
Koopman Kernel Regression

Petar Bevanda
TU Munich
petar.bevanda@tum.de

Max Beier
TU Munich
max.beier@tum.de

Armin Lederer
TU Munich
armin.lederer@tum.de

Stefan Sosnowski
TU Munich
sosnowski@tum.de

Eyke Hüllermeier
LMU Munich
eyke@ifi.lmu.de

Sandra Hirche
TU Munich
hirche@tum.de

Abstract

Many machine learning approaches for decision making, such as reinforcement learning, rely on simulators or predictive models to forecast the time-evolution of quantities of interest, e.g., the state of an agent or the reward of a policy. Forecasts of such complex phenomena are commonly described by highly nonlinear dynamical systems, making their use in optimization-based decision-making challenging. Koopman operator theory offers a beneficial paradigm for addressing this problem by characterizing forecasts via linear time-invariant (LTI) ODEs, turning multi-step forecasts into sparse matrix multiplication. Though there exists a variety of learning approaches, they usually lack crucial learning-theoretic guarantees, making the behavior of the obtained models with increasing data and dimensionality unclear. We address the aforementioned by deriving a universal Koopman-invariant reproducing kernel Hilbert space (RKHS) that solely spans transformations into LTI dynamical systems. The resulting *Koopman Kernel Regression (KKR)* framework enables the use of statistical learning tools from function approximation for novel convergence results and generalization error bounds under weaker assumptions than existing work. Our experiments demonstrate superior forecasting performance compared to Koopman operator and sequential data predictors in RKHS.

1 Introduction

Dynamical systems theory is a fundamental paradigm for understanding and modeling the time evolution of a phenomenon governed by certain underlying laws. Such a perspective has been successful in describing countless real-world phenomena, ranging from engineering mechanics [1] and human movement modeling [2] to molecular and quantum systems [3, 4]. However, as the laws governing dynamical systems are often unknown, modeling and understanding the underlying phenomena may have to rely on data rather than first principles. In this regard, machine learning methods, which have shown immense potential in tackling complex tasks in domains such as language models [5] and computer vision [6], are coming to the fore. Though powerful, state-of-the-art neural vector fields [7] or flows [8] commonly compose highly nonlinear maps for forecast, i.e. computing

$$x(t) = x(0) + \int_0^t f(x(t))dt \quad (1)$$

for, e.g. a scalar ODE $\dot{x} = f(x)$. Hence, it is often challenging to use such models in optimization-based decision making that relies on simulators or predictive models, e.g., reinforcement learning [9–11]. A particularly beneficial perspective for dealing with the aforementioned comes from Koopman operator theory [12–15]. Through a point-spectral decomposition of Koopman operators, forecasts become superpositions of solution curves of a set of linear ODEs $\{\dot{z}_j = \lambda_j z_j\}_{j=1}^D$

$$x(t) = \sum_{k=1}^D e^{\lambda_k t} z_k(0), \quad \{x \xrightarrow{\text{lift } g_j} z_j\}_{j=1}^D \quad (2)$$

where a vector valued function $\text{span}(\{g_j\}_{j=1}^D)$ “lifts” x onto a manifold $\mathbb{Z} := \text{span}(\{z_j\}_{j=1}^D)$. Throughout, we refer to these models as *linear time-invariant (LTI) predictors*. The learning objective of such representations is twofold: spanning system trajectories by the learned manifold \mathbb{Z} and constraining the LTI dynamics to it. The latter is a long-standing challenge of Koopmanism [16–20], as manifold dynamics of existing approaches “leak-out” [21] and limit predictive performance.

To tackle the aforesaid, we connect the representation theories of reproducing kernel Hilbert spaces (RKHS) and Koopman operators. As a first in the literature, we derive a universal kernel whose RKHS exclusively spans manifolds invariant under the dynamics, as depicted in Figure 1. A key corollary of

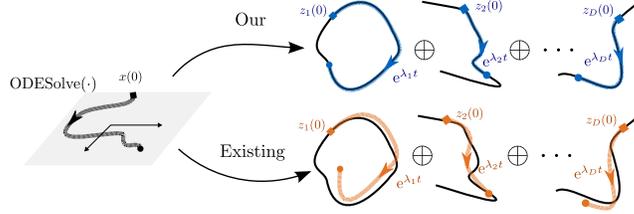


Figure 1: Illustration of on-manifold dynamics of LTI predictors.

unconstrained manifold dynamics is the lack of essential learning-theoretic guarantees, making the behavior of existing learned models unclear for increasing data and dimensionality. To address this, we utilize equivalences to function regression in RKHS to formalize a statistical learning framework for learning LTI predictors from sample trajectories of a dynamical system. This, in turn, enables the use of statistical learning tools from function approximation for novel convergence results and generalization error bounds under weaker assumptions than before [22–24]. Thus, we believe that our Koopman Kernel Regression (KKR) framework takes the best of both RKHS and Koopmanism by leveraging modular kernel learning tools to build provably effective LTI predictors.

The remainder of this paper is structured as follows: We briefly introduce LTI predictors and discuss related work in Section 2. The derivation of the KKR framework, including the novel Koopman RKHS, is presented in Section 3. In Section 4, we show the novel learning guarantees in terms of convergence and generalization error bounds. They are validated in comparison to the state-of-the-art through numerical experiments in Section 5.

Notation Lower/upper case bold symbols \mathbf{x}/\mathbf{X} denote spatial vector/matrix-valued quantities. A *trajectory* defines a curve $\mathbf{x}_T \subset \mathbb{X}$ traced out by the flow over time $\mathbb{T} = [0, T]$ from any $(\tau, \mathbf{x}) \in \mathbb{T} \times \mathbb{X}$. In discretizing \mathbb{T} , collection of points $\mathbf{x}_H \subset \mathbb{X}$ from discrete time steps $\mathbb{H} = \{t_0 \cdots t_H\}$ is considered. The state/output trajectory spaces are denoted as $\mathbb{X}_T \subseteq L^2(\mathbb{T}, \mathbb{X}) / \mathbb{Y}_T \subseteq L^2(\mathbb{T}, \mathbb{Y})$, with discrete-time analogues $\mathbb{X}_H \subseteq \ell^2(\mathbb{H}, \mathbb{X}) / \mathbb{Y}_H \subseteq \ell^2(\mathbb{H}, \mathbb{Y})$ with domain and co-domain separated by “/”. The vector space of continuous functions on \mathbb{X}_T endowed with the topology of uniform convergence on compact domain subsets is denoted $C(\mathbb{X}_T)$. The collection of bounded linear operators from \mathbb{Y}_T to \mathbb{Y}_T is denoted as $\mathcal{B}(\mathbb{Y}_T)$. The adjoint of $\mathcal{A} \in \mathcal{B}(\cdot)$ is \mathcal{A}^* . Discrete-time eigenvalues read $\mu := e^{\lambda \Delta t}$, $\lambda \in \mathbb{C}$. A random variable X defined on a probability space $(\Omega, \mathcal{A}, \rho)$ has expectation $\mathbb{E}[X] = \int_{\Omega} X(\omega) \rho(\omega)$.

2 Problem Statement and Related Work

To begin, we formalize our problem statement and put our work into context with existing work.

2.1 Problem Statement

Consider a forward-complete system¹ comprising a nonlinear state-space model

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x}_0 = \mathbf{x}(0), \quad (3a)$$

$$y = q(\mathbf{x}), \quad (3b)$$

on a compact domain $\mathbb{X} \subset \mathbb{R}^d$ with a quantity of interest $q \in C(\mathbb{X})$. The above system class includes all systems with Lipschitz flow $\mathbf{F}^t(\mathbf{x}_0) := \int_0^t \mathbf{f}(\mathbf{x}(\tau)) d\tau$, e.g., mechanical systems [27].

