

ERROR ANALYSIS OF KERNEL EDMD FOR PREDICTION AND CONTROL IN THE KOOPMAN FRAMEWORK

FRIEDRICH PHILIPP, MANUEL SCHALLER, KARL WORTHMANN, SEBASTIAN PEITZ, AND FELIKS NÜSKE

ABSTRACT. Extended Dynamic Mode Decomposition (EDMD) is a popular data-driven method to approximate the Koopman operator for deterministic and stochastic (control) systems. This operator is linear and encompasses full information on the (expected stochastic) dynamics. In this paper, we analyze a kernel-based EDMD algorithm, known as kEDMD, where the dictionary consists of the canonical kernel features at the data points. The latter are acquired by i.i.d. samples from a user-defined and application-driven distribution on a compact set. We prove bounds on the prediction error of the kEDMD estimator when sampling from this (not necessarily ergodic) distribution. The error analysis is further extended to control-affine systems, where the considered invariance of the Reproducing Kernel Hilbert Space is significantly less restrictive in comparison to invariance assumptions on an a-priori chosen dictionary.

1. INTRODUCTION

Extended Dynamic Mode Decomposition (EDMD, see, e.g., [56]) and its predecessor DMD [48] have received much attention in recent years. The reason is the capability to predict so-called *observables* along the flow of dynamical (control) systems in a data-driven manner. These observable functions are quantities of interests, which can be used for prediction [20], computation and analysis of key characteristics like attractors [27] or metastable sets [35, 21, 34] as well as control [18], cf. the book [31] and the recent overview article [7] for further applications. The particular appeal of the EDMD approach is due to several reasons: First, the implementation of the learning scheme is comparatively simple. Second, the resulting model is linear and, last, the Koopman framework provides a mathematically-sound foundation.

Application of the EDMD method usually requires two steps: First, the linear Koopman operator [23] (or its generator) is restricted to a subspace \mathcal{D} spanned by a finite set of observable functions. Second, the projection of this finite-rank linear operator to \mathcal{D} , also called *compression*, is approximated using samples, i.e., evaluations of the observables along the flow of the dynamical system. The first step usually requires the a-priori construction of a dictionary consisting of functions on the state space such as, e.g., monomials, which inevitably leads to very large dictionaries for high-dimensional state spaces. Consequently, the respective data requirements to learn the compression of the Koopman operator grows rapidly due to the curse of dimensionality. Data-informed dictionaries using kernel methods have proven their potential to alleviate this issue. In kernel EDMD (kEDMD), the dictionary is built up by kernel feature evaluations on the data set [57, 22] which are contained in an infinite-dimensional ambient Reproducing Kernel Hilbert Space (RKHS, see, e.g., [39, 49]). Thus, the a-priori choice and tuning of the dictionary reduces to the mere choice of a suitable kernel. As another advantage, one gains access to powerful tools from the rich theory of RKHSs such as the kernel trick. For further recent works on the use of reproducing kernels in the context of dynamical systems see also [5, 12, 25, 26].

To derive data-driven approximations of Koopman operator or generator, two fundamentally different sampling techniques are common. The first one is to draw independent-and-identically-distributed (i.i.d.) initial values and to propagate the flow by one time step. In this way, the dynamics is learned by means of more and more i.i.d. samples and results from, e.g., Monte-Carlo integration, can be invoked. The second approach, which is very appealing from a practical point of view, is to acquire data from one

(or several) long trajectories. This approach is called ergodic sampling and is widely applied to systems stemming, for instance, from statistical physics problems, see [28, 29]. As these measurements are clearly not independent, the exploration of the state space has to be ensured by means of properties of the dynamical system such as ergodicity. Ergodic sampling is often mandated if the target measure is a high-dimensional multi-modal distribution, where efficient algorithms to generate i.i.d samples are unavailable. It is also appealing due to its ease of implementation and its ability to reveal unknown ergodic measures and long-term behavior of the system. However, i.i.d. sampling also has a particular advantage: In contrast to ergodic sampling, where one usually obtains characteristics of the behavior through the lens of the ergodic invariant measure, e.g., on an attractor of the system, i.i.d. sampling allows for learning the dynamics on any desired region of the state space.

For an a-priori chosen dictionary, i.e., the original EDMD approach, the approximation error was analyzed in [37, 59] for systems governed by ordinary and stochastic differential equations. However, for kEDMD, this analysis is so far only available if some invariant measure exists on the whole state space, see [42]. In this work, we complement these works by extending the results of [42] on the kernel- and data-based approximation of the Koopman operator to a much more flexible setting without (potentially restrictive) invariance assumptions on the sets and/or measures under consideration. Our main tool is a probabilistic error bound for (cross-)covariance operators in the Hilbert-Schmidt norm using Hoeffding's inequality. The non-invariance of the considered subset of the state space leads to Koopman operators which no longer map the function space into itself, but onto another space in general. Thus, the Koopman operators do not form a semigroup as it is the case under invariance assumptions. Hence, we generalize the definition of the Koopman generator to establish our findings without modifying the kernel-based EDMD algorithm.

To model control systems in the Koopman framework, a popular approach consists of seeking a linear surrogate model in a lifted state space by means of augmenting the state by the control, cf. [24]. Bi-linear models are an alternative [55, 53]. This is motivated by the fact that the Koopman generator inherits control affinity and can, thus, be constructed using Koopman generators of autonomous systems. While loosing linearity of the data-driven surrogate model, the bi-linear approach has proven to be favorable for nonlinear systems with state-control couplings [40]. For EDMD, a respective analysis of the approximation error for dynamical control systems can be conducted by splitting the error into its two sources, cf. [37] for the estimation error and [47] for the projection error. In these works, error bounds for the generator approximation of autonomous systems were leveraged exploiting the control-affine structure. In addition, for set-point stabilization, an error bound proportional to the size of the state and the control was established in [6]. On the one hand, these error bounds were leveraged in [52] to design a robust controller of the bi-linear system using techniques based on linear matrix inequalities. On the other hand, they were exploited in [6] to rigorously ensure practical asymptotic stability of EDMD-based Model Predictive Control using cost controllability [10] of the original system. In this paper, we provide the first error bounds for the approximation of control-affine systems using kernel-based EDMD. Whereas many existing results concerning EDMD for control require the Koopman invariance of the finite-dimensional linear span of the pre-defined dictionary [16, 6], we are able to improve our approximation bounds by assuming that the Koopman operator leaves the (usually infinite-dimensional) ambient RKHS invariant. In contrast to the standard EDMD case with a finite-dimensional dictionary space, this assumption appears to be much more naturally satisfied in many relevant cases.

2. SUMMARY OF THE KEY RESULTS

We briefly highlight our two main findings. Our first result considers error bounds for kernel EDMD (kEDMD) of systems governed by a stochastic differential equation

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t$$

for initial values contained in a compact set \mathcal{X} . The Koopman operator is defined by the conditional expectation of observables along the flow, which is assumed to be contained in possibly unbounded set \mathcal{Y} . More precisely, for an observable function $\psi : \mathcal{Y} \rightarrow \mathbb{R}$ and for all $x \in \mathcal{X}$, we define

$$(K^t \psi)(x) = \mathbb{E}[\psi(X_t) | X_0 = x]. \quad (2.1)$$

We explicitly allow for $\sigma \equiv 0$ such that ordinary differential equations are included. To approximate the Koopman operator from data, we collect samples $(x_k, y_k)_{k=0}^{m-1}$, where $x_k \in \mathcal{X}$ is distributed according to a chosen Borel measure μ on \mathcal{Y} such that $\mu(\mathcal{X}) = 1$ holds and $y_k \sim \rho_t(x_k, \cdot)$ denotes the corresponding successor state. Here, $\rho_t(x, \cdot)$ is the probability distribution associated with the transition kernel of the process, i.e., $\rho_t(x, A) = \mathbb{P}(X_t \in A | X_0 = x)$. As we are interested in approximations of the Koopman operator on Reproducing Kernel Hilbert Spaces (RKHS) on \mathcal{X} and \mathcal{Y} denoted by $\mathbb{H}_{\mathcal{X}}$ and $\mathbb{H}_{\mathcal{Y}}$, respectively, the data matrices $\Psi(X)$ and $\Psi(Y)$ are then formed by kernel evaluations at the data points, i.e.,

$$\Psi(X) = (k(x_i, x_j))_{i,j=0}^{m-1} \quad \text{and} \quad \Psi(Y) = (k(x_i, y_j))_{i,j=0}^{m-1}.$$

Correspondingly, the kEDMD estimator $\widehat{K}_r^{m,t}$ of the Koopman operator K^t on the function space \mathcal{D} spanned by the kernel features, i.e., $\mathcal{D} = \text{span}\{k(x_0, \cdot), \dots, k(x_{m-1}, \cdot)\}$, is given by

$$\widehat{K}_r^{m,t} : \mathcal{D} \rightarrow \mathcal{D} \quad \text{with matrix representation} \quad \widehat{M}_r^{m,t} = [\Psi(X)]_r^\dagger \Psi(Y)^\top,$$

where $[\Psi(X)]_r^\dagger$ is a pseudo-inverse of a rank- r truncation of $\Psi(X)$ to avoid ill-conditioning.

First main result: error bound for autonomous systems, cf. Theorem 4.2. Let an error bound $\varepsilon > 0$ and a probabilistic tolerance $\delta \in (0, 1)$ be given. Then, under suitable assumptions to be specified later, there is a sufficient amount of data $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$,

$$\|P_r K^t - \widehat{K}_r^{m,t}\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow L^2(\mathcal{X}; \mu)} \leq c_r \varepsilon \quad (2.2)$$

holds with probability $1 - \delta$. Here, $c_r \geq 0$ is a constant depending on the truncation order $r \in \mathbb{N}$, and P_r is an orthogonal projection onto the first r components of the Mercer orthogonal basis of $L^2(\mathcal{X}; \mu)$.

Further, if the chosen RKHS is invariant under the Koopman flow, then, with probability at least $1 - \delta$,

$$\|K^t - \widehat{K}_r^{m,t}\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow L^2(\mathcal{X}; \mu)} \leq \sqrt{\lambda_{r+1}} \|K^t\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow \mathbb{H}_{\mathcal{X}}} + c_r \varepsilon$$

with Mercer eigenvalues $\lambda_r \rightarrow 0$ for $r \rightarrow \infty$.

As a second result, we provide bounds for kEDMD-based predictions of control-affine systems

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n_u} u_i(t) g_i(x(t)) \quad (2.3)$$

with a piecewise constant control function taking values in a compact set $\mathcal{U} \subset \mathbb{R}^{n_u}$, resulting e.g. from sampling with zero-order hold. For a constant control function $u(t) \equiv u \in \mathcal{U}$, the counterpart of the Koopman operator (2.1) is given by

$$(K_u^t \psi)(x^0) := \psi(x(t; x^0, u)), \quad (2.4)$$

where $x^0 \in \mathcal{X}$ and $x(\cdot; x^0, u)$ is the solution to (2.3) with $x(0) = x^0$. Again, \mathcal{Y} is chosen in a way such that $x(t; x^0, u) \in \mathcal{Y}$ for $x^0 \in \mathcal{X}$. As the generator of the Koopman operator (2.4) inherits the control affine structure of the underlying control system (2.3), we show that the Koopman operator is approximately control affine (for small times $t \geq 0$). Then, we define its approximation by

$$\widehat{K}_{r,u}^{m,t} := \widehat{K}_r^{m,t} + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\widehat{K}_{r,\gamma_i e_i}^{m,t} - \widehat{K}_r^{m,t}), \quad (2.5)$$

where $\widehat{K}_{r,\gamma_i e_i}^{m,t}$, $\widehat{K}_r^{m,t}$ are estimators of the Koopman operators corresponding to (2.3) for the constant controls $u = \gamma_i e_i$, $i = 1, \dots, n_u$, and $u = 0$, respectively, using the kEDMD method as outlined above for autonomous systems.

Second main result: error bound for control systems, cf. Theorem 5.1. Let an error bound $\varepsilon > 0$, a probabilistic tolerance $\delta \in (0, 1)$, and a control $u \in \mathcal{U} \subset \mathbb{R}^m$, \mathcal{U} compact be given. Then, under suitable assumptions to be specified later, there is a sufficient amount of data $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$,

$$\|P_r K_u^t - \widehat{K}_{r,u}^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X}; \mu)} \leq c_r \varepsilon + t^2 \cdot C_k \widehat{\lambda}_r^{-1}.$$

The first term resembles the right-hand side of (2.2) and the second term stems from the *approximate control-affinity* of the Koopman operator.

The outline of this paper is as follows. In the following section, the reader is introduced to the setting and fundamentals of this work, such as stochastic differential equations, Reproducing Kernel Hilbert spaces, Koopman operators, and kEDMD including the main assumptions. In Section 4 we present and prove our main result, Theorem 4.2, on the approximation error on the kEDMD estimator for autonomous systems. In Section 5 we first recall the affine dependence of the Koopman generator on the control for control-affine systems and show that the same holds approximately (for small sampling times) for the Koopman operator, resulting in a natural estimator via an affine combination of autonomous operators. Our second main result, Theorem 5.1, provides an error bound on this estimator. Finally, Section 6 concludes the paper with an exhibition of numerical results.

3. SETTING

In the following, we introduce the reader to the setting considered in this paper which involves stochastic differential equations (Section 3.1), Koopman operators (Section 3.3), symmetric positive definite kernels and their associated reproducing kernel Hilbert spaces (Section 3.4), and the kEDMD approximation method (Section 3.5). In Section 3.2 we provide our assumptions and discuss them in detail.

3.1. Stochastic differential equations. Let a stochastic differential equation (SDE) with drift vector field $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and diffusion matrix field $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ be given, i.e.,

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad (3.1)$$

where W_t is d -dimensional Brownian motion. We assume that both b and σ are Lipschitz-continuous. Then [38, Theorem 5.2.1] guarantees the existence of a unique solution $(X_t)_{t \geq 0}$ to (3.1). The case $\sigma \equiv 0$ will be referred to as “the deterministic case” as the SDE then reduces to the ODE $\dot{x} = b(x)$.

The solution $(X_t)_{t \geq 0}$ of the SDE (3.1) is a continuous time-homogeneous Markov process, whose transition kernel is denoted by $\rho_t : \mathbb{R}^d \times \mathcal{B} \rightarrow \mathbb{R}$, where \mathcal{B} is the Borel σ -algebra on \mathbb{R}^d . Then $\rho_t(x, \cdot)$ is a probability measure for all $x \in \mathbb{R}^d$, and for each $A \in \mathcal{B}$ we have that $\rho_t(x, A)$ is a representative of the conditional probability for A containing X_t , given $X_0 = x$, i.e.,

$$\rho_t(x, A) = \mathbb{P}(X_t \in A | X_0 = x) \quad \text{for } \mathbb{P}^{X_0}\text{-a.e. } x \in \mathbb{R}^d,$$

where \mathbb{P}^{X_0} denotes the law of X_0 . Note that the law of X_t is determined by

$$\mathbb{P}^{X_t}(A) = \int \rho_t(x, A) d\mathbb{P}^{X_0}(x), \quad A \in \mathcal{B}. \quad (3.2)$$

It is well known [3, Section 1.3] that the transition kernel satisfies the so-called *Chapman-Kolmogorov equation*

$$\rho_{t+s}(x, A) = \int \rho_s(y, A) \rho_t(x, dy), \quad A \in \mathcal{B}, \quad x \in \mathbb{R}^d, \quad s, t \geq 0. \quad (3.3)$$

As it is usual, we agree to write $\rho_t(x, dy)$ for $d\rho_t(x, \cdot)(y)$.

