Extended Dynamic Mode Decomposition: Sharp bounds on the sample efficiency

Friedrich M. Philipp [1] **Manuel Schaller** [1] **Septimus Boshoff** [2] **Sebastian Peitz** [2] **Feliks Nüske** [3]
Karl Worthmann [1]

Abstract

We rigorously derive novel and sharp finite-data error bounds for highly sample-efficient Extended Dynamic Mode Decomposition (EDMD) for both i.i.d. and ergodic sampling. In particular, we show all results in a very general setting removing most of the typically imposed assumptions such that, among others, discrete- and continuous-time stochastic processes as well as nonlinear partial differential equations are contained in the considered system class. Besides showing an exponential rate for i.i.d. sampling, we prove, to the best of our knowledge, the first superlinear convergence rates for ergodic sampling of deterministic systems. We verify sharpness of the derived error bounds by conducting numerical simulations for highly-complex applications from molecular dynamics and chaotic flame propagation.

1. Introduction

Extended Dynamic Mode Decomposition (EDMD; (Williams et al., 2015)) is one of the most commonly used machine-learning methods for identifying highly-nonlinear and, in addition, possibly infinite-dimensional dynamical systems from data. At its heart, EDMD provides a data-driven approach to learn the Koopman operator (Koopman, 1931) propagating observable functions along the flow, which results in a purely data-driven and well-interpretable surrogate model for analysis, prediction, and control, e.g., based on identified symmetries for data augmentation (Weissenbacher et al., 2022) or deep learning (Han et al., 2021). Since the Koopman operator is a linear, it serves as a powerful tool to leverage well-established concepts from approximation, ergodic, operator, and statistical-learning theory in certifiable machine learning also in safety-critical applications.

EDMD and Koopman-based methods have successfully enabled data-driven simulations and analysis of various highly complex applications, such as molecular dynamics (Schütte et al., 2016; Wu et al., 2017; Klus et al., 2018), nonlinear partial differential equations including turbulent flows (Giannakis et al., 2018; Mezić, 2013), quantum mechanics (Klus et al., 2022), neuroscience (Brunton et al., 2016), deep learning (Dogra & Redman, 2020), electrocardiography (Golany et al., 2021) or climate prediction (Azencot et al., 2020) to name just a few. For further applications and Koopman-related learning architectures, we refer to Kutz et al. (2016); Mauroy et al. (2020); Brunton et al. (2022); Retchin et al. (2023).

Convergence of EDMD to the Koopman operator in the infinite data limit was proven in Korda & Mezić (2018). However, despite the enormous success of EDMD, error bounds depending on the number of data samples are still scarce. First finite-data error bounds for a class of finite-dimensional systems were given in Zhang & Zuazua (2023) for systems governed by nonlinear ordinary differential equations (ODEs) based on i.i.d. (independently and identically distributed) sampling and in Nüske et al. (2023) for nonlinear stochastic differential equations (SDEs) based on i.i.d. and ergodic sampling, including an extension to control systems. However, for ergodic sampling, the restrictive assumption of exponential stability was imposed on the Koopman semigroup, which significantly limited the range of applications. For observable functions contained in a Reproducing Kernel Hilbert Space (RKHS), bounds were only recently provided for prediction (Philipp et al., 2023a) and control (Philipp et al., 2023b) of continuous-time systems, and by (Kostic et al., 2023) and (Kostic et al., 2022) for i.i.d. sampling from an invariant measure. In conclusion, so far only finite-data error bounds exist for continuous-time

[1]Technische Universität Ilmenau, Optimization-based Control Group, Institute of Mathematics, Ilmenau, Germany [2]Paderborn University, Department of Computer Science, Paderborn, Germany [3]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany. Correspondence to: Friedrich Philipp <friedrich.philipp@tu-ilmenau.de>.

systems and, if ergodic sampling is considered, under quite restrictive conditions. Moreover, all bounds decay at most linearly in the amount of data used for estimation.

