ERROR BOUNDS FOR KERNEL-BASED APPROXIMATIONS OF THE KOOPMAN OPERATOR

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

ABSTRACT. We consider the data-driven approximation of the Koopman operator for stochastic differential equations on reproducing kernel Hilbert spaces (RKHS). Our focus is on the estimation error if the data are collected from long-term ergodic simulations. We derive both an exact expression for the variance of the kernel cross-covariance operator, measured in the Hilbert-Schmidt norm, and probabilistic bounds for the finite-data estimation error. Moreover, we derive a bound on the prediction error of observables in the RKHS using a finite Mercer series expansion. Further, assuming Koopman-invariance of the RKHS, we provide bounds on the full approximation error. Numerical experiments using the Ornstein-Uhlenbeck process illustrate our results.

1. INTRODUCTION

The Koopman operator [23] has become an essential tool in the modeling process of complex dynamical systems based on simulation or measurement data. The philosophy of the Koopman approach is that for a (usually non-linear) dynamical system on a finite-dimensional space, the time-evolution of expectation values of observable functions satisfies a linear differential equation. Hence, after "lifting" the dynamical system into an infinite-dimensional function space of observables, linear methods become available for its analysis. The second step is then to notice that traditional Galerkin approximations of the Koopman operator can be consistently estimated from simulation or measurement data, establishing the fundamental connection between the Koopman approach and modern data science. Koopman methods have found widespread application in system identification [4], control [24, 42, 25, 17, 49], sensor placement [31], molecular dynamics [50, 44, 35, 36, 18, 56], and many other fields. We refer to [19, 33, 5] for comprehensive reviews of the state of the art.

The fundamental numerical method for the Koopman approach is *Extended Dynamic Mode Decomposition* (EDMD) [54], which allows to learn a Galerkin approximation of the Koopman operator from finite (simulation or measurement) data on a subspace spanned by a finite set of observables, often called dictionary. An appropriate choice of said dictionary is a challenging problem. In light of this issue, representations of the Koopman operator on large approximation spaces have been considered in recent years, including deep neural networks [29, 32], tensor product spaces [21, 37], and *reproducing kernel Hilbert spaces* (RKHS) [55, 11, 20]. In the work [20] it was shown that by means of the integral operator associated to an RKHS, it is possible to construct a type of Galerkin approximation of the Koopman operator. The central object are (cross-)covariance operators, which can be estimated from data, using only evaluations of the feature map. Due to the relative simplicity of the resulting numerical algorithms on the one hand, and the rich approximation properties of reproducing kernels on the other hand, kernel methods have emerged as a promising candidate to overcome the fundamental problem of dictionary selection.

A key question is the quantification of the estimation error for (compressed[1]) Koopman operators. For finite dictionaries and independent, identically distributed (i.i.d.) samples, error estimates were provided in [26, 38], see also [58] for the ODE case and [49] for an extension to control-affine systems. The

[1] A compression of a linear operator T to a subspace M is given by $PT|_M$, where P denotes a projection onto M.

estimation error for cross-covariance operators on kernel spaces was considered in [34], where general concentration inequalities were employed. The data were also allowed to be correlated, and mixing coefficients were used to account for the lack of independence. In this article, we take a different route and follow the approach of our previous paper [38], where we, in addition, also derived error estimates for the Koopman generator and operator for finite dictionaries and data collected from long-term, ergodic trajectories. This setting is relevant in many areas of science, where sampling i.i.d. from an unknown stationary distribution is practically infeasible, e.g., in fluid or molecular dynamics. The centerpiece of our results was an exact expression for the variance of the finite-data estimator, which can be bounded by an asymptotic variance. The asymptotic variance by itself is a highly interesting dynamical quantity, which can also be described in terms of Poisson equations for the generator [27, Section 3].

We consider the Koopman semigroup $(K^t)_{t\geq 0}$ generated by a stochastic differential equation on the space L^2_μ, where μ is a probability measure which is invariant w.r.t. the associated Markov process. We study the action of K^t on observables in an RKHS $\mathbb{H}$ which is densely and compactly embedded in L^2_μ. If this action is considered through the “lens” of the kernel integral operator $\mathcal{E} : L^2_\mu \to \mathbb{H}$ (see Section 2.2), we arrive at a family of operators $C^t_{\mathbb{H}} = \mathcal{E}K^t\mathcal{E}^*$ (cf. Figure 1). The action of $C^t_{\mathbb{H}} : \mathbb{H} \to \mathbb{H}$ is that of a cross-covariance operator:

$$C^t_{\mathbb{H}}\psi = \int (K^t\psi)(x)k(x,\cdot)\,d\mu(x), \qquad \psi \in \mathbb{H},$$

where $k(\cdot,\cdot)$ is the kernel generating the RKHS $\mathbb{H}$. These operators possess canonical empirical estimators based on finite simulation data, which only require evaluations of the feature map.

$$\begin{tikzcd}
L^2_\mu \arrow[r, "K^t"] \arrow[d, "\mathcal{E}", leftarrow, swap] & L^2_\mu \arrow[d, "\mathcal{E}"] \\
\mathbb{H} \arrow[r, "C^t_{\mathbb{H}}"] & \mathbb{H}
\end{tikzcd}$$

FIGURE 1. Diagram illustrating the different operators involved

Our contribution, illustrated in Figure 2, is two-fold. In our first main result, Theorem 3.1, we provide an exact formula for the Hilbert-Schmidt variance of the canonical empirical estimator $\widehat{C}^{m,t}_{\mathbb{H}}$ of the cross-covariance operator $C^t_{\mathbb{H}}$, for m data points sampled from a long ergodic simulation. This result extends the findings in [38] to the kernel setting and no longer depends on the dictionary size (which would be infinite, at any rate). Due to the infinite-dimensional setting, additional assumptions are required, in particular, a spectral decomposition of the Koopman generator. Our result allows for probabilistic estimates for the error $\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|_{HS}$, see Proposition 3.4.

As a second main result, we propose an empirical estimator for the restriction of the Koopman operator K^t to $\mathbb{H}$, truncated to finitely many terms of its estimated Mercer series expansion, and prove a probabilistic bound for the resulting estimation error in Theorem 4.1, measured in the operator norm for bounded linear maps from $\mathbb{H}$ to L^2_μ. This result can be seen as a bound on the prediction error for the RKHS-based Koopman operator due to the use of finite data. In the situation where the RKHS is invariant under the Koopman operator we are able to complement the preceding error analysis with a bound on the full approximation error in Theorem 4.5.

Finally, we illustrate our results for a one-dimensional Ornstein-Uhlenbeck (OU) process. For this simple test case, all quantities appearing in our error estimates are known analytically and can be well approximated numerically. Therefore, we are able to provide a detailed comparison between the error bound obtained from our results and the actual errors observed for finite data. Our experiments show that our bounds for the estimation error of the cross-covariance operator are accurate, and that the corrections we introduced to account for the inter-dependence of the data are indeed required. Concerning the prediction error, we find our theoretical bounds still far too conservative, which reflects the problem of accounting for the effect of inverting the mass matrix in traditional EDMD. This finding indicates that additional research is required on this end.

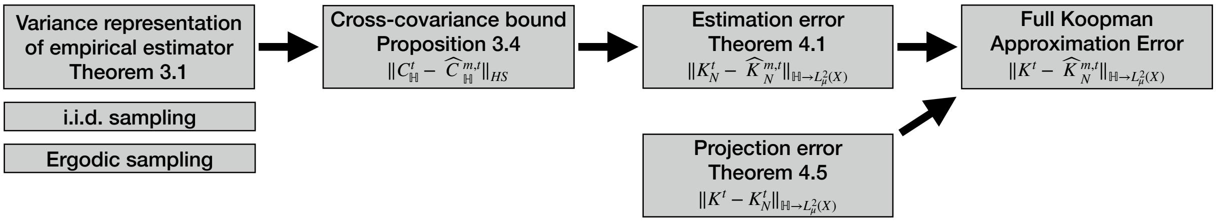
Variance representation of empirical estimator Theorem 3.1
i.i.d. sampling
Ergodic sampling
Cross-covariance bound Proposition 3.4
$\|C^t_{\mathbb{H}} - \widehat{C}^{m,t}_{\mathbb{H}}\|_{HS}$
Estimation error Theorem 4.1
$\|K^t_N - \widehat{K}^{m,t}_N\|_{\mathbb{H}\to L^2_\mu(X)}$
Full Koopman Approximation Error
$\|K^t - \widehat{K}^{m,t}_N\|_{\mathbb{H}\to L^2_\mu(X)}$
Projection error Theorem 4.5
$\|K^t - K^t_N\|_{\mathbb{H}\to L^2_\mu(X)}$

FIGURE 2. Illustration of main results

The paper is structured as follows: the setting is introduced in Section 2. The result concerning the variance of the empirical cross-covariance operator, Theorem 3.1, is presented and proved in Section 3, while our bound for the prediction error is part of Theorem 4.1 in Section 4. Numerical experiments are shown in Section 5, conclusions are drawn in Section 6.

2. PRELIMINARIES

In this section, we provide the required background on stochastic differential equations (Section 2.1), reproducing kernel Hilbert spaces (Section 2.2), Koopman operators (Section 2.3), and their representations on an RKHS (Section 2.4).

2.1. Stochastic differential equations.

Let $\mathcal{X} \subset \mathbb{R}^d$ and let a stochastic differential equation (SDE) with drift vector field $b : \mathcal{X} \to \mathbb{R}^d$ and diffusion matrix field $\sigma : \mathcal{X} \to \mathbb{R}^{d\times d}$ be given, i.e.,

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t, \tag{2.1}$$

where W_t is d-dimensional Brownian motion. We assume that both b and σ are Lipschitz-continuous and that $(1 + \|\cdot\|_2)^{-1}[\|b\|_2 + \|\sigma\|_F]$ is bounded on $\mathcal{X}$. Then [39, Theorem 5.2.1] guarantees the existence of a unique solution $(X_t)_{t\geq 0}$ to (2.1).

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

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

Throughout, we will assume the existence of an *invariant* (Borel) *probability measure* μ for the Markov process $(X_t)_{t\geq 0}$, i.e., we have

$$\int \rho_t(x, A)\,d\mu(x) = \mu(A) \tag{2.2}$$

for all $t \geq 0$.

In addition to being invariant, we will often assume that μ is *ergodic*, meaning that for any $t > 0$ every ρ_t-invariant set A (that is, $\rho_t(x, A) = 1$ for all $x \in A$) satisfies $\mu(A) \in \{0, 1\}$. In this case, the Birkhoff ergodic theorem [15, Theorem 9.6] (see also (D.1)) and its generalizations apply, and allow us to calculate expectations w.r.t. μ using long-time averages over simulation data.

We let $\|\cdot\|_p$ denote the $L^p_\mu(\mathcal{X})$-norm, $1 \leq p < \infty$. In the particular case $p = 2$, scalar product and norm on the Hilbert space $L^2_\mu(\mathcal{X})$ will be denoted by $\langle\cdot\,,\cdot\rangle_\mu$ and $\|\cdot\|_\mu$, respectively.

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

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

for all choices of $x_1, \dots, x_m \in \mathcal{X}$ and $c_1, \dots, c_m \in \mathbb{R}$. It is well known that k generates a so-called *reproducing kernel Hilbert space* (RKHS) [1, 6, 40] $(\mathbb{H}, \langle \cdot\,, \cdot \rangle)$ of continuous functions, such that for $\psi \in \mathbb{H}$ the *reproducing property*

$$\psi(x) = \langle \psi, \Phi(x) \rangle, \qquad x \in \mathcal{X}, \tag{2.3}$$

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

$$\Phi(x) = k(x, \cdot), \qquad x \in \mathcal{X}.$$

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

$$\varphi(x) = k(x, x), \qquad x \in \mathcal{X}.$$

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

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = \langle k(x, \cdot), k(x, \cdot) \rangle = k(x, x) = \varphi(x).$$

We shall frequently make use of the following estimate:

$$|k(x, y)| = |\langle \Phi(x), \Phi(y) \rangle| \leq \|\Phi(x)\|\|\Phi(y)\| = \sqrt{\varphi(x)\varphi(y)}.$$

In particular, it shows that k is bounded if and only if its diagonal φ is bounded.

By $\mathcal{L}^p_\mu(\mathcal{X})$, $p \in [1, \infty)$, we denote the space of all *functions* (not equivalence classes) on $\mathcal{X}$ with a finite p-norm $\|\cdot\|_p$. Henceforth, we shall impose the following

Compatibility Assumptions:

(A1) $\varphi \in \mathcal{L}^2_\mu(\mathcal{X})$.
(A2) If $\psi \in L^2_\mu(\mathcal{X})$ such that $\int\int k(x, y)\psi(x)\psi(y)\, d\mu(x)\, d\mu(y) = 0$, then $\psi = 0$.
(A3) If $\psi \in \mathbb{H}$ such that $\psi(x) = 0$ for μ-a.e. $x \in \mathcal{X}$, then $\psi(x) = 0$ for all $x \in \mathcal{X}$.

Many of the statements in this subsection can also be found in [52, Chapter 4]. However, as we aim to present the contents in a self-contained way, we provide the proofs in Appendix A.

The following lemma explains the meaning of the compatibility assumptions (A1) and (A2).

Lemma 2.1. *Under the assumption that $\varphi \in \mathcal{L}^1_\mu(\mathcal{X})$ (in particular, under assumption* (A1)*), we have that $\mathbb{H} \subset \mathcal{L}^2_\mu(\mathcal{X})$ with*

$$\|\psi\|_\mu \leq \sqrt{\|\varphi\|_1} \cdot \|\psi\|, \qquad \psi \in \mathbb{H}, \tag{2.4}$$

and assumption (A2) *is equivalent to the density of $\mathbb{H}$ in $\mathcal{L}^2_\mu(\mathcal{X})$.*

We have meticulously distinguished between functions and equivalence classes as there might be distinct functions $\phi, \psi \in \mathbb{H}$, which are equal μ-almost everywhere[2], i.e., $\phi = \psi$ in $L^2_\mu(\mathcal{X})$. The compatibility assumption (A3) prohibits this situation so that $\mathbb{H}$ can in fact be seen as a subspace of $L^2_\mu(\mathcal{X})$, which is then densely and continuously embedded.

[2] For example, if $\mu = \delta_a$ and $\phi(a) = \psi(a)$

Remark 2.2. (a) Condition (A1) implies $k \in L^4_{\mu\otimes\mu}(\mathcal{X}\times\mathcal{X})$, where $\mu\otimes\mu$ is the product measure on $\mathcal{X}\times\mathcal{X}$.

(b) The density of $\mathbb{H}$ in $L^2_\mu(\mathcal{X})$ is strongly related to the term *universality* in the literature, see [53].

(c) Condition (A3) holds if $\operatorname{supp}\mu = \mathcal{X}$, cf. [52, Exercise 4.6].

It immediately follows from

$$\int |\psi(x)|\|\Phi(x)\|\,d\mu(x) \le \|\psi\|_\mu \|\varphi\|_1^{1/2}, \tag{2.5}$$

for $\psi \in L^2_\mu(\mathcal{X})$ that the linear operator $\mathcal{E} : L^2_\mu(\mathcal{X}) \to \mathbb{H}$, defined by

$$\mathcal{E}\psi := \int \psi(x)\Phi(x)\,d\mu(x), \qquad \psi \in L^2_\mu(\mathcal{X}),$$

is well defined (as a Bochner integral in $\mathbb{H}$) and bounded with operator norm not larger than $\|\varphi\|_1^{1/2}$.

