# Koopman-Equivariant Gaussian Processes

**Petar Bevanda*** TU Munich
**Max Beier*** TU Munich
**Alex Capone** CMU Robotics Institute
**Stefan Sosnowski** TU Munich
**Sandra Hirche** TU Munich
**Armin Lederer** ETH Zürich

## Abstract

We propose a family of Gaussian processes (GP) for dynamical systems with linear time-invariant responses, which are nonlinear only in initial conditions. This linearity allows us to tractably quantify forecasting and representational uncertainty, simultaneously alleviating the challenge of computing the distribution of trajectories from a GP-based dynamical system and enabling a new probabilistic treatment of learning Koopman operator representations. Using a trajectory-based equivariance – which we refer to as *Koopman equivariance* – we obtain a GP model with enhanced generalization capabilities. To allow for large-scale regression, we equip our framework with variational inference based on suitable inducing points. Experiments demonstrate on-par and often better forecasting performance compared to kernel-based methods for learning dynamical systems.

## 1 INTRODUCTION

Learning predictive models for forecasting dynamic systems is a challenging task due to complex and often unknown interactions between quantities of interest (Brunton and Kutz, 2019). The great utility of such models helps advance various different fields such as fluid mechanics (Kundu et al., 2015), molecular biology (Lindorff-Larsen et al., 2011), robotics (Billard et al., 2022) or safety-constrained decision making (Hewing et al., 2020b; Brunke et al., 2022). Dynamical system descriptions commonly require simulation for forecasting and uncertainty propagation, which can be difficult for non-parametric data-driven models (Hewing et al., 2020a; Beckers and Hirche, 2022). In most real-world applications involving dynamical systems, measurements often come in the form of sequential one-step transition data that is sampled arbitrarily and potentially non-uniformly. Furthermore, there is often a certain regularity in the evolution of quantities of interest (Biloš et al., 2023) across domains (Sezer et al., 2020; Deb et al., 2017; Lim and Zohren, 2021), making it important to impose structure that discourages temporal fluctuations. To account for these different challenges in modeling dynamical systems, the choice of *representations* when learning from data becomes a deciding factor in the difficulty of forecasting as well as inference, especially when modeling complex phenomena (Mezić and Banaszuk, 2004) or long time-series (Gu et al., 2022). In this paper, we focus on non-parametric learning paradigms, emphasizing *uncertainty quantification* and *forecasting simplicity*. In particular, we study the interplay between Gaussian processes (Rasmussen and Williams, 2006) and effective dynamical system linearizations based on Koopman operators (Mauroy et al., 2020; Brunton et al., 2022). A more exhaustive account of related work is delegated to the supplementary material.

Table 1: Nonlinear dynamics modeling from data

| Approach | LTI forecast | End-to-end | Bayesian |
|---|---|---|---|
| GP | ✗ | ✓ | ✓ |
| Koopman | ✓ | ✗ | ✗ |
| **KE-GP** (ours) | ✓ | ✓ | ✓ |

**Gaussian processes.** Gaussian processes (GPs) (Rasmussen and Williams, 2006) have the capability of inferring models with little structural prior knowledge: either by using so-called universal kernels (Micchelli et al., 2006) or placing a prior on a set of kernels and optimizing their likelihood (Duvenaud, 2014). In particular, their ability to quantify epistemic uncertainty has led to a common application in safety-critical control

*Equal contribution. Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

problems (Berkenkamp and Schoellig, 2015; Sui et al., 2015; Berkenkamp et al., 2017; Curi et al., 2022; Baumann et al., 2021; Khosravi et al., 2023; As et al., 2024; Polymenakos et al., 2020; Lederer et al., 2021). Commonly employed as single-step predictors, GP models necessitate approximations for predicting probability distributions that go beyond a single time-step into the future. Thus, dealing with multi-step prediction often relies on iterative sampling-based approaches (Bradford et al., 2020; Hewing et al., 2020a; Beckers and Hirche, 2022) that are generally computationally expensive. Alternatively, one can employ methods of reduced computational complexity, such as Taylor approximations (Girard et al., 2003) or exact moment matching (Deisenroth and Rasmussen, 2011). However, such approaches deliver no accuracy guarantees for long-term forecasts. Notably, one can avoid approximate uncertainty propagation via multitask GPs models (Bonilla et al., 2007) that use a collection of "condensed models", one for each of the prediction steps (Bradford et al., 2020; Pfefferkorn et al., 2022), or employ a single contextual kernel defined over a joint spatio-temporal domain (Zenati et al., 2022; Li et al., 2024).

**Koopman operator-based learning.** The linearity of Koopman operators and the forecasting simplicity of *linear time-invariant* (LTI) models stemming from their eigendecompositions has led to their increasing popularity in learning dynamical systems (Bevanda et al., 2021; Otto and Rowley, 2021; Brunton et al., 2022). Nevertheless, existing LTI predictors based on operator regression are limited to dissecting long-term components of ergodic dynamics (Korda and Mezić, 2018; Klus et al., 2020; Kostic et al., 2022, 2023). While this approach is extremely powerful for stationary data and reversible dynamics, almost all real-world dynamical systems are irreversible and often even nonstationary (Wu and Noé, 2020). Thus, an increasing amount of methods considers kernels that are *dynamics-informed* (Zhao and Giannakis, 2016; Berry and Sauer, 2016; Banisch and Koltai, 2017; Alexander and Giannakis, 2020; Burov et al., 2021; Dufée et al., 2024). By plugging samples of the dynamics from sequential data into the kernel itself, eigenfunctions of Koopman operators can be directly accessed for both ergodic (Dufée et al., 2024) and transient settings (Bevanda et al., 2023). While the latter has generalization and consistency guarantees, fully tractable representational uncertainty is impossible due to a two-stage regression approach (Angrist and Pischke, 2009; Wang et al., 2022). Still, the existing Koopman operator-based learning approaches offer no epistemic uncertainty bounds, principled model selection or handling of observation noise.

In this work, we present **K**oopman-**E**quivariant **G**aussian **P**rocesses (KE-GPs), the first universal GP models with fully tractable and closed-form confidence bounds for multi-step prediction. By leveraging latent dynamics, our model provides simple LTI responses as a nonlinear function of the initial condition. Strikingly, our GP model provides enhanced generalization compared to existing methods due to intrinsic symmetries (Koopman-equivariants). Furthermore, it delivers continuous-time posteriors without requiring time-derivative data. KE-GPs allow for tractable *simultaneous characterization of both forecasting and representational uncertainty* – alleviating a traditional challenge of GPs and enabling a novel probabilistic treatment of learning dynamics representations.

**Organization.** In Section 2 we introduce the necessary preliminaries together with our problem statement. Section 3 includes the derivation of Koopman-equivariant Gaussian process models, including representation theory and dynamical properties. We then analyze the sample-complexity of our approach through an information-theoretic[1] lens, in Section 4. To handle large datasets, in Section 5, we present our Koopman-equivariant inducing variables for scalable GP-based modeling using variational inference. In Section 6 we demonstrate the utility of our KE-GP approach through a comparison to existing GP and Koopman approaches for learning dynamical systems including predator-prey ODE, datasets from realistic robotic simulators as well as real-world weather data. Finally, in Section 7, we conclude and mention the limitations of the approach.

## 2 PROBLEM SETTING

Our work builds upon the extensive literature on GPs, their interplay with linear operators, and the concept of Koopman-equivariance, which extracts informative "latent states" of dynamics, i.e., Koopman operator eigenfunctions, based on trajectory data. The following covers the necessary prerequisites for setting up the interplay between GPs, linear operators, and intrinsic dynamical system symmetries.

**Notation.** For non-negative integers $n$ and $m$, $[m, n] = \{m, m+1, \ldots, n\}$ with $n \geq m$ gives an interval set of integers. We use the shorthand $[n]{:=}[1, n]$. With a slight abuse of notation, adjoints of operators as well as (conjugate) transposes of matrices are denoted as $(\cdot)^*$ for simplicity. Lower/upper case symbols denote functions/operators while bold symbols are reserved for matrices and vectors. We denote the joint data distribution as $P(dz \times dx \times dy)$, its marginal distributions as $P(dx), P(dz)$, etc., and their support as $\mathbb{X}, \mathbb{Z}$. For functions of observed variables (e.g., $x$ or $z$), $\|\cdot\|_2$

[1]Proofs of theoretical results are in the supplemental.

denotes the $L^2$ norm w.r.t. the respective marginal data distribution. $\|\cdot\|_\infty$ denotes the $L^\infty$ norm. For any kernel $k$, $\mathcal{GP}(0,k)$ refers to the "standard Gaussian process" (van der Vaart and van Zanten, 2008) with zero mean, and covariance defined by $k$. $\lesssim, \gtrsim, \asymp$ represent (in)equalities up to constants; the hidden constants will not depend on any sample size. $\tilde{\mathcal{O}}(\cdot)$ denotes inequality up to logarithm factors.

### 2.1 Gaussian Process Regression

A Gaussian process is a generalization of the Gaussian distribution. It specifies a distribution, such that any finite collection of random variables follows a joint Gaussian distribution, which can be interpreted as a distribution over functions $f : \mathbb{R}^n \to \mathbb{R}$ commonly denoted by $f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$ (Rasmussen and Williams, 2006). This distribution is defined using a prior mean function $m : \mathbb{R}^n \to \mathbb{R}$ and a covariance function $k : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}_{0,+}$. The mean function $m(\cdot)$ includes prior models and is often set to 0 in the absence of such information, which we also assume in the following. The covariance function $k(\cdot,\cdot)$ encodes more abstract prior knowledge, such as symmetries and smoothness of the sample functions.

Given a dataset $\mathbb{D}_N = \{\boldsymbol{z}^{(i)}, y^{(i)}\}_{i\in[N]}$ with training targets $y^{(i)} = \mathsf{y}(\boldsymbol{z}^{(i)}) + \omega^{(i)}$ perturbed by i.i.d. Gaussian noise $\omega^{(i)} \sim \mathcal{N}(0, \sigma_{\mathsf{on}}^2)$, we place a Gaussian process prior $\mathcal{GP}(0, k(\cdot,\cdot))$ on the unknown function $f(\cdot)$ to infer a model. This is straightforwardly achieved by computing the posterior distribution given the training data, which is Gaussian at each test point $\boldsymbol{z} \in \mathbb{R}^n$. We can then compactly express the posterior as $p(\mathsf{y}(\boldsymbol{z})|\mathbb{D}_N) = \mathcal{N}(\mu(\boldsymbol{z}), \sigma^2(\boldsymbol{z}))$, where

$$\mu(\boldsymbol{z}) = \boldsymbol{k}^\intercal(\boldsymbol{z})(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y}, \tag{1}$$

$$\sigma^2(\boldsymbol{z}) = k(\boldsymbol{z},\boldsymbol{z}) - \boldsymbol{k}^\intercal(\boldsymbol{z})(\boldsymbol{K} + \sigma_{\mathsf{on}}^2 \boldsymbol{I}_N)^{-1}\boldsymbol{k}(\boldsymbol{z}), \tag{2}$$

with $k_i(\boldsymbol{z}) = k(\boldsymbol{z}, \boldsymbol{z}^{(i)})$, $K_{ij} = k(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)})$ and $\boldsymbol{y}^\intercal = [y^{(1)} \ \cdots \ y^{(N)}]$. In addition to inferring the posterior distribution, we use the training data to optimize the hyperparameters that arise from the kernel parameterization. This is enabled by the probabilistic approach to the regression problem, which allows us to choose the hyperparameters by minimizing the negative log-likelihood $-\log(p(\boldsymbol{y}|\boldsymbol{Z})) = {}^1\!/_2 \boldsymbol{y}^\intercal(\boldsymbol{K} + \sigma_{\mathsf{on}}^2 \boldsymbol{I}_N)^{-1}\boldsymbol{y} + {}^1\!/_2 \log(\det(\boldsymbol{K} + \sigma_{\mathsf{on}}^2 \boldsymbol{I}_N)) + {}^N\!/_2 \log(2\pi)$.

### 2.2 System Class & Modeling Approach

**System class.** We consider state-space models

$$\text{(dynamics)} \qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^n, \tag{SSMa}$$

$$\text{(output)} \qquad y = h(\boldsymbol{x}) \in \mathbb{R}, \tag{SSMb}$$

with a well-defined flow $\boldsymbol{F}_t(\boldsymbol{x}_0) := \int_0^t \boldsymbol{f}(\boldsymbol{x}(\tau))d\tau$ that requires local Lipschitz continuity of $\boldsymbol{f}$, which is natural to physical systems that often evolve "smoothly". The canonical forecasting model for (SSM) is $y(t, \boldsymbol{x}_0) := h_t(\boldsymbol{x}_0) \equiv h \circ \boldsymbol{F}_t(\boldsymbol{x}_0)$. In practice, a numerical integration scheme is usually required to solve the integral for a shorter time-interval $\Delta t$, such that the actual forecast becomes an $H$-fold composition of nonlinear maps $y(t, \boldsymbol{x}_0) \approx h \circ \boldsymbol{F}_{\Delta t} \circ \cdots \circ \boldsymbol{F}_{\Delta t}(\boldsymbol{x}_0)$ with $H = t/\Delta t \in \mathbb{N}$.

**Spectral Models of Dynamics.** To decompose nonlinear dynamics into simple linear factors and avoid approximate integration schemes, one can utilize the fact that the composition of a function $h$ with the flow $\boldsymbol{F}_t$ can be replaced by a linear, *Koopman*, operator $\mathcal{A}_t : \mathcal{H}' \to \mathcal{H}$ with $[\mathcal{A}_t h](\boldsymbol{x}_0) := h_t(\boldsymbol{x}_0) := h(\boldsymbol{x}_t)$ (Koopman, 1931; Cvitanović et al., 2016). The usefulness of linear operators lies in their ability to forecast any $h \in \mathcal{H}$ in terms of a spectral decomposition (Weidmann, 1980)

$$[\mathcal{A}_t h](\boldsymbol{x}_0) = \sum_{j=1}^{\infty} \underbrace{\mathrm{e}^{\lambda_j t}}_{\text{dynamics}} \underbrace{\langle g_j', h\rangle}_{\text{mode}} \underbrace{g_j(\boldsymbol{x}_0)}_{\text{(eigen)feature}}, \tag{KMD}$$

where the dynamics are parameterized by eigenvalues $\{\lambda_j(\mathcal{A}_t)\}_{j=1}^\infty \in \mathbb{C}$ while $\mathcal{H}' := \operatorname{span}(\{g_j'\}_{j=1}^\infty)$ span an auxiliary and $\mathcal{H} := \operatorname{span}(\{g_j\}_{j=1}^\infty)$ the main representation hypothesis. Under mild conditions *Koopman mode decomposition* (KMD) exists and is dense in $C(\mathbb{X})$ (Korda and Mezić, 2020), cf. Bevanda et al. (2023) and references therein. Learning a finite (KMD) from data, up to a re-scaling of modes, amounts to learning a $D$-dimensional representation $\mathcal{H}_D := \operatorname{span}(\{g_j'\}_{j=1}^D)$.[2]

### 2.3 Problem Statement

Given no knowledge of the Koopman operator or (SSM), our goal is to learn a finite-dimensional model for (KMD) from initial-state and timestep pairs to future output values

$$\mathbb{D}_N = \{(\boldsymbol{x}_0^{(i)}, t^{(i)}),\ y^{(i)}\}_{i\in[N]}, \tag{3}$$

Our model should satisfy the following properties:

- **(D) Trajectory distributions in closed-form:** It corresponds to a Gaussian process framework that models (KMD) based on (3).
- **(E) Data-efficient for dynamical systems:** Allows a sample complexity reduction through the equivariance of (KMD) w.r.t. past state trajectories.