3.2. Assumptions. In many works it is assumed that there is some set $\mathcal{X} \subseteq \mathbb{R}^d$ which is invariant under the flow of the SDE and that there is an invariant distribution π on \mathcal{X} such that the solution process X_t is stationary with respect to π , i.e., $\mathbb{P}^{X_t} = \pi$ for all $t \geq 0$. In view of (3.2), this is equivalent to

$$\int \rho_t(x, A) d\pi(x) = \pi(A), \quad A \in \mathcal{B}_{\mathcal{X}}, t \geq 0.$$

Here, $\mathcal{B}_{\mathcal{X}}$ denotes the Borel sigma algebra on \mathcal{X} .

However, although the Krylov–Bogolyubov Theorem [11] ensures the existence of such an invariant measure π for a large class of SDEs, both the set \mathcal{X} as the measure π are often unknown in practice; even if π is known, generating independent and identically distributed (i.i.d) samples with respect to it remains, in general, a challenging problem.

In this paper, we relax both conditions in a way that the objects become both more flexible and more accessible for applications:

- $\mathcal{X} := \overline{\Omega}$, where $\Omega \subset \mathbb{R}^d$ is a bounded and connected open set;
- $T > 0$ is a finite time horizon;
- $\mathcal{Y} \subset \mathbb{R}^d$ is a closed set such that $\mathcal{X} \subset \mathcal{Y}$ and

$$\forall t \geq 0 \forall x \in \mathcal{X} : \rho_t(x, \mathcal{Y}) = 1. \quad (3.4)$$

- μ is a finite Borel measure on \mathcal{Y} such that $\mu(\mathcal{X}) = 1$ and there exists a constant $L > 0$ such that

$$\int_{\mathcal{X}} \rho_t(x, A) d\mu(x) \leq L \cdot \mu(A), \quad A \in \mathcal{B}_{\mathcal{Y}}, t \in [0, T]. \quad (3.5)$$

The set \mathcal{X} will later serve as the set, which initial data for the process is sampled from. The assumption (3.4) then assures that a subsequent sample after time $t \in (0, T]$ of the process remains in \mathcal{Y} almost surely. Clearly, $\mathcal{Y} = \mathbb{R}^d$ is a possible choice.

Let us discuss the above assumptions in the following remark.

Remark 3.1. (a) In the case $\mathcal{Y} = \mathcal{X}$, the set \mathcal{X} is invariant under the stochastic flow of the SDE (3.1), i.e.,

$$\forall t \geq 0 \forall x \in \mathcal{X} : \rho_t(x, \mathcal{X}) = 1.$$

A sufficient condition on the coefficients of the SDE (3.1) for the invariance of \mathcal{X} under the SDE flow is given by

$$\sigma(x) = 0 \quad \text{for all } x \in \mathbb{R}^d \setminus \mathcal{X} \text{ and} \quad (3.6a)$$

$$b(x)^\top \nu(x) \leq 0 \quad \text{for all } x \in \partial\mathcal{X} \text{ and all outer normal vectors } \nu(x) \in \mathbb{R}^n, \quad (3.6b)$$

where an outer normal vector of \mathcal{X} at $x \in \partial\mathcal{X}$ is any vector $\nu \in \mathbb{R}^d$, $\nu \neq 0$, such that the open ball with center $x + \nu$ and radius $\|\nu\|_2$ and \mathcal{X} are disjoint¹. This follows by an application of the Bony-Brezis (also known as Nagumo) theorem, which is proven in [54, Theorem 10.XVI, p. 117f] for the deterministic case under the *tangent or interior pointing*-condition (3.6b). In the stochastic setting, continuity of solutions to the SDE ensures that any trajectory leaving \mathcal{X} has to cross the boundary. The condition (3.6a) implies that the noise has no influence at the boundary (σ is continuous) and in the exterior of \mathcal{X} such that the SDE reduces to an ordinary differential equation. Hence, the proof in [54] for the deterministic setting also applies to the SDE case.

(b) Note that (3.5) means that the push-forward measures ν_t of μ under the dynamics, defined on \mathcal{Y} by $\nu_t(A) = \int_{\mathcal{X}} \rho_t(x, A) d\mu(x)$, $t \in [0, T]$, are uniformly Lipschitz-continuous w.r.t. μ , i.e., $\nu_t(A) \leq$

¹If Ω has a C^1 -boundary, the outer unit vector is unique. In case of Lipschitz continuous boundary, as a consequence of Rademacher's theorem [13, Theorem 3.2], it is unique up to a set of Lebesgue measure zero on the boundary.

$L \cdot \mu(A)$ for all $A \in \mathcal{B}_{\mathcal{Y}}$, cf. [41, Definition 3.1]. In consequence, μ -integrable functions on \mathcal{Y} are also ν_t -integrable, and for $\psi \in L^1(\mathcal{Y}; \mu)$ we have

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} |\psi(y)| \rho_t(x, dy) d\mu(x) = \int_{\mathcal{Y}} |\psi| d\nu_t \leq L \int_{\mathcal{Y}} |\psi| d\mu. \quad (3.7)$$

(c) In the deterministic case (i.e., $\sigma \equiv 0$), (3.4) means that $F_t(x) \in \mathcal{Y}$ for all $t \in [0, T]$ and all $x \in \mathcal{X}$, where F_t denotes the flow of the ODE $\dot{x} = b(x)$. Moreover, the condition (3.5) is equivalent to

$$\mu(F_t^{-1}(A)) \leq L \cdot \mu(A), \quad A \in \mathcal{B}_{\mathcal{Y}}, t \in [0, T]. \quad (3.8)$$

In Section A we show that this condition is satisfied for a large class of systems and measures.

3.3. Koopman operators. For $p \in [1, \infty]$, we let $\|\cdot\|_{p, \mathcal{Y}}$ denote the $L^p(\mathcal{Y}; \mu)$ -norm. In particular, for $p < \infty$,

$$\|\psi\|_{p, \mathcal{Y}} = \left(\int_{\mathcal{Y}} |\psi(y)|^p d\mu(y) \right)^{1/p}.$$

Moreover, we define $\langle \phi, \psi \rangle_{2, \mathcal{Y}} := \int_{\mathcal{Y}} \phi(y) \overline{\psi(y)} d\mu(y)$ for $\phi, \psi \in L^2(\mathcal{Y}; \mu)$. Similarly, we define the norm $\|\cdot\|_{p, \mathcal{X}}$ on $L^p(\mathcal{X}; \mu)$ and the scalar product $\langle \cdot, \cdot \rangle_{2, \mathcal{X}}$ on $L^2(\mathcal{X}; \mu)$, respectively.

Let $B(\mathcal{Y})$ denote the set of all bounded Borel-measurable functions on \mathcal{Y} . For $t \in [0, T]$, we define the *Koopman operator* K^t on $B(\mathcal{Y})$ associated with the SDE (3.1) by

$$(K^t \psi)(x) = \int_{\mathcal{Y}} \psi(y) \rho_t(x, dy) = \mathbb{E}[\psi(X_t) | X_0 = x], \quad x \in \mathcal{X},$$

for $\psi \in B(\mathcal{Y})$.

Remark 3.2. (a) For the definition of K^t to carry over to spaces of equivalence classes of functions that coincide a.e. on \mathcal{Y} , it must be ensured that $K^t \psi_1 = K^t \psi_2$ μ -a.e. on \mathcal{X} whenever $\psi_1 = \psi_2$ μ -a.e. on \mathcal{Y} . This is equivalent to the absolute continuity of the measure ν_t w.r.t. μ (see Remark 3.1), which is a relaxation of (3.5). However, it is easy to see that K^t maps $L^1(\mu)$ (boundedly) into itself if and only if $\nu_t \leq L\mu$ for some $L > 0$, i.e., if (3.5) holds.

(b) Usually, the Koopman operator of a dynamical system on \mathcal{X} is considered as a linear operator from a space of functions on \mathcal{X} *into itself*. The price we pay by relaxing the invariance condition on \mathcal{X} is that our version of the Koopman operator is a linear map between *two different spaces*.

(c) Let us briefly consider the deterministic case. Here, we have $\rho_t(x, A) = \delta_{F_t(x)}(A)$, and thus $(K^t \psi)(x) = \int_{\mathcal{Y}} \psi(y) d\delta_{F_t(x)}(y) = \psi(F_t(x))$ for $t \in [0, T]$, cf. (3.4). This complies with the usual definition of the Koopman operator for deterministic dynamical systems as a composition operator.

A proof of the following proposition can be found in the Appendix Section B.

Proposition 3.3. *Let $p \in [1, \infty]$. For every $t \geq 0$, the operator K^t extends uniquely to*

- a contraction from $L^\infty(\mathcal{Y}; \mu)$ to $L^p(\mathcal{X}; \mu)$;
- a bounded operator from $L^p(\mathcal{Y}; \mu)$ to $L^p(\mathcal{X}; \mu)$.

If $p < \infty$, for $\psi \in L^p(\mathcal{Y}; \mu)$ we have $K^t \psi \rightarrow \psi|_{\mathcal{X}}$ in $L^p(\mathcal{X}; \mu)$ as $t \rightarrow 0$.

Remark 3.4. If \mathcal{X} is invariant under the SDE flow (i.e., $\mathcal{Y} = \mathcal{X}$), the operators K^t satisfy the semigroup property $K^{t+s} = K^t K^s$ for $s, t \geq 0$, which is a simple consequence of the Chapman-Kolmogorov equation (3.3). If μ is an invariant measure, $(K^t)_{t \geq 0}$ is a C_0 -semigroup of contractions on² $L(L^p(\mathcal{X}; \mu))$,

²In what follows, $L(\mathcal{H}, \mathcal{K})$ denotes the set of all bounded linear operators between Hilbert spaces \mathcal{H} and \mathcal{K} . As usual, we also set $L(\mathcal{H}) := L(\mathcal{H}, \mathcal{H})$.

where $p \in [1, \infty)$. However, if μ is not invariant but at least satisfies (3.5) on each interval $[0, T]$ (with the constants L depending on T), the semigroup $(K^t)_{t \geq 0}$ might not be uniformly bounded. An example is given in Appendix C.

3.4. Reproducing kernel Hilbert spaces. In what follows, let $k : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a continuous and bounded symmetric positive definite kernel, that is, we have $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{Y}$ and

$$\sum_{i,j=1}^m k(x_i, x_j) c_i c_j \geq 0$$

for all choices of $x_1, \dots, x_m \in \mathcal{Y}$ and $c_1, \dots, c_m \in \mathbb{R}$. It is well known that k generates a so-called *reproducing kernel Hilbert space* (RKHS) [2, 4, 39] $(\mathbb{H}_{\mathcal{Y}}, \langle \cdot, \cdot \rangle_{\mathbb{H}_{\mathcal{Y}}})$ of bounded continuous functions on \mathcal{Y} , such that for $\psi \in \mathbb{H}_{\mathcal{Y}}$ the *reproducing property*

$$\psi(x) = \langle \psi, \Phi_{\mathcal{Y}}(x) \rangle_{\mathbb{H}_{\mathcal{Y}}}, \quad x \in \mathcal{Y}, \quad (3.9)$$

holds, where $\Phi_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathbb{H}_{\mathcal{Y}}$ denotes the so-called *feature map* corresponding to the kernel k , i.e.,

$$\Phi_{\mathcal{Y}}(x) = k(x, \cdot), \quad x \in \mathcal{Y}.$$

In the sequel, we shall denote the norm on $\mathbb{H}_{\mathcal{Y}}$ by $\| \cdot \|_{\mathbb{H}_{\mathcal{Y}}}$ and the kernel diagonal by φ :

$$\varphi(x) = k(x, x), \quad x \in \mathcal{Y}.$$

Then for $x \in \mathcal{Y}$ we have

$$\|\Phi_{\mathcal{Y}}(x)\|_{\mathbb{H}_{\mathcal{Y}}}^2 = \langle \Phi_{\mathcal{Y}}(x), \Phi_{\mathcal{Y}}(x) \rangle_{\mathbb{H}_{\mathcal{Y}}} = \langle k(x, \cdot), k(x, \cdot) \rangle_{\mathbb{H}_{\mathcal{Y}}} = k(x, x) = \varphi(x).$$

We shall frequently make use of the following estimate:

$$|k(x, y)| = |\langle \Phi_{\mathcal{Y}}(x), \Phi_{\mathcal{Y}}(y) \rangle_{\mathbb{H}_{\mathcal{Y}}}| \leq \|\Phi_{\mathcal{Y}}(x)\|_{\mathbb{H}_{\mathcal{Y}}} \|\Phi_{\mathcal{Y}}(y)\|_{\mathbb{H}_{\mathcal{Y}}} = \sqrt{\varphi(x)\varphi(y)}.$$

For example, it shows that $\|k\|_{\infty} = \|\varphi\|_{\infty}$.

In contrast to other works, where the kernel is induced by an observation map [12] or is specifically designed by taking samples of the dynamics into account [5], we admit any continuous kernel on \mathcal{Y} which satisfies the following

Compatibility assumptions:

- (C1) If $\psi \in L^2(\mathcal{Y}; \mu)$ such that $\int_{\mathcal{Y}} \int_{\mathcal{Y}} k(x, y) \psi(x) \psi(y) d\mu(x) d\mu(y) = 0$, then $\psi = 0$ μ -a.e.
- (C2) If $\psi \in \mathbb{H}_{\mathcal{Y}}$ such that $\psi(y) = 0$ for μ -a.e. $y \in \mathcal{Y}$, then $\psi = 0$.

Since the functions in $\mathbb{H}_{\mathcal{Y}}$ are continuous, note that condition (C2) holds if the measure μ is absolutely continuous with respect to Lebesgue measure with positive density. The condition allows us to consider $\mathbb{H}_{\mathcal{Y}}$ as a subspace of $L^2(\mathcal{Y}; \mu)$, i.e., the inclusion operator $S_{\mathcal{Y}} : \mathbb{H}_{\mathcal{Y}} \rightarrow L^2(\mathcal{Y}; \mu)$, defined by

$$S_{\mathcal{Y}}\psi := [\psi], \quad \psi \in \mathbb{H}_{\mathcal{Y}},$$

is injective. Here, $[\psi]$ denotes the equivalence class w.r.t. a.e. equality. It is easily seen that $S_{\mathcal{Y}}$ is bounded with norm $\|S_{\mathcal{Y}}\| \leq \|\varphi\|_{1, \mathcal{Y}}^{1/2}$. Hence, $\mathbb{H}_{\mathcal{Y}}$ is continuously embedded in $L^2(\mathcal{Y}; \mu)$, i.e.,

$$\|\psi\|_{2, \mathcal{Y}} \lesssim \|\psi\|_{\mathbb{H}_{\mathcal{Y}}}, \quad \psi \in \mathbb{H}_{\mathcal{Y}}.$$

Furthermore, the condition (C1) is equivalent to the density of $\mathbb{H}_{\mathcal{Y}}$ in $L^2(\mathcal{Y}; \mu)$, see [42]. Hence, $\mathbb{H}_{\mathcal{Y}}$ is continuously and densely embedded in $L^2(\mathcal{Y}; \mu)$.

