LEARNING INVARIANT REPRESENTATIONS OF TIME-HOMOGENEOUS STOCHASTIC DYNAMICAL SYSTEMS

Vladimir R. Kostic*
Istituto Italiano di Tecnologia
University of Novi Sad
vladimir.kostic@iit.it

Pietro Novelli*
Istituto Italiano di Tecnologia
pietro.novelli@iit.it

Riccardo Grazzi
Istituto Italiano di Tecnologia
University College London

Karim Lounici
CMAP École Polytechnique

Massimiliano Pontil
Istituto Italiano di Tecnologia
University College London

ABSTRACT

We consider the general class of time-homogeneous stochastic dynamical systems, both discrete and continuous, and study the problem of learning a representation of the state that faithfully captures its dynamics. This is instrumental to learning the transfer operator or the generator of the system, which in turn can be used for numerous tasks, such as forecasting and interpreting the system dynamics. We show that the search for a good representation can be cast as an optimization problem over neural networks. Our approach is supported by recent results in statistical learning theory, highlighting the role of approximation error and metric distortion in the learning problem. The objective function we propose is associated with projection operators from the representation space to the data space, overcomes metric distortion, and can be empirically estimated from data. In the discrete-time setting, we further derive a relaxed objective function that is differentiable and numerically well-conditioned. We compare our method against state-of-the-art approaches on different datasets, showing better performance across the board.

1 INTRODUCTION

Dynamical systems are a mathematical framework describing the evolution of state variables over time. These models, often represented by nonlinear differential equations (ordinary or partial) and possibly stochastic, have broad applications in science and engineering, ranging from climate sciences (Cannarsa et al., 2020; Fisher et al., 2009), to finance (Pascucci, 2011), to atomistic simulations (Mardt et al., 2018; McCarty and Parrinello, 2017; Schütte et al., 2001), and to open quantum system dynamics (Gorini and Kossakowski, 1976; Lindblad, 1976), among others. However, it is usually the case that no analytical models of the dynamics are available and one must resort to data-driven techniques to characterize a dynamical system. Two powerful paradigms have emerged: deep neural networks (DNN) and kernel methods. The latter are backed up by solid statistical guarantees (Steinwart and Christmann, 2008) determining when linearly parameterized models can be learned efficiently. Yet, selecting an appropriate kernel function may be a hard task, requiring a significant amount of experimentation and expertise. In comparison, the former are very effective in learning complex data representations (Goodfellow et al., 2016), and benefit from a solid ecosystem of software tools, making the learning process feasible on large-scale systems. However, their statistical analysis is still in its infancy, with only a few proven results on their generalization properties.

Kernel-based methods hinge on the powerful idea of characterizing dynamical systems by lifting their definition over a Hilbert space of functions and then studying the associated *transfer operators*. They describe the average evolution of functions of the state (*observables*) over time and for deterministic systems are also known as *Koopman operators*. Furthermore, transfer operators are *linear*, and under additional assumptions admit a spectral decomposition, which provides a valuable tool to interpret and analyze the behavior of non-linear systems (see e.g. Brunton et al., 2022; Kutz et al., 2016, and

*Equal contribution, corresponding authors.

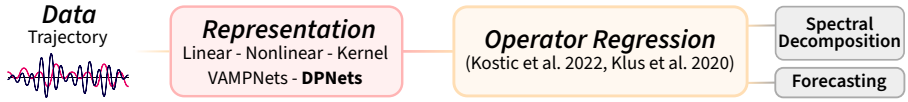

Figure 1: Pipeline for learning dynamical systems. DPNets learn a data representation to be used with standard operator regression methods. In turn, these are employed to solve downstream tasks such as forecasting and interpreting dynamical systems via spectral decomposition.

references therein). The usefulness of this approach critically relies on an appropriate choice of the *observable space*. In particular, in order to fully benefit from the spectral decomposition, it is crucial to find an observable space $\mathcal{F}$ which linearizes the dynamics and is *invariant* under the action of the transfer operator. In kernel-based methods, such a space is implicitly linked to the kernel function and gives a clear mathematical meaning to the problem of choosing a "good" kernel. Unfortunately, the analytical form of transfer operators is often intractable or unavailable, especially in complex or poorly understood systems, posing challenges in constructing invariant representations.

In this paper, we build synergistically upon both kernel and DNN paradigms: we first employ DNNs to learn an invariant representation that fully captures the system dynamics, and then forward this representation to kernel-based algorithms for the actual transfer operator regression task. This general framework is illustrated in Fig. 1. Our method, named Deep Projection Networks (DPNets), addressing the challenge of providing good representations to the operator regression algorithms, can be cast as an optimization problem over neural networks and can benefit from a differentiable and numerically well-conditioned score functional, enhancing the stability of the training process.

Previous work Extensive research has been conducted on learning dynamical systems from data. The monographs (Brunton et al., 2022; Kutz et al., 2016) are standard references in this field. To learn transfer operators we mention the works (Alexander and Giannakis, 2020; Bouvrie and Hamzi, 2017; Das and Giannakis, 2020; Kawahara, 2016; Klus et al., 2019; Kostic et al., 2022; Williams et al., 2015b) presenting kernel-based algorithms, and (Azencot et al., 2020; Bevanda et al., 2021; Fan et al., 2021; Lusch et al., 2018; Morton et al., 2018) based on deep learning schemes. Finding meaningful representations of the state of the system to be used in conjunction with transfer operator learning is a critical challenge, tackled by many authors. We mention the works (Azencot et al., 2020; Federici et al., 2024; Lusch et al., 2018; Morton et al., 2018; Otto and Rowley, 2019) where a representation is learned via DNN schemes, as well as (Kawahara, 2016; Li et al., 2017; Mardt et al., 2019; 2018; Tian and Wu, 2021; Yeung et al., 2019; Wu and Noé, 2019) addressing the problem of learning invariant subspaces of the transfer operator. Mostly related to our methodology is (Andrew et al., 2013), which introduced deep canonical correlation analysis and, especially, VAMPnets (Mardt et al., 2018; Wu and Noé, 2019), which repurposed this approach to learn transfer operators. Within our theoretical framework, indeed, we will show that VAMPNets can be recovered as a special case in the setting of discrete dynamics. More in general, operator learning with DNN approaches is reviewed e.g. in Kovachki et al. (2023), with a specific focus on PDEs.

Contributions Our contributions are: **1)** Leveraging recent results in statistical learning theory, we formalize the problem of representation learning for dynamical systems and design a score function based on the orthogonal projections of the transfer operator in data spaces (Sec. 2); **2)** We show how to reliably solve the corresponding optimization problem. In the discrete case, our score leads to optimal representation learning under the mild assumption of compactness of the transfer operator (Thm. 1). In the continuous case, our method applies to time reversal invariant dynamics (Thm. 2) including the important case of Langevin dynamics; **3)** We show how the learned representation can be used within the framework of operator regression in a variety of settings. Numerical experiments illustrate the versatility and competitive performance of our approach against several baselines.

Notation We let $\mathbb{N}$ be the set of natural numbers and $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$. For $m \in \mathbb{N}$ we denote $[m] := \{1, \dots, m\}$. If $\mathcal{T}$ is a compact linear operator on two Hilbert spaces we let $\sigma_i(\mathcal{T})$ be its i-th largest singular value, we let $(\mathcal{T})^\dagger$ be the Moore-Penrose pseudo-inverse, and $\mathcal{T}^*$ the adjoint operator. We denote by $\|\cdot\|$ and $\|\cdot\|_{\mathrm{HS}}$ the operator and Hilbert-Schmidt norm, respectively.

2 Representation Learning for Dynamical Systems

In this section, we give some background on transfer operators for dynamical systems, discuss key challenges in learning them from data, and propose our representation learning approach.

Background Let $(X_t)_{t\in\mathbb{T}}$ be a stochastic process taking values in some state space $\mathcal{X}$, where the time index t can be either discrete ($\mathbb{T} = \mathbb{N}_0$) or continuous ($\mathbb{T} = [0, +\infty)$), and denote by μ_t the law of X_t. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a prescribed space of real-valued functions, henceforth referred to as observables space. By letting $s, t \in \mathbb{T}$, with $s \leq t$, the *forward transfer operator* $\mathcal{T}_{s,t} \colon \mathcal{F} \to \mathcal{F}$ evolves an observable $f : \mathcal{X} \to \mathbb{R}$ from time $s \in \mathbb{T}$ to time t, by the conditional expectation

$$[\mathcal{T}_{s,t}(f)](x) := \mathbb{E}[f(X_t)\,|X_s = x],\ x \in \mathcal{X}. \tag{1}$$

For a large class of stochastic dynamical systems, these *linear* operators are *time-homogeneous*, that is they only depend on the time difference $t - s$. In this case $\mathcal{T}_{s,t} = \mathcal{T}_{0,t-s} =: \mathcal{T}_{t-s}$. Further, we will use the shorthand notation $\mathcal{T} := \mathcal{T}_1$. Time-homogeneous transfer operators at different times are related through the Chapman-Kolmogorov equation (Allen, 2007), $\mathcal{T}_{s+t} = \mathcal{T}_t\,\mathcal{T}_s$, implying that the family $(\mathcal{T}_t)_{t\in\mathbb{T}}$ forms a semigroup. The utility of transfer operators is that, when the space $\mathcal{F}$ is suitably chosen, they *linearize* the process. Two key requirements are that $\mathcal{F}$ is *invariant* under the action of $\mathcal{T}_t$, that is $\mathcal{T}_t[\mathcal{F}] \subseteq \mathcal{F}$ for all t (a property that we tacitly assumed above), and *rich enough* to represent the flow of the process. Two common choices for $\mathcal{F}$ that fulfill both requirements are the space of bounded Borel-measurable functions and the space of square-integrable functions $L^2_\pi(\mathcal{X})$ with respect to the invariant distribution π, when one exists. To ease our presentation we first focus on the latter case, and then extend our results to non-stationary time-homogeneous processes. Formally, the process $(X_t)_{t\in\mathbb{T}}$ admits an *invariant distribution* π when $\mu_s = \pi$ implies $\mu_t = \pi$ for all $s \leq t$ (Ross, 1995). This in turn allows one to define the transfer operator on $\mathcal{F} = L^2_\pi(\mathcal{X})$.

While in the discrete-time setting the whole process can be studied only through $\mathcal{T} = \mathcal{T}_1$, when time is continuous the process is characterized by the infinitesimal generator of the semigroup $(\mathcal{T}_t)_{t\geq 0}$, defined as

$$\mathcal{L} := \lim_{t\to 0^+} (\mathcal{T}_t - I)/t. \tag{2}$$

$\mathcal{L}$ can be properly defined on the Sobolev space $W^{1,2}_\pi(\mathcal{X}) \subset L^2_\pi(\mathcal{X})$ formed by functions with square-integrable gradients (see Lasota and Mackey, 1994; Ross, 1995).

Learning transfer operators In practice, dynamical systems are only observed, and neither $\mathcal{T}$ nor its domain $\mathcal{F} = L^2_\pi(\mathcal{X})$ are known, providing a key challenge to learn them from data. The most popular algorithms (Brunton et al., 2022; Kutz et al., 2016) aim to learn the action of $\mathcal{T} \colon \mathcal{F} \to \mathcal{F}$ on a predefined Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, forming a subset of functions in $\mathcal{F}$. This allows one, via the kernel trick, to formulate the problem of learning the *restriction* of $\mathcal{T}$ to $\mathcal{H}$, $\mathcal{T}_{|\mathcal{H}} \colon \mathcal{H} \to \mathcal{F}$, via empirical risk minimization (Kostic et al., 2022). However, recent theoretical advances (Korda and Mezić, 2017; Klus et al., 2019; Nüske et al., 2023), proved that such algorithms are statistically consistent only to $P_{\mathcal{H}}\mathcal{T}_{|\mathcal{H}}$, where $P_{\mathcal{H}}$ is the orthogonal projection onto the closure of $\mathcal{H}$ in $\mathcal{F}$. The projection $P_{\mathcal{H}}$ constrains the *evolved* observables back inside $\mathcal{H}$, thereby, in general, altering the dynamics of the system. Therefore, to assure that one properly learns the dynamics, two major requirements on $\mathcal{H}$ are needed: i) $\mathcal{T}_{|\mathcal{H}}$ needs to approximate well $\mathcal{T}$, i.e. the space $\mathcal{H}$ needs to be big enough relative to the domain of $\mathcal{T}$; ii) the difference between the projected restriction and the true one, i.e. the approximation error $\left\|[I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}}\right\|$, needs to be small.

When $\mathcal{H}$ is an infinite-dimensional *universal* RKHS, both the above requirements are satisfied (Kostic et al., 2022), i.e. $\mathcal{H}$ is dense in $\mathcal{F}$ and the approximation error is zero, leading to an arbitrarily good approximation of dynamics with enough data. Still, another key issue arises: the norms on the a-priori chosen $\mathcal{H}$ and the unknown $\mathcal{F}$ do not coincide, since the latter depends on the process itself. This *metric distortion* phenomenon has been recently identified as the source of spurious estimation of the spectra of $\mathcal{T}$ (Kostic et al., 2023), limiting the utility of the learned transfer operators. Indeed, even if $\mathcal{T}$ is self-adjoint, that is the eigenfunctions are orthogonal in $\mathcal{F}$, the estimated ones will not be orthogonal in $\mathcal{H}$, giving rise to spectral pollution (Kato, 1976). This motivates one to additionally require that iii) $\mathcal{H}$ is a *subspace* of $\mathcal{F}$, i.e. both spaces have the same norm.

To summarize, the desired optimal $\mathcal{H}$ is the *leading invariant subspace* of $\mathcal{T}$, that is the subspace corresponding to the largest (in magnitude) eigenvalues of $\mathcal{T}$. This subspace $\mathcal{H}$ achieves zero approximation error, eliminates metric distortion and best approximates (in the dynamical system sense) the operator $\mathcal{T}$. Since any RKHS $\mathcal{H}$ is entirely described by a *feature map*, learning a leading invariant subspace $\mathcal{H}$ from data is, fundamentally, a *representation learning* problem.

Approach We start by formalizing the problem of learning a good finite dimensional representation space for $\mathcal{T}$, and then address the same for the generator $\mathcal{L}$. We keep the discussion less formal here, and state our main results more precisely in the next section. Our approach is inspired by the

following upper and lower bounds on the approximation error, a direct consequence of the norm change from $\mathcal{H}$ to $\mathcal{F}$,

$$\left\|[I - P_{\mathcal{H}}]\mathcal{T}P_{\mathcal{H}}\right\|^2 \lambda^+_{\min}(C_{\mathcal{H}}) \leq \left\|[I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}}\right\|^2 \leq \left\|[I - P_{\mathcal{H}}]\mathcal{T}P_{\mathcal{H}}\right\|^2 \lambda_{\max}(C_{\mathcal{H}}), \tag{3}$$

where $C_{\mathcal{H}}$ is the covariance operator on $\mathcal{H}$ w.r.t. the measure π, while $\lambda^+_{\min}$ and $\lambda_{\max}$ are the smallest and largest non-null eigenvalues, respectively. Note that the norms on the hypothetical domain $\mathcal{H}$ and true domain $L^2_\pi(\mathcal{X})$ coincide if and only if $C_{\mathcal{H}} = I$, in which case equalities hold in (3) and the approximation error becomes $\left\|[I - P_{\mathcal{H}}]\mathcal{T}P_{\mathcal{H}}\right\|$.

When the operator $\mathcal{T}$ is known, the latter quantity can be directly minimized by standard numerical algorithms for spectral computation to find invariant subspaces (see e.g. Golub and Van Loan, 2013). Unfortunately, in our stochastic setting $\mathcal{T}$ is unknown since we cannot compute the conditional expectation in (1). To overcome this issue we propose a learning approach to recover the invariant space $\mathcal{H}$, which is rooted in the singular value decomposition, holding under the mild assumption that $\mathcal{T}$ is a compact operator[1]. The main idea is that the subspace made of the leading r *left singular functions* of $\mathcal{T}$ serves as a good approximation of the desired leading invariant subspace of $\mathcal{T}$. Namely, due to the orthonormality of the singular functions, we have that $C_{\mathcal{H}} = I$ and $P_{\mathcal{H}}\mathcal{T}$ becomes the r-truncated SVD of $\mathcal{T}$, that is, its best rank-r approximation. Therefore, according to (3), the approximation error is at most $\sigma_{r+1}(\mathcal{T})$, which can be made arbitrarily small by rising r. Moreover, the distance of the subspace of left singular functions to the desired leading invariant subspace is determined by the "*normality*" of $\mathcal{T}$ (Trefethen and Embree, 2020). If the operator $\mathcal{T}$ is *normal*, that is $\mathcal{T}\mathcal{T}^* = \mathcal{T}^*\mathcal{T}$, then both its left and right singular spaces are invariant subspaces of $\mathcal{T}$, resulting in zero approximation error irrespectively of r. This leads us to the following optimization problem

$$\max_{\mathcal{H},\mathcal{H}' \subset L^2_\pi(\mathcal{X})} \left\{ \left\|P_{\mathcal{H}}\mathcal{T}P_{\mathcal{H}'}\right\|^2_{\mathrm{HS}} \mid C_{\mathcal{H}} = C_{\mathcal{H}'} = I, \dim(\mathcal{H}) \leq r,\, \dim(\mathcal{H}') \leq r \right\}. \tag{4}$$

Relying on the application of Eckart-Young-Mirsky's Theorem (Kato, 1976), we can show that the desired representation space $\mathcal{H}$ can be computed by solving (4). Note that, in general, the auxiliary space $\mathcal{H}'$ is needed to capture right singular functions, while if we have prior knowledge that $\mathcal{T}$ is normal without loss of generality one can set $\mathcal{H} = \mathcal{H}'$ in (4).

The same reasoning cannot be applied straight away to the generator $\mathcal{L}$ of a continuous dynamical system, which in is not even guaranteed to be bounded (Lasota and Mackey, 1994), let alone compact. For *time-reversal invariant* processes, however, $\mathcal{T}$ and $\mathcal{L}$ are self-adjoint, that is $\mathcal{T} = \mathcal{T}^*$ and $\mathcal{L} = \mathcal{L}^*$. This includes the important case of Langevin dynamics, which are of paramount importance for molecular dynamics (see, e.g., Tuckerman, 2010). For such systems we show (see Theorem 2) that the leading r-dimensional subspace of the generator $\mathcal{L}$ can be found by solving the following optimization problem featuring the *partial trace* objective

$$\max_{\mathcal{H} \subset W^{1,2}_\pi(\mathcal{X})} \left\{ \mathrm{tr}\left(P_{\mathcal{H}}\,\mathcal{L}\,P_{\mathcal{H}}\right) \mid C_{\mathcal{H}} = I, \dim(\mathcal{H}) \leq r \right\}, \tag{5}$$

where now the projector is $P_{\mathcal{H}} \colon W^{1,2}_\pi(\mathcal{X}) \to W^{1,2}_\pi(\mathcal{X})$. We stress that, since we assume $\mathcal{L} = \mathcal{L}^*$, we can relax the above assumption on the compactness of the transfer operators $(\mathcal{T}_t)_{t\geq 0}$, and still show that the leading invariant subspace of $\mathcal{L}$ is the optimal $\mathcal{H}$ in (5). So, on such $\mathcal{H}$ we can estimate well $\mathcal{L}$ via generator operator regression, see e.g. (Hou et al., 2023; Klus et al., 2020).