[2] We can set $\mathcal{H}_D' = \mathcal{H}_D$ w.l.o.g. following (Korda and Mezić, 2020).

**(S) Scales to large-scale data:** Admits variational inference techniques for (KMD) based on suitable inducing points.

The closed-form trajectory distributions (**D**) of our proposed framework allow for **1)** continuous epistemic uncertainty over entire time-intervals (an important challenge in utilizing GP models for dynamical systems (Ridderbusch et al., 2023)) and **2)** tractable Bayesian model selection – both of which are absent in existing spectral dynamics modeling (Brunton et al., 2022). Additionally, successful inference on large datasets strongly depends on the availability of informative inducing points (Titsias and Lawrence, 2010), which is particularly hard for high-dimensional inputs (Moss et al., 2023). We propose to utilize the timeseries structure to induce an equivariant covariance using past trajectories that does not increase input dimensionality. Our approach can reduce the maximal information gain (**E**) (Srinivas et al., 2010) and allows for effective variational inference for large-scale GP regression (**S**).

# 3 KOOPMAN SPECTRAL GAUSSIAN PROCESSES

Here we introduce a GP that respects the (KMD) structure, building on the extensive literature on GPs (Rasmussen and Williams, 2006; Duvenaud, 2014), generalized additive models (Krause and Ong, 2011; Mutný, 2024) and their interplay with linear operators (Matsumoto and Sullivan, 2024).

## 3.1 GP-Based Koopman Mode Decomposition

Given a finite set of eigenvalues $\{\lambda_j\}_{j=1}^{|D|}$, the spectral decomposition (KMD) induced by the Koopman operator can be straightforwardly translated into a structured GP model. For this, we assume independent GP priors $g_j(\cdot) \sim \mathcal{GP}(0, k_{g_j}(\cdot,\cdot))$ for the eigenfeatures $g_j(\cdot)$ and exploit the linearity of (KMD) with the modes $\langle g_j', h\rangle$ equal to constant values, such that $y(\cdot,\cdot)$ follows a distribution $y(t, \boldsymbol{x}_0) \sim \mathcal{GP}(0, k_y((t,\boldsymbol{x}),(t',\boldsymbol{x}')))$ with

$$k_y((t,\cdot),(t',\cdot')) := \sum_{j\in[D]} a_j(t,t') k_{g_j}(\cdot,\cdot'), \qquad (\mathsf{cov_{SD}})$$

where $a_j(t,t') = \mathrm{e}^{\lambda_j t}\,\mathrm{e}^{\lambda_j^* t'}$, and $k_{g_j}(\cdot,\cdot)$ can be arbitrary kernels. Conceptually, the kernel ($\mathsf{cov_{SD}}$) is akin to a simulation-induced kernel for linear systems (Chen, 2018), but now captures nonlinear dynamics (SSM). It exhibits the intuitive property that the spatial kernels $k_{g_j}(\cdot,\cdot)$ capture the representational uncertainty due to lifting of the dynamics to a higher dimensional space, in which the forecasting uncertainty evolves linearly according to the LTI features $\{a_j(t,t')\}_{j\in[D]}$. A temporal covariance $a_j(t,t')$ with decay $|\lambda|$ close to zero will result in models with uniform uncertainty over time, whereas taking negative or positive decays will result in models with contracting or expanding variance over time, respectively. This allows for a straightforward encoding of prior knowledge about the temporal evolution of systems, e.g., stability.

**Spectral Hyperprior.** While we generally do not have direct access to a sequence of eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$, it is well known that this spectrum can be effectively covered by sampling a random distribution (Bevanda et al., 2023). We can parameterize a spectral distribution, such that a high-likelihood representation for a finite series in ($\mathsf{cov_{SD}}$) can be obtained by integrating its parameters into Bayesian model selection. To adopt such a spectral prior, we use the noise transfer (outsourcing) trick by (Kallenberg, 1997, Theorem 5.10) to model the eigenvalue distribution $p(\lambda) \approx \rho(\boldsymbol{\vartheta})$. This choice limits the number of required parameters since $\boldsymbol{\vartheta}$ has fewer parameters (degrees of freedom) than the number of eigenspaces $\|\boldsymbol{\vartheta}\|_0 \ll |D|$. Furthermore, it allows for the use of log-likelihood maximization just like with any other set of hyperparameters. Note that the exact Koopman operator $\mathcal{A}_t$ can be approximated with arbitrary accuracy using a sufficiently large finite sequence $\{\lambda_j(\mathcal{A}_t)\}_{j=1}^{D}$ (Bevanda et al., 2023).

## 3.2 Koopman-Equivariant Kernels

While we can use arbitrary kernels for $k_{g_j}(\cdot,\cdot)$ in ($\mathsf{cov_{SD}}$), such a dynamics-agnostic formulation does not exploit any properties of the Koopman operator underlying the spectral decomposition (KMD) which gives rise to ($\mathsf{cov_{SD}}$). In particular, Koopman operators $\mathcal{A}_t$ allow us to reverse the order of forward simulation and measurement function $h(\cdot)$ when determining the output $y$ at a time $t$, i.e., $h(\boldsymbol{F}_t(\boldsymbol{x}_0)) = [\mathcal{A}_t h](\boldsymbol{x}_0)$. Considering only a single eigenvalue $\lambda_j$ of the spectral decomposition (KMD) of the Koopman operator $\mathcal{A}_t$, this equivalence of representations induces a special class of functions, which we refer to as Koopman-equivariant.

**Definition 3.1** (Koopman-equivariance)**.** *Let $[\tau_s,\tau_e] \subset \mathbb{R}$ be a compact subset of the time axis and $\mathcal{M}$ a manifold. A map $\phi_\lambda : \mathcal{M} \mapsto \mathbb{C}$ is called* $([\tau_s,\tau_e],\lambda)$-Koopman-equivariant *if*

$$\mathcal{A}_t\phi_\lambda := \phi_\lambda \circ \boldsymbol{F}_t = \mathrm{e}^{\lambda t}\,\phi_\lambda \qquad (4)$$

*on $\mathcal{M}$ for any $t \in [\tau_s,\tau_e]$.*

To ensure that the prior $\mathcal{GP}(0, k_{g_j})$ over eigenfeatures $g_j(\cdot)$ encodes Koopman-equivariance, observe that Definition 3.1 is a special case of the more general concept of subgroup equivariance (Satorras et al., 2021) adapted to

Koopman operator open eigenfunctions (Mezić, 2020). Since equivariance can be interpreted as a form of symmetry, this allows for the application of well-known techniques for the symmetrization of functions to ensure obtaining Koopman-equivariant functions. We follow the agnostic symmetrization approach of (Kim et al., 2023; Nguyen et al., 2023). For this, we embed past state trajectories from time $\tau_s$ to $\tau_e$ into the input data, i.e.,

$$\mathbb{D}_N^{[\tau_s,\tau_e]} = \{(\boldsymbol{x}_{[\tau_s,\tau_e]}^{(i)}, t^{(i)}), y^{(i)}\}_{i\in[N]}, \tag{5}$$

and exploit Definition 3.1 to obtain projections onto Koopman-equivariant subspaces of our hypothesis space. Notably, we can satisfy Koopman-equivariance in a simple and constructive manner by averaging, or taking an expectation, as we may normalize a measure w.l.o.g. to construct a probability measure on a compact subset of the time-axis. This is formalized in the following.

**Definition 3.2.** *The average w.r.t. a measure $\mu$ on a local*[3] *group $\mathbb{G}_\lambda := \left\{\mathrm{e}^{-\lambda t}\mathcal{A}_t\right\}_{t\in[\tau_s,\tau_e]}$ reads*

$$\mathcal{E}_\lambda^{[\tau_s,\tau_e]} g \mapsto \mathbb{E}_{t\sim\mu(\mathbb{G}_\lambda)}\left[\mathrm{e}^{-\lambda t}\, g \circ \boldsymbol{F}_t(\boldsymbol{x})\right] \tag{6}$$

*which we refer to as $([\tau_s,\tau_e],\lambda)$-equivariance operator.*

By assuming the compactness of the group, we have that the group is unimodular (Folland, 2016, Corollary 2.28) and admits a unitary representation. Thus, we have a self-adjoint symmetrization operator by (Elesedy, 2023, Proposition 3.7), enabling us to transfer the analysis therein to the case of Koopman-equivariance.

**Theorem 3.3.** *Let the action of $\mathbb{G}_\lambda$ be bounded on the compact neighborhood where the local group structure is valid and the symmetrization operator $\mathcal{E}_\lambda^{[\tau_s,\tau_e]} : L_\mu^2 \to L_\mu^2$ w.r.t. to a normalized local Haar measure $\mu$ be well-defined and self-adjoint. Then, it defines the unique solution to*

$$\underset{\psi\in\mathcal{S}_\lambda:=\{g\in L_\mu^2:\ \mathcal{E}_\lambda^{[\tau_s,\tau_e]}g=g\}}{\text{minimize}} \|g-\psi\|_{L_\mu^2}^2 = \|g-\mathcal{E}_\lambda^{[\tau_s,\tau_e]}g\|, \tag{7}$$

*mapping $g$ to its closest (in norm) Koopman-equivariant function $\phi_\lambda = \arg\min_{\psi\in\mathcal{S}_\lambda}\|g-\psi\|_{L_\mu^2}^2$.*

**Remark 3.4.** *Admittedly, the assumption of a normalized measure may be hard to enforce for an unknown group. Without normalization, we lose the orthogonal (minimum norm) projection property, but remark that any representation of a compact group is equivalent to a unitary representation (Reisert and Burkhardt, 2007). Hence, while possible in some cases Kim et al. (2023), we will not focus on building such a measure in practice.*

[3]We have the restriction $\mathbb{G}_\lambda \ni \mathfrak{g}_1+\mathfrak{g}_2$ for all $\mathfrak{g}_1,\mathfrak{g}_2\in\mathbb{G}_\lambda$.

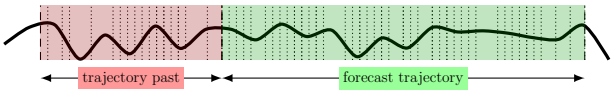


Figure 1: Backward time equivariance interval (red) and the simulation-induced prediction horizon (green).

By construction, the above symmetrization operator from Definition 3.2 renders every base function $g$ $\mu$-almost-everywhere Koopman-equivariant over a time intervals $[\tau_s,\tau_e]$, such that it allows the straightforward design of Koopman-equivariant features $\phi_{\lambda_j}(\cdot)$. To ensure causality of these features, we restrict ourselves to past trajectories, i.e., intervals $[\tau_s,0]$, which results in

$$\phi_{\lambda_j}(\boldsymbol{x}_{[\tau_s,0]}) = \left[\mathcal{E}_\lambda^{[\tau_s,\tau_e]}g\right](\boldsymbol{x}_{[\tau_s,0]}). \tag{8}$$

Finally, since $\mathcal{E}_{\lambda_j}^{[\tau_s,0]}$ is a linear operator, we can exploit the closedness of Gaussian processes under linear operators (Matsumoto and Sullivan, 2024) by placing a GP prior $g(\cdot)\sim\mathcal{GP}(0,k_g(\cdot,\cdot))$ on $g(\cdot)$ with arbitrary kernel $k_g(\cdot,\cdot)$, such that we obtain the Koopman-equivariant prior $\phi_{\lambda_j}(\cdot)\sim\mathcal{GP}(0,k_{\phi_{\lambda_j}}(\cdot,\cdot))$ with $k_{\phi_{\lambda_j}}(\cdot,\cdot) := \mathcal{E}_{\lambda_j}k_g(\cdot,\cdot')\mathcal{E}_{\lambda_j}^*$, inducing a Koopman-equivariant spectral decomposition kernel

$$k_y^{\mathsf{KE}}((t,\cdot),(t',\cdot')) := \sum_{j\in[D]} a_j(t,t')k_{\phi_{\lambda_j}}(\cdot,\cdot), \quad (\mathsf{cov}_{\mathsf{KESD}})$$

that exploits the full information in past trajectories in a structured way to allow predictions of the future evolution using (KMD) as illustrated in Figure 1.

**Practical Considerations.** In practice, our resolution of a trajectory is commonly limited by a sampling time, so we only have access to an empirical measure $\hat{\mu}$ for the expectation in (16). Nevertheless, in most practical considerations and sufficiently regular trajectories, we will get a good sample-based approximation using quadrature so that $\|\hat{\mathcal{E}}_\lambda^{[\tau_s,0]}g-\mathcal{E}_\lambda^{[\tau_s,0]}g\|\approx 0$.

## 4 ANALYSIS OF SAMPLE COMPLEXITY

**Information Gain.** To analyze the sample complexity of regression (E), we use the notion of *information gain*, classical in the analysis of Gaussian processes (Srinivas et al., 2012). Our analysis allows us to put into perspective the sample complexity gains of using the proposed operator-theoretic GP w.r.t. more generic and less structured nonlinear models for dynamical systems. Thus, the following complexity study is a first in the literature. Given the generalized additive structure of our *spectral decomposition covariance* ($\mathsf{cov}_{\mathsf{SD}}$), we quantify the sample-complexity of learning using the

well-established notion of maximal information gain

$$\gamma_N^\sigma(k) := \sup_{\boldsymbol{x}_N \subseteq \mathbb{X}} I\left(\boldsymbol{y}_N; y\right) = \frac{1}{2}\log|\boldsymbol{I}_N + \sigma^{-2}\boldsymbol{K}_N|, \quad (9)$$

that measures the interaction between the data, observation noise, and kernel. This quantity frequently appears in the analysis of the generalization or worst-case estimation error of Gaussian processes (Krause and Ong, 2011; Vakili et al., 2021). The less complex the feature map of the kernel on the same $\mathbb{X}$, the smaller (9) will be, implying better statistical efficiency.

### 4.1 Mercer Eigenvalues as a Proxy to Information Gain

To study the effect of general kernels on the complexity of learning, we will rely on Mercer's theorem (Mercer, 1909) which states that for a well-behaved $k_x$, it can be expressed via the series expansion

$$k_x\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \textstyle\sum_{j=1}^{\infty} \mu_j \varphi_j(\boldsymbol{x}) \varphi_j\left(\boldsymbol{x}'\right), \quad (10)$$

such that $\{\sqrt{\mu_j}\varphi_j\}_{j=1}^{\infty}$ form an orthornormal basis of $L^2(\mathbb{X})$ with respect to a finite Borel measure[4]. The complexity bounds we derive in this work will depend on *how rapidly the eigenvalues* $\{\mu_j\}_{j=1}^{\infty} \subseteq \mathbb{R}_+$ *decay*. The decay of these eigenvalues is closely related to the complexity of the nonparametric model as well as the generalization properties of the posterior (Micchelli and Wahba, 1979). Generally, these eigenvalues decay faster for covariate distributions that are concentrated in a small volume and for kernels that give smooth mean predictors (Widom, 1963, 1964). Thus, the bounds we prove here verify the intuition that our Koopman-equivariant covariance can provide an improved finite-sample performance (**E**). To analyze the effects of the induced Koopman equivariance on the sample complexity, some assumptions are needed:

(HR) *Regularity of the hypothesis:* **a)** $k_x$ is a Mercer kernel (Mercer, 1909). **b)** $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}, |k_x\left(\boldsymbol{x}, \boldsymbol{x}'\right)| \leq \bar{k}$, for some $\bar{k} > 0$ **c)** $\forall j \in \mathbb{N}, \forall \boldsymbol{x} \in \mathbb{X}$, $|\varphi_j(\boldsymbol{x})| \leq r$, for some $r > 0$.