In fact, the embedding is even compact as $S_{\mathcal{Y}}$ is a *Hilbert-Schmidt operator* with Hilbert-Schmidt norm $\|S_{\mathcal{Y}}\|_{HS}^2 = \|\varphi\|_{1, \mathcal{Y}}$, see [42]. For this, recall that an operator $T \in L(\mathcal{H})$ on a Hilbert space \mathcal{H} is *trace class* if for some (and hence for each) orthonormal basis (ONB) $(e_j)_{j \in \mathbb{N}}$ of \mathcal{H} we have $\sum_{j=1}^{\infty} \langle (T^*T)^{1/2} e_j, e_j \rangle < \infty$. An operator $S \in L(\mathcal{H}, \mathcal{K})$ between Hilbert spaces \mathcal{H} and \mathcal{K} is said to be

Hilbert-Schmidt [14, Chapter III.9] if S^*S is trace class, i.e., $\|S\|_{HS}^2 := \sum_{j=1}^{\infty} \|Se_j\|^2 < \infty$ for some (and hence for each) ONB $(e_j)_{j \in \mathbb{N}}$ of \mathcal{H} . Trace-class and Hilbert-Schmidt operators are compact operators. In particular, their spectrum consists of discrete eigenvalues with the origin as the only accumulation point. As $S_{\mathcal{Y}}$ is Hilbert-Schmidt, it follows that the *covariance operator*

$$C_{\mathcal{Y}} := S_{\mathcal{Y}}^* S_{\mathcal{Y}} \in L(\mathbb{H}_{\mathcal{Y}})$$

is trace-class. Mercer's theorem shows the existence of a particular ONB $(e_j)_{j=1}^{\infty}$ of $L^2(\mathcal{Y}; \mu)$ composed of eigenfunctions of $S_{\mathcal{Y}} S_{\mathcal{Y}}^*$, which we shall henceforth call the *Mercer basis* corresponding to the kernel k .

Theorem 3.5 (Mercer's Theorem [45]). *There exists an ONB $(e_j)_{j=1}^{\infty}$ of $L^2(\mathcal{Y}; \mu)$ comprised of eigenfunctions of $S_{\mathcal{Y}} S_{\mathcal{Y}}^*$ with corresponding eigenvalues $\lambda_j > 0$ such that $\sum_{j=1}^{\infty} \lambda_j = \|\varphi\|_{1, \mathcal{Y}} < \infty$. Furthermore, $(f_j)_{j=1}^{\infty}$ with $f_j = \sqrt{\lambda_j} e_j$ constitutes an ONB of $\mathbb{H}_{\mathcal{Y}}$ consisting of eigenfunctions of $C_{\mathcal{Y}}$ with corresponding eigenvalues λ_j . Moreover, for all $x, y \in \mathcal{Y}$,*

$$k(x, y) = \sum_j f_j(x) f_j(y) = \sum_j \lambda_j e_j(x) e_j(y),$$

where the series converges absolutely.

The restriction $k_{\mathcal{X}}$ of the kernel k to $\mathcal{X} \times \mathcal{X}$ is certainly also a continuous and bounded symmetric positive definite kernel. Hence, the discussion above applies with \mathcal{Y} replaced by \mathcal{X} , which introduces the symbols $\mathbb{H}_{\mathcal{X}}$, $\Phi_{\mathcal{X}}$, $S_{\mathcal{X}}$, and $C_{\mathcal{X}}$. Note that condition (C1) for \mathcal{Y} implies (C1) for \mathcal{X} . However, the analogous implication does not apply to (C2), which we explicitly assume here for \mathcal{X} :

(C2 $_{\mathcal{X}}$) If $\psi \in \mathbb{H}_{\mathcal{X}}$ such that $\psi(x) = 0$ for μ -a.e. $x \in \mathcal{X}$, then $\psi = 0$ on \mathcal{X} .

Then all the statements above also hold for \mathcal{Y} replaced by \mathcal{X} . Again, (C2 $_{\mathcal{X}}$) holds if the measure μ is absolutely continuous with respect to Lebesgue measure with positive density. By [39, Corollary 5.8],

$$\mathbb{H}_{\mathcal{X}} = \{\psi|_{\mathcal{X}} : \psi \in \mathbb{H}_{\mathcal{Y}}\}$$

and

$$\|\psi\|_{\mathbb{H}_{\mathcal{X}}} = \inf\{\|\eta\|_{\mathbb{H}_{\mathcal{Y}}} : \eta|_{\mathcal{X}} = \psi\}.$$

Covariance and cross-covariance operators. For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ define the rank-one operator $\Phi(x) \otimes \Phi(y) : \mathbb{H}_{\mathcal{Y}} \rightarrow \mathbb{H}_{\mathcal{X}}$ by

$$[\Phi(x) \otimes \Phi(y)] \psi := \langle \psi, \Phi(y) \rangle_{\mathbb{H}_{\mathcal{Y}}} \Phi(x) = \psi(y) \Phi(x), \quad \psi \in \mathbb{H}_{\mathcal{Y}}.$$

Clearly, if $x \in \mathcal{X}$, then $\Phi(x) \otimes \Phi(x) : \mathbb{H}_{\mathcal{X}} \rightarrow \mathbb{H}_{\mathcal{X}}$. The *covariance operator* $C_{\mathcal{X}} = S_{\mathcal{X}}^* S_{\mathcal{X}}$ then has the representation

$$C_{\mathcal{X}} = \int_{\mathcal{X}} \Phi(x) \otimes \Phi(x) d\mu(x).$$

For $t \geq 0$ and $\psi \in \mathbb{H}_{\mathcal{Y}}$ we further define the *cross-covariance operator* $C_{\mathcal{X}\mathcal{Y}}^t : \mathbb{H}_{\mathcal{Y}} \rightarrow \mathbb{H}_{\mathcal{X}}$ by

$$C_{\mathcal{X}\mathcal{Y}}^t := \int_{\mathcal{X}} \int_{\mathcal{Y}} \Phi(x) \otimes \Phi(y) \rho_t(x, dy) d\mu(x) = S_{\mathcal{X}}^* K^t S_{\mathcal{Y}}.$$

As the product of the two Hilbert-Schmidt operators $S_{\mathcal{X}}^* K^t$ and $S_{\mathcal{Y}}$, the operator $C_{\mathcal{X}\mathcal{Y}}^t$ is trace class for all $t \geq 0$ (cf. [19, p. 521]). Note that for $t = 0$ we have $C_{\mathcal{X}\mathcal{Y}}^t \psi = S_{\mathcal{X}}^* S_{\mathcal{X}} \psi|_{\mathcal{X}} = C_{\mathcal{X}} \psi|_{\mathcal{X}}$. Moreover, for all $\eta \in \mathbb{H}_{\mathcal{X}}$ and $\psi \in \mathbb{H}_{\mathcal{Y}}$ we have

$$\langle \eta, C_{\mathcal{X}\mathcal{Y}}^t \psi \rangle_{\mathbb{H}_{\mathcal{X}}} = \langle \eta, K^t \psi \rangle_{2, \mathcal{X}}. \quad (3.10)$$

3.5. Data samples and Kernel EDMD. Next, we briefly recall the Koopman approximation method called *Extended Dynamic Mode Decomposition* (EDMD), where certain (usually continuous) functions from $L^2(\mathcal{Y}; \mu)$ are evaluated on a set of sampled data.

Assumption 3.6. The data (x_k, y_k) , $k = 0, \dots, m-1$, are drawn i.i.d. from the joint distribution

$$d\mu_{0,t}(x, y) := \rho_t(x, dy) d\mu(x)$$

on $\mathcal{X} \times \mathcal{Y}$, where the time-step $t > 0$ is fixed. This then implies that the x_k are sampled i.i.d. with respect to μ and y_k is drawn dependently of x_k from the probability distribution $\rho_t(x_k, \cdot)$.

The EDMD Algorithm [56]. In EDMD, one chooses functions in $L^2(\mathcal{Y}; \mu)$ from a so-called *dictionary* $\mathcal{D} = \{\psi_1, \dots, \psi_N\}$ spanning the finite-dimensional space $\mathbb{V} = \text{span}\{\psi_1, \dots, \psi_N\} \subset L^2(\mathcal{Y}; \mu)$ with the goal of approximating the compression $P_{\mathbb{V}}K^t|_{\mathbb{V}}$ of K^t to \mathbb{V} . Here, $P_{\mathbb{V}}$ denotes the orthogonal projection onto the space \mathbb{V} in $L^2(\mathcal{Y}; \mu)$. The matrix representation of the compression operator with respect to the basis \mathcal{D} of \mathbb{V} is then given by $C^{-1}A$, where $C = (\langle \psi_i, \psi_j \rangle_{\mu})_{i,j=1}^N$ is the so-called mass matrix and $A = (\langle \psi_i, K^t \psi_j \rangle_{\mu})_{i,j=1}^N$ is called the stiffness matrix. In order to define empirical estimators for C and A , we replace the scalar products in the matrices' entries by sums of $\psi_i(x_k)\psi_j(x_k)$ and $\psi_i(x_k)\psi_j(y_k)$, respectively, which leads to the matrices

$$\widehat{C}_m = \frac{1}{m} \Psi(X) \Psi(X)^{\top} \quad \text{and} \quad \widehat{A}_m = \frac{1}{m} \Psi(X) \Psi(Y)^{\top},$$

where $\Psi = [\psi_1, \dots, \psi_N]^{\top}$ and

$$\Psi(X) = [\Psi(x_0), \dots, \Psi(x_{m-1})] \quad \text{and} \quad \Psi(Y) = [\Psi(y_0), \dots, \Psi(y_{m-1})].$$

Hence, the EDMD approximation of $P_{\mathbb{V}}K^t|_{\mathbb{V}}$ is given by the estimator $\widehat{M}^{m,t} = \widehat{C}_m^{-1} \widehat{A}_m$. In [37] it was proved that $\widehat{M}^{m,t}$ indeed converges in probability to $C^{-1}A$ as $m \rightarrow \infty$, and bounds on the accuracy were provided.

The kEDMD Algorithm [57, 22]. In contrast to traditional EDMD, in *kernel EDMD* (kEDMD) the dictionary is not chosen a priori, but is given by the features sampled on the data points, that is, $\mathcal{D} = \{\Phi(x_0), \dots, \Phi(x_{m-1})\}$. The data matrices $\Psi(X)$ and $\Psi(Y)$ are then easily seen to be given by

$$\Psi(X) = (k(x_i, x_j))_{i,j=0}^{m-1} \quad \text{and} \quad \Psi(Y) = (k(x_i, y_j))_{i,j=0}^{m-1}.$$

Hence, since $\Psi(X)$ is symmetric,

$$\widehat{C}_m^{-1} \widehat{A}_m = \left[\frac{1}{m} \Psi(X) \Psi(X)^{\top} \right]^{-1} \left[\frac{1}{m} \Psi(X) \Psi(Y)^{\top} \right] = \Psi(X)^{-1} \Psi(Y)^{\top}.$$

However, since the eigenvalues of $\Psi(X)$ typically decay fast, in practice one often replaces the inverse of the ill-conditioned matrix $\Psi(X)$ by a pseudo-inverse of a rank- r ($r < m$) eigendecomposition truncation of $\Psi(X)$. Hence, the kEDMD estimator for the Koopman operator compression becomes

$$\widehat{M}_r^{m,t} = [\Psi(X)]_r^{\dagger} \Psi(Y)^{\top}. \quad (3.11)$$

Now, the empirical estimators for the operators $C_{\mathcal{X}}$ and $C_{\mathcal{X}\mathcal{Y}}^t$ are given by

$$\widehat{C}_{\mathcal{X}}^m = \frac{1}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes \Phi(x_k) \quad \text{and} \quad \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t} = \frac{1}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes \Phi(y_k).$$

We have $\widehat{C}_{\mathcal{X}}^m : \mathbb{H}_{\mathcal{X}} \rightarrow \mathbb{H}_{\mathcal{X}}$ and $\widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t} : \mathbb{H}_{\mathcal{Y}} \rightarrow \mathbb{H}_{\mathcal{X}}$. In fact, both operators map to $\mathbb{V} = \text{span } \mathcal{D} \subset \mathbb{H}_{\mathcal{X}}$. Since

$$\widehat{C}_{\mathcal{X}}^m \Phi(x_j) = \frac{1}{m} \sum_{k=0}^{m-1} [\Phi(x_k) \otimes \Phi(x_k)] \Phi(x_j) = \frac{1}{m} \sum_{k=0}^{m-1} k(x_j, x_k) \Phi(x_k),$$

$$\widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}\Phi(x_j) = \frac{1}{m} \sum_{k=0}^{m-1} [\Phi(x_k) \otimes \Phi(y_k)]\Phi(x_j) = \frac{1}{m} \sum_{k=0}^{m-1} k(x_j, y_k)\Phi(x_k),$$

the matrix representations of $\widehat{C}_{\mathcal{X}}^m|_{\mathbb{V}}$ and $\widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}|_{\mathbb{V}}$ with respect to the basis $\mathcal{D} = \{\Phi(x_0), \dots, \Phi(x_{m-1})\}$ are given by $\frac{1}{m}\Psi(X)$ and $\frac{1}{m}\Psi(Y)^\top$, respectively. Thus, the kEDMD estimator matrix $\widehat{M}_r^{m,t}$ in (3.11) represents the operator $[\widehat{C}_{\mathcal{X}}^m]_r^\dagger \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}|_{\mathbb{V}}$. In the next section, we shall give probabilistic bounds on the $\|\cdot\|_{\mathbb{H}_Y \rightarrow L^2(\mathcal{X}; \mu)}$ -error between $P_r K^t$ and

$$\widehat{K}_r^{m,t} := [\widehat{C}_{\mathcal{X}}^m]_r^\dagger \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}, \quad (3.12)$$

which then serves as a quantification of the prediction error made by kEDMD. Here, P_r denotes the orthogonal projection in $L^2(\mathcal{X}; \mu)$ onto the span of eigenvectors corresponding to the largest r eigenvalues of $C_{\mathcal{X}}$.

4. BOUND ON THE KOOPMAN PREDICTION ERROR

Before we present our main result, let us bound the error made by estimating the operators $C_{\mathcal{X}}$ and $C_{\mathcal{X}\mathcal{Y}}$ by their empirical estimators $\widehat{C}_{\mathcal{X}}^m$ and $\widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}$. Since $C_{\mathcal{X}}$ and $C_{\mathcal{X}\mathcal{Y}}$ are Hilbert-Schmidt, we may quantify the error in the Hilbert-Schmidt norm.

Proposition 4.1. *For every $t \geq 0$, the following probabilistic bounds on the Hilbert-Schmidt estimation errors holds:*

$$\begin{aligned} \mathbb{P}(\|C_{\mathcal{X}\mathcal{Y}}^t - \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}\|_{HS} > \varepsilon) &\leq 2e^{-\frac{m\varepsilon^2}{8\|k\|_\infty^2}} \\ \mathbb{P}(\|C_{\mathcal{X}} - \widehat{C}_{\mathcal{X}}^m\|_{HS} > \varepsilon) &\leq 2e^{-\frac{m\varepsilon^2}{8\|k\|_\infty^2}}. \end{aligned}$$

Proof. For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ let us abbreviate $C_x := \Phi(x) \otimes \Phi(x)$ and $C_{xy} := \Phi(x) \otimes \Phi(y)$. If $(f_i) \subset \mathbb{H}_Y$ denotes the Mercer ONB of \mathbb{H}_Y corresponding to k on $\mathcal{Y} \times \mathcal{Y}$, for $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$ we have

$$\langle C_{xy}, C_{x'y'} \rangle_{HS} = \sum_i \langle C_{xy} f_i, C_{x'y'} f_i \rangle_{\mathbb{H}_X} = \sum_i f_i(y) f_i(y') k(x, x') = k(x, x') k(y, y').$$

This proves $\|C_{xy}\|_{HS}^2 = \varphi(x)\varphi(y)$. Moreover, it yields

$$\|C_{\mathcal{X}\mathcal{Y}}^t\|_{HS}^2 = \left\| \int C_{xy} d\mu_{0,t}(x, y) \right\|_{HS}^2 = \int \int k(x, x') k(y, y') d\mu_{0,t}(x, y) d\mu_{0,t}(x', y') \leq \|k\|_\infty^2.$$

Hence, since the C_{x_k, y_k} are independent, $\mathbb{E}[C_{\mathcal{X}\mathcal{Y}}^t - C_{x_k, y_k}] = 0$, and

$$\|C_{\mathcal{X}\mathcal{Y}}^t - C_{x_k, y_k}\|_{HS} \leq \|C_{\mathcal{X}\mathcal{Y}}^t\|_{HS} + \|C_{x_k, y_k}\|_{HS} \leq 2\|k\|_\infty,$$

we may apply Hoeffding's inequality for Hilbert space-valued random variables [43, Theorem 3.5] (see also [32, Theorem A.5.2]) to $C_{\mathcal{X}\mathcal{Y}}^t - C_{x_k, y_k}$ and obtain

$$\mathbb{P}(\|C_{\mathcal{X}\mathcal{Y}}^t - \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m,t}\|_{HS} > \varepsilon) = \mathbb{P}\left(\left\| \frac{1}{m} \sum_{k=0}^{m-1} (C_{\mathcal{X}\mathcal{Y}}^t - C_{x_k, y_k}) \right\|_{HS} > \varepsilon\right) \leq 2e^{-\frac{m\varepsilon^2}{8\|k\|_\infty^2}}.$$

The proof of the second claim follows analogous lines. \square

Let (e_j) be the Mercer orthonormal basis of $L^2(\mathcal{X}; \mu)$ corresponding to the kernel k and let $\lambda_j = \|S_{\mathcal{X}} S_{\mathcal{X}}^* e_j\|_{2, \mathcal{X}}$ as well as $f_j := \sqrt{\lambda_j} e_j \in \mathbb{H}_{\mathcal{X}}$ (cf. Theorem 3.5). We arrange the Mercer eigenvalues λ_j of $C_{\mathcal{X}}$ in a non-increasing way, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \dots$$

and denote by P_r the orthogonal projection in $L^2(\mathcal{X}; \mu)$ onto $\text{span}\{e_1, \dots, e_r\}$.