In this work, we provide a complete analysis of the estimation error of EDMD for a much larger class of Markov processes in Polish spaces. In particular, we remove the (restrictive) requirement of exponential stability and the restriction to systems governed by stochastic differential equations in (Nüske et al., 2023). This broadens the class of systems covered by our results such that, in addition, nonlinear partial differential equations and discrete-time Markov processes are included. In particular, all systems considered in (Rozwood et al., 2023) are covered rendering the proposed techniques accessible for a sophisticated error analysis w.r.t. the number of used data samples. We provide *sharp* bounds on the convergence rate and, thus, the sampling efficiency of EDMD, i.e., Koopman-based machine learning for both i.i.d. data and ergodic sampling. Since data can be collected from a single (sufficiently-long) trajectory, which considerably facilitates the data collection process, ergodic sampling is of particular interest for many practical applications as demonstrated in our examples.

Our contribution in this work is three-fold:

(1) We severely weaken assumptions made in previous works for ergodic sampling of stochastic systems and prove sharp error bounds with a linear rate.
(2) We derive the first error bounds showing superlinear convergence for ergodic sampling.
(3) We establish all our results for continuous- and discrete-time systems – for i.i.d. and ergodic sampling.

Notation: We denote the constant function $x \mapsto 1$ by $\mathbb{1}$. Furthermore, the notation $\langle \cdot, \cdot \rangle_F$ and $\|\cdot\|_F$ is used for the Frobenius scalar product on $\mathbb{R}^{n\times m}$ and its corresponding norm, respectively. We use the notation $[n : m] := \mathbb{Z} \cap [n, m]$. For a probability measure μ, the scalar product and the norm on $L^2(\mu)$ are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. The orthogonal projection w.r.t. a closed subspace M in $L^2(\mu)$ is denoted by P_M.

2. EDMD: Data-driven prediction of nonlinear dynamics

Let $(X_n)_{n\in\mathbb{N}_0}$ be a time-homogeneous discrete-time Markov process taking its values in a Polish space $\mathcal{X}$. One representative system class is given by *discrete-time* dynamical systems

$$x_{n+1} = T(x_n) + \varepsilon_n \tag{DTDS}$$

with independently and identically distributed (i.i.d.) noise ε_n. Another system class contained in our setting are processes, which arise from samples of a time-homogeneous *continuous-time* Markov process $(Y_t)_{t\ge 0}$, i.e., $X_n = Y_{n\Delta t}$ with sampling period $\Delta t > 0$. The process Y_t might, e.g., be the solution of a stochastic differential equation

$$dY_t = f(Y_t)\,dt + \sigma(Y_t)\,dW_t. \tag{SDE}$$

This includes their deterministic analogues, i.e., (DTDS) with $\varepsilon_n \equiv 0$ and (SDE) with $\sigma \equiv 0$.

Let $\rho : \mathcal{X} \times \mathfrak{B}(\mathcal{X}) \to [0, 1]$ denote the transition kernel associated with (X_n), where $\mathfrak{B}(\mathcal{X})$ denotes the Borel sigma algebra on $\mathcal{X}$, i.e., $\rho(x, A) = \mathbb{P}(X_{n+1} \in A \mid X_n = x)$. Then we have $\mathbb{P}^{X_{n+1}}(A) = \int \rho(x, A)\, d\mathbb{P}^{X_n}(x)$, where $\mathbb{P}^X$ denotes the law of a random variable X. Using the notation $\rho_0(x, A) := \delta_x(A)$ with the Dirac measure δ_x, we iteratively define $\rho_{n+1}(x, A) := \int \rho_n(y, A)\, \rho(x, dy)$ for $x \in \mathcal{X}$ and $A \in \mathfrak{B}(\mathcal{X})$, where $\rho(x, dy)$ stands for $d\rho(x, \cdot)(y)$. Then, $\rho = \rho_1$ and $\mathbb{P}^{X_n}(A) = \int \rho_n(x, A)\, d\mathbb{P}^{X_0}(x)$ hold. For discrete-time deterministic dynamics, i.e., (DTDS) with $\varepsilon_n \equiv 0$, we have $\rho(x, A) = \delta_x(T^{-1}(A))$.

