of the full term. Nevertheless, this penalty term provides a closer approximation to the full term in comparison to Eq. (15), for any $\kappa \leq m$. Finally, we emphasize that our discussion also holds for the case $\kappa \ll m$.

5. Experiments

To evaluate our proposed consistent dynamic Koopman AE, we perform a comprehensive study using various datasets and compare to state-of-the-art Koopman-based approaches as well as other baseline sequential models. Our network minimizes Eq. (13) with a decaying learning rate initially set to 0.01. We fix the loss weights to $\lambda_{\text{id}} = \lambda_{\text{fwd}} = 1$, $\lambda_{\text{bwd}} = 0.1$, and $\lambda_{\text{con}} = 0.01$, for the AE, forward forecast, backward prediction and consistency, respectively. We use $\lambda_s = 8$ prediction steps forward and backward in time. We provide additional details in the SM. Our code is available at github.com/erichson/koopmanAE.

5.1. Baselines

Our comparison is mainly performed against the state-of-the-art method of Lusch et al. (2018), henceforth referred to as the Dynamic AE (DAE) model. Their approach may be viewed as a special case of our network by setting $\lambda_{\text{bwd}} = \lambda_{\text{con}} = 0$. While we use this work as a baseline, other models such as (Takeishi et al., 2017; Morton et al., 2018) could be also considered. The main difference between DAE and the latter techniques is the least squares solution for the evolution matrix C per training iteration. In our experience,

this change leads to delicate training procedures, and thus it is less favorable. Unless said otherwise, both models are trained using the same parameters, where DAE does not have the regularizing penalties.

We additionally compare against a feed-forward model and a recurrent neural network. The feed-forward network simply learns a nonlinear function $\zeta : f_k \rightarrow f_{k+1}$, where during inference we take \hat{f}_{k+1} as input for predicting \hat{f}_{k+2} , and so on. The recurrent neural network is similar to the feed-forward model but adds an hidden state h_k such that $h_k = \sigma(Uf_k + Wh_{k-1} + b)$, and the prediction is obtained via $\hat{f}_k = Vh_k + c$. We performed a parameter search when comparing with these baselines.

5.2. Nonlinear Pendulum with no Friction

The nonlinear (undamped) pendulum (Hirsch et al., 1974) is a classic textbook example for dynamical systems, which is also used for benchmarking deep models (e.g., Greydanus et al., 2019; Bertalan et al., 2019; Chen et al., 2020). This problem can be modeled as a second order ODE by

$$\frac{d^2\theta}{dt^2} + \frac{g}{l} \sin \theta = 0, \quad (18)$$

where the angular displacement from an equilibrium is denoted by $\theta \in [0, 2\pi)$. We use l and g to denote the length and gravity, respectively, with $l = 1$ and $g = 9.8$ in practice. We consider the following initial conditions $\theta(0) = \theta_0$ and $\dot{\theta}(0) = 0$. The motion of the pendulum is approximately harmonic for a small amplitude of the oscillation $\theta_0 \ll 1$. However, the problem becomes inherently nonlinear for large amplitudes of the oscillations.

We experiment with oscillation angles $\theta_0 = 0.8$ and $\theta_0 = 2.4$ on the time interval $t = [0, 170]$. The data are generated using a time step $\Delta t = 0.1$, yielding $T = 1700$ equally spaced points $x_1, \dots, x_T \in \mathbb{R}^2$. In addition, we map the sequence $\{x_t\}$ to a high-dimensional space via a random orthogonal transformation to obtain the training snapshots, i.e., $\mathbf{P} \in \mathbb{R}^{64 \times 2}$ such that $f_t = \mathbf{P} x_t$ for any t . Finally, we split the new sequence into a training set of 600 points and leave the rest for the test set. We show in Fig. 6 examples