Theorem 4.2. *Fix an arbitrary $r \in \mathbb{N}$ and assume that the first $r + 1$ eigenvalues λ_j of $C_{\mathcal{X}}$ are simple, i.e., $\lambda_{j+1} < \lambda_j$ for all $j = 1, \dots, r$, and let*

$$\delta_r = \min_{j=1, \dots, r} \frac{\lambda_j - \lambda_{j+1}}{2} \quad \text{and} \quad c_r = \frac{1}{\sqrt{\lambda_r}} + \frac{r+1}{\delta_r \lambda_r} (1 + \|\varphi\|_{1, \mathcal{Y}}) \|\varphi\|_{1, \mathcal{X}}^{1/2}. \quad (4.1)$$

Further, let $\varepsilon \in (0, \delta_r)$ and $\delta \in (0, 1)$ be arbitrary, fix some $m \geq \max\{r, \frac{8\|k\|_{\infty}^2 \ln(4/\delta)}{\varepsilon^2}\}$ and define $\widehat{K}_r^{m, t}$ as in (3.12). Then with probability at least $1 - \delta$, we have that

$$\|P_r K^t - \widehat{K}_r^{m, t}\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow L^2(\mathcal{X}; \mu)} \leq c_r \varepsilon.$$

If, furthermore, $K^t \mathbb{H}_{\mathcal{Y}} \subset \mathbb{H}_{\mathcal{X}}$, then, with probability at least $1 - \delta$,

$$\|K^t - \widehat{K}_r^{m, t}\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow L^2(\mathcal{X}; \mu)} \leq \sqrt{\lambda_{r+1}} \|K^t\|_{\mathbb{H}_{\mathcal{Y}} \rightarrow \mathbb{H}_{\mathcal{X}}} + c_r \varepsilon. \quad (4.2)$$

Proof of Theorem 4.2. Let $\widehat{\lambda}_1 \geq \dots \geq \widehat{\lambda}_r$ denote the largest r eigenvalues of $\widehat{C}_{\mathcal{X}}^m$ in descending order and let $\widehat{e}_1, \dots, \widehat{e}_r$ be corresponding eigenfunctions, respectively, such that $\|\widehat{e}_j\|_{\mathbb{H}_{\mathcal{X}}} = \widehat{\lambda}_j^{-1/2}$ for $j = 1, \dots, m$. If we set $\widehat{f}_j = \widehat{\lambda}_j^{1/2} \cdot \widehat{e}_j$, then for $\psi \in \mathbb{H}_{\mathcal{Y}}$ we have

$$\widehat{K}_r^{m, t} \psi = [\widehat{C}_{\mathcal{X}}^m]_r^\dagger \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m, t} \psi = \sum_{j=1}^r \widehat{\lambda}_j^{-1} \langle \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m, t} \psi, \widehat{f}_j \rangle_{\mathbb{H}_{\mathcal{X}}} \widehat{f}_j = \sum_{j=1}^r \langle \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m, t} \psi, \widehat{e}_j \rangle_{\mathbb{H}_{\mathcal{X}}} \widehat{e}_j,$$

which estimates (see (3.10))

$$\sum_{j=1}^r \langle C_{\mathcal{X}\mathcal{Y}}^t \psi, e_j \rangle_{\mathbb{H}_{\mathcal{X}}} e_j = \sum_{j=1}^r \langle K^t \psi, e_j \rangle_{2, \mathcal{X}} e_j = P_r K^t \psi.$$

Hence, we have to estimate the deviation of $\sum_{j=1}^r \langle \widehat{C}_{\mathcal{X}\mathcal{Y}}^{m, t} \psi, \widehat{e}_j \rangle_{\mathbb{H}_{\mathcal{X}}} \widehat{e}_j$ from $\sum_{j=1}^r \langle C_{\mathcal{X}\mathcal{Y}}^t \psi, e_j \rangle_{\mathbb{H}_{\mathcal{X}}} e_j$, which was done in [42] for the special case of $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. The proof for the present case follows analogous lines. Also the estimate in the case $K^t \mathbb{H}_{\mathcal{Y}} \subset \mathbb{H}_{\mathcal{X}}$ can be proved by following the lines of the proof of the counterpart in [42]. \square

Remark 4.3. The assumption $K^t \mathbb{H}_{\mathcal{Y}} \subset \mathbb{H}_{\mathcal{X}}$ is not too exotic. In fact, we prove in Section D in the Appendix that the condition holds for the Ornstein-Uhlenbeck process, if $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}$ is any compact set with non-empty interior, and k is a Gaussian RBF kernel.

5. EXTENSION TO DETERMINISTIC CONTROL-AFFINE SYSTEMS

In this part, we extend the error estimates to deterministic control systems of the form

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n_u} u_i(t) g_i(x(t)), \quad t \in [0, T], \quad (5.1)$$

with vector fields $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and $g_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $i \in [1 : n_u] := \{1, \dots, n_u\}$, $n_u \in \mathbb{N}$, and a prediction horizon $T > 0$. We let \mathcal{X} and \mathcal{Y} be compact sets in \mathbb{R}^d and choose specifically

$$d\mu(x) = \frac{1}{|\mathcal{X}|} dx,$$

where $|\mathcal{X}|$ denotes the Lebesgue measure of \mathcal{X} on \mathbb{R}^d . We furthermore assume that the restrictions of the vector fields f and g_1, \dots, g_{n_u} to \mathcal{Y} are C^1 . Admissible controls u in (5.1) are measurable functions $u : [0, T] \rightarrow \mathcal{U}$, where $\mathcal{U} \subset \mathbb{R}^{n_u}$ is a compact, convex set satisfying $0 \in \text{int}(\mathcal{U})$. To avoid excessive notation, we assume that for all $x^0 \in \mathcal{X}$ and all admissible controls $u : [0, T] \rightarrow \mathcal{U}$,

- $[0, T]$ is contained in the existence interval for the solution $x(\cdot; x^0, u)$ to (5.1) and
- $x(t; x^0, u) \in \mathcal{Y}$ for each $t \in [0, T]$.

If the first assumption is satisfied and \mathcal{X} is contained in the interior of \mathcal{Y} , then by continuity of the flow in time and the compactness of both \mathcal{X} and \mathcal{U} , there exists a time horizon T_0 such that the second assumption is satisfied for all $T \in (0, T_0)$.

Here, we consider sampled-data systems with Zero-order Hold (ZoH), i.e., for a given, typically small sampling period $t > 0$, $u(s) \equiv u$ on $[kt, (k+1)t)$, $k \in \mathbb{N}_0$. This system class is well established in systems and control, see, e.g., [8, 33] and the references therein. Sampled-data systems can be used, e.g., for output-feedback stabilization of nonlinear systems [30] or in the redesign of continuous-time control signals [17]. In particular, there are many results showing that system-theoretic properties like stabilizability, or even cost controllability, are preserved under sampling [58]. The latter even is inherited by the EDMD-based surrogate model in the Koopman framework as recently shown in [6] analyzing EDMD-based Model Predictive Control. Hence, we also restrict our consideration to sampled-data systems knowing that the analysis on one sampling period suffices to obtain similar results based on the concatenation of the control signal for multiple sampling periods.

For a constant control $u \in \mathcal{U}$ we let K_u^t be the Koopman operator for the autonomous system (5.1), that is,

$$(K_u^t \psi)(x^0) := \psi(x(t; x^0, u)), \quad x^0 \in \mathcal{X}, \psi \in L^2(\mathcal{Y}).$$

In order to control the system by piecewise constant controls, in principle, one could compute the kEDMD estimator for K_u^t for every applied control u , which, however, would require a lot of computation time in each step. As this is infeasible in practice, we circumvent this issue by approximating K_u^t by a linear combination of $n_u + 1$ Koopman operators $K_{u_i}^t$ with quadratic precision. In this course, we leverage the affine representation of the Koopman generator.

The Koopman generator \mathcal{L}^u associated with (5.1) is given by

$$\mathcal{L}^u \psi = \lim_{t \rightarrow 0} \frac{K_u^t \psi - \psi|_{\mathcal{X}}}{t} \quad (5.2)$$

for $\psi \in \text{dom } \mathcal{L}^u$, where the limit is taken in $L^2(\mathcal{X})$, and $\text{dom } \mathcal{L}^u$ consists of those functions $\psi \in L^2(\mathcal{Y})$ for which this L^2 -limit exists. It is easy to see that $C^1(\mathcal{Y}) \subset \text{dom } \mathcal{L}^u$ for each $u \in \mathcal{U}$ such that for $\psi \in C^1(\mathcal{Y})$ and $x^0 \in \mathcal{X}$,

$$\begin{aligned} (\mathcal{L}^u \psi)(x^0) &= \lim_{t \rightarrow 0} \frac{1}{t} ((K_u^t \psi)(x^0) - \psi(x^0)) = \frac{d}{dt} (K_u^t \psi)(x^0) \Big|_{t=0} = \frac{d}{dt} \psi(x(t; x^0, u)) \Big|_{t=0} \\ &= \nabla \psi(x^0)^\top \left(f(x^0) + \sum_{i=1}^{n_u} u_i g_i(x^0) \right). \end{aligned} \quad (5.3)$$

Let $\{e_i\}_{i=1}^{n_u}$ denote the standard basis of \mathbb{R}^{n_u} and let

$$\gamma_i := \inf\{\gamma > 0 : \gamma e_i \notin \mathcal{U}\}, \quad i = 1, \dots, n_u.$$

Then the γ_i are positive numbers thanks to $0 \in \text{int}(\mathcal{U})$, and we have $\gamma_i e_i \in \partial \mathcal{U} \subset \mathcal{U}$. Now, if $u = \sum_{i=1}^{n_u} u_i e_i$ is the coordinate representation of the vector $u \in \mathcal{U}$, the representation (5.3) reveals that, on

$C^1(\mathcal{Y})$,

$$\mathcal{L}^u = \mathcal{L}^0 + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\mathcal{L}^{\gamma_i e_i} - \mathcal{L}^0). \quad (5.4)$$

To transfer the error bounds from autonomous systems derived in this work to control systems, we will show that the restriction of the Koopman operator to $C^2(\mathcal{Y})$ is also affine in the control—at least approximately for small $t > 0$. For this, set

$$f_u(x) := f(x) + \sum_{i=1}^{n_u} u_i g_i(x)$$

and let $\psi \in C^2(\mathcal{Y})$, $u \in \mathcal{U}$, and $x^0 \in \mathcal{X}$. We apply Taylor's theorem to $\psi \circ x(\cdot; x^0, u)$ at $t_0 = 0$ and observe that

$$\begin{aligned} (K_u^t \psi - \psi|_{\mathcal{X}})(x^0) &= \psi(x(t; x^0, u)) - \psi(x(0; x^0, u)) \\ &= t \nabla \psi(x^0) \cdot f_u(x^0) + \frac{1}{2} r(\xi) t^2, \end{aligned}$$

for some $\xi \in [0, t]$, where, abbreviating $x(s) = x(s; x^0, u)$,

$$r(s) = f_u(x(s))^\top \nabla^2 \psi(x(s)) f_u(x(s)) + \nabla \psi(x(s))^\top D f_u(x(s)) f_u(x(s)).$$

By compactness of \mathcal{U} , \mathcal{X} , and \mathcal{Y} and the continuity of all functions involved, we can bound the remainder term uniformly in $u \in \mathcal{U}$, $x^0 \in \mathcal{X}$ and $s \in [0, T]$. More specifically,

$$\max_{s \in [0, T]} |r(s)| \leq c(f, g_1, \dots, g_{n_u}) \cdot \|\psi\|_{C^2(\mathcal{Y})}, \quad (5.5)$$

where $\|\psi\|_{C^2(\mathcal{Y})} := \|\psi\|_{C(\mathcal{Y})} + \|\nabla \psi\|_{C(\mathcal{Y})^d} + \|\nabla^2 \psi\|_{C(\mathcal{Y})^{d \times d}}$. Thus,

$$K_u^t \psi = \psi + t \mathcal{L}^u \psi + \mathcal{O}(t^2) \quad (5.6)$$

where the constant in $\mathcal{O}(t^2)$ as defined in (5.5) can be chosen independently of u .

Invoking control-affinity of the generator (5.4), we thus get

$$K_u^t \psi = K^t \psi + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (K_{\gamma_i e_i}^t - K^t) \psi + \mathcal{O}(t^2), \quad (5.7)$$

which means that the Koopman operator is control affine up to second order in time. This is Equation (3.14) in [40], where we stress however that the remainder term here depends on the chosen dictionary function.

The error bound (4.2) obtained in Theorem 4.2 can now be transferred to the controlled case. Thus, for $u \in \mathcal{U}$ we define the estimator

$$\widehat{K}_{r,u}^{m,t} := \widehat{K}_r^{m,t} + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\widehat{K}_{r,\gamma_i e_i}^{m,t} - \widehat{K}_r^{m,t}),$$

where the $\widehat{K}_{r,\gamma_i e_i}^{m,t}$ are defined as in (3.12) for the system $\dot{x} = f(x) + \gamma_i g_i(x)$.