Inspired by the spectral decomposition of Koopman operators, we look to replace the nonlinear state-space model (3) by an *LTI predictor*

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}, \quad \mathbf{z}_0 = \mathbf{g}(\mathbf{x}_0), \quad (4a)$$

$$y = \mathbf{c}^\top \mathbf{z}, \quad (4b)$$

¹Although we outline the scalar output case for ease of exposition, expanding to a vector-valued case is possible w.l.o.g. If required, forward completeness can be relaxed to unboundedness observability [25, 26].

with $\bar{D} \in \mathbb{N}$ and \mathbf{g} a \bar{D} -dimensional function approximator dense in $C(\mathbb{X})$. Then, from initial conditions $\mathbb{X}_0 \subseteq \mathbb{X}$ that form a *non-recurrent* domain \mathbb{X}_T , (4) admits a universal approximation of the flow of (3) such that $\forall \varepsilon > 0, \exists \bar{D}$ so that $\sup_{\mathbf{x} \in \mathbb{X}_0} |y_T(t) - \mathbf{c}^\top e^{A t} \mathbf{g}(\mathbf{x})| < \varepsilon, \forall t \in [0, T]$ [28]². In this work, we aim to find a solution to the following constrained, functional optimization problem

(OR) *Output reconstruction:*

$$\min_{\mathbf{c}, \mathbf{g}, \mathbf{A}} \|\mathbf{y}_T - \mathbf{c}^\top \mathbf{g}_T\|_{\mathbb{Y}_T}, \quad (5a)$$

(KI) *Koopman-invariance:*

$$\text{such that } \mathbf{g}(\mathbf{x}(t)) = e^{A t} \mathbf{g}(\mathbf{x}(0)), \quad \forall t \in [0, T]. \quad (5b)$$

Although the sought-out model (4) is simple, the above problem is non-trivial and much of the existing body of work utilizes different simplifications that often lead to undesirable properties. In the following, we elaborate on these properties and motivate our novel sample-based solution to (5), which remains relatively simple but nonetheless ensures a well-defined solution with strong learning guarantees.

2.2 Related Work

Koopman operator regression in RKHS Equipped with a rich set of estimators, operator regression in RKHS seeks a sampled-data solution to

$$\min_{\mathbf{A}} \|\mathbf{g}(\mathbf{x}(t)) - \mathbf{A}^* \mathbf{g}(\mathbf{x}(0))\|_{L^2}, \quad (6)$$

with \mathbf{A} a Hilbert-Schmid operator [29] — commonly known as KRR, and EDMD (PCR) or RRR when under different fixed-rank constraints [23, 24]. The choice of RKHS \mathbf{g} is commonly one that is dense in a suitable L^2 space, e.g. that of the RBF kernel. By an additional projection, a quantity of interest can be predicted via a mode decomposition of the estimated operator, leading to a model akin to (4). In the light of (5), the feature map \mathbf{g} is predetermined while violating **(KI)** is merely minimized for a single time-instant t . As a consequence, such approaches are oblivious to the time-series structure — offering limited predictive power over the time interval $[0, T]$ of a trajectory as displayed in Figure 1. The extent to which **(KI)** is violated due to spectral properties [30] or estimator bias [24] is known as spectral pollution [31]. The strong implications of this phenomena, motivate regularization [32] and spectral bias measures [24] to reduce its effects. Due to the above challenges, guarantees for Koopman operator regression (KOR) have only recently gained increased attention. Often, however, existing theoretical results [29, 33] are generally not applicable to nonlinear dynamics [34] due to the commonly unavoidable misspecification [35] of the problem (6) incurred by neglecting **(KI)**. The first more general statistical learning results [23, 24] are derived in a stochastic setting under the assumption that the underlying operator is compact and self-adjoint. In stark contrast, the same set of assumptions is restrictive for the deterministic setting [35]: compactness only holds for affine deterministic dynamics [36, 37] while self-adjointness is known to generally not hold for Koopman operators [13, 38, 39]. Regardless of the setting, however, the state-of-the-art exhibits alarming properties: forecasting error not necessarily vanishing with LTI predictor (4) rank [23, Theorem 1] and risk based on a single time-instant.

Learning via Koopman eigenspaces Geared towards LTI predictors and closer to our own problem setting (2), another distinct family of approaches aims to directly learn the operator’s invariant subspaces [28, 40–42]. The goal is to fit $\mathbf{g}(\cdot)$ based on approximate Koopman operator eigenfunctions that still fit the output of interest **(OR)**. However, existing data-driven approaches in this line of work rely on ad-hoc choices and lack essential learning-theoretic properties such as feasibility and uniqueness of solutions — prohibiting provably accurate and automated LTI predictor learning.

Kernels for sequential data Motivated by the lack of priors that naturally incorporate streaming and sequential data, there is an increasing interest in *signatures* [43]. They draw from the rich theory of controlled differential equations (CDEs) [44, 45] and build models that depend on a time-varying observation history. An RKHS suitable for sequence modeling is induced by a signature transformation of a base/static RKHS. Generally, if the latter is universal, so are the signature kernels [46]. While arguably more general and well-versed for discriminative and generative tasks [47], forecasting using signature kernels [48] comes at a price, as their nonlinear dependence on observation streams leads to a significant complexity increase compared to LTI predictors.

²Background on prerequisite Koopman operator theory can be found in the supplementary material.

Motivated by the restrictions of existing Koopman-based predictors, we propose a *function approximation* approach that exploits time-series data and Koopman operator theory to provably learn LTI predictors. Through a novel *invariance transform* we can satisfy **(KI)** by construction and directly minimize the forecasting risk over an entire time-interval **(OR)**. In simple terms: Koopman operator regression fixes $g(\cdot)$ and regresses \mathbf{A} and \mathbf{c} in (5), whereas our KKR approach selects \mathbf{A} to jointly regress \mathbf{c} and $g(\cdot)$. Similar in spirit to generalized Laplace analysis [21, 49], our approach allows the construction of eigenmodes from data without inferring the operator itself. Crucially, we demonstrate that selecting \mathbf{A} requires no prior knowledge as confirmed by our theoretical results and experiments. To facilitate learning LTI predictors, we derive *universal* RKHSs that are *guaranteed* to satisfy **(KI)** over trajectories — a first in the literature. The resulting equivalences to function regression in RKHS allow for more general and complete learning guarantees in terms of consistency and risk bounds that are free of restrictive operator-theoretic assumptions.

3 Koopman Kernel Regression

With the optimization (5) being prohibitively hard due to nonlinear and possibly high dimensional constraints, we eliminate the constraints (5b) by enforcing the feature map $g(\cdot)$ to have the dynamics of intrinsic LTI coordinates associated with Koopman operators, i.e., their (open) eigenfunctions [21].

Definition 1. A Koopman eigenfunction $\phi_{\lambda} \in C(\mathbb{X})$ satisfies $\phi_{\lambda}(\mathbf{x}) = e^{-\lambda t} \phi_{\lambda}(\mathbf{F}^t(\mathbf{x}))$, $\forall t \in [0, T]$.

It is proven that Koopman eigenfunctions from Definition 1 are universal approximators of continuous functions [28] — making them a viable replacement for the feature map $g(\cdot)$ in (4). However, following their definition, it is evident that Koopman eigenfunctions are by no means arbitrary due to their inherent dependence on the dynamics’ flow. Using the well-established fact that Koopman operators compose a function with the flow, i.e., $\mathcal{K}^t g(\cdot) = g(\mathbf{F}^t(\cdot))$, it becomes evident the eigenfunctions from Definition 1 are (semi)group invariants, as they remain unchanged after applying $\{e^{-\lambda t} \mathcal{K}^t\}_{t=0}^T$. Thus, inspired by the seminal work of Hurwitz on constructing invariants [50], we can equivalently reformulate (5) as an unconstrained problem and jointly optimize over eigenfunctions³.

Lemma 1 (Invariance transform). Consider a function $g \in C(\mathbb{X}_0)$ over a set of initial conditions $\mathbb{X}_0 \subseteq \mathbb{X}$ that form a non-recurrent domain \mathbb{X}_T . The invariance transform \mathcal{I}_{λ}^T transforms g into an Koopman eigenfunction $\phi_{\lambda} \in C(\mathbb{X}_T)$ for (3a) with LTI dynamics described by $\lambda \in \mathbb{C}$

$$\phi_{\lambda}(\mathbf{x}_T) = \mathcal{I}_{\lambda}^T g(\mathbf{x}_0) := \int_{\tau=0}^T e^{-\lambda(\tau-t)} g(\mathbf{F}^{\tau}(\mathbf{x}_0)) d\tau. \quad (7)$$

The above Lemma 1 is a key stepping stone towards deriving a representer theorem for LTI predictors. However, it is also interesting in its own right as it provides an explicit expression for the flow of an eigenfunction from any point in the state space. Thus, it provides a recipe to obtain a function space that fulfills **(KI)** by construction. As we show in the following, a sufficiently rich set of eigenvalues [28] and Lemma 1 will allow for a reformulation of (5) into an unconstrained problem

$$\min_M \|y_T - M(\mathbf{x}_T)\|_{\mathbb{Y}_T}. \quad (8)$$

where the operator $M(\cdot) := \mathbf{1}^{\top} [\phi_{\lambda_1}(\cdot) \cdots \phi_{\lambda_D}(\cdot)]^{\top}$ is universal and consisting of Koopman-invariant functions.