(WS) *The past trajectory interval* $[\tau_s, \tau_e]$ *and the set of initial conditions form a non-recurrent domain.*

(OR) *The operator* $A_t := \sum_{j=1}^{\infty} \mathrm{e}^{\lambda_j t}\, \mathcal{E}_{\lambda_j}^{[\tau_s, 0]}$ *is compact.*

Generally, (HR) is a mild requirement and is fulfilled for continuous kernels on compact domains (Wang et al., 2022), while (OR) is classical for limiting the ill-posedness of inverse problems (Cavalier, 2008). (WS) is a mild technical assumption and allows for a well-specified symmetrization, as it ensures the existence of uncountably many functions satisfying Definition 3.1 for any eigenvalue (Bevanda et al., 2023, Appendix A). For our technical results, we differentiate between *mildly* and *severely* ill-posed setting, based on (generalized) *exponential* and *polynomial* decay rates, respectively.

**Remark 4.1** (Strict complexity reduction)**.** *A direct consequence of well-specified equivariance (WS) is a guaranteed* strict *reduction in the effective dimension (Elesedy, 2021), which is known to equal the information gain up to logarithmic factors (Zenati et al., 2022).*

To study information gain rates for a general class of kernels (that includes our own), we will rely on recent results based on spectral decay properties of kernels (Vakili et al., 2021) and can state the following.

**Theorem 4.2.** *Consider the Mercer eigenvalues* $\{\mu_j\}_{j=1}^{\infty}$ *for* $k_x$ *and let Assumptions (HR),(WS) and (OR) hold. Then* $\exists \theta \geq 1$ *for*

(Poly) $\lambda_j(A_t) \lesssim j^{-p} \wedge \mu_j \lesssim j^{-a}$, $a > 1$ *or*
(Exp) $\lambda_j(A_t) \lesssim \mathrm{e}^{-j^p} \wedge \mu_j \lesssim \mathrm{e}^{-j^b}$, $b > 0$ *so that*

$$\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) \lesssim \tilde{\mathcal{O}}((\gamma_N^\sigma(k_x))^{1/\theta})$$

*where* $\theta = \frac{\max\{2p,a\}}{a}$ *(Poly) and* $\theta = \frac{\max\{2p,b\}}{b}$ *(Exp).*

The above result summarizes the rate gains from the Koopman-equivariant Gaussian process. In case the equivariance operator has a sufficiently strong singular value decay, i.e., $\theta > 1$, the information gain of our Koopman-equivariant GP with covariance ($\mathsf{cov}_{\mathsf{KESD}}$) may be much smaller than for ($\mathsf{cov}_{\mathsf{SD}}$). Crucially, $\theta \geq 1$ is guaranteed, so a slow decay of the operator eigenvalues values will not deteriorate the already existing eigenvalue decay of $\{\mu_j\}_{j=1}^{\infty}$. As (OR) plays the role of a feature extractor, our result suggests one could obtain a significantly improved rate when $\lambda_j(A_t^* A_t)$ has a fast decay, signaling an induced RKHS with low complexity.

**Discussion.** The significance of the asymptotic rates for the maximum information gain when using ($\mathsf{cov}_{\mathsf{KESD}}$) provided by Theorem 4.2 becomes clear when comparing the rates to the ones of other kernels as summarized in Table 2. For example, when using a naïve contextual (spatio-temporal) kernel $k^{\mathsf{SE}}(t,t') \otimes k^{\mathsf{SE}}(\boldsymbol{x}_0, \boldsymbol{x}_i')$ (Zenati et al., 2022) defined over a joint spatio-temporal domain (Li et al., 2024) via RBF kernels[5] $k^{\mathsf{SE}}$, it is well known that the maximum information gain behaves as $\tilde{\mathcal{O}}(\log(N)^{n+2})$. Due to the LTI features $a_j$ in ($\mathsf{cov}_{\mathsf{SD}}$) for describing temporal

[4] Generalization to more general input spaces is straightforward (Steinwart and Scovel, 2012).

[5] The SE kernel is used for ease of exposition, but our results cover large classes of Mercer kernels.

Table 2: Worst-case information gain (w/o log factors) for universal RBF base kernel $k^{\mathsf{SE}}$. Under mild conditions, $\theta \geq 1$ guarantees reduced sample complexity. The number of discretization steps necessary to handle trajectory inputs in naïve kernels and ($\mathsf{cov}_{\mathsf{SD}}$) is denoted by $|G|$.

| $\gamma_N^{\sigma}(\cdot)$ | naïve | ($\mathsf{cov}_{\mathsf{SD}}$) | ($\mathsf{cov}_{\mathsf{KESD}}$) |
|---|---|---|---|
| $\boldsymbol{x}_0$ | $\tilde{\mathcal{O}}(\log(N)^{n+2})$ | $\tilde{\mathcal{O}}(\log(N)^{n+1})$ | — |
| $\boldsymbol{x}_{\tau_s,0}$ | $\tilde{\mathcal{O}}(\log(N)^{\lvert G\rvert n+2})$ | $\tilde{\mathcal{O}}(\log(N)^{\lvert G\rvert n+1})$ | $\tilde{\mathcal{O}}(\log(N)^{\frac{n}{\theta}+1})$ |

correlations, the information gain for $\mathsf{cov}_{\mathsf{SD}}$ reduces to $\tilde{\mathcal{O}}(\log(N)^{n+1})$ (Mutný, 2024). In contrast, our proposed kernel $\mathsf{cov}_{\mathsf{KESD}}$ can exploit the inherent structure imposed by dynamical systems through Koopman-equivariance, such that a complexity of $\tilde{\mathcal{O}}(\log(N)^{\frac{n}{\theta}+1})$ is guaranteed when using SE kernels as the basis for $k_{\phi_{\lambda_j}}$ in ($\mathsf{cov}_{\mathsf{KESD}}$). Hence, for $\theta > 1$, we virtually counteract the curse of spatial dimensionality that comes from the generic and measure-agnostic bounds on the eigenvalue decay for popular kernels (Belkin, 2018). Note that, due to employing Koopman-equivariance (Theorem 3.3), the sample complexity is not impeded by the length or time-discretization of a continuous-time trajectory, which sets ($\mathsf{cov}_{\mathsf{KESD}}$) apart from kernels agnostic of the dynamical systems properties as illustrated in Table 2.

# 5 VARIATIONAL INFERENCE FOR KOOPMAN-EQUIVARIANT GPs

GP model scale poorly with the dataset size, requiring $\mathcal{O}(N^3)$ computations and $\mathcal{O}(N^2)$ memory during training. To address this, in the following we present a sparse GP approximation that uses a variational inference approach. Our approach closely follows stochastic variational inference with sparse GPs (Hensman et al., 2013; van der Wilk et al., 2018), with some additional modifications to the selection and optimization of inducing points. As discussed at the end of this section, this choice allows considerable scalability during training ($\mathsf{S}$), and presents desirable properties when used in conjunction with our equivariant covariance function.

## 5.1 Variational Inference with Sparse GPs

During training, computational complexity stems mainly from the inversion of the data covariance matrix $\boldsymbol{K}_{\boldsymbol{yy}}$, where $[\boldsymbol{K}_{\boldsymbol{yy}}]_{nn'} = k_y^{\mathsf{KE}}((t,\tilde{\boldsymbol{z}}^{(n)}),(t',\tilde{\boldsymbol{z}}^{(n')})) =: k_{y(t,t')}^{\mathsf{KE}}(\tilde{\boldsymbol{z}}^{(n)},\tilde{\boldsymbol{z}}^{(n')})$. To address these problems, we resort to variational inference using *inducing variables* (Quiñonero-Candela and Rasmussen, 2005; Hensman et al., 2013). We obtain the sparse GP by considering $M \ll N$ inducing observations $\boldsymbol{m}$, corresponding to the inducing trajectories $\{\tilde{\boldsymbol{z}}^{(m)} := \boldsymbol{x}_{[\tau_s,\tau_e]}^{(m)}\}_{m=1}^{M} = \tilde{\boldsymbol{Z}}$. Instead of employing the GP prior for the trajectories $\boldsymbol{m}$ corresponding to $\tilde{\boldsymbol{Z}}$, we place a simpler Gaussian prior $q(\cdot)$ over $\boldsymbol{m}$, specified by a mean $\mathbf{m}$ and covariance $\boldsymbol{S}$. By leveraging a variational inference argument (Hensman et al., 2013), we then obtain the approximate Gaussian process posterior $\mathcal{GP}(\tilde{\mu}(\cdot),\tilde{\sigma}^2(\cdot,\cdot))$ with mean and variance

$$\begin{aligned}\tilde{\mu}(\cdot)&=\boldsymbol{k}_{\boldsymbol{u}}^{\mathsf{T}}(\cdot)\,\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\mathbf{m},\\ \tilde{\sigma}^2(\cdot,\cdot)&=k(\cdot,\cdot)-\boldsymbol{k}_{\boldsymbol{u}}^{\mathsf{T}}(\cdot)\,\boldsymbol{K}_{\boldsymbol{uu}}^{-1}[\boldsymbol{K}_{\boldsymbol{uu}}-\boldsymbol{S}]\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{k}_{\boldsymbol{u}}(\cdot)\end{aligned} \tag{11}$$

where $[\boldsymbol{K}_{\boldsymbol{uu}}]_{ij} = k_{y(t,t')}^{\mathsf{KE}}(\tilde{\boldsymbol{z}}^{(i)},\tilde{\boldsymbol{z}}^{(j)})$ and $\boldsymbol{k}_{\boldsymbol{u}}(\cdot) = [k_{y(t,t')}^{\mathsf{KE}}(\tilde{\boldsymbol{z}}^{(m)},\cdot)]_{m=1}^{M}$. The shape of the posterior can be adjusted by changing the values $\tilde{\boldsymbol{Z}}$ and output mean $\boldsymbol{m}$ and variance $\boldsymbol{S}$ of the inducing outputs. Here, we follow an approach similar to Hensman et al. (2013), which allows us to minimize by sampling batches of data instead of computing the full gradient, improving memory complexity to $\mathcal{O}(BM+M^2)$. We choose the hyperparameters and inducing points jointly by minimizing the loss

$$\begin{aligned}\sum_{i=1}^{N}\Big(&-\tfrac{1}{2}\log(2\pi\sigma_{\mathsf{on}}^2)-\tfrac{\sigma_{\mathsf{on}}^2}{2}(y_i-\boldsymbol{k}_i^{\top}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\mathbf{m})^2\\ &-\tfrac{1}{2}\sigma_{\mathsf{on}}^2\tilde{k}_{i,i}-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{S}\boldsymbol{\Lambda}_i\right)\Big)-\mathrm{KL}\left(q(\boldsymbol{u})\|p(\boldsymbol{u})\right),\end{aligned} \tag{12}$$

where $\boldsymbol{k}_i = [k_{y(t,t')}^{\mathsf{KE}}(\tilde{\boldsymbol{z}}^{(m)},\boldsymbol{x}_{[\tau_s,\tau_e]}^{(i)})]_{m=1}^{M}$, $\Lambda_i = \boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{k}_i\boldsymbol{k}_i^{\top}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}$, $\tilde{k}_{i,i} = [\boldsymbol{K}_{\boldsymbol{yy}}-\boldsymbol{K}_{\boldsymbol{yu}}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{K}_{\boldsymbol{yu}}]_{ii}$, and $[\boldsymbol{K}_{\boldsymbol{yu}}]_{ij} = k_{y(t,t')}^{\mathsf{KE}}(\boldsymbol{x}_{[\tau_s,\tau_e]}^{(i)},\tilde{\boldsymbol{z}}^{(j)})$. However, unlike Hensman et al. (2013), we only optimize inducing trajectories and avoid sampling any time/context-related inducing points, which is due to the structure of the spectral decomposition ($\mathsf{KMD}$). This allows a significant reduction in training complexity compared, e.g., to the generic contextual kernel $k^{\mathsf{C}}(\cdot,\cdot) := k^{\mathsf{SE}}(t,t')\otimes k^{\mathsf{SE}}(\boldsymbol{x}_0,\boldsymbol{x}_0')$, where inducing points representing time are also optimized.

**Empirical Information Gain** Due to the structure of our Koopman-equivariant construction, and the resulting benefits in information gain presented in Section 4, our approach is also more robust to a lack of correlation between points, an issue commonly observed in conventional sparse GP approximations (Murray and Adams, 2010; Hensman et al., 2015). In particular, a lower information gain implies that less inducing points are required than with conventional GPs to accurately represent the full posterior (Burt et al., 2019), which further highlights the benefits of the structured prior induced by our construction of the Koopman-equivariant kernel ($\mathsf{cov}_{\mathsf{KESD}}$).

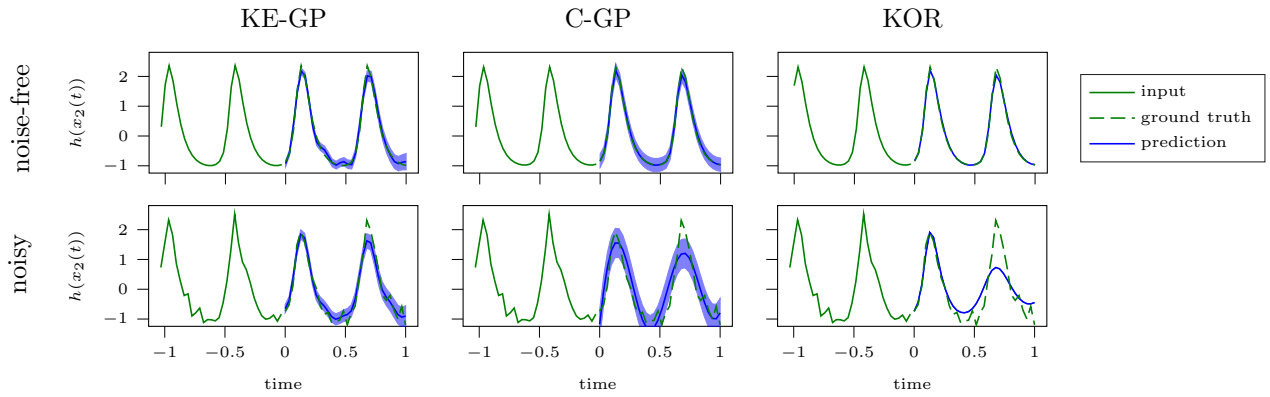


Figure 2: Multi-step mean and 2-sigma interval of the prediction for predator population from the predator-prey dynamics for our proposed Koopman-equivariant GP (KE-GP), a generic contextual kernel (C-GP), and a Koopman operator regression approach (KOR) for noise-free (top) and noisy (bottom) training data.