Remark 2.3. The so-called *kernel mean embedding* $\mathcal{E}_k$, mapping probability measures ν on $\mathcal{X}$ to the RKHS $\mathbb{H}$, is defined by $\mathcal{E}_k\nu = \int \Phi(x)\,d\nu(x)$, see, e.g., [51]. Hence, we have $\mathcal{E}\psi = \mathcal{E}_k\nu$ with $d\nu = \psi\,d\mu$.

Note that the operator $\mathcal{E}$ is not an embedding in strict mathematical terms. The terminology *embedding* rather applies to its adjoint $\mathcal{E}^*$. Indeed, the operator $\mathcal{E}$ enjoys the simple but important property:

$$\langle \mathcal{E}\psi, \eta\rangle = \int \psi(x)\langle \Phi(x), \eta\rangle\,d\mu(x) = \int \psi(x)\eta(x)\,d\mu(x) = \langle \psi, \eta\rangle_\mu \tag{2.6}$$

for $\psi \in L^2_\mu(\mathcal{X})$ and $\eta \in \mathbb{H}$. This implies that the adjoint operator $\mathcal{E}^* : \mathbb{H} \to L^2_\mu(\mathcal{X})$ is the inclusion operator from $\mathbb{H}$ into $L^2_\mu(\mathcal{X})$, i.e.,

$$\mathcal{E}^*\eta = \eta, \qquad \eta \in \mathbb{H}. \tag{2.7}$$

We shall further define the covariance operator[3]

$$C_{\mathbb{H}} := \mathcal{E}\mathcal{E}^* \in L(\mathbb{H}).$$

Recall that a linear operator $T \in L(\mathcal{H})$ on a Hilbert space $\mathcal{H}$ is *trace class* if for some (and hence for each) orthonormal basis $(e_j)_{j\in\mathbb{N}}$ of $\mathcal{H}$ we have that $\sum_{j=1}^\infty \langle (T^*T)^{1/2}e_i, e_i\rangle < \infty$. A linear operator $S \in L(\mathcal{H},\mathcal{K})$ between Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$ is said to be *Hilbert-Schmidt* [12, Chapter III.9] if S^*S is trace class, i.e., $\|S\|_{HS}^2 := \sum_{j=1}^\infty \|Se_i\|^2 < \infty$ for some (and hence for each) orthonormal basis $(e_j)_{j\in\mathbb{N}}$.

Lemma 2.4. *Let the Compatibility Assumptions* (A1)–(A3) *be satisfied. Then the following hold.*

(a) *The operator $\mathcal{E}$ is an injective Hilbert-Schmidt operator with*

$$\|\mathcal{E}\|_{HS}^2 = \|\varphi\|_1.$$

(b) *The space $\mathbb{H}$ is densely and compactly embedded in $L^2_\mu(\mathcal{X})$.*

(c) *The operator $C_{\mathbb{H}}$ is an injective non-negative selfadjoint trace class operator.*

The next theorem is due to Mercer and can be found in, e.g., [45]. It shows the existence of a particular orthonormal basis $(e_j)_{j=1}^\infty$ of $L^2_\mu(\mathcal{X})$ composed of eigenfunctions of $\mathcal{E}^*\mathcal{E}$, which we shall henceforth call the *Mercer basis* corresponding to the kernel k. Again for the sake of self-containedness, we give a short proof in Appendix A.

[3] In what follows, by $L(\mathcal{H},\mathcal{K})$ we denote the set of all bounded (i.e., continuous) linear operators between Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$. As usual, we also set $L(\mathcal{H}) := L(\mathcal{H},\mathcal{H})$.

Theorem 2.5 (Mercer's Theorem)**.** *There exists an orthonormal basis $(e_j)_{j=1}^{\infty}$ of $L^2_\mu(\mathcal{X})$ consisting of eigenfunctions of $\mathcal{E}^*\mathcal{E}$ with corresponding eigenvalues $\lambda_j > 0$ such that $\sum_{j=1}^{\infty} \lambda_j = \|\varphi\|_1 < \infty$. Furthermore, $(f_j)_{j=1}^{\infty}$ with $f_j = \sqrt{\lambda_j} e_j$ constitutes an orthonormal basis of $\mathbb{H}$ consisting of eigenfunctions of $C_{\mathbb{H}}$ with corresponding eigenvalues λ_j. Moreover, for all $x, y \in \mathcal{X}$,*

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

the series converges absolutely.

2.3. **The Koopman semigroup.** The *Koopman semigroup* $(K^t)_{t\geq 0}$ associated with the SDE (2.1) is defined by

$$(K^t\psi)(x) = \mathbb{E}[\psi(X_t)|X_0 = x] = \int \psi(y)\, \rho_t(x, dy),$$

for $\psi \in B(\mathcal{X})$, the set of all bounded Borel-measurable functions on $\mathcal{X}$, and $\rho_t(x, dy) = d\rho_t(x, \cdot)(y)$. It is easy to see that the invariance of μ is equivalent to the identity

$$\int K^t\psi\, d\mu = \int \psi\, d\mu \tag{2.8}$$

for all $t \geq 0$ and $\psi \in B(\mathcal{X})$ (which easily extends to functions $\psi \in L^1_\mu(\mathcal{X})$, see Proposition 2.7).

Remark 2.6. Note that in the case $\sigma = 0$ the SDE (2.1) reduces to the deterministic ODE $\dot{x} = b(x)$. Then (2.8) implies $\int |\psi(\phi(t,x))|^2\, d\mu(x) = \int |\psi(x)|^2\, d\mu(x)$ for all $t \geq 0$ and all $\psi \in B(\mathcal{X})$, where $\phi(\cdot, x)$ is the solution of the initial value problem $\dot{y} = b(y)$, $y(0) = x$. Hence, the composition operator $K^t : \psi \mapsto \psi \circ \phi(t, \cdot)$ is unitary in $L^2_\mu(\mathcal{X})$. However, we shall require below (see Theorem 3.1) that K^t has its spectrum in the interior of the unit circle. Therefore, we assume throughout that $\sigma \neq 0$.

The proofs of the following two propositions can be found in Appendix A.

Proposition 2.7. *For each $p \in [1, \infty]$ and $t \geq 0$, K^t extends uniquely to a bounded operator from $L^p_\mu(\mathcal{X})$ to itself with operator norm $\|K^t\|_{L^p_\mu \to L^p_\mu} \leq 1$.*

By $C_b(\mathcal{X})$ we denote the set of all bounded continuous functions on $\mathcal{X}$. As the measure μ is finite, we have $C_b(\mathcal{X}) \subset B(\mathcal{X}) \subset L^p_\mu(\mathcal{X})$ for all $p \in [1, \infty]$. In fact, $C_b(\mathcal{X})$ is dense in each $L^p_\mu(\mathcal{X})$, $p \in [1, \infty)$, see [48, Theorem 3.14].

Proposition 2.8. *$(K^t)_{t\geq 0}$ is a C_0-semigroup of contractions in $L^p_\mu(\mathcal{X})$ for each $p \in [1, \infty)$.*

The *infinitesimal generator* of the C_0-semigroup $(K^t)_{t\geq 0}$ is the (in general unbounded) operator in $L^2_\mu(\mathcal{X})$, defined by

$$\mathcal{L}\psi = L^2_\mu\text{-}\lim_{t\to 0} \frac{K^t\psi - \psi}{t}, \tag{2.9}$$

whose domain $\operatorname{dom}\mathcal{L}$ is the set of all $\psi \in L^2_\mu(\mathcal{X})$ for which the above limit exists. By Proposition 2.8 and the Lumer-Phillips theorem (see [28]), the operator $\mathcal{L}$ is densely defined, closed[4], dissipative (i.e., $\operatorname{Re}\langle \mathcal{L}\psi, \psi\rangle_\mu \leq 0$ for all $\psi \in \operatorname{dom}\mathcal{L}$), and its spectrum is contained in the closed left half-plane.

Lemma 2.9. *The constant function $\mathbb{1}$ is contained in $\operatorname{dom}\mathcal{L}$ and $\mathcal{L}\mathbb{1} = 0$. Moreover, if $M := \operatorname{span}\{\mathbb{1}\} \subset L^2_\mu(\mathcal{X})$, then both M and $M^\perp$ are invariant under $\mathcal{L}$ and all K^t, $t \geq 0$.*

[4] Recall that a linear operator T, defined on a subspace $\operatorname{dom} T$ of a Hilbert space $\mathcal{H}$, which maps to a Hilbert space $\mathcal{K}$, is closed if its graph is closed in $\mathcal{H} \times \mathcal{K}$.

Proof. It is easy to see that $K^t\mathbb{1} = \mathbb{1}$ for each $t \geq 0$ and hence $\mathbb{1} \in \operatorname{dom}\mathcal{L}$ with $\mathcal{L}\mathbb{1} = 0$. Hence $K^tM \subset M$ for all $t \geq 0$ and $\mathcal{L}M \subset M$. Now, if $\psi \perp \mathbb{1}$, then $\langle K^t\psi, \mathbb{1}\rangle_\mu = \int K^t\psi\, d\mu = \int \psi\, d\mu = \langle \psi, \mathbb{1}\rangle_\mu = 0$, which shows that also $K^tM^\perp \subset M^\perp$. The relation $\mathcal{L}M^\perp \subset M^\perp$ follows from (2.9). □

2.4. **Representation of Koopman Operators on the RKHS.** Using the integral operator $\mathcal{E}$, it is possible to represent the Koopman operator with the aid of a linear operator on $\mathbb{H}$, which is based on kernel evaluations. This construction mimics the well-known kernel trick used frequently in machine learning. To begin with, for any $x, y \in \mathcal{X}$ define the rank-one operator $C_{xy} : \mathbb{H} \to \mathbb{H}$ by

$$C_{xy}\psi := \langle \psi, \Phi(y)\rangle\Phi(x) = \psi(y)\Phi(x).$$

For $t \geq 0$ and $\psi \in \mathbb{H}$ we further define the cross-covariance operator $C_{\mathbb{H}}^t : \mathbb{H} \to \mathbb{H}$ by

$$C_{\mathbb{H}}^t\psi := \int\int C_{xy}\psi\, \rho_t(x, dy)\, d\mu(x) = \int (K^t\psi)(x)\Phi(x)\, d\mu(x) = \mathcal{E}K^t\psi = \mathcal{E}K^t\mathcal{E}^*\psi.$$

Thus, we have

$$C_{\mathbb{H}}^t = \mathcal{E}K^t\mathcal{E}^*. \tag{2.10}$$

In other words, the cross-covariance operator $C_{\mathbb{H}}^t$ represents the action of the Koopman semigroup through the lens of the RKHS integral operator $\mathcal{E}$ (see [20] for details). Being the product of the two Hilbert-Schmidt operators $\mathcal{E}K^t$ and $\mathcal{E}^*$, the operator $C_{\mathbb{H}}^t$ is trace class for all $t \geq 0$ (cf. [16, p. 521]).

Note that due to $\rho_0(x, \cdot\,) = \delta_x$, for $t = 0$ this reduces to the already introduced covariance operator

$$\int\int C_{xy}\, \rho_0(x, dy)\, d\mu(x) = \int C_{xx}\, d\mu(x) = \mathcal{E}\mathcal{E}^* = C_{\mathbb{H}}.$$

The identity (2.10) shows that for all $\eta, \psi \in \mathbb{H}$ we have

$$\langle \eta, C_{\mathbb{H}}^t\psi\rangle = \langle \eta, K^t\psi\rangle_\mu, \tag{2.11}$$

which shows that the role of $C_{\mathbb{H}}^t$ is analogous to that of the stiffness matrix in a traditional finite-dimensional approximation of the Koopman operator. In this analogy, the covariance operator $C_{\mathbb{H}}$ plays the role of the mass matrix.

2.5. **Empirical estimators.** Next, we introduce empirical estimators for $C_{\mathbb{H}}^t$ based on finite data (x_k, y_k), $k = 1, \ldots, m$. We consider two sampling scenarios for fixed $t > 0$:

(1) The x_k are drawn i.i.d. from μ, and each $y_k \sim \mu$ is obtained from the conditional distribution $\rho_t(x_k, \cdot)$, i.e., $y_k|(x_k = x) \sim \rho_t(x, \cdot)$ for μ-a.e. $x \in \mathcal{X}$. For example, y_k can be obtained by simulating the SDE (2.1) starting from x_k until time t.

(2) μ is ergodic and both x_k and y_k are obtained from a single (usually long-term) simulation of the dynamics X_t at discrete integration time step $\Delta t > 0$, using a sliding-window estimator, i.e.,

$$x_0 = X_0 \sim \mu, \quad x_k = X_{k\Delta t}, \quad \text{and} \quad y_k = X_{k\Delta t + t}.$$

Moreover, we assume that there exists a Riesz basis $(\psi_j)_{j=0}^\infty$ of $L^2_\mu(\mathcal{X})$ consisting of eigenfunctions of the generator $\mathcal{L}$ with corresponding eigenvalues μ_j satisfying $\sum_{j=0}^\infty e^{2(\operatorname{Re}\mu_j)\Delta t} < \infty$.

Remark 2.10. It easily follows from the discussion in Appendix B that the last assumption on the generator $\mathcal{L}$ and on the decay of its eigenvalues μ_j is equivalent to the similarity of $\mathcal{L}$ to an (unbounded) normal operator N such that $e^{N\Delta t} \in L(L^2_\mu(\mathcal{X}))$ is Hilbert-Schmidt. If the assumption holds with $\psi_j = Se_j$, where (e_j) is an orthonormal basis of $L^2_\mu(\mathcal{X})$, the operator N is given by $N = \sum_j \mu_j\langle\,\cdot\,, e_j\rangle e_j$ with $\operatorname{dom} N = \{\psi : (\mu_j\langle\psi, e_j\rangle) \in \ell^2\}$ and $\mathcal{L} = SNS^{-1}$. The condition $\sum_{j=0}^\infty e^{2(\operatorname{Re}\mu_j)\Delta t} < \infty$ then obviously means that the eigenvalues of $e^{N\Delta t}$ form an ℓ^2 sequence.

Recall that the joint distribution of two random variables X and Y is given by

$$dP_{X,Y}(x,y) = dP_{Y|X=x}(y) \cdot dP_X(x).$$

Set $X = x_k$ and $Y = y_k$. Then, in both cases **(1)** and **(2)**, we have $P_X = \mu$ and

$$P_{Y|X=x}(B) = P(y_k \in B | x_k = x) = P(X_t \in B | X_0 = x) = \rho_t(x,B).$$

In other words, for the joint distribution $\mu_{0,t}$ of x_k and y_k we have

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

More explicitly,

$$\mu_{0,t}(A \times B) = \int_A \rho_t(x,B)\, d\mu(x).$$

Now, since

$$C^t_{\mathbb{H}} = \int\int C_{xy}\, \rho_t(x,dy)\, d\mu(x) = \int C_{xy}\, d\mu_{0,t}(x,y) = \mathbb{E}\big[C_{x_k,y_k}\big],$$

for the empirical estimator for $C^t_{\mathbb{H}}$ we choose the expression

$$\widehat{C}^{m,t}_{\mathbb{H}} = \frac{1}{m}\sum_{k=0}^{m-1} C_{x_k,y_k}. \tag{2.12}$$

3. Variance of the Empirical Estimator

In case **(1)**, the law of large numbers [3, Theorem 2.4] and, in case **(2)**, ergodicity [2] ensures the expected behavior

$$\lim_{m\to\infty} \|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|_{HS} = 0 \qquad \text{a.s.}$$

However, this is a purely qualitative result, and nothing is known a priori on the rate of this convergence. The main result of this section, Theorem 3.1, contains an exact expression for the Hilbert-Schmidt variance of the empirical estimator $\widehat{C}^{m,t}_{\mathbb{H}}$ based on m data points, which then yields probabilistic estimates for the expression $\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|_{HS}$, see Proposition 3.4. Here, our focus is on the estimation from a single ergodic trajectory, i.e., case **(2)** above. While the broader line of reasoning partially resembles that of our previous paper [38], we require additional steps due to the infinite-dimensional setting introduced by the RKHS.