3.1 Functional Regression Problem

Notice that the problem reformulation (8) is still intractable, as a closed-form expression for the flow map is generally unavailable even for known ODEs. This requires integration schemes that can introduce inaccuracies over a time interval $[0, T]$. Thus, to make the above optimization problem tractable, data samples are used — ubiquitous in learning dynamical systems.

Assumption 1. A collection of N pairs of trajectories $\mathbb{D}_N = \{\mathbf{x}_T^{(i)}, y_T^{(i)}\}_{i=1}^N \in (\mathbb{X}_T \times \mathbb{Y}_T)^N$ is available.

By aggregating different invariance transformations (7) into the *mode decomposition operator*

$$M(\cdot) \equiv \sum_{j=1}^{\bar{D}} \phi_{\lambda_j}(\cdot): \mathbb{X}_T \mapsto \mathbb{Y}_T, \quad (9)$$

³Proofs for all theoretical results can be found in the supplementary material.

we can formulate a supervised learning approach in the following.

Learning Problem With Assumption 1 and Lemma 1, the sample-based approximation of problem (8) reduces to solving

$$\min_M \sum_{i=1}^N \|y_T^{(i)} - M(\mathbf{x}_T^{(i)})\|_{\mathbb{Y}_T}. \quad (10)$$

while preserving the mode decomposition structure (9). To realize the above learning problem, we resort to the theory of reproducing kernels [51, 52] and look for an operator $\hat{M} \in \mathcal{H}$, where \mathcal{H} is an RKHS. A well-established approach using RKHS theory is to select \hat{M} as a solution to the *regularized least squares problem*

$$\hat{M} = \arg \min_{M \in \mathcal{H}} \sum_{i=1}^N \|y_T^{(i)} - M(\mathbf{x}_T^{(i)})\|_{\mathbb{Y}_T}^2 + \gamma \|M\|_{\mathcal{H}}^2, \quad (11)$$

with $\gamma \in \mathbb{R}_+$ and $\|\cdot\|_{\mathcal{H}}$ a corresponding RKHS norm. As our target is a function-valued mapping $M(\cdot)$ – an operator – $\|\cdot\|_{\mathcal{H}}$ is induced by an *operator-valued kernel* $K: \mathbb{X}_T \times \mathbb{X}_T \mapsto \mathcal{B}(\mathbb{Y}_T)$ mapping to the space of bounded operators over the output space [53]. The salient feature of the above formulation (11) is its well-posedness: its solution exists and is unique for any \mathcal{H} , expressed as

$$\hat{M}(\cdot) = \sum_{i=1}^N K(\cdot, \mathbf{x}_T^{(i)}) \beta_i, \quad \beta_i \in \mathbb{Y}_T \quad (12)$$

through a representer theorem [54]. Still, due to the Koopman-invariant structure (9) from Lemma 1, the choice of the RKHS \mathcal{H} for \hat{M} is not arbitrary. Thus, the question is how to craft \mathcal{H} so the solution \hat{M} is decomposable into Koopman operator eigenfunctions (9), forming an *LTI predictor*.

Firstly, it is obvious that (9) consists of summands that may lie in different RKHS, denoted as $\{\mathcal{H}^{\lambda_j}\}_{j=1}^{\bar{D}}$. Then, \mathcal{H} is constructed from the following direct sum of Hilbert spaces [55]:

$$\tilde{\mathcal{H}} = \mathcal{H}^{\lambda_1} \oplus \dots \oplus \mathcal{H}^{\lambda_{\bar{D}}} \quad \text{so that} \quad \mathcal{H} = \text{range}(\mathcal{S}) := \{f_1 + \dots + f_{\bar{D}} : f_1 \in \mathcal{H}^{\lambda_1}, \dots, f_{\bar{D}} \in \mathcal{H}^{\lambda_{\bar{D}}}\} \quad (13)$$

with $\mathcal{S}: \tilde{\mathcal{H}} \rightarrow \mathcal{H}, (f_1 \dots f_{\bar{D}}) \mapsto f_1 + \dots + f_{\bar{D}}$ the summation operator [56]. Thus, to construct \mathcal{H} , a specification of the RKHS collection $\{\mathcal{H}^{\lambda_j}\}_{j=1}^{\bar{D}}$ is required, so that it represents Koopman eigenfunctions from (9).

Theorem 1 (Koopman eigenfunction kernel). *Consider trajectory data $\{\mathbf{x}_T^{(i)}\}_{i=1}^N$ from Assumption 1, a $\lambda \in \mathbb{C}$ and a universal (base) kernel $k: \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$. Then, the kernel $K^\lambda: \mathbb{X}_T \times \mathbb{X}_T \mapsto \mathcal{B}(\mathbb{Y}_T)$*

$$K^\lambda(\mathbf{x}_T, \mathbf{x}'_T) = \int_{\tau=0}^T \int_{\tau'=0}^T e^{-\lambda(\tau-t)} k(\mathbf{x}_T(\tau), \mathbf{x}'_T(\tau')) e^{-\lambda^*(\tau'-t)} d\tau d\tau', \quad (14)$$

- (i) defines an RKHS \mathcal{H}^λ ,
- (ii) is universal for every eigenfunction of Definition 1 corresponding to λ ,
- (iii) induces a data-dependent function space $\text{span}\{K^\lambda(\cdot, \mathbf{x}_T^{(1)}), \dots, K^\lambda(\cdot, \mathbf{x}_T^{(N)})\}$ that is Koopman-invariant over trajectory-data $\{\mathbf{x}_T^{(i)}\}_{i=1}^N$.

In Theorem 1, we derive an eigenfunction RKHS by defining its corresponding kernel that embeds the invariance transformation (7) over data samples. Also, we would like to highlight that the above result addresses a long-standing open challenge in the Koopman operator community [19–21], i.e., defining universal function spaces that are guaranteed to be Koopman-invariant. Now, we are ready to introduce the *Koopman kernel* as the kernel obtained by combining “eigen-RKHS” as described in (13).

Proposition 1 (Koopman kernel). *Consider trajectory data \mathbb{D}_N of Assumption 1 and a set of kernels $\{K^{\lambda_j}\}_{j=0}^{\bar{D}}$ from Theorem 1. Then, the kernel $K: \mathbb{X}_T \times \mathbb{X}_T \mapsto \mathcal{B}(\mathbb{Y}_T)$ given by*

$$K(\mathbf{x}_T, \mathbf{x}'_T) = \sum_{j=1}^{\bar{D}} K^{\lambda_j}(\mathbf{x}_T, \mathbf{x}'_T) \quad (15)$$

- (i) defines an RKHS $\mathcal{H} := \mathcal{S}(\mathcal{H}^{\lambda_1} \oplus \dots \oplus \mathcal{H}^{\lambda_{\bar{D}}})$,
- (ii) is universal for any output (3b), provided a sufficient amount⁴ of eigenspaces \bar{D} .

Above, we have derived the “Koopman-RKHS” \mathcal{H} for solving the problem (11) with a universal RKHS spanning Koopman eigenfunctions. Thus, the sample-data solution for an eigenfunction flow follows from the functional regression problem (11) and takes the form $\phi_{\lambda_j}(\cdot) = \sum_{i=1}^N K^{\lambda_j}(\cdot, \mathbf{x}_T^{(i)}) \beta_i, \beta_i \in \mathbb{Y}_T$ — providing a basis for the LTI predictor.

⁴Sufficient amount is a rich enough set of eigenvalues $\{e^{\lambda_j[0,T]}\}_{j=1}^{\bar{D}}$ from $\overline{\mathbb{B}_1(\mathbf{0})}$ in \mathbb{C} [57, Theorem 3.0.2].

3.2 Practicable LTI Predictor Regression

As a functional approximation problem, the solution of (11) is not parameterized by vector-valued coefficients, but rather functions of time. Although there are a few options to deal with function-valued solutions [53], we consider a vector-valued solution. A common drawback of such a discretization involves the loss of the inter-sample relations along the continuous signal. Crucially, this problem does not apply in our case, as the inter-sample relationships remain modeled for the discrete-time “Koopman kernel” due to its causal structure. Importantly, the vector-valued solution allows us to preserve all of the desirable properties derived in the continuous case.