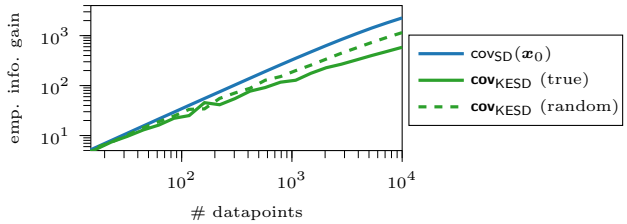


Figure 3: Empirical information gain $\hat{\gamma}$ for a 2D linear system scaled to remove effects of constants. The improved rates confirm our theoretical results for Koopman-equivariant GPs, leading to a lower information gain compared to their non-equivariant counterpart ($\mathsf{cov}_{\mathsf{SD}}$), even when a randomly sampled eigenvalue spectrum $\{\lambda_j\}_{j=1}^{D}$ is used instead of the true spectrum.

Table 3: Comparable nonparametric frameworks.

| Method | (D) | (E) | (S) |
|---|---|---|---|
| C-GP (Li et al., 2024) | (✓) | ✗ | ✓ |
| KOR (Kostic et al., 2022) | ✗ | (✓) | ✓ |
| **KE-GP** (ours) | ✓ | ✓ | ✓ |

## 6 NUMERICAL EXPERIMENTS

To demonstrate the applicability of KE-GPs to realistic data, we perform qualitative and quantitative studies on a set of benchmark examples. As a classical dynamical systems example, we choose the predator-prey model; from the robotics domain, we consider expert demonstrations on the halfcheetah environment from D4RL (Fu et al., 2020) and forecast the first state and action; as a high uncertainty example we choose temperature data from the Monash TSF benchmark (Godahewa et al., 2021) taken at *Oikolab* – demonstrating the usefulness of building in (KMD) structure as a prior for highly complex weather dynamics. Since these datasets provide a single long trajectory, we split off the last chunk as test data and partition the trajectory into $N$ input-task pairs to comply with our model structure. We publish accompanying code at `https://github.com/TUM-ITR/koopman-equivariant-gp`.

**Baselines.** To put our novel algorithm into perspective, we compare to two standard approaches: Gaussian Processes with the time-dependent context (C-GP) by Li et al. (2024), and operator regression for dynamical systems (KOR) (Kostic et al., 2022) from the `kooplearn` package and equip it with SciPy's (Virtanen et al., 2020) *minimize* for hyperparameter tuning. As summarized in Table 3, these two methods exhibit some of the important properties discussed in Section 2, such that they are valuable baselines. We implement C-GPs as well as our KE-GPs in GPJax (Pinder and Dodd, 2022) and equip them with state-of-the-art inference techniques (Leibfried et al., 2020). For Koopman operator regression, we utilized the PCR estimator with a randomized solver from Turri et al. (2023). While Nyström KOR (Meanti et al., 2023) stands as the operator regression's analog to inducing point methods, its performance was uncompetitive in comparison.. In addition, we compare against Gaussian Process ordinary differential equations (GPODE) Hegde et al. (2022) in the extended comparison of Appendix F.1.

**Qualitative Comparison.** We first qualitatively compare the different approaches on the predator-pray model as illustrated in Figure 2. It can be clearly seen that all methods allow for accurate prediction of the future trajectory when given noise-free data from the dynamical system. However, when the state trajectories are perturbed by noise as commonly encountered in practice, significant differences between the predictions become apparent. While our proposed KE-GP maintains a high accuracy and reasonably small confidence

Table 4: Simulations on small subsets of the Predator-Prey (PP), D4RL Half-Cheetah (D4RL), and Oikolab Temparature (OT) datasets. We report RMSE in mean and standard deviation for 5 runs. Training data are $N$ past trajectories over a unit-normalized interval, discretized using $H$ equidistant points.

| | $N \times H$ | **KE-GP** | C-GP | KOR |
|---|---|---|---|---|
| PP | 32×32 | 0.28±0.0 | 0.60±0.0 | **0.27**±0.0 |
| D4RL | 32×16 | 0.46±0.0 | 0.98±0.0 | **0.44**±0.0 |
| OT | 32×16 | **0.63**±0.0 | 0.68±0.0 | 0.86±0.0 |

Table 5: Simulations on large subsets of the Predator-Prey (PP), D4RL Half-Cheetah (D4RL), and Oikolab Temparature (OL) datasets. Training data are $N$ past trajectories over a unit-normalized interval, discretized using $H$ equidistant points.

| | $N \times H$ | **KE-GP** | C-GP | KOR |
|---|---|---|---|---|
| PP | 512×32 | **0.26**±0.0 | 0.42±0.0 | 0.53±0.0 |
| D4RL | 3000×16 | 0.48±0.02 | 0.66±0.07 | **0.44**±0.0 |
| OT | 4000×16 | **0.54**±0.03 | 0.60±0.02 | 0.71±0.0 |

intervals, the estimated uncertainty of the C-GP considerably grows and the prediction accuracy for longer horizons significantly drops for the Koopman operator regression (KOR) approach from Kostic et al. (2022). This high accuracy of KE-GPs can be attributed to their strong generalization capabilities captured by the information gain as discussed in Section 4. When empirically comparing this value, we can immediately see an improvement over non-equivariant kernels, cf. Figure 3.

**Quantitative Evaluation.** We perform two evaluations for each model run and dataset: a small subset for which exact inference is possible and a large subset handled using variational inference. We observe that KE-GP performs robustly on all datasets and sizes as depicted in Tables 4 and 5. While it consistently outperforms the C-GP, the KOR baseline is better on some datasets but the difference in accuracy is marginal. Importantly, KE-GPs provide a significant improvement over KOR for the other data sets. This is fully in line with our qualitative comparison, which shows that KOR can be sensitive to noise with a severe impact on its performance. In addition, we want to stress here that the KOR method does not come with methods for automated model selection, such that manual parameter tuning was necessary to make it competitive. Therefore, this comparison clearly demonstrates the improved generalization ability achieved by embedding the operator-theoretic foundations in our KE-GP approach. Appendix F also includes wall-clock-times, asymptotic complexity for of our own as well as comparison methods.

# 7 CONCLUSION

We presented a novel approach to incorporate an operator-theoretic dynamical system structure into Gaussian process regression. Our framework enables a tractable probabilistic treatment of continuous-time dynamical models not present in existing literature. Utilizing a symmetrization tailored to dynamical systems, based on the concept of Koopman-equivariance (KE), we achieve a sample-complexity reduction compared to a contextual kernel without our proposed (KMD) structure. In scaling to large datasets we exploit our model structure to avoid sampling any time/context-related inducing points. Hence, it does not suffer from a lack of correlation between inducing points, which is common for conventional sparse GP. Through numerical experiments, we show the utility of our KE-GP, demonstrating superior prediction performance to vanilla contextual GPs and on par or better than Koopman operator learning.

## Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and the ERC Consolidator grant "CO-MAN" (ID 864686). A. Lederer acknowledges the support by NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40 225155).

### References

Romeo Alexander and Dimitrios Giannakis. Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques. *Physica D: Nonlinear Phenomena*, 409:132520, 2020. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2020.132520.

Joshua Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edition, 2009. URL `https://EconPapers.repec.org/RePEc:pup:pbooks:8769`.

Yarden As, Bhavya Sukhija, and Andreas Krause. Safe exploration using bayesian world models and log-barrier optimization. 5 2024. URL `http://arxiv.org/abs/2405.05890`.

Ralf Banisch and Péter Koltai. Understanding the geometry of transport: Diffusion maps for Lagrangian trajectory data unravel coherent sets. *Chaos: An*

*Interdisciplinary Journal of Nonlinear Science*, 27 (3):035804, 02 2017.

Dominik Baumann, Alonso Marco, Matteo Turchetta, and Sebastian Trimpe. Gosafe: Globally optimal safe robot learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4452–4458, 2021. doi: 10.1109/ICRA48506.2021.9560738.

Thomas Beckers and Sandra Hirche. Prediction with Approximated Gaussian Process Dynamical Models. *IEEE Transactions on Automatic Control*, 67(12): 6460–6473, 2022. doi: 10.1109/TAC.2021.3131988.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1348–1361. PMLR, 06–09 Jul 2018.

Felix Berkenkamp and Angela P. Schoellig. Safe and robust learning control with Gaussian processes. In *2015 European Control Conference (ECC)*, pages 2496–2501, 2015. doi: 10.1109/ECC.2015.7330913.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tyrus Berry and Timothy Sauer. Local kernels and the geometric structure of data. *Applied and Computational Harmonic Analysis*, 40(3):439–469, 2016. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2015.03.002.

Petar Bevanda, Stefan Sosnowski, and Sandra Hirche. Koopman operator dynamical models: Learning, analysis and control. *Annual Reviews in Control*, 52:197–212, 2021.

Petar Bevanda, Max Beier, Armin Lederer, Stefan Sosnowski, Eyke Hüllermeier, and Sandra Hirche. Koopman Kernel Regression. In *Advances in Neural Information Processing Systems*, volume 37, 2023.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer New York, 1997. ISBN 978-1-4612-6857-4. doi: 10.1007/978-1-4612-0653-8.

Aude Billard, Sina Mirrazavi, and Nadia Figueroa. *Learning for Adaptive and Reactive Robot Control: A Dynamical Systems Approach*. MIT Press, 2022.

Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2452–2470. PMLR, 23–29 Jul 2023.

Erik 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*, 100, 9 2021. ISSN 10075704. doi: 10.1016/j.cnsns.2021.105833.

Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Eric Bradford, Lars Imsland, Dongda Zhang, and Ehecatl Antonio del Rio Chanona. Stochastic data-driven model predictive control using gaussian processes. *Computers & Chemical Engineering*, 139, 2020.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022. doi: 10.1146/annurev-control-042920-020211.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering*. Cambridge University Press, 1 2019. ISBN 9781108380690. doi: 10.1017/9781108380690.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340, 2022.

Marko Budišić, Ryan Mohr, and Igor Mezić. Applied Koopmanism. *Chaos*, 22(4), 10 2012. ISSN 10541500. doi: 10.1063/1.4772195.

Dmitry Burov, Dimitrios Giannakis, Krithika Manohar, and Andrew Stuart. Kernel analog forecasting: Multiscale test problems. *Multiscale Modeling & Simulation*, 19(2):1011–1040, 2021. doi: 10.1137/20M1338289.

David R. Burt, Carl E. Rasmussen, and Mark van der Wilk. Rates of Convergence for Sparse Variational Gaussian Process Regression. 3 2019. URL `http://arxiv.org/abs/1903.03571`.

Edoardo Caldarelli, Antoine Chatalic, Adrià Colomé, Cesare Molinari, Carlos Ocampo-Martinez, Carme Torras, and Lorenzo Rosasco. Linear quadratic control of nonlinear systems with Koopman operator learning and the Nyström method. *arXiv preprint arXiv:2403.02811*, 2024.

Laurent Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, may 2008. doi: 10.1088/0266-5611/24/3/034004.

Tianshi Chen. On kernel design for regularized LTI system identification. *Automatica*, 90:109–122, 2018. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2017.12.039.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *Advances in Neural Information Processing Systems*, 2020.

Sebastian Curi, Armin Lederer, Sandra Hirche, and Andreas Krause. Safe reinforcement learning via confidence-based filters. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 3409–3415, 2022. doi: 10.1109/CDC51059.2022.9992470.

Predrag Cvitanović, Roberto Artuso, Ronnie Mainieri, Gregor Tanner, and Gábor Vattay. *Chaos: Classical and Quantum*. Niels Bohr Inst., Copenhagen, 2016. URL `http://ChaosBook.org/`.

Andreas C Damianou, Michalis K Titsias, Neil D Lawrence, and Amos Storkey. Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes. *Journal of Machine Learning Research*, 17:1–62, 2016.

Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74: 902–924, 2017. ISSN 1364-0321.

Marc Peter Deisenroth and Carl Edward Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. Technical report, 2011.

Benjamin Dufée, Bérenger Hug, Étienne Mémin, and Gilles Tissot. Ensemble forecasts in reproducing kernel hilbert space family. *Physica D: Nonlinear Phenomena*, 459:134044, 2024. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2023.134044.

Nelson Dunford. Spectral Theory. I Convergence to Projections. *Transactions of the American Mathematical Society*, 54(2):185, 9 1943. ISSN 00029947. doi: 10.2307/1990329.

David Kristjanson Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

Stefanos Eleftheriadis, Tom Nicholson, Marc Deisenroth, and James Hensman. Identification of Gaussian process state space models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In *Advances in Neural Information Processing Systems*, volume 34, pages 17273–17283, 2021.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2959–2969. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/elesedy21a.html`.

Byrn Elesedy. *Symmetry and Generalisation in Machine Learning*. PhD thesis, University of Oxford, 2023.

Xuhui Fan, Edwin V. Bonilla, Terence J. O'Kane, and Scott A. Sisson. Free-form variational inference for gaussian process state-space models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Gerald B Folland. *A course in abstract harmonic analysis*. CRC press, 2016.

Roger Frigola, Yutian Chen, and Carl E Rasmussen. Variational gaussian process state-space models. In *Advances in Neural Information Processing Systems*, 2014.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Agathe Girard, Carl Edward Rasmussen, Joaquin Quinonero Candela, and Roderick Murray-Smith. Gaussian process priors with uncertain inputs-application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems*, pages 545–552, 2003. ISBN 0-262-02550-7.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Pashupati Hegde, Çağatay Yıldız, Harri Lähdesmäki, Samuel Kaski, and Markus Heinonen. Variational multiple shooting for bayesian odes with gaussian processes. In *Uncertainty in Artificial Intelligence*, pages 790–799. PMLR, 2022.

Markus Heinonen, Cagatay Yildiz, Henrik Mannerström, Jukka Intosalmi, and Harri Lähdesmäki. Learning unknown ode models with gaussian processes. In *International conference on machine learning*, pages 1959–1968. PMLR, 2018.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of*

*the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282–290, Arlington, Virginia, USA, 2013. AUAI Press.

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. MCMC for variationally sparse Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Lukas Hewing, Elena Arcari, Lukas P. Fröhlich, and Melanie N. Zeilinger. On simulation and trajectory prediction with Gaussian process dynamics. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120, pages 424–434. PMLR, 10–11 Jun 2020a.

Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, 2020b. doi: 10.1146/annurev-control-090419-075625.

Joel L. Horowitz. Ill-posed inverse problems in economics. *Annual Review of Economics*, 6:21–51, 8 2014. ISSN 1941-1383. doi: 10.1146/annurev-economics-080213-041213.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla Bayesian optimization performs great in high dimensions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 20793–20817. PMLR, 21–27 Jul 2024.

Masahiro Ikeda, Isao Ishikawa, and Corbinian Schlosser. Koopman and Perron–Frobenius operators on reproducing kernel Banach spaces. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(12):123143, 12 2022. ISSN 1054-1500. doi: 10.1063/5.0094889.

Isao Ishikawa, Yuka Hashimoto, Masahiro Ikeda, and Yoshinobu Kawahara. Koopman operators with intrinsic observables in rigged reproducing kernel hilbert spaces. 3 2024. URL `http://arxiv.org/abs/2403.02524`.

Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, 1997. ISBN 0-387-94957-7. doi: 10.1007/b98838. URL `http://link.springer.com/10.1007/b98838`.

Takahiro Kawashima and Hideitsu Hino. Gaussian Process Koopman Mode Decomposition. *Neural Computation*, 35(1):82–103, 01 2023.

Mohammad Khosravi, Christopher König, Markus Maier, Roy S. Smith, John Lygeros, and Alisa Rupenyan. Safety-aware cascade controller tuning using constrained bayesian optimization. *IEEE Transactions on Industrial Electronics*, 70(2):2128–2138, 2023. doi: 10.1109/TIE.2022.3158007.