3 Deep projection score functionals

In this section we show how to solve the representation learning problems (4) and (5) using DNNs. In the discrete case, we consider a generalized version of problem (4), which encompasses non-stationary processes, for which the probability distributions change along the trajectory. Namely, let X and X' be two $\mathcal{X}$-valued random variables with probability measures μ and μ', respectively. Because of time-homogeneity, w.l.o.g. X models the state at time 0 and X' its evolution after some time t. Then, the transfer operator can be defined on the data-dependent domains $\mathcal{T}_t \colon L^2_{\mu'}(\mathcal{X}) \to L^2_\mu(\mathcal{X})$. Replacing $P_{\mathcal{H}'}$ with $P'_{\mathcal{H}'} \colon L^2_{\mu'}(\mathcal{X}) \to L^2_{\mu'}(\mathcal{X})$ in (4), and using the covariances $C_{\mathcal{H}}$ and $C_{\mathcal{H}'}$ w.r.t.

[1]This property is fulfilled by a large class of Markov processes (see e.g. Kostic et al., 2022) and is weaker than requiring the operator being Hilbert-Schmidt as in (Mardt et al., 2018).

the measures μ and μ', we obtain the appropriate representation learning problem for non-stationary process; see App. B for more details. Within this general setting we optimize two feature maps

$$\psi_w : \mathcal{X} \to \mathbb{R}^r, \text{ and } \psi'_w : \mathcal{X} \to \mathbb{R}^r, \tag{6}$$

parameterized by w taking values in some set $\mathcal{W}$. Next, defining the two RKHSs[2] $\mathcal{H}_w := \text{span}(\psi_{w,j})_{j\in[r]}$ and $\mathcal{H}'_w := \text{span}(\psi'_{w,j})_{j\in[r]}$, both equipped with the standard inner product, we aim to solve (4) by optimizing over the weights w. To avoid solving the constrained optimization problem, we further propose to relax the hard constraints in (4) through a *metric distortion loss* $\mathcal{R}\colon \mathbb{R}^{r\times r} \to \mathbb{R}_+$ that is zero if and only if it is applied to the identity matrix. Our choice in Section 5, well defined for SPD matrices C, is

$$\mathcal{R}(C) := \text{tr}(C^2 - C - \ln(C)).$$

Finally, in Lem. 3, App. C we show that (4) can be solved by maximizing, for $\gamma > 0$

$$\mathcal{P}^\gamma(w) = \left\|(C_X^w)^{\dagger/2} C_{XX'}^w (C_{X'}^w)^{\dagger/2}\right\|_{\text{HS}}^2 - \gamma\left(\mathcal{R}(C_X^w) + \mathcal{R}(C_{X'}^w)\right), \tag{7}$$

where we introduced the uncentered covariance matrices

$$C_X^w := \mathbb{E}\,\psi_w(X)\psi_w(X)^\top,\ C_{X'}^w := \mathbb{E}\,\psi'_w(X')\psi'_w(X')^\top \text{ and } C_{XX'}^w := \mathbb{E}\,\psi_w(X)\psi'_w(X')^\top. \tag{8}$$

Notice that if $\gamma = 0$ and the covariances C_X^w and $C_{X'}^w$ are nonsingular, then the score reduces to $\left\|(C_X^w)^{-\frac{1}{2}} C_{XX'}^w (C_{X'}^w)^{-\frac{1}{2}}\right\|_{\text{HS}}^2$ which is called VAMP-2 score in the VAMPNets approach of (Mardt et al., 2018). Moreover, if in (7) the HS norm is replaced by the nuclear norm, the score becomes the objective of canonical correlation analysis (CCA) (Harold, 1936; Hardoon et al., 2004) in feature space. When DNNs are used to parametrize (6), such score is the objective of Deep-CCA (Andrew et al., 2013), known as VAMP-1 score (Mardt et al., 2018) in the context of molecular kinetics. Therefore, by maximizing $\mathcal{P}^\gamma$ we look for the strongest *linear correlation* between $\mathcal{H}_w$ and $\mathcal{H}'_w$. We stress that differently from Deep-CCA and VAMPNets the crucial addition of $\mathcal{R}$ is to *decorrelate* the features within each space, guiding the maximization towards a solution in which the norms of $\mathcal{H}$ and $\mathcal{H}'$ coincide with those of $L^2_\mu(\mathcal{X})$ and $L^2_{\mu'}(\mathcal{X})$, overcoming metric distortion.

While in the optimal representation the covariances C_X^w and $C_{X'}^w$ are the identity, in general they are non-invertible, making the score $\mathcal{P}^\gamma$ non-differentiable during optimization. Indeed, unless the rank of both covariances is stable for every $w \in \mathcal{W}$, we might have exploding gradients (Golub and Pereyra, 1973). Even ignoring differentiability issues, the use of the pseudo-inverse, as well as the use of the inverse in the non-singular case, can introduce severe numerical instabilities when estimating $\mathcal{P}^\gamma$ and its gradients during the training process. More precisely, the numerical conditioning of evaluating $\mathcal{P}^\gamma$ using (7) is determined by the covariance condition numbers $\lambda_1(C_X^w)/\lambda_r(C_X^w)$ and $\lambda_1(C_{X'}^w)/\lambda_r(C_{X'}^w)$ that can be very large in practical situations (see e.g. the fluid dynamics example in Sec. 5) To overcome this issue, we introduce the relaxed score

$$\mathcal{S}^\gamma(w) := \left\|C_{XX'}^w\right\|_{\text{HS}}^2 / \left(\left\|C_X^w\right\|\left\|C_{X'}^w\right\|\right) - \gamma\left(\mathcal{R}(C_X^w) + \mathcal{R}(C_{X'}^w)\right), \tag{9}$$

which, as implied by Theorem 1 below, is a lower bound for the score in (7). A key advantage of this approach is that the score $\mathcal{S}^\gamma$ is both differentiable (apart from the trivial case $C_X^w, C_{X'}^w = 0$) and has stable gradients, since we avoid matrix inversion. Indeed, computing $\mathcal{S}^\gamma$ is always well-conditioned, as the numerical operator norm has conditioning equal to one.

The following theorem provides the theoretical backbone of our representation learning approach.

Theorem 1. *If $\mathcal{T}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ is compact, $\mathcal{H}_w \subseteq L^2_\mu(\mathcal{X})$ and $\mathcal{H}'_w \subseteq L^2_{\mu'}(\mathcal{X})$, then for all $\gamma \geq 0$*

$$\mathcal{S}^\gamma(w) \leq \mathcal{P}^\gamma(w) \leq \sigma_1^2(\mathcal{T}) + \cdots + \sigma_r^2(\mathcal{T}). \tag{10}$$

Moreover, if $(\psi_{w,j})_{j\in[r]}$ and $(\psi'_{w,j})_{j\in[r]}$ are the leading r left and right singular functions of $\mathcal{T}$, respectively, then both equalities in (10) *hold. Finally, if the operator $\mathcal{T}$ is Hilbert-Schmidt, $\sigma_r(\mathcal{T}) > \sigma_{r+1}(\mathcal{T})$ and $\gamma > 0$, then the "only if" relation is satisfied up to unitary equivalence.*

This result establishes that the relaxed score is a tight lower bound, in the optimization sense, for (7). More precisely, in general $|\mathcal{S}^\gamma(w) - \mathcal{P}^\gamma(w)|$ is arbitrarily small for small enough $\mathcal{R}(w)$, while on the feasible set in (4) both scores coincide. For instance, for a deterministic linear system this happens as

[2]Here, $\psi_{w,j}$ and $\psi'_{w,j}$ are the j-th component of ψ_w and ψ'_w, respectively.

Algorithm 1 DPNets Training

Input: Data $\mathcal{D}$; metric loss parameter γ; DNNs $\psi_w, \psi'_w : \mathcal{X} \to \mathbb{R}^r$; optimizer U; # of steps k.

1: Initialize DNN weights to w_1.
2: **for** $j = 1$ to k **do**;
3: $\hat{C}_X^{w_j}, \hat{C}_{X'}^{w_j}, \hat{C}_{XX'}^{w_j} \leftarrow$ Covariances for ψ_{w_j} and ψ'_{w_j} from a mini-batch of $m \leq n$ samples.
4: $F(w_j) \leftarrow -\widehat{\mathcal{S}}_m^{\gamma}(w_j)$ from (9) and the covariances in Step 3.
5: $w_{j+1} \leftarrow U(w_j, \nabla F(w_j))$ where $\nabla F(w_j)$ is computed using backpropagation.
6: **end for**
7: **return** representations ψ_{w_K}, ψ'_{w_K}.

soon as the classes of features ψ_w and ψ'_w include linear functions. In general, we can appeal to the universal approximation properties of DNN (Cybenko, 1989) to fulfill the hypotheses of Theorem 1.

Generator learning We now consider how to learn the representation of continuous time-homogeneous processes arising from stochastic differential equations (SDE) of the form

$$dX_t = A(X_t)\, dt + B(X_t)\, dW_t, \tag{11}$$

where A and B are the drift and diffusion terms, respectively, and W_t is a Wiener process. The solution of (11) is a stochastic process that is time-reversal invariant w.r.t. its invariant distribution π, (Arnold, 1974). To show that solving (5) leads to learning the leading invariant subspace of $\mathcal{L}$, we restrict to the case $\mathcal{L} = \mathcal{L}^*$ and consider the partial trace of $\mathcal{L}$ w.r.t. to the subspace $\mathcal{H}_w := \text{span}(\psi_{w,j})_{j\in[r]} \subseteq W_\pi^{1,2}(\mathcal{X})$ as our objective function. Being $\mathcal{L}$ self-adjoint, we can relax the compactness assumption of Thm. 1, requiring only the much weaker condition that $\mathcal{L}$ has the largest eigenvalues separated from the *essential spectrum* (Kato, 1976). In particular, it holds $\text{tr}\left(P_{\mathcal{H}_w}\, \mathcal{L}\, P_{\mathcal{H}_w}\right) = \text{tr}\left((C_X^w)^\dagger C_{X\partial}^w\right)$, where $C_{X\partial}^w = \mathbb{E}[\psi_w(X)\, d\psi_w(X)^\top]$ is the continuous version of the cross-covariance and $d\psi_w(\cdot)$ is given by the Itō formula (see e.g. (Arnold, 1974; Klus et al., 2020; Hou et al., 2023))

$$d\psi_{w,i}(x) := \mathbb{E}[\nabla\psi_{w,i}(X)^\top dX/dt + \tfrac{1}{2}(dX/dt)^\top \nabla^2\psi_{w,i}(X)(dX/dt) \,|\, X = x] \tag{12}$$

$$= \nabla\psi_{w,i}(x)^\top A(X) + \tfrac{1}{2}\,\text{tr}(B(x)^\top \nabla^2\psi_{w,i}(x)B(x)). \tag{13}$$

The following result, the proof of which is given in App. E, then justifies our continuous DPNets.

Theorem 2. *If $\mathcal{H}_w \subseteq W_\pi^{1,2}(\mathcal{X})$, and $\lambda_1(\mathcal{L}) \geq \cdots \geq \lambda_{r+1}(\mathcal{L})$ are eigenvalues of $\mathcal{L}$ above its essential spectrum, then for every $\gamma \geq 0$ it holds*

$$\mathcal{P}_\partial^\gamma(w) := \text{tr}\left((C_X^w)^\dagger C_{X\partial}^w\right) - \gamma\mathcal{R}(C_X^w) \leq \lambda_1(\mathcal{L}) + \cdots + \lambda_r(\mathcal{L}), \tag{14}$$

and the equality is achieved when $\psi_{w,j}$ is the eigenfunction of $\mathcal{L}$ corresponding to $\lambda_j(\mathcal{L})$, for $j \in [r]$.

Notice how $\text{tr}\left((C_X^w)^\dagger C_{X\partial}^w\right)$ in Eq. (14) is the sum of the (finite) eigenvalues of the symmetric eigenvalue problem $C_{X\partial}^w - \lambda C_X^w$. Thus, in non-pathological cases, its value and gradient can be computed efficiently in a numerically stable way, see e.g. (Andrew and Tan, 1998).

4 METHODS

Learning the representation To optimize a DPNet representation one needs to replace the population covariances and cross-covariance in Eqs (7), (9) and (14) with their empirical counterparts. In practice, given $\psi_w, \psi'_w \colon \mathcal{X} \to \mathbb{R}^r$ and a dataset $\mathcal{D} = (x_i, x'_i)_{i\in[n]}$ consisting of samples from the joint distribution of (X, X') one estimates the needed covariances C_X^w, $C_{X'}^w$ and $C_{XX'}^w$ as

$$\widehat{C}_X^w = \frac{1}{n}\sum_{i\in[n]} \psi_w(x_i)\psi_w(x_i)^\top, \widehat{C}_{X'}^w = \frac{1}{n}\sum_{i\in[n]} \psi'_w(x'_i)\psi'_w(x'_i)^\top \text{ and } \widehat{C}_{XX'}^w = \frac{1}{n}\sum_{i\in[n]} \psi_w(x_i)\psi'_w(x'_i)^\top.$$

In App. D.1 we prove that the empirical scores $\widehat{\mathcal{P}}_n^\gamma \colon \mathcal{W} \to \mathbb{R}$ and $\widehat{\mathcal{S}}_n^\gamma \colon \mathcal{W} \to \mathbb{R}$, that is the scores computed using empirical covariances, concentrate around the true ones with high probability for any fixed w. Obtaining uniform guarantees over w requires advanced tools from empirical processes and regularity assumptions on the representations which could be the subject of future work. The difficulties arising from the estimation of $\mathcal{P}^\gamma$ during optimization are also addressed.

In Alg. 1 we report the training procedure for DPNets-relaxed (using score (9)), which can also be applied to (7) and (14) suitably modifying step 4. When using mini-batch SGD methods to compute $\nabla\widehat{\mathcal{P}}^{\gamma}_m(w)$ or $\nabla\widehat{\mathcal{S}}^{\gamma}_m(w)$, the batch size m should be taken sufficiently large to mitigate the impact of biased stochastic estimators. The time complexity of the training algorithm is addressed in App. D.4.

Operator regression Once a representation ψ_w is learned, it can be used within the framework of operator regression, see Fig. 1. Following the reasoning in (Kostic et al., 2022) one sees that any model $\widehat{\mathcal{T}}_w \colon \mathcal{H}_w \to \mathcal{H}_w$ of the transfer operator $\mathcal{T}$ acts on functions in $\mathcal{H}_w$ as

$$h_z := \psi_w(\cdot)^\top z \quad \mapsto \quad \widehat{\mathcal{T}}_w h_z := \psi_w(\cdot)^\top \widehat{T} z, \; z \in \mathbb{R}^r, \tag{15}$$

where $\widehat{T} \in \mathbb{R}^{r\times r}$ is a matrix in the representation space. For example, denoting the data matrices $\widehat{\Psi}_w = [\psi_w(x_1) \,|\, \cdots \,|\, \psi_w(x_n)]$, $\widehat{\Psi}'_w = [\psi_w(x'_1) \,|\, \cdots \,|\, \psi_w(x'_n)] \in \mathbb{R}^{r\times n}$, the ordinary least square estimator (OLS) minimizes the empirical risk $\left\|\widehat{\Psi}'_w - \widehat{T}^\top \widehat{\Psi}_w\right\|^2_{\mathrm{HS}}$, yielding $\widehat{T} := (\widehat{\Psi}_w^\top)^\dagger \widehat{\Psi}'^\top_w$.

Downstream tasks Once $\widehat{T}$ is obtained, it can be used predict the next expected state of the dynamics given the initial one, compute the modal decomposition of an observable (Kutz et al., 2016), estimate the spectrum of the transfer operator (Kostic et al., 2023) or controlling the dynamics (Proctor et al., 2016). Indeed, recalling that X' is the one step ahead evolution of X, we can use (15) to approximate $\mathbb{E}[h_z(X') \,|\, X = x]$ as $\psi_w(x)^\top \widehat{T} z$. Moreover, relying on the reproducing property of $\mathcal{H}_w$, the predictions can be extended to functions $f \colon \mathcal{X} \to \mathbb{R}^\ell$ as $\mathbb{E}[f(X') \,|\, X = x] \approx \psi_w(x)^\top (\widehat{\Psi}_w^\top)^\dagger \widehat{F}'$, where $\widehat{F}' = [f(x'_1) \,|\, \ldots \,|\, f(x'_n)]^\top \in \mathbb{R}^{n\times \ell}$ is the matrix of the observations of the evolved data. Clearly, when the observable is the state itself (i.e. $f(x) = x$), we obtain one step ahead predictions.

Next, observe that eigenvalue decomposition of the matrix $\widehat{T}$ leads to the eigenvalue decomposition of the operator $\widehat{\mathcal{T}}$. Namely, let $(\widehat{\lambda}_i, \widehat{u}_i, \widehat{v}_i) \in \mathbb{C} \times \mathbb{C}^d \times \mathbb{C}^d$ be an eigen-triplet made of eigenvalue, left eigenvector and right eigenvector of $\widehat{T}$, that is $\widehat{T}\widehat{v}_i = \widehat{\lambda}_i \widehat{v}_i$, $\widehat{u}_i^* \widehat{T} = \widehat{\lambda}_i \widehat{u}_i^*$ and $\widehat{u}_i^* \widehat{v}_k = \delta_{i,k}$, $i, k \in [r]$, we directly obtain, using (15), the spectral decomposition

$$\widehat{\mathcal{T}}_w = \textstyle\sum_{i\in[r]} \widehat{\lambda}_i \, \widehat{f}_i \otimes \widehat{g}_i, \quad \text{where} \quad \widehat{f}_i(x) := \psi_w(x)^\top \widehat{v}_i \;\text{ and }\; \widehat{g}_i(x) := (\widehat{u}_i)^* \psi_w(x). \tag{16}$$

Finally, we can use the spectral decomposition of $\widehat{T}$ to efficiently forecast observables for several time-steps in the future via $\mathbb{E}[h_z(X_t) \,|\, X_0 = x] \approx \sum_{i\in[r]} \widehat{\lambda}_i^t \, \langle \widehat{g}_i, h_z\rangle_{\mathcal{H}_w} \, \widehat{f}_i(x)$, which is known as the extended dynamic mode decomposition (EDMD), see e.g. (Brunton et al., 2022). Noticing that $\langle \widehat{g}_i, h_z\rangle_{\mathcal{H}_w} = (\widehat{u}_i)^* z = (\widehat{u}_i)^* \widehat{T} z / \widehat{\lambda}_i$, when $\widehat{\lambda}_i \neq 0$, and, again, using the reproducing property, we can extend this approximation to vector-valued observables $f \colon \mathcal{X} \to \mathbb{R}^\ell$ as

$$\mathbb{E}[f(X_t) \,|\, X_0 = x] \approx \textstyle\sum_{i\in[r]} \widehat{\lambda}_i^{t-1} \left(\psi_w(x)^\top \widehat{v}_i\right) \left(\widehat{u}_i^* (\widehat{\Psi}_w^\top)^\dagger \widehat{F}'\right) \in \mathbb{R}^\ell. \tag{17}$$

5 Experiments

In this experimental section we show that DPNets (i) learn reliable representations of non-normal dynamical systems, (ii) can be used in its continuous-dynamics form (14) to approximate the eigenvalues of a Langevin generator, and (iii) outperforms kernel methods and auto-encoders in large-scale and/or structured (e.g. graphs and images) data settings. To have a fair comparison, every neural network model in these experiments has been trained on the same data splits, batch sizes, number of epochs, architectures and seeds. The learning rate, however, has been optimized for each one separately. We defer every technical detail, as well as additional results to App. F. The code to reproduce the examples can be found at `https://pietronvll.github.io/DPNets/`, and it heavily depends on Kooplearn `https://kooplearn.readthedocs.io/`.