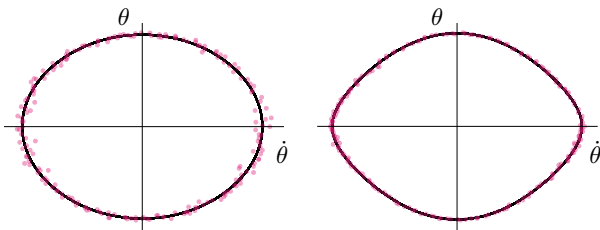


Figure 6. We show the pendulum’s trajectories for the initial conditions of amplitude oscillations $\theta_0 = 0.8$ (left) and $\theta_0 = 2.4$ (right). The discrete sampled data points are slightly perturbed.

of the clean and noisy trajectories for these data.

Experimental results. Fig. 7 shows the pendulum results for initial conditions $\theta_0 = 0.8$ (top row) and $\theta_0 = 2.4$ (bottom row). We used a bottleneck $\kappa = 6$ and $\alpha = 0.5$ for the DAE and our models. (The parameter α controls the width of the network, i.e., the number of neurons used for the hidden layers.) The relative forecasting error is computed at each time step via $\|f_t - \hat{f}_t\|_2 / \|f_t\|_2$, where \hat{f}_t is the high-dimensional estimated prediction, see also Eq. (7). We forecast over a time horizon of 1000 steps, and we average the error over 30 different initial observations f_t , where the shaded areas represent the ± 1 standard deviations. Overall, our model yields the best results in all the cases we explored. The RNN obtains good measures in the clean case and for short prediction times, but its performance deteriorates in the noisy setup and when forecasting is required for long horizons. The DAE model (Lusch et al., 2018) recovers the pendulum dynamics in the linear regime but struggles when the nonlinearity increases. We also outperform the Hamiltonian NN (Greydanus et al., 2019) in all settings.

Tab. 1 lists the prediction error at the final time step $\|f_T - \hat{f}_T\|_2 / \|f_T\|_2$ for both the DAE and our model. For each case we list the minimum, maximum and average prediction error evaluated for models that were trained with 18 different seed values.

model	θ	noise	prediction error			#parms
			min	max	avg.	
DAE	0.8	-	0.016	0.254	0.102	0.03M
Ours	0.8	-	0.011	0.080	0.034	0.03M
DAE	0.8	true	0.062	0.563	0.232	0.03M
Ours	0.8	true	0.045	0.362	0.131	0.03M
DAE	2.4	-	0.042	0.211	0.112	0.03M
Ours	2.4	-	0.027	0.171	0.074	0.03M
DAE	2.4	true	0.191	0.521	0.316	0.03M
Ours	2.4	true	0.063	0.592	0.187	0.03M

Table 1. Summary of results for the nonlinear pendulum with no friction. Our consistent Koopman model has a significantly reduced variance, and it outperforms the DAE in all situations. Each model was trained with 18 different seed values.

5.3. High-dimensional Fluid Flows

Next, we consider two challenging fluid flow examples. The first instance is a periodic flow past a cylinder that exhibits vortex shedding from boundary layers. This flow is commonly used in physically-based machine learning studies (Takeishi et al., 2017; Morton et al., 2018). The data are generated by numerically solving the Navier–Stokes equations given here in their vorticity form $\partial_t \omega = -\langle v, \nabla \omega \rangle + \frac{1}{Re} \Delta \omega$, where ω is the vorticity taken as the curl of the velocity, $\omega = \text{curl}(v)$. We employ an immersed boundary projec-