Jinwoo Kim, Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. 36:18582–18612, 2023.

Franz J Király and Harald Oberhauser. Kernels for Sequentially Ordered Data. *Journal of Machine Learning Research*, 20:1–45, 2019.

Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

Stefan Klus, Ingmar Schuster, and Krikamol Muandet. Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020. ISSN 1432-1467. doi: 10.1007/s00332-019-09574-z.

B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.

B. T. Knapik, B. T. Szabó, A. W. van der Vaart, and J. H. van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164: 771–813, 4 2016. doi: 10.1007/s00440-015-0619-7.

Jonathan Ko, Daniel J Klein, Dieter Fox, and Dirk Haehnel. Gp-ukf: Unscented kalman filters with gaussian process prediction and observation models. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2007.

B. O. Koopman. Hamiltonian Systems and Transformation in Hilbert Space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 5 1931. ISSN 0027-8424. doi: 10.1073/pnas.17.5.315.

Milan Korda and Igor Mezić. On Convergence of Extended Dynamic Mode Decomposition to the Koopman Operator. *Journal of Nonlinear Science*, 28(2):687–710, 4 2018. ISSN 14321467. doi: 10.1007/s00332-017-9423-0.

Milan Korda and Igor Mezić. Optimal Construction of Koopman Eigenfunctions for Prediction and Control. *IEEE Transactions on Automatic Control*, 65(12): 5114–5129, 12 2020. ISSN 15582523. doi: 10.1109/TAC.2020.2978039.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning Dynamical Systems via Koopman Operator Regression in Reproducing Kernel Hilbert Spaces. In *Advances in Neural Information Processing Systems*, pages 4017–4031, 2022.

Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Sharp spectral rates for koopman operator learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 32328–32339, 2023.

Vladimir R Kostic, Pietro Novelli, Riccardo Grazzi, Karim Lounici, and massimiliano pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=twSnZwiOIm.

Andreas Krause and Cheng Ong. Contextual Gaussian Process Bandit Optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/f3f1b7fc5a8779a9e618e1f23a7b7860-Paper.pdf.

Henrik Kreidler. Compact operator semigroups applied to dynamical systems. *Semigroup Forum*, 97(3): 523–547, 12 2018. ISSN 00371912. doi: 10.1007/s00233-018-9958-x.

Pijush K Kundu, Ira M Cohen, and David R Dowling. *Fluid mechanics*. Academic press, 2015.

Kari Küster. The Koopman Linearization of Dynamical Systems, 2015. URL https://homepages.laas.fr/henrion/mfo16/kari-kuester.pdf.

Armin Lederer, Alejandro J Ordóñez Conejo, Korbinian A Maier, Wenxin Xiao, Jonas Umlauft, and Sandra Hirche. Gaussian process-based real-time learning for safety critical applications. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6055–6064. PMLR, 18–24 Jul 2021.

Darrick Lee and Harald Oberhauser. The Signature Kernel. 5 2023. URL http://arxiv.org/abs/2305.04625.

Felix Leibfried, Vincent Dutordoir, ST John, and Nicolas Durrande. A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V Bonilla, Theodoros Damoulas, and Terry J Lyons. SigGPDE: Scaling Sparse Gaussian Processes on Sequential Data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6233–6242. PMLR, 10 2021.

J. Li, M. Zagorowska, G. De Pasquale, A. Rupenyan, and John Lygeros. Safe time-varying optimization based on gaussian processes with spatio-temporal kernel. In *Proceedings of NeurIPS 2024*, 2024.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal Rates for Regularized Conditional Mean Embedding Learning. 8 2022. URL http://arxiv.org/abs/2208.01711.

Yingzhao Lian and Colin N. Jones. On Gaussian process based koopman operators. In *IFAC-PapersOnLine*, volume 53, pages 449–455. Elsevier B.V., 2020. doi: 10.1016/j.ifacol.2020.12.217.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379:20200209, 4 2021. ISSN 1364-503X.

Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011. doi: 10.1126/science.1208351.

Kartik Loya, Jake Buzhardt, and Phanindra Tallapragada. Koopman Operator Based Predictive Control With a Data Archive of Observables. *ASME Letters in Dynamic Systems and Control*, 3(3):031009, 10 2023.

Tadashi Matsumoto and T. J. Sullivan. Images of gaussian and other stochastic processes under closed, densely-defined, unbounded linear operators. *Analysis and Applications*, 22:619–633, 4 2024. doi: 10.1142/S0219530524400025.

Alexandre Mauroy and Igor Mezic. Analytic extended dynamic mode decomposition. 5 2024. URL http://arxiv.org/abs/2405.15945.

Alexandre Mauroy, Igor Mezić, and Yoshihiko Susuki. *The Koopman Operator in Systems and Control*, volume 484 of *Lecture Notes in Control and Information Sciences*. Springer International Publishing, Cham, 2020.

Giacomo Meanti, Antoine Chatalic, Vladimir Kostic, Pietro Novelli, Massimiliano Pontil, and Lorenzo Rosasco. Estimating koopman operators with sketching to provably learn large scale dynamical systems. In *Advances in Neural Information Processing Systems*, volume 36, pages 77242–77276, 2023.

J. Mercer. Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458): 415–446, 1 1909. ISSN 0264-3952, 2053-9258. doi: 10.1098/rsta.1909.0016.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal rates for vector-valued spectral regularization learning algorithms. 5 2024. URL http://arxiv.org/abs/2405.14778.

Igor Mezić. Spectrum of the Koopman Operator, Spectral Expansions in Functional Spaces, and State-

Space Geometry. *Journal of Nonlinear Science*, 30 (5):2091–2145, 2020.

Igor Mezić and Andrzej Banaszuk. Comparison of systems with complex behavior. *Physica D: Nonlinear Phenomena*, 197(1):101–133, 2004.

Charles A. Micchelli and Grace Wahba. Design problems for optimal surface interpolation. Technical report, Department Of Statistics, University of Wisconsin Madison, 1979.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal Kernels. Technical report, 2006.

Henry B. Moss, Sebastian W. Ober, and Victor Picheny. Inducing point allocation for sparse gaussian processes in high-throughput bayesian optimisation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 5213–5230. PMLR, 25–27 Apr 2023.

Iain Murray and Ryan Prescott Adams. Slice sampling covariance hyperparameters of latent gaussian models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, page 1732–1740, 2010.

Mojmír Mutný. *Modern Adaptive Experiment Design: Machine Learning Perspective*. Doctoral thesis, ETH Zurich, Zurich, 2024.

Tien Dat Nguyen, Jinwoo Kim, Hongseok Yang, and Seunghoon Hong. Learning symmetrization for equivariance with orbit distance minimization. 11 2023. URL `http://arxiv.org/abs/2311.07143`.

Samuel E Otto and Clarence W Rowley. Koopman Operators for Estimation and Control of Dynamical Systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:2021, 2021.

Maik Pfefferkorn, Michael Maiworm, and Rolf Findeisen. Exact Multiple-Step Predictions in Gaussian Process-based Model Predictive Control: Observations, Possibilities, and Challenges. In *2022 American Control Conference (ACC)*, pages 2829–2836, 2022. doi: 10.23919/ACC53348.2022.9867259.

Thomas Pinder and Daniel Dodd. Gpjax: A gaussian process framework in jax. *Journal of Open Source Software*, 7(75):4455, 2022. doi: 10.21105/joss.04455. URL `https://doi.org/10.21105/joss.04455`.

Kyriakos Polymenakos, Luca Laurenti, Andrea Patane, Jan-Peter Calliess, Luca Cardelli, Marta Kwiatkowska, Alessandro Abate, and Stephen Roberts. Safety guarantees for iterative predictions with gaussian processes. In *Proceedings of the IEEE Conference on Decision and Control*, pages 3187–3193, 2020. ISBN 9781728174471.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006.

Marco Reisert and Hans Burkhardt. Learning equivariant functions with matrix valued kernels. *Journal of Machine Learning Research*, 8(3), 2007.

Steffen Ridderbusch, Sina Ober-Blöbaum, and Paul Goulart. The past does matter: Correlation of subsequent states in trajectory predictions of gaussian process models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 1752–1761, 11 2023. URL `http://arxiv.org/abs/2211.11103`.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The Signature Kernel Is the Solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 1 2021. ISSN 2577-0187. doi: 10.1137/20M1366794.

Simo Sarkka, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013. doi: 10.1109/MSP.2013.2246292.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9323–9332. PMLR, 18–24 Jul 2021.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, 90: 106181, 2020. ISSN 1568-4946.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1015–1022. Omnipress, 2010.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 5 2012. ISSN 0018-9448. doi: 10.1109/TIT.2011.2182033.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 997–1005, Lille, France, 07–09 Jul 2015. PMLR.

Michalis Titsias and Neil D. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Csaba Toth and Harald Oberhauser. Bayesian learning from sequential data using Gaussian processes with signature covariances. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9548–9560. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/toth20a.html`.

Giacomo Turri, Vladimir Kostic, Pietro Novelli, and Massimiliano Pontil. A randomized algorithm to solve reduced rank operator regression. *arXiv preprint arXiv:2312.17348*, 2023.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 82–90. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/vakili21a.html`.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36, 6 2008. ISSN 0090-5364. doi: 10.1214/009053607000000613.

Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Ziyu Wang, Yuhao Zhou, and Jun Zhu. Fast instrument learning with faster rates. In *Advances in Neural Information Processing Systems*, volume 35, pages 16596–16611, 2022.

Joachim Weidmann. *Linear Operators in Hilbert Spaces*, volume 68. Springer New York, 1980. doi: 10.1007/978-1-4612-6027-1.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. I. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. II. *Archive for Rational Mechanics and Analysis*, 17(3):215–229, 1964.

Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *Journal of Nonlinear Science*, 30:23–66, 2 2020. ISSN 0938-8974. doi: 10.1007/s00332-019-09567-y.

Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized ucb for contextual bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 5689–5720. PMLR, 28–30 Mar 2022.

Zhexuan Zeng and Ye Yuan. A Generalized Nyquist-Shannon Sampling Theorem Using the Koopman Operator. 3 2023. URL `http://arxiv.org/abs/2303.01927`.

Zhizhen Zhao and Dimitrios Giannakis. Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, 29(9):2888, aug 2016. doi: 10.1088/0951-7715/29/9/2888.

Mauricio Álvarez, David Luengo, and Neil D. Lawrence. Latent force models. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 9–16. PMLR, 16–18 Apr 2009.

## Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**, we provide this as part of the supplementary material.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [We publish accompanying code at `https://github.com/TUM-ITR/koopman-equivariant-gp`]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**.]
   (b) Complete proofs of all theoretical results. [**Yes**.]
   (c) Clear explanations of any assumptions. [**Yes**.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**.]
   (b) The license information of the assets, if applicable. [**Not Applicable**.]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**.]
   (d) Information about consent from data providers/curators. [**Not Applicable**.]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable.**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**.]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**.]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**.]

# Supplementary Materials for "Koopman-Equivariant Gaussian Processes"

The supplementary materials are organized as follows.

- Appendix A expands on various aspects of related work from the main paper in greater detail.
- Appendix B contains background on standing assumptions and spectral theory of Koopman operators.
- To complement the sample complexity analysis, we address our framework's representation/approximation power in Appendix C.
- The proofs of theoretical results are found in Appendix D.
- Details on the setup of numerical experiments are found in Appendix E.
- Finally, Appendix F includes additional experiments and ablation studies.

## A EXPANDED RELATED WORK

**Vanilla GPs** Gaussian process (GP) regression (Rasmussen and Williams, 2006) has attracted attention for learning nonlinear dynamical systems due to its capability of inferring models with little structural prior knowledge: either by using so-called universal kernels (Micchelli et al., 2006) or placing a prior on a set of kernels and optimizing their likelihood of explaining the data (Duvenaud, 2014). In particular, their ability to quantify epistemic uncertainty has led to a common application in safety-critical control problems (Berkenkamp and Schoellig, 2015; Sui et al., 2015; Berkenkamp et al., 2017; Curi et al., 2022). However, commonly used GP models are single-step predictors, such that uncertainty propagation necessitates approximations when predicting probability distributions more than a single time step in the future. Uncertainty propagation often relies on iterative approaches, in which the previous predictions are used as uncertain inputs to the GP model. This can be exploited in a sampling-based fashion by randomly drawing states (Bradford et al., 2020) or using the unscented transform (Ko et al., 2007). While the computational complexity of sampling-based approaches can be reduced through further approximations (Hewing et al., 2020a; Beckers and Hirche, 2022), it generally remains high. Approximating the predictive distributions, e.g., using a Taylor approximation (Girard et al., 2003) or through exact moment matching (Deisenroth and Rasmussen, 2011) can reduce the complexity, there are no accuracy guarantees of these approximations for long-term forecasts. Direct solutions to these challenges include, e.g., direct modeling of uncertainty intervals (Polymenakos et al., 2020; Curi et al., 2020) or using "a GP per time-step" of prediction (Pfefferkorn et al., 2022), which suffers from a lack of time-correlation and forecast non-linearity.

**State-space GPs** A variety of works considers models with task correlation. Considering modeling dynamical systems, latent variable state-space models have the ability to decouple the model into the dynamics (process) and static (output) structures (Wang et al., 2005; Titsias and Lawrence, 2010; Frigola et al., 2014; Damianou et al., 2016; Eleftheriadis et al., 2017). Like the recent work of (Fan et al., 2023), these models are limited to settings where a single trajectory is available and do not exploit any time-series structure. Still, they require posterior approximations due to the nonlinearity of latent dynamics. Aiding tractability, some works consider linear time-invariant (LTI) models (Álvarez et al., 2009; Sarkka et al., 2013), but come with strong prior-knowledge requirements and unclear representational power. In particular, our KE-GPs can be considered as a continuous contextual GP for dynamical systems with (KMD) structure.

**Koopman operator-based approaches** While operator regression (Williams et al., 2015; Klus et al., 2020; Kostic et al., 2022; Li et al., 2022; Ishikawa et al., 2024; Mauroy and Mezic, 2024; Meunier et al., 2024) could be applied to build an LTI predictor (KMD), this comes with inherent limitations. Namely, the recovery of normal spectra and eigenspaces of $\mathcal{A}_t$ using operator regression in an infinite-dimensional RKHS is an *ill-posed inverse problem* (Knapik et al., 2011, 2016; Horowitz, 2014). Spectral estimation gets increasingly hard with the eigenvalue decay of the covariance (Klebanov et al., 2020), limiting the utility of estimated spectra and eigenspaces. To mitigate these effects, Kostic et al. (2023) suggests using low-rank estimators and empirically estimated RKHSs (Kostic et al., 2024) to control the degree of ill-posedness. However, there is no guarantee such a low-rank representation would span an observable of interest and form an LTI predictor (KMD). In stark contrast, our KE-GP regression bypasses this ill-posedness by construction and directly learns a universal representation of (KMD) in a probabilistic fashion using Bayesian principles. Furthermore, *our approximation-based complexity bounds* (in terms of information gain) *are measure-independent and do not require any i.i.d.-type sampling assumptions.* This is in stark contrast to state-of-the-art concentration results in Koopman operator learning by Kostic et al. (2022, 2023) that are dependent on measures, cf. (Belkin, 2018) for a discussion.