Baselines We compared our methods DPNets ($\mathcal{P}^\gamma$) and DPNets-relaxed ($\mathcal{S}^\gamma$), where appropriate, with the following baselines: Dynamic AE (DAE) of (Lusch et al., 2018), Consistent AE of (Azencot et al., 2020), DMD of (Schmid, 2010), KernelDMD and ExtendedDMD of (Williams et al., 2015a;b), and VAMPNets of (Mardt et al., 2018).

Logistic Map We study the dynamical system $X_{t+1} = (4X_t(1 - X_t) + \xi_t) \bmod 1$, for $\mathcal{X} = [0, 1)$ and ξ_t being i.i.d. trigonometric noise (Ostruszka et al., 2000) . The associated transfer operator $\mathcal{T}$ is *non-normal*, making the learning of its spectral decomposition particularly challenging (Kostic et al., 2023). Since $\mathcal{T}$ can be computed exactly, we can sidestep the problem of operator regression (see

Representation	Spectral Error	Optimality Gap
DPNets	0.28 (0.03)	**0.64** (**0.01**)
DPNets-relaxed	**0.06** (**0.05**)	1.19 (0.04)
VAMPNets	0.21 (0.09)	0.97 (0.22)
Cheby-T	0.20	1.24
NoiseKernel	0.19	2.17

Table 1: *Logistic Map*. Comparison of DPNets to relevant baselines. Mean and standard deviation are over 20 independent runs. We used $r = 7$.

Model	$\mathcal{P}$	Transition	Enthalpy ΔH
DPNets	**12.84**	**17.59 ns**	**-1.97 kcal/mol**
Nys-PCR	7.02	5.27 ns	-1.76 kcal/mol
Nys-RRR	2.22	0.89 ns	-1.44 kcal/mol
Reference	-	40 ns	-6.1 kcal/mol

Table 2: *Chignolin*. Comparison between DPNets and kernel methods with Nyström sampling (Meanti et al., 2023).

Fig. 1), and focus directly on evaluating the quality of the representation encoded by $\mathcal{H}_w$. We evaluate DPNets on two different metrics: (i) the optimality gap $\sum_{i=1}^{r} \sigma_i^2(\mathcal{T}) - \mathcal{P}^0(w)$ for $r = 3$ and (ii) the spectral error, given by $\max_i \min_j |\lambda_i(P_{\mathcal{H}}\mathcal{T}_{|\mathcal{H}}) - \lambda_j(\mathcal{T})|$. While (i) informs on how close one is to solve (4), (ii) measures how well the true eigenvalues of $\mathcal{T}$ can be recovered within the representation space $\mathcal{H}_w$. In Tab. 1 we compare the DPNets representation against VAMPNets, ExtendedDMD with a feature map of Chebyshev polynomials and the feature defining the trigonometric noise Ostruszka et al. (2000) of the process itself. Notice that in this challenging setting recovering eigenvalues via singular vectors is generally problematic, so a moderate optimality gap may lead to a larger spectral error. In Fig. 2 we report the evolution of the spectral error during training for DPNets and VAMPNets, while we defer to App. F.1 an in-depth analysis of the role of the feature dimension. Notice that DPNets and DPNets-relaxed excel in both metrics.

Continuous dynamics We investigate a one-dimensional *continuous* SDE describing the stochastic motion of a particle into the Schwantes potential (Schwantes and Pande, 2015). The invariant distribution for this process is the Boltzmann distribution $\pi(dx) \propto e^{-\beta V(x)}dx$, where $V(x)$ is the potential at state x. The non-null eigenvalues of $\mathcal{L}$ hold physical significance, as their absolute value represents the average rate at which particles cross one of the system's potential barriers (Kramers, 1940); our objective is to accurately estimate them. Here X_t is sampled *non-uniformly* according to a geometric law. In the lower panel of Fig. 2 we report the estimated transition rates, of $\mathcal{L}$ along the DPNets training loop. Notice that the embedding ψ_w progressively improves eigenvalue estimation, indicating that the invariant subspaces of the generator $\mathcal{L}$ are well captured.

Ordered MNIST Following Kostic et al. (2022), we create a stochastic dynamical system by randomly sampling images from the MNIST dataset according to the rule that X_t should be an image of the digit $t \pmod 5$ for all $t \in \mathbb{N}_0$. Given an image from the dataset with label c, a model for the transfer operator $\mathcal{T}$ of this system should then be able to produce an MNIST-alike image of the next digit in the cycle. In the upper panel of Fig. 3 we thus evaluate DPNets and a number of baselines by how accurate is an "oracle" supervised MNIST classifier (test accuracy for in-distribution $\geq 99\%$) in predicting the correct label $c + t \pmod 5$ after t steps of evolution. DPNets consistently retain an accuracy above 95%, while for every other method it degrades. The "reference" line corresponds to random guessing, while the "Oracle-Feature" baseline is an operator regression model (EDMD) using, as the dictionary of functions, the output logits of the oracle, and despite having been trained with the true labels, its performance degrades drastically after $t \geq 5$.

Fluid dynamics We study the classical problem of the transport of a passive scalar field by a 2D fluid flow past a cylinder (Raissi et al., 2020). Each data point comprises a regular 2D grid that encompasses fluid variables, including velocity, pressure, and scalar field concentration at each grid point. This system is non-stationary and is also known to exhibit non-normal dynamics (see Trefethen and Embree, 2020). We evaluate each trained baseline by feeding to it the last snapshot of the train trajectory and evaluating the relative RMSE (that is, the RMSE normalized by the data variance) between the forecasts and the subsequent test snapshots. In this experiment, we also have a *physics-informed* (PINN) baseline not related to transfer operator learning, which is however the model for which this dataset was created in the first place. Remarkably, the forecasting error of DPNets does not grow sensibly with time as it does for every other method.

The Metastable states of Chignolin In our last experiment, we study the dynamics of Chignolin, a folding protein, from a $106\,\mu$s long molecular dynamics simulation sampled every $200\,$ps, totalling over $500,000$ data points (Lindorff-Larsen et al., 2011). We focus on the leading eigenfunctions of the transfer operator, which are known (Schütte et al., 2001) to provide a simple characterization of the slowest (and usually most important) physical processes occurring along the dynamics. From the leading left eigenfunctions of $\widehat{\mathcal{T}}$, one can indeed construct the free energy surface (see Fig. 4),

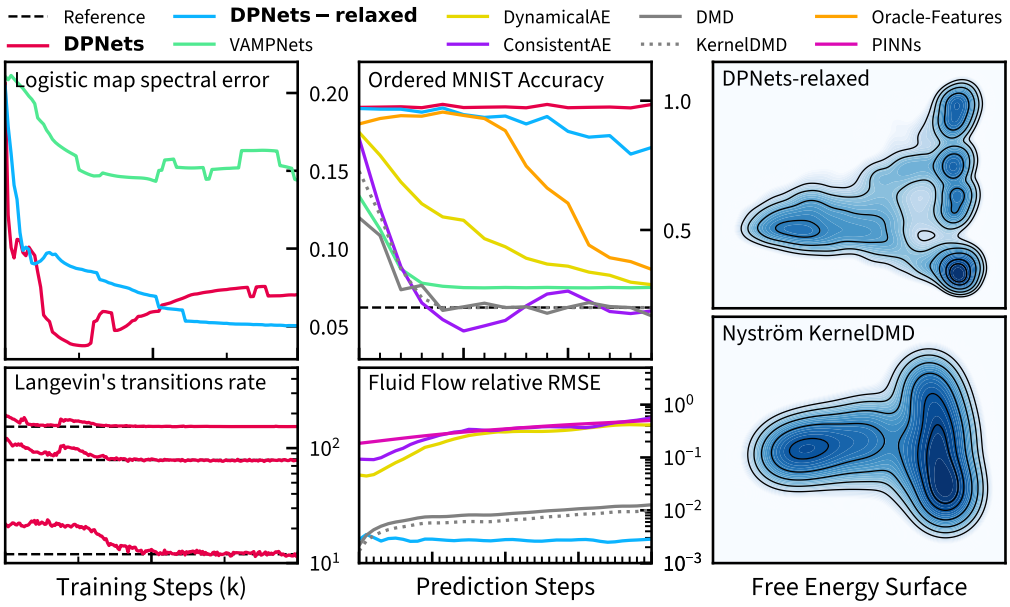

Figure 2: *Eigenvalue error decays during training*. Upper: spectral error for the logistic map. Lower: DPNets-estimated eigenvalues of $\mathcal{L}$ for the Langevin dynamics.

Figure 3: *Forecasting with DPNets*. Upper: classification accuracy over time for ordered MNIST. Lower: forecasting RMSE over time for fluid dynamic example.

Figure 4: Free energy surface of the 2 slowest modes of Chignolin,estimated by DPNets and Nystr̈om PCR. To be compared with Bonati et al. (2021).

whose local minima identify the metastable states of the system. The free energy surface, indeed, is proportional to the negative log-pdf of finding the system in a given state while at thermodynamic equilibrium, meaning that a state of *low* free energy is highly probable, hence metastable. For Chignolin, the metastable states have been thoroughly studied (see e.g. Novelli et al., 2022). The leading left eigenfunction of $\widehat{\mathcal{T}}$ encodes the folding-unfolding transition and the typical transition time is the *implied timescale* (Mardt et al., 2019) associated with its eigenvalue. The difference between the local minima of the free energy surface encodes the enthalpy ΔH of the transition. In Tab. 2 we compare these quantities to the reference values reported in Lindorff-Larsen et al. (2011). We trained a GNN-based DPNet-relaxed, as both DPNets unrelaxed and VAMPNets failed to converge, possibly due to the large scale of the data. We compared it to a KernelDMD estimator trained with the recent Nyström sketching technique (Meanti et al., 2023) as classical kernel methods are completely intractable at this scale[3]. Notice that DPNets succeed in finding additional meta-stable states of the system, which match the analysis of (Bonati et al., 2021); see App. F for more discussion.

6 Conclusions

We propose a framework for learning a representation of dynamical systems, based on orthogonal projections in data-spaces. It captures a leading invariant subspace of the transfer operator and can be applied to both discrete and continuous dynamical systems. In the discrete case, the representation is learned through the optimization of a smooth and numerically well conditioned objective function. Extensive numerical experiments demonstrate the effectiveness and generality of DPNets in various settings, suggesting that they are a promising tool for data-driven dynamical systems. A limitation of this work is that the score functional for the continuous systems might be unstable since it leverages covariance matrix inversion. Moreover, a future direction would be to study the statistical learning properties of the algorithm presented here.

[3]A back-of-the-envelope calculation shows that 450 GBs would be needed just to store kernel matrices in single precision.

REFERENCES

Alexander, R. and Giannakis, D. (2020). Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques. *Physica D: Nonlinear Phenomena*, 409:132520.

Allen, E. (2007). *Modeling with Itô stochastic differential equations*, volume 22. Springer Science & Business Media.

Andrew, A. L. and Tan, R. C. E. (1998). Computation of derivatives of repeated eigenvalues and the corresponding eigenvectors of symmetric matrix pencils. *SIAM Journal on Matrix Analysis and Applications*, 20(1):78–100.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255. PMLR.

Arnold, L. (1974). *Stochastic Differential Equations: Theory and Applications*, volume 2. John Wiley & Sons.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Azencot, O., Erichson, N. B., Lin, V., and Mahoney, M. (2020). Forecasting sequential data using consistent koopman autoencoders. In *International Conference on Machine Learning*, pages 475–485. PMLR.

Bevanda, P., Beier, M., Kerz, S., Lederer, A., Sosnowski, S., and Hirche, S. (2021). KoopmanizingFlows: Diffeomorphically Learning Stable Koopman Operators. *arXiv preprint arXiv.2112.04085*.

Bonati, L., Piccini, G., and Parrinello, M. (2021). Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118(44).

Bouvrie, J. and Hamzi, B. (2017). Kernel Methods for the Approximation of Nonlinear Systems. *SIAM Journal on Control and Optimization*, 55(4):2460–2492.

Brunton, S. L., Budišić, M., Kaiser, E., and Kutz, J. N. (2022). Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340.

Cannarsa, P., Mansutti, D., and Provenzale, A. (2020). *Mathematical Approach to Climate Change and Its Impacts: MAC2I*, volume 38. Springer.

Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L., and Ulissi, Z. (2021). Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Das, S. and Giannakis, D. (2020). Koopman spectra in reproducing kernel Hilbert spaces. *Applied and Computational Harmonic Analysis*, 49(2):573–607.

Fan, F., Yi, B., Rye, D., Shi, G., and Manchester, I. R. (2021). Learning Stable Koopman Embeddings. *arXiv preprint arXiv.2110.06509*.

Federici, M., Forré, P., Tomioka, R., and Veeling, B. S. (2024). Latent representation and simulation of markov processes via time-lagged information bottleneck. In *The Twelfth International Conference on Learning Representations*.

Fisher, M., Nocedal, J., Trémolet, Y., and Wright, S. J. (2009). Data assimilation in weather forecasting: a case study in pde-constrained optimization. *Optimization and Engineering*, 10(3):409–426.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(14):361–383.

Ghorbani, M., Prasad, S., Klauda, J. B., and Brooks, B. R. (2022). GraphVAMPNet, using graph neural networks and variational approach to markov processes for dynamical modeling of biomolecules. *The Journal of Chemical Physics*, 156(18):184103.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. JHU press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Gorini, V. and Kossakowski, A. (1976). Completely positive dynamical semigroups of N-level systems. *Journal of Mathematical Physics*, 17(5):821.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, pages 63–77. Springer.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Harold, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321.

Hou, B., Sanjari, S., Dahlin, N., Bose, S., and Vaidya, U. (2023). Sparse learning of dynamical systems in rkhs: An operator-theoretic approach. *ICML2023*.

Kato, T. (1976). *Perturbation theory for linear operators; 2nd ed.* Grundlehren der mathematischen Wissenschaften : a series of comprehensive studies in mathematics. Springer, Berlin.

Kawahara, Y. (2016). Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis. In *Advances in Neural Information Processing Systems*, volume 29.

Klus, S., Nüske, F., and Hamzi, B. (2020). Kernel-based approximation of the koopman generator and schrödinger operator. *Entropy*, 22(7):722.

Klus, S., Schuster, I., and Muandet, K. (2019). Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315.

Korda, M. and Mezić, I. (2017). On Convergence of Extended Dynamic Mode Decomposition to the Koopman operator. *Journal of Nonlinear Science*, 28(2):687–710.

Kostic, V., Lounici, K., Novelli, P., and Pontil, M. (2023). Sharp spectral rates for koopman operator learning. *Advances in Neural Information Processing Systems*, 36.

Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L., and Pontil, M. (2022). Learning dynamical systems via Koopman operator regression in reproducing kernel hilbert spaces. In *Advances in Neural Information Processing Systems*.

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2023). Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97.

Kramers, H. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L. (2016). *Dynamic Mode Decomposition*. Society for Industrial and Applied Mathematics.

Lasota, A. and Mackey, M. C. (1994). *Chaos, Fractals, and Noise*, volume 97 of *Applied Mathematical Sciences*. Springer New York.

Li, Q., Dietrich, F., Bollt, E. M., and Kevrekidis, I. G. (2017). Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2022). Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*.

Lindblad, G. (1976). On the generators of quantum dynamical semigroups. *Communications in Mathematical Physics*, 48(2):119–130.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science*, 334(6055):517–520.

Lusch, B., Kutz, J. N., and Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1).

Mardt, A., Pasquali, L., Noé, F., and Wu, H. (2019). Deep learning markov and koopman models with physical constraints. In *Mathematical and Scientific Machine Learning*.

Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). Vampnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5.

McCarty, J. and Parrinello, M. (2017). A variational conformational dynamics approach to the selection of collective variables in metadynamics. *The Journal of Chemical Physics*, 147(20):204109.

Meanti, G., Chatalic, A., Kostic, V., Novelli, P., Pontil, M., and Rosasco, L. (2023). Estimating koopman operators with sketching to provably learn large scale dynamical systems. *Advances in Neural Information Processing Systems*, 36.

Minsker, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119.

Morton, J., Jameson, A., Kochenderfer, M. J., and Witherden, F. (2018). Deep dynamical modeling and control of unsteady fluid flows. *Advances in Neural Information Processing Systems*, 31.

Novelli, P., Bonati, L., Pontil, M., and Parrinello, M. (2022). Characterizing metastable states with the help of machine learning. *Journal of Chemical Theory and Computation*, 18(9):5195–5202.

Nüske, F., Peitz, S., Philipp, F., Schaller, M., and Worthmann, K. (2023). Finite-data error bounds for koopman-based prediction and control. *Journal of Nonlinear Science*, 33(1):14.

Ostruszka, A., Pakoński, P., Słomczyński, W., and Życzkowski, K. (2000). Dynamical entropy for systems with stochastic perturbation. *Physical Review E*, 62(2):2018–2029.

Otto, S. E. and Rowley, C. W. (2019). Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593.

Pascucci, A. (2011). *PDE and Martingale Methods in Option Pricing*. Springer Milan.

Proctor, J. L., Brunton, S. L., and Kutz, J. N. (2016). Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161.

Raissi, M., Yazdani, A., and Karniadakis, G. E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030.

Ross, S. M. (1995). *Stochastic Processes*. John Wiley & Sons.

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28.

Schütt, K. T., Hessmann, S. S. P., Gebauer, N. W. A., Lederer, J., and Gastegger, M. (2023). SchNetPack 2.0: A neural network toolbox for atomistic machine learning. *The Journal of Chemical Physics*, 158(14):144801.

Schütt, K. T., Kessel, P., Gastegger, M., Nicoli, K. A., Tkatchenko, A., and Müller, K.-R. (2019). SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *Journal of Chemical Theory and Computation*, 15(1):448–455.

Schütte, C., Huisinga, W., and Deuflhard, P. (2001). Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems. In *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pages 191–223. Springer Berlin Heidelberg.

Schwantes, C. R. and Pande, V. S. (2015). Modeling Molecular Kinetics with tICA and the Kernel Trick. *Journal of Chemical Theory and Computation*, 11(2):600–608.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer New York.

Tian, W. and Wu, H. (2021). Kernel embedding based variational approach for low-dimensional approximation of dynamical systems. *Computational Methods in Applied Mathematics*, 21(3):635–659.

Trefethen, L. N. and Embree, M. (2020). *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. Technical report.

Tuckerman, M. (2010). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford university press.

Ullah, E. and Arora, R. (2023). Generalization bounds for kernel canonical correlation analysis. *Transactions on Machine Learning Research*.

Williams, M. O., Kevrekidis, I. G., and Rowley, C. W. (2015a). A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346.

Williams, M. O., Rowley, C. W., and Kevrekidis, I. G. (2015b). A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265.

Wu, H. and Noé, F. (2019). Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science*, 30(1):23–66.

Yeung, E., Kundu, S., and Hodas, N. (2019). Learning deep neural network representations for koopman operators of nonlinear dynamical systems. In *2019 American Control Conference (ACC)*. IEEE.

Appendix

The appendix is organized as follows.

- App. A reviews the notation used throughout the paper and basic notions of stochastic processes and reproducing kernel Hilbert spaces.
- In App. B we present the theoretical background for material presented in Sec. 2.
- In App. C we prove the first main result of the paper, Thm. 1.
- App. D adds value to Sec. 4. We first give theoretical justification for using the empirical score instead of the population one, and then provide details on how to address downstream tasks.
- In App. E we prove the second main result of the paper, Thm. 2 about DPNets for continuous time stochastic processes.
- Finally, App. F contains more information about the experiments and additional results.

A Background and notation

A.1 Stochastic processes tools

Let $\mathcal{X}$ be a state space and $\Sigma_{\mathcal{X}}$ be the Borel σ-algebra on $\mathcal{X}$. Let $\mathcal{M}$ denote the space of all finite measures on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. In what follows, we consider a random process $(X_t)_{t\in\mathbb{T}}$ with values in a state space $\mathcal{X}$.