Theorem 5.1. *Let the assumptions in Theorem 4.2 be satisfied. For fixed $r \in \mathbb{N}$, let δ_r and c_r be as in (4.1). Further, let $\varepsilon \in (0, \delta_r)$ and $\delta \in (0, 1)$ be arbitrary and fix some $m \geq \max\{r, \frac{8\|k\|_\infty^2 \ln(4(n_u+1)/\delta)}{\varepsilon^2}\}$. Then, for each $\psi \in \mathbb{H}_{\mathcal{Y}} \cap C^2(\mathcal{X})$, there is a constant $c_\psi \geq 0$ such that with probability at least $1 - \delta$ for all $u \in \mathcal{U}$ we have*

$$\|P_r K_u^t \psi - \widehat{K}_{r,u}^{m,t} \psi\|_{L^2(\mathcal{X}; \mu)} \leq (1 + 2\|\Gamma^{-1} u\|_1) c_r \varepsilon \|\psi\|_{\mathbb{H}_{\mathcal{Y}}} + c_\psi t^2. \quad (5.8)$$

where $\Gamma = \text{diag}_{i=1, \dots, n_u} \gamma_i$. In the case where $K_u^t \mathbb{H}_y \subset \mathbb{H}_X$ for all $u \in \mathcal{U}_\partial := \{0, \gamma_1 e_1, \dots, \gamma_{n_u} e_{n_u}\}$, we obtain

$$\|K_u^t \psi - \widehat{K}_{r,u}^{m,t} \psi\|_{L^2(\mathcal{X}; \mu)} \leq (1 + 2\|\Gamma^{-1}u\|_1)(\sqrt{\lambda_{r+1}}M + c_r \varepsilon) \|\psi\|_{\mathbb{H}_y} + c_\psi t^2, \quad (5.9)$$

where $M = \max\{\|K_u^t\|_{\mathbb{H}_y \rightarrow \mathbb{H}_X} : u \in \mathcal{U}_\partial\}$.

If the kernel k enjoys C^2 -regularity, for $m \geq \max\{r, \frac{8\|k\|_\infty^2 \ln(4/\delta)}{\varepsilon^2}\}$ data points we have with probability at least $1 - \delta$ that

$$\|P_r K_u^t - \widehat{K}_{r,u}^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X}; \mu)} \leq c_r \varepsilon + t^2 \cdot C_k \widehat{\lambda}_r^{-1}$$

with a constant C_k depending only on the kernel k . Here, $\widehat{\lambda}_r$ denotes the r -th largest eigenvalue of the operator \widehat{C}_X^m . Again, if $K_u^t \mathbb{H}_y \subset \mathbb{H}_X$ for all $u \in \mathcal{U}_\partial$, then

$$\|K_u^t - \widehat{K}_{r,u}^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X}; \mu)} \leq \sqrt{\lambda_{r+1}}M + c_r \varepsilon + t^2 \cdot C_k \widehat{\lambda}_r^{-1}.$$

Proof. Using the approximate control-affinity (5.7), we compute

$$(P_r K_u^t - \widehat{K}_{r,u}^{m,t})\psi = (P_r K^t - \widehat{K}_r^{m,t})\psi + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} \left(P_r K_{\gamma_i e_i}^t - \widehat{K}_{r, \gamma_i e_i}^{m,t} - P_r K^t + \widehat{K}_r^{m,t} \right) \psi + t^2 \phi,$$

where $\|\phi\|_{L^\infty(\mathcal{X})} \leq c_\psi$ for all choices of $u \in \mathcal{U}$ and $t \in [0, T]$. By the choice of m , the respective bounds in Theorem 4.2 on $\|P_r K^t - \widehat{K}_r^{m,t}\|$ and $\|P_r K_{\gamma_i e_i}^t - \widehat{K}_{r, \gamma_i e_i}^{m,t}\|$, $i \in \{1, \dots, n_u\}$, hold with probability at least $1 - \delta(n_u + 1)^{-1}$ each. Hence, the probability for all $n_u + 1$ bounds to hold exceeds $1 - \delta$, cf. [37, Lemma 22]. In that event,

$$\begin{aligned} & \left\| (P_r K^t - \widehat{K}_r^{m,t})\psi + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} \left(P_r K_{\gamma_i e_i}^t - \widehat{K}_{r, \gamma_i e_i}^{m,t} - P_r K^t + \widehat{K}_r^{m,t} \right) \psi \right\|_{L^2(\mathcal{X})} \\ & \leq \|P_r K^t - \widehat{K}_r^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X})} \|\psi\|_{\mathbb{H}_y} \\ & \quad + \left(\max_{i \in \{1, \dots, n_u\}} \|P_r K_{\gamma_i e_i}^t - \widehat{K}_{r, \gamma_i e_i}^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X})} + \|P_r K^t - \widehat{K}_r^{m,t}\|_{\mathbb{H}_y \rightarrow L^2(\mathcal{X})} \right) \|\psi\|_{\mathbb{H}_y} \sum_{i=1}^{n_u} \frac{|u_i|}{\gamma_i} \\ & \leq (1 + 2\|\Gamma^{-1}u\|_1) c_r \varepsilon \|\psi\|_{\mathbb{H}_y}, \end{aligned}$$

which proves the given estimate. The proof for the case $K_u^t \mathbb{H}_y \subset \mathbb{H}_X$ for all $u \in \mathcal{U}_\partial$ follows analogous lines, making use of (4.2).

Now, let $k \in C^2$ and denote the flow of $\dot{x} = f_v(x)$ by F_v , $v \in \mathcal{U}$. Also, let $F := F_0$. In addition to the random variables $y_k = F^t(x_k)$, $k = 0, \dots, m-1$, let $y_{k,v} = F_v^t(x_k)$, $i = 1, \dots, n_u$. Moreover, let $\widehat{C}_{\mathcal{X}^y, v}^{m,t}$ be the empirical estimator of the cross-covariance operator $C_{\mathcal{X}^y, v}^t$ for the system $\dot{x} = f_v(x)$. Then we have

$$\begin{aligned} \widehat{K}_{r,u}^{m,t} &= \widehat{K}_r^{m,t} + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\widehat{K}_{r, \gamma_i e_i}^{m,t} - \widehat{K}_r^{m,t}) = [\widehat{C}_X^m]_r^\dagger \left[\widehat{C}_{\mathcal{X}^y}^{m,t} + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\widehat{C}_{\mathcal{X}^y, \gamma_i e_i}^{m,t} - \widehat{C}_{\mathcal{X}^y}^{m,t}) \right] \\ &= [\widehat{C}_X^m]_r^\dagger \frac{1}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes \left[\Phi(F^t(x_k)) + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\Phi(F_{\gamma_i e_i}^t(x_k)) - \Phi(F^t(x_k))) \right] \\ &= [\widehat{C}_X^m]_r^\dagger \frac{1}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes \left[\mathbb{K}^t \Phi(x_k) + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\mathbb{K}_{\gamma_i e_i}^t \Phi(x_k) - \mathbb{K}^t \Phi(x_k)) \right], \end{aligned}$$

where $\mathbb{K}_v^t : L^2(\mathcal{Y}, \mathbb{H}_y) \rightarrow L^2(\mathcal{X}, \mathbb{H}_y)$ for $v \in \mathcal{U}$ is defined by $\mathbb{K}_v^t \Psi(x) = \Psi(F_v^t(x))$, $\Psi \in L^2(\mathcal{Y}, \mathbb{H}_y)$, and $\mathbb{K}^t := \mathbb{K}_0^t$. As $\Phi \in C^2$, we can, analogously as in the discussion preceding the theorem, see that

$$\mathbb{K}^t \Phi + \sum_{i=1}^{n_u} \frac{u_i}{\gamma_i} (\mathbb{K}_{\gamma_i e_i}^t - \mathbb{K}^t) \Phi = \mathbb{K}_u^t \Phi + t^2 \Psi_{u,t}$$

with some $\Psi_{u,t} \in L^2(\mathcal{X}, \mathbb{H}_y)$ such that $C_\Phi = \sup\{\|\Psi_{u,t}\|_{L^\infty(\mathcal{X}, \mathbb{H}_y)} : u \in \mathcal{U}, t \in [0, T]\} < \infty$. Hence, we obtain

$$\begin{aligned} \widehat{K}_{r,u}^{m,t} &= [\widehat{C}_{\mathcal{X}}^m]_r^\dagger \frac{1}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes [\Phi(F_u^t(x_k)) + t^2 \Psi_{u,t}(x_k)] \\ &= [\widehat{C}_{\mathcal{X}}^m]_r^\dagger \left[\widehat{C}_{\mathcal{X}y,u}^{m,t} + \frac{t^2}{m} \sum_{k=0}^{m-1} \Phi(x_k) \otimes \Psi_{u,t}(x_k) \right]. \end{aligned}$$

Note that $[\widehat{C}_{\mathcal{X}}^m]_r^\dagger \widehat{C}_{\mathcal{X}y,u}^{m,t}$ is the estimator for K_u^t as defined in (3.12). Here, we shall denote it by $\widetilde{K}_{r,u}^{m,t}$. Now, if $\psi \in \mathbb{H}_y$, then

$$\begin{aligned} \|P_r K^t \psi - \widehat{K}_{r,u}^{m,t} \psi\|_{L^2(\mathcal{X})} &\leq \|P_r K^t \psi - \widetilde{K}_{r,u}^{m,t} \psi\|_{L^2(\mathcal{X})} + \|\widetilde{K}_{r,u}^{m,t} \psi - \widehat{K}_{r,u}^{m,t} \psi\|_{L^2(\mathcal{X})} \\ &\leq c_r \varepsilon \|\psi\|_{\mathbb{H}_y} + \frac{t^2}{m} \left\| \sum_{k=0}^{m-1} [\widehat{C}_{\mathcal{X}}^m]_r^\dagger [\Phi(x_k) \otimes \Psi_{u,t}(x_k)] \psi \right\|_{L^2(\mathcal{X})} \\ &\leq c_r \varepsilon \|\psi\|_{\mathbb{H}_y} + \frac{t^2}{m} \sum_{k=0}^{m-1} |\langle \psi, \Psi_{u,t}(x_k) \rangle_{\mathbb{H}_y}| \cdot \|[\widehat{C}_{\mathcal{X}}^m]_r^\dagger \Phi(x_k)\|_{L^2(\mathcal{X})} \\ &\leq c_r \varepsilon \|\psi\|_{\mathbb{H}_y} + \|\varphi\|_{1,\mathcal{X}}^{1/2} \cdot \frac{t^2}{m} \sum_{k=0}^{m-1} |\langle \psi, \Psi_{u,t}(x_k) \rangle_{\mathbb{H}_y}| \cdot \|[\widehat{C}_{\mathcal{X}}^m]_r^\dagger \Phi(x_k)\|_{\mathbb{H}_\mathcal{X}} \\ &\leq c_r \varepsilon \|\psi\|_{\mathbb{H}_y} + \|k\|_\infty^{1/2} \cdot \frac{t^2}{m} \sum_{k=0}^{m-1} \|\psi\|_{\mathbb{H}_y} \|\Psi_{u,t}(x_k)\|_{\mathbb{H}_y} \widehat{\lambda}_r^{-1} \sqrt{\varphi(x_k)} \\ &\leq c_r \varepsilon \|\psi\|_{\mathbb{H}_y} + \widehat{\lambda}_r^{-1} C_\Phi \|k\|_\infty \cdot t^2 \|\psi\|_{\mathbb{H}_y}, \end{aligned}$$

which proves the claim with $C_k = C_\Phi \|k\|_\infty$. \square

6. NUMERICAL RESULTS

In this part, we present two numerical examples. First we consider an autonomous stochastic system to illustrate the results of Section 4. Second, we present numerical results for a controlled deterministic system to showcase the bounds deduced in Section 5.

6.1. Estimation Error for Ornstein-Uhlenbeck Process. We consider a one-dimensional Ornstein-Uhlenbeck (OU) process obeying the stochastic differential equation

$$dX_t = -\alpha X_t + \sqrt{2\beta^{-1}} dW_t, \quad (6.1)$$

with $\alpha = 1$ and $\beta = 2$ in the numerical tests. The initial domain is chosen as $\mathcal{X} = [-1.5, 1.5]$, the final domain is simply the entire real line $\mathcal{Y} = \mathbb{R}$. The measure μ is chosen as the invariant distribution of the OU process on \mathcal{Y} , which is normal with mean zero and standard deviation $(\alpha\beta)^{-1/2}$. We then use the restriction of μ to the finite interval $[-1.5, 1.5]$ to generate initial conditions for the SDE (6.1).

Mirroring the analysis in [42], we consider the approximation of the cross-covariance operator $C_{\mathcal{X}\mathcal{Y}}^t$ at lag time $t = 0.05$, corresponding to five elementary time steps at integration time step $\Delta_t = 0.01$. We use the Gaussian radial basis function kernel

$$k(x, y) = \exp \left[-\frac{(x - y)^2}{\sigma^2} \right],$$

for bandwidths $\sigma \in \{0.05, 0.1, 0.5\}$. Setting the confidence level to $1 - \delta = 0.9$, we use Proposition 4.1 to bound the minimal guaranteed error that is achieved with this confidence, for a range of data sizes m between $m = 20$ and $m = 5 \cdot 10^4$. As a comparison, we estimate the actual incurred error for fifty independent collections of m pairs (x_k, y_k) each, and plot the $(1 - \delta)$ -percentile for this error as a function of m . To do so, we first observe that the Hilbert-Schmidt norm of the empirical covariance operator $\widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m}$ can be calculated using the Mercer features and eigenvalues of the Gaussian kernel:

$$\|\widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m}\|_{\text{HS}}^2 = \sum_{i,j=1}^{\infty} \lambda_i \lambda_j (M_{ij}^m)^2, \quad (6.2)$$

using the following matrix elements for the Mercer features:

$$M_{ij}^m := \frac{1}{m} \sum_{l=1}^m e_i(x_l) e_j(y_l). \quad (6.3)$$

In practice, we truncate the infinite sum in (6.2) after $i_{\max} = j_{\max} = 30$ terms. We then first generate ten large reference collections (x_l, y_l) , $1 \leq l \leq m_{\max}$, where $m_{\max} = 10^7$, and average the matrix elements $M_{ij}^{m_{\max}}$ over these large reference samples. The actual error for $m < m_{\max}$ data pairs (x_k, y_k) , $1 \leq k \leq m$ is then estimated by the formula

$$\|C_{\mathcal{X}\mathcal{Y}}^t - \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m}\|_{\text{HS}}^2 \approx \|\widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m_{\max}} - \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m}\|_{\text{HS}}^2 = \|\widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m_{\max}}\|_{\text{HS}}^2 - 2 \left\langle \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m_{\max}}, \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m} \right\rangle_{\text{HS}} + \|\widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m}\|_{\text{HS}}^2.$$

The Hilbert-Schmidt inner product can be computed by the formula

$$\left\langle \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m_{\max}}, \widehat{C}_{\mathcal{X}\mathcal{Y}}^{t,m} \right\rangle_{\text{HS}} = \sum_{i,j=1}^{\infty} \lambda_i \lambda_j M_{ij}^{m_{\max}} M_{ij}^m.$$

The results are shown in Figure 1. On average, the bounds from Proposition 4.1 over-estimate the actual incurred error by about one order of magnitude.

6.2. Duffing Oscillator. To conclude this study, we illustrate that kernel EDMD, combined with state-of-the-art model validation and low-rank approximation techniques, enables accurate prediction of a non-linear control system with very few necessary design choices on the modeler's end. We study the two-dimensional Duffing oscillator, governed by the ordinary differential equation

$$\frac{d}{dt} \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} z_2 \\ -\alpha z_1 u - 2\beta z_1^3 \end{bmatrix}, \quad (6.4)$$

which has been used as a model system in numerous previous studies. We set $\alpha = -1$, $\beta = 1$. By $z(t; z^0, u)$, we denote the state at time $t \geq 0$ emanating from the initial condition z^0 at time zero and the control $u : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}$. We generate $m = 10^4$ uniformly sampled initial conditions $\{x_k\}_{k=1}^m$ from the square $\mathcal{X} = [-1.5, 1.5]^2$. Next, for fixed control inputs $\bar{u} \in \{0, 1\}$, we compute the successor state $y_k^{\bar{u}} = z(t; x_k, \bar{u})$, $k = 1, \dots, m$ for $t = 0.025$ via numerical integration with time step $\Delta_t = 0.005$. The data pairs $\{(x_k, y_k^{\bar{u}})\}_{k=1}^m$ serve as training data for the kEDMD algorithm corresponding to the flow with constant control $\bar{u} \in \{0, 1\}$. As shown in Figure 2, the time-shifted points $y_k^{\bar{u}}$ can indeed exit the initial domain \mathcal{X} , such that invariance of \mathcal{X} is not given. This in particular underlines the necessity of incorporating an additional set \mathcal{Y} as done in the previous analysis.