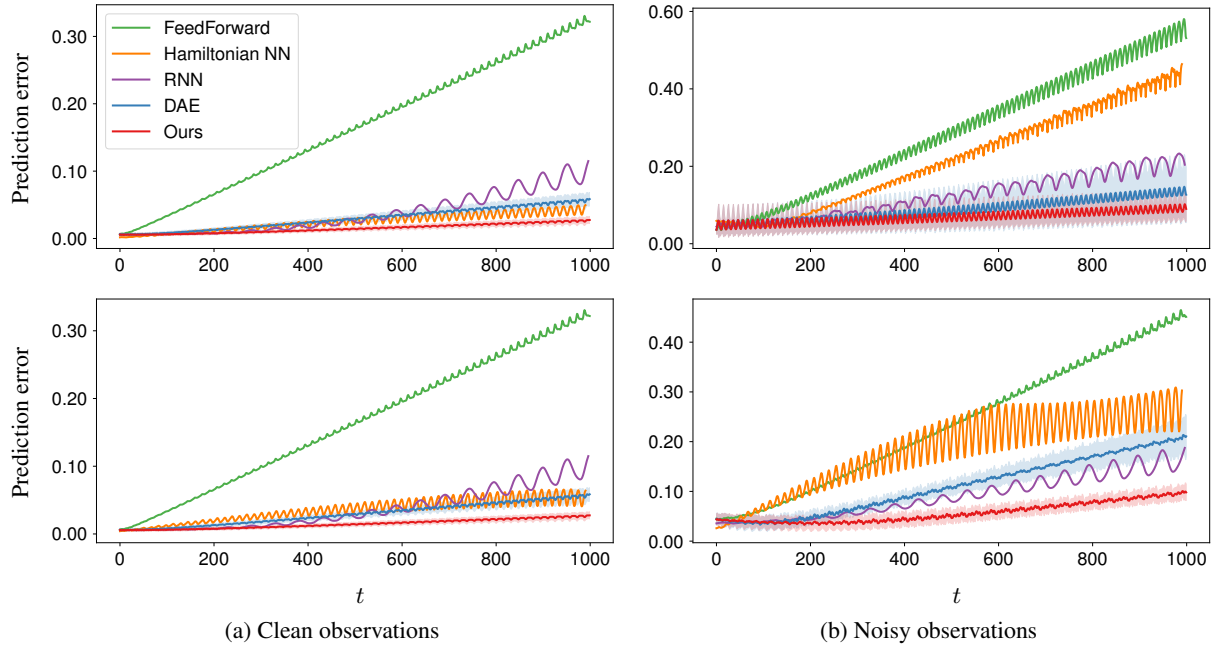


Figure 7. Prediction errors, over a time horizon of 1000 steps, for clean and noisy observations from a pendulum with initial conditions $\theta_0 = 0.8$ (top row) and $\theta_0 = 2.4$ (bottom row). Our model outperforms the DAE and the Hamiltonian NN in all settings.

tion solver (Taira & Colonius, 2007) with $Re = 100$. Our simulation yields 300 snapshots of 192×199 grid points, sampled at regular intervals in time, spanning five periods of vortex shedding. We split the data in half for training and testing. Our second example is an inviscid flow, i.e., $Re = \infty$, of a vortex pair travelling over a curved domain of a sphere given as a triangle mesh with 2562 nodes. We facilitate the intrinsic solver (Azencot et al., 2014) by producing 600 snapshots of which we use 550 for training.

Experimental results. It can be shown that the cylinder dynamics evolve on a low-dimensional attractor (Noack et al., 2003), which can be viewed as a nonlinear oscillator with a state-dependent damping (Loiseau & Brunton, 2018). As the dynamics are within the linear regime, we expect that on clean data, both DAE and ours will obtain good prediction results. We compare both models using width $\alpha = 2$ and bottleneck $\kappa = 10$, and we list in Tab. 2 the obtained predictions at $T = 100$. Indeed, for the clean data

model	noise	prediction error			#parms
		min	max	avg.	
DAE	-	0.078	0.348	0.156	2.49M
Ours	-	0.064	0.196	0.112	2.49M
DAE	true	0.125	0.982	0.343	2.49M
Ours	true	0.114	0.218	0.151	2.49M

Table 2. Summary of results for the flow past the cylinder. Each model was trained with 18 different seed values.

the results are similar in nature, whereas for the noisy case, our model provides more robust predictions. In general, our model outperforms the DAE when the data are noisy and the prediction horizon is long, illustrating the regularizing effect of our loss terms.