Definition 1 (Markov process.)**.** *Let $\mathcal{F}_t := \Sigma(\{X_s : 0 \leq s \leq t\})$ be the sigma-algebra generated by the elements of the process up to time t. The random process $\mathbf{X} = (X_t)_{t\in\mathbb{T}}$ is called a Markov process if for all $t, \tau \geq 0$ and $B \in \Sigma_{\mathcal{X}}$ it holds*

$$\mathbb{P}(X_{t+\tau} \in B|\mathcal{F}_t) = \mathbb{P}(X_{t+\tau} \in B|X_t), \tag{18}$$

that is, conditioning on the full history of the process $\mathcal{F}_t$ up to time t is equivalent to condition only on the state of the process at time t. For this reason, Markov process are sometimes called "memoryless".

Definition 2. *The transition density function $p_\tau : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ of a time-homogeneous process $\mathbf{X}$ is defined by*

$$\mathbb{P}(X_{t+\tau} \in B|X_t = x) = \int_B p_\tau(x, y)dy,$$

for every measurable set $B \in \Sigma_{\mathcal{X}}$.

The distribution of a time-homogeneous stochastic process $(X_t)_t$ with transition density functions $(p_\tau)_{\tau>0}$ can be described by the semigroup of transfer operators $(\mathcal{T}_\tau)_{\tau\geq 0}$ usually defined on $L^\infty(\mathcal{X})$.

Definition 3 (Transfer operator)**.** *For any $\tau \geq 0$, the Koopman transfer operator $\mathcal{T}_\tau : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ is defined by*

$$\mathcal{T}_\tau(f) = \mathbb{E}[f(X_{t+\tau})|X_t = \cdot], \quad \forall f \in L^\infty(\mathcal{X}).$$

Definition 4 (Feller process)**.** *Let $(\mathcal{T}_\tau)_{\tau\geq 0}$ be the semigroup of transfer operators of a homogeneous Markov process. Then, $(\mathcal{T}_\tau)_{\tau\geq 0}$ is said to be a* Feller *semigroup when the following condition holds:*

1. *$\mathcal{T}_\tau(\mathcal{C}_0(\mathcal{X})) \subseteq \mathcal{C}_0(\mathcal{X})$ for all $\tau \geq 0$.*
2. *$\lim_{\tau\to 0} \|\mathcal{T}_\tau f - f\| = 0$ for all $f \in \mathcal{C}_0(\mathcal{X})$.*

Here $\mathcal{C}_0(\mathcal{X})$ is the Banach space of all continuous functions vanishing at infinity.

Definition 5 (Infinitesimal generator of a semigroup)**.** *Let $(\mathcal{T}_\tau)_{\tau\geq 0}$ be a Feller semigroup. We define its infinitesimal generator $\mathcal{L} : \mathcal{C}_0(\mathcal{X}) \to \mathcal{C}_0(\mathcal{X})$*

$$\mathcal{L}f = \lim_{\tau\to 0} \frac{1}{\tau}(\mathcal{T}_\tau f - f),$$

for any $f \in \mathcal{C}_0(\mathcal{X})$ such that the above limit is well-defined.

The above definitions can be lifted to $L^2_\pi(\mathcal{X})$ if the dynamical system has an invariant distribution, or if we are interested in the action of $\mathcal{T}_\tau$ on a specific couple of states X, X' of the process, as discussed in the main text.

A.2 Reproducing kernel Hilbert spaces

In this section we review few basic concepts on kernel based approaches to learning transfer operators; for more details on reproducing kernel Hilbert spaces (RKHS) we refer reader to (Aronszajn, 1950).

Let $k_X \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_{X'} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be two bounded kernels, and let $\mathcal{H}$ and $\mathcal{H}'$ be their respective reproducing kernel Hilbert spaces (RKHS). Denote the canonical feature maps by $\psi(x) := k_X(x, \cdot)$, $x \in \mathcal{X}$, and $\psi'(x') := k_{X'}(x', \cdot)$, $x' \in \mathcal{X}$.

Next, consider two probability measures μ and μ' on $\mathcal{X}$, and their associated L^2 spaces $L^2_\mu(\mathcal{X})$ and $L^2_{\mu'}(\mathcal{X})$. If $k_X(\cdot,\cdot)$ is square-integrable w.r.t. measure μ and $k_{X'}(\cdot,\cdot)$ is square-integrable w.r.t. the measure μ', then we can define injection (or the evaluation) operators $S_\mu \colon \mathcal{H} \hookrightarrow L^2_\mu(\mathcal{X})$ and $S_{\mu'} \colon \mathcal{H}' \hookrightarrow L^2_{\mu'}(\mathcal{X})$. Then, their adjoints $S^*_\mu \colon L^2_\mu(\mathcal{X}) \to \mathcal{H}$ and $S^*_{\mu'} \colon L^2_{\mu'}(\mathcal{X}) \to \mathcal{H}'$ are given by

$$S^*_\mu f = \int_{\mathcal{X}} f(x)\psi(x)\mu(dx) \in \mathcal{H} \quad \text{and} \quad S^*_{\mu'} f' = \int_{\mathcal{X}} f'(x')\psi'(x')\mu'(dx') \in \mathcal{H}',$$

where $f \in L^2_\mu(\mathcal{X})$ and $f' \in L^2_{\mu'}(\mathcal{X})$.

Using injections and their adjoints one can introduce covariance operators $C_X \colon \mathcal{H} \to \mathcal{H}$ and $C_{X'} \colon \mathcal{H}' \to \mathcal{H}'$ by

$$C_X := S^*_\mu S_\mu = \mathbb{E}_{X\sim\mu}\psi(X) \otimes \psi(X), \quad \text{and } C_{X'} := S^*_{\mu'} S_{\mu'} = \mathbb{E}_{X'\sim\mu'}\psi(X') \otimes \psi(X'),$$

as well as the cross-covariance $C_{XX'} \colon \mathcal{H}' \to \mathcal{H}$ operator

$$C_{XX'} := S^*_\mu \mathcal{T} S_{\mu'} = \mathbb{E}_{(X,X')\sim\rho}\psi(X) \otimes \psi'(X'),$$

where ρ is the joint measure of (X, X') and $\mathcal{T} : L^2_{\mu'}(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ is a transfer operator corresponding to the evolution of $X \sim \mu$ to $X' \sim \mu'$.

Different estimators for the problem of learning transfer operators from data $\mathcal{D} = (x_i, x'_i)_{i\in[n]}$ have been considered, see e.g. (Kostic et al., 2022; Li et al., 2022), which are all of the form $\widehat{T}_W = \widehat{S}'^* W \widehat{S}$, where $W \in \mathbb{R}^{n\times n}$ and $\widehat{S} \colon \mathcal{H} \to \mathbb{R}^n$ and $\widehat{S}' \colon \mathcal{H}' \to \mathbb{R}^n$ are sampling operators given by $\widehat{S} f = n^{-\frac{1}{2}}[f(x_1) \ldots f(x_n)]^\top$, $f \in \mathcal{H}$, and $\widehat{S}' f' = n^{-\frac{1}{2}}[f'(x'_1) \ldots f'(x'_n)]^\top$, $f' \in \mathcal{H}'$. In particular, kernel methods usually consider *universal kernels* that generate infinite-dimensional RKHS spaces that are dense in L^2. In such a way, one can approximate transfer operators via operator (vector-valued) regression arbitrarily well with enough data. However, one important aspect has been recently reported in (Kostic et al., 2022; 2023). Namely, from the application perspective, utility of transfer operators is largely relying on the ability to estimate well their eigenvalues and eigenfunctions, which lead to the notion of modal decomposition of observables (Brunton et al., 2022). In that context, the difference in the norms between RKHS and L^2 spaces can produce spectral pollution and lead to bad estimation of the eigenvalues and eigenfunctions. We recall that the difference in norms is due to the covariance, i.e. for every $h \in \mathcal{H}$ its L2 norm is

$$\int_{\mathcal{X}} |h(x)|^2 \mu(dx) = \langle S_\mu h, S_\mu h\rangle_{L^2_\mu(\mathcal{X})} = \langle h, C_X h\rangle_{\mathcal{H}}. \tag{19}$$

While this difference in norms is unavoidable for universal kernels (indeed for universal kernels generating infinite-dimensional RKHS spaces, since covariance is a trace class operator, we have $\lambda_j(C_X) \to 0$ as $j \to \infty$ potentially leading to $\langle h, C_X h\rangle_{\mathcal{H}} \ll \langle h, h\rangle_{\mathcal{H}}$ for some $h \in \mathcal{H}$), for finite dimensional kernels one can hope to avoid metric distortion between spaces by aiming to find features for which $C_X = I$.

B Representation Learning: problem and approach

In this section we prove some of the statements made in Sec. 2. In particular we show that (3) holds true, and show how the objectives (4) and (5) are related to the desired space $\mathcal{H}$.

We first note that $\mathcal{T}_{|\mathcal{H}} := \mathcal{T} S_\pi \colon \mathcal{H} \to L^2_\pi(\mathcal{X})$, while $\mathcal{T} P_{\mathcal{H}} = \mathcal{T} S_\pi C_X^\dagger S^*_\pi$. Hence, $\left\|[I - P_{\mathcal{H}}]\mathcal{T} P_{\mathcal{H}}\right\|^2 = \left\|[I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}} C_X^{\dagger/2}\right\|^2 = \left\|C_X^{\dagger/2}([I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}})^*\right\|^2$. Therefore, using that $\mathrm{Im}\left(([I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}})^*\right) = \mathrm{Im}(S^*_\pi \mathcal{T}^*[I - P_{\mathcal{H}}]) \subseteq \mathrm{Im}(C_X^\dagger)$ (3) follows.

Next, we show how Eckart-Young's theorem justifies (4), which is the basis of Thm. 1.

Lemma 1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces, and $A\colon \mathcal{H}_1 \to \mathcal{H}_2$ be a compact operator. Let $r \in \mathbb{N}$ and $\mathcal{P}_k^r := \{P\colon \mathcal{H}_k \to \mathcal{H}_k \,|\, P^* = P^2 = P, \mathrm{rank}(P) \leq r\}$ denote the set of rank-r orthogonal projectors in $\mathcal{H}_k$, $k = 1, 2$. It holds that*

(i) $\max\limits_{(P_1,P_2)\in\mathcal{P}_1^r\times\mathcal{P}_2^r} \|P_2AP_1\|_{\mathrm{HS}}^2 = \sum\limits_{i\in[r]} \sigma_i^2(A)$,

(ii) Let $U_r\Sigma_rV_r^*$ be the r-truncated SVD of A, then $(V_rV_r^*, U_rU_r^*) \in \operatorname*{arg\,max}\limits_{(P_1,P_2)\in\mathcal{P}_1^r\times\mathcal{P}_2^r} \|P_2AP_1\|_{\mathrm{HS}}^2$.

(iii) If $\sigma_{r+1}(A) < \sigma_r(A)$ and $\|A\|_{\mathrm{HS}} < \infty$, then $\operatorname*{arg\,max}\limits_{(P_1,P_2)\in\mathcal{P}_1^r\times\mathcal{P}_2^r} \|P_2AP_1\|_{\mathrm{HS}}^2$ is singleton.

Proof. This lemma is a direct consequence of the Eckart-Young theorem for compact operators (Kato, 1976). Recall, since A is compact, there exists its SVD $A = U\Sigma V^*$ and Eckart-Young theorem states that if A is Hilbert-Schmidt, for every $r \in \mathbb{N}$ and every $B\colon \mathcal{H}_1 \to \mathcal{H}_2$ such that $\mathrm{rank}(B) \leq r$, it holds $\|A - B\|_{\mathrm{HS}} \geq \|A - A_r\|_{\mathrm{HS}}$, where $A_r := U_r\Sigma_rV_r^*$ denotes the r-truncated SVD of A. Moreover, if $\sigma_{r+1}(A) < \sigma_r(A)$, then equality implies $B = A_r$.

Hence, for arbitrary $P_k \in \mathcal{P}_k^r$, $k = 1, 2$, using that the norm of a projection is 1 and Pythagoras theorem, for every $m \geq r$ we have that

$$\begin{aligned}\|P_2A_mP_1\|_{\mathrm{HS}}^2 &\leq \|P_2A_m\|_{\mathrm{HS}}^2\|P_1\|^2 = \|P_2A_m\|_{\mathrm{HS}}^2 = \|A_m\|_{\mathrm{HS}}^2 - \|[I - P_2]A_m\|_{\mathrm{HS}}^2 \\ &= \|A_m\|_{\mathrm{HS}}^2 - \|A_m - P_2A_m\|_{\mathrm{HS}}^2 \leq \|A_m\|_{\mathrm{HS}}^2 - \|A_m - A_r\|_{\mathrm{HS}}^2 = \|A_r\|_{\mathrm{HS}}^2,\end{aligned}$$

and, similarly, $\|P_2A_mP_1\|_{\mathrm{HS}}^2 \leq \|A_m\|_{\mathrm{HS}}^2 - \|A_m - A_mP_1\|_{\mathrm{HS}}^2 \leq \|A_r\|_{\mathrm{HS}}^2$. However, since

$$\begin{aligned}\big|\|P_2AP_1\|_{\mathrm{HS}} - \|P_2A_mP_1\|_{\mathrm{HS}}\big| &\leq \|P_2(A - A_m)P_1\|_{\mathrm{HS}} \leq \sqrt{r}\|P_2(A - A_m)P_1\| \\ &\leq \sqrt{r}\|A - A_m\| \leq \sqrt{r}\sigma_{m+1}(A),\end{aligned}$$

we obtain $\|P_2AP_1\|_{\mathrm{HS}}^2 \leq \|A_r\|_{\mathrm{HS}}^2 + \sqrt{r}\sigma_{m+1}(A)$ for all $m \geq r$. Thus, since $\|P_2AP_1\|_{\mathrm{HS}}^2 = \|A_r\|_{\mathrm{HS}}^2$ obviously holds for $P_1 = V_rV_r^*$ and $P_2 = U_rU_r^*$, letting $m \to \infty$ we obtain (i) and (ii).

Finally, assume that A is an Hilbert-Schmidt operator such that $0 \leq \sigma_{r+1}(A) < \sigma_r(A)$. Then, similarly to the above now working with A instead of A_m,

$$\|P_2AP_1\|_{\mathrm{HS}}^2 \leq \|A\|_{\mathrm{HS}}^2 - \max\{\|A - AP_1\|_{\mathrm{HS}}^2, \|A - P_2A\|_{\mathrm{HS}}^2\} \leq \|A_r\|_{\mathrm{HS}}^2.$$

So, if

$$(P_1, P_2) \in \operatorname*{arg\,max}_{(P_1',P_2')\in\mathcal{P}_1^r\times\mathcal{P}_2^r} \|P_2'AP_1'\|_{\mathrm{HS}}^2,$$

then

$$\max\{\|A - AP_1\|_{\mathrm{HS}}^2, \|A - P_2A\|_{\mathrm{HS}}^2\} = \|A - A_r\|_{\mathrm{HS}}^2.$$

Now, the uniqueness result of the Eckart-Young theorem implies that $AP_1 = P_2A = A_r$, and, consequently $\mathrm{rank}(P_1) = \mathrm{rank}(P_2) = r$, $A_r^\dagger AP_1 = A_r^\dagger A_r$ and $P_2AA_r^\dagger = A_rA_r^\dagger$. Therefore, $V_rV_r^*P_1 = V_rV_r^*$ and $P_2U_rU_r^* = U_rU_r^*$, imply $\mathrm{Im}(V_r) \subseteq \mathrm{Im}(P_1)$ and $\mathrm{Im}(U_r) \subseteq \mathrm{Im}(P_2)$. But since P_1 and P_2 have exactly rank r, we obtain $P_1 = V_rV_r^*$ and $P_2 = U_rU_r^*$. □

We also discuss how to extend part of the analysis to the important setting of stationary deterministic systems for which the transfer operator is not compact.

Remark 1. *The compactness assumption on the the transfer operator, does not hold for purely deterministic dynamical systems. However, our approach is still applicable to the study of deterministic systems, that is when $X_{t+1} = F(X_t)$, for a deterministic map $F\colon \mathcal{X} \to \mathcal{X}$, and $X_t \sim \pi$ for all $t \in \mathbb{T}$, where π is the invariant measure defined on the attractor. Then, we have that $\mathbb{E}[f(X_{t+1}) \,|\, X_t] = F \circ f$ and $\mathcal{T}\colon L_\pi^2(\mathcal{X}) \to L_\pi^2(\mathcal{X})$ is unitary, see e.g. (Brunton et al., 2022). Thus, $\mathcal{T}$ is not a compact operator, however it is a normal one. But then, Pythagoras theorem gives that*

$$\inf_{w\in\mathcal{W}} \|[I - P_{\mathcal{H}_w}]\mathcal{T}P_{\mathcal{H}_w}]\|_{\mathrm{HS}}^2 = \inf_{w\in\mathcal{W}} \left(\|\mathcal{T}P_{\mathcal{H}_w}\|_{\mathrm{HS}}^2 - \|P_{\mathcal{H}_w}\mathcal{T}P_{\mathcal{H}_w}\|_{\mathrm{HS}}^2\right) = r - \sup_{w\in\mathcal{W}} \mathcal{P}(w),$$

where the second equality holds since the HS-norm is unitarily invariant. Therefore, when w is such that $\mathcal{T}\psi_{w,j} = \lambda_j \psi_{w,j}$, since $\mathcal{T}$ is unitary we have $\mathcal{P}(w) = \sum_{j=1}^{r} |\lambda_j|^2 = r$ and the above inf is zero. Consequently, we have identified an r-dimensional invariant subspace of $\mathcal{T}$.

Finally, we conclude this section with the lemma that is the basis of generator learning via optimization problem (5) whose objective is defined through a partial trace.

Lemma 2. *Let $\mathcal{H}$ be a separable Hilbert space and let $\mathcal{H}_r \subseteq \mathcal{H}$ be a (finite) r-dimensional subspace. If $A\colon \mathcal{H} \to \mathcal{H}$ is a self-adjoint operator having at least $r+1$ eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_r(A) \geq \lambda_{r+1}(A)$ above its essential spectrum, then*

$$\mathrm{tr}(AP_{\mathcal{H}_r}) \leq \lambda_1(A) + \lambda_2(A) + \cdots + \lambda_r(A),$$

and equality holds when $\mathcal{H}_r$ is spanned by eigenfunctions of A corresponding to eigenvalues $\lambda_1(A), \ldots, \lambda_r(A)$.

Proof. First, let $(v_i)_{i\in[r]}$ denote eigenvectors of A corresponding to eigenvalues λ_i, that is $Av_i = \lambda_i(A)v_i$, and let $(h_i)_{i\in\mathbb{N}}$ be an ortho-normal basis of $\mathcal{H}$ such that $(h_i)_{i\in[r]}$ is the ortho-normal basis of $\mathcal{H}_r$. Then, clearly

$$\mathrm{tr}(AP_{\mathcal{H}_r}) = \mathrm{tr}(P_{\mathcal{H}_r} A P_{\mathcal{H}_r}) = \sum_{i\in\mathbb{N}} \langle h_i, (P_{\mathcal{H}_r} A P_{\mathcal{H}_r}) h_i \rangle = \sum_{i\in[r]} \langle h_i, Ah_i \rangle, \tag{20}$$

which is clearly equal to $\lambda_1(A) + \lambda_2(A) + \ldots + \lambda_r(A)$ whenever $h_i = v_i$, $i \in [r]$.