The Koopman operator and EDMD. Let μ be a Borel probability measure on $\mathcal{X}$ satisfying

$$\int \rho(x, A)\, d\mu(x) \;\le\; L^2 \mu(A), \qquad A \in \mathfrak{B}(\mathcal{X}), \tag{1}$$

with a constant $L \ge 0$, see (Philipp et al., 2023b) for a detailed discussion. For $p \in [1, \infty)$, the *linear*, but infinite-dimensional Koopman operator $K_p : L^p(\mu) \to L^p(\mu)$ of the nonlinear process X_n is defined by the identity

$$(K_p\psi)(x) = \mathbb{E}\big[\psi(X_1) \,|\, X_0 = x\big] = \int \psi(y)\, \rho(x, dy) \tag{2}$$

for all $\psi \in L^p(\mu)$.[1] In the deterministic case (DTDS) with $\varepsilon = 0$, this reduces to $K_p\psi = \psi \circ T$. An iterative application of (2) yields $(K_p^n\psi)(x) = \int \psi(y)\, \rho_n(x, dy)$ for $n \in \mathbb{N}$. We set $K := K_2$.

Let us briefly recall the well-known extended dynamic mode decomposition (EDMD, see Williams et al. (2015)), which aims at approximating the Koopman operator. To this end, let a dictionary $\mathcal{D} = \{\psi_1, \ldots, \psi_N\} \subset L^2(\mu)$ of μ-linearly independent (cf. Definition C.1) continuous functions on $\mathcal{X}$ be given. If we define the N-dimensional subspace $\mathbb{V} := \operatorname{span}\mathcal{D}$ and the matrices $C, C_+ \in \mathbb{R}^{N\times N}$ by

$$C = \big(\langle \psi_i, \psi_j\rangle\big)_{i,j=1}^N \quad \text{and} \quad C_+ = \big(\langle \psi_i, K\psi_j\rangle\big)_{i,j=1}^N,$$

then C is invertible, and the matrix representation $K_{\mathbb{V}}$ of the compression $P_{\mathbb{V}}K|_{\mathbb{V}}$ w.r.t. the basis $\mathcal{D}$ is given by

$$K_{\mathbb{V}} = C^{-1}C_+, \tag{3}$$

as rigorously shown in Lemma (C.2). In EDMD, the matrix $K_{\mathbb{V}}$ is learned by using evaluations of the dictionary observables on data samples $(x_k, y_k) \in \mathcal{X} \times \mathcal{X}$, $k \in [0 : m-1]$, which are collected in the $N \times m$ data matrices

$$\Psi_X = \big[\Psi(x_k)\big]_{k=0}^{m-1} \quad \text{and} \quad \Psi_Y = \big[\Psi(y_k)\big]_{k=0}^{m-1},$$

where $\Psi = [\psi_1, \ldots, \psi_N]^\top$. Then, the matrix representation of the Koopman operator estimator is given by

$$\widehat{K}_m = \widehat{C}^{-1}\widehat{C}_+, \tag{4}$$

using the empirical estimators of C and C_+, respectively, i.e., $\widehat{C} = \frac{1}{m}\Psi_X\Psi_X^\top$ and $\widehat{C}_+ = \frac{1}{m}\Psi_X\Psi_Y^\top$.

In this work, we distinguish between two different sampling schemes.

(S1) Ergodic sampling $\mu = \pi$. We assume the existence of an invariant probability measure π for X_n, i.e.,

$$\int \rho(x, A)\, d\pi(x) = \pi(A), \qquad A \in \mathfrak{B}(\mathcal{X}). \tag{5}$$

In the deterministic case of (DTDS), invariance of π corresponds to $\pi(T^{-1}(A)) = \pi(A)$ for all $A \in \mathfrak{B}(\mathcal{X})$, i.e., T is measure-preserving. Further, π is assumed to be ergodic, i.e., whenever $A \in \mathfrak{B}(\mathcal{X})$ is such that $\rho(x, A) = 1$ for all $x \in A$, then $\pi(A) \in \{0, 1\}$.

In this case, the Koopman operator K_p is a contraction in $L^p(\pi)$ and even an isometry in the deterministic case. The EDMD data consists of samples $x_k = X_k$ from a single trajectory of X_n with $x_0 \sim \pi$ and $y_k = x_{k+1}$. To ensure a.s. invertibility of $\widehat{C}$, we shall assume that for each $(N-1)$-dimensional subspace $M \subset \mathbb{R}^N$ and $x \in \Psi^{-1}(M)$, we have $\rho(x, \Psi^{-1}(M)) = 0$, see Appendix C for a proof of this sufficient condition.