In Theorem 3.1 and its proof, we will be concerned with evolving kernels $k_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined by

$$k_t(x,x') := \int\int k(y,y')\, \rho_t(x,dy)\, \rho_t(x',dy').$$

We have

$$k_t(x,x') = \int\int \langle\Phi(y),\Phi(y')\rangle\, \rho_t(x,dy)\, \rho_t(x',dy') = \left\langle \int \Phi(y)\, \rho_t(x,dy), \int \Phi(y')\, \rho_t(x',dy') \right\rangle.$$

The integrals in the last expression are well defined as limits in $\mathbb{H}$ for μ-a.e. $x, x' \in \mathcal{X}$ as

$$\int\int \|\Phi(y)\|\, \rho_t(x,dy)\, d\mu(x) = \int\int \sqrt{\varphi(y)}\, \rho_t(x,dy)\, d\mu(x) = \int \sqrt{\varphi(x)}\, d\mu(x) \;\le\; \|\varphi\|_1^{1/2},$$

see (2.8). This shows that k_t is well defined ($(\mu \otimes \mu)$-a.e.) and that it is a positive definite kernel on its domain. Moreover, $k_0 = k$ and

$$|k_t(x,x')| \le \int \sqrt{\varphi(y')} \int \sqrt{\varphi(y)}\, \rho_t(x,dy)\, \rho_t(x',dy') = (K^t\sqrt{\varphi})(x) \cdot (K^t\sqrt{\varphi})(x').$$

In particular, $k_t \in L^2_{\mu\otimes\mu}(\mathcal{X}^2)$ with $\|k_t\|_{L^2_{\mu\otimes\mu}} \le \|\varphi\|_1$. By Φ_t we denote the corresponding feature map, i.e.,

$$\Phi_t(x) = k_t(x,\cdot).$$

Note that not necessarily $\Phi_t(x) \in \mathbb{H}$. Finally, we define

$$\Phi_{t,x} := \Phi(x)\Phi_t(x).$$

We are now in the position to formulate our first main result.

Theorem 3.1. *Setting $z_k = (x_k, y_k)$, $k = 1, \ldots, m$, the Hilbert-Schmidt variance of the empirical estimator can be written as*

$$\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \frac{1}{m}\left[\mathbb{E}_0(t) + 2\sum_{k=1}^{m-1} \tfrac{m-k}{m}\cdot \mathbb{E}\big[\langle C_{z_k} - C^t_{\mathbb{H}}, C_{z_0} - C^t_{\mathbb{H}}\rangle_{HS}\big]\right], \tag{3.1}$$

where

$$\mathbb{E}_0(t) := \mathbb{E}\big[\|C_{z_0} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \langle K^t\varphi, \varphi\rangle_\mu - \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}}.$$

In case **(1)**, $\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \frac{1}{m}\mathbb{E}_0(t)$, *whereas in case* **(2)** *we have*

$$\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \frac{1}{m}\left[\mathbb{E}_0(t) + 2\sum_{j=1}^{\infty}\frac{d_{j,t}q_j}{1-q_j}\left(1 - \frac{1}{m}\cdot\frac{1-q_j^m}{1-q_j}\right)\right], \tag{3.2}$$

with

$$q_j = e^{\mu_j \Delta t}, \quad d_{j,t} = \langle c_{j,t}, \psi_j\rangle_\mu, \quad \textit{and} \quad c_{j,t}(x) = \langle \Phi_{t,x}, \widetilde{\psi}_j\rangle_\mu.$$

Before proving Theorem 3.1 in Subsection 3.1 below, let us comment on its statements and draw some conclusions.

Remark 3.2. (a) Note that, by ergodicity of the invariant measure μ, the generator $\mathcal{L}$ has no eigenvalues on the imaginary axis, except the simple zero eigenvalue (see Proposition D.1 in the Appendix). In contrast, if we drop the ergodicity assumption, we have

$$\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \frac{1}{m}\left[\mathbb{E}_0(t) + 2\sum_{j=\nu_0}^{\infty}\frac{d_{j,t}q_j}{1-q_j}\left(1 - \frac{1}{m}\cdot\frac{1-q_j^m}{1-q_j}\right)\right] + \frac{m-1}{m}\sum_{j=1}^{\nu_0-1} d_{j,t},$$

where $\nu_0 = \#\{j : \mu_j \in \frac{2\pi i}{\Delta t}\mathbb{Z}\}$ is the number of eigenvalues of $\mathcal{L}$ of the form $\frac{2k\pi i}{\Delta t}$, $k \in \mathbb{Z}$, counting multiplicities. Obviously, the last term does not decay to zero as $m \to \infty$ if $\sum_{j=1}^{\nu_0-1} d_{j,t} \neq 0$.

(b) The definition of $c_{j,t}$ requires $\Phi_{t,x}$ to be in $L^2_\mu(\mathcal{X})$ for μ-a.e. $x \in \mathcal{X}$. This will in fact be proved in Lemma 3.6 below.

In the following, we let

$$\sigma_m^2 := \mathbb{E}_0(t) + 2\sum_{j=1}^{\infty}\frac{d_{j,t}q_j}{1-q_j}\left(1 - \frac{1}{m}\cdot\frac{1-q_j^m}{1-q_j}\right) \qquad \text{and} \qquad \sigma_\infty^2 := \mathbb{E}_0(t) + 2\sum_{j=1}^{\infty}\frac{d_{j,t}q_j}{1-q_j}.$$

Then

$$\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] = \frac{\sigma_m^2}{m}$$

and $\sigma_m^2 \to \sigma_\infty^2$ as $m \to \infty$. Both infinite series converge absolutely as $(q_j) \in \ell^2$ by assumption, and $(d_{j,t}) \in \ell^2$ as shown in the proof of Theorem 3.1. We can therefore interpret σ_∞^2 as *asymptotic variance* of the estimator $\widehat{C}^{m,t}_{\mathbb{H}}$, similar to our previous results in [38, Lemma 6].

An upper bound on the variance can be obtained as follows:

Corollary 3.3. *In case* **(2)***, for all $m \in \mathbb{N}$ we have*

$$\sigma_m^2 \le \langle K^t\varphi, \varphi\rangle_\mu \left[1 + \frac{4B}{A\delta_q}\|q\|_{\ell^2}\right], \tag{3.3}$$

where A and B denote the lower and upper Riesz bounds of (ψ_j), respectively,

$$q = (q_j)_{j=1}^\infty, \qquad \textit{and} \qquad \delta_q = \inf_{j\ge 1}|1 - q_j| > 0.$$

Proof. First of all, by Lemma 3.6,

$$\mathbb{E}_0(t) = \langle K^t\varphi, \varphi\rangle_\mu - \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}} \le \langle K^t\varphi, \varphi\rangle_\mu.$$

We have $|1 - q_j| \ge \delta_q$ and $|q_j| \le 1$ for all $j \ge 1$ and hence

$$\frac{1}{|1-q_j|}\cdot\left|1 - \frac{1}{m}\cdot\frac{1-q_j^m}{1-q_j}\right| \le \frac{1}{\delta_q}\left(1 + \frac{1}{m}\sum_{k=0}^{m-1}|q_j|^k\right) \le \frac{2}{\delta_q}.$$

This and (3.7) imply (3.3). □

Proposition 3.4. *We have the following probabilistic bound on the estimation error:*

$$\mathbb{P}\big(\|C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}\|_{HS} > \varepsilon\big) \le \begin{cases} \dfrac{\sigma_m^2}{m\varepsilon^2}, & \textit{in case } \textbf{(2)}, & (3.4)\\[2mm] \dfrac{E_0(t)}{m\varepsilon^2}, & \textit{in case } \textbf{(1)}, & (3.5)\\[2mm] 2\,e^{-\frac{m\varepsilon^2}{8\|k\|_\infty^2}}, & \textit{in case } \textbf{(1)} \textit{ with bounded kernel.} & (3.6)\end{cases}$$

In particular, the above also holds upon replacing the left-hand side by $\mathbb{P}\big(\|\mathcal{E}K^t\psi - \widehat{C}_{\mathbb{H}}^{m,t}\psi\| > \varepsilon\big)$ for $\psi \in \mathbb{H}$, $\|\psi\| = 1$.

Proof. The inequalities (3.4) and (3.5) are an immediate consequence of Markov's inequality, applied to the random variable $\|C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}\|_{HS}^2$. The inequality (3.6) follows from $C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t} = \frac{1}{m}\sum_{k=0}^{m-1}(C_{\mathbb{H}}^t - C_{z_k})$, Hoeffding's inequality for Hilbert space-valued random variables [43, Theorem 3.5] (see also [30, Theorem A.5.2]), and (cf. Lemma 3.6 below)

$$\|C_{\mathbb{H}}^t - C_{xy}\|_{HS} \le \|C_{\mathbb{H}}^t\|_{HS} + \|C_{xy}\|_{HS} = \sqrt{\langle k, k_t\rangle_{L^2_{\mu\otimes\mu}}} + \sqrt{\varphi(x)\varphi(y)} \le 2\|k\|_\infty,$$

since also $\|k_t\|_\infty \le \|k\|_\infty$. The estimate

$$\|\mathcal{E}K^t\psi - \widehat{C}_{\mathbb{H}}^{m,t}\psi\| = \|\mathcal{E}K^t\mathcal{E}^*\psi - \widehat{C}_{\mathbb{H}}^{m,t}\psi\| = \|(C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t})\psi\| \le \|C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}\|_{HS}$$

finally yields the last claim. □

Remark 3.5. Under additional assumptions (boundedness of the kernel, mixing, etc.), other concentration inequalities than Markov's, such as, e.g., [3, Theorem 2.12] (α-mixing) or [46, Théorème 3.1] (β-mixing), might lead to better estimates than (3.4).

3.1. **Proof of Theorem 3.1.**

Lemma 3.6. *Let $t \ge 0$. Then $\Phi_{t,x} \in L^2_\mu(\mathcal{X})$ for μ-a.e. $x \in \mathcal{X}$ with*

$$\|\Phi_{t,x}\|_\mu^2 \le \varphi(x)(K^t\varphi)(x)\cdot\langle K^t\varphi, \varphi\rangle_\mu.$$

Moreover, for every $t \ge 0$ we have

$$\|C_{xy}\|_{HS}^2 = \varphi(x)\varphi(y) \qquad \textit{and} \qquad \|C_{\mathbb{H}}^t\|_{HS}^2 = \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}} = \int\!\!\int \Phi_{t,x}(y)\,d\mu(y)\,d\mu(x).$$

Proof. We estimate

$$\begin{aligned}|\Phi_{t,x}(x')|^2 = |k(x,x')k_t(x,x')|^2 &\le \varphi(x)\varphi(x')(K^t\sqrt{\varphi})^2(x)\cdot(K^t\sqrt{\varphi})^2(x')\\ &\le \varphi(x)(K^t\varphi)(x)\cdot\varphi(x')(K^t\varphi)(x'),\end{aligned}$$

where we have applied Jensen's inequality to $(K^t\sqrt{\varphi})(x)$. This proves the first inequality. Next, if $(f_j)\subset\mathbb{H}$ denotes the Mercer basis corresponding to k, then

$$\langle C_{xy}, C_{x'y'}\rangle_{HS} = \sum_i \langle C_{xy}f_i, C_{x'y'}f_i\rangle = \sum_i f_i(y)f_i(y')k(x,x') = k(x,x')k(y,y')$$

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

$$\begin{aligned}\|C_{\mathbb{H}}^t\|_{HS}^2 &= \left\|\int C_{xy}\,d\mu_{0,t}(x,y)\right\|_{HS}^2 = \int\int k(x,x')k(y,y')\,d\mu_{0,t}(x,y)\,d\mu_{0,t}(x',y')\\ &= \int\int k(x,x')\left[\int\int k(y,y')\,\rho_t(x,dy)\,\rho_t(x',dy')\right]d\mu(x')\,d\mu(x) = \langle k,k_t\rangle_{L^2_{\mu\otimes\mu}},\end{aligned}$$

as claimed. □

Proof of Theorem 3.1. First of all, we have

$$\begin{aligned}\mathbb{E}\big[\|\widehat{C}_{\mathbb{H}}^{m,t} - C_{\mathbb{H}}^t\|_{HS}^2\big] &= \mathbb{E}\Big[\Big\|\frac{1}{m}\sum_{k=0}^{m-1}(C_{z_k} - C_{\mathbb{H}}^t)\Big\|_{HS}^2\Big] = \mathbb{E}\Big[\frac{1}{m^2}\sum_{k,\ell=0}^{m-1}\langle C_{z_k} - C_{\mathbb{H}}^t, C_{z_\ell} - C_{\mathbb{H}}^t\rangle_{HS}\Big]\\ &= \mathbb{E}\left[\frac{1}{m^2}\sum_{k=0}^{m-1}\|C_{z_k} - C_{\mathbb{H}}^t\|_{HS}^2 + \frac{2}{m^2}\sum_{k=0}^{m-1}\sum_{\ell=k+1}^{m-1}\langle C_{z_k} - C_{\mathbb{H}}^t, C_{z_\ell} - C_{\mathbb{H}}^t\rangle_{HS}\right]\\ &= \frac{1}{m}\mathbb{E}\big[\|C_{z_0} - C_{\mathbb{H}}^t\|_{HS}^2\big] + \frac{2}{m^2}\sum_{k=1}^{m-1}(m-k)\mathbb{E}\big[\langle C_{z_k} - C_{\mathbb{H}}^t, C_{z_0} - C_{\mathbb{H}}^t\rangle_{HS}\big].\end{aligned}$$

where we exploited that $\mathbb{E}[\langle C_{z_k} - C_{\mathbb{H}}^t, C_{z_\ell} - C_{\mathbb{H}}^t\rangle_{HS}]$ only depends on the difference $\ell - k$.

Let us compute the first term. Since $\mathbb{E}[C_{z_0}] = C_{\mathbb{H}}^t$ and thus $\mathbb{E}[\langle C_{z_0}, C_{\mathbb{H}}^t\rangle_{HS}] = \|C_{\mathbb{H}}^t\|_{HS}^2$,

$$\mathbb{E}\big[\|C_{z_0} - C_{\mathbb{H}}^t\|_{HS}^2\big] = \mathbb{E}\big[\|C_{z_0}\|_{HS}^2\big] - \|C_{\mathbb{H}}^t\|_{HS}^2.$$

For $\psi\in\mathbb{H}$ we have

$$\|C_{z_0}\psi\|^2 = \|\psi(y_0)\Phi(x_0)\|^2 = \psi(y_0)^2\varphi(x_0).$$

Using the Mercer basis $(f_i)\subset\mathbb{H}$ corresponding to k in $\mathbb{H}$ (cf. Theorem 2.5), we obtain

$$\mathbb{E}\big[\|C_{z_0}\|_{HS}^2\big] = \mathbb{E}\Big[\sum_i\|C_{z_0}f_i\|^2\Big] = \mathbb{E}\Big[\sum_i f_i(y_0)^2\varphi(x_0)\Big] = \mathbb{E}[\varphi(x_0)\varphi(y_0)].$$

Note that the latter equals ($\varphi(x) = k(x,x)$ by definition)

$$\mathbb{E}[\varphi(x_0)\varphi(y_0)] = \int\varphi(x)\int\varphi(y)\,\rho_t(x,dy)\,d\mu(x) = \int\varphi(x)(K^t\varphi)(x)\,d\mu(x) = \langle K^t\varphi,\varphi\rangle_\mu.$$

We obtain

$$\mathbb{E}\big[\|C_{z_0} - C_{\mathbb{H}}^t\|_{HS}^2\big] = \mathbb{E}[\varphi(x_0)\varphi(y_0)] - \langle k,k_t\rangle_{L^2_{\mu\otimes\mu}} = \langle K^t\varphi,\varphi\rangle_\mu - \langle k,k_t\rangle_{L^2_{\mu\otimes\mu}} = E_0(t)$$

and thus (3.1).