The upper bound we prove by induction. First, for $r = 1$ we have that $\langle h_i, Ah_i \rangle \leq \lambda_1(A)$ follows directly from the Courant-Fischer max-min theorem for operators which claims that

$$\lambda_i(A) = \max_{h_1,\ldots,h_i} \min\{\langle h, Ah \rangle \mid h \in \mathrm{span}(h_j)_{j\in[i]}, \|h\| = 1\},\ i \in [r+1]$$

Next, assuming that (20) holds for arbitrary A and $r \leq m-1$, we will prove that it holds for $r = m$.

Start by observing that there exists $\mathcal{H}' \subseteq \mathcal{H}_m$ such that $\dim(\mathcal{H}') \leq m-1$ and $\mathcal{H}' \perp v_1$. Indeed, taking $g_i = \sum_{j\in[m]} b_{ij} h_j$, $i \in [m-1]$, for some $B = [b_{ij}] \in \mathbb{R}^{(m-1)\times m}$, we have that $g_i \perp v_1$ for all $i \in [m-1]$ iff $B\beta = 0$, where the vector $\beta \in \mathbb{R}^m$ is given by $\beta_j = \langle h_j, v_1 \rangle$. Now, if $\beta = 0$, then $\mathcal{H}_m \perp v_1$ and the claim trivially follows. Otherwise, since $\dim(\mathrm{span}(\beta)^\perp) = m-1$, there exists a matrix $B \in \mathbb{R}^{(m-1)\times m}$ so that $B\beta = 0$. Consequently, the space $\mathcal{H}'$ spanned by $(g_i)_{i\in[m-1]}$ is such that $\dim(\mathcal{H}') \leq m-1$ and $\mathcal{H}' \subseteq \mathrm{span}(v_1)^\perp$. Without the loss of generality, assume that $(g_i)_{i\in[m-1]}$ are orthonormal basis of $\mathcal{H}'$ and that g_m is such that $(g_i)_{i\in[m]}$ are orthonormal basis of $\mathcal{H}$.

First, clearly $\langle g_m, Ag_m \rangle \leq \lambda_1$. On the other hand, define $A' := A + (\lambda_{r+1} - \lambda_1)\, v_1 \otimes v_1$. This is an operator obtained by deflating (moving) the first eigenvalue into the $(r+1)$-th one. Namely, we have that $\lambda_i(A') = \lambda_{i+1}(A)$ for $i \in [m-1]$. Moreover, it holds that $A'P_{\mathcal{H}'} = AP_{\mathcal{H}'}$. Hence, according to the inductive hypothesis, we have

$$\mathrm{tr}(AP_{\mathcal{H}'}) = \mathrm{tr}(A'P_{\mathcal{H}'}) \leq \sum_{i\in[m-1]} \lambda_i(A') = \sum_{i=2}^{m} \lambda_i(A),$$

and, consequently,

$$\mathrm{tr}(AP_{\mathcal{H}}) = \sum_{i\in[r]} \langle g_i, Ag_i \rangle = \mathrm{tr}(AP_{\mathcal{H}'}) + \langle g_m, Ag_m \rangle \leq \sum_{i=2}^{m} \lambda_i(A) + \lambda_1(A) = \sum_{i\in[m]} \lambda_i(A),$$

which completes the proof. □

C SCORE FUNCTIONAL

We start by extending problem (4) to the representation learning for stable non-stationary processes $(X_t)_{t\in\mathbb{N}_0}$. Since for the operator regression problem we typically collect data samples uniformly at random along trajectory, we have that $X \sim \mu := \frac{1}{n}\sum_{i\in[n]} \mu_{i-1}$, recalling that μ_t is the law

of X_t and n is the sample size. Then, the one step ahead evolution of a random variable X is $X' \sim \mu' = \frac{1}{n}\sum_{i\in[n]} \mu_i$. Since the transfer operator is properly defined as $\mathcal{T}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_{\mu}(\mathcal{X})$, we obtain the following version of (3)

$$\left\|[I - P_{\mathcal{H}}]\mathcal{T}P'_{\mathcal{H}}\right\|^2 \lambda^+_{\min}(C'_{\mathcal{H}}) \leq \left\|[I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}}\right\|^2 \leq \left\|[I - P_{\mathcal{H}}]\mathcal{T}P'_{\mathcal{H}}\right\|^2 \lambda_{\max}(C'_{\mathcal{H}}), \tag{21}$$

where, $P_{\mathcal{H}}\colon L^2_{\mu}(\mathcal{X}) \to L^2_{\mu}(\mathcal{X})$ and $P'_{\mathcal{H}}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_{\mu'}(\mathcal{X})$ are the orthogonal projections onto $\mathcal{H}$ as a subspace of $L^2_{\mu}(\mathcal{X})$ and $L^2_{\mu'}(\mathcal{X})$, respectively, and covariance $C'_{\mathcal{H}}$ is now w.r.t. measure μ'. Now, assuming that the feature map of $\mathcal{H}$ is bounded by some constant $c_{\mathcal{H}}$, we have that

$$\left\|C'_{\mathcal{H}} - C_{\mathcal{H}}\right\| \leq \int_{\mathcal{X}} \left\|\psi(x) \otimes \psi(x)\right\| |\mu' - \mu|(dx) \leq 2c^2_{\mathcal{H}} \left\|\mu' - \mu\right\|_{TV} = \frac{2c^2_{\mathcal{H}}}{n} \left\|\mu_n - \mu_0\right\|_{TV}.$$

Thus, if the dynamics is stable, the total variation norm $\left\|\mu_n - \mu_0\right\|_{TV}$ is bounded w.r.t. $n \to \infty$, and we conclude that for large enough sample size n one can replace $C'_{\mathcal{H}}$ by $C_{\mathcal{H}}$ in (21) to obtain tight approximation of the approximation error $\left\|[I - P_{\mathcal{H}}]\mathcal{T}_{|\mathcal{H}}\right\|^2$ by controlling $\left\|[I - P_{\mathcal{H}}]\mathcal{T}P'_{\mathcal{H}}\right\|$. This motivates the optimization problem for non-stationary dynamics

$$\max_{\mathcal{H}\subset L^2_{\mu}(\mathcal{X}), \mathcal{H}'\subset L^2_{\mu'}(\mathcal{X})} \left\{ \left\|P_{\mathcal{H}}\mathcal{T}P'_{\mathcal{H}'}\right\|^2_{\mathrm{HS}} \mid C_{\mathcal{H}} = C'_{\mathcal{H}'} = I, \dim(\mathcal{H}) \leq r,\ \dim(\mathcal{H}') \leq r \right\}. \tag{22}$$

Next, we prove (7), relating the projection score with the covariance operators.

Lemma 3. *Let X and X' be two $\mathcal{X}$-valued random variables distributed according to probability measures μ and μ', respectively. Given $w \in \mathcal{W}$ let for every $j \in [r]$ functions $\psi_{w,j}\colon \mathcal{X} \to \mathbb{R}$ and $\psi'_{w,j}\colon \mathcal{X} \to \mathbb{R}$ be square integrable w.r.t measures μ and μ', respectively. If $P_{\mathcal{H}_w}\colon L^2_{\mu}(\mathcal{X}) \to L^2_{\mu}(\mathcal{X})$ and $P'_{\mathcal{H}'_w}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_{\mu'}(\mathcal{X})$ are orthogonal projections onto subspaces $\mathcal{H}_w := \mathrm{span}(\psi_{w,j})_{j\in[r]}$ and $\mathcal{H}'_w := \mathrm{span}(\psi'_{w,j})_{j\in[r]}$, respectively, then*

$$\left\|P_{\mathcal{H}_w}\mathcal{T}P'_{\mathcal{H}'_w}\right\|^2_{\mathrm{HS}} = \left\|(C^w_X)^{\dagger/2} C^w_{XX'} (C^w_{X'})^{\dagger/2}\right\|^2_{\mathrm{HS}}. \tag{23}$$

Proof. The proof directly follows from the notion of finite-dimensional RKHS. Let $k^w_X\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $k^w_{X'}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be two kernels given by

$$k^w_X(x,y) := \psi_w(x)^\top \psi_w(y) \quad \text{and} \quad k^w_{X'}(x,y) := \psi'_w(x)^\top \psi'_w(y),\ x,y \in \mathcal{X}.$$

Then $\mathcal{H}_w$ and $\mathcal{H}'_w$ are the reproducing kernel Hilbert spaces (RKHS) associated with kernels k^w_X and $k^w_{X'}$, respectively.

Now, due to square integrability of the embeddings $\psi_{w,j}$ and $\psi'_{w,j}$, $j \in [r]$, we have that the injection operators of the two RKHS spaces into their respective L^2 spaces are well-defined: $S_\mu\colon \mathcal{H}_w \hookrightarrow L^2_{\mu}(\mathcal{X})$ and $S_{\mu'}\colon \mathcal{H}'_w \hookrightarrow L^2_{\mu'}(\mathcal{X})$. Moreover, observing that $\mathcal{H}_w$ and $\mathcal{H}'_w$ are isometrically isomorphic to $\mathbb{R}^r$, we have that $S^*_\mu S_\mu\colon \mathcal{H}_w \to \mathcal{H}_w$ and $S^*_{\mu'} S_{\mu'}\colon \mathcal{H}'_w \to \mathcal{H}'_w$ can be identified with $C^w_X \in \mathbb{R}^{r\times r}$ and $C^w_{X'} \in \mathbb{R}^{r\times r}$, respectively, that is

$$S^*_\mu S_\mu = Q C^w_X Q^* \quad \text{and} \quad S^*_{\mu'} S_{\mu'} = Q' C^w_{X'} (Q')^*.$$

where $Q\colon \mathbb{R}^r \to \mathcal{H}_w$ and $Q'\colon \mathbb{R}^r \to \mathcal{H}'_w$ are partial isometries, meaning that QQ^* and $Q'(Q')^*$ are identity operators on $\mathcal{H}_w$ and Q^*Q and $(Q')^*Q'$ are identity matrices in $\mathbb{R}^{r\times r}$.

As a consequence, the polar decompositions of finite rank operators S_μ and $S_{\mu'}$ can be written as

$$S_\mu = UQ(C^w_X)^{1/2}Q^* \quad \text{and} \quad S_{\mu'} = U'Q'(C^w_{X'})^{1/2}(Q')^*, \tag{24}$$

where $U\colon \mathcal{H}_w \to L^2_{\mu}(\mathcal{X})$ and $U'\colon \mathcal{H}'_w \to L^2_{\mu'}(\mathcal{X})$ are partial isometries.

But then, since $\mathcal{H}_w$ as a subspace of $L^2_{\mu}(\mathcal{X})$ is identified as $\mathrm{Im}(S_\mu)$, and $\mathcal{H}'_w$ as a subspace of $L^2_{\mu'}(\mathcal{X})$ is identified as $\mathrm{Im}(S_{\mu'})$, using adjoints $S^*_\mu\colon L^2_{\mu}(\mathcal{X}) \to \mathcal{H}_w$ and $S^*_{\mu'}\colon L^2_{\mu'}(\mathcal{X}) \to \mathcal{H}'_w$, we can write the aforementioned orthogonal projections as

$$P_{\mathcal{H}_w} = S_\mu (S^*_\mu S_\mu)^\dagger S^*_\mu \quad \text{and} \quad P_{\mathcal{H}'_w} = S_{\mu'} (S^*_{\mu'} S_{\mu'})^\dagger S^*_{\mu'}, \tag{25}$$

which, using (24) and the fact that $(C_X^w)^{1/2}(C_X^w)^\dagger = (C_X^w)^{\dagger/2}$, is equivalent to

$$P_{\mathcal{H}_w} = UQ(C_X^w)^{\dagger/2}Q^*S_\mu^* \quad \text{and} \quad P_{\mathcal{H}'_w} = U'Q'(C_{X'}^w)^{\dagger/2}(Q')^*S_{\mu'}^*.$$

Finally, since for every $v \in \mathbb{R}^r$ we obtain

$$\begin{aligned}(Q^*S_\mu^*\mathcal{T}S_{\mu'}Q')v &= Q^*S_\mu^*(\mathcal{T}S_{\mu'}Q'v) = Q^*\int_{\mathcal{X}} \mu(dx)\phi_X^w(x)(\mathcal{T}S_{\mu'}Q'v)(x) \\ &= Q^*\int_{\mathcal{X}}\int_{\mathcal{X}} \mu(dx)p(x,dx')\phi_X^w(x)(Q'v)(x') \\ &= Q^*\int_{\mathcal{X}\times\mathcal{X}} \rho(dx,dx')\phi_X^w(x)(Q'v)(x')\end{aligned}$$

where ρ is joint measure of (X, X') and ϕ_X^w is the canonical feature map of k_X^w, that is $\phi_X^w(x) = k_X^w(\cdot, x)$, $x \in \mathcal{X}$. Next, using the reproducing property

$$\begin{aligned}(Q^*S_\mu^*\mathcal{T}S_{\mu'}Q')v &= Q^*\int_{\mathcal{X}\times\mathcal{X}} \rho(dx,dx')\phi_X^w(x)\langle\phi_{X'}^w(x'), Q'v\rangle_{\mathcal{H}'_w} \\ &= \left[\int_{\mathcal{X}\times\mathcal{X}} \rho(dx,dx')Q^*\phi_X^w(x)\otimes((Q')^*\phi_{X'}^w(x'))\right]v \\ &= \left[\int_{\mathcal{X}\times\mathcal{X}} \rho(dx,dx')\psi_w(x)\otimes\psi'_w(x')\right]v = C_{XX'}^w v,\end{aligned}$$

where $\phi_{X'}^w$ is the canonical feature map of $k_{X'}^w$ and we used that $Q^*\phi_X^w(x) = \psi_w(x)$ and $(Q')^*\phi_{X'}^w(x') = \psi'_w(x')$.

Therefore, using (25) we obtain

$$P_{\mathcal{H}_w}\mathcal{T}P'_{\mathcal{H}'_w} = UQ(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2}(Q')^*(U')^*,$$

which using that U, U' and Q, Q' are partial isometries, implies (23). □

Now, using Lem. 1, we prove one of the main theoretical results of the paper. From this point forward, we abbreviate $\mathcal{P}(w) := \mathcal{P}^0(w)$, $\mathcal{S}(w) := \mathcal{S}^0(w)$, $\widehat{\mathcal{P}}_n(w) := \mathcal{P}_n^0(w)$ and $\widehat{\mathcal{S}}_n(w) := \mathcal{P}_n^0(w)$.

Theorem 1. *If $\mathcal{T}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ is compact, $\mathcal{H}_w \subseteq L^2_\mu(\mathcal{X})$ and $\mathcal{H}'_w \subseteq L^2_{\mu'}(\mathcal{X})$, then for all $\gamma \geq 0$*

$$\mathcal{S}^\gamma(w) \leq \mathcal{P}^\gamma(w) \leq \sigma_1^2(\mathcal{T}) + \cdots + \sigma_r^2(\mathcal{T}). \tag{10}$$

Moreover, if $(\psi_{w,j})_{j\in[r]}$ and $(\psi'_{w,j})_{j\in[r]}$ are the leading r left and right singular functions of $\mathcal{T}$, respectively, then both equalities in (10) *hold. Finally, if the operator $\mathcal{T}$ is Hilbert-Schmidt, $\sigma_r(\mathcal{T}) > \sigma_{r+1}(\mathcal{T})$ and $\gamma > 0$, then the "only if" relation is satisfied up to unitary equivalence.*

Proof. Recall that $\mathcal{P}^\gamma(w) = \mathcal{P}(w) - \gamma\left(\mathcal{R}(C_X^w) + \mathcal{R}(C_{X'}^w)\right)$ with $\mathcal{P}(w) = \left\|P_{\mathcal{H}_w}\mathcal{T}P'_{\mathcal{H}'_w}\right\|_{\mathrm{HS}}^2 = \left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2}\right\|_{\mathrm{HS}}^2$, where the last equality follows from Lemma 3, and that $\mathcal{S}(w) := \left\|C_{XX'}^w\right\|_{\mathrm{HS}}^2/\left(\left\|C_X^w\right\|\left\|C_{X'}^w\right\|\right)$. The first inequality in (10)holds thanks to a standard matrix norm inequality, while the second holds by applying Lemma 1(i) and noting that $\mathcal{R}(C_X^w) + \mathcal{R}(C_{X'}^w) \geq 0$.

Now assume that $(\psi_{w,j})_{j\in[r]}$ and $(\psi'_{w,j})_{j\in[r]}$ are some leading r left and right singular functions of $\mathcal{T}$, respectively. Then, since singular functions form ortho-normal systems in L^2 spaces we have that

$$(C_X^w)_{i,j} = \langle\psi_{w,i}, \psi_{w,j}\rangle_{L^2_\mu(\mathcal{X})} = \delta_{i,j} \;\text{ and }\; (C_{X'}^w)_{i,j} = \left\langle\psi'_{w,i}, \psi'_{w,j}\right\rangle_{L^2_\mu(\mathcal{X})} = \delta_{i,j},\; i,j \in [r],$$

that is $C_X^w = C_{X'}^w = I_r$, and, therefore, $\mathcal{R}(C_X^w) = \mathcal{R}(C_{X'}^w) = 0$ and $\mathcal{S}(w) = \mathcal{P}(w)$. Thus, using Lemma 1(ii), we obtain that (10) holds with equalities in place of inequalities.

Next, assume that the operator $\mathcal{T}$ is Hilbert-Schmidt, $\sigma_r(\mathcal{T}) > \sigma_{r+1}(\mathcal{T})$, $\gamma > 0$ and $\mathcal{S}^\gamma(w) = \sigma_1^2(\mathcal{T}) + \ldots + \sigma_r^2(\mathcal{T})$. Then, clearly $\mathcal{R}(C_X^w) = \mathcal{R}(C_{X'}^w) = 0$, which implies that $C_X^w = C_{X'}^w = I_r$, that is $(\psi_{w,j})_{j\in[r]}$ and $(\psi'_{w,j})_{j\in[r]}$ form orthonormal systems. Consequently,

$$P_{\mathcal{H}_w} = \sum_{j\in[r]} \psi_{w,j}\otimes\psi_{w,j} \quad \text{and} \quad P_{\mathcal{H}'_w} = \sum_{j\in[r]} \psi'_{w,j}\otimes\psi'_{w,j},$$

and

$$\mathcal{P}(w) = \mathcal{P}^{\gamma}(w) = \mathcal{S}^{\gamma}(w) = \sigma_1^2(\mathcal{T}) + \ldots + \sigma_r^2(\mathcal{T}).$$

But then, Lemma 1(iii) implies that $P_{\mathcal{H}_w}$ and $P_{\mathcal{H}'_w}$ are orthogonal projectors on the leading r left and right singular spaces of $\mathcal{T}$. In other words, up to unitary changes of basis, $(\psi_{w,j})_{j\in[r]}$ and $(\psi'_{w,j})_{j\in[r]}$ are the leading r left and right singular functions of $\mathcal{T}$, which completes the proof. □

We conclude this section with a remark on the importance of regularization.