**Works connecting Koopmanism and GPs** Previous attempts at connecting Koopmanism to GPs (Lian and Jones, 2020; Kawashima and Hino, 2023; Loya et al., 2023) rely on heuristics and ad-hoc chocies, lacking theoretical justification as well as rigorous representational considerations. Furthermore, they hinge on heuristics by applying subspace-identification or dynamic mode decomposition before applying Gaussian process regression. In contrast, we offer a principled and fully-tractable approach with provable representational and learning guarantees.

**Signature kernels** Also geared towards sequential data, there is a recent rise in popularity of so-called *signature kernels* (Király and Oberhauser, 2019; Lee and Oberhauser, 2023; Salvi et al., 2021; Lemercier et al., 2021) that also use a symmetrization to be rendered time-reparametrization invariant. This prohibits the extraction of dynamical system representations related to transfer operators and their eigenfunctions. Crucially, time-reparametrization-invariance allows them to excel at discriminative tasks (Lemercier et al., 2021; Salvi et al., 2021) but not at generative tasks such as long-term forecasting (Bevanda et al., 2023).

# B KOOPMAN OPERATOR MODELS FOR DETERMINISTIC DYNAMICS

**Remark B.1** (Operator boundedness)**.** *Consider a forward complete system on a compact set $\mathbb{X}$ and a continuous flow $\boldsymbol{F}_t$. It is well-known that a time-t Koopman operator $\mathcal{A}_t$ is then a contraction semigroup on $C(\mathbb{X})$ (Kreidler, 2018). Due to forward completeness of the flow, we therefore obtain a Banach algebra $C(\mathbb{X})$ with a bounded semigroup $\{\mathcal{A}_t\}_{t\geq 0} \in \mathcal{B}(C(\mathbb{X}))$.*

**Definition B.2** (Non-recurrent domain)**.** *Let time $T \in (0, \infty)$ be given. A set $\mathbb{X}_0 \subset \mathbb{X}$ is called nonrecurrent if*

$$\boldsymbol{x} \in \mathbb{X}_0 \Longrightarrow \boldsymbol{F}_t(\boldsymbol{x}) \notin \mathbb{X}_0 \quad \forall t \in (0, T].$$

*A non-recurrent domain is the image $\mathbb{X}_T$ of non-recurrent set of initial conditions $\mathbb{X}_0$ traced out by the flow map $\boldsymbol{F}_t(\cdot)$*

$$\mathbb{X}_T = \bigcup_{t\in[0,T]} \boldsymbol{F}_t(\mathbb{X}_0) = \bigcup_{t\in[0,T]} \{\boldsymbol{F}_t(\boldsymbol{x}_0) \mid \boldsymbol{x}_0 \in \mathbb{X}_0\}.$$

Less formally, one can think of the non-recurrent domain as the domain *where flow does not intersect itself.*

Practically, non-recurrence is commonly ensured by a choice of the time interval $[0, T]$ so no periodicity is exhibited. Note that it does not mean the system's behavior is not allowed the be periodic, but our perception of it via data does. Effectively this prohibits the multi-valuedness of eigenfunctions – allowing them to define an injective feature map. Thus, non-recurrence is a certain but general condition that bounds the time-horizon $T$ in which it is feasible to completely describe the nonlinear system's flow via an LTI predictor.

Note that our Assumption (WS) requires the existence of a nonrecurrent set that allows for a nonrecurernt domain. It makes for a less-restrictive and intuitive condition compared to existing RKHS approaches (Kostic et al., 2022, 2023) that rely on the self-adjointness and compactness of the actual Koopman operator, which is rarely fulfilled for deterministic dynamics (SSM) and hard to verify without prior knowledge.

### B.1 Koopman Mode Decomposition (KMD)

As in the main text, when referring to *Koopman Mode Decomposition* (KMD), we let the eigenfunctions absorb the spatial mode coefficients $\langle g_j', h\rangle$ (possible w.l.o.g.) as they correspond to eigenfunctions $g_j$ and not eigenvalues $\lambda_j$ (Budišić et al., 2012, Definition 9).

**Lemma B.3** (Universality of (KMD))**.** *Consider a quantity of interest $h \in C(\mathbb{X})$, a forward-complete system flow $\boldsymbol{F}_t(\cdot)$ on a non-recurrent domain $\mathbb{X}$ (Definition B.2) of a compact set $\mathbb{X}$. Then, the output trajectory $y(t) = h(\boldsymbol{x}(t)), \forall t \in [0,T]$ is arbitrarily closely described by the eigenpairs $\{\mathrm{e}^{\lambda_j t}, g_j\}_{j\in\mathbb{N}} \subseteq (\mathbb{C}\times C(\mathbb{X}))$ of the Koopman operator semigroup $\{\mathcal{A}_t\}_{t=0}^T$ so that $\forall \varepsilon > 0, \exists \bar{D} \in \mathbb{N}$*

$$|h(\boldsymbol{x}(t)) - \textstyle\sum_{j=1}^{\bar{D}} \mathrm{e}^{\lambda_j t}\, g_j(\boldsymbol{x}_0)| < \varepsilon, \forall t \in [0,T]. \tag{13}$$

*Proof.* With continuous eigenfunctions for continuous systems proved valid in (Mezić, 2020, Lemma 5.1),(Korda and Mezić, 2020, Theorem 1), the space of continuous functions over a compact set is naturally the space of interest. On a non-recurrent domain, there exist uniquely defined non-trivial eigenfunctions and, by (Küster, 2015, Theorem 3.0.2), the spectrum is rich – with any eigenvalue in the closed complex unit disk legitimate (Ikeda et al., 2022). Further, by (Korda and Mezić, 2020, Theorem 2), this richness is inherited by the Koopman eigenfunctions — making them universal approximators of continuous functions. □

**Intuition on spectral sampling** One may wonder if sampling spectra from a set enclosing the true spectrum may be enough to represent the spectral decomposition of the Koopman operator. Recalling that the spectral decomposition consists of projections to eigenspaces, we remark on a well-known result.

**Remark B.4.** *The choice of our measure of integration might seem arbitrary, and it indeed is. Since we, in general, do not assume knowledge of the spectrum of the Koopman-semigroup, we have to make an approximation. To this end, an educated guess on where the (point-) spectrum might be located is helpful. As elaborated above, the Hille-Yosida-Theorem provides a convenient way to connect the practically attainable growth rates to bounds on the spectrum. The Riesz projection operator $P_\lambda : \mathcal{C} \mapsto \{g \in \mathcal{C} : \mathcal{A}g = \lambda g\}$ to an eigenspace of $\mathcal{A}$ can be represented by*

$$P_\lambda = \frac{1}{2\pi i}\int_{\gamma_\lambda} \frac{\mathrm{d}s}{s - \mathcal{A}},$$

*where $\gamma_\lambda$ is a Jordan curve enclosing $\lambda$ and no other point in $\sigma(\mathcal{A})$ (Dunford, 1943). Obviously $\bigcup_{\lambda\in\sigma(\mathcal{A})} \mathrm{range}(P_\lambda) = \mathcal{C}$, iterating on the fact that we can represent the operator $T$ by its spectral components. It becomes apparent that sampling from a set enclosing $\sigma(\lambda)$ can be seen as sampling curves, eventually enclosing sufficient spectral components. And as stated, one can choose arbitrary measures on $\mathbb{C}$ as long as one ensures they enclose the spectrum.*

## C REPRESENTATIONAL POWER OF KOOPMAN SPECTRAL KERNELS

When the Koopman operator is spectral, e.g., on a non-recurrent domain, the canonical representation of a Koopman operator acting on a well-specified observable $h \in \mathcal{H}$ remains well-specified.

**Lemma C.1.** *Denote by $[\mathsf{y}_t^{KE}]_\sim$ the $L^2$ equivalence class of $\mathsf{y}_t^{KE}$ and denote*

$$\mathsf{A}_t^{[\tau_s,0]} = \sum_{j=1}^{\infty} \mathrm{e}^{\lambda_j t}\, \mathcal{E}_{\lambda_j}^{[\tau_s,0]} \tag{14}$$

*as the canonical spectral representation of a Koopman operator on the time-interval $[\tau_s, 0]$. If $h_0 \in \mathcal{H}$, there exists a kernel $k_y^{KE}$, with integral operator $\mathcal{T}_{k_y^{KE}} = \mathsf{A}_t^{[\tau_s,0]} \mathcal{T}_{k_x} \mathsf{A}_t^{[\tau_s,0]^*}$, s.t. for $\mathsf{h} \sim \mathcal{GP}(0, k_x)$, $\mathsf{y}_t^{KE} \sim \mathcal{GP}\left(0, k_y^{KE}\right)$, $[\mathsf{y}_t^{KE}]_\sim$ has the same distribution as $\mathsf{A}_t^{[\tau_s,0]}[\mathsf{h}]_\sim$.*

*Proof.* Lemma 3.1 (Wang et al., 2022). □

**Remark C.2.** *Note that the above holds for the ($\mathsf{cov}_{\mathsf{SD}}$) when setting the individual equivariance operators to the identity so that $\mathsf{A}_t^{[0,0]} = \sum_{j=1}^{\infty} \mathrm{e}^{\lambda_j t}\, I_j$*

The above infinite sum may seem concerning. However, under mild conditions (WS), there always exists a finite rank representation that is dense in the space of continuous function equipped with the supremum norm (universal). This is formalized in the following.

**Lemma C.3** (Universality). *Let (WS) hold and consider a universal base kernel $k_x$ so that $\{k_{g_j} = k_x\}_{j=1}^D$. Then the induced kernels ($\mathsf{cov}_{\mathsf{SD}}$) and ($\mathsf{cov}_{\mathsf{KESD}}$) are universal, given a sufficiently rich spectral components $\{\lambda_j \in \mathbb{C}\}_j^D$.*

*Proof.* The universality of ($\mathsf{cov}_{\mathsf{SD}}$) follows directly by Korda and Mezić (2020, Theorem 2). With a well-specified symmetrization by (WS), the universality for functions satisfying Koopman-equivariance is inherited by applying Bevanda et al. (2023, Theorem 1 (ii)) component-wise. □

In the above lemma, the "sufficiently rich" can be understood as a set of eigenvalues that enclose the true spectrum. This is straightforwardly achieved by sampling eigenvalues from a distribution with support that encloses the true spectra (Bevanda et al., 2023, Proposition 3).

**Remark C.4** (Uncountable eigenpairs). *Under Assumption (WS) any and all eigenvalues are legitimate, and, for each $\lambda_j$, there are at least uncountably infinitely many eigenfunction-eigenvalue pairs (Bollt, 2021, Corollary 3).*

While naïvely, one would be tempted to optimize a large set of individual eigenvalues, this may be a highly ill-posed problem; as indicated by Remark C.4 and limits the optimization to a very few eigenvalues in practice (Korda and Mezić, 2020; Caldarelli et al., 2024). This is a key motivation in our likelihood optimization of the eigenvalue distribution with only a few degrees of freedom, allows us to efficiently *choose the most likely set of eigenvalues amongst the infinite possibilities.*

# D PROOFS OF THEORETICAL RESULTS

In the following, we restate the definition of Koopman equivariance for completeness.

**Definition D.1** (Definition 3.1 restated). *Let $[\tau_s, \tau_e] \subset \mathbb{R}$ be a compact subset of the time axis and $\mathcal{M}$ a manifold. A map $\phi_\lambda : \mathcal{M} \mapsto \mathbb{C}$ is called* $([\tau_s, \tau_e], \lambda)$-Koopman-equivariant *if*

$$\phi_\lambda \circ \boldsymbol{F}_t = \mathrm{e}^{\lambda t}\, \phi_\lambda \tag{15}$$

*on $\mathcal{M}$ for any $t \in [\tau_s, \tau_e]$.*

## D.1 Symmetrization Based on Koopman Equivariance

**Theorem D.2** (Theorem 3.3 restated & expanded). *Let the action of $\mathbb{G}_\lambda$ be bounded on the compact neighborhood where the local group structure is valid and the symmetrization operator $\mathcal{E}_\lambda^{[\tau_s,\tau_e]} : L_\mu^2 \to L_\mu^2$*

$$\mathcal{E}_\lambda^{[\tau_s,\tau_e]} g \mapsto \mathbb{E}_{t\sim\mu(\mathbb{G}_\lambda)} \left[\mathrm{e}^{-\lambda t}\, g\left(\boldsymbol{x}(t)\right)\right] \tag{16}$$

*w.r.t. to a normalized local Haar measure $\mu$ be well-defined and self-adjoint. Then,*

*i. a function $f \in L_\mu^2$ is* $([\tau_s, \tau_e], \lambda)$-equivariant *if and only if $\mathcal{E}_\lambda^{[\tau_s,\tau_e]}[f] = f$, implying $\mathcal{E}_\lambda^{[\tau_s,\tau_e]}$ is a projection operator so $\|\mathcal{E}_\lambda^{[\tau_s,\tau_e]}\| = 1$ if $L_\mu^2$ contains any* $([\tau_s, \tau_e], \lambda)$-equivariant *functions ($\|\mathcal{E}_\lambda^{[\tau_s,\tau_e]}\| = 0$ otherwise);*

*ii. $L_\mu^2$ decomposes into symmetric and anti-symmetric part $L_\mu^2 = \mathcal{S}_\lambda \oplus \mathcal{S}_\lambda^\perp$ where $\mathcal{S}_\lambda = \big\{g \in L_\mu^2 : g$ is* $([\tau_s, \tau_e], \lambda)$-equivariant$\big\}$ *and $\mathcal{S}_\lambda^\perp = \{g \in L_\mu^2 : \mathcal{E}_\lambda^{[\tau_s,\tau_e]} g = 0\}$;*

*iii. the symmetrization operator $\mathcal{E}_\lambda^{[\tau_s,\tau_e]}$ maps $g$ to the unique solution of*

$$\phi_\lambda = \underset{\psi\in\mathcal{S}_\lambda}{\arg\min} \|g - \psi\|_{L_\mu^2}^2. \tag{17}$$

*Proof.* The first claim i. follows by (Elesedy and Zaidi, 2021, Proposition 24), the second ii. by (Elesedy and Zaidi, 2021, Proposition 25), and iii. by following the proof of (Elesedy, 2023, Proposition 3.9). □

### D.2 New Information Gain Rates

**Technical Lemmas** As our technical results rely on a spectral representation of the base hypothesis space, we state the following technical lemma on Mercer representations considered in this work.

**Lemma D.3** (Mercer representation)**.** *Let $\mathcal{H}$ be any RKHS with kernel $k_x$ s.t. $\int P(d\boldsymbol{x})k_x(\boldsymbol{x},\boldsymbol{x}) < \infty$. Then*

*i. $\mathcal{H}$ can be embedded into $L^2(P(d\boldsymbol{x}))$, and the natural inclusion operator $\iota_x : \mathcal{H} \to L^2(P(d\boldsymbol{x}))$ and $\iota_x^\top$ are Hilbert-Schmidt; the map $\mathcal{T}_{k_x} : h \mapsto \int P(d\boldsymbol{x})k_x(\boldsymbol{x},\cdot)h(\boldsymbol{x})$ defines a positive, self-adjoint and trace-class operator; $\mathcal{T}_{k_x} = \iota_x\iota_x^\top$.*

*ii. $\mathcal{T}_{k_x}$ has the decomposition*

$$\mathcal{T}_{k_x}h = \sum_{i\in I}\mu_i\langle\bar{\varphi}_i, h\rangle_2\bar{\varphi}_i,$$

*where the index set $I \subset \mathbb{N}$ is at most countable, and $\{\bar{e}_i\}$ is an orthonormal system in $L^2(P(d\boldsymbol{x}))$.*

*iii. There exists an orthogonal system $\{e_i : i \in I\}$ of $\mathcal{H}$ s.t. $[e_i]_\sim = \sqrt{\lambda_i}\bar{\varphi}_i$.*

*iv. If $k_x$ is additionally bounded and continuous, $\{e_i : i \in I\}$ will define a Mercer's representation whose convergence is absolute and uniform.*

*Proof.* (Steinwart and Scovel, 2012, Lemma 2.3, 2.2 (for i), 2.12 (for ii-iii), Corollary 3.5 (for iv)). □

We state the following technical Lemma that will help prove a result on information gain rates.