(S2) I.i.d. sampling $\mu = \nu$. Let ν be any probability distribution on $\mathcal{X}$ satisfying (1) with some constant $L \geq 0$. In this case, the EDMD data consists of i.i.d. samples $x_k \in \mathcal{X}$ and $y_k|(x_k = x) \sim \rho(x, \cdot)$. Here, we assure the almost sure invertibility of $\widehat{C}$ by assuming that $m \geq N$ and $\psi_1, \ldots, \psi_N$ are strongly μ-linearly independent, see Appendix C for details including a proof of this characterization.

In the remainder of the manuscript, we tacitly assume $\varphi := \sum_{j=1}^N \psi_j^2 \in L^2(\mu)$ in both cases $\mu = \nu$ and $\mu = \pi$ to ensure the existence of the variances of $\widehat{C}_+$ and $\widehat{C}$.

3. Certifiable and efficient machine learning

In this section, we provide a concise overview on our main results. In particular, we present novel and sharp error bounds depending on the amount of data samples m. The key tool enabling us to derive the error estimates and the respective convergence rates are adroitly-composed formulas to represent the variance, which are shown – together with the in-depth analysis – in the subsequent Section 4.

[1] In Appendix B, we show that $\psi \in L^p(\mu)$ is $\rho(x, \cdot)$-integrable for μ-a.e. $x \in \mathcal{X}$. Hence, the Koopman operator is well defined. In fact, condition (1) is both necessary and sufficient for K_p to be well-defined and bounded (with $\|K_p\| \leq L^{2/p}$).

		Sampling		
		i.i.d.∼ probability distribution ν $\int \rho(x, A)\, \mathrm{d}\nu(x) \leq L^2 \nu(A)$	invariant and ergodic measure π $\int \rho(x, A)\, \mathrm{d}\pi(x) = \pi(A)$	
Stochastic	DTDS	✓ (Theorem 4.6)	✓ (Theorem 4.2)	
Stochastic	SDE	(Nüske et al., 2023): ✓	(Nüske et al., 2023): exp. stability ✓ (Theorem 4.2)	
Deterministic	DTDS ($\varepsilon = 0$)	✓ (Theorem 4.6)	✓ (Theorem 4.5)	Superlinear rate
Deterministic	SDE ($\sigma = 0$)	(Zhang & Zuazua, 2023): ✓	✓ (Theorem 4.5)	Superlinear rate

Table 1. Fundamental contributions of this work.

We motivate our findings by showing numerically approximated convergence rates while referring to Section 5 for a detailed description of the numerical experiments. To be slightly more precise, we consider a reversible stochastic system modeling the folding kinetics of a protein and a deterministic nonlinear partial differential equation given by the Kuramoto-Sivashinsky equation for chaotic flame propagation. In Figure 1, we depict the convergence rate of the learning error $\|K_{\mathbb{V}} - \widehat{K}_m\|_F$ of EDMD in terms of the amount of data m used for ergodic sampling **(S1)**. We clearly observe that the convergence rate of the error is linear for the molecular dynamics example (note that we depict the root mean square error $[\mathbb{E}[\|\widehat{C} - C\|_F^2]]^{1/2}$ with corresponding rate $m^{-1/2}$) and superlinear for the deterministic nonlinear partial differential equation (PDE).