Case (1). In this case, z_k and z_ℓ are independent for $k\neq\ell$, so that

$$\mathbb{E}\big[\langle C_{z_k} - C_{\mathbb{H}}^t, C_{z_\ell} - C_{\mathbb{H}}^t\rangle_{HS}\big] = 0.$$

Hence, the statement of the theorem for case **(1)** follows.

Case (2). Here, the cross terms do not vanish. In fact,

$$\begin{aligned}\mathbb{E}\big[\langle C_{z_k} - C^t_{\mathbb{H}}, C_{z_0} - C^t_{\mathbb{H}}\rangle_{HS}\big] &= \mathbb{E}[\langle C_{z_k}, C_{z_0}\rangle_{HS}] - \|C^t_{\mathbb{H}}\|^2_{HS} = \mathbb{E}\Big[\sum_i \langle C_{z_k} f_i, C_{z_0} f_i\rangle\Big] - \|C^t_{\mathbb{H}}\|^2_{HS} \\ &= \mathbb{E}\Big[\Big(\sum_i f_i(y_k) f_i(y_0)\Big) k(x_k, x_0)\Big] - \|C^t_{\mathbb{H}}\|^2_{HS} \\ &= \mathbb{E}\big[k(y_k, y_0) k(x_k, x_0)\big] - \|C^t_{\mathbb{H}}\|^2_{HS}.\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}\big[k(y_k, y_0) k(x_k, x_0)\big] &= \int\int\int\int k(y', y) k(x', x)\, \rho_t(x', dy')\, \rho_{k\Delta t}(x, dx')\, \rho_t(x, dy)\, d\mu(x) \\ &= \int\int k(x, x') \Big[\int\int k(y, y')\, \rho_t(x, dy)\, \rho_t(x', dy')\Big]\, \rho_{k\Delta t}(x, dx')\, d\mu(x) \\ &= \int\int k(x, x') k_t(x, x')\, \rho_{k\Delta t}(x, dx')\, d\mu(x) \\ &= \int \Big[\int [\Phi(x)\Phi_t(x)](x')\, \rho_{k\Delta t}(x, dx')\Big]\, d\mu(x) \\ &= \int [K^{k\Delta t}\Phi_{t,x}](x)\, d\mu(x).\end{aligned}$$

Hence,

$$\mathbb{E}\big[\langle C_{z_k} - C^t_{\mathbb{H}}, C_{z_0} - C^t_{\mathbb{H}}\rangle_{HS}\big] = \int (K^{k\Delta t}\Phi_{t,x})(x)\, d\mu(x) - \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}}.$$

Let us now exploit the assumptions on the spectral properties of the generator $\mathcal{L}$ in case **(2)**. For μ-a.e. $x \in \mathcal{X}$, we have

$$\Phi_{t,x} = \sum_{j=0}^{\infty} c_{j,t}(x)\psi_j,$$

the series converging in $L^2_\mu(\mathcal{X})$. Therefore,

$$K^s\Phi_{t,x} = \sum_{j=0}^{\infty} c_{j,t}(x) K^s\psi_j = \sum_{j=0}^{\infty} c_{j,t}(x) e^{\mu_j s}\psi_j,$$

and thus (for $k \geq 1$)

$$\int \big(K^{k\Delta t}\Phi_{t,x}\big)(x)\, d\mu(x) = \int \sum_{j=0}^{\infty} c_{j,t}(x) e^{\mu_j k\Delta t}\psi_j(x)\, d\mu(x) = \sum_{j=0}^{\infty} d_{j,t}\cdot e^{\mu_j k\Delta t} = \sum_{j=0}^{\infty} d_{j,t}\cdot q_j^k.$$

This series converges absolutely for each $t \geq 0$ due to our assumption that $\sum_j |q_j|^2 < \infty$ and since for each $j \in \mathbb{N}_0$ we have by Lemma 3.6 that

$$\begin{aligned}\sum_{j=0}^{\infty} |d_{j,t}|^2 \leq B^2 \sum_{j=0}^{\infty} \|c_{j,t}\|^2_\mu &= B^2 \int \sum_{j=0}^{\infty} |\langle \Phi_{t,x}, \widetilde{\psi}_j\rangle_\mu|^2\, d\mu(x) \\ &\leq \frac{B^2}{A^2}\int \|\Phi_{t,x}\|^2_\mu\, d\mu(x) \leq \frac{B^2}{A^2}\langle K^t\varphi, \varphi\rangle^2_\mu,\end{aligned} \tag{3.7}$$

where A and B are the Riesz bounds of (ψ_j).

Without loss of generality, we may assume that $\mu_0 = 0$ with $\psi_0 = \mathbb{1}$ and $\mu_1, \ldots, \mu_{\nu_0 - 1} \in \frac{2\pi i}{\Delta t}\mathbb{Z}$, and $\psi_k \in \mathbb{1}^\perp$ for $k \geq 1$, see Lemma 2.9. The duality relations then imply $\widetilde{\psi}_0 = \mathbb{1}$. Now, $c_{0,t}(x) = \langle \Phi_{t,x}, \mathbb{1}\rangle_\mu = \int k(x,y)k_t(x,y)\,d\mu(y)$ and hence

$$d_{0,t} = \langle c_{0,t}, \mathbb{1}\rangle_\mu = \int c_{0,t}(x)\,d\mu(x) = \int\int k(x,y)k_t(x,y)\,d\mu(y)\,d\mu(x) = \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}}. \tag{3.8}$$

This implies

$$\mathbb{E}\big[\langle C_{z_k} - C^t_{\mathbb{H}}, C_{z_0} - C^t_{\mathbb{H}}\rangle_{HS}\big] = \sum_{j=0}^{\infty} d_{j,t}\cdot q_j^k - \langle k, k_t\rangle_{L^2_{\mu\otimes\mu}} = \sum_{j=1}^{\infty} d_{j,t}\cdot q_j^k$$

and therefore

$$\begin{aligned}\mathbb{E}\big[\|\widehat{C}^{m,t}_{\mathbb{H}} - C^t_{\mathbb{H}}\|^2_{HS}\big] &= \frac{1}{m}\mathbb{E}_0(t) + \frac{2}{m}\sum_{j=1}^{\infty} d_{j,t}\sum_{k=1}^{m-1}(1-\tfrac{k}{m})q_j^k\\ &= \frac{1}{m}\left[\mathbb{E}_0(t) + 2\sum_{j=1}^{\nu_0-1} d_{j,t}\sum_{k=1}^{m-1}(1-\tfrac{k}{m})q_j^k + 2\sum_{j=\nu_0}^{\infty} d_{j,t}\sum_{k=1}^{m-1}(1-\tfrac{k}{m})q_j^k\right].\end{aligned}$$

The identity

$$\sum_{k=1}^{m-1}(1-\tfrac{k}{m})q^k = \begin{cases}\frac{q}{1-q}\left(1 - \frac{1}{m}\cdot\frac{1-q^m}{1-q}\right) & \text{if } q \neq 1\\ \frac{m-1}{2} & \text{if } q = 1\end{cases}$$

finally yields (3.2). □

4. Bound on the Koopman prediction error

The kernel cross-covariance operator $C^t_{\mathbb{H}}$ can also be used to approximate the predictive capabilities of the Koopman operator, for observables in $\mathbb{H}$. Approximating the full Koopman operator involves the inverse of the co-variance operator, which becomes an unbounded operator on a dense domain of definition in the infinite-dimensional RKHS case. Moreover, its empirical estimator $\widehat{C}^m_{\mathbb{H}}$ is finite-rank and thus not even injective. While Fukumizu et al. tackle this problem in [10] by means of a regularization procedure, we choose to use pseudo-inverses instead (cf. Remark 4.2). We truncate the action of the Koopman operator using N terms of the Mercer series expansion and derive a bound for the prediction error for fixed truncation parameter N. While we use similar ideas as presented in [11], we heavily rely on our new results on the cross-covariance operator, cf. Section 3. Afterwards, we deal with the case of Koopman-invariance of the RKHS [22]. Here, we establish an estimate for the truncation error, which then yields a bound on the deviation from the full Koopman operator. We emphasize that this error bound is extremely useful in comparison to its prior counterparts based on the assumption that the space spanned by a finite number of so-called observables (dictionary) is invariant under the Koopman operator. The latter essentially requires to employ only Koopman eigenfunctions as observables, see, e.g., [25, 14].

Let (e_j) be the Mercer orthonormal basis of $L^2_\mu(\mathcal{X})$ corresponding to the kernel k and let $\lambda_j = \|\mathcal{E}e_j\|_\mu$ as well as $f_j := \sqrt{\lambda_j}e_j$ (cf. Theorem 2.5). We arrange the Mercer eigenvalues in a non-increasing way, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \ldots.$$

Let $\psi \in \mathbb{H}$. Then

$$K^t\psi = \sum_{j=1}^{\infty}\langle K^t\psi, e_j\rangle_\mu e_j = \sum_{j=1}^{\infty}\langle C^t_{\mathbb{H}}\psi, e_j\rangle e_j = \sum_{j=1}^{N}\langle C^t_{\mathbb{H}}\psi, e_j\rangle e_j + \sum_{j=N+1}^{\infty}\langle C^t_{\mathbb{H}}\psi, e_j\rangle e_j. \tag{4.1}$$

4.1. **Estimation error.** In the next theorem, we estimate the probabilistic error between the first summand

$$K_N^t\psi = \sum_{j=1}^N \langle C_{\mathbb{H}}^t\psi, e_j\rangle e_j, \qquad \psi \in \mathbb{H},$$

and its empirical estimator, which is of the form $\sum_{j=1}^N \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, \widehat{e}_j\rangle \widehat{e}_j$ with approximations $\widehat{e}_j$ of the e_j.

Theorem 4.1. *Assume that the eigenvalues λ_j of $C_{\mathbb{H}}$ are simple, i.e., $\lambda_{j+1} > \lambda_j$ for all j. Fix an arbitrary $N \in \mathbb{N}$ and let*

$$\delta_N = \min_{j=1,\ldots,N} \frac{\lambda_j - \lambda_{j+1}}{2}. \tag{4.2}$$

Further, let $\varepsilon \in (0, \delta_N)$ and $\delta \in (0,1)$ be arbitrary and fix some[5] *$m \ge \max\{N, \frac{2\sigma_m^2}{\varepsilon^2\delta}\}$. Let now $\widehat{\lambda}_1 \ge \ldots \ge \widehat{\lambda}_m$ denote the largest m eigenvalues of $\widehat{C}_{\mathbb{H}}^m$ in descending order and let $\widehat{e}_1, \ldots, \widehat{e}_m$ be corresponding eigenfunctions, respectively, such that $\|\widehat{e}_j\| = \widehat{\lambda}_j^{-1/2}$ for $j = 1, \ldots, m$. If we define*

$$\widehat{K}_N^{m,t}\psi = \sum_{j=1}^N \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, \widehat{e}_j\rangle \widehat{e}_j, \qquad \psi \in \mathbb{H}, \tag{4.3}$$

then, with probability at least $1-\delta$, we have that

$$\|K_N^t - \widehat{K}_N^{m,t}\|_{\mathbb{H}\to L_\mu^2(\mathcal{X})} \le \left[\frac{1}{\sqrt{\lambda_N}} + \frac{N+1}{\delta_N\lambda_N}(1+\|\varphi\|_1)\|\varphi\|_1^{1/2}\right]\varepsilon. \tag{4.4}$$

All of the above statements equally apply to case **(1)** *upon replacing σ_m by $E_0(t)$.*

Remark 4.2. (a) If we set $\widehat{f}_j = \widehat{\lambda}_j^{1/2} \cdot \widehat{e}_j$, then

$$\widehat{C}_{\mathbb{H}}^m = \sum_{j=1}^m \widehat{\lambda}_j \langle\,\cdot\,, \widehat{f}_j\rangle \widehat{f}_j,$$

and thus

$$\sum_{j=1}^N \langle\,\cdot\,, \widehat{e}_j\rangle \widehat{e}_j = \sum_{j=1}^N \frac{1}{\widehat{\lambda}_j}\langle\,\cdot\,, \widehat{f}_j\rangle \widehat{f}_j = (\widehat{C}_{\mathbb{H}}^m)^\dagger \widehat{Q}_N,$$

where $\widehat{Q}_N = \sum_{j=1}^N \langle\,\cdot\,, \widehat{f}_j\rangle \widehat{f}_j$ is the orthogonal projector onto the span of the first N eigenfunctions of $\widehat{C}_{\mathbb{H}}^m$ in $\mathbb{H}$. Therefore,

$$\widehat{K}_N^{m,t}\psi = \sum_{j=1}^m \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, \widehat{e}_j\rangle \widehat{e}_j = (\widehat{C}_{\mathbb{H}}^m)^\dagger \widehat{Q}_N \widehat{C}_{\mathbb{H}}^{m,t}\psi.$$

In particular, for $N = m$ we have $\widehat{K}_N^{m,t} = (\widehat{C}_{\mathbb{H}}^m)^\dagger \widehat{C}_{\mathbb{H}}^{m,t}$, which surely is one of the first canonical choices for an empirical estimator of K^t.

(b) The functions $\widehat{e}_j$ have unit length in the empirical L_μ^2-norm:

$$\frac{1}{m}\sum_{k=1}^m \widehat{e}_j(x_k)\widehat{e}_j(x_k) = \left\langle \widehat{C}_{\mathbb{H}}^m \widehat{e}_j,\, \widehat{e}_j\right\rangle = 1.$$

Therefore, projecting onto the first N empirical Mercer features is the *whitening transformation* commonly used in traditional EDMD [19].

[5] By Corollary 3.3, an amount of at least $m \ge \max\left\{N\,,\ \frac{2\|\varphi\|_\mu^2}{\varepsilon^2\delta}\left[1 + \frac{4B}{A\delta_q}\|q\|_{\ell^2}\right]\right\}$ data points suffices.

Proof of Theorem 4.1. By Proposition 3.4, both events $\|C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}\|_{HS} \le \varepsilon$ and $\|C_{\mathbb{H}} - \widehat{C}_{\mathbb{H}}^m\|_{HS} \le \varepsilon$ occur with probability at least $1 - \delta/2$, respectively. Hence, they occur simultaneously with probability at least $1 - \delta$.

In the remainder of this proof we assume that both events occur. Then all the statements deduced in the following hold with probability at least $1 - \delta$.