Consider sampling $[0, T]$ at $H=T/\Delta t$ regular intervals to yield a discrete-time dataset from Assumption 1, discretized at points $\mathbb{H} \equiv \{t_0 \cdots t_H\}$. As a discretization of a function over time, with a slight abuse of notation, we denote the target vectors as $\mathbf{y}_H = [y(t_0) \cdots y(t_H)]^\top$. Thus, we are solving the time- and data-discretized version of the problem (5) that takes the form of a linear coregionalization model [58, 59].

Corollary 1 (Time-discrete Koopman kernel). *Consider trajectory data $\{\mathbf{x}_H^{(i)}\}_{i=1}^N$ and let $\mu_j := e^{\lambda_j \Delta t}$, $\boldsymbol{\mu}_j^\top := [\mu_j^0 \cdots \mu_j^H]$. Then, the scalar-induced matrix kernel $\mathbf{K}^{\mu_j}: \mathbb{X}_H \times \mathbb{X}_H \mapsto \mathcal{B}(\mathbb{Y}_H)$*

$$\mathbf{K}^{\mu_j}(\mathbf{x}_H, \mathbf{x}_H') = \boldsymbol{\mu}_j \boldsymbol{\mu}_j^{*\top} \underbrace{\frac{1}{(H+1)^2} \sum_{m=0}^H \sum_{n=0}^H \mu_j^{-m} k^j(\mathbf{x}_H(t_m), \mathbf{x}_H'(t_n)) \mu_j^{*-n}}_{k^{\mu_j}(\mathbf{x}_H, \mathbf{x}_H')}, \quad (16)$$

satisfies the properties (i)–(iii) from Theorem 1 over \mathbb{H} , so that it defines an RKHS \mathcal{H}^{μ_j} , is universal per Definition 1 over \mathbb{H} with $\text{span}\{\mathbf{K}^{\mu_j}(\cdot, \mathbf{x}_H^{(1)}), \dots, \mathbf{K}^{\mu_j}(\cdot, \mathbf{x}_H^{(N)})\}$ (KI) over $\{\mathbf{x}_H^{(i)}\}_{i=1}^N$. Given a collection of kernels $\{\mathbf{K}^{\mu_j}\}_{j=0}^D$, the matrix Koopman kernel $\mathbf{K}^{\mu_j}: \mathbb{X}_H \times \mathbb{X}_H \mapsto \mathcal{B}(\mathbb{Y}_H)$

$$\mathbf{K}(\mathbf{x}_H, \mathbf{x}_H') = \sum_{j=1}^D \mathbf{K}^{\mu_j}(\mathbf{x}_H, \mathbf{x}_H'), \quad (17)$$

satisfies the properties (i)–(ii) from Proposition 1 over \mathbb{H} , defining RKHS $\mathcal{H}^{\Delta t} := \mathcal{S}(\mathcal{H}^{\mu_1} \oplus \cdots \oplus \mathcal{H}^{\mu_D})$.

Now, we are fully equipped to obtain the time-discrete solution to our initial problem (5) provided a dataset of trajectories. Before presenting the solution to Koopman Kernel Regression, we introduce some helpful shorthand notation. We use the following kernel matrix abbreviations: $k_{\mathbf{X}\mathbf{X}} = [k(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})]_{a,b=1}^N$, $k(\mathbf{x}, \mathbf{X}) = [k(\mathbf{x}, \mathbf{x}^{(b)})]_{b=1}^N$, $\mathbf{K}_{\mathbf{X}\mathbf{X}} = [\mathbf{K}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})]_{a,b=1}^N$ and $\mathbf{K}(\mathbf{x}, \mathbf{X}) = [\mathbf{K}(\mathbf{x}, \mathbf{x}^{(b)})]_{b=1}^N$.

Proposition 2 (KKR). *Consider a discrete-time dataset of Assumption 1, $\mathbb{D}_N^{\Delta t} = \{\mathbf{x}_H^{(i)}, \mathbf{y}_H^{(i)}\}_{i=1}^N$, and let $\mathbf{y}_H^\top = [y_H^{(1)\top} \cdots y_H^{(N)\top}]$ with \otimes the Kronecker product. Then,*

$$\boldsymbol{\alpha}_j = k_{\mathbf{X}_0 \mathbf{X}_0}^{-1} k_{\mathbf{X}_H \mathbf{X}_H}^{\mu_j} (\mathbf{I}_N \otimes \boldsymbol{\mu}_j^{*\top}) \boldsymbol{\beta}, \quad \underline{\boldsymbol{\beta}} = (\mathbf{K}_{\mathbf{X}_H \mathbf{X}_H} + \gamma \mathbf{I}_{H+1} \otimes \mathbf{I}_N)^{-1} \mathbf{y}_H \quad (18)$$

defines a unique time-sampled solution to (11) in terms of eigenfunctions $\hat{\phi}(\mathbf{x}_0) = [k_{\mathbf{x}_0 \mathbf{X}_0}^j \boldsymbol{\alpha}_j]_{j=1}^D$, determining an LTI predictor⁵ with $\boldsymbol{\Lambda} = \text{diag}([\mu_1 \cdots \mu_D])$,

$$\mathbf{z}^+ = \boldsymbol{\Lambda} \mathbf{z}, \quad \mathbf{z}_0 = \hat{\phi}(\mathbf{x}_0), \quad (19a)$$

$$\hat{\mathbf{y}} = \mathbf{1}^\top \mathbf{z}. \quad (19b)$$

Notice how in (18), we re-scale the trajectory domain to that of the state-space. This enables us to write the forecast of (19), with a slight abuse of notation, using an extended observability matrix [60]

$$\hat{\mathbf{y}}_H = \boldsymbol{\Gamma} \hat{\phi}(\mathbf{x}_0), \quad \boldsymbol{\Gamma} := [\mathbf{1}^\top \mathbf{1}^\top \boldsymbol{\Lambda} \cdots \mathbf{1}^\top \boldsymbol{\Lambda}^H]^\top. \quad (20)$$

The confinement to a non-recurrent domain plays a crucial role in making the base kernel RKHSs isometric to “eigen-RKHSs” $\mathcal{H}^{k^j} \cong \mathcal{H}^{k^{\mu_j}}$ via invariance transforms, guaranteeing a feasible return from the time-series domain \mathbb{X}_T to the state-space domain $\mathbb{X}_0 \subseteq \mathbb{X}$ for evaluating the model over initial conditions.

Remark 1. *The salient feature of our proposed KKR framework compared to existing methods is the fact that Koopman-invariance (KI) over data samples is independent from the outcome of an optimization algorithm, e.g. minimizing the forecasting risk to compute $\boldsymbol{\beta}$ in (18). Thus, we are able to directly optimize for a downstream task (forecasting) (OR) given a suitably rich set of eigenvalues.*

⁵For discrete-time predictors, we omit the time-step specification and denote the next state with “ $(\cdot)^+$ ”.

3.3 Selecting Eigenvalues

Until now, we have used the sufficient cardinality $\bar{D} \in \mathbb{N}$ of an eigenvalue set that encloses [61] or is the true spectrum. However, we have provided no insight regarding the selection of \bar{D} spectral components or how they can be estimated. Here, we go beyond the learning-independent and non-constructive existence result of [28] and provide a consistency guarantee and relate it to sampling eigenvalues without the knowledge of the true spectrum.

Proposition 3. *Consider the oracle Koopman kernel $\mathbf{K}(\mathbf{x}_H, \mathbf{x}_H')$ and a dense set $\{\mu_j\}_{j=1}^\infty$ in $\overline{\mathbb{B}_1(\mathbf{0})}$. Then, $\|\mathbf{K}(\mathbf{x}_H, \mathbf{x}_H') - \sum_{j=1}^D \mathbf{K}^{\mu_j}(\mathbf{x}_H, \mathbf{x}_H')\|_{\mathcal{B}(\mathbb{Y}_H)} \rightarrow 0, \forall \mathbf{x}_H, \mathbf{x}_H' \in \mathbb{X}_H$ as $D \rightarrow \infty$.*

As shown in Proposition 3, even if we do not know the *oracle* kernel, we can arbitrarily approximate it by sampling from a dense set supported on the closed complex unit disk $\overline{\mathbb{B}_1(\mathbf{0})}$ [57, Theorem 3.0.2] with the error vanishing in the limit $D \rightarrow \infty$. There is no loss of generality when considering the unit disk as any finite radius disk can be scaled in the interval $[0, T]$. Furthermore, approximation of the oracle kernel by sampling a distribution over $\overline{\mathbb{B}_1(\mathbf{0})}$ leads to an almost sure $\mathcal{O}(1/\sqrt{D})$ convergence rate. It is conceivable that faster rates can be obtained in practice by including prior knowledge to shape the spectral distribution, e.g. using well-known concepts such as leverage-scores or subspace orthogonality [62, 63]. Based on spectral priors one can include a more biased sampling technique by precomputing components of the operator spectrum, e.g. computing Fourier averages [64], to determine the phases ω_j of complex-conjugate pairs $\mu_{j,\pm} = |\mu_j| e^{\pm i\omega_j}$ and sample the modulus from another physics-informed distribution. However, rigorous considerations of optimized and efficient sampling are beyond the scope of this paper and rather a topic of future work.