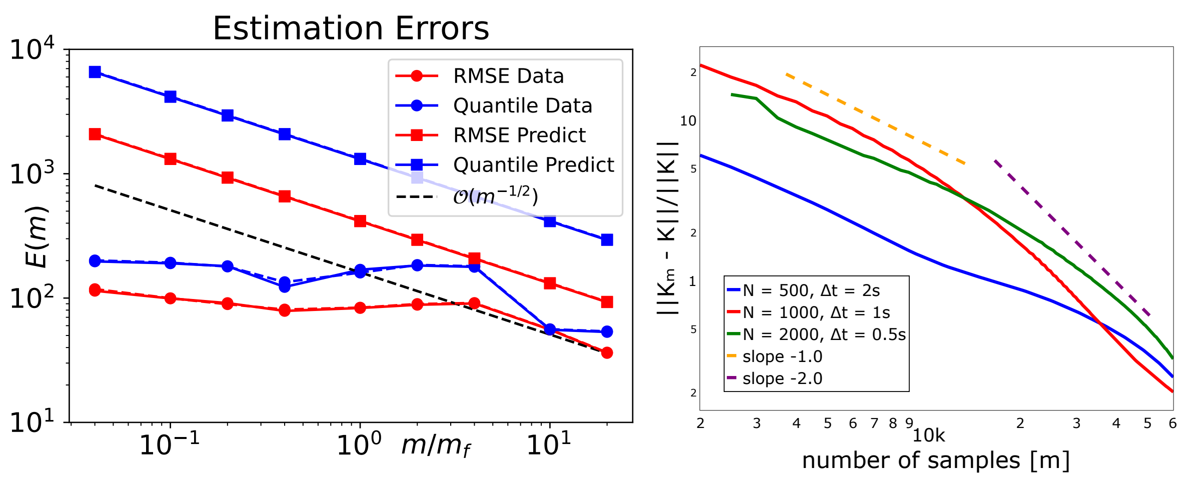

Figure 1. Convergence rate for ergodic sampling of a stochastic system for protein folding (left) and of a deterministic partial differential equation modeling chaotic flame propagation (right).

The main contribution are rigorously derived sharp error bounds for data-driven learning in the Koopman framework based on EDMD. As a byproduct, we obtain the first rigorous error analysis directly applicable to a broad class of highly-complex systems, which contains – among many others – the two considered challenging applications. Our key findings are, in addition, summarized in Table 3.

Case (S1): Ergodic sampling of stochastic systems, see Subsection 4.1. For this sampling strategy and the special case (SDE), error bounds with linear rate were already given in (Nüske et al., 2023). The proof relied on the assumption that the Koopman semigroup on $\mathbb{1}^\perp$ is exponentially stable. Our first major result Theorem 4.2 does not require this assumption and provides a linear rate for a wide class of stochastic systems, including the molecular dynamics application as well as discrete-time systems like (DTDS). We show that there is a constant $c \geq 0$ such that for all $\varepsilon > 0$,

$$\mathbb{P}\big(\|K_{\mathbb{V}} - \widehat{K}_m\|_F > \varepsilon\big) \;\leq\; \frac{c}{m\varepsilon^2}$$

holds. The corresponding convergence rate depicted in the upper plot of Figure 1 reveals that this linear rate is sharp.

Case (S1): Ergodic sampling of deterministic systems, e.g., (DTDS) with $\varepsilon_n \equiv 0$ or (SDE) with $\sigma \equiv 0$, see Subsection 4.2 for details. The main condition for obtaining the above error bound for stochastic systems is that $\lambda = 1$ is an isolated eigenvalue of the Koopman operator. Although this assumption is much more general than that of exponential stability of the Koopman semigroup on $\mathbb{1}^\perp$ imposed in (Nüske et al., 2023), it excludes a broad class of *deterministic cases*, cf. Kakutani & Petersen (1981). However, leveraging advanced tools from operator theory, we prove the first superlinear convergence rates for EDMD with ergodic sampling of deterministic systems (such as nonlinear PDEs) as a second major result in Theorem 4.5. That is, we prove that there are constants $c \geq 0$ and $\alpha \geq 1$ such that for all $\varepsilon > 0$

$$\mathbb{P}\big(\|K_{\mathbb{V}} - \widehat{K}_m\|_F > \varepsilon\big) \leq \frac{c}{m^\alpha \varepsilon^2}.$$

This rate is observed in the right plot of Figure 1 illustrating EDMD for the Kuramoto-Sivashinsky equation.

Case (S2): i.i.d. sampling of, e.g., (DTDS) or (SDE), see Subsection 4.3 for details. The last major result considers the case of i.i.d. sampling. Going beyond Nüske et al. (2023), and under very general assumptions, we provide an exponential convergence rate for EDMD with i.i.d. sampling using Hoeffding's inequality in Theorem 4.6. We show that there are constants $c_1, c_2 > 0$ such that for all $\varepsilon > 0$

$$\mathbb{P}\big(\|K_{\mathbb{V}} - \widehat{K}_m\|_F > \varepsilon\big) \leq c_1 \exp(-c_2 m \varepsilon^2).$$

In the upcoming Section 4, we provide the previse statements of the error bounds presented above. Moreover, in Section 5, we revisit the molecular dynamics and flame propagation examples by means of an extensive case study.