**Lemma D.4** ((Bhatia, 1997))**.** *Let $\mathcal{A}, \mathcal{B}$ be any two operators, $\|\cdot\|$ denote the operator norm and $s_j$ the $j$-th largest singular value. Then*

$$s_j(\mathcal{A}\mathcal{B}) \leq \min\{\|\mathcal{B}\|s_j(\mathcal{A}), \|\mathcal{A}\|s_j(\mathcal{B})\}$$

The above results will be used to bound the eigenvalue decay i.e.

$$\lambda_j(\mathcal{A}\mathcal{B}\mathcal{B}^*\mathcal{A}^*) = s_j(\mathcal{A}\mathcal{B})^2.$$

**Theorem D.5** (Theorem 4.2 restated)**.** *Consider the Mercer eigenvalues $\{\mu_j\}_{j=1}^\infty$ for $k_x$ and let Assumptions (HR),(WS) and (OR) hold. Then $\exists\theta \geq 1$ for*

(Poly)

$$\lambda_j(A_t) \lesssim j^{-p} \wedge \mu_j \lesssim j^{-a}, a > 1$$

*or*

(Exp)

$$\lambda_j(A_t) \lesssim \mathrm{e}^{-j^p} \wedge \mu_j \lesssim \mathrm{e}^{-j^b}, b > 0$$

*so that*

$$\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) \in \tilde{\mathcal{O}}((\gamma_N^\sigma(k_x))^{1/\theta})$$

*where $\theta = \frac{\max\{2p,a\}}{a}$ (Poly) and $\theta = \frac{\max\{p,b\}}{b}$ (Exp).*

*Proof.* We prove the results considering the following two eigendecay profiles:

(Poly) First, based on the information gain results from (Vakili et al., 2021, Corollary 1), we can use a simplified (free of constants) information gain

$$\gamma_N^\sigma\left(k_x\right) \in \mathcal{O}\left(N^{\frac{1}{a}}(\log N)^{1-\frac{1}{a}}\right) \tag{18}$$

for the base kernel $k_x$ with a polynomial decay of Mercer eigenvalues

$$\mu_j \lesssim j^{-a}, a > 1. \tag{19}$$

Secondly, by boundedness of $A_t$, we have an $A_t$-induced decay $\lambda_j(A_t\mathcal{T}_{k_x}A_t^*) \lesssim j^{-a'}$ for some $a'$. This allows us to define a ratio between them and the native decay rate of Mercer eigenvalues $\theta := \frac{a'}{a} > 0$. To uncover the information gain differences, we can equivalently define $a' = \theta a$, such that expressing everything in terms of $\theta$ and $a$ we get

$$\begin{aligned}
\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) &\in \mathcal{O}\left(N^{\frac{1}{a\theta}}(\log N)^{(1-\frac{1}{a\theta})}\right) && \text{(20a)}\\
&\in \mathcal{O}\left(N^{\frac{1}{a}(\frac{1}{\theta})}(\log N)^{\left(\frac{1}{\theta}+(1-\frac{1}{\theta})-\frac{1}{a\theta}\right)}\right) && \text{(20b)}\\
&\in \mathcal{O}\left(N^{\frac{1}{a}(\frac{1}{\theta})}(\log N)^{(1-\frac{1}{a})(\frac{1}{\theta})}(\log N)^{1-\frac{1}{\theta}}\right) && \text{(20c)}\\
&\in \mathcal{O}\left(\left(N^{\frac{1}{a}}(\log N)^{(1-\frac{1}{a})}\right)^{\frac{1}{\theta}}(\log N)^{1-\frac{1}{\theta}}\right) && \text{(20d)}\\
&\in \gamma_N^\sigma(k_x)^{\frac{1}{\theta}}\,\mathcal{O}\left((\log N)^{1-\frac{1}{\theta}}\right) && \text{(20e)}\\
&\in \tilde{\mathcal{O}}\left(\left(\gamma_N^\sigma(k_x)\right)^{\frac{1}{\theta}}\right). && \text{(20f)}
\end{aligned}$$

Using $\lambda_j(A_t) \lesssim j^{-p}$ and $\lambda_j(\iota_x\iota_x^*) \overset{\text{Lem. D.3}}{=} \lambda_j(\mathcal{T}_{k_x}) := \mu_j \lesssim j^{-a}, a > 1$ and invoking Lemma D.4, we obtain

$$\begin{aligned}
\lambda_j\left(A_1\iota_x\iota_x^*A_1^*\right) = \left(s_j\left(A_1\iota_x\right)\right)^2 &\leq \left(\min\{\|\iota_x\|s_j(A_t), \|A_t\|s_j(\iota_x)\}\right)^2 && \text{(21)}\\
&\lesssim \left(\min\{s_j(A_t), s_j(\iota_x)\}\right)^2 && \text{(22)}\\
&\lesssim \left(\min\left\{j^{-p}, \sqrt{j^{-a}}\right\}\right)^2 && \text{(23)}\\
&\lesssim \left(j^{-\max\{p,\frac{a}{2}\}}\right)^2 && \text{(24)}\\
&\lesssim j^{-\max\{2p,a\}} && \text{(25)}
\end{aligned}$$

leading to

$$\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) \in \tilde{\mathcal{O}}\left(\gamma_N^\sigma(k_x)^{\frac{a}{\max\{2p,a\}}}\right). \tag{26}$$

where identifying $\theta := \frac{\max\{2p,a\}}{a} \geq 1$ proves the (Poly) part of the result.

(Exp) First, based on the information gain results from (Vakili et al., 2021, Corollary 1), we can use a simplified (free of constants) information gain

$$\gamma_N^\sigma(k_x) \in \mathcal{O}\left((\log N)^{1+\frac{1}{b}}\right) \tag{27}$$

for the base kernel $k_x$ with an exponential decay of Mercer eigenvalues

$$\mu_j \lesssim \mathrm{e}^{-j^b}, b > 0. \tag{28}$$

Secondly, by boundedness of $A_t$, we have an $A_t$-induced decay $\lambda_j(A_t\mathcal{T}_{k_x}A_t^*) \lesssim \mathrm{e}^{-j^{b'}}$. This allows us to define a ratio between them and the native decay rate of Mercer eigenvalues $\theta := \frac{b'}{b} > 0$. To uncover the information gain differences, we can equivalently define $b' = \theta b$, such that expressing everything in terms

of $\theta$ and $b$ we get

$$\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) \in \mathcal{O}\left((\log N)^{(1+\frac{1}{b\theta})}\right) \tag{29a}$$
$$\in \mathcal{O}\left((\log N)^{(\frac{1}{\theta}+(1-\frac{1}{\theta})+\frac{1}{b\theta})}\right) \tag{29b}$$
$$\in \mathcal{O}\left((\log N)^{(1+\frac{1}{b})\frac{1}{\theta}}(\log N)^{1-\frac{1}{\theta}}\right) \tag{29c}$$
$$\in \mathcal{O}\left(\left((\log N)^{(1+\frac{1}{b})}\right)^{\frac{1}{\theta}}(\log N)^{1-\frac{1}{\theta}}\right) \tag{29d}$$
$$\in \mathcal{O}\left((\log N)^{(1+\frac{1}{b})}\right)^{\frac{1}{\theta}}\mathcal{O}(\log N)^{1-\frac{1}{\theta}} \tag{29e}$$
$$\in \gamma_N^\sigma\left(k_x\right)^{\frac{1}{\theta}}\mathcal{O}(\log N)^{1-\frac{1}{\theta}} \tag{29f}$$
$$\in \tilde{\mathcal{O}}\left(\gamma_N^\sigma\left(k_x\right)^{\frac{1}{\theta}}\right). \tag{29g}$$

Using $\lambda_j(A_t) \lesssim \mathrm{e}^{-j^p}$ and $\lambda_j(\iota_x\iota_x^*) \overset{\text{Lem. D.3}}{=} \lambda_j\left(\mathcal{T}_{k_x}\right) := \mu_j \lesssim \mathrm{e}^{-j^b}, b > 0$ and invoking Lemma D.4, we obtain

$$\lambda_j\left(A_1\iota_x\iota_x^*A_1^*\right) = \left(s_j\left(A_1\iota_x\right)\right)^2 \leq \left(\min\{\|\iota_x\|s_j(A_t), \|A_t\|s_j(\iota_x)\}\right)^2 \tag{30}$$
$$\lesssim \left(\min\{s_j(A_t), s_j(\iota_x)\}\right)^2 \tag{31}$$
$$\lesssim \left(\mathrm{e}^{-\max\{j^p, \frac{1}{2}j^b\}}\right)^2 \tag{32}$$
$$\lesssim \mathrm{e}^{-\max\{2j^p, j^b\}} \tag{33}$$
$$\lesssim \mathrm{e}^{-j^{\max\{p,b\}}} \tag{34}$$

leading to

$$\gamma_N^\sigma\left(k_y^{\mathsf{KE}}\right) \in \tilde{\mathcal{O}}\left(\gamma_N^\sigma\left(k_x\right)^{\frac{b}{\max\{p,b\}}}\right). \tag{35}$$

where identifying $\theta := \frac{\max\{p,b\}}{b} \geq 1$ proves the (Exp) part of the result, finishing the proof.

□

**Remark D.6** (Base SE kernel in ($\mathsf{cov}_{\mathsf{KESD}}$))**.** *By plugging in the known input-dimension dependent information gain for the SE kernel in (29f), we see that the decay speed-up $\theta > 1 \equiv 2p > b$ can counteract the curse of input dimensionality, leading to to*

$$\gamma_N^\sigma\left(k_y^{\mathit{KE}}\right) \in \mathcal{O}\left((\log N)^{\frac{n}{\theta}+1}\right).$$

**Intuition on time-independent $A_t$** We defined the operator $A_t$ in (OR) as time-independent for ease of exposition; the family of $\{A_t\}_{t=0}^T$ is uniquely defined by an infinitesimal generator. This is due to the confinement to a non-recurrent domain (WS) that prescribes finite frequencies of oscillation and finite growth rates, allowing for the infinitesimal generator bounded (Zeng and Yuan, 2023). In practice, we can always take the worst-case time-exponent for analysis by scaling the time appropriately.

# E DETAILS ON NUMERICAL EXPERIMENTS

## E.1 Implementation Details

We implemented GP regression using `GPJax` (Pinder and Dodd, 2022). All of the experiments were performed on machines with 2TB of RAM, 8 NVIDIA Tesla P100 16GB GPUs and 4 AMD EPYC 7542 CPUs.

### E.1.1 Spectral Hyperprior

To tractably optimize over a spectral prior, we use the noise transfer (outsourcing) trick by (Kallenberg, 1997, Theorem 5.10) to model the eigenvalue distribution $p(\lambda) \approx \rho(\boldsymbol{\vartheta})$. This choice limits the number of required parameters since $\boldsymbol{\vartheta}$ has fewer parameters (degrees of freedom) than the number of eigenspaces $\|\boldsymbol{\vartheta}\|_0 \ll |D|$. Furthermore, it allows for the use of log-likelihood maximization just like with any other set of hyperparameters. We use a uniform distribution on $\{\lambda_j = s_j + \mathrm{i}\omega_j : s \in [-\vartheta_s, \vartheta_s] + \vartheta_{\overline{s}},\ \omega \in [-\vartheta_\omega, \vartheta_\omega] + \vartheta_{\overline{\omega}}\}$, $\boldsymbol{\vartheta}=[\vartheta_s, \vartheta_{\overline{s}}, \vartheta_\omega, \vartheta_{\overline{\omega}}]^\top$. To obtain equivariant features, we compute the expectation (16) wrt. a uniform underlying distribution in time, which in case of equally spaced points in time leads to a trapezoid rule.

### E.1.2 Preprocessing and Initialization

We standardize all data trajectories such that the target has to have zero mean and unit variance and the forecast time is between zero and one. This allows us to choose similar parameters for all datasets. We initialize the generative parameters as follows for KE-GP and C-GP:

- init.i) *Prior mean*: $\mu(\boldsymbol{x}) = 0$ and keep it fixed;
- init.ii) *Lengthscale*: $\frac{\sqrt{n_x}}{2}\,\mathrm{std}(\mathbb{X}_{\text{input}})$ following Hvarfner et al. (2024);
- init.iii) *Signal variance*: $\sigma_s^2 = 1$ and fix it, since the data was standardized. Is advised based on the results of Hvarfner et al. (2024);
- init.iv) *Observation noise variance*: $\sigma_{\mathsf{on}}^2 = 1$;
- init.v) *Spectral prior*: $\vartheta_\omega = 15, \vartheta_{\overline{\omega}} = 0$, $\vartheta_s = 1$, $\vartheta_{\overline{s}} = 0$ for scale and bias parameters of the uniform distribution.
- init.vi) *Inducing trajectories*: $N$=32 for exact GP and variational inference, sampled form the train split.

### E.1.3 Variational Inference

As variational inference is notoriously sensitive to initial guesses, we employ a scheme to robustly optimize the generative and variational parameters. To this end we first get reasonable guesses as described above, sample a set of inducing points $(\tilde{Z}, \tilde{Y})$ from the training data and train an exact GP on the inducing inputs via marginal log-likelihood. This yields a set of generative parameters, in particular parameters for the spectral distribution that fit the data. The so obtained posterior is used to initialize the variational GP: $m = \overline{\mu}_{\text{MLL}}(\tilde{Z})$ and $S = \overline{\sigma}_{\text{MLL}}(\tilde{Z}, \tilde{Z})$. This means the initial VI model is the exact posterior on inducing inputs, making the optimization easier, as the initial gradients are smaller. To optimize the variational GP we follow Toth and Oberhauser (2020) and start by first optimizing the variational parameters only, then optimize all parameters jointly and finally optimize the variational parameters on the joint training and validation dataset.

Due to the structure of our Koopman-equivariant construction, and the resulting benefits in information gain presented in Section 4, we found KE-GPs more robust to a lack of correlation between points than C-GP, an issue commonly observed in conventional sparse GP approximations (Murray and Adams, 2010; Hensman et al., 2015). In particular, a lower information gain implies that less inducing points are required than with conventional GPs to accurately represent the full posterior (Burt et al., 2019).

**C-GP** For the contextual GP (Li et al., 2024) with covariance $k^{\mathsf{SE}}(t, t') \otimes k^{\mathsf{SE}}(\boldsymbol{x}_0, \boldsymbol{x}_i')$, we perform the same type of varionational inference as described above, but consider inducing points $\boldsymbol{z} = [\boldsymbol{x}_0^\top, t^\top]^\top$.