Let us define the intermediate approximation

$$\widetilde{K}_N^{m,t}\psi = \sum_{j=1}^N \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, e_j\rangle e_j, \qquad \psi \in \mathbb{H}.$$

Let $\psi \in \mathbb{H}$ be arbitrary. Setting $C := C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}$, we have

$$\begin{aligned}\|K_N^t\psi - \widetilde{K}_N^{m,t}\psi\|_\mu^2 &= \left\|\sum_{j=1}^N \langle C\psi, e_j\rangle e_j\right\|_\mu^2 = \sum_{j=1}^N \left|\langle C\psi, e_j\rangle\right|^2 = \sum_{j=1}^N \left|\langle \psi, C^* e_j\rangle\right|^2\\ &\le \|\psi\|^2 \sum_{j=1}^N \|C^* e_j\|^2 \le \|\psi\|^2 \sum_{j=1}^N \frac{1}{\lambda_j}\|C^* f_j\|^2 \le \frac{\|\psi\|^2}{\lambda_N}\sum_{j=1}^N \|C^* f_j\|^2\\ &\le \frac{\|\psi\|^2}{\lambda_N}\sum_{j=1}^\infty \|C^* f_j\|^2 = \frac{\|\psi\|^2}{\lambda_N}\cdot \|C_{\mathbb{H}}^t - \widehat{C}_{\mathbb{H}}^{m,t}\|_{HS}^2,\end{aligned}$$

and thus,

$$\|K_N^t\psi - \widetilde{K}_N^{m,t}\psi\|_\mu \le \frac{\|\psi\|}{\sqrt{\lambda_N}}\cdot\varepsilon.$$

Next, we aim at estimating the remaining error

$$\begin{aligned}\widetilde{K}_N^{m,t}\psi - \widehat{K}_N^{m,t}\psi &= \sum_{j=1}^N \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, e_j\rangle e_j - \sum_{j=1}^N \langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, \widehat{e}_j\rangle \widehat{e}_j\\ &= \sum_{j=1}^N \lambda_j^{-1}\langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, f_j\rangle f_j - \sum_{j=1}^N \widehat{\lambda}_j^{-1}\langle \widehat{C}_{\mathbb{H}}^{m,t}\psi, \widehat{f}_j\rangle \widehat{f}_j\\ &= \sum_{j=1}^N \lambda_j^{-1}\langle f, f_j\rangle f_j - \sum_{j=1}^N \widehat{\lambda}_j^{-1}\langle f, \widehat{f}_j\rangle \widehat{f}_j\\ &= \sum_{j=1}^N \left[\lambda_j^{-1}P_j f - \widehat{\lambda}_j^{-1}\widehat{P}_j f\right]\\ &= \sum_{j=1}^N \lambda_j^{-1}(P_j - \widehat{P}_j)f + \sum_{j=1}^N (\lambda_j^{-1} - \widehat{\lambda}_j^{-1})\widehat{P}_j f,\end{aligned}$$

where $f = \widehat{C}_{\mathbb{H}}^{m,t}\psi$,

$$P_j f = \langle f, f_j\rangle f_j \qquad \text{and} \qquad \widehat{P}_j f = \langle f, \widehat{f}_j\rangle \widehat{f}_j.$$

By (2.4), it suffices to estimate the above error in the $\|\cdot\|$-norm. By Theorem C.3, the first summand can be estimated as

$$\left\|\sum_{j=1}^N \lambda_j^{-1}(P_j - \widehat{P}_j)f\right\| \le \sum_{j=1}^N \frac{1}{\lambda_j}\|P_j - \widehat{P}_j\|\|f\| \le \frac{N\cdot\|C_{\mathbb{H}} - \widehat{C}_{\mathbb{H}}^m\|}{\lambda_N\delta_N}\|f\| \le \frac{N}{\lambda_N\delta_N}\|f\|\varepsilon.$$

For the second summand we have

$$\Big\|\sum_{j=1}^{N}(\lambda_j^{-1}-\widehat{\lambda}_j^{-1})\widehat{P}_j f\Big\|^2 = \sum_{j=1}^{N}|\lambda_j^{-1}-\widehat{\lambda}_j^{-1}|^2\|\widehat{P}_j f\|^2 = \sum_{j=1}^{N}\frac{|\lambda_j-\widehat{\lambda}_j|^2}{\lambda_j^2\widehat{\lambda}_j^2}\|\widehat{P}_j f\|^2.$$

Now, note that $\epsilon < \delta_N$ by assumption and therefore $\|C_{\mathbb{H}} - \widehat{C}_{\mathbb{H}}^m\|_{HS} \le \delta_N \le \frac{\lambda_N-\lambda_{N+1}}{2} \le \frac{\lambda_N}{2}$. For $j = 1,\dots,N$, according to Theorem C.1 this implies

$$\widehat{\lambda}_j \ge \lambda_j - |\lambda_j - \widehat{\lambda}_j| \ge \lambda_j - \|C_{\mathbb{H}} - \widehat{C}_{\mathbb{H}}^m\|_{HS} \ge \lambda_j - \tfrac{\lambda_N}{2} \ge \tfrac{\lambda_j}{2}.$$

Hence,

$$\Big\|\sum_{j=1}^{N}(\lambda_j^{-1}-\widehat{\lambda}_j^{-1})\widehat{P}_j f\Big\|^2 \le 4\sum_{j=1}^{N}\frac{|\lambda_j-\widehat{\lambda}_j|^2}{\lambda_j^4}\|\widehat{P}_j f\|^2 \le 4\frac{\|C_{\mathbb{H}}-\widehat{C}_{\mathbb{H}}^m\|_{HS}^2}{\lambda_N^4}\|\widehat{Q}_N f\|^2,$$

and thus,

$$\Big\|\sum_{j=1}^{N}(\lambda_j^{-1}-\widehat{\lambda}_j^{-1})\widehat{P}_j f\Big\| \le \frac{2}{\lambda_N^2}\|f\|\varepsilon \le \frac{1}{\lambda_N\delta_N}\|f\|\varepsilon.$$

From

$$\|\widehat{C}_{\mathbb{H}}^{m,t}\| \le \|\widehat{C}_{\mathbb{H}}^{m,t} - C_{\mathbb{H}}^t\| + \|C_{\mathbb{H}}^t\| \le \|\widehat{C}_{\mathbb{H}}^{m,t} - C_{\mathbb{H}}^t\|_{HS} + \|\mathcal{E}K^t\mathcal{E}^*\| \le \varepsilon + \|\varphi\|_1$$

we conclude

$$\big\|\widetilde{K}_N^{m,t}\psi - \widehat{K}_N^{m,t}\psi\big\| \le \frac{N+1}{\lambda_N\delta_N}\|\widehat{C}_{\mathbb{H}}^{m,t}\psi\|\varepsilon \le \frac{N+1}{\lambda_N\delta_N}(\varepsilon+\|\varphi\|_1)\|\psi\|\varepsilon.$$

All together, we obtain (recall (2.4))

$$\begin{aligned}\|K_N^t\psi - \widehat{K}_N^{m,t}\psi\|_\mu &\le \|K_N^t\psi - \widetilde{K}_N^{m,t}\psi\|_\mu + \|\varphi\|_1^{1/2}\|\widetilde{K}_N^{m,t}\psi - \widehat{K}_N^{m,t}\psi\|\\ &\le \frac{\|\psi\|}{\sqrt{\lambda_N}}\cdot\varepsilon + \frac{N+1}{\lambda_N\delta_N}(\varepsilon+\|\varphi\|_1)\|\varphi\|_1^{1/2}\|\psi\|\varepsilon\\ &= \left[\frac{1}{\sqrt{\lambda_N}} + \frac{N+1}{\delta_N\lambda_N}(1+\|\varphi\|_1)\|\varphi\|_1^{1/2}\right]\varepsilon\cdot\|\psi\|,\end{aligned}$$

which implies (4.4). □

4.2. **Projection error in case of Koopman-invariance of the RKHS.** In the preceeding section, we have seen that the empirical operator $\widehat{K}_N^{m,t}$ can be written as $(\widehat{C}_{\mathbb{H}}^m)^\dagger\widehat{C}_{\mathbb{H}}^{m,t}$ if $m = N$. In the limit $m \to \infty$, we would arrive at the operator $C_{\mathbb{H}}^{-1}C_{\mathbb{H}}^t$, which is not even well-defined for all $\psi \in \mathbb{H}$, in general. However, if the RKHS is invariant under K^t, the above operator limit is well-defined as a bounded operator on $\mathbb{H}$. In this situation we are able to extend Theorem 4.1 to an estimate on the full error made by our empirical estimator.

We start by defining the operator

$$K_{\mathbb{H}}^t := C_{\mathbb{H}}^{-1}C_{\mathbb{H}}^t$$

on its natural domain

$$\operatorname{dom} K_{\mathbb{H}}^t := \{\psi \in \mathbb{H} : C_{\mathbb{H}}^t\psi \in \operatorname{ran} C_{\mathbb{H}}\}. \tag{4.5}$$

We consider $K_{\mathbb{H}}^t$ as an operator from $\mathbb{H}$ into itself (with domain of definition in $\mathbb{H}$).

Lemma 4.3. *We have*

$$\operatorname{dom} K_{\mathbb{H}}^t = \{\psi \in \mathbb{H} : K^t\psi \in \mathbb{H}\}, \tag{4.6}$$

and $K_{\mathbb{H}}^t$ is closed.

Proof. Note that $C_{\mathbb{H}}^t\psi \in \operatorname{ran} C_{\mathbb{H}}$ if and only if $\mathcal{E}K^t\psi = C_{\mathbb{H}}\phi$ for some $\phi \in \mathbb{H}$. Since $\mathbb{C}_{\mathbb{H}}\phi = \mathcal{E}\phi$ and $\ker \mathcal{E} = \{0\}$, the latter is equivalent to $K^t\psi = \phi \in \mathbb{H}$, which proves the representation of the domain. As to the closedness of $K_{\mathbb{H}}^t$, let $(\psi_n) \subset \operatorname{dom} K_{\mathbb{H}}^t$ and $\phi \in \mathbb{H}$ such that $\psi_n \to \psi$ in $\mathbb{H}$ and $K_{\mathbb{H}}^t\psi_n \to \phi$ in $\mathbb{H}$ as $n \to \infty$. The latter implies $C_{\mathbb{H}}^t\psi_n \to C_{\mathbb{H}}\phi$, while the first implies $C_{\mathbb{H}}^t\psi_n \to C_{\mathbb{H}}^t\psi$ in $\mathbb{H}$ as $n \to \infty$, from which we conclude that $C_{\mathbb{H}}^t\psi = C_{\mathbb{H}}\phi$, i.e., $\psi \in \operatorname{dom} K_{\mathbb{H}}^t$ and $K_{\mathbb{H}}^t\psi = \phi$. □

If the Koopman operator leaves the RKHS $\mathbb{H}$ invariant (i.e., $K^t\mathbb{H} \subset \mathbb{H}$), $K_{\mathbb{H}}^t$ is defined on all of $\mathbb{H}$. Moreover, since the canonical inclusion map $\mathcal{E}^* : \mathbb{H} \to L^2(\mu)$ is injective, it possesses an unbounded inverse on its range $\mathbb{H}$, and therefore:

$$C_{\mathbb{H}}^{-1}C_{\mathbb{H}}^t\phi = C_{\mathbb{H}}^{-1}\mathcal{E}K^t\mathcal{E}^*\phi = (\mathcal{E}\mathcal{E}^*)^{-1}\mathcal{E}\mathcal{E}^*(\mathcal{E}^*)^{-1}K^t\mathcal{E}^*\phi = (\mathcal{E}^*)^{-1}K^t\mathcal{E}^*\phi. \tag{4.7}$$

Remarkably, invariance of $\mathbb{H}$ under the Koopman operator implies that the left-hand side not only reproduces the Koopman operator on $\mathbb{H}$, but actually defines a bounded operation.

Parts of the next proposition can be found in [22, Theorem 5.3] and [8, Theorem 1].

Proposition 4.4. *For $t > 0$, the following statements are equivalent:*

(i) $K^t\mathbb{H} \subset \mathbb{H}$.
(ii) $K_{\mathbb{H}}^t \in L(\mathbb{H})$.
(iii) $\operatorname{ran} C_{\mathbb{H}}^t \subset \operatorname{ran} C_{\mathbb{H}}$.

Proof. With regard to the two representations (4.5) and (4.6) of the domain, it is immediate that both (i) and (iii) are equivalent to $\operatorname{dom} K_{\mathbb{H}}^t = \mathbb{H}$. The equivalence of the latter to (ii) follows from the closed graph theorem. □

Note that if one of (i)–(iii) holds, then $K_{\mathbb{H}}^t = K^t|_{\mathbb{H}}$.

Theorem 4.5. *In addition to the assumptions in Theorem 4.1, assume that $\mathbb{H}$ is invariant under the Koopman operator K^t. For fixed $N \in \mathbb{N}$, let δ_N be as in (4.2), choose ε, δ, and m as in Theorem 4.1 and define the empirical estimator $\widehat{K}_N^{m,t}$ as in (4.3). Then, with probability at least $1-\delta$ we have that*

$$\|K^t - \widehat{K}_N^{m,t}\|_{\mathbb{H}\to L^2_\mu(\mathcal{X})} \le \sqrt{\lambda_{N+1}}\,\|K_{\mathbb{H}}^t\| + \left[\frac{1}{\sqrt{\lambda_N}} + \frac{N+1}{\delta_N\lambda_N}(1+\|\varphi\|_1)\|\varphi\|_1^{1/2}\right]\varepsilon. \tag{4.8}$$

Proof. First of all, Theorem 4.1 implies that

$$\begin{aligned}\|K^t - \widehat{K}_N^{m,t}\|_{\mathbb{H}\to L^2_\mu(\mathcal{X})} &\le \|K^t - K_N^t\|_{\mathbb{H}\to L^2_\mu(\mathcal{X})} + \|K_N^t - \widehat{K}_N^{m,t}\|_{\mathbb{H}\to L^2_\mu(\mathcal{X})}\\ &\le \|K^t - K_N^t\|_{\mathbb{H}\to L^2_\mu(\mathcal{X})} + \left[\frac{1}{\sqrt{\lambda_N}} + \frac{N+1}{\delta_N\lambda_N}(1+\|\varphi\|_1)\|\varphi\|_1^{1/2}\right]\varepsilon.\end{aligned}$$

Now, for $\psi \in \mathbb{H}$,

$$\begin{aligned}\|K^t\psi - K_N^t\psi\|_\mu^2 &= \left\|\sum_{j=N+1}^\infty \langle C_{\mathbb{H}}^t\psi, e_j\rangle e_j\right\|_\mu^2 = \sum_{j=N+1}^\infty |\langle C_{\mathbb{H}}^t\psi, e_j\rangle|^2 = \sum_{j=N+1}^\infty \frac{1}{\lambda_j}|\langle C_{\mathbb{H}}^t\psi, f_j\rangle|^2\\ &= \sum_{j=N+1}^\infty \frac{1}{\lambda_j}|\langle K_{\mathbb{H}}^t\psi, C_{\mathbb{H}}f_j\rangle|^2 = \sum_{j=N+1}^\infty \lambda_j|\langle K_{\mathbb{H}}^t\psi, f_j\rangle|^2 \le \lambda_{N+1}\|K_{\mathbb{H}}^t\psi\|^2,\end{aligned}$$

which proves the theorem. □

We have just proved that the projection error $\|K^t\psi - K_N^t\psi\|_\mu$ decays at least as fast as the square roots of the eigenvalues of $C_{\mathbb{H}}$. Recall that $(\lambda_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$, since $C_{\mathbb{H}}$ is trace class with $\sum_{j=1}^\infty \lambda_j = \operatorname{Tr}(C_{\mathbb{H}}) = \|\mathcal{E}^*\|_{HS}^2 = \|\varphi\|_1$, see Lemma 2.4(c).

5. Illustration with the Ornstein-Uhlenbeck process

For the numerical illustration of our results, we consider the Ornstein-Uhlenbeck (OU) process on $\mathcal{X} = \mathbb{R}$, which is given by the SDE

$$dX_t = -\alpha X_t \, dt + dW_t,$$

where $\alpha > 0$ is a positive parameter.

5.1. Analytical Results.

Since all relevant properties of the OU process are available in analytical form, we can exactly calculate all of the terms appearing in our theoretical error bounds. Moreover, we can also compute the exact estimation and prediction errors for finite data in closed form. Let us begin by recapping the analytical results required for our analysis, which can be found in [41].