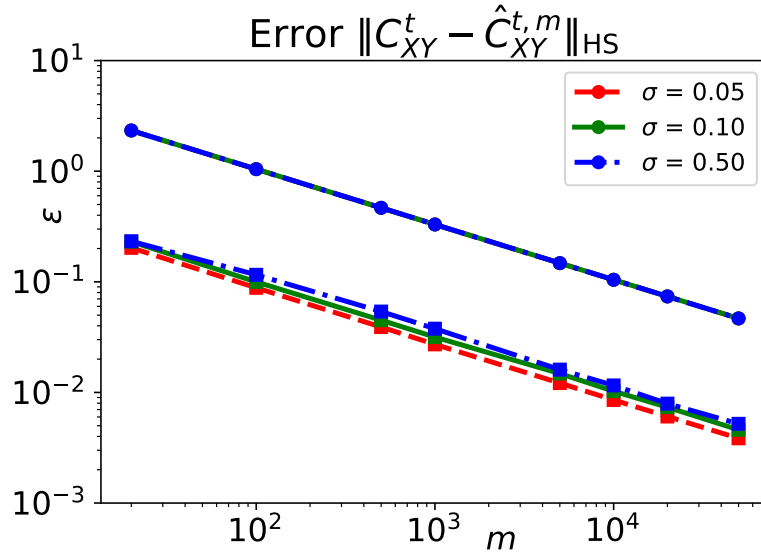


FIGURE 1. Ornstein-Uhlenbeck process: Error bound from Prop. 4.1 (circles) versus $(1 - \delta)$ -percentile of the actual estimation error estimated from fifty independent trials (squares), for three different bandwidths σ indicated by different colors and linestyles.

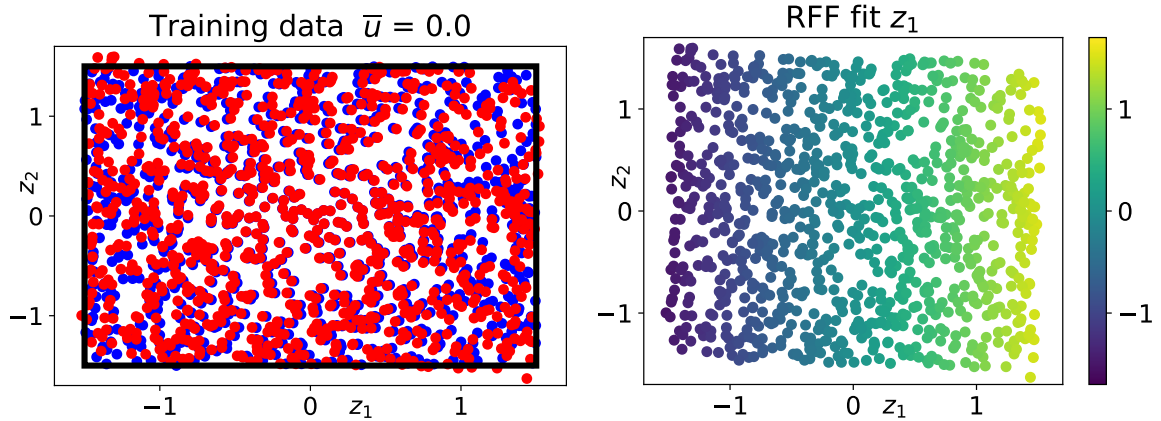


FIGURE 2. Left: initial training data x_k (blue) and corresponding time shifted data y_k (red) for fixed input $\bar{u} = 0$. The black square indicates the initial domain $\mathcal{X} = [-1.5, 1.5]^2$. Right: Evaluation of the state variable x_k on the data $y_k^{\bar{u}}$, approximated in terms of the RFF basis at $\sigma = 1.0$ and $p = 500$.

As reproducing kernel, we again select the Gaussian RBF kernel. To reduce the computational effort, we approximate the resulting feature maps using random Fourier features (RFFs) [44]. It was shown in [36] that this is equivalent to applying plain EDMD with a randomly selected basis set of complex exponentials

$$\psi_i(x) = e^{i\omega_i^\top x}, \quad 1 \leq i \leq p \in \mathbb{N},$$

where $\omega_i \in \mathbb{R}^d$, $1 \leq i \leq p$, are random frequencies drawn from the spectral measure associated with the kernel. In our case, the spectral measure is a normal distribution with standard deviation σ^{-1} , i.e. the

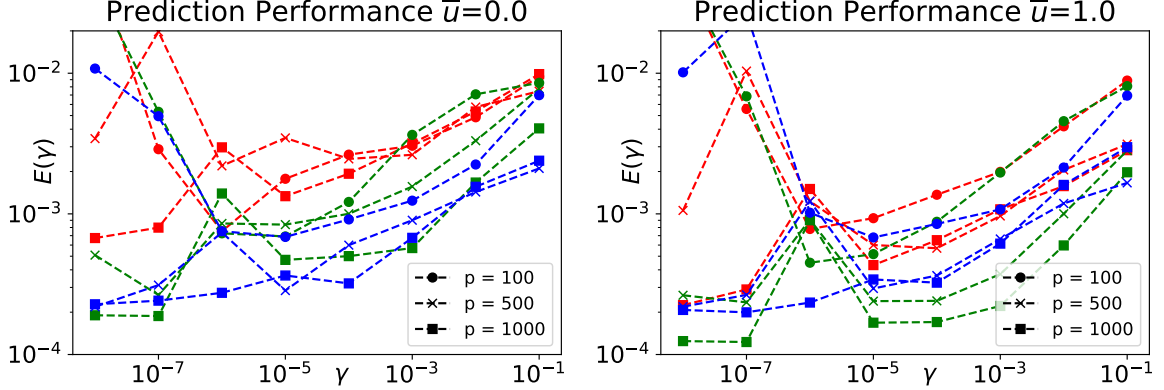


FIGURE 3. Duffing oscillator: Mean-squared prediction error of the trained Koopman model for $\bar{u} = 0.0$ (left) and $\bar{u} = 1.0$ (right), tested on 100 new trajectories of length $T_{\text{test}} = 20t = 0.5$. The error is shown as a function of the regularization strength $\gamma > 0$ for different feature sizes $p \in \mathbb{N}$ (dots, x-es, squares) and bandwidths $\sigma \in \{0.5, 1.0, 1.5\}$ (blue, green, red).

inverse of the kernel bandwidth. Using the feature matrices $\mathbf{M}, \mathbf{M}_t^{\bar{u}} \in \mathbb{C}^{p \times m}$ defined by

$$\mathbf{M}_{i,k} = e^{i\omega_i^\top x_k}, \quad (\mathbf{M}_t^{\bar{u}})_{i,k} = e^{i\omega_i^\top y_k^{\bar{u}}}, \quad 1 \leq i \leq p, \quad 1 \leq k \leq m,$$

the RFF Koopman matrix corresponding to the flow with constant control $\bar{u} \in \{0, 1\}$ amounts to

$$\mathbf{K}_{\bar{u}} = (\mathbf{M}\mathbf{M}^H + \gamma\text{Id})^{-1} \mathbf{M}(\mathbf{M}_t^{\bar{u}})^H,$$

where $\gamma > 0$ is a regularization parameter.

We begin by tuning the bandwidth $\sigma > 0$, the feature size $p \in \mathbb{N}$, and the tolerance $\gamma > 0$. To this end, we draw a test ensemble of 100 random initial conditions, from which we generate simulations of the ODE (6.4) until time $T_{\text{test}} = 20t = 0.5$. We also use the learned Koopman models to predict the system state over the same time horizon, starting from the same initial conditions, and compute the relative mean-squared error between the true and predicted states over all time steps. Note that state prediction with the Koopman operator requires an approximation of the state variables z_1 and z_2 as linear combinations of the RFF basis functions ψ_i . These approximations are computed by a least-squares fit along with each Koopman model. An example is shown in Figure 2.

The results of the model validation are shown for different inputs $u : [0, T_{\text{test}}] \rightarrow \mathbb{R}$ and different hyper-parameters in Figure 3. We conclude that the prediction error generally decreases with regularization if $\gamma \geq 10^{-5}$, but seems to destabilize if the regularization is decreased further. We therefore identify $\sigma = 1.0$, $p = 500$ and $\gamma = 10^{-5}$ as suitable model parameters.

We then evaluate the long-time prediction performance of the bi-linear Koopman surrogate model

$$\mathbf{K}_{u(t)} \mathbf{w} = [\mathbf{K}_0 + (\mathbf{K}_1 - \mathbf{K}_0)u(t)] \mathbf{w} \quad (6.5)$$

for the lifted state $\mathbf{w} \in \mathbb{C}^p$, generating 50 trajectories for (a) constant control input $u \equiv 0$; (b) constant control input $u \equiv 1$; (c) time-dependent control input $u(t) = \cos(t)$. To obtain predictions corresponding to the real-valued system (6.4), we project the propagated complex-valued lifted state onto its real part. The time horizon covered by these trajectories is $T_{\text{long}} = 500t = 12.5$, and we use the project-and-lift scheme [15] every 25 steps to improve performance. Figure 4 shows the mean relative error over the real part of these trajectories as a function of simulation time. We also show exemplary trajectories for each of the three scenarios in the same figure, and observe excellent performance of the kernel-based bi-linear Koopman model.

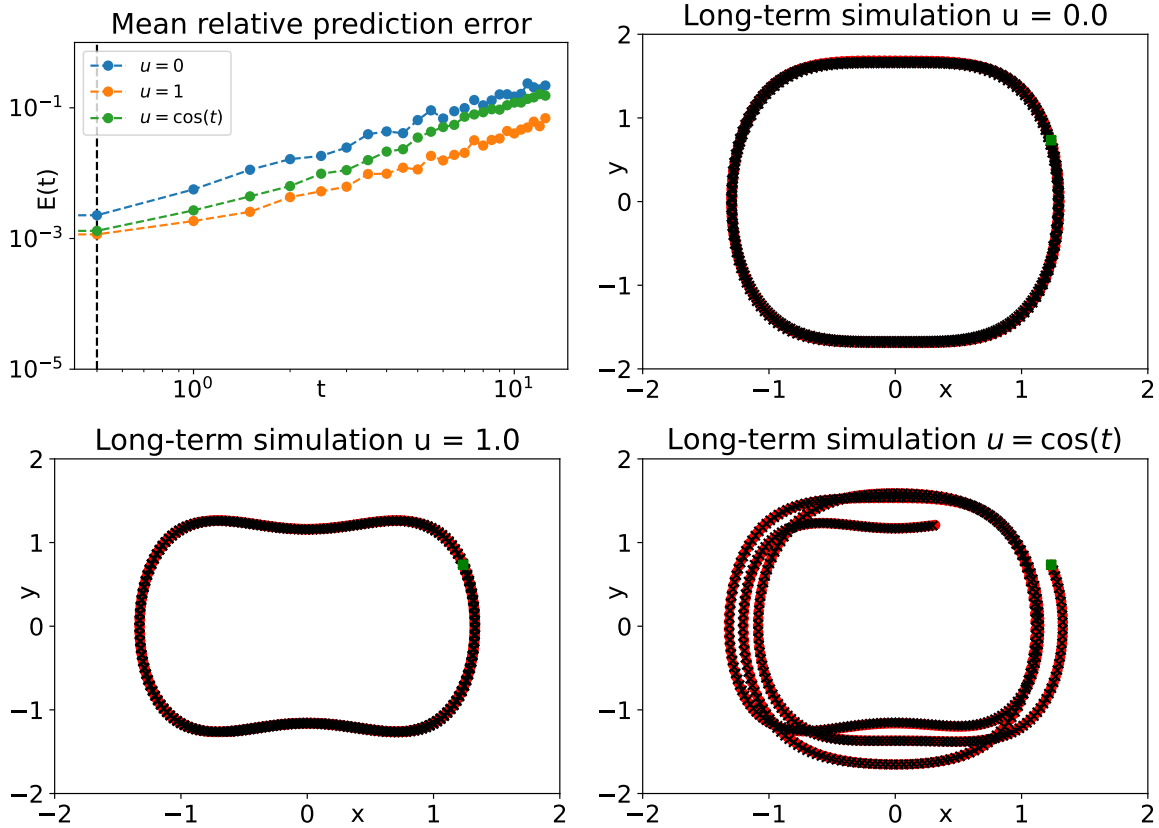


FIGURE 4. Upper left: Mean-squared prediction error for the Duffing oscillator predicted by the bi-linear Koopman model (6.5) for different control inputs, evaluated on 50 new trajectories of length $T_{\text{long}} = 500t = 12.5$. The black dashed line indicates the time horizon T_{test} used for model validation in Figure 3. The remaining panels show exemplary trajectories of the same length T_{long} for different control inputs. Red dots mark the predicted solution, the true ODE solution is shown in black. The initial value is highlighted in green.

7. SUPPLEMENTS

Acknowledgments. F. Philipp was funded by the Carl Zeiss Foundation within the project DeepTurb–Deep Learning in and from Turbulence. He was further supported by the free state of Thuringia and the German Federal Ministry of Education and Research (BMBF) within the project THInKI–Thüringer Hochschulinitiative für KI im Studium. K. Worthmann gratefully acknowledges funding by the German Research Foundation (DFG grant WO 2056/14-1, project no 507037103).

Data Availability. Codes and data required for reproducing the numerical results in this study are available at <https://doi.org/10.5281/zenodo.10378153>.

REFERENCES

- [1] H. Amann, Ordinary differential equations, An introduction to nonlinear analysis, de Gruyter, Berlin, New York, 1990.
- [2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950), 337–404.
- [3] D. Bakry, I. Gentil, and M. Ledoux, Analysis and Geometry of Markov Diffusion Operators, Grundlehren der mathematischen Wissenschaften, Springer International Publishing Switzerland, 2014.