3.4 Numerical Algorithm and Time-Complexity

Algorithm 1 Regression and LTI Forecasts using KKR

Data $\mathbb{D} = \{\mathbf{x}_H^{(i)}, y_H^{(i)}\}_{i=1}^N$, Eigenvalues $\{\mu_j\}_{j=1}^D$
function REGRESS($\mathbb{D}, \{\mu_j\}_{j=1}^D$)
 form Gramians $k_{\mathbf{X}_0 \mathbf{X}_0}, \{k_{\mathbf{X}_H \mathbf{X}_H}^{\mu_j}\}_{j=1}^D, \mathbf{K}_{\mathbf{X}_H \mathbf{X}_H}$
 fit mode operator $\hat{M}(\cdot): \mathbb{X}_H \mapsto \mathbb{Y}_H$ (18, right)
 recover eigenfunctions $\hat{\phi}(\cdot): \mathbb{X}_0 \mapsto \mathbb{Z}_0$ (18, left)
 construct $\Gamma: \mathbb{Z}_0 \mapsto \mathbb{Y}_H$ (20, right)
 return LTI predictor $\Gamma \hat{\phi}(\cdot): \mathbb{X}_0 \mapsto \mathbb{Y}_H$
end function
function FORECAST(\mathbf{x}_0)
 “lift” $\mathbf{z}_0 = \hat{\phi}(\mathbf{x}_0)$
 rollout $\hat{y}_H = \Gamma \mathbf{z}_0$
 return trajectory \hat{y}_H
end function

For a better overview, the pseudocode for regression and forecasting of our method are shown in Algorithm 1. We also put the time-complexity of our algorithm into perspective w.r.t. Koopman operator regression of PCR/RRR [23] and ridge regression using state-of-the-art signature kernels [48] (RR-Sig-PDE) in Table 1. The training complexity of our KKR is comparable to that of RR-Sig-PDE regression and generally better than that of PCR/RRR. Given that accurate LTI forecasts require higher-rank predictors, the seemingly mild quadratic dependence makes $D^2 > NH$ and leads to a more costly matrix inversion. Furthermore, our LTI predictor also has a slightly better forecast complexity due to

not depending on trajectory length. Obviously, due to a mere matrix multiplication after an initial nonlinear map, LTI predictors have a significantly lower evaluation complexity than the nonlinear predictor of Sig-PDE’s. Due to requiring updated observation sequences as inputs, Sig-PDE kernels introduce a raw evaluation complexity that is also quadratic in sequence length.

Method	Training	H -step forecast	N	H	D	l	d
KKR (ours)	$\mathcal{O}(N^3 H^3 + DN^2 H^2 d)$	$\mathcal{O}(DH + DNd)$					
PCR/RRR	$\mathcal{O}(D^2 N^2 H^2 + N^2 H^2 d)$	$\mathcal{O}(DH + DNHd)$					
RR-Sig-PDE	$\mathcal{O}(N^3 H^3 + N^2 H^2 l^2 d)$	$\mathcal{O}(NH^2 l^2 d)$					

Table 1:
Time complexities.

4 Learning Guarantees

With a completely defined KKR estimator, we assess its essential learning-theoretic properties, i.e., the behavior of the learned functions w.r.t. to the ground truth with increasing dataset size.

4.1 Consistency

Although well-established in most function approximation settings [65–67], the setting of Koopman-based LTI predictor learning for nonlinear systems is void of consistency guarantees. Here we use a definition of universal consistency from [68] that describes the uniform convergence of the learned function to the target function as the sample size goes to infinity for any compact input space \mathbb{X} and every target function $q \in C(\mathbb{X})$. The existing convergence results for Koopman-based LTI predictors [69] are in the sense of strong operator topology — allowing the existence of empirical eigenvalues that are not guaranteed to be close to true ones even with increasing data [70]. This lack of spectral convergence has a cascaded effect in Koopman operator regression as, in turn, the convergence of eigenfunctions and mode coefficients is not guaranteed. Here, the convergence of modes is replaced by the convergence of eigenfunctions, and convergence of spectra is replaced by the convergence of (20) to the mode decomposition operator $\hat{M} \equiv \Gamma \hat{\phi} \rightarrow M \equiv \Gamma \phi$ with the estimate denoted by $(\hat{\cdot})$.

Theorem 2 (Universal consistency). *Consider a universal kernel \mathbf{K} (17) and a data distribution supported on $\mathbb{X}_H \times \mathbb{Y}_H$. Then, as $N \rightarrow \infty$, $\|M - \hat{M}\|_{\mathbb{Y}_H} \rightarrow 0$ and $\|\phi_{\mu_j} - \hat{\phi}_{\mu_j}\|_{\mathbb{Y}_H} \rightarrow 0, \forall j=1, \dots, D$.*

4.2 Generalization Gap: Uniform Bounds

Due to formulating the LTI predictor learning problem as a function regression problem in an RKHS, we can utilize well-established concepts from statistical learning to provide bounds on the generalization capabilities of KKR. Given a dataset of trajectories, the following *empirical risk* is minimized

$$\hat{\mathcal{R}}_N(\hat{M}) := \frac{1}{N} \sum_{i \in [N]} \|y_H^{(i)} - \hat{M}(\mathbf{x}_H^{(i)})\|_{\mathbb{Y}_H}^2$$

which is “in-sample” mean square error (MSE) w.r.t. a trajectory-data generating distribution ρ_D of i.i.d. initial conditions. The *true risk*/generalization error of an estimator is the “out-of-sample” MSE of the model on the entire domain and denoted as $\mathcal{R}(\cdot)$. Those quantities are, in essence, the model’s performance on test and training data, respectively. Allowing for statements on the test performance with an increasing amount of data by means of training performance is a desirable feature in data-driven learning. Hence, we analyze our model in terms of the *generalization gap*

$$|\mathcal{R}(\hat{M}) - \hat{\mathcal{R}}_N(\hat{M})| = \left| \mathbb{E}_{(\mathbf{x}_H, y_H) \sim \rho_D} [\|y_H - \hat{M}(\mathbf{x}_H)\|_{\mathbb{Y}_H}^2] - \frac{1}{N} \sum_{i=1}^N \|y_H^{(i)} - \hat{M}(\mathbf{x}_H^{(i)})\|_{\mathbb{Y}_H}^2 \right|. \quad (21)$$

To ensure a well-specified problem, we require models in the hypothesis to admit a bounded norm.

Assumption 2 (Bounded RKHS Norm). *The unknown function M has a bounded norm in the RKHS $\mathcal{H}^{\Delta t}$ attached to the Koopman kernel $\mathbf{K}(\cdot, \cdot)$, i.e., $\|M\|_{\mathcal{H}^{\Delta t}} \leq B$ for some $B \in \mathbb{R}_+$.*

The above smoothness assumption is mild, e.g., satisfied by band-limited continuous trajectories [71] and computable from data [72, 73]. In stark contrast, well-specified Koopman operator regression [23] requires the operator to map the RKHS onto itself, which is a very strong assumption [34, 35].

To derive the main result of this section, we utilize the framework of Rademacher random variables for measuring complexity of our model’s hypothesis space, a concept generally explored in [74] and more particularly for classes of operator-valued kernels in [75]. Conveniently, the derivation is, in terms of the RKHS $\mathcal{H}^{\Delta t}$, similar to standard methods on RKHS-based complexity bounds [74]. We use well-known results based on concentration inequalities to provide high probability bounds on a model’s generalization gap in terms of those complexities. Finally, we upper bound any constant with quantities specified in our assumptions and can state the following result.