The invariant measure μ, and the density of the stochastic transition kernel ρ_t, are given by

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

with $v_t^2 = (1 - e^{-2\alpha t})/2\alpha$. The Koopman operators K^t are self-adjoint in $L^2_\mu(\mathbb{R})$, their eigenvalues and corresponding eigenfunctions are given by

$$q_j = e^{-\alpha j t} \qquad \text{and} \qquad \psi_j(x) = \frac{1}{\sqrt{2^j \alpha^j j!}} H_j(\sqrt{2\alpha} x), \quad j \in \mathbb{N}_0,$$

where H_j are the physicist's Hermite polynomials.

We consider the Gaussian radial basis function (RBF) kernel with bandwidth $\sigma > 0$, i.e.,

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

Let us quickly verify that this choice of the kernel satisfies the compatibility assumptions (A1)–(A3). Indeed, (A1) is trivial as $k(x,x) = 1$ and (A3) follows easily from the continuity of the functions in $\mathbb{H}$. To see that $\mathbb{H}$ is dense in $L^2_\mu(\mathbb{R})$ (i.e., (A2)), let $\psi \in L^2_\mu(\mathbb{R})$ be such that $\langle \psi, \Phi(y) \rangle_\mu = 0$ for all $y \in \mathbb{R}$. The latter means that $\phi * \varphi_\sigma = 0$, where $\phi(x) = \psi(x) e^{-\alpha x^2}$ and $\varphi_\sigma(x) = e^{-x^2/\sigma^2}$. We apply the Fourier transform and obtain $\widehat{\phi} \cdot \widehat{\varphi_\sigma} = 0$. Noting that the Fourier transform of a Gaussian is again a Gaussian, we get $\widehat{\phi} = 0$ and thus $\psi = 0$.

The Mercer eigenvalues and features with respect to the invariant measure μ of the OU process, i.e., the eigenvalues and eigenfunctions of the integral operator $\mathcal{E}^*\mathcal{E}$ in $L^2_\mu(\mathbb{R})$, are also available in analytical form [9]. They are given by

$$\lambda_i = \sqrt{\frac{\alpha}{C_1}} \left[\frac{1}{\sigma^2 C_1} \right]^i \qquad \text{and} \qquad \varphi_i(x) = \gamma_i e^{-\zeta^2 x^2} H_i\left(\sqrt{\alpha} \eta x \right), \quad i \in \mathbb{N}_0,$$

using the following constants:

$$\eta = \left[1 + \frac{4}{\alpha \sigma^2} \right]^{1/4}, \qquad \gamma_i = \left[\frac{\eta}{2^i \Gamma(i+1)} \right]^{1/2}, \qquad \zeta^2 = \frac{\alpha}{2}(\eta^2 - 1), \qquad C_1 = \alpha + \zeta^2 + \sigma^{-2}.$$

With these results, we can compute the variance of the empirical estimator for $C^t_{\mathbb{H}}$ as described in Theorem 3.1. The eigenvalues q_j were already given above. The coefficients $d_{j,t}$ can be calculated using Mercer's theorem as

$$d_{j,t} = \int \int k(x,x') k(y,y') \psi_j(x) \psi_j(x') \, d\mu_{0,t}(x,y) \, d\mu_{0,t}(x',y')$$

$$= \sum_{k,\ell} \lambda_k \lambda_\ell \left[\int \varphi_k(x)\varphi_\ell(y)\psi_j(x)\, d\mu_{0,t}(x,y) \right]^2 .$$

The series needs to be truncated at a finite number of terms and the integrals can be calculated by numerical integration. As $d_{0,t} = \langle k, k_t \rangle_{L^2_{\mu\otimes\mu}} = \|C^t_{\mathbb{H}}\|^2_{HS}$ (cf. (3.8)), and hence

$$\|C^t_{\mathbb{H}}\|^2_{HS} = \sum_{k,\ell} \lambda_k \lambda_\ell \left[\int \varphi_k(x)\varphi_\ell(y)\, d\mu_{0,t}(x,y) \right]^2 , \tag{5.1}$$

the Hilbert-Schmidt norm of the cross-covariance operator $C^t_{\mathbb{H}}$ can be computed similarly. Since, for the Gaussian RBF kernel, we have $\varphi(x) = k(x,x) = 1$ for all x, we therefore find

$$\mathbb{E}_0(t) = \left\langle K^t\varphi,\, \varphi \right\rangle_\mu - \|C^t_{\mathbb{H}}\|^2_{HS} = 1 - \|C^t_{\mathbb{H}}\|^2_{HS},$$

completing the list of terms required by Theorem 3.1. In addition, we notice that upon replacing either one or two of the integrals in (5.1) by finite-data averages, we can also calculate $\|\hat{C}^{m,t}_{\mathbb{H}}\|^2_{HS}$ and $\langle C^t_{\mathbb{H}}, \hat{C}^{m,t}_{\mathbb{H}} \rangle_{HS}$. Therefore, the estimation error for finite data $\{(x_k, y_k)\}_{k=1}^m$ can be obtained by simply expanding the inner product

$$\|C^t_{\mathbb{H}} - \hat{C}^{m,t}_{\mathbb{H}}\|^2_{HS} = \|C^t_{\mathbb{H}}\|^2_{HS} + \|\hat{C}^{m,t}_{\mathbb{H}}\|^2_{HS} - 2\langle \hat{C}^{m,t}_{\mathbb{H}}, C^t_{\mathbb{H}} \rangle_{HS},$$

allowing us to precisely compare the estimation error to the error bounds obtained in Theorem 3.1.

Besides the estimation error for $C^t_{\mathbb{H}}$, we are also interested in the prediction error, which is bounded according to Theorem 4.1. We will compare these bounds to the actual error $\|(K^t_N - \hat{K}^{m,t}_N)\phi\|_{L^2_\mu(\mathcal{X})}$, for a specific observable $\phi \in \mathbb{H}$ and a fixed number of N Mercer features. For the OU process, it is again beneficial to consider Gaussian observables ϕ:

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(x-m_0)^2}{2\sigma_0^2} \right] .$$

Application of the Koopman operator leads to yet another, unnormalized Gaussian observable, which is given by

$$K^t\phi(x) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{(m_0 - e^{-\alpha t}x)^2}{2\sigma_t^2} \right] , \qquad \sigma_t^2 = \sigma_0^2 + v_t^2 .$$

The inner products of $K^t\phi$ with the Mercer eigenfunctions φ_i can be evaluated by numerical integration, providing full access to the truncated observable $K^t_N\phi$. On the other hand, the empirical approximation $\hat{K}^{m,t}_N\phi$ can be computed directly based on the data. We note that

$$\hat{K}^{m,t}_N\phi = \sum_{j=1}^N \left\langle \hat{C}^{m,t}_{\mathbb{H}}\phi,\, \hat{e}_j \right\rangle \hat{e}_j = \frac{1}{m}\sum_{k=1}^m \phi(y_k) \sum_{j=1}^N \langle \Phi(x_k),\, \hat{e}_j \rangle\, \hat{e}_j = \frac{1}{m}\sum_{k=1}^m \phi(y_k) \sum_{j=1}^N \hat{e}_j(x_k)\hat{e}_j .$$

The functions $\hat{e}_j$ can be obtained from the eigenvalue decomposition of the standard kernel Gramian matrix

$$\frac{1}{m}K_{\mathcal{X}} := \frac{1}{m}\left[k(x_k, x_l)\right]_{k,l=1}^m ,$$

as the latter is the matrix representation of the empirical covariance operator $\hat{C}^m_{\mathbb{H}}$ on the subspace $\operatorname{span}\{\Phi(x_k)\}_{k=1}^m$. If $\frac{1}{m}K_{\mathcal{X}} = V\Lambda V^\top$ is the spectral decomposition of the Gramian, then

$$\hat{e}_j = \frac{1}{m^{1/2}\hat{\lambda}_j} \sum_{l=1}^m V_{lj}\Phi(x_l)$$

are the correctly normalized eigenfunctions according to Theorem 4.1. Plugging this into the above, we find

$$\begin{aligned}\hat{K}_N^{m,t}\phi(x) &= \frac{1}{m}\sum_{k=1}^{m}\phi(y_k)\sum_{j=1}^{N}\frac{1}{m^{1/2}\hat{\lambda}_j}\sum_{l=1}^{m}V_{lj}k(x_l,x_k)\frac{1}{m^{1/2}\hat{\lambda}_j}\sum_{r=1}^{m}V_{rj}k(x_r,x)\\ &= \frac{1}{m}\phi(Y)^\top\frac{1}{m}K_{\mathcal{X}}\left[V_N\Lambda_N^{-2}V_N^\top\right]K_{\mathcal{X},x}\\ &= \frac{1}{m}\phi(Y)^\top V_N\Lambda_N^{-1}V_N^\top K_{\mathcal{X},x},\end{aligned}$$

where $\phi(Y)=[\phi(y_k)]_{k=1}^m$, $K_{\mathcal{X},x}=[k(x_k,x)]_{k=1}^m$, $V_N=V[I_N\ 0_{m-N}]^\top$, $\Lambda_N=\operatorname{diag}(\widehat{\lambda}_j)_{j=1}^N$.

5.2. **Numerical Results.** For the actual numerical experiments, we set $\alpha=1$, choose the elementary integration time step as $\Delta_t=10^{-2}$, and set the lag time to $t=0.05$. We compute the exact variance $\mathbb{E}[\|C_{\mathbb{H}}^t-\hat{C}_{\mathbb{H}}^{m,t}\|_{HS}^2]$ by the expression given in Theorem 3.1, and also the coarser estimate for the variance given in Corollary 3.3. We test three different kernel bandwidths, $\sigma\in\{0.05,0.1,0.5\}$. All Mercer series are truncated after the first 10 terms for $\sigma\in\{0.1,0.5\}$, and 20 terms for $\sigma=0.05$, while Koopman eigenfunction expansions are truncated after 15 terms.

In the first set of experiments, we use Chebyshev's inequality to compute the maximal estimation error $\|C_{\mathbb{H}}^t-\hat{C}_{\mathbb{H}}^{m,t}\|_{HS}$ that can be guaranteed with confidence $1-\delta=0.9$, for a range of data sizes m between $m=20$ and $m=50.000$. As a comparison, we generate 200 independent simulations of length $m+\frac{t}{\Delta_t}$, corresponding to the sliding-window estimator with m data points, for each data size. We then compute the resulting estimation error using the expressions given in the previous section. We extract the $1-\delta$-percentile of the estimation error for all trajectories, i.e., the maximal error that is not exceeded by $100*(1-\delta)$ percent of the trajectories. In addition, we also use Chebyshev's inequality with the i.i.d. variance $\frac{1}{m}\mathbb{E}_0(t)$ to predict the estimation error. The comparison of these results for all data sizes m and the different kernel bandwidths is shown in Figure 3. We observe that the bound from Theorem 3.1 is quite accurate, over-estimating the actual error by about a factor three, and captures the detailed qualitative dependence of the estimation error on m. The coarser bound from Corollary 3.3, however, appears to discard too much information, it over-estimates the error by one to two orders of magnitude, and also does not capture the initial slope for small m. Finally, we note that for the larger kernel bandwidths, the i.i.d. variance is indeed too small, leading to an under-estimation of the error. This observation confirms that it is indeed necessary to take the effect of the correlation between data points into account.

In a second set of experiments, we test the performance of our theoretical bounds concerning the prediction of expectations for individual observables, obtained in Theorem 4.1. For the same three Gaussian RBF kernels as in the first set of experiments, we consider the observable $\phi=\varphi_0$, i.e., the first Mercer feature, and choose $N=10$ in the Mercer series expansion $K_N^t\phi$ and its empirical approximation $\hat{K}_N^{m,t}\phi$. Note that ϕ is a different observable depending on the bandwidth. Again, we set $1-\delta=0.9$, and use the bound from Theorem 4.1 to bound the L_μ^2-error between $K_N^t\phi$ and $\hat{K}_N^{m,t}\phi$. As a comparison, we compute the actual L_μ^2-error by numerical integration, using the fact that we can evaluate $K_N^t\phi$ and $\hat{K}_N^{m,t}\phi$ based on the discussion above. We repeat this procedure 15 times and provide average errors and standard deviations. The results for all three kernels are shown in Figure 4, and we find that our theoretical bounds are much too pessimistic in all cases. This finding highlights our previous observation that bounding the prediction error outside the RKHS still requires more in-depth research.

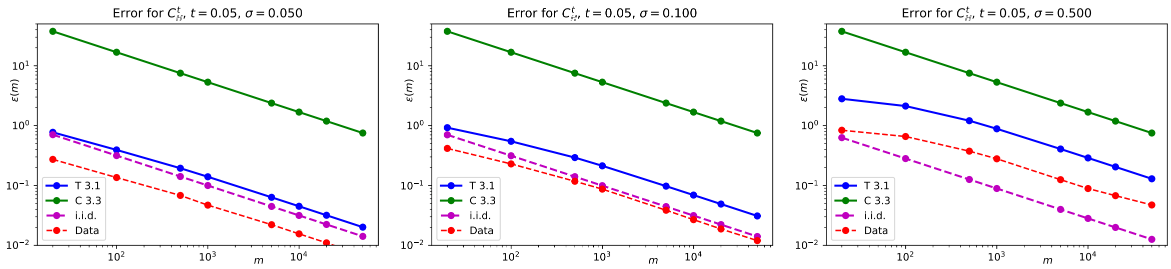

FIGURE 3. Probabilistic error estimates for $C^t_{\mathbb{H}}$ associated to the OU process, at lag time $t = 0.05$, and the Gaussian RBF kernel with different bandwidths $\sigma \in \{0.05, 0.1, 0.05\}$ (corresponding to left, center and right panels). The blue and green curves show the estimated error using the fine and coarse bounds from Theorem 4.1 and Corollary 3.3, respectively, while the purple curves represent the bound obtained from the i.i.d.-variance $\frac{1}{m}\mathbb{E}_0(t)$. The red curve shows the 0.9-percentile of the estimation error based on 200 independent simulations.

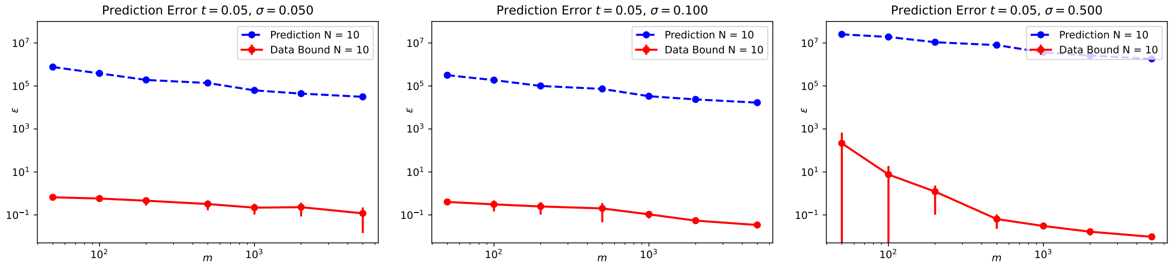

FIGURE 4. Comparison of the theoretical bound on the prediction error $\|K^t_N\phi - \hat{K}^{m,t}_N\phi\|_\mu$, if ϕ is chosen as the first Mercer feature φ_0, using $N = 10$ in the Mercer series representation. The predicted error is shown in blue, error bars for the actual error obtained from 15 independent data sets are shown in red. Different panels correspond to the same kernel bandwidths as in Figure 3 above.