## E.2 Benchmark Dynamics

We perform our quantitative study on the following examples of varying complexity. From the robotics domain, we consider expert demonstrations from D4RL (Fu et al., 2020) from the halfcheetah environment and forecast the first state and action. We take temperature data from the Monash TSF benchmark (Godahewa et al., 2021) as a sample for highly complex weather dynamics. Since the latter datasets provide a single long trajectory, we split off the last chunk as test data and partition the trajectory into $\#N$ dataparis pairs to comply with (5).

Table 6: Comparable nonparametric frameworks.

| Method | (**D**) | (**E**) | (**S**) |
|---|---|---|---|
| C-GP (Li et al., 2024) | (✓) | ✗ | ✓ |
| GPODE (Hegde et al., 2022) | ✗ | ✗ | ✓ |
| KOR (Kostic et al., 2022) | ✗ | (✓) | ✓ |
| **KE-GP** (ours) | ✓ | ✓ | ✓ |

Table 7: Simulations on large subsets of the Predator-Prey (PP), D4RL Half-Cheetah (D4RL), and Oikolab Temperature (OL) datasets. Training data are $N$ past trajectories over a unit-normalized interval, discretized using $H$ equidistant points. We report RMSE in mean and standard deviation for five runs.

| | $N \times H$ | **KE-GP** | C-GP | GPODE | KOR |
|---|---|---|---|---|---|
| PP | 512×32 | **0.26**±0.0 | 0.42±0.0 | 0.35±0.03 | 0.53±0.0 |
| D4RL | 3000×16 | 0.48±0.02 | 0.66±0.07 | 0.60±0.02 | **0.44**±0.0 |
| OT | 4000×16 | **0.54**±0.03 | 0.60±0.02 | 0.78±0.03 | 0.71±0.0 |

**Predator-Prey ODE** We use the predator-prey model:

$$\dot{x}_1 = r_1 x_1 + c_i \gamma_1 x_1 x_1, \quad \dot{x}_2 = r_2 x_2 + c_i \gamma_2 x_1 x_2. \tag{36}$$

where $r_1, r_2$, $\gamma$, $c_1, c_2$ are reproduction rates, interaction effects and frequency, respectively. We choose parameters $r_1$=0.2, $\gamma_1$=0.4, $r_2$=0.25, $\gamma_2$=0.2, $c_i = 2$. We create a dataset by simulating the system for $H = 64$ steps with $\Delta t = 3$ for $N = 1024$ trajectories from initial conditions in $[0, 2] \times [0, 1]$.

**Linear ODE** To validate the information gain results on a simple example we use a 2-dimensional linear system $\dot{x_1} = -6x_2, \dot{x_2} = 6x_1$. We create a dataset by simulating the system for $H = 16$ steps with $\Delta t = 0.06$ for $N = 1000$ trajectories from initial conditions in the unit box. The eigenvalues of this system are $\lambda_{1,2} = \pm 6j$. Which we use to compare a randomly sampled enclosing vs. a matching spectral distribution.

**D4RL** The Datasets for Deep Data-Driven Reinforcement Learning (D4RL) (Fu et al., 2020) provides a trajectory collection of reinforcement learning agents interacting with the environments defined in the OpenAI gym (Brockman et al., 2016). We pick the environment halfcheetah with the expert policy. For the halfcheetah we use all actions and observations as our state, demonstrating that our method works in a high dimensional input space. The task for the halfcheetah is to forecast the first action. Input trajectory have 16 steps, the target is to be forecast for 16 steps.

**Oikolab Temparature** As a final benchmark, we draw on the *monash_tsf* dataset collection (Godahewa et al., 2021) and choose their OIKOLAB weather station dataset to test our KE-GP method on the task temperature forecasting. To this end, we provide the models with state trajectories consisting of 8 quantities, including temperature for the past 32 hours and set the task as forecasting the temperature for 16 hours.

# F ADDITIONAL EXPERIMENTS

## F.1 Extended Comparison

We extend our comparison by the recent GPODE models Heinonen et al. (2018); Hegde et al. (2022). Though this model does not fit our requirements, see Table 6, it represents the most popular modeling approach to dynamical systems data. We adopt the codebase of Hegde et al. (2022) with efficient multiple shooting for computing the ELBO. Table 7 contains the simulation results.

In our simulations, we observed that while the method converges and captures the vector field for the simple system and D4RL, the predictive mean is less accurate than our baselines. A reason might be the complexity of learning a high-dimensional vector field that needs to be accurate in every dimension to produce good predictions. As such, GPODE needs a dimensionality reduction step (PCR) to achieve acceptable performance.

Table 8: Cost of solving full estimation problems. $N$: number of context trajectories, $H$: length of context trajectories, $d$: dimensionality of context trajectories, $S$: length of trajectory to forecast.

| Method | **KE-GP** | C-GP | GPODE | KOR |
|---|---|---|---|---|
| single time Gramian $\mathcal{O}$ | $DN^2H^2d$ | $N^2H + N^2d$ | $N^2H^2d^2$ | $N^2(H+S-1)^2d$ |
| multiple times $\mathcal{O}$ | $+DN^2S^2$ | $+N^2S^2$ | | |
| inverse $\mathcal{O}$ | $N^3S^3$ | $N^3S^3$ | $N^3H^3d^3$ | $N^3(H+S-1)^3$ |

Table 9: Wall-clock Times on the Predatory-Prey dataset.

| | | **KE-GP** | C-GP | GPODE | KOR |
|---|---|---|---|---|---|
| Training | epoch [s] | 16 | 10 | 61 | 1.1 |
| | total [min] | 25 | 18 | 55 | 20 |
| Prediciton | time [s] | 1.0 | 1.5 | 30 | 4.5 |

### F.2 Time Complexities

We analyze the asymptotic time complexities of individual steps in the algorithms, cf. Table 8.

KEGP shifts the computational burden from performing the matrix inverse to constructing the kernel. KEGP gramians are more expensive to compute than CGP's by a linear factor $D$, representing the number of eigenfunctions we sample, and quadratically with $H$, the number of past timesteps. However, the computational bottleneck in inference is the inversion of a gram matrix, which naively scales $\mathcal{O}(N^3)$, the standard complexity of a GP. Using spatiotemporal decomposability, computing the Gramian for multiple query times is cheaper.

As the trajectories to forecast are the same for KEGP and CGP, their complexity is not different once the Gramians are computed. The quadratic complexity in $H$ was not an issue in our experiments, as there is no significant performance increase when increasing $H$ beyond a certain point, cf. Appendix F.6 - Input Trajectory Length.

### F.3 Wallclock Times

We provide wall-clock times on the predator-prey dataset for training and prediction for KEGP, CGP, KOR, and GPODE. The datasets consist of 512 training trajectories, 256 validation trajectories, and 256 test trajectories, with context length, forecast length, and input dimension. KOR is run on the CPU (2xAMD EPYC 7542), whereas the other methods are run on one GPU (1xNVIDIA Tesla P100 16G GPU).

**Training.** For the GPs, the most expensive training step is maximizing the ELBO over variational and generative parameters. One epoch consists of gradient updates based on 512 training trajectories and inference on 256 validation trajectories.

The most expensive step for KOR is computing an SVD with a randomized solver that runs for 61 seconds. Gradient-free optimization queries the solver and runs for 55 minutes. Note that the longer computation times for KOR experiments are primarily due to us not running them with inducing points, as the Nyström method was less competitive.

**Prediction.** The KEGP, CGP, and KOR predictions are fast by utilizing spatiotemporal decomposition and the task kernel structure. GPODE takes seconds (an order of magnitude longer) for evaluation since (a) the model is sequential and (b) it needs multiple samples for variance estimates. Thus, motivating the use of operator-theoretic dynamical models.

### F.4 Empirical Information Gain vs C-GP

In this subsection, we demonstrate the implications of our information gain analysis. Fast decaying singular values in the Mercer decomposition indicate that the kernel is statistically efficient for a problem. This, on the

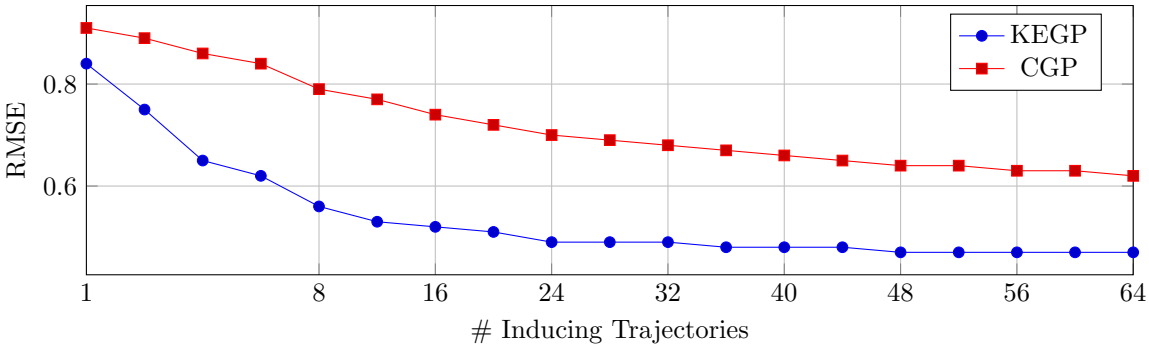


Figure 4: RMSE performance of KEGP and CGP for varied cardinality of inducing trajectories.

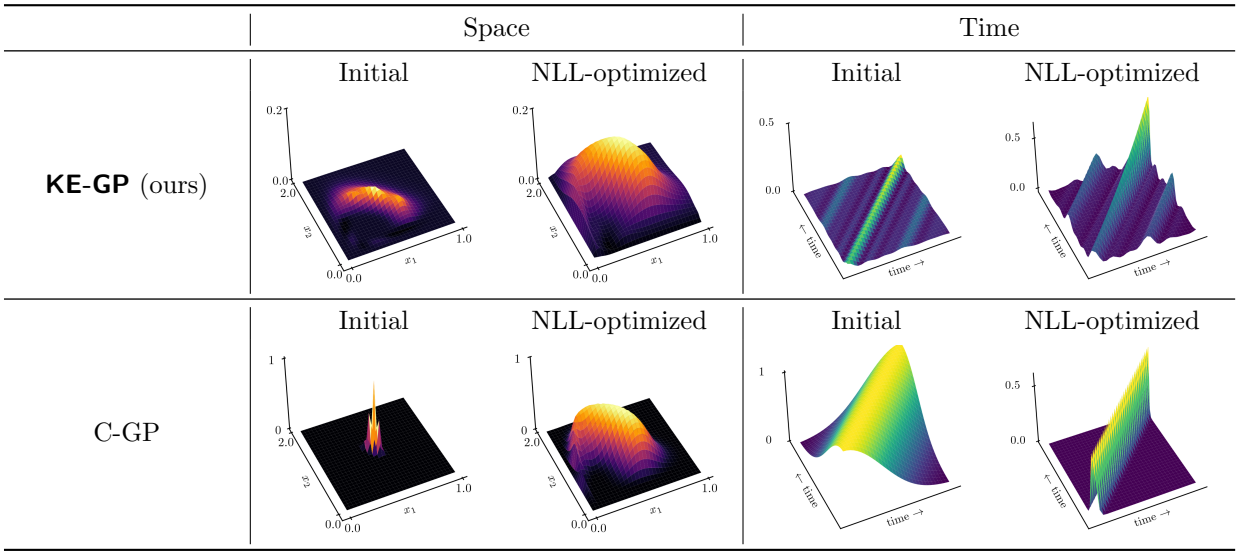


Figure 5: Visualization of the GP covariances in space and time. The spatial, KE-GP prior already strongly indicates the shape of the NLL-optimized covariance.

other hand, enables accurate low-rank approximations such as inducing point methods. As such, an immediate corollary of fast singular value decay (by bounding their tail) is that a small amount of inducing points delivers a similar performance to the full GP estimate.

We assess this effect in Figure 4. We report the performance of unoptimized exact Gaussian Process (GP) models with varying numbers of inducing points. The performance of KEGP reaches saturation much more quickly than CGP's, suggesting that a smaller number of inducing trajectories is sufficient to effectively address the task.

### F.5 Covariance Visualisation vs C-GP

We compare initial and marginal log-likelihood optimized covariances for our KE-GP and C-GP. The spatial covariance corresponds to that of a trajectory. The temporal covariance is separately displayed for the same trajectory. As Figure 5 shows, the learned covariances for our KE-GP and C-GP are of similar shape, the initial spatial covariance of KEGP is less local than that of MTGP, already encoding the trajectory structure before hyperparameter optimization. Further, the KEGP temporal covariance delivers a considerably simple forecasting model, as it is the superposition of multiple one-dimensional LTI systems (KMD), as opposed to a spatially and temporally nonlinear covariance of C-GP.

Notably, since the spectral distribution is a parametrized uniform distribution, KE-GP has only seven parameters, whereas C-GP has 100.

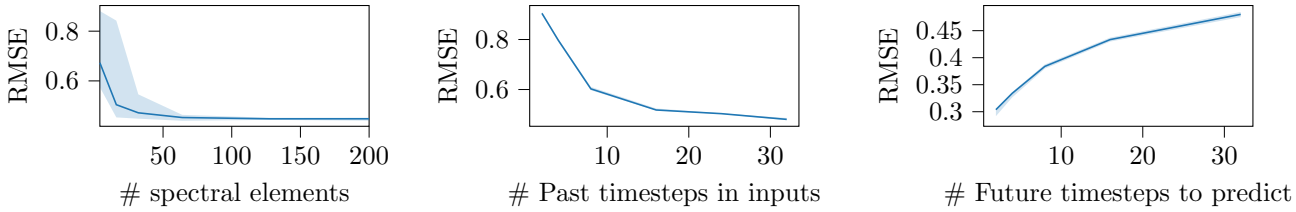


Figure 6: Ablation Studies: test RMSE with varying data and spectral parameters; we report mean and interquartile range over 10 runs.

### F.6 Ablation Studies

As the KE-GP model performance depends on parameters that are absent in naive models, we perform ablation studies to shed light on the role of the new components. In particular, we take a closer look at the effect of the number of modes, the effect of varying length input trajectories, and the effect of varying length forecast time. To this end, we take the predator-prey dynamics (36), and build exact KE-GP models on $N_{\text{train}} = 32$ trajectories and evaluate the model on $N_{\text{test}} = 256$ validation trajectories. We do not optimize the base kernel $k_{g_j}$ hyperparameters, i.e., to enable a consistent comparison between models.

**Eigenspace Dimensionality.** As displayed in Figure 6 (left), the performance improves significantly when increasing the number of eigenspaces sampled up to about $D = 64$; by increasing $D$ further, the performance increase saturates. This behavior resembles the increase in resolution we gain by increasing the number of eigenspaces sampled. A reasonable $D$ is dependent on the quality of the learned hyperprior (Subsection E.1.1), which in turn depends on the dynamical system at hand. Crucially, there is no loss of performance when there are too many spectral elements.

**Input Trajectory Length.** Figure 6 (middle) shows that a longer input trajectory leads to the equivariance operator introducing more information about the dynamics into the prior covariance. As prescribed by representation theory of Section C and discussed in F.5, this leads to a better prior even without the need for optimization.

**Target Trajectory Length.** As displayed in Figure 6 (right), the forecasting problem gets progressively harder as we want to forecast for a longer time, the RMSE increases at a rate of approximately $\sqrt{\mathsf{steps}}$, which is expected for RMSE.