Theorem 3 (Generalization Gap of KKR). *Let $\mathbb{D}_N^{\Delta t} = \{\mathbf{x}_H^{(i)}, y_H^{(i)}\}_{i=1}^N$ be a dataset as in Assumption 1 consistent with a Lipschitz system on a non-recurrent domain. Then the generalization gap (21) of a model \hat{M} from Proposition 2 under Assumption 2 is, with probability $1 - \delta$, upper bounded by*

$$|\mathcal{R}(\hat{M}) - \hat{\mathcal{R}}_N(\hat{M})| \leq 4RB \sqrt{\frac{\kappa H^2}{N}} + \sqrt{\frac{8 \log \frac{2}{\delta}}{N}} \in \mathcal{O}\left(\frac{H}{\sqrt{N}}\right), \quad (22)$$

where R is an upper bound on the loss in the domain, and κ the supremum of the base kernel.

We observe an overall dependence of order $\mathcal{O}(1/\sqrt{N})$ w.r.t. data points, resembling the regular Monte Carlo rate to be expected when working with Rademacher complexities. Remarkably, an increase in the order of the predictor D cannot widen the generalization gap but will eventually decrease the

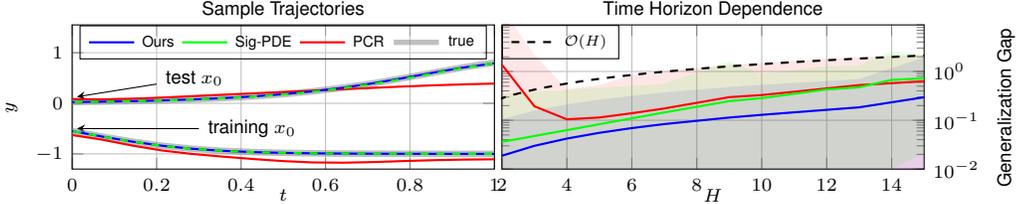


Figure 2: Forecasting performance (48 i.i.d. runs) for the bi-stable system for $H=14$ and $N=50$ for respectively optimal $D_{\text{KKR}}=100$, $D_{\text{PCR}}=10$ and 15 delays for Sig-PDEs. **Left:** Exemplary trajectories showing the advantage of learning with time-series kernels. **Right:** The generalization gap with an increasing forecast horizon, demonstrating generalization advantages of KKR.

empirical risk due to the consistency of eigenspaces (Proposition 3). Combined, our findings are a substantial improvement, both quantitatively and in terms of interpretability, over existing risk bounds on forecasting error [23, Theorem 1]. Additionally, our intuitive non-recurrence requirement is easily verifiable from data. In contrast, the Koopman operator regression in RKHS comes with various strong assumptions [35] that require commonly unavailable expert knowledge. Also, the generalization of existing Koopman-based statistical learning approaches depends on rank while ours is rank-independent. The significant implications of our results are demonstrated in the following.

5 Numerical Experiments

In our experiments⁶, we report the squared error of the forecast vector for the length of data trajectories averaged over multiple repetitions with corresponding min-max intervals. We validate our theoretical guarantees and compare to state-of-the-art operator and time-series approaches in RKHS. For fairness, the same kernel and hyperparameters are chosen for our KKR, PCR (EDMD), RRR [23] and regression with signature kernels (Sig-PDE) [48]. Note, PCR and RRR are provided with the same trajectory data split into one-step data pairs while the time and observation time-delays are fed as data to the Sig-PDE regressor due to its recurrent structure. Along with code for reproduction of our experiments, we provide a JAX [76] reliant Python module implementing a sklearn [77] compliant KKR estimator at <https://github.com/TUM-ITR/koopcore>.

Bi-stable system Consider an ODE $\dot{x} = ax + bx^3$ that arises in modeling of nonlinear friction. The parameters are $a = 4$, $b = -4$, making for a bi-stable system at fixed points ± 1 . The numerical results are depicted in Figure 2. Sample trajectories both on training and testing data indicate the utility of the forecast risk minimization of KKR. While EDMD correctly captures the initial trend of most trajectories it fails to match the accuracy of Sig-PDE or our KKR predictors that utilize time-series structure. Furthermore, the behavior of KKR’s generalization gap for an increasing time horizon $T = H\Delta t$, $\Delta t = 1/14s$ closely matches our theoretical analysis.

Van der Pol oscillator Consider an ODE $\ddot{x} = \dot{x}(2 - 10x^2) - 0.8x$ describing a dissipative system whose nonlinear damping induces a stable limit cycle — a phenomenon present in various dynamics. In Figure 3 two fundamental effects are validated: the generalization gap with increasing data and consistency with test risk that does not deteriorate for increasing eigenspace cardinality. The performance of PCR/RRR is strongly tied to predictor rank while Sig-PDE’s less so w.r.t. delay length.

Table 2: Average risk (20 runs) $[\times 10^{-2}]$ for Van der Pol for various *spectral sampling* and *lengthscales*, $N=200$, $H=14$.

$\rho(\mu)$	uniform		boundary-biased		physics-informed	
	16	200	16	200	16	200
$\mathcal{R}_{\ell=10^1}$	13.7	5.38	11.2	5.38	5.60	5.58
$\mathcal{R}_{\ell=10^0}$	6.46	0.78	4.10	0.78	0.97	0.92
$\mathcal{R}_{\ell=10^{-1}}$	7.12	1.74	4.33	1.74	1.83	1.80

Crucially, our KKR approach does not require a careful choice of the eigenspace cardinality to perform for a specific amount of data. Although the eigenvalues that determine the eigenspaces are randomly chosen from a uniform distribution in the unit ball, KKR consistently outperforms PCR/RRR. In Table 2 we show the spectral sampling and hyperparameter effects. We employ the following strategies: *uniform* - uniform distribution on the complex unit disk, *boundary-biased* - a distribution on the complex unit disk skewed towards the unit circle, *physics-informed* - eigenvalues of various vector field Jacobians. As expected, physics-informed performs well with lower rank

⁶Additional details on the numerical experiments can be found in the supplementary material.

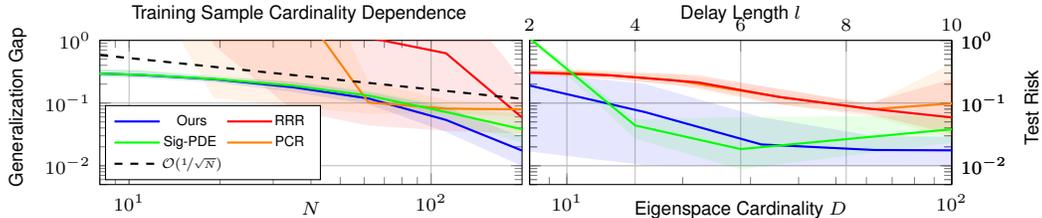


Figure 3: Forecasting risks (20 i.i.d. runs) for the Van der Pol system over a time-horizon $H = 14$ ($T = 1s$). **Left:** Generalization gap for the best D/l (ours 500, PCR 62, RRR 100, RR-Sig-PDE 10) is depicted with a growing number of data points. **Right:** Test risk behavior with an increasing amount of eigenspaces is shown for $N = 200$. Shaded areas depict min-max risk intervals.

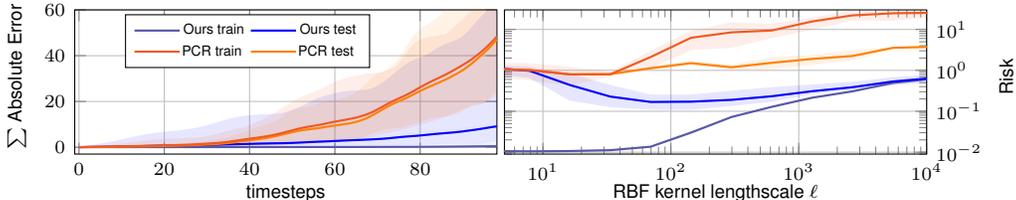


Figure 4: Cumulative error and forecast risks (5 train-test splits) for flow past cylinder data and $H = 99$. Our KKR with orders-of-magnitude greater usable ℓ -range and accuracy. **Left:** Cumulative absolute error for the best D/l (ours 200/70, PCR 200/35) is depicted over timesteps. **Right:** Forecast risk for 99 steps within a range of RBF lengthscales. Shaded areas depict min-max intervals.

compared to uninformed approaches. However, it is outperformed by unit-ball sampling approaches for higher rank due to a lack of coverage. Table 3 includes CPU timings for completeness.