6. CONCLUSIONS

We have analyzed the finite-data estimation error for data-driven approximations of the Koopman operator on reproducing kernel Hilbert spaces. More specifically, we have provided an exact expression for the variance of empirical estimators for the cross-covariance operator, if a sliding-window estimator is applied to a long ergodic trajectory of the dynamical system. This setting is relevant for many complex systems, such as molecular dynamics simulations. Our results present a significant improvement over the state of the art, since they concern a setting where the notorious problem of dictionary selection can be circumvented, and therefore no longer depend on the dictionary size. We have also extended the concept of asymptotic variance to an infinite-dimensional approximation space for the Koopman operator. Our numerical study on the Ornstein Uhlenbeck process has shown that, even using a simple mass concentration inequality, accurate bounds on the estimation error can be obtained.

In our second result, we have extended our estimates to a uniform bound on the prediction error for observables in the RKHS. Thereby, we have circumvented dealing with an unbounded inverse of the covariance operator by applying a finite-dimensional truncation of the associated Mercer series. In case of Koopman-invariance of the RKHS, however, we were able to find a bound on the truncation error which then yields estimates for the full approximation error.

Still, the resulting error bounds have proven very conservative in the numerical examples. Therefore, obtaining sharper bounds on the prediction error constitutes a primary goal for future research.

REFERENCES

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950), 337–404.

[2] A. Beck, and J. T. Schwartz, A vector-valued random ergodic theorem, Proc. Amer. Math. Soc. 8(6) (1957), 1049–1059.

[3] D. Bosq, Linear Processes in Function Spaces – Theory and Applications, Springer-Verlag New York, Inc., 2000.

[4] S. L. Brunton, J.L. Proctor, and J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (2016), 3932–3937.

[5] S. L. Brunton, M. Budisic, E. Kaiser, J. N. Kutz, Modern Koopman theory for dynamical systems, SIAM Rev. 64(2) (2022), 229–340.

[6] A. Berlinet and C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Kluwer Academic Publishers, 2004.

[7] O. Christensen, An Introduction to Frames and Riesz Bases, 2nd ed., Springer International Publishing Switzerland, 2016.

[8] R.G. Douglas, On majorization, factorization, and range inclusion of operators on Hilbert space, Proc. Amer. Math. Soc. 17 (1966), 413–415.

[9] G. E. Fasshauer and M. J. McCourt, Stable Evaluation of Gaussian Radial Basis Function Interpolants SIAM J. Sci. Comput. 2012 34:2, A737–A762.

[10] K. Fukumizu, L. Song, and A. Gretton, Kernel Bayes' Rule: Bayesian inference with positive definite kernels, J. Mach. Learn. Res. 14 (2013), 3753–3783.

[11] D. Giannakis, Data-driven spectral decomposition and forecasting of ergodic dynamical systems, Appl. Comput. Harmon. Anal, 47(2) (2019), 338–396.

[12] 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.

[13] M. Hairer, Ergodic properties of Markov processes, Lecture notes, https://www.hairer.org/notes/Markov.pdf

[14] M. Haseli and J. Cortés, Learning Koopman eigenfunctions and invariant subspaces from data: Symmetric subspace decomposition, IEEE Trans. Automat. Control 67(7) (2022), 3442-3457.

[15] O. Kallenberg, Foundations of modern probability, Springer-Verlag, New York, 1997.

[16] T. Kato, Perturbation Theory for Linear Operators, Springer-Verlag Berlin Heidelberg, 1995.

[17] E. Kaiser, J. N. Kutz, S. L. Brunton, Data-driven discovery of Koopman eigenfunctions for control, Mach. Learn. Sci. Technol. 2 (2021), 035023.

[18] S. Klus, P. Koltai, and C. Schütte, On the numerical approximation of the Perron–Frobenius and Koopman operator, J. Comput. Dyn. 1(3) (2016), 51–79.

[19] 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.

[20] S. Klus, I. Schuster, K. Muandet, Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces, J. Nonlinear Sci. 30(1) (2020), 283–315.

[21] S. Klus, and C. Schütte, Towards tensor-based methods for the numerical approximation of the Perron–Frobenius and Koopman operator, J. Comput. Dyn. 3(2) (2016), 139–161.

[22] II. Klebanov, I. Schuster, and T.J. Sullivan, A rigorous theory of conditional mean embeddings, SIAM J. Math. Data Sci. 2 (2020), 583–606.

[23] B. O. Koopman, Hamiltonian Systems and Transformations in Hilbert Space, Proc. Natl. Acad. Sci. 5 (17) (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] M. Korda and I. Mezić, Optimal construction of Koopman eigenfunctions for prediction and control, IEEE Trans. Automat. Control 65 (12) (2020), 5114–5129.

[26] A. J. Kurdila and P. Bobade, Koopman theory and linear approximation spaces, Preprint, arXiv:1811.10809 (2018).

[27] T. Lelièvre, and G. Stoltz, Partial differential equations and stochastic methods in molecular dynamics, Acta Numer. 25 (2016), 681–880.

[28] G. Lumer and R.S. Phillips, Dissipative operators in a Banach space, Pacific J. Math. 11 (1961), 679–698.

[29] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nat. Commun. 9(1) (2018), 1–10.

[30] M. Mollenhauer, On the Statistical Approximation of Conditional Expectation Operators, Dissertation, Freie Universität Berlin, 2021.

[31] K. Manohar, B. W. Brunton, J. N. Kutz, S. L. Brunton, Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. IEEE Control Systems Magazine, 38(3) (2018), 63–86.

[32] A. Mardt, L. Pasquali, H. Wu, and F. Noé, VAMPnets for deep learning of molecular kinetics, Nat. Commun. 9, 5 (2018).

[33] A. Mauroy, Y. Susuki, and I. Mezić, Koopman operator in systems and control, Springer International Publishing, Berlin, 2020.

[34] M. Mollenhauer, S. Klus, C. Schütte, and P. Koltai, Kernel autocovariance operators of stationary processes: Estimation and convergence, J. Mach. Learn. Res. 23(327) (2022), 1–34.

[35] F. Noé and F. Nüske, A variational approach to modeling slow processes in stochastic dynamical systems, Multiscale Model. Simul. 11 (2013), 635–655.

[36] 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.

[37] F. Nüske, P. Gelß, S. Klus, and C. Clementi, Tensor-based computation of metastable and coherent sets, Physica D 427 (2021), 133018.

[38] 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.

[39] B. Øksendal, Stochastic Differential Equations, An Introduction with Applications, Fifth Edition, Corrected Printing, Springer-Verlag Heidelberg New York, 2000.

[40] V. Paulsen, An introduction to the theory of reproducing kernel Hilbert spaces, https://www.math.uh.edu/ vern/rkhs.pdf

[41] G. A. Pavliotis, Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, Texts in Applied Mathematics, vol. 60, Springer, 2014.

[42] S. Peitz, and S. Klus, Koopman operator-based model reduction for switched-system control of PDEs, Automatica 106 (2019), 184–191.

[43] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, The Annals of Probability 22 (1994), 1679–1706.

[44] J. H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, Markov models of molecular kinetics: Generation and validation, J. Chem. Phys. 17(134) (2011), 174105.

[45] F. Riesz and B. Nagy, Functional Analysis, Blackie & Son Ltd., Glasgow, Bombay, Toronto, 1955.

[46] N. Rhomari, Approximation et inégalités exponentielles pour les sommes des vecteurs aléatoires dépendants, Comptes rendus de l'Académie des science, Série 1, 334 (2002), 149–154.

[47] W. Rudin, Functional Analysis, Second edition, McGraw-Hill, Inc., 1991.

[48] W. Rudin, Real and Complex Analysis, Third edition, McGraw-Hill, Inc., 1987.

[49] M. Schaller, K. Worthmann, F. Philipp, S. Peitz, and F. Nüske, Towards reliable data-based optimal and predictive control using extended DMD, IFAC PapersOnLine, to appear, arXiv preprint arXiv:2202.09084.

[50] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, A direct approach to conformational dynamics based on hybrid Monte Carlo, J. Comput. Phys. 1(151) (1999), 146–168.

[51] A. Smola, A. Gretton, L. Song, and B. Schölkopf, A Hilbert Space Embedding for Distributions, in: M. Hutter, R.A. Servedio, and E. Takimoto (Eds.): ALT 2007, Lecture Notes in Computer Science, vol. 4754, pp. 13–31, 2007. Springer, Berlin, Heidelberg.

[52] I. Steinwart and A. Christmann, Support Vector Machines, Springer Science+Business Media, LLC, 2008.

[53] B.K. Sriperumbudur, K. Fukumizu, and G.R.G. Lanckriet, Universality, characteristic kernels and RKHS embedding of measures, J. Mach. Learn. Res. 12 (2011), 2389–2410.

[54] 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(6) (2015), 1307–1346

[55] 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.

[56] H. Wu, and F. Noé, Variational approach for learning Markov processes from time series data, J. Nonlinear Sci. 30(1) (2020), 23–66.

[57] Y. Yu, T. Wang, and R. J. Samworth, A useful variant of the Davis-Kahan theorem for statisticians, Biometrika 102 (2015), 315–323.

[58] 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. Proofs

Proof of Lemma 2.1. Let $\psi \in \mathbb{H}$. Then (2.4) follows from

$$\int |\psi(x)|^2 \, d\mu(x) = \int |\langle \psi, \Phi(x)\rangle|^2 \, d\mu(x) \;\le\; \|\psi\|^2 \int \varphi(x)\, d\mu(x) = \|\psi\|^2 \|\varphi\|_1.$$

Assume that (A2) holds and that $\psi \in L^2_\mu(\mathcal{X})$ is such that $\langle \psi, \Phi(x)\rangle_\mu = 0$ for all $x \in \mathcal{X}$. Then

$$0 = \int \langle \psi, \Phi(x)\rangle_\mu \psi(x)\, d\mu(x) = \int\!\!\int k(x,y)\psi(x)\psi(y)\, d\mu(x)\, d\mu(y).$$

Hence, $\psi = 0$ by (A2). Conversely, assume that $\mathbb{H}$ is dense in $L^2_\mu(\mathcal{X})$. Let $\psi \in L^2_\mu(\mathcal{X})$ such that $\int\!\int k(x,y)\psi(x)\psi(y)\, d\mu(x)\, d\mu(y) = 0$. Since the integrand equals $\langle \psi(x)\Phi(x), \psi(y)\Phi(y)\rangle$ and the integral $\int \psi(x)\Phi(x)\, d\mu(x)$ exists by (2.5), we obtain $\int \psi(x)\Phi(x)\, d\mu(x) = 0_{\mathbb{H}}$. This implies that $\langle \psi, \Phi(y)\rangle_\mu = \int \psi(x) k(x,y)\, d\mu(x) = 0$ for each $y \in \mathcal{X}$. Hence, $\langle \psi, \phi\rangle_\mu = 0$ for each $\phi \in \mathcal{H} := \operatorname{span}\{\Phi(x) : x \in \mathcal{X}\}$. Now, let $\phi \in \mathbb{H}$. Then there exists a sequence $(\phi_n) \subset \mathcal{H}$ such that $\|\phi_n - \phi\| \to 0$ as $n \to \infty$. Therefore,

$$|\langle \psi, \phi\rangle_\mu| = |\langle \psi, \phi - \phi_n\rangle_\mu| \;\le\; \|\psi\|_\mu \|\phi - \phi_n\|_\mu \;\le\; \|\psi\|_\mu \sqrt{\|\varphi\|_1}\|\phi - \phi_n\|.$$

Hence, $\langle \psi, \phi\rangle_\mu = 0$, and the density of $\mathbb{H}$ in $L^2_\mu(\mathcal{X})$ implies $\psi = 0$. □

Proof of Lemma 2.4. (a) For $\psi \in L^2_\mu(\mathcal{X})$ we have

$$\|\mathcal{E}\psi\|^2 = \int\!\!\int \psi(x)\psi(y)\langle \Phi(x), \Phi(y)\rangle\, d\mu(x)\, d\mu(y) = \int\!\!\int k(x,y)\psi(x)\psi(y)\, d\mu(x)\, d\mu(y).$$

Hence, the injectivity of $\mathcal{E}$ follows from (A2). If (e_i) is an orthonormal basis of $\mathbb{H}$, then

$$\sum_i \|\mathcal{E}^* e_i\|_\mu^2 = \sum_i \|e_i\|_\mu^2 = \sum_i \int |e_i(x)|^2\, d\mu(x) = \sum_i \int |\langle \Phi(x), e_i\rangle|^2\, d\mu(x) = \int \|\Phi(x)\|^2\, d\mu(x).$$

The claim is now a consequence of $\|\Phi(x)\|^2 = \varphi(x)$.

(b) By Lemma 2.1, $\mathbb{H}$ is dense in $L^2_\mu(\mathcal{X})$. Moreover, $\mathcal{E}^*$ is compact by (a) and Schauder's theorem [47, Theorem 4.19].

(c) This follows from (a) and $\ker C_{\mathbb{H}} = \ker \mathcal{E}\mathcal{E}^* = \ker \mathcal{E}^* = \{0\}$ by (A3). □

Proof of Theorem 2.5. By Lemma 2.4, the operator $\mathcal{E}^*\mathcal{E} \in \mathcal{B}(L^2_\mu(\mathcal{X}))$ is a positive self-adjoint trace-class operator. Hence, by the well known spectral theory of compact operators (see, e.g., [12]) there exists an orthonormal basis $(e_j)_{j=1}^\infty$ of $L^2_\mu(\mathcal{X})$ consisting of eigenfunctions of $\mathcal{E}^*\mathcal{E}$ corresponding to a summable sequence $(\lambda_j)_{j=1}^\infty$ of strictly positive eigenvalues. Since $\mathcal{E}^*\psi = \psi$ for $\psi \in \mathbb{H}$, we have $\mathcal{E}e_j = \lambda_j e_j$ and thus $e_j \in \mathbb{H}$ for all j and $C_{\mathbb{H}} e_j = \mathcal{E}\mathcal{E}^* e_j = \mathcal{E}e_j = \lambda_j e_j$. Moreover, $\langle f_i, f_j\rangle = \sqrt{\lambda_j/\lambda_i}\langle \mathcal{E}e_i, e_j\rangle = \sqrt{\lambda_j/\lambda_i}\langle e_i, e_j\rangle_\mu = \delta_{ij}$ by (2.6) so that the f_j indeed form an orthonormal system in $\mathbb{H}$. The completeness of (f_j) in $\mathbb{H}$ follows from the injectivity of $\mathcal{E}$. Finally, $\sum_{j=1}^\infty \lambda_j = \operatorname{Tr} C_{\mathbb{H}} = \|\varphi\|_1$ and

$$k(x,y) = \langle \Phi(x), \Phi(y)\rangle = \sum_j \langle \Phi(x), f_j\rangle\langle f_j, \Phi(y)\rangle = \sum_j f_j(x) f_j(y),$$

which completes the proof. □

Proof of Proposition 2.7. Let $\psi \in B(\mathcal{X})$. For $p = \infty$ we have $|(K^t\psi)(x)| = |\mathbb{E}^x[\psi(X_t)]| \le \mathbb{E}^x[|\psi(X_t)|] \le \|\psi\|_\infty$. If $p < \infty$, by Jensen's inequality, for every convex $\phi : \mathbb{R} \to \mathbb{R}$ we have $\phi \circ K^t\psi \le K^t(\phi \circ \psi)$ and thus $|K^t\psi|^p \le K^t|\psi|^p$, which, by invariance of μ, leads to

$$\|K^t\psi\|_p^p = \int |K^t\psi|^p\, d\mu \le \int K^t|\psi|^p\, d\mu = \int |\psi|^p\, d\mu = \|\psi\|_p^p.$$

The claim now follows by density of $B(\mathcal{X})$ in $L^p_\mu(\mathcal{X})$. □