In Tab. 3, we show a few estimated predictions for the flow over a sphere. At the top part of the table we compare our method to DAE. Again, the results are consistent with previous examples where our approach yields better error and robustness estimates. In addition, we show a negative result of our method at the bottom of the table where we add increasing levels of diffusion to the Euler equation, making it non-invertible. Indeed, higher diffusion leads to inferior predictions when compared to the non-diffusive test.

model	diffusion	prediction error			#parms
		min	max	avg.	
DAE	0	0.071	1.558	0.715	0.34M
Ours	0	0.062	1.291	0.532	0.34M
Ours	0.0001	0.123	1.122	0.349	0.34M
Ours	0.001	0.44	1.386	0.725	0.34M
Ours	0.01	0.572	1.47	0.995	0.34M

Table 3. Summary of results for the flow over a sphere. Each model was trained with 18 different seed values.

5.4. Sea Surface Temperature Data

Our last dataset includes complex climate data representing the daily average sea surface temperature measurements

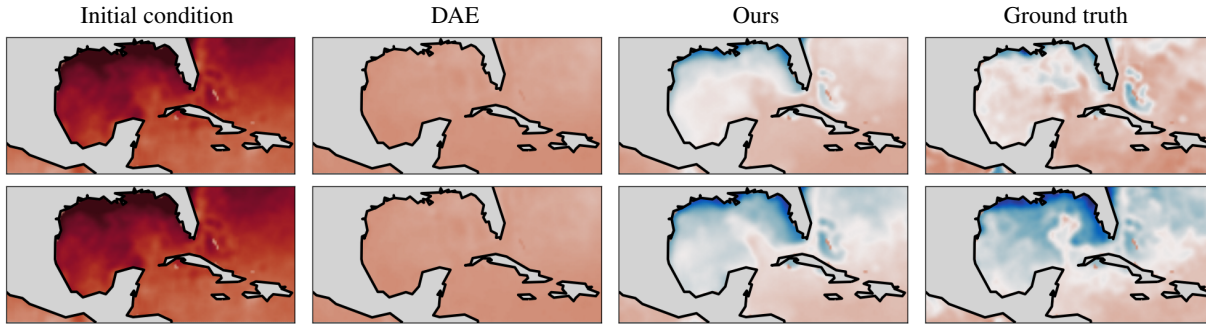


Figure 8. Starting from the same initial conditions, we use our model and DAE to forecast the sea surface temperature for the 120th day (top row) and 175th day (bottom row). The results can be visually compared to the ground truth data (right column). Our model generally attains predictions that are much closer to the ground truth compared to the results obtained by DAE.

around the Gulf of Mexico. This dataset is used in climate sciences to study the intricate dynamics between the oceans and the atmosphere. Climate prediction is generally a hard task involving challenges such as irregular heat radiation and flux as well as uncertainty in the wind behavior. Nevertheless, the input dynamics exhibit non-stationary periodic structures and are empirically low-dimensional, suggesting that Koopman-based methods can be employed. We extract a subset of the NOAA OI SST V2 High Resolution Dataset hereafter SST, and we refer to (Reynolds et al., 2007) for additional details. In our experiments, we use data with a spatial resolution of 100×180 spanning a time horizon of 1,305 days, of which 1,095 snapshots are used for training.