Remark 2. *Recalling the proof of the previous theorem, note that when $\gamma = 0$, equality in* (10) *is achieved whenever $\mathcal{H}_w$ and $\mathcal{H}'_w$ are* spanned *by leading r left and right singular functions of $\mathcal{T}$, respectively. Meaning that after the change of basis $Q, Q' \in \mathbb{R}^{r\times r}$ so that $(\sum_{i\in[r]} Q_{i,j}\psi_{w,i})_{j\in[r]}$ and $(\sum_{i\in[r]} Q'_{i,j}\psi'_{w,i})_{j\in[r]}$ are the leading left and right singular functions, respectively. Under the additional conditions, also the "only if" part holds for some changes of basis. Indeed, we can take the change of basis to be $Q = (C_X^w)^{-1/2}$ and $Q' = (C_{X'}^w)^{-1/2}$, which without regularization term, need not be unitary. This, as we see in App. D.2, highly impacts on the stability of computation of the transfer operator estimators, and, therefore, impacts their practical use.*

D METHODS

D.1 STATISTICAL LEARNING GUARANTEES

Before we study the statistical learning guarantees for our novel score $\mathcal{S}$, we first discuss a fundamental limitation of using the score $\mathcal{P}$ to learn the invariant representation. In order to maximize the score $\mathcal{P}$ one can use standard ridge regularization on the empirical covariances. This approach considered in DeepCCA (Andrew et al., 2013) and VAMPNets (Mardt et al., 2019), typically requires a large number of training samples n, and the rates are governed by the choice of the regularization hyperparameter, typically ranging from $O(n^{-1/3})$ to $O(n^{-1/2})$. Namely, in this approach one uses score $\left\|(\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\|_{\mathrm{HS}}$ instead of $\mathcal{P}$, where λ is the regularization parameter. So the main concern is now to measure how close the regularized empirical score is to the true score? To do this, we first study the operator norm deviation:

$$\begin{aligned}
&\left|\left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2}\right\| - \left\|(\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\|\right| \\
&\quad \le \left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2} - (\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\| \\
&\quad \le \left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2} - (C_X^w + \lambda I)^{-1/2}C_{XX'}^w(C_{X'}^w + \lambda I)^{-1/2}\right\| \\
&\qquad + \left\|(C_X^w + \lambda I)^{-1/2}C_{XX'}^w(C_{X'}^w + \lambda I)^{-1/2} - (\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\|.
\end{aligned}$$

Using Lemma 4 in (Ullah and Arora, 2023), we get the following control on the "bias term" of the previous display:

$$\left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2} - (C_X^w + \lambda I)^{-1/2}C_{XX'}^w(C_{X'}^w + \lambda I)^{-1/2}\right\| \le 8\sqrt{\lambda}.$$

Next, using Lemma 6 in (Fukumizu et al., 2007) gives the following asymptotic rate of convergence on the "variance" part. Assume that $\lambda = \lambda_n \to 0$ as $n \to \infty$, then

$$\left\|(C_X^w + \lambda I)^{-1/2}C_{XX'}^w(C_{X'}^w + \lambda I)^{-1/2} - (\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{\lambda^3\, n}}\right).$$

Combining the last three displays and for the optimal choice $\lambda = \lambda_n = n^{-1/4}$, we obtain

$$\left|\left\|(C_X^w)^{\dagger/2}C_{XX'}^w(C_{X'}^w)^{\dagger/2}\right\| - \left\|(\widehat{C}_X^w + \lambda I)^{-1/2}\widehat{C}_{XX'}^w(\widehat{C}_{X'}^w + \lambda I)^{-1/2}\right\|\right| = O_{\mathbb{P}}(n^{-1/3}).$$

According again to (Ullah and Arora, 2023), in the most favorable scenario of finite-dimensional spaces with well-conditioned covariance matrices, we can obtain an improved control on the "variance" part of the order of magnitude $O_{\mathbb{P}}(\lambda^{-1/2}n^{-1/2})$. Consequently, taking $\lambda = \lambda_n \asymp n^{-1/2} \to 0$ as $n \to \infty$, the estimation error improves from $O_{\mathbb{P}}(n^{-1/3})$ to $O_{\mathbb{P}}(n^{-1/2})$ in the best-case scenario. While these guarantees were obtained for the spectral norm, one can directly deduce such guarantees also for the HS-norm when the latent space is low-dimensional. This analysis highlights a fundamental limitation of using the score $\mathcal{P}(w)$. On the one hand, the Ridge regularization parameter λ

cannot be set too small to maintain the numerical stability of the method. On the other hand, the statistical analysis requires that $\lambda = \lambda_n$ converges sufficiently fast to 0 as $n \to \infty$ to guarantees that the empirical score $\widehat{\mathcal{P}}$ approximate the true objective $\mathcal{P}$. These two constraints are antagonistic and it is not clear that they can be simultaneously satisfied in practice. Hence, balancing between numerical stability and convergence to the true objective presents a challenging trade-off when using $\mathcal{P}$ for learning invariant representations. By contrast, our score $\mathcal{S}$ and its relaxed version do not suffer from this limitation. Namely, there is no need to sacrifice numerical stability for statistical accuracy or vice versa since $|\mathcal{S}(w) - \widehat{\mathcal{S}}(w)| = O_{\mathbb{P}}(n^{-1/2})$ in all situations, as we prove it in Theorem 3 below.

Our goal is to derive concentration guarantees for the empirical score $\widehat{\mathcal{S}}$ from the true score $\mathcal{S}$. We focus on time-homogeneous Markovian dynamical systems in the stationary regime with invariant measure π which was proposed in (Lasota and Mackey, 1994; Kostic et al., 2022).

Recall the definition of the true and empirical scores

$$\mathcal{S}(w) := \frac{\left\|C^w_{XX'}\right\|^2_{\mathrm{HS}}}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|} \quad \text{and} \quad \widehat{\mathcal{S}}(w) := \frac{\left\|\widehat{C}^w_{XX'}\right\|^2_{\mathrm{HS}}}{\left\|\widehat{C}^w_X\right\|\left\|\widehat{C}^w_{X'}\right\|}, \tag{26}$$

where, as before,

$$\widehat{C}^w_X := \tfrac{1}{n}\sum_{i\in[n]} \psi_w(x_i)\psi_w(x_i)^\top, \widehat{C}^w_{X'} := \tfrac{1}{n}\sum_{i\in[n]} \psi'_w(x'_i)\psi'_w(x'_i)^\top \text{ and } \widehat{C}^w_{XX'} := \tfrac{1}{n}\sum_{i\in[n]} \psi_w(x_i)\psi'_w(x'_i)^\top.$$

Denote by ρ the joint distribution of (X, X').

We assume that the embeddings are bounded almost surely, that is there exists an absolute constant c such that

$$\operatorname*{ess\,sup}_{x\sim\mu} \|\psi_w(x)\|^2 \le c, \qquad \operatorname*{ess\,sup}_{x'\sim\mu'} \|\psi'_w(x')\|^2 \le c. \tag{27}$$

For any fixed w and any fixed $\delta \in (0,1)$, we assume that n is large enough such that

$$\frac{4c}{3\left\|C^w_X\right\| n}\log\left(12r\delta^{-1}\right) + \sqrt{\frac{2}{\left\|C^w_X\right\| n}\log(12r\delta^{-1})} < \frac{1}{3}. \tag{28}$$

Define

$$\varepsilon_n(\delta) = \frac{4c}{3\left\|C^w_X\right\| n}\log\left(12r\delta^{-1}\right) + \sqrt{\frac{2}{\left\|C^w_X\right\| n}\log(12r\delta^{-1})},$$

and

$$\varepsilon''_n(\delta) := c^2\sqrt{\frac{5\log(18\delta^{-1})}{n}} + c^2\frac{C}{n},$$

where C is some absolute constant.

Theorem 3. *Let Conditions* (27) *and* (28) *be satisfied. Then we get with probability at least* $1-\delta$

$$\left|\mathcal{S}(w) - \widehat{\mathcal{S}}(w)\right| \le \mathcal{S}(w)\,\frac{3\varepsilon_n(\delta)}{1-3\varepsilon_n(\delta)} + \frac{\varepsilon''_n(\delta)}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|(1-3\varepsilon_n(\delta))}. \tag{29}$$

Proof. By definition of $\mathcal{S}(w)$ and $\widehat{\mathcal{S}}(w)$, we have

$$\begin{aligned}
\left|\widehat{\mathcal{S}}(w) - \mathcal{S}(w)\right| &\le \widehat{\mathcal{S}}(w)\frac{\left|\left\|\widehat{C}^w_X\right\|\left\|\widehat{C}^w_{X'}\right\| - \left\|C^w_X\right\|\left\|C^w_{X'}\right\|\right|}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|} + \frac{\left|\left\|\widehat{C}^w_{XX'}\right\|^2_{\mathrm{HS}} - \left\|C^w_{XX'}\right\|^2_{\mathrm{HS}}\right|}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|} \\
&\le |\widehat{\mathcal{S}}(w) - \mathcal{S}(w)|\frac{\left|\left\|\widehat{C}^w_X\right\|\left\|\widehat{C}^w_{X'}\right\| - \left\|C^w_X\right\|\left\|C^w_{X'}\right\|\right|}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|} + \mathcal{S}(w)\frac{\left|\left\|\widehat{C}^w_X\right\|\left\|\widehat{C}^w_{X'}\right\| - \left\|C^w_X\right\|\left\|C^w_{X'}\right\|\right|}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|} \\
&\quad + \frac{\left|\left\|\widehat{C}^w_{XX'}\right\|^2_{\mathrm{HS}} - \left\|C^w_{XX'}\right\|^2_{\mathrm{HS}}\right|}{\left\|C^w_X\right\|\left\|C^w_{X'}\right\|}.
\end{aligned}$$

Using (34), (36) below, we get with probability at least $1-2\delta$,

$$\frac{\left|\|\widehat{C}_X^w\|\|\widehat{C}_{X'}^w\| - \|C_X^w\|\|C_{X'}^w\|\right|}{\|C_X^w\|\|C_{X'}^w\|} \le \varepsilon_n(\delta) + \varepsilon_n'(\delta) + \varepsilon_n(\delta)\varepsilon_n'(\delta).$$

Next, (37) gives with probability at least $1-\delta$

$$\frac{\left|\|\widehat{C}_{XX'}^w\|_{\mathrm{HS}}^2 - \|C_{XX'}^w\|_{\mathrm{HS}}^2\right|}{\|C_X^w\|\|C_{X'}^w\|} \le \frac{\varepsilon_n''(\delta)}{\|C_X^w\|\|C_{X'}^w\|}.$$

Under Conditions 27 and 28, we get

$$\varepsilon_n(\delta) = \varepsilon_n'(\delta) = \frac{4c}{3\|C_X^w\|\,n}\log\left(4r\delta^{-1}\right) + \sqrt{\frac{2}{\|C_X^w\|\,n}\log(4r\delta^{-1})} < \frac{1}{3},$$

and

$$\varepsilon_n''(\delta) := c^2\sqrt{\frac{5\log(6\delta^{-1})}{n}} + c^2\frac{C}{n}.$$

An union bound gives with probability at least $1-3\delta$,

$$\left|\widehat{\mathcal{S}}(w) - \mathcal{S}(w)\right| \le \mathcal{S}(w)\,\frac{3\varepsilon_n(\delta)}{1-3\varepsilon_n(\delta)} + \frac{\varepsilon_n''(\delta)}{\|C_X^w\|\|C_{X'}^w\|(1-3\varepsilon_n(\delta))}. \tag{30}$$

Replacing δ with $\delta/3$, we get the result with probability $1-\delta$. □

To control the operator norm deviation of the empirical covariances $\widehat{C}_X^w$ and $\widehat{C}_{X'}^w$ from their population counterparts, we use the following dimension-free version of (Minsker, 2017) of the non-commutative Bernstein inequality (see also Theorem 7.3.1 in (Tropp, 2012) for an easier to read and slightly improved version) as well as an extension to self-adjoint Hilbert-Schmidt operators on separable Hilbert spaces.

Proposition 1 ((Minsker, 2017) and Theorem 7.3.1 in (Tropp, 2012))**.** *Let A_i, $i \in [n]$ be i.i.d copies of a random Hilbert-Schmidt operator A on separable Hilbert spaces. Let $\|A\| \le c$ almost surely, $\mathbb{E}A = 0$ and let $\mathbb{E}[A^2] \preceq V$ for some trace class operator V. Then with probability at least $1-\delta$*

$$\left\|\frac{1}{n}\sum_{i\in[n]} A_i\right\| \le \frac{2c}{3n}\mathcal{L}_A(\delta) + \sqrt{\frac{2\|V\|}{n}\mathcal{L}_A(\delta)}, \tag{31}$$

where

$$\mathcal{L}_A(\delta) := \log\frac{4}{\delta} + \log\frac{\mathrm{tr}(V)}{\|V\|}.$$

Proposition 2. *Assume that $c_\psi := \sup_x\left\{\|\psi_w(x)\|^2\right\} < \infty$. Given $\delta > 0$, with probability in the i.i.d. draw of $(x_i)_{i=1}^n$ from μ, it holds that*

$$\mathbb{P}\{\|\widehat{C}_X^w - C_X^w\|/\|C_X^w\| \le \varepsilon_n(\delta)\} \ge 1-\delta, \tag{32}$$

where

$$\varepsilon_n(\delta) := \frac{4c_\psi}{3\|C_X^w\|\,n}\mathcal{L}(\delta) + \sqrt{\frac{2}{\|C_X^w\|\,n}\mathcal{L}(\delta)} \quad \textit{and} \quad \mathcal{L}(\delta) := \log\frac{4r}{\delta}. \tag{33}$$

Proof of Proposition 2. Proof follows directly from Proposition 1 applied to operators $\psi_w(x_i)\otimes\psi_w(x_i)$ using the fact that $C_X^w = \mathbb{E}\,\psi_w(x_i)\otimes\psi_w(x_i)$, where we recall that $\psi_w(x) := (\psi_{w,1}(x),\ldots,\psi_{w,r}(x)) \in \mathbb{R}^r$. Hence, we have the obvious upper bound $\frac{\mathrm{tr}(C_X^w)}{\|C_X^w\|} \le r$. □

We deduce from (32), with probability at least $1-\delta$

$$(1-\varepsilon_n(\delta))\left\|C_X^w\right\| \leq \left\|\widehat{C}_X^w\right\| \leq \left\|C_X^w\right\|(1+\varepsilon_n(\delta))\left\|C_X^w\right\|, \tag{34}$$

A similar result is also valid for $\widehat{C}_{X'}^w$ provided that $c_{\psi'} := \sup_{w,x}\left\{\left\|\psi'_w(x)\right\|^2\right\} < \infty$. Define

$$\varepsilon'_n(\delta) := \frac{4c_{\psi'}}{3\left\|C_X^w\right\| n}\mathcal{L}(\delta) + \sqrt{\frac{2}{\left\|C_X^w\right\| n}\mathcal{L}(\delta)} \quad \text{and} \quad \mathcal{L}(\delta) := \log\frac{4r}{\delta}. \tag{35}$$

Then, with probability at least $1-\delta$

$$(1-\varepsilon'_n(\delta))\left\|C_{X'}^w\right\| \leq \left\|\widehat{C}_{X'}^w\right\| \leq (1+\varepsilon'_n(\delta))\left\|C_{X'}^w\right\|, \tag{36}$$

We study now the deviation of $\left\|\widehat{C}_{XX'}^w\right\|_{\mathrm{HS}}^2$ from $\left\|C_{XX'}^w\right\|_{\mathrm{HS}}^2$. We can essentially apply Theorem 3 in (Gretton et al., 2005) with kernels $k_w(x,x') = \langle\psi_w(x),\psi_w(x')\rangle$ and $l_w(y,y') = \langle\psi'_w(y),\psi'_w(y')\rangle$. Note that these two kernels are essentially bounded. Indeed we have

$$\operatorname*{ess\,sup}_{x,x'}|k_w(x,x')| \leq \sup_x\|\psi_w(x)\|^2 \leq c_\psi \quad \text{and} \quad \operatorname*{ess\,sup}_{y,y'}|l_w(y,y')| \leq \sup_x\|\psi'_w(y)\|^2 \leq c_{\psi'}.$$

Hence, for any $n \geq 2$ and $\delta > 0$, we get with probability at least $1-\delta$,

$$\left|\left\|\widehat{C}_{XX'}^w\right\|_{\mathrm{HS}}^2 - \left\|C_{XX'}^w\right\|_{\mathrm{HS}}^2\right| \leq \epsilon''_n(\delta), \tag{37}$$

where

$$\epsilon''_n(\delta) := c_\psi c_{\psi'}\sqrt{\frac{5\log(6\delta^{-1})}{n}} + c_\psi c_{\psi'}\frac{C}{n},$$

for some absolute constant $C > 0$.

D.2 Operator regression and prediction

We next discuss how to design an estimator of the transfer operator $\mathcal{T}\colon L^2_{\mu'}(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ using the learned subspaces $\mathcal{H}_w$ and $\mathcal{H}'_w$. Namely, we estimate $\mathcal{T} \approx \widehat{\mathcal{T}}_w\colon \mathcal{H}'_w \to \mathcal{H}_w$. The purpose of such estimation is to, given a initial state $x \in \mathcal{X}$, predict the average evolution $\mathbb{E}[f(X') \mid X = x]$ of an observable $f \in L^2_{\mu'}(\mathcal{X})$. Remark that in the main text, we have just discussed the task when one takes $\mathcal{H}'_w = \mathcal{H}_w$.

In what follows, let us assume that after the training we obtained $w \in \mathcal{W}$ such that C_X^w and $C_{X'}^w$ are invertible, that is that $(\psi_{w,j})_{j\in[r]}$ $(\psi'_{w,j})_{j\in[r]}$ form basis of the spaces $\mathcal{H}_w$ and $\mathcal{H}'_w$, respectively. This means that the operators $E_{w,}\colon \mathbb{R}^r \mapsto \mathcal{H}_w$ and $E'_w\colon \mathbb{R}^r \mapsto \mathcal{H}'_w$ can be properly defined as partial isometries by $E_w v = \psi_w(\cdot)^\top v$ and $E'_w v = \psi'_w(\cdot)^\top v$. So, every estimator can be written in the form $\widehat{\mathcal{T}}_w = E_w\widehat{T}(E'_w)^*$ for some $\widehat{T} \in \mathbb{R}^{r\times r}$.

Different estimators can then computed from data $\mathcal{D}_n := (x_i, x'_i)_{i\in[n]}$ (either seen or unseen during training time). To elaborate on this, let us, as usual for kernel methods, define the sampling operators $\widehat{S}\colon \mathcal{H}_w \to \mathbb{R}^n$ and $\widehat{S}'\colon \mathcal{H}'_w \to \mathbb{R}^n$, given by $\widehat{S}h = n^{-\frac{1}{2}}[h(x_1)\ \dots\ h(x_n)]^\top$, $f \in \mathcal{H}_w$, and $\widehat{S}'g = n^{-\frac{1}{2}}[g(x'_1)\ \dots\ g(x'_n)]^\top$, $g \in \mathcal{H}'_w$. Notice that we can extend the domain of definition of these operators via interpolation to arbitrary functions $\mathcal{X} \to \mathbb{R}$ that can be evaluated on a dataset, respectively. Hence, without possible confusion, when evaluating we can use $\widehat{S}f$ and $\widehat{S}'f'$ even when $f \notin \mathcal{H}_w$ or $f' \notin \mathcal{H}'_w$.

Now, as shown in (Kostic et al., 2022), the empirical estimator $\widehat{\mathcal{T}}_w$ of the transfer operator $\mathcal{T}$ using dataset $\mathcal{D}_n$ can be obtained via operator regression by minimizing the empirical risk

$$\left\|\widehat{S}' - \widehat{S}\widehat{\mathcal{T}}_w\right\|_{\mathrm{HS}}^2 = \left\|\widehat{S}' - \widehat{S}E_w\widehat{T}(E'_w)^*\right\|_{\mathrm{HS}}^2 = \left\|\widehat{S}'E'_w - \widehat{S}E_w\widehat{T}\right\|_{\mathrm{HS}}^2$$

where the last equality holds since E'_w is a partial isometry. Therefore, the simple least square (LS) estimator is than obtained as $\widehat{T} := (E_w^*\widehat{S}^*\widehat{S}E_w)^\dagger(E_w^*\widehat{S}^*\widehat{S}'E'_w) = (\widehat{C}_X^w)^\dagger\widehat{C}_{XX'}^w$, or, equivalently, as $\widehat{\mathcal{T}}_w = E_w\widehat{T}(E'_w)^* = \widehat{S}^*(\widehat{S}\widehat{S}^*)^\dagger\widehat{S}'$

Once the regression is performed, recalling that X' is a $\Delta t = 1$ step ahead evolution of X we can use it to approximate $\mathbb{E}[f(X') \mid X = x] \approx (\widehat{\mathcal{T}}_w f)(x)$ for $f\colon \mathcal{X} \to \mathbb{R}$, as the following result shows.