#data = $N \times H = 200 \times 14$	KKR	PCR	RRR	Sig-PDE	Table 3: Computation times for Van der Pol.
Training [s]/Forecast [ms]	8.0/ 54	90/ 84	88/150	8.6/5900	

Flow past a cylinder We consider high-dimensional data of velocity magnitudes in a Kármán vortex street under varying initial cylinder placement, as illustrated in Figure 5. The cylinder position is varied on a 7×7 grid in a 50×100 -dimensional space and the flow is recorded over $H=99$. The quantity of interest is a velocity magnitude sensor placed in the wake of the cylinder. In forecasting from an initial velocity field, KKR outperforms PCR by orders-of-magnitude as shown in Figure 4. We omit Sig-PDE regression due to persistent divergence after ≈ 20 steps. The latter is hardly surprising, given that Sig-PDE models iterate one step predictions based only on the shapes of time-delays while LTI models directly output time-series from initial conditions.

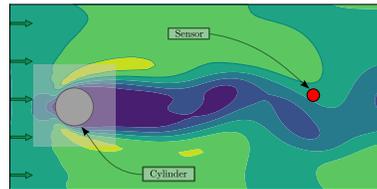


Figure 5: Flow illustration. Area of initial cylinder positions shaded.

6 Conclusion

We presented a novel statistical learning framework for learning LTI predictors using trajectories of a dynamical system. The method is rooted in the derivation of a novel RKHS over trajectories, which solely consists of universal functions that have LTI dynamics. Equivalences with function regression in RKHS allow us to provide consistency guarantees not present in previous literature. Another key contribution is a novel rank-independent generalization bound for i.i.d. sampled trajectories that directly describes forecasting performance. The significant implications of the proposed approach are confirmed in experiments, leading to superior performance compared to Koopman operator and sequential data predictors in RKHS. In this work, we confined our forecasts to a non-recurrent domain for a specific length of trajectory data, where the choice of spectra is arbitrary. However, exploring more efficacious spectral sampling schemes is a natural next step for extending our results to asymptotic regimes that include, e.g., periodic and quasi-periodic behavior. It has to be noted that vector-valued kernel methods have limited scalability with a growing number of training data and output dimensionality. Therefore, exploring solutions that improve scalability is an important topic for future work. Furthermore, to enable the use of LTI predictors in safety-critical domains, the quantification of the forecasting error is essential. Hence, deriving uniform prediction error bounds for KKR is of great interest.

Acknowledgements

The authors acknowledge the financial support of the EU Horizon 2020 research and innovation programme “SeaClear” (ID 871295) and the ERC Consolidator grant “CO-MAN” (ID 864686). Petar Bevanda also thanks Jan Brüdigam for feedback, and Vladimir Kostić and Pietro Novelli for useful discussions on operator regression.

References

- [1] J. L. Meriam, L. G. Kraige, and J. N. Bolton, *Engineering Mechanics: Dynamics*. John Wiley & Sons, 2020.
- [2] A. Billard, S. Mirrazavi, and N. Figueroa, *Learning for Adaptive and Reactive Robot Control: A Dynamical Systems Approach*. MIT Press, 2022.
- [3] V. May and O. Kühn, *Charge and Energy Transfer Dynamics in Molecular Systems*. Wiley, 2011.
- [4] J. Johansson, P. Nation, and F. Nori, “QuTiP: An open-source Python framework for the dynamics of open quantum systems,” *Computer Physics Communications*, vol. 183, no. 8, pp. 1760–1772, 8 2012.
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.
- [7] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural Ordinary Differential Equations,” in *Advances in Neural Information Processing Systems*, 2018.
- [8] M. Biloš, J. Sommer, S. S. Rangapuram, T. Januschowski, and S. Günnemann, “Neural Flows: Efficient Alternative to Neural ODEs,” *Advances in Neural Information Processing Systems*, 2021.
- [9] M. Janner, Q. Li, and S. Levine, “Offline Reinforcement Learning as One Big Sequence Modeling Problem,” in *Advances in Neural Information Processing Systems*, 2021, pp. 1273–1286.
- [10] K. Chua, R. Calandra, R. McAllister, and S. Levine, “Deep reinforcement learning in a handful of trials using probabilistic dynamics models,” in *Advances in Neural Information Processing Systems*, 2018.
- [11] E. A. Theodorou, J. Buchli, and S. Schaal, “A Generalized Path Integral Control Approach to Reinforcement Learning,” *Journal of Machine Learning Research*, vol. 11, pp. 3137–3181, 2010.
- [12] M. Budišić, R. Mohr, and I. Mezić, “Applied Koopmanism,” *Chaos*, vol. 22, no. 4, 10 2012.
- [13] A. Mauroy, I. Mezić, and Y. Susuki, *The Koopman Operator in Systems and Control*, ser. Lecture Notes in Control and Information Sciences. Cham: Springer International Publishing, 2020, vol. 484.
- [14] I. Mezić and A. Banaszuk, “Comparison of systems with complex behavior,” *Physica D: Nonlinear Phenomena*, vol. 197, no. 1, pp. 101–133, 2004.
- [15] I. Mezić, “Spectral Properties of Dynamical Systems, Model Reduction and Decompositions,” *Nonlinear Dynamics*, vol. 41, no. 1-3, pp. 309–325, 8 2005.

- [16] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, “Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control,” *PLOS ONE*, vol. 11, no. 2, p. e0150171, 2 2016.
- [17] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering*. Cambridge University Press, 1 2019.
- [18] S. E. Otto and C. W. Rowley, “Koopman Operators for Estimation and Control of Dynamical Systems,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, p. 2021, 2021.
- [19] P. Bevanda, S. Sosnowski, and S. Hirche, “Koopman operator dynamical models: Learning, analysis and control,” *Annual Reviews in Control*, vol. 52, pp. 197–212, 2021.
- [20] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz, “Modern Koopman Theory for Dynamical Systems,” *SIAM Review*, vol. 64, no. 2, pp. 229–340, 2022.
- [21] I. Mezić, “Spectrum of the Koopman Operator, Spectral Expansions in Functional Spaces, and State-Space Geometry,” *Journal of Nonlinear Science*, vol. 30, no. 5, pp. 2091–2145, 2020.
- [22] S. Klus, I. Schuster, and K. Muandet, “Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces,” *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 283–315, 2020.
- [23] V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, and M. Pontil, “Learning Dynamical Systems via Koopman Operator Regression in Reproducing Kernel Hilbert Spaces,” in *Advances in Neural Information Processing Systems*, 2022, pp. 4017–4031.
- [24] V. Kostic, K. Lounici, P. Novelli, and M. Pontil, “Koopman Operator Learning: Sharp Spectral Rates and Spurious Eigenvalues,” 2 2023. [Online]. Available: <http://arxiv.org/abs/2302.02004>
- [25] D. Angeli and E. D. Sontag, “Forward completeness, unboundedness observability, and their Lyapunov characterizations,” *Systems & Control Letters*, vol. 38, no. 4-5, pp. 209–217, 12 1999.
- [26] V. Andrieu and L. Praly, “On the existence of a Kazantzis-Kravaris/Luenberger observer,” *SIAM Journal on Control and Optimization*, vol. 45, no. 2, pp. 422–456, 2006.
- [27] M. Krstic, “Forward-Complete Systems,” in *Delay Compensation for Nonlinear, Adaptive, and PDE Systems*. Boston: Birkhäuser, 2009, pp. 171–190.
- [28] M. Korda and I. Mezić, “Optimal Construction of Koopman Eigenfunctions for Prediction and Control,” *IEEE Transactions on Automatic Control*, vol. 65, no. 12, pp. 5114–5129, 12 2020.
- [29] S. Klus, I. Schuster, and K. Muandet, “Eigendecompositions of transfer operators in reproducing kernel hilbert spaces,” *Journal of Nonlinear Science*, vol. 30, pp. 283–315, 2 2020.
- [30] M. J. Colbrook, L. J. Ayton, and M. Szöke, “Residual Dynamic Mode Decomposition: Robust and verified Koopmanism,” Tech. Rep.
- [31] M. Lewin and E. Sere, “Spectral pollution and how to avoid it,” *Proceedings of the London Mathematical Society*, vol. 100, no. 3, pp. 864–900, 5 2010.
- [32] M. Khosravi, “Representer theorem for learning koopman operators,” *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 2995–3010, 2023.
- [33] F. Philipp, M. Schaller, K. Worthmann, S. Peitz, and F. Nüske, “Error bounds for kernel-based approximations of the Koopman operator,” 1 2023. [Online]. Available: <http://arxiv.org/abs/2301.08637>
- [34] S. Das and D. Giannakis, “Koopman spectra in reproducing kernel Hilbert spaces,” *Applied and Computational Harmonic Analysis*, vol. 49, no. 2, pp. 573–607, 9 2020.
- [35] C. Valva and D. Giannakis, “Consistent spectral approximation of Koopman operators using resolvent compactification,” 9 2023. [Online]. Available: <http://arxiv.org/abs/2309.00732>