Experimental results. We demonstrate in Fig. 8 a visual comparison of the estimated predictions as obtained from the DAE model and ours. Here we are using width $\alpha = 6$ and bottleneck $\kappa = 10$. In both cases, forecasting for the 120th day (top) and 175th day (bottom), our results are closer to the ground truth by a large margin. A more detailed analysis is provided in Tab. 4 where we show the prediction errors averaged $\frac{1}{T} \sum_{t=1}^T |\hat{f}_t - f_t| / |f_t|$ over a time horizon of 180 days. (Here we use the average prediction error to better capture prediction errors that corresponds to seasonal effects.) The prediction results are significant, as predicting climate data is in general a challenging problem.

model	prediction error			#parms
	min	max	avg.	
DAE	0.315	1.062	0.701	2.05M
Ours	0.269	1.028	0.632	2.05M

Table 4. Summary of results for the sea surface temperature data. Each model was trained with 18 different seed values.

6. Ablation Study

To support our empirical results, we have also conducted an ablation study to quantify the effect of our additional loss terms when weighted differently. To this end, we revisit the

noisy pendulum flow with an initial condition $\theta_0 = 0.8$. Our results are summarized in Tab. 5 where we explore various values for λ_{bwd} and λ_{con} which balance the back prediction and consistency penalties, respectively. Our model generally outperforms other baselines for all the parameters we checked, measured via the average error (4th column) and most distant prediction error (5th column).

Model	λ_{bwd}	λ_{con}	$\frac{1}{T} \sum_{t=1}^T \frac{\ \hat{f}_t - f_t\ _2}{\ f_t\ _2}$	$\frac{\ \hat{f}_T - f_T\ _2}{\ f_T\ _2}$
DAE	0.0	0.0	0.10	0.15
Ours	1e-1	0.0	0.06	0.11
Ours (*)	1e-1	1e-2	0.06	0.08
Ours	1e-1	1e-1	0.07	0.12
Ours	2e-1	1e-2	0.05	0.06
Ours	2e-1	1e-1	0.06	0.08

Table 5. Ablation study for the pendulum with fixed seed 1. The star indicates the setting that is used during our experiments above.

7. Discussion

In this paper, we have proposed a novel physically constrained learning model for processing high-dimensional time series data. Our method is based on Koopman theory as we approximate dynamical systems via linear evolution matrices. Key to our approach is that we consider the backward dynamics during prediction, and we promote the consistency of the forward and backward systems. These modifications may be viewed as relaxing strict reversibility and stability constraints, while still regularizing the parameter space. We evaluate our method on several challenging datasets and compare with a state-of-the-art Koopman-based network as well as other baselines. Our approach notably outperforms the other models on noisy data and for long time predictions. We believe that our work can be extended in many ways, and in the future, we plan on considering our setup within a recurrent neural network design.

Acknowledgements

We are grateful to the generous support from Amazon AWS. OA would like to acknowledge the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 793800. NBE would like to acknowledge CLTC for providing support of this work. MWM would like to acknowledge ARO, DARPA, NSF and ONR for providing partial support of this work. We would like to thank J. Nathan Kutz, Steven L. Brunton, Lionel Mathelin and Alejandro Queiruga for valuable discussions about dynamical systems and Koopman theory. Further, we would like to acknowledge the NOAA for providing the SST data (<https://www.esrl.noaa.gov/psd/>).