Proposition 3. *Let $\widehat{\mathcal{T}}\colon \mathcal{H}'_w \to \mathcal{H}_w$ be a LS estimator of $\mathcal{T}$, then for every $x \in \mathcal{X}$ and $f \in \mathcal{H}_w$*

$$(\widehat{\mathcal{T}}_w f)(x) = \psi_w(x)^\top \, (\widehat{C}_X^w)^\dagger \widehat{\Psi}_w [f(x'_1) \,|\, \cdots \,|\, f(x'_n)]^\top, \tag{38}$$

where $\widehat{\Psi}_w := [\psi_w(x_1) \,|\, \cdots \,|\, \psi_w(x_n)] \in \mathbb{R}^{r \times n}$.

Proof. The proof follows directly observing that $\widehat{\Psi}_w^\top = \sqrt{n}[\widehat{S}\psi_{w,1} \,|\, \ldots \,|\, \widehat{S}\psi_{w,r}] = \sqrt{n}\widehat{S}E_w$. Namely, then

$$\widehat{\Psi}_w [f(x'_1) \,|\, \cdots \,|\, f(x'_n)]^\top = (\widehat{S}E_w)^* \widehat{S}' f = E_w^* \widehat{S}^* \widehat{S}' f = E_w^* \widehat{S}^* \widehat{S}' E_w E_w^* f = \widehat{C}_{XX'}^w E_w^* f$$

and, hence

$$\psi_w(x)^\top \, (\widehat{C}_X^w)^\dagger \widehat{\Psi}_w [f(x'_1) \,|\, \cdots \,|\, f(x'_n)]^\top = E_w (\widehat{C}_X^w)^\dagger \widehat{C}_{XX'}^w E_w^* f = \widehat{\mathcal{T}}_w f.$$

□

We remark that the previous result formally holds for $f \in \mathcal{H}'_w$, but it can be easily extended to functions in $L^2_{\mu'}(\mathcal{X})$ via interpolation.

D.3 Dynamics mode decomposition and forecasting

Now we consider the problem of forecasting the process for several time steps in future using what is known as (extended) dynamic mode decomposition, which is based on the estimated eigenvalues and eigenfunctions of the transfer operator. As observed in the main body, this is meaningful only if the operator is an endomorphism on a function space, that is if it maps the space into itself.

Hence, after training DPNet we will use just one representation ψ_w and its r-dimensional space of functions $\mathcal{H}_w := \operatorname{span}(\psi_{w,j})_{j \in [r]}$ to perform the operator regression, as explained in the previous section, and obtain an estimator $\widehat{\mathcal{T}}_w = E_w \widehat{T} E_w^* \colon \mathcal{H}_w \to \mathcal{H}_w$, for some matrix $\widehat{T} \in \mathbb{R}^{r \times r}$. Then, if $(\widehat{\lambda}_i, \widehat{u}_i, \widehat{v}_i)_{i \in [r]} \subset \mathbb{C} \times \mathbb{C}^r \times \mathbb{C}^r$ is a spectral decomposition of $\widehat{T}$, then $(\widehat{\lambda}_i, E_w \widehat{u}_i, E_w \widehat{v}_i)_{i \in [r]}$ is a spectral decomposition of $\widehat{\mathcal{T}}_w$. In the following result we show how to compute dynamic mode decomposition of $\mathcal{T}$ based on the estimator $\widehat{\mathcal{T}}_w$ and use it for forecasting by approximating $\mathbb{E}[f(X_t) \,|\, X_0 = x] \approx ((\widehat{\mathcal{T}}_w)^t f)(x)$, for $f \colon \mathcal{X} \to \mathbb{R}$, $x \in \mathcal{X}$.

Proposition 4. *Let $\widehat{\mathcal{T}}_w = E_w \widehat{T} E_w \colon \mathcal{H}_w \to \mathcal{H}_w$ be rank r LS estimator of $\mathcal{T}_{\Delta t}$, for $\Delta t = 1$. If $\widehat{T} = \sum_{i \in [r]} \widehat{\lambda}_i \widehat{v}_i \, \widehat{u}_i^*$ is the spectral decomposition of $\widehat{T}$, and $\widehat{f}_i(x) := \psi_w(x)^\top \widehat{v}_i$ and $\widehat{g}_i(x) := (\widehat{u}_i)^* \psi_w(x)$, $i \in [r]$, then for every $t \in \mathbb{N}$, every $x \in \mathcal{X}$ and every $f \in \mathcal{H}_w$ it holds that*

$$((\widehat{\mathcal{T}}_w)^t f)(x) = \sum_{i \in [r]} \widehat{\lambda}_i^t \, \widehat{f}_i(x) \, \widehat{u}_i^* D_w(f), \tag{39}$$

where $D_w(f) := \widehat{\Lambda}^{-1} (\widehat{C}_X^w)^\dagger \widehat{\Psi}_w [f(x'_1) \,|\, \ldots \,|\, f(x'_n)]^\top \in \mathbb{R}^r$.

Proof. First observe that

$$(\widehat{\mathcal{T}}_w)^t f = E_w \widehat{T}^t E_w^* f = E_w \widehat{T}^{t-1} E_w^* E_w \widehat{T} E_w^* f = \sum_{i \in [r]} \widehat{\lambda}_i^{t-1} \, (\widehat{f}_i \otimes \widehat{g}_i) E_w \widehat{T} E_w^* f.$$

Hence,

$$((\widehat{\mathcal{T}}_w)^t f)(x) = \sum_{i \in [r]} \widehat{\lambda}_i^{t-1} \, \widehat{f}_i(x) (\widehat{u}_i^* \widehat{T} E_w^* f) = \sum_{i \in [r]} \widehat{\lambda}_i^t \, \widehat{f}_i(x) (\widehat{\lambda}^{-1} \widehat{u}_i^* \widehat{T} E_w^* f)$$

and the rest of the proof follows as in Prop. 3. □

We remark that $\widehat{u}_i^* D_w(f)$ is known as i-th Koopman mode of the observable f, and that in comparison to Encoder-Decoder approaches $D_w(f)$ can be considered as a decoder when forecasting function $f \colon \mathcal{X} \to \mathbb{R}$. Clearly this is easily extended to vector valued functions, and, hence, we can forecast the states by using $\widehat{u}_i^* D_w(I) := \widehat{\Lambda}^{-1} (\widehat{C}_X^w)^\dagger \widehat{\Psi}_w [x'_1 \,|\, \ldots \,|\, x'_n]^\top \in \mathbb{R}^d$, where $\mathcal{X} \subset \mathbb{R}^d$.

Algorithm 2 DPNets Training (Discrete)

Input: data $\mathcal{D}_n = (x_1, \dots, x_n)$, $\mathcal{D}'_n = (x'_1, \dots, x'_n)$, metric distortion loss $\mathcal{R}$; optimizer U; DNNs $\psi_w, \psi'_w : \mathcal{X} \to \mathbb{R}^r$; metric loss coefficient γ; # of steps K; minibatch size m.

Initialize DNN weights w_1
for $k = 1$ to K **do**,
  Sample minibatches $(y_1, \dots, y_m)$ from $\mathcal{D}_n$, and $(y'_1, \dots, y'_m)$ from $\mathcal{D}'_n$.
  Compute empirical covariance matrices
  $\widehat{C}^{w_k}_X \leftarrow \frac{1}{n}\sum_{i=1}^m \psi_{w_k}(y_i)\psi_{w_k}(y_i)^\top, \widehat{C}^{w_k}_{X'} \leftarrow \frac{1}{n}\sum_{i=1}^m \psi'_{w_k}(y'_i)\psi'_{w_k}(y'_i)^\top$
  $\widehat{C}^{w_k}_{XX'} \leftarrow \frac{1}{n}\sum_{i=1}^m \psi_{w_k}(y_i)\psi'_{w_k}(y'_i)^\top$
  if DPNets-relaxed **then**
    $F(w_k) \leftarrow \widehat{\mathcal{S}}^\gamma_m(w_k) := \frac{\left\|\widehat{C}^{w_k}_{XX'}\right\|^2_{\mathrm{HS}}}{\left\|\widehat{C}^{w_k}_X\right\|\left\|\widehat{C}^{w_k}_{X'}\right\|} - \gamma\big(\mathcal{R}(\widehat{C}^{w_k}_X) + \mathcal{R}(\widehat{C}^{w_k}_{X'})\big)$
  else
    $F(w_k) \leftarrow \widehat{\mathcal{P}}^\gamma_m(w_k) := \left\|(\widehat{C}^{w_k}_X)^{\frac{\dagger}{2}}\widehat{C}^{w_k}_{XX'}(\widehat{C}^{w_k}_{X'})^{\frac{\dagger}{2}}\right\|^2_{\mathrm{HS}} - \gamma\big(\mathcal{R}(\widehat{C}^{w_k}_X) + \mathcal{R}(\widehat{C}^{w_k}_{X'})\big)$
  end if
  $w_{k+1} \leftarrow U(w_k, \nabla F(w_k))$ where $\nabla F(w_k)$ is computed via backpropagation
end for
return representations ψ_{w_K}, ψ'_{w_K}

D.4 Extended Algorithm and Training Time

Algorithm 2 is an extended version of Algorithm 1. The time complexity of computing the empirical scores $\widehat{\mathcal{P}}^\gamma_m(w)$, $\widehat{\mathcal{S}}^\gamma_m(w)$ and (thanks to backpropagation) their gradients, is $O(m\,\mathrm{Cost}(\psi_w)+mr^2+r^3)$ and $O(m\,\mathrm{Cost}(\psi_w)+mr^2)$ respectively, where $\mathrm{Cost}(\psi_w)$ is the cost of one evaluation of ψ_w. Namely, computing the embeddings for m samples costs $O(m\,\mathrm{Cost}(\psi_w))$, computing the (cross)covariance matrices for m samples costs $O(mr^2)$, computing the pseudoinverse via eignevalue decomposition costs $O(r^3)$, while computing the operator norm of the covariance matrices using e.g. the Arnoldi iteration method costs $O(r^2)$. We note that the cost of training VAMPNets (Mardt et al., 2018) is the of the same order as DPNets without relaxation, since evaluating the metric distortion loss is relatively cheap once we have the covariance matrices.

E SDE learning

We next prove the second main result on the optimization problem (5).

Theorem 2. *If $\mathcal{H}_w \subseteq W^{1,2}_\pi(\mathcal{X})$, and $\lambda_1(\mathcal{L}) \geq \cdots \geq \lambda_{r+1}(\mathcal{L})$ are eigenvalues of $\mathcal{L}$ above its essential spectrum, then for every $\gamma \geq 0$ it holds*

$$\mathcal{P}^\gamma_\partial(w) := \mathrm{tr}\left((C^w_X)^\dagger C^w_{X\partial}\right) - \gamma\mathcal{R}(C^w_X) \leq \lambda_1(\mathcal{L}) + \cdots + \lambda_r(\mathcal{L}), \tag{14}$$

and the equality is achieved when $\psi_{w,j}$ is the eigenfunction of $\mathcal{L}$ corresponding to $\lambda_j(\mathcal{L})$, for $j \in [r]$.

Proof. In view of Lem. 2, we only need to prove that $\mathrm{tr}(P_{\mathcal{H}_w}\mathcal{L}P_{\mathcal{H}_w}) = \mathrm{tr}\left((C^w_X)^\dagger C^w_{X\partial}\right)$. To that end, we reason as in the proof of Lem. 3 to obtain that $P_{\mathcal{H}_w} = UQ(C^w_X)^{\dagger/2}Q^*S^*_\pi$, where $S_\pi\colon \mathcal{H}_w \hookrightarrow W^{1,2}_\pi(\mathcal{X})$ is an injection and $Q\colon \mathbb{R}^r \to \mathcal{H}_w$ and $U\colon \mathcal{H}_w \to W^{1,2}_\pi(\mathcal{X})$ are partial isometries. Moreover, recalling (2), we have that

$$\begin{aligned}
(Q^*S^*_\pi\mathcal{L}S_\pi Q) &= \lim_{\Delta t\to 0^+}\frac{Q^*S^*_\pi(\mathcal{T}_{\Delta t} - I)S_\pi Q}{\Delta t} \\
&= \lim_{\Delta t\to 0^+} Q^*\left[\int_{\mathcal{X}}\pi(dx)\phi^w_X(x)\otimes\frac{\int_{\mathcal{X}}p_{\Delta t}(x,dx')(Q^*\phi^w_X(x') - Q^*\phi^w_X(x))}{\Delta t}\right] \\
&= \int_{\mathcal{X}}\pi(dx)\psi_w(x)\otimes\left(\lim_{\Delta t\to 0^+}\frac{\int_{\mathcal{X}}p_{\Delta t}(x,dx')(\psi_w(x') - \psi_w(x))}{\Delta t}\right) \\
&= \mathbb{E}_{X\sim\pi}\left[\psi_w(X)\otimes d\psi_w(X)\right] = C^w_{X\partial},
\end{aligned}$$

	DPNets	DPNets-relax	VAMPNets	DynAE	ConsAE
Logistic	1×10^{-4}	1×10^{-4}	3×10^{-6}	-	-
Fluid	Failed	1×10^{-4}	Failed	9×10^{-4}	9×10^{-4}
MNIST	9×10^{-4}	9×10^{-4}	9×10^{-4}	9×10^{-4}	9×10^{-4}
Chignolin	Failed	1×10^{-3}	Failed	-	-
Langevin	1×10^{-3}	-	-	-	-

Table 3: Best learning rates found by a grid search for each (experiment, method) pair. For all experiments, the grid was made by 100 equally spaced values in the interval $(10^{-6}, 10^{-2})$.

where the last line follows from Itō formula (see e.g. Arnold, 1974). Hence, using $P_{\mathcal{H}_w} = UQ(C_X^w)^{\dagger/2}Q^* S_\pi^*$ we have that

$$\mathrm{tr}(P_{\mathcal{H}_w}\mathcal{L}P_{\mathcal{H}_w}) = \mathrm{tr}((C_X^w)^{\dagger/2}C_{X\partial}^w(C_X^w)^{\dagger/2}) = \mathrm{tr}((C_X^w)^{\dagger}C_{X\partial}^w),$$

which completes the proof □

To conclude this section, we remark that in the continuous setting the estimator $\widehat{\mathcal{L}}_w \colon \mathcal{H}_w \to \mathcal{H}_w$ of $\mathcal{L}$ on the learned space $\mathcal{H}_w$ can be obtained via operator regression in a similar way as discussed in Sec. 4. So, the LS estimator is given by matrix $\widehat{L} = (\widehat{C}_X^w)^\dagger \widehat{C}_{X\partial}^w$, and its spectral decomposition is

$$\widehat{\mathcal{L}}_w = \textstyle\sum_{i\in[r]} \widehat{\lambda}_i\, \widehat{f}_i \otimes \widehat{g}_i, \quad \text{where} \quad \widehat{f}_i(x) := \psi_w(x)^\top \widehat{v}_i \;\text{ and }\; \widehat{g}_i(x) := (\widehat{u}_i)^* \psi_w(x), \tag{40}$$

where $\widehat{L} = \sum_{i\in[r]} \widehat{\lambda}_i \widehat{v}_i\, \widehat{u}_i^\top$ is the spectral decomposition of the matrix $\widehat{L}$.

Hence, using that $\lambda_i(\mathcal{T}_{\Delta t}) = \exp(\lambda_i(\mathcal{L}))$, we directly obtain the modal decomposition in the continuous time,

$$\mathbb{E}[f(X_t) \,|\, X_0 = x] \approx \textstyle\sum_{i\in[r]} \exp(\widehat{\lambda}_i\, t)\, \widehat{f}_i(x)\, \widehat{u}_i^* D_w(f), \quad t \in [0, +\infty) \tag{41}$$

where $D_w(f) = (\widehat{C}_X^w)^\dagger \widehat{\Psi}_w[f(x_1)\,|\cdots|\,f(x_n)]^\top \in \mathbb{R}^r$ are the coefficients of the LS estimator of f in $\mathcal{H}_w$. As a final remark, note that in (41) we regress function f onto $\mathcal{H}_w$ using least squares with data $(x_i, f(x_i))$.

F EXPERIMENTS

Hardware The experiments were performed on a workstation equipped with an Intel(R) Core™i9-9900X CPU @ 3.50GHz, 48GB of RAM and a NVIDIA GeForce RTX 2080 Ti GPU. Due to RAM insufficiency, the Nyström baseline reported in Table 2 was performed on a CPU node of a cluster with 2x AMD EPYC 7713 @ 2.0GHz and 512GB of RAM.

Software All experiments and baselines have been implemented in Python 3.11 and Pytorch 2.0, the only exception being the PINNs baseline for the fluid flow experiment, for which we relied on the original implementation of the code. All the code to reproduce the experiments will be made openly available.

General remarks Every algorithm has been performed in `float32` single precision, fixing the random number generator seed where appropriate. We made sure that each algorithm was trained on the same combinations of input-output data, and for neural network models we used the same batches and number of epochs. The learning rate for each method was tuned independently by running a small number of steps at 100 equally spaced learning rates in the interval $(10^{-6}, 10^{-2})$ and selecting the best out of these. The best learning rates for each experiment are reported in Table 3

Once a representation was learned, we always used the Ordinary Least Squares estimator described in 4 to perform the subsequent evaluation tasks.

F.1 LOGISTIC MAP

Data We generated the data as explained in Kostic et al. (2022) for a value $N = 20$ of the trigonometric noise. To train DPNets and VAMPNets we have sampled a trajectory of $2^{14} \approx 16000$ points. We used a batch size of 2^{13} points and trained for 500 epochs.

Optimization Adam with learning rate tuned as explained in the general remarks.

Architecture Multi layer perceptron of shape `Linear[64]` $\rightarrow$ `Linear[128]` $\rightarrow$ `Linear[64]` $\rightarrow$ `Linear[feature_dim]` with Leaky ReLU activations. We have set the dimension of the feature map to $r = 7$ as the minimal dimension allowing to meaningfully learn the first three eigenvalues.

Additional results: the role of feature dimension In the left panel of Fig. 5 we compare the eigenvalues and the singular values of $\mathcal{T}$ for the noisy logistic map, while in the center and right panels we show the metrics reported in Tab. 1 of the main text as a function of the feature dimension r. Notice how the singular values decay significantly slower than eigenvalues, a consequence of the fact that the transfer operator is not normal, i.e. $\mathcal{T}\mathcal{T}^* \neq \mathcal{T}^*\mathcal{T}$. Non-normality makes the estimation of the spectra of $\mathcal{T}$ particularly sensitive, as captured by the pseudospectra, see Trefethen and Embree (2020). In Fig. 6 we show how small estimation errors in operator norm (label on the contour lines) incur larger errors in the eigenvalue estimation (distance of the true eigenvalues to the contour lines in the complex plane). This means that for non-normal operators the estimation error needs to be typically much smaller than the modulus of the eigenvalues one wants to recover. For $r < 7$, the spectral error of every model is of the same order of the eigenvalues to approximate, as unequivocally shown in Fig. 6. At $r = 7$, DPNets-relaxed already give a decent approximation of the three leading eigenvalues, see also Fig. 6. From $r > 7$ onward, every model progressively yields reasonable estimations, with DPNets and Cheby-T quickly catching up with DPNets-relaxed. For the optimality gap, every model shows an improving trend by increasing the feature dimension r, with DPNets showing the strongest performance.