Proof of Proposition 2.8. Let $\psi \in C_b(\mathcal{X})$ and fix $x \in \mathcal{X}$. Denote the stochastic solution process of the SDE (2.1) with initial value x by X_t^x. Since $X_t^x(\omega)$ is continuous in t for $\mathbb{P}$-a.e. $\omega \in \Omega$ (see [39, Theorem 5.2.1]), $\psi(X_t^x(\omega)) \to \psi(X_0^x(\omega)) = \psi(x)$ as $t \to 0$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. Hence, by dominated convergence,

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

as $t \to 0$. It now follows from Proposition 2.7 and, again, dominated convergence that $\|K^t\psi - \psi\|_p \to 0$ as $t \to 0$. If $\psi \in L^p_\mu(\mathcal{X})$ and $\varepsilon > 0$, there exists $\eta \in C_b(\mathcal{X})$ such that $\|\psi - \eta\|_p < \varepsilon/3$. Choose $\delta > 0$ such that $\|K^t\eta - \eta\|_p < \varepsilon/3$ for $t < \delta$. Then

$$\|K^t\psi - \psi\|_p \le \|K^t(\psi - \eta)\|_p + \|K^t\eta - \eta\|_p + \|\eta - \psi\|_p < \varepsilon$$

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

Appendix B. Riesz bases

Recall that a Riesz basis [7] of a Hilbert space $\mathcal{H}$ is a sequence $(\psi_j) \subset \mathcal{H}$ satisfying $\overline{\operatorname{span}}\{\psi_j\} = \mathcal{H}$ and for which there exist $A, B > 0$ such that for all $c \in \ell^2$,

$$A\|c\|_2 \;\le\; \Big\|\sum_j c_j\psi_j\Big\|_{\mathcal{H}} \;\le\; B\|c\|_2.$$

The constant A (B, resp.) is called a lower (upper, resp.) Riesz bound of the basis. Also recall that to every Riesz basis (ψ_j) there exists a dual Riesz basis $(\widetilde{\psi}_j)$ such that $\langle \psi_j, \widetilde{\psi}_k\rangle_{\mathcal{H}} = \delta_{jk}$. If (ψ_j) has the bounds A and B, then $(\widetilde{\psi}_j)$ has bounds $1/B$ and $1/A$. Every element f of $\mathcal{H}$ admits a representation $f = \sum_j \langle f, \widetilde{\psi}_j\rangle_{\mathcal{H}}\psi_j = \sum_j \langle f, \psi_j\rangle_{\mathcal{H}}\widetilde{\psi}_j$ and

$$A^2\|f\|_{\mathcal{H}}^2 \le \sum_j |\langle f, \psi_j\rangle|^2 \le B^2\|f\|_{\mathcal{H}}^2 \qquad \text{and} \qquad B^{-2}\|f\|_{\mathcal{H}}^2 \le \sum_j |\langle f, \widetilde{\psi}_j\rangle|^2 \le A^{-2}\|f\|_{\mathcal{H}}^2.$$

It can furthermore be easily seen that a sequence $(\psi_j) \subset \mathcal{H}$ is a Riesz basis of $\mathcal{H}$ if and only if there exists a boundedly invertible linear operator $S \in L(\mathcal{H})$ and an orthonormal basis (e_j) of $\mathcal{H}$ such that $\psi_j = Se_j$ for all j. Then $\widetilde{\psi}_j = (S^{-1})^*e_j$ for all j, $B = \|S\|$, and $A = \|S^{-1}\|^{-1}$.

Appendix C. Some facts from spectral theory

In this section, let $\mathcal{H}$ be a Hilbert space. If P is an orthogonal projection in $\mathcal{H}$, we set $P^\perp = I - P$. For $v \in \mathcal{H}$, $\|v\| = 1$, denote by P_v the rank-one orthogonal projection onto $\operatorname{span}\{v\}$.

We say that a linear operator on $\mathcal{H}$ is *non-negative* if it is self-adjoint and its spectrum is contained in $[0, \infty)$. For a non-negative compact operator T on $\mathcal{H}$ we denote by $\lambda_1(T) \ge \lambda_2(T) \ge \dots$ the eigenvalues of T in descending order (counting multiplicities). We set $\lambda_j(T) = 0$ if $j > \operatorname{rank}(T)$. Moreover, if T has only simple eigenvalues[6], we let $P_j(T)$ denote the orthogonal projection onto the eigenspace $\ker(T - \lambda_j(T))$ and $Q_n(T) = \sum_{j=1}^n P_j(T)$ the spectral projection corresponding to the n largest eigenvalues of T.

Theorem C.1 ([12, Cor. II.2.3]). *If T and $\widehat{T}$ are two non-negative compact operators on $\mathcal{H}$, then for all $j \in \mathbb{N}$,*

$$|\lambda_j(T) - \lambda_j(\widehat{T})| \;\le\; \|T - \widehat{T}\|.$$

[6] i.e., $\dim\ker(T - \lambda) = 1$ for each eigenvalue λ of T

Lemma C.2. *For $v, w \in \mathcal{H}$ with $\|v\| = \|w\| = 1$ we have*

$$\|P_v - P_w\| = \|P_w^\perp P_v\| = \sqrt{1 - |\langle v, w\rangle|^2}. \tag{C.1}$$

Proof. First of all, the second equation in (C.1) is clear, since

$$\|P_w^\perp P_v f\|^2 = \|\langle f, v\rangle P_w^\perp v\|^2 = |\langle f, v\rangle|^2(1 - \|P_w v\|^2) = |\langle f, v\rangle|^2(1 - |\langle v, w\rangle|^2).$$

Second, if $P_{v,w}$ denotes the orthogonal projection onto $\mathcal{H}_{v,w} := \operatorname{span}\{v, w\}$, we have

$$\|P_v - P_w\| = \|(P_v - P_w)P_{v,w}\| = \|(P_v - P_w)|_{\mathcal{H}_{v,w}}\| = \sup_{x\in\mathcal{H}_{v,w},\,\|x\|=1} \|(P_v - P_w)x\|,$$

which is a two-dimensional problem in $\mathcal{H}_{v,w}$. Now, if $x \in \mathcal{H}_{v,w}$, $\|x\| = 1$, we write $x = av + bw$ and obtain $a^2 + 2ab\gamma + b^2 = 1$, where $\gamma = \langle v, w\rangle$. Moreover, $\langle x, v\rangle = a + b\gamma$, $\langle x, w\rangle = a\gamma + b$ and so

$$\begin{aligned}\|(P_v - P_w)x\|^2 &= \|\langle x, v\rangle v - \langle x, w\rangle w\|^2 = \|(a + b\gamma)v - (a\gamma + b)w\|^2 \\ &= (a + b\gamma)^2 - 2(a + b\gamma)(a\gamma + b)\gamma + (a\gamma + b)^2 \\ &= a^2 + 2ab\gamma + b^2\gamma^2 - 2\gamma(a^2\gamma + ab\gamma^2 + ab + b^2\gamma) + a^2\gamma^2 + 2ab\gamma + b^2 \\ &= (1 - \gamma^2)a^2 + 2ab\gamma - 2ab\gamma^3 + b^2(1 - \gamma^2) \\ &= (1 - \gamma^2)(a^2 + b^2 + 2ab\gamma) \\ &= 1 - |\langle v, w\rangle|^2.\end{aligned}$$

Hence, the objective function is constant on $\{x \in \mathcal{H}_{v,w} : \|x\| = 1\}$ and (C.1) is proved. □

The next theorem is a variant of the Davis-Kahan $\sin(\Theta)$ theorem (cf. [57]).

Theorem C.3. *Let T and $\widehat{T}$ be non-negative Hilbert-Schmidt operators on $\mathcal{H}$, let $n \in \mathbb{N}$, assume that the largest $n + 1$ eigenvalues of T are simple, and set*

$$\delta = \min_{j=1,\dots,n} \frac{\lambda_j(T) - \lambda_{j+1}(T)}{2}.$$

If $\|T - \widehat{T}\|_{HS} < \delta$, then for $j = 1, \dots, n$ we have

$$\|P_j(T) - P_j(\widehat{T})\| \le \frac{\|T - \widehat{T}\|}{\delta}.$$

Proof. For $j \in \mathbb{N}$ put $\lambda_j = \lambda_j(T)$, $P_j = P_j(T)$, $\widehat{\lambda}_j = \lambda_j(\widehat{T})$, and $\widehat{P}_j = P_j(\widehat{T})$. By Theorem C.1, we have $|\lambda_j - \widehat{\lambda}_j| \le \|T - \widehat{T}\|_{HS} < \delta$ for all j, hence $\widehat{\lambda}_j$ is contained in the interval $I_j = (\lambda_j - \delta, \lambda_j + \delta)$ for $j = 1, \dots, n + 1$. By assumption, $\sup I_{j+1} \le \inf I_j$ for $j = 1, \dots, n$. In particular, the intervals $I_1, \dots, I_{n+1}$ are pairwise disjoint.

Now, let $j \in \{1, \dots, n\}$. Then for $k \in \mathbb{N} \setminus \{j\}$ we have $|\widehat{\lambda}_k - \lambda_j| > \delta$. Therefore, we have $\operatorname{dist}(\lambda_j, \sigma(\widehat{T})\backslash\{\widehat{\lambda}_j\}) \ge \delta$ and thus, for $f \in \widehat{P}_j^\perp\mathcal{H}$,

$$\|(\widehat{T} - \lambda_j)f\| \ge \operatorname{dist}\big(\lambda_j, \sigma(\widehat{T}|_{\widehat{P}_j^\perp\mathcal{H}})\big)\|f\| = \operatorname{dist}(\lambda_j, \sigma(\widehat{T})\backslash\{\widehat{\lambda}_j\})\|f\| \ge \delta\|f\|.$$

As $TP_j = \lambda_j P_j$ and $\widehat{P}_j^\perp\widehat{T} = \widehat{T}\widehat{P}_j^\perp$, we obtain

$$\|T - \widehat{T}\| \ge \|\widehat{P}_j^\perp(\widehat{T} - T)P_j\| = \|\widehat{P}_j^\perp\widehat{T}P_j - \widehat{P}_j^\perp TP_j\| = \|(\widehat{T} - \lambda_j)\widehat{P}_j^\perp P_j\| \ge \delta\|\widehat{P}_j^\perp P_j\|.$$

The claim now follows from Lemma C.2. □

Appendix D. Ergodicity and the generator

In this section, we prove the following proposition on the spectral properties of the generator $\mathcal{L}$ under the ergodicity assumption.

Proposition D.1. *Assume that the invariant measure μ is ergodic. Then* $\ker\mathcal{L} = \operatorname{span}\{\mathbb{1}\}$ *and* $\ker(\mathcal{L} - i\omega I) = \{0\}$ *for* $\omega \in \mathbb{R}\backslash\{0\}$.

Proof. First of all, it is worth mentioning that $\mathcal{L}\psi = 0$ implies $K^t\psi = \psi$ for all $t \geq 0$ and that $\mathcal{L}\psi = i\omega\psi$, $\omega \in \mathbb{R} \setminus \{0\}$, implies $K^{2\pi/\omega}\psi = \psi$. Therefore, it suffices to show that $K^t\psi = \psi$ for some $t > 0$ and $\psi \in L^2_\mu(\mathcal{X})$ is only possible for constant ψ. For this, we consider the Markov process $(X_{nt})_{n=0}^\infty$. For convenience, we assume w.l.o.g. that $t = 1$ holds. By invariance of μ, the process $(X_n)_{n=0}^\infty$ is stationary, i.e., $(X_n)_{n=0}^\infty$ and $(X_{n+1})_{n=0}^\infty$ are equally distributed as $\mathcal{X}^{\mathbb{N}_0}$-valued random variables. According to [15, Lemma 9.2] there exist $\mathcal{X}$-valued random variables X_{-k}, $k \in \mathbb{N}$, such that $X := (X_n)_{n\in\mathbb{Z}}$ is also stationary. By P_μ denote the law of the $\mathcal{X}^{\mathbb{Z}}$-valued random variable X.

On $S := \mathcal{X}^{\mathbb{Z}}$ define the left shift $T : S \to S$ by $T(x_n)_{n\in\mathbb{Z}} := (x_{n+1})_{n\in\mathbb{Z}}$. Stationarity of X means that also $TX \sim P_\mu$.

A set $\mathcal{A} \in \mathcal{B}_{\mathcal{X}}^{\mathbb{Z}} := \bigotimes_{k\in\mathbb{Z}} \mathcal{B}_{\mathcal{X}}$ is called shift-invariant if $T^{-1}\mathcal{A} = \mathcal{A}$. It is easy to see that the set of shift-invariant sets forms a sub-σ-algebra $\mathcal{I}$ of $\mathcal{B}_{\mathcal{X}}^{\mathbb{Z}}$. Now, by [13, Corollary 5.11] and the ergodicity of μ we have $P_\mu(\mathcal{A}) \in \{0, 1\}$ for any $\mathcal{A} \in \mathcal{I}$. Now, Birkhoff's Ergodic Theorem [15, Theorem 9.6] states that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k X) = \mathbb{E}\big[f(X)|X^{-1}\mathcal{I}\big] \tag{D.1}$$

almost surely and in $L^1(\Omega)$ for any $f \in L^1(S)$. Given $\psi \in L^1_\mu(\mathcal{X})$, let us apply this theorem to the function $f = \psi \circ \pi_0$, where the projection $\pi_0 : S \to X$ is defined by $\pi_0(x_n)_{n\in\mathbb{Z}} = x_0$. First of all,

$$\int |f|\, dP_\mu = \int |\psi(x_0)|\, dP_\mu((x_n)_{n\in\mathbb{Z}}) = \int |\psi(x)|\, d\mu(x) < \infty$$

as $P_\mu \circ \pi_0^{-1} = \mu$. Hence, we have $f \in L^1(S)$. Furthermore, we compute $f(T^kX) = \psi(\pi_0(T^kX)) = \psi(X_k)$. For $\mathcal{A} \in \mathcal{I}$ we have $\mathbb{P}(X^{-1}\mathcal{A}) = P_\mu(\mathcal{A}) \in \{0, 1\}$. Thus, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \psi(X_k) = \mathbb{E}[f(X)] = \int f\, dP_\mu = \int \psi \circ \pi_0\, dP_\mu = \int \psi\, d\mu$$

almost surely and in $L^1(\Omega)$.

Therefore, if $\psi \in L^2_\mu(\mathcal{X})$ such that $K^t\psi = \psi$, then $K^{kt}\psi = \psi$ for all $k \in \mathbb{N}_0$, hence for μ-a.e. $x \in X$ we have

$$\begin{aligned}\psi(x) &= \frac{1}{n}\sum_{k=0}^{n-1}\psi(x) = \frac{1}{n}\sum_{k=0}^{n-1}K^{kt}\psi(x) = \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[\psi(X_{kt})|X_0 = x] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\psi(X_{kt})\,\middle|\, X_0 = x\right] \overset{n\to\infty}{\longrightarrow} \int \psi\, d\mu.\end{aligned}$$

Thus, ψ must indeed be (μ-essentially) constant. □

AUTHOR AFFILIATIONS

F. Philipp TECHNISCHE UNIVERSITÄT ILMENAU, INSTITUTE FOR MATHEMATICS, WEIMARER STRASSE 25, D-98693 ILMENAU, GERMANY

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

M. Schaller TECHNISCHE UNIVERSITÄT ILMENAU, INSTITUTE FOR MATHEMATICS, WEIMARER STRASSE 25, D-98693 ILMENAU, GERMANY

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

K. Worthmann TECHNISCHE UNIVERSITÄT ILMENAU, INSTITUTE FOR MATHEMATICS, WEIMARER STRASSE 25, D-98693 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

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