References

- Arjovsky, M., Shah, A., and Bengio, Y. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pp. 1120–1128, 2016.
- Azencot, O., Weißmann, S., Ovsjanikov, M., Wardetzky, M., and Ben-Chen, M. Functional fluids on surfaces. In *Computer Graphics Forum*, volume 33, pp. 237–246. Wiley Online Library, 2014.
- Azencot, O., Yin, W., and Bertozzi, A. Consistent dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 18(3):1565–1585, 2019.
- Bengio, Y., Simard, P., and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
- Bertalan, T., Dietrich, F., Mezić, I., and Kevrekidis, I. G. On learning Hamiltonian systems from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):121107, 2019.
- Chang, B., Chen, M., Haber, E., and Chi, E. H. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, 2019.
- Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6571–6583, 2018.
- Chen, Z., Zhang, J., Arjovsky, M., and Bottou, L. Symplectic recurrent neural networks. In *International Conference on Learning Representations*, 2020.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.
- Eisner, T., Farkas, B., Haase, M., and Nagel, R. *Operator theoretic aspects of Ergodic theory*, volume 272. Springer, 2015.
- Elman, J. L. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.
- Erichson, N. B., Muehlebach, M., and Mahoney, M. W. Physics-informed autoencoders for Lyapunov-stable fluid flow prediction. *arXiv preprint arXiv:1905.10866*, 2019.
- Graves, A. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 5–13. Springer, 2012.
- Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, pp. 15353–15363, 2019.
- Hirsch, M. W., Devaney, R. L., and Smale, S. *Differential equations, dynamical systems, and linear algebra*, volume 60. Academic press, 1974.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1989–1998, 2018.
- Kerg, G., Goyette, K., Touzel, M. P., Gidel, G., Vorontsov, E., Bengio, Y., and Lajoie, G. Non-normal recurrent neural network (nnRNN): Learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems*, pp. 13591–13601, 2019.
- Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, 17(5):315, 1931.
- Laurent, T. and von Brecht, J. A recurrent neural network without chaos. *arXiv preprint arXiv:1612.06212*, 2016.
- Li, Y., He, H., Wu, J., Katabi, D., and Torralba, A. Learning compositional Koopman operators for model-based control. In *International Conference on Learning Representations*, 2020.
- Loiseau, J.-C. and Brunton, S. L. Constrained sparse Galerkin regression. *Journal of Fluid Mechanics*, 838:42–67, 2018.

- Lusch, B., Kutz, J. N., and Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.
- Lutter, M., Ritter, C., and Peters, J. Deep Lagrangian networks: Using physics as model prior for deep learning. In *International Conference on Learning Representations*, 2019.
- MacKay, M., Vicol, P., Ba, J., and Grosse, R. B. Reversible recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 9029–9040, 2018.
- Mardt, A., Pasquali, L., Wu, H., and Noé, F. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):1–11, 2018.
- Melzi, S., Ren, J., Rodola, E., Sharma, A., Wonka, P., and Ovsjanikov, M. Zoomout: Spectral upsampling for efficient shape correspondence. *arXiv preprint arXiv:1904.07865*, 2019.
- Mezić, I. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, 2005.
- Miller, J. and Hardt, M. Stable recurrent models. In *International Conference on Learning Representations*, 2019.
- Morton, J., Jameson, A., Kochenderfer, M. J., and Witherden, F. Deep dynamical modeling and control of unsteady fluid flows. In *Advances in Neural Information Processing Systems*, pp. 9258–9268, 2018.
- Morton, J., Witherden, F. D., and Kochenderfer, M. J. Deep variational Koopman models: Inferring Koopman observations for uncertainty-aware dynamics modeling and control. *arXiv preprint arXiv:1902.09742*, 2019.
- Noack, B. R., Afanasiev, K., Morzyński, M., Tadmor, G., and Thiele, F. A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. *Journal of Fluid Mechanics*, 497:335–363, 2003.
- Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., and Guibas, L. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 31(4):30, 2012.
- Pan, S. and Duraisamy, K. Physics-informed probabilistic learning of linear embeddings of nonlinear dynamics with guaranteed stability. *SIAM Journal on Applied Dynamical Systems*, 19(1):480–509, 2020.
- Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318, 2013.
- Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007.
- Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656: 5–28, 2010.
- Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- Sussillo, D. and Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.
- Taira, K. and Colonius, T. The immersed boundary method: A projection approach. *Journal of Computational Physics*, 225(2):2118–2137, 2007.
- Takeishi, N., Kawahara, Y., and Yairi, T. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, pp. 1130–1140, 2017.
- Zhong, Y. D., Dey, B., and Chakraborty, A. Symplectic ODE-net: Learning Hamiltonian dynamics with control. In *International Conference on Learning Representations*, 2020.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international Conference on Computer Vision*, pp. 2223–2232, 2017.