- [36] R. K. Singh and A. Kumar, “Compact Composition Operators,” *Journal of the Australian Mathematical Society*, vol. 28, no. 3, pp. 309–314, 1979.
- [37] M. Ikeda, I. Ishikawa, and Y. Sawano, “Boundedness of composition operators on reproducing kernel Hilbert spaces with analytic positive definite functions,” *Journal of Mathematical Analysis and Applications*, vol. 511, no. 1, p. 126048, 7 2022.
- [38] P. Cvitanović, R. Artuso, R. Mainieri, G. Tanner, and G. Vattay, *Chaos: Classical and Quantum*. Copenhagen: Niels Bohr Inst., 2016. [Online]. Available: <http://ChaosBook.org/>
- [39] I. Mezić, “On Numerical Approximations of the Koopman Operator,” *Mathematics*, vol. 10, no. 7, p. 1180, 4 2022.
- [40] P. Bevanda, J. Kirmayr, S. Sosnowski, and S. Hirche, “Learning the Koopman Eigendecomposition: A Diffeomorphic Approach,” in *2022 American Control Conference (ACC)*, 2022, pp. 2736–2741.
- [41] P. Bevanda, M. Beier, S. Kerz, A. Lederer, S. Sosnowski, and S. Hirche, “Diffeomorphically Learning Stable Koopman Operators,” *IEEE Control Systems Letters*, vol. 6, pp. 3427–3432, 2022.
- [42] E. M. Bollt, “Geometric considerations of a good dictionary for Koopman analysis of dynamical systems: Cardinality, “primary eigenfunction,” and efficient representation,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 100, 9 2021.
- [43] F. J. Király and H. Oberhauser, “Kernels for Sequentially Ordered Data,” *Journal of Machine Learning Research*, vol. 20, pp. 1–45, 2019.
- [44] T. J. Lyons, “Differential equations driven by rough signals.” *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.
- [45] P. K. Friz and N. B. Victoir, *Multidimensional Stochastic Processes as Rough Paths*. Cambridge University Press, 2 2010.
- [46] D. Lee and H. Oberhauser, “The Signature Kernel,” 5 2023. [Online]. Available: <http://arxiv.org/abs/2305.04625>
- [47] M. Lemerrier, C. Salvi, T. Cass, E. V. Bonilla, T. Damoulas, and T. J. Lyons, “SigGPDE: Scaling Sparse Gaussian Processes on Sequential Data,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 10 2021, pp. 6233–6242.
- [48] C. Salvi, T. Cass, J. Foster, T. Lyons, and W. Yang, “The Signature Kernel Is the Solution of a Goursat PDE,” *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 3, pp. 873–899, 1 2021.
- [49] I. Mezić, “Analysis of fluid flows via spectral properties of the koopman operator,” *Annual Review of Fluid Mechanics*, vol. 45, no. 1, pp. 357–378, 2013.
- [50] A. Hurwitz, “über die Erzeugung der Invarianten durch Integration,” *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, vol. 1897, pp. 71–72, 1897.
- [51] B. Schölkopf and A. J. Smola, *Learning with Kernels*. The MIT Press, 2018.
- [52] Ingo Steinwart and Andreas Christmann, *Support Vector Machines*, 1st ed., ser. Information Science and Statistics. New York, NY: Springer, 2008.
- [53] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren, “Operator-valued Kernels for Learning from Functional Response Data,” *Journal of Machine Learning Research*, vol. 17, no. 20, pp. 1–54, 2016.
- [54] C. A. Micchelli and M. Pontil, “On Learning Vector-Valued Functions,” *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.

- [55] N. Aronszajn, “Theory of Reproducing Kernels,” *Transactions of the American Mathematical Society*, vol. 68, no. 3, p. 337, 1950.
- [56] T. Hotz and F. J. E. Telschow, “Representation by Integrating Reproducing Kernels,” 2 2012. [Online]. Available: <http://arxiv.org/abs/1202.4443>
- [57] K. Küster, “The Koopman Linearization of Dynamical Systems,” 2015. [Online]. Available: <https://homepages.laas.fr/henrion/mfo16/kari-kuester.pdf>
- [58] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, “Kernels for vector-valued functions: A review,” pp. 195–266, 2011.
- [59] A. Lederer, A. Capone, T. Beckers, J. Umlauf, and S. Hirche, “The Impact of Data on the Stability of Learning-Based Control,” in *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, 2021, pp. 623–635.
- [60] S. J. Qin, “An overview of subspace identification,” *Computers and Chemical Engineering*, vol. 30, no. 10-12, pp. 1502–1513, 9 2006.
- [61] N. Dunford, “Spectral Theory. I Convergence to Projections,” *Transactions of the American Mathematical Society*, vol. 54, no. 2, p. 185, 9 1943.
- [62] F. Bach, “On the equivalence between kernel quadrature rules and random feature expansions,” *Journal of Machine Learning Research*, vol. 18, no. 21, pp. 1–38, 2017.
- [63] Z. Li, J.-F. Ton, D. Oglic, and D. Sejđinovic, “Towards a unified analysis of random fourier features,” *Journal of Machine Learning Research*, vol. 22, no. 1, 2021.
- [64] A. Mauroy and I. Mezić, “On the use of Fourier averages to compute the global isochrons of (quasi)periodic dynamics,” *Chaos*, vol. 22, no. 3, 7 2012.
- [65] A. Caponnetto and E. De Vito, “Optimal rates for the regularized least-squares algorithm,” *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 7 2007.
- [66] C. Carmeli, E. De Vito, and A. Toigo, “Vector Valued Reproducing Kernel Hilbert Spaces of Integrable Functions and Mercer Theorem,” *Analysis and Applications*, vol. 4, no. 4, pp. 377–408, 2006.
- [67] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità, “Vector Valued Reproducing Kernel Hilbert Spaces and Universality,” *Analysis and Applications*, vol. 08, no. 01, pp. 19–61, 1 2010.
- [68] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, “Universal Multi-Task Kernels,” *Journal of Machine Learning Research*, vol. 9, no. 52, pp. 1615–1646, 2008.
- [69] M. Korda and I. Mezić, “On Convergence of Extended Dynamic Mode Decomposition to the Koopman Operator,” *Journal of Nonlinear Science*, vol. 28, no. 2, pp. 687–710, 4 2018.
- [70] J. A. Rosenfeld, R. Kamalapurkar, L. F. Gruss, and T. T. Johnson, “Dynamic mode decomposition for continuous time systems with the liouville operator,” *Journal of Nonlinear Science*, vol. 32, p. 5, 2 2022.
- [71] M. Kanagawa, P. Hennig, D. Sejđinovic, and B. K. Sriperumbudur, “Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences,” 7 2018. [Online]. Available: <http://arxiv.org/abs/1807.02582>
- [72] P. Scharnhorst, E. T. Maddalena, Y. Jiang, and C. N. Jones, “Robust Uncertainty Bounds in Reproducing Kernel Hilbert Spaces: A Convex Optimization Approach,” 4 2021. [Online]. Available: <http://arxiv.org/abs/2104.09582>
- [73] B. C. Csaji and B. Horvath, “Nonparametric, Nonasymptotic Confidence Bands With Paley-Wiener Kernels for Band-Limited Functions,” *IEEE Control Systems Letters*, vol. 6, pp. 3355–3360, 2022.
- [74] P. L. Bartlett and S. Mendelson, “Rademacher and Gaussian Complexities: Risk Bounds and Structural Results,” *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 3 2002.

- [75] R. Huusari and H. Kadri, “Entangled Kernels - Beyond Separability,” *Journal of Machine Learning Research*, vol. 22, no. 24, pp. 1–40, 2021.
- [76] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, “JAX: composable transformations of Python+NumPy programs,” 2018. [Online]. Available: <http://github.com/google/jax>
- [77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.