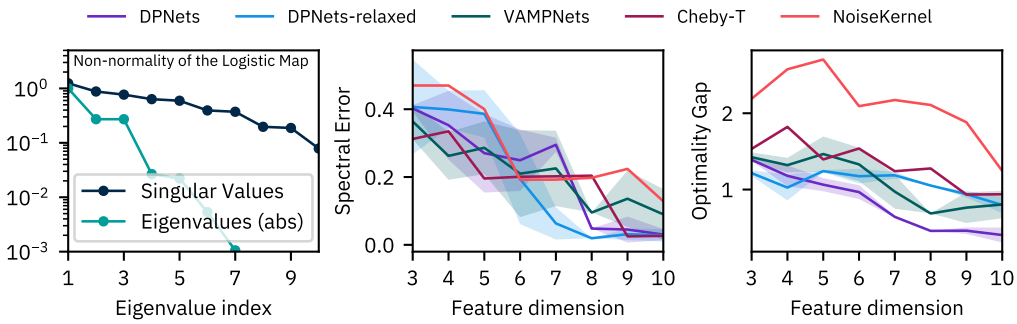

Figure 5: (Left) Decay of the singular values $\sigma_i(\mathcal{T})$ and of the the eigenvalues $|\lambda_i(\mathcal{T})|$ for the logistic map example. (Center, Right) Spectral error and Optimality gap as a function of the feature dimension r for the baselines considered in Tab. 1.

Hyperparameters Metric loss coefficient $\gamma = 1$.

The reported results concern the average over 20 independent seed initializations.

F.2 Fluid Flow past a cylinder

The data for this experiment are equally spaced sampled solutions of the Navier-Stokes equations (Trefethen and Embree, 2020) for an incompressible Newtonian fluid coupled with the transport equation

$$\partial_t c + \boldsymbol{u} \cdot \nabla c = \mathrm{Pec}^{-1} \nabla^2 c.$$

Here, Pec is the Péclet number, and $c : \mathbb{R}^3 \rightarrow \mathbb{R}$ is a field representing the concentration $c(t; x, y)$ of a scalar quantity which is transported by the fluid flow without influencing the fluid motion itself (e.g. a dye dissolved in water). These partial derivative equations are solved over a 100×200 regular grid.

In this experiment we have only been able to train DPNets-relaxed. Indeed both DPNets (unrelaxed) and VAMPNets failed due to linear algebra errors arising in the back propagation step for the pseudo-inverse matrix.

Data Available at `https://github.com/maziarraissi/HFM`. It consists of 201 snapshots: 160 used for training, the rest for testing. Has been standardized: each channel with its own mean and std. In Fig. 8 we show two snapshots from the training dataset.

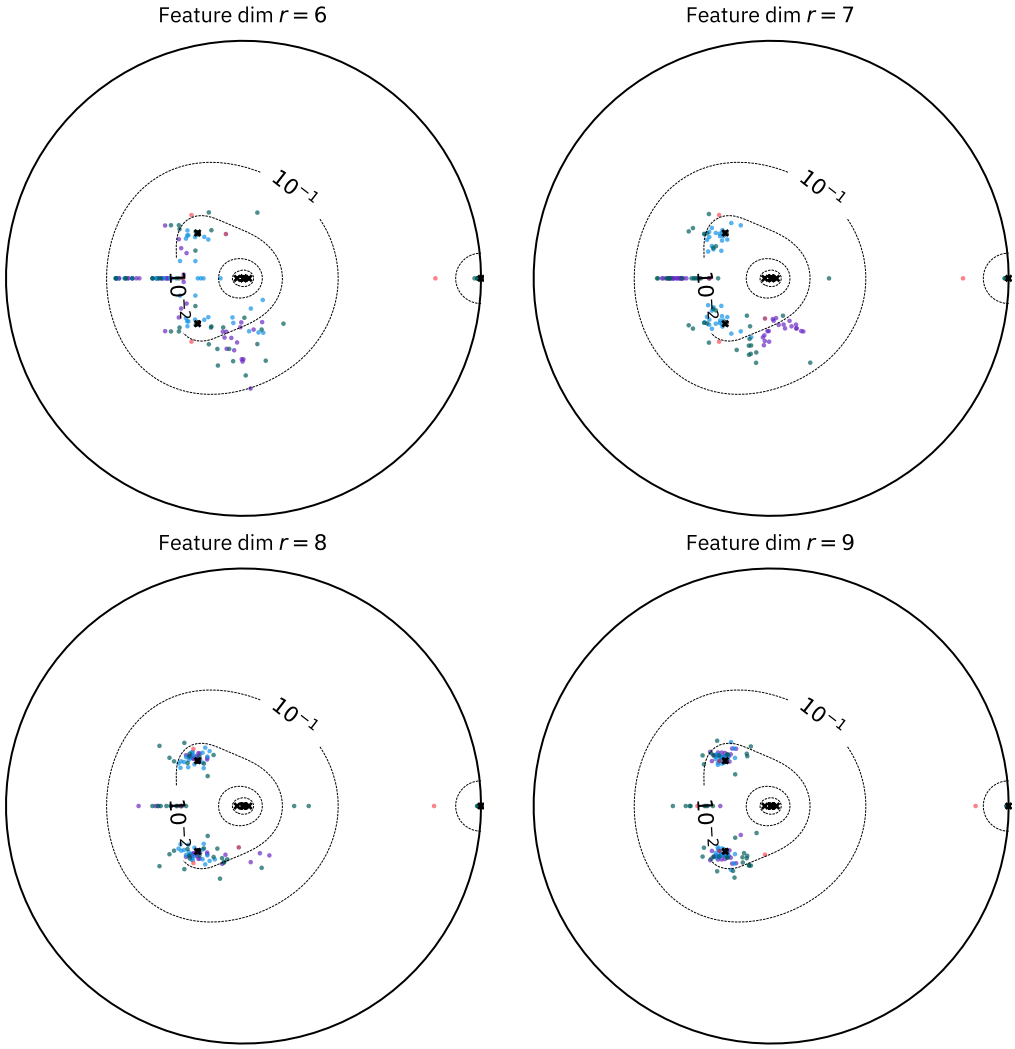

Figure 6: Estimation of the eigenvalues of the logistic map at different feature dimensions. For each model, we plot the eigenvalues estimated by 20 independent initial seeds. The color in each scatter plot follows the same color-coding of Fig. 5. Black $\times$ represent reference eigenvalues, while the dashed contour lines show the pseudospectral regions when the estimation error ranges from 10^{-1} (outermost) to 10^{-4} (innermost). Note that for this problem pseudospectra indicates that the leading eigenvalue is easy to recover, while recovering eigenvalues close to zero is very hard.

Optimization Adam with learning rate tuned as explained in the general remarks. Full-batch training. 5000 total training iterations/epochs. The training time of a full batch $\approx$ 39 mins.

Architecture MLP with layers of width `[128, 512, 1024, 512, 256, 64]` and ReLU activation function

Hyperparameters Metric loss coefficient $\gamma = 1$;

F.3 Continuous dynamics

The implementation of this experiment is straightforward, and our results can be reproduced using the following informations.

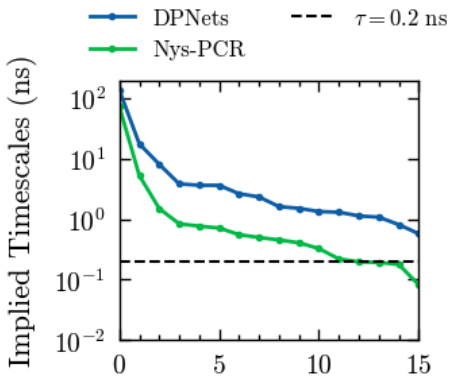

Figure 7: The implied timescale (Mardt et al., 2018) associated to the eignvalues of the transfer operator. Everything which is below the simulation lagtime $\tau = 0.2$ ns, is just numerical noise.

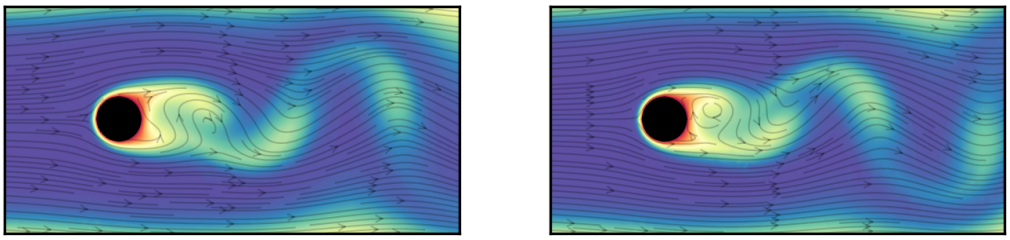

Figure 8: Two snapshots of the passive scalar concentration for the flow past a cylinder.

Data Produced in-house with JAX MD. Both the dataset and the script to produce new trajectories will be released. The dataset consists of 10^5 snapshots: 70% used for training, 10% for validation, 20% for testing.

Optimization Adam with learning rate 10^{-3}. Other parameters are the predefined values in `Optax`'s implementation. Batch size: 8192. 500 epochs. Training time is ≈ 2 mins.

Architecture Multi layer perceptron with CeLu activation function. Dimension of the hidden layer: `[32, 64, 128, 128, 64, 4]`.

Hyperparameters Metric loss coefficient $\gamma = 50$.

F.4 Ordered MNIST

Data out of the full MNIST dataset we generated a trajectory of 1000 steps as discussed in the main text, and evaluated the forecasting accuracy over 1000 different test initial conditions. In Tab. 4 we report the training time for DPNets, DPNets-relaxed and the baselines used. We observe that both our methods are the fastest during training. See also Fig. 10 for the generated sequences of digits by the compared methods.

Optimization Adam with learning rate tuned as explained in the general remarks. Batches of 128 samples trained over 150 epochs.

Architecture `Conv2d[16]` $\rightarrow$ `ReLU` $\rightarrow$ `MaxPool[2]` $\rightarrow$ `Conv2d[32]` $\rightarrow$ `ReLU` $\rightarrow$ `MaxPool[2]` $\rightarrow$ `Linear[5]`.

For the Auto-Encoder baselines we used this architecture for the encoder and the "reversed" network constructed with transposed convolutions for the decoder.

Hyperparameters Metric loss coefficient $\gamma = 1$.

Table 4: Ordered MNIST Training times. Each model is run on CPU and we report mean $\pm$ std for 20 runs.

Model	Fit Time (s)	Time per Epoch (s)
DynamicalAE	0.571 ± 0.034	0.057 ± 0.003
Oracle-Features	0.098	-
DMD	0.333	-
KernelDMD-Poly3	4.914	-
KernelDMD-AbsExp	0.832	-
KernelDMD-RBF	0.776	-
DPNets	0.046 ± 0.004	0.049 ± 0.001
DPNets-relaxed	0.043 ± 0.004	0.051 ± 0.004

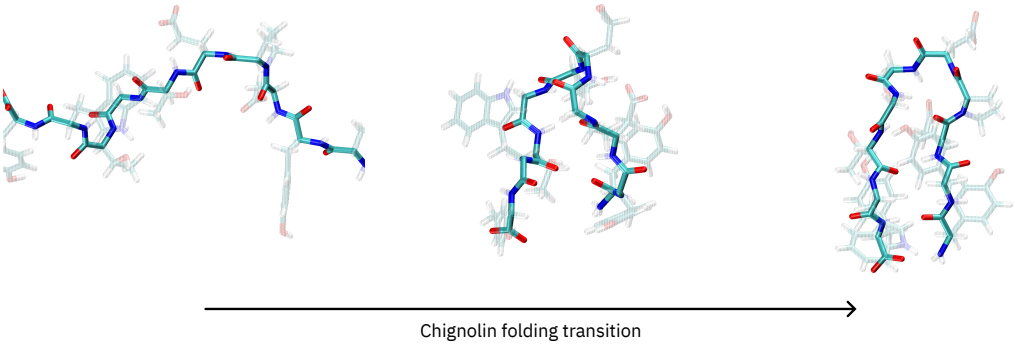

Figure 9: Three snapshots of Chignolin while undergoing a folding transition.

F.5 The metastable states of Chignolin

In this experiment we build from the work in (Ghorbani et al., 2022) by employing our framework to learn the leading eigenfunctions associated to the dynamics of Chignolin, a folding protein, from a $106\,\mu$s long molecular dynamics simulation sampled every $200\,$ps, totalling over half a million data points (Lindorff-Larsen et al., 2011). We consider every heavy atom for a total of 93 nodes as well as a cutoff radius of 6 Angstroms giving an average of 30 neighbours for each atom. Compared to (Ghorbani et al., 2022), which selects only the ≈ 20 C_α atoms each with its first 5 neighbours, our experiment has therefore a much larger scale. Indeed, (Ghorbani et al., 2022) reports being able to train directly with the objective $\mathcal{P}^0$, while in our case we always encountered numerical errors, and we were able to only succesfully train the $\mathcal{S}^\gamma$ objective. We parametrized the feature map with a graph neural network (GNN) model. GNNs currently are the state of the art in modeling atomistic systems (Chanussot et al., 2021), and allow one to elegantly incorporate the roto-translational and permutational symmetries prescribed by physics. Specifically, we train a SchNet (Schütt et al., 2019; 2023) model with 3 interaction blocks, where in each block the latent atomic environment is 64-dimensional and the inter-atomic distances used for the message-passing step are expanded over 20 radial basis functions. After the last interaction block, each latent atomic environment is forwarded to a linear layer and then aggregated via averaging. The model has been trained for 100 epochs with an Adam optimizer and a learning rate of 10^{-3}. We analyzed the eigenfunctions using the technique described in (Novelli et al., 2022), which links each metastable state to physically interpretable conformational descriptors. Our analysis aligns perfectly with (Novelli et al., 2022), where the slowest metastable state corresponds to the folding-unfolding transition and is linked to the distance between residues (1, 10) and (2, 9) located at opposite ends of the protein. Additionally, the immediately faster metastable state represents a conformational change within the folded state, characterized by the relative angle between residues 6 and 8.

In Fig. 9 we plot how the structure of Chignolin changes while performing a folding transition, while in Fig. 7 we plot the implied timescales of the dynamical modes of Chignolin as estimated by DPNets-relaxed and Nyström-PCR.

This is by far the heaviest experiment of the paper, and we made use of the package `SchNetPack` (Schütt et al., 2019; 2023) on multiple instances. In particular, we have used `SchNetPack` dataloaders and preprocessing transformations (casting to 32 bit precision and on-the fly computation of the distance matrix). Further, we have used `SchNetPack`'s implementation of the SchNet interaction block. To reproduce our results, the following informations may prove useful.

Data The data was presented for the first time in (Lindorff-Larsen et al., 2011) and is freely available for non-commercial use upon request to DeShaw research. Dataset of 524743 snapshots. Each graph is composed by the 93 heavy atoms. The edges are formed only if two atoms are less than 5 Angstroms distant. The average number of edges is 28.

Optimization Adam with learning rate 10^{-3}. Other parameters are the predefined values in `Torch`'s implementation. Batch size: 192. 100 epochs. Training time: ≈ 11 hrs.

Architecture SchNet with 3 blocks, 20 RBF functions expansions, 64 latent dimension. At the output of SchNet, the hidden variables associated to the nodes are averaged and forwarded to a dense layer with 16 final output features.

Hyperparameters Metric loss coefficient $\gamma = 0.01$. For the Nyström baseline we used $M = 5000$ inducing points.

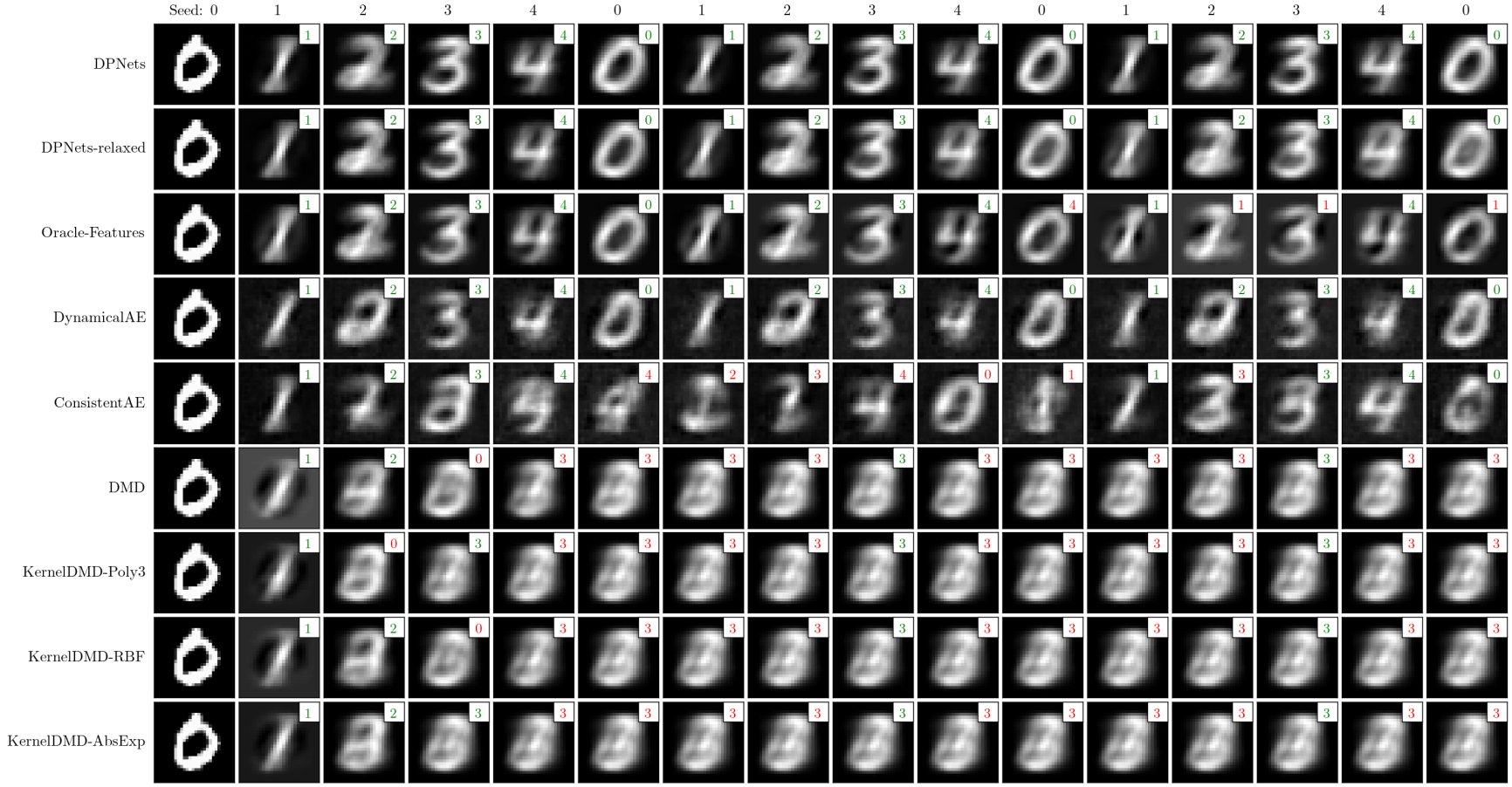

Figure 10: Sample of the states predicted by the different models starting from a seed image of the digit 0. The green label in the upper right corner of each image indicates proper classification, while the red one means miss-classification.