- [4] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, 2004.
- [5] P. Bevanda, M. Beier, A. Lederer, S. Sosnowski, E. Hüllermeier, S. Hirche, *Koopman kernel regression*, Preprint, arXiv:2305.16215.
- [6] L. Bold, L. Grüne, M. Schaller, and K. Worthmann, *Practical asymptotic stability of data-driven model predictive control using extended DMD*, preprint arXiv:2308.00296 (2023).
- [7] S. L. Brunton, M. Budisic, E. Kaiser, and J. N. Kutz, *Modern Koopman theory for dynamical systems*, *SIAM Rev.* 64(2) (2022), 229–340.
- [8] B. Castillo, S. Di Gennaro, S. Monaco, and D. Normand-Cyrot, *On regulation under sampling*, *IEEE Transactions on Automatic Control* 42 (1997), 864–868.
- [9] Y. Chen and U. Vaidya, *Sample complexity for non-linear stochastic systems*, 2019 American Control Conference (ACC), Philadelphia, USA, July 10–12, 2019, pp. 3526–3531.
- [10] J.-M. Coron and L. Grüne and K. Worthmann, *Model predictive control, cost controllability, and homogeneity*, *SIAM Journal on Control and Optimization* 58(5) (2020), 2979–2996.
- [11] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, London Mathematical Society Lecture Note Series, Cambridge University Press, 1996.
- [12] S. Das, D. Giannakis, and J. Slawinska, *Reproducing kernel Hilbert space compactification of unitary evolution groups*, *Appl. Comput. Harmon. Anal.* 54 (2021), 75–136.
- [13] L.C. Evans and R.F. Gariepy, *Measure theory and fine properties of functions*, Taylor & Francis Group, CRC Press, 2015.
- [14] I.C. Gohberg and M.G. Krein, *Introduction To The Theory of Linear Nonselfadjoint Operators*, volume 18 of *Translations of Mathematical Monographs*. American Mathematical Society, 1969.
- [15] P. van Goor, R. Mahony, M. Schaller, K. Worthmann, *Reprojection methods for Koopman-based modelling and prediction*, preprint arXiv:2307.16188, 2023.
- [16] D. Goswami and D.A. Paley, *Bilinearization, reachability, and optimal control of control-affine nonlinear systems: A Koopman spectral approach*, *IEEE Transactions on Automatic Control* 67(6) (2021), 2715–2728.
- [17] L. Grüne and K. Worthmann, *Sampled-data redesign for nonlinear multi-input systems*, *Geometric Control And Non-smooth Analysis: In Honor of the 73rd Birthday of H. Hermes and of the 71st Birthday of R.T. Rockafellar*, *World Scientific* (2008), 206–227.
- [18] E. Kaiser, J.N. Kutz, and S.L. Brunton, *Data-driven discovery of Koopman eigenfunctions for control*, *Mach. Learn. Sci. Technol.* 2 (2021), 035023.
- [19] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag Berlin Heidelberg, 1995.
- [20] S. Klus, P. Koltai, and C. Schütte, *On the numerical approximation of the Perron–Frobenius and Koopman operator*, *J. Comput. Dyn.* 1 (2016), 51–79.
- [21] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, *Data-driven model reduction and transfer operator approximation*, *J. Nonlinear Sci.* 28(3) (2018), 985–1010.
- [22] S. Klus, I. Schuster, K. Muandet, *Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces*, *J. Nonlinear Sci.* 30 (2020), 283–315.
- [23] B.O. Koopman, *Hamiltonian Systems and transformations in Hilbert space*, *Proc. Natl. Acad. Sci.* 5 (1931), 315–318.
- [24] M. Korda and I. Mezić, *Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control*, *Automatica* 93 (2018), 149–160.
- [25] V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, and M. Pontil, *Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces*, *Adv. Neural. Inf. Process. Syst.* 35 (2022), 4017–4031.
- [26] V. Kostic, K. Lounici, P. Novelli, and M. Pontil, *Sharp spectral rates for Koopman operator learning*, Preprint, arXiv:2302.02004, 2023.
- [27] V. Kühner, *What can Koopmanism do for attractors in dynamical systems?*, *The Journal of Analysis*, 29 (2021), 449–471.
- [28] T. Lelièvre, M. Rousset and G. Stoltz, *Free Energy Computations*, Imperial College Press (2010)
- [29] T. Lelièvre and G. Stoltz, *Partial differential equations and stochastic methods in molecular dynamics*, *Acta Numer.* 25 (2016), 681–880.
- [30] Z. Li and J. Zhao, *Output feedback stabilization for a general class of nonlinear systems via sampled-data control*, *International Journal of Robust and Nonlinear Control* 28:7 (2018), 2853–2867.
- [31] A. Mauroy, Y. Susuki, and I. Mezić, *Koopman operator in systems and control*, Springer International Publishing, Berlin, 2020.
- [32] M. Mollenhauer, *On the Statistical Approximation of Conditional Expectation Operators*, Dissertation, Freie Universität Berlin, 2021.
- [33] D. Nešić and A.R. Teel, *Sampled-data control of nonlinear systems: an overview of recent results*, In: Moheimani, S.R. (eds) *Perspectives in robust control*. Lecture Notes in Control and Information Sciences 268. Springer, London (2007), 221–239.

- [34] F. Nüske, P. Gelß, S. Klus, and C. Clementi, Tensor-based computation of metastable and coherent sets, *Physica D* 427 (2021), 133018.
- [35] F. Nüske, B.G. Keller, G. Pérez-Hernández, A.S.J.S. Mey, and F. Noé, Variational approach to molecular kinetics, *J. Chem. Theory Comput.* 10 (2014), 1739–1752.
- [36] F. Nüske and S. Klus, Efficient approximation of molecular kinetics using random Fourier features, *J. Chem. Phys.* 159 (7): 074105
- [37] F. Nüske, S. Peitz, F. Philipp, M. Schaller, and K. Worthmann, Finite-data error bounds for Koopman-based prediction and control, *J. Nonlinear Sci.* 33 (2023), 1–34.
- [38] B. Øksendal, *Stochastic Differential Equations, An Introduction with Applications*, Fifth Edition, Corrected Printing, Springer-Verlag Heidelberg New York, 2000.
- [39] V.I. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*, Cambridge Studies in Advanced Mathematics 152, Cambridge University Press, 2016.
- [40] S. Peitz, S.E. Otto, and C.W. Rowley, Data-driven model predictive control using interpolated Koopman generators, *SIAM Journal on Applied Dynamical Systems*, 19(3), 2162–2193 (2020).
- [41] F. Philipp, Bessel orbits of normal operators, *J. Math. Anal. Appl.* 448 (2017), 767–785.
- [42] F. Philipp, M. Schaller, K. Worthmann, S. Peitz and F. Nüske, Error bounds for kernel-based approximations of the Koopman operator. Submitted, preprint arXiv:2301.08637 (2023).
- [43] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *The Annals of Probability* 22 (1994), 1679–1706.
- [44] A. Rahimi and B. Recht, Random features for large-scale kernel machines, *Advances in Neural Information Processing Systems* 20 (2007)
- [45] F. Riesz and B. Nagy, *Functional Analysis*, Blackie & Son Ltd., Glasgow, Bombay, Toronto, 1955.
- [46] W. Rudin, *Real and Complex Analysis*, Third edition, McGraw-Hill, Inc., 1987.
- [47] M. Schaller, K. Worthmann, F. Philipp, S. Peitz, and F. Nüske, Towards reliable data-based optimal and predictive control using extended DMD, to appear in *IFAC PapersOnLine*, arXiv preprint arXiv:2202.09084.
- [48] P.J. Schmid, Dynamic mode decomposition of numerical and experimental data, *Journal of fluid mechanics*, 656, 5–28, 2010.
- [49] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science+Business Media, LLC, 2008.
- [50] I. Steinwart, D. Hush, and C. Scovel, An explicit description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF kernels, *IEEE Transactions on Information Theory* 52 (2006), 4635–4643.
- [51] R. Strässer, J. Berberich, and F. Allgöwer, Robust data-driven control for nonlinear systems using the Koopman operator, preprint arXiv:2304.03519, 2023.
- [52] R. Strässer, M. Schaller, K. Worthmann, J. Berberich, F. Allgöwer Koopman-based feedback design with stability guarantees, preprint arXiv:2312.01441, 2023.
- [53] A. Surana, Koopman operator based observer synthesis for control-affine nonlinear systems, In 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 6492–6499. IEEE, 2016.
- [54] W. Walter, *Ordinary differential equations*, Springer Science + Business Media, 1998.
- [55] M.O. Williams, M.S. Hemati, S. Dawson, I.G. Kevrekidis, and C. W. Rowley, Extending data-driven Koopman analysis to actuated systems, *IFAC-PapersOnLine* 49, no. 18 (2016), 704–709.
- [56] M.O. Williams, I.G. Kevrekidis, and C.W. Rowley, A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition, *J. Nonlinear Sci.* 25 (2015), 1307–1346.
- [57] M.O. Williams, C.W. Rowley, and I. Kevrekidis, A kernel-based method for data-driven Koopman spectral analysis, *J. Comput. Dyn.* 2(2) (2015), 247–265.
- [58] K. Worthmann, M. Reble, L. Grüne, and F. Allgöwer, The role of sampling for stability and performance in unconstrained nonlinear model predictive control, *SIAM Journal on Control and Optimization* 52:1 (2014), 581–605.
- [59] C. Zhang and E. Zuazua, A quantitative analysis of Koopman operator methods for system identification and predictions, *Comptes Rendus. Mécanique*, Online first (2023), 1–31.

APPENDIX A. THE CONDITION (3.5) FOR DETERMINISTIC SYSTEMS

Let $\mathcal{X} = \overline{\Omega}$, where $\Omega \subset \mathbb{R}^d$ is open and bounded. We consider the flow $F : [0, T] \times \mathcal{X} \rightarrow \mathbb{R}^d$ of a given autonomous ODE $\dot{x} = b(x)$. Let $\mathcal{Y} \subset \mathbb{R}^d$ such that

$$\mathcal{Z} := F([0, T] \times \mathcal{X})$$

is contained in \mathcal{Y} . In this section, we provide conditions on the flow and a finite measure μ on \mathcal{Y} which guarantee that condition (3.8) holds, i.e.,

$$\mu(F_t^{-1}(A)) \leq L \cdot \mu(A) \quad \forall A \in \mathcal{B}(\mathcal{Y}), \forall t \in [0, T].$$

First of all, we note that it is sufficient to prove the above for \mathcal{Y} replaced by \mathcal{Z} . Indeed, then for $A \in \mathcal{B}(\mathcal{Y})$ we have

$$\mu(F_t^{-1}(A)) = \mu(F_t^{-1}(A \cap \mathcal{Z})) \leq L \cdot \mu(A \cap \mathcal{Z}) \leq L \cdot \mu(A).$$

Proposition A.1. *The condition (3.8) is satisfied whenever the following conditions are met:*

- The flow $F : [0, \tilde{T}] \times \mathcal{X} \rightarrow \mathbb{R}^d$ of the ODE $\dot{x} = b(x)$ is defined for some $\tilde{T} > T$ and is C^1 .
- $d\mu(x) = \varphi(x) dx$ with a positive density $\varphi \in L^\infty(\mathcal{X})$ satisfying $1/\varphi \in L^\infty(\mathcal{Z})$.

Note that the first assumption is satisfied if $b \in C^1(\mathcal{Y})$, cf. [1, Theorem 10.3].

Proof of Proposition A.1. First of all, we show that

$$\delta := \inf \{ \det F'_t(x) : x \in \mathcal{X}, t \in [0, T] \} > 0. \quad (\text{A.1})$$

Since $F_0(z) = z$ for all $z \in \mathcal{Z}$, we have $F'_0(z) = I_d$ for all $z \in \mathcal{Z}$. Let $t_1 := \tilde{T} - T$. By assumption, the flow $F : [0, t_1] \times \mathcal{Z} \rightarrow \mathbb{R}^d$ is defined on \mathcal{Z} and is C^1 . Hence, there exists some $t_0 \in (0, t_1)$ such that $\det F'_t(z) \geq \frac{1}{2}$ for all $t \in [0, t_0]$ and all $z \in \mathcal{Z}$. Choose t_0 such that $T = nt_0$ for some $n \in \mathbb{N}$. We prove by induction that for all $k \in \{0, \dots, n-1\}$ we have

$$\forall t \in [0, t_0] \forall x \in \mathcal{X} : \det F'_{kt_0+t}(x) \geq 2^{-k-1}. \quad (\text{A.2})$$

Then (A.1) holds with $\delta = 2^{-n}$. We have proved (A.2) already for $k = 0$. So, let (A.2) hold for some $k < n-1$, and let $t \in [0, t_0]$ and $x \in \mathcal{X}$. Then

$$\det F'_{(k+1)t_0+t}(x) = \det D_x F_{t_0}(F_{kt_0+t}(x)) = \det F'_{t_0}(F_{kt_0+t}(x)) \cdot \det F'_{kt_0+t}(x) \geq \frac{1}{2} \cdot \frac{1}{2^{k+1}} = \frac{1}{2^{k+2}},$$

which concludes the proof of (A.1).

In order to prove (3.8), we may restrict the flow to $F : [0, T] \times \Omega \rightarrow \mathbb{R}^d$ as $\mu(\partial\Omega) = 0$. Let $t \in [0, T]$, set $\mathcal{Z}_t := F_t(\Omega)$, and let $U \subset \mathbb{R}^d$ be open. Note that \mathcal{Z}_t is open due to the invariance of domain theorem³. Hence, also $U_t = U \cap \mathcal{Z}_t$ is open. By the change of variables formula we have

$$\begin{aligned} \mu(F_t^{-1}(U)) &= \mu(F_t^{-1}(U_t)) = \int_{F_t^{-1}(U_t)} \varphi(x) dx = \int_{U_t} \varphi(F_t^{-1}(y)) |\det DF_t^{-1}(y)| dy \\ &= \int_{U_t} \frac{\varphi(F_t^{-1}(y))}{|\det F'_t(F_t^{-1}(y))| \varphi(y)} \varphi(y) dy \leq \frac{\|\varphi\|_\infty \|\frac{1}{\varphi}\|_\infty}{\delta} \mu(U_t) \leq \frac{\|\varphi\|_\infty \|\frac{1}{\varphi}\|_\infty}{\delta} \mu(U). \end{aligned}$$

The claim now follows from the outer regularity of the measures μ and $\mu \circ F_t^{-1}$. \square

APPENDIX B. PROOF OF PROPOSITION 3.3

Let $\psi \in B(\mathcal{Y})$. Then $|(K^t\psi)(x)| \leq \int_{\mathcal{Y}} |\psi(y)| \rho_t(x, dy) \leq \|\psi\|_\infty$, hence $K^t\psi \in B(\mathcal{X})$ with $\|K^t\psi\|_\infty \leq \|\psi\|_\infty$ and $\|K^t\psi\|_p \leq \|\psi\|_\infty$. Let now $1 \leq p < \infty$. Then, by Hölder's inequality and (3.7),

$$\int_{\mathcal{X}} |(K^t\psi)(x)|^p d\mu(x) \leq \int_{\mathcal{X}} \int_{\mathcal{Y}} |\psi(y)|^p \rho_t(x, dy) d\mu(x) \leq L \int_{\mathcal{Y}} |\psi(y)|^p d\mu(y),$$

and hence $\|K^t\psi\|_{p, \mathcal{X}} \leq L^{1/p} \|\psi\|_{p, \mathcal{Y}}$.

Let $\psi \in C_c(\mathcal{Y})$ (the set of continuous functions on \mathcal{Y} with compact support). For fixed $x \in \mathcal{X}$, denote the stochastic solution process of the SDE (3.1) with initial value x by X_t^x . Since $X_t^x(\omega)$ is continuous

³or simply the inverse function theorem thanks to (A.1)

in t a.s. (see [38, Theorem 5.2.1]) and $X_t^x \in \mathcal{Y}$ a.s., we have $\psi(X_t^x(\omega)) \rightarrow \psi(X_0^x(\omega)) = \psi(x)$ as $t \rightarrow 0$ a.s. Hence, by dominated convergence,

$$K^t \psi(x) = \mathbb{E}[\psi(X_t^x)] = \int \psi(X_t^x(\omega)) d\mathbb{P}(\omega) \rightarrow \psi(x)$$

as $t \rightarrow 0$. By the first part of this proof,

$$|K^t \psi(x) - \psi(x)| \leq |K^t \psi(x)| + |\psi(x)| \leq 2\|\psi\|_\infty, \quad x \in \mathcal{X},$$

hence, dominated convergence implies that $\|K^t \psi - \psi\|_{p, \mathcal{X}} \rightarrow 0$ as $t \rightarrow 0$.