4. Sharp convergence rates for EDMD-based machine learning

Our strategy to prove the three estimates presented in Section 3 consists of three major steps. First, we provide a representation formula for the variances of the empirical estimators $\widehat{C}$ and $\widehat{C}_+$ in terms of the m sample points. Second, this representation is combined with concentration inequalities, i.e., Markov's and Hoeffding's inequality, to deduce a probabilistic bound on the errors $\hat{C} - C$ and $\widehat{C}_+ - C_+$. In a last step, we combine these to a obtain a bound on the learning error $K_{\mathbb{V}} - \widehat{K}_m$ in view of (3) and (4).

4.1. Case (S1): Ergodic sampling. Error bounds – almost without assumptions

In this subsection, we let $\mu = \pi$ and draw the data samples from long ergodic trajectories of the process, according to case **(S1)**.

A key step in our error analysis consists of deducing representations of the variances of $\widehat{C}$ and $\widehat{C}_+$ which are well-suited for further analysis. For the formulation of our next result, we note that always $K\mathbb{1} = \mathbb{1}$, hence, $K\,\mathrm{span}\{\mathbb{1}\} \subset \mathrm{span}\{\mathbb{1}\}$, and also $KL^2_0(\mu) \subset L^2_0(\mu)$, where $L^2_0(\mu) = \mathrm{span}\{\mathbb{1}\}^\perp$. In what follows, we set $K_0 := K|_{L^2_0(\mu)}$, which is a contractive linear operator from $L^2_0(\mu)$ into itself. Moreover, we define the constants

$$\mathbb{E}_+ := \langle K\varphi, \varphi\rangle - \|C_+\|_F^2 \qquad \text{and} \qquad \mathbb{E}_0 := \|\varphi\|^2 - \|C\|_F^2.$$

Theorem 4.1. (Variance representation) *Define the quantities*

$$\sigma^2_{m,+} := \mathbb{E}_+ + \sum_{i,j=1}^N \langle p_m(K_0)Qg_{ij}, Qg^*_{ji}\rangle \qquad \textit{and} \qquad \sigma^2_{m,0} := \mathbb{E}_0 + \sum_{i,j=1}^N \langle K_0 p_m(K_0)Q\psi_{ij}, Q\psi_{ij}\rangle,$$

where $\psi_{ij} = \psi_i\psi_j$, $g_{ij} = \psi_i \cdot K\psi_j$, $g_{ji}^* = \psi_j \cdot K^*\psi_i$, $Q = P_{L_0^2(\mu)}$*, and* p_m *is the polynomial*

$$p_m(z) = 2\sum_{k=1}^{m-1}(1 - \tfrac{k}{m})z^{k-1}.$$

Then the variances of $\widehat{C}_+$ *and* $\widehat{C}$ *admit the following representations:*

$$\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] = \frac{\sigma_{m,+}^2}{m} \qquad \textit{and} \qquad \mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] = \frac{\sigma_{m,0}^2}{m}.$$

Next, a thorough analysis of the expressions $\sigma_{m,+}^2$ and $\sigma_{m,0}^2$ in the variance representations leads to the following bounds:

$$\sigma_{m,+}^2 \le [1 + \|p_m(K_0)\|]\,\mathbb{E}_+ \qquad \text{and} \qquad \sigma_{m,0}^2 \le [1 + \|K_0 p_m(K_0)\|]\,\mathbb{E}_0.$$

So far, the results hold in full generality. However, if we further assume that $\lambda = 1$ is an isolated simple eigenvalue[2] of K, we may further estimate the above bounds independently of m:

$$\sigma_{m,+}^2 \le \big[1 + 4\|(I - K_0)^{-1}\|\big]\,\mathbb{E}_+ \qquad \text{and} \qquad \sigma_{m,0}^2 \le \big[1 + 4\|K_0(I - K_0)^{-1}\|\big]\,\mathbb{E}_0. \tag{6}$$

An application of Markov's inequality in combination with Lemma C.5 immediately yields the following theorem, which is the main result of this subsection.

Theorem 4.2. *Assume that* $\lambda = 1$ *is an isolated simple eigenvalue of* K*. For* $\varepsilon > 0$*, define the constant*

$$\alpha := \big[1 + 4\|(I - K_0)^{-1}\|\big] \cdot \big(2\|C^{-1}\|_F\|C_+\|_F + \varepsilon\big)^2 \cdot \big[\big(\|C_+\|_F^{-2} + \|C^{-1}\|_F^2\big)\|\varphi\|^2 - 2\big]\,.$$

Then we have

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \;\le\; \frac{\alpha}{m\varepsilon^2}.$$

In particular, if $\delta \in (0,1)$*, then for* $m \ge \frac{\alpha}{\delta\varepsilon^2}$ *ergodic samples, with probability at least* $1-\delta$ *we have that* $\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F \le \varepsilon$*.*

Remark 4.3. **(a)** If K is normal (e.g., self-adjoint or unitary), we have

$$\|(I - K_0)^{-1}\| = \frac{1}{\operatorname{dist}(1, \sigma(K_0))},$$

where $\sigma(K_0)$ denotes the spectrum of the operator K_0. If there exist eigenvalues of K close to $\lambda = 1$, the above distance is small (hence $\|(I - K_0)^{-1}\|$ is large) and there exist so-called *meta-stable* sets, which are almost invariant; that is, trajectories of X_n remain in these sets for a long time (Davies, 1982). In this case, lots of measurements m are needed to gather sufficient information on the process, which is reflected in Theorem 4.2.

(b) If there exist $M \ge 1$ and $\omega > 0$ such that $\|K_0^n\| \le Me^{-\omega n}$ (which is equivalent to K_0 having spectral radius smaller than one[3]), then, setting $q = e^{-\omega} < 1$, we have

$$\|p_m(K_0)\| = 2\left\|\sum_{k=1}^{m-1} \tfrac{m-k}{m}\cdot K_0^{k-1}\right\| \le 2M\sum_{k=1}^{m-1}\tfrac{m-k}{m}\cdot q^{k-1} = Mp_m(q) = \frac{2M}{1-q}\left(1 - \frac{1-q^m}{m(1-q)}\right) \;\le\; \frac{2M}{1-q}$$

and $\|K_0 p_m(K_0)\| \le \frac{2Mq}{1-q}$. Especially, if $M = 1$, we obtain $\sigma_{m,+}^2 \le \frac{3-q}{1-q}\cdot\mathbb{E}_+$ and $\sigma_{m,0}^2 \le \frac{1+q}{1-q}\cdot\mathbb{E}_0$. For example, if K is a non-negative self-adjoint operator with an isolated simple eigenvalue at $\lambda = 1$, we have $M = 1$ and $q = \max\sigma(K_0)$.

[2]This condition is equivalent to $\lambda \notin \sigma(K_0)$.

[3]This condition was assumed in Nüske et al. (2023) and in particular excludes the deterministic case.

4.2. Case (S1): Ergodic sampling of deterministic systems. Superlinear convergence

Let us consider the deterministic subcase of case **(S1)**, where K is a unitary composition operator with a bijective measure-preserving map $T : \mathcal{X} \to \mathcal{X}$, i.e., $Kf = f \circ T$. The key result is the next theorem, which shows that the variances of $\widehat{C}_+$ and $\widehat{C}$ exhibit a direct link to mean ergodicity.

Theorem 4.4. (Variance representation) *Let K be unitary. Then for the variances of $\widehat{C}_+$ and $\widehat{C}$, respectively, we have*

$$\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] = \sum_{i,j=1}^{N} \left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k Q g_{ij} \right\|^2 \qquad \textit{and} \qquad \mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] = \sum_{i,j=1}^{N} \left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k Q \psi_{ij} \right\|^2.$$

By the mean ergodic theorem (see, e.g., Krengel (1985)), we know that for every single $f \in L_0^2(\mu)$ we have that $T_m f := \frac{1}{m}\sum_{k=0}^{m-1} K_0^k f \to 0$ as $m \to \infty$ (in norm). However, this convergence can be arbitrarily slow and is, in addition, qualitatively bounded from above by $