If $\psi \in L^p(\mathcal{Y}; \mu)$ and $\varepsilon > 0$, there exists $\eta \in C_c(\mathcal{Y})$ such that $\|\psi - \eta\|_{p, \mathcal{Y}} < \frac{\varepsilon}{2}(1 + L^{1/p})^{-1}$, see [46, Theorem 3.14]. Further, choose $\delta > 0$ such that $\|K^t \eta - \eta\|_{p, \mathcal{X}} < \frac{\varepsilon}{2}$ for $t < \delta$. Then

$$\|K^t \psi - \psi\|_{p, \mathcal{X}} \leq \|K^t(\psi - \eta)\|_{p, \mathcal{X}} + \|K^t \eta - \eta\|_{p, \mathcal{X}} + \|\eta - \psi\|_{p, \mathcal{Y}} < \varepsilon$$

for $t < \delta$, which proves the claim. \square

APPENDIX C. AN EXAMPLE: LACK OF CONTRACTIVITY OF THE KOOPMAN SEMIGROUP

Consider the deterministic case with the scalar ODE

$$\dot{x} = x(2 - x) \quad \text{on } \mathcal{X} = [1, 2].$$

The solution of this ODE is given by

$$x(t; x_0) = \frac{2e^{2t}x_0}{2 - x_0 + e^{2t}x_0}, \quad x_0 \in \mathcal{X},$$

which is non-decreasing and approaches 2 as $t \rightarrow \infty$ for any $x_0 \in \mathcal{X}$. In particular, \mathcal{X} is invariant under the flow of the ODE. Since \mathcal{X} is compact, the convergence towards the equilibrium $x = 2$ is uniform in $x_0 \in \mathcal{X}$. We will use this fact to show that the Koopman semigroup is not a bounded semigroup on $L^2(\mathcal{X})$, i.e., it does not satisfy $\|K^t\|_{L(L^2(\mathcal{X}))} \leq M$ for some $M \geq 0$ and all $t \geq 0$. To show this, it suffices to find $\psi \in L^2(1, 2)$ such that for any $M \geq 0$, there is $t \geq 0$ such that $\|K^t \psi\|_{L^2(\mathcal{X})} \geq M\|\psi\|_{L^2(\mathcal{X})}$.

Let $\psi(x) := \frac{1}{(2-x)^{1/4}}$. Then $\psi \in L^2(\mathcal{X})$ as

$$\int_1^2 \psi^2(x) dx = \int_1^2 \frac{1}{\sqrt{2-x}} dx = \int_0^1 \frac{1}{\sqrt{x}} dx = 2\sqrt{x} \Big|_0^1 = 2 < \infty.$$

Further, we have that $\psi(x) \rightarrow \infty$ as $x \rightarrow 2$ and ψ is non-decreasing such that

$$\forall M \geq 0 \exists \varepsilon > 0 : \psi(x) \geq M \quad \forall x \in (2 - \varepsilon, 2). \quad (\text{C.1})$$

Moreover, as the flow is non-decreasing towards 2:

$$\forall \varepsilon \geq 0 \exists t^* \geq 0 \forall t \geq t^* \forall x_0 \in \mathcal{X} \setminus \{2\} : x(t; x_0) \in (2 - \varepsilon, 2). \quad (\text{C.2})$$

Hence, for given M , choose ε as in (C.1) and correspondingly t^* as in (C.2). Then $x(t; x_0) \in (2 - \varepsilon, 2)$ for all $t \geq t^*$ and $x_0 \in \mathcal{X} \setminus \{2\}$, hence

$$\forall t \geq t^* : \|K^t \psi\|_{L^2(\mathcal{X})}^2 = \int_1^2 |\psi(x(t; x_0))|^2 dx_0 \geq M^2 = \frac{1}{2}M^2 \|\psi\|_{L^2(\mathcal{X})}^2.$$

APPENDIX D. ON THE CONDITION $K^t \mathbb{H}_{\mathcal{Y}} \subset \mathbb{H}_{\mathcal{X}}$ IN CASE OF THE GAUSSIAN KERNEL AND THE ORNSTEIN-UHLENBECK PROCESS

The Ornstein-Uhlenbeck (OU) process on $X = \mathbb{R}$ is the solution of the SDE $dX_t = -\alpha X_t dt + dW_t$, where $\alpha > 0$. The invariant measure μ and the Markov transition kernel ρ_t , $t > 0$, are known and given by

$$d\mu(x) = \sqrt{\frac{\alpha}{\pi}} \cdot e^{-\alpha x^2} dx \quad \text{and} \quad \rho_t(x, dy) = \sqrt{\frac{c_t}{\pi}} \cdot \exp\left[-c_t(y - e^{-\alpha t}x)^2\right] dy,$$

where

$$c_t = \frac{\alpha}{1 - e^{-2\alpha t}}.$$

We consider the Gaussian radial basis function (RBF) kernels on \mathbb{R} with bandwidth $\sigma > 0$, i.e.,

$$k^\sigma(x, y) = \exp\left[-\frac{(x - y)^2}{\sigma^2}\right].$$

For $y \in \mathbb{R}$ and a kernel k on \mathbb{R} set $k_y(x) := k(x, y)$, $x \in \mathbb{R}$.

Let $X \subset \mathbb{R}$ be any set. By $\mathbb{H}_\sigma(X)$ we denote the RKHS generated by the kernel $k_X^\sigma = k^\sigma|_{X \times X}$ on X . Hilbert space norm and scalar product on $\mathbb{H}_\sigma(X)$ will be denoted by $\|\cdot\|_{\sigma, X}$ and $\langle \cdot, \cdot \rangle_{\sigma, X}$, respectively. For two positive definite kernels on X we write $k_1 \preceq k_2$ if

$$\sum_{i, j=1}^n \alpha_i \alpha_j k_1(x_i, x_j) \leq \sum_{i, j=1}^n \alpha_i \alpha_j k_2(x_i, x_j)$$

for any choice of $n \in \mathbb{N}$ and $\alpha_j \in \mathbb{R}$, $x_j \in X$, $j = 1, \dots, n$. We also write $V \xrightarrow{c} W$ for two normed vector spaces V and W if $V \subset W$ is continuously embedded in W .

Lemma D.1. *Let $X \subset \mathbb{R}$ and $0 < \sigma_1 < \sigma_2$. Then we have $\mathbb{H}_{\sigma_2}(X) \xrightarrow{c} \mathbb{H}_{\sigma_1}(X)$ with*

$$\|\psi\|_{\sigma_1, X} \leq \sqrt{\frac{\sigma_2}{\sigma_1}} \cdot \|\psi\|_{\sigma_2, X}, \quad \psi \in \mathbb{H}_{\sigma_2}(X),$$

and $k_X^{\sigma_2} \preceq \frac{\sigma_2}{\sigma_1} k_X^{\sigma_1}$.

Proof. The first claim is Corollary 6 in [50]. The second follows from Aronszajn's inclusion theorem [39, Theorem 5.1]. \square

Theorem D.2. *Let $\mathcal{Y} = \mathbb{R}$ and let $\mathcal{X} \subset \mathbb{R}$ be any compact set with non-empty interior. Then for each $t \geq 0$ and all $\alpha, \sigma > 0$ we have*

$$K^t \mathbb{H}_\sigma(\mathcal{Y}) \subset \mathbb{H}_\sigma(\mathcal{X}) \quad \text{with} \quad \|K^t\|_{\mathbb{H}_\sigma(\mathcal{Y}) \rightarrow \mathbb{H}_\sigma(\mathcal{X})} \leq e^{\frac{\alpha}{2}t}.$$

Proof. First of all, it follows from [50, Corollaries 4&5] that to each $\psi \in \mathbb{H}_\sigma(\mathcal{X})$ there exists a unique extension $\tilde{\psi} \in \mathbb{H}_\sigma(\mathcal{Y})$ such that $\langle \tilde{\psi}, \tilde{\phi} \rangle_{\sigma, \mathcal{Y}} = \langle \psi, \phi \rangle_{\sigma, \mathcal{X}}$ for all $\psi, \phi \in \mathbb{H}_\sigma(\mathcal{X})$. Conversely, we have

$$\mathbb{H}_\sigma(\mathcal{X}) = \{\psi|_{\mathcal{X}} : \psi \in \mathbb{H}_\sigma(\mathcal{Y})\} \quad \text{and} \quad \langle \psi, \phi \rangle_{\sigma, \mathcal{Y}} = \langle \psi|_{\mathcal{X}}, \phi|_{\mathcal{X}} \rangle_{\sigma, \mathcal{X}}, \quad \psi, \phi \in \mathbb{H}_\sigma(\mathcal{Y}). \quad (\text{D.1})$$

For $t = 0$ the claim is obviously true. Hence, suppose that $t > 0$. For $z \in \mathbb{R}$ let us compute $K^t k_z^\sigma$. For this, we define

$$\tau = \sqrt{\frac{c_t \sigma^2}{1 + c_t \sigma^2}}, \quad \text{and} \quad \nu = \frac{e^{\alpha t}}{\tau} \sigma > \frac{\sigma}{\tau} > \sigma.$$

Since for $\sigma_1, \sigma_2 > 0$ and $z, w \in \mathbb{R}$ we have

$$\int_{-\infty}^{\infty} k_z^{\sigma_1}(x) k_w^{\sigma_2}(x) dx = \sqrt{\pi} \cdot \frac{\sigma_1 \sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \cdot \exp\left(-\frac{(z - w)^2}{\sigma_1^2 + \sigma_2^2}\right),$$

with $\sigma_t = 1/\sqrt{c_t}$ we obtain for $x \in \mathcal{X}$

$$\begin{aligned} (K^t k_z^\sigma)(x) &= \int_{-\infty}^{\infty} k_z^\sigma(y) \rho_t(x, dy) = \sqrt{\frac{c_t}{\pi}} \int_{-\infty}^{\infty} k_z^\sigma(y) \exp\left[-c_t(y - e^{-\alpha t}x)^2\right] dy \\ &= \sqrt{\frac{c_t}{\pi}} \int_{-\infty}^{\infty} k_z^\sigma(y) k_{e^{-\alpha t}x}^{\sigma_t}(y) dy = \sqrt{c_t} \cdot \frac{\sigma \sigma_t}{\sqrt{\sigma^2 + \sigma_t^2}} \cdot \exp\left(-\frac{(z - e^{-\alpha t}x)^2}{\sigma^2 + \sigma_t^2}\right) \\ &= \frac{\sigma}{\sqrt{\sigma^2 + 1/c_t}} \cdot \exp\left(-\frac{(e^{\alpha t}z - x)^2}{e^{2\alpha t}(\sigma^2 + 1/c_t)}\right) = \tau \cdot k_{e^{\alpha t}z}^\nu(x). \end{aligned}$$

Since $\nu > \sigma$, it follows from (D.1) and Lemma D.1 that

$$K^t k_z^\sigma = \tau \cdot k_{e^{\alpha t}z}^\nu|_{\mathcal{X}} \in \mathbb{H}_\nu(\mathcal{X}) \subset \mathbb{H}_\sigma(\mathcal{X}).$$

Note that

$$k^\nu(e^{\alpha t}x, e^{\alpha t}y) = \exp\left(-\frac{e^{2\alpha t}(x-y)^2}{\nu^2}\right) = \exp\left(-\frac{\tau^2(x-y)^2}{\sigma^2}\right) = k^{\sigma/\tau}(x, y).$$

Now, let $n \in \mathbb{N}$ and $\alpha_j, x_j \in \mathbb{R}$, $j = 1, \dots, n$, be arbitrary and let $\psi = \sum_{j=1}^n \alpha_j k_{x_j}^\sigma \in \mathbb{H}_\sigma(\mathcal{Y})$. Then, again by (D.1) and Lemma D.1,

$$\begin{aligned} \|K^t \psi\|_{\sigma, \mathcal{X}}^2 &\leq \frac{\nu}{\sigma} \|K^t \psi\|_{\nu, \mathcal{X}}^2 = \frac{\nu}{\sigma} \left\| \sum_{j=1}^n \alpha_j K^t k_{x_j}^\sigma \right\|_{\nu, \mathcal{X}}^2 = \frac{\nu \tau^2}{\sigma} \left\| \sum_{j=1}^n \alpha_j k_{e^{\alpha t}x_j}^\nu \right\|_{\nu, \mathcal{Y}}^2 \\ &= \frac{\nu \tau^2}{\sigma} \sum_{i,j=1}^n \alpha_i \alpha_j k^\nu(e^{\alpha t}x_i, e^{\alpha t}x_j) = \frac{\nu \tau^2}{\sigma} \sum_{i,j=1}^n \alpha_i \alpha_j k^{\sigma/\tau}(x_i, x_j) \\ &\leq \frac{\nu \tau^2}{\sigma} \cdot \frac{1}{\tau} \sum_{i,j=1}^n \alpha_i \alpha_j k^\sigma(x_i, x_j) = \frac{\nu \tau}{\sigma} \|\psi\|_{\sigma, \mathcal{Y}}^2 = e^{\alpha t} \|\psi\|_{\sigma, \mathcal{Y}}^2. \end{aligned}$$

This shows that K^t maps $\mathbb{H}_{0,\sigma}(\mathcal{Y}) := \text{span}\{k_y^\sigma : y \in \mathcal{Y}\} \subset \mathbb{H}_\sigma(\mathcal{Y})$ boundedly into $\mathbb{H}_\sigma(\mathcal{X})$. Since $\mathbb{H}_{0,\sigma}(\mathcal{Y})$ is dense in $\mathbb{H}_\sigma(\mathcal{Y})$, it follows that $K^t|_{\mathbb{H}_{0,\sigma}}$ extends to a bounded operator $T : \mathbb{H}_\sigma(\mathcal{Y}) \rightarrow \mathbb{H}_\sigma(\mathcal{X})$. In order to see that $T\psi = K^t\psi$ for $\psi \in \mathbb{H}_\sigma(\mathcal{Y})$, let $(\psi_n) \subset \mathbb{H}_{0,\sigma}(\mathcal{Y})$ such that $\psi_n \rightarrow \psi$ in $\mathbb{H}_\sigma(\mathcal{Y})$. Then $K^t\psi_n = T\psi_n \rightarrow T\psi$ in $\mathbb{H}_\sigma(\mathcal{X})$. Since $\mathbb{H}_\sigma(\mathcal{Y}) \xrightarrow{c} L_\mu^2(\mathcal{Y})$, we have $\psi_n \rightarrow \psi$ in $L_\mu^2(\mathcal{Y})$ and thus $K^t\psi_n \rightarrow K^t\psi$ in $L_\mu^2(\mathcal{X})$. Also, $K^t\psi_n \rightarrow T\psi$ in $L_\mu^2(\mathcal{X})$. Hence, $K^t\psi = T\psi$ μ -a.e. on \mathbb{R} . But as both $K^t\psi$ and $T\psi$ are continuous and μ is absolutely continuous w.r.t. Lebesgue measure with a positive density, we conclude that $K^t\psi = T\psi \in \mathbb{H}_\sigma$. \square

AUTHOR AFFILIATIONS

F. Philipp TECHNISCHE UNIVERSITÄT ILMENAU, OPTIMIZATION-BASED CONTROL GROUP, INSTITUTE FOR MATHEMATICS, ILMENAU, GERMANY

Email address: friedrich.philipp@tu-ilmenau.de

M. Schaller TECHNISCHE UNIVERSITÄT ILMENAU, OPTIMIZATION-BASED CONTROL GROUP, INSTITUTE FOR MATHEMATICS, ILMENAU, GERMANY

Email address: manuel.schaller@tu-ilmenau.de

K. Worthmann OPTIMIZATION-BASED CONTROL GROUP, INSTITUTE OF MATHEMATICS, TECHNISCHE UNIVERSITÄT ILMENAU, ILMENAU, GERMANY

Email address: karl.worthmann@tu-ilmenau.de

S. Peitz PADERBORN UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE, DATA SCIENCE FOR ENGINEERING, GERMANY

Email address: sebastian.peitz@upb.de

F. Nüske MAX PLANCK INSTITUTE FOR DYNAMICS OF COMPLEX TECHNICAL SYSTEMS, MAGDEBURG, GERMANY, AND DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, FREIE UNIVERSITÄT BERLIN, GERMANY

Email address: nueske@mpi-magdeburg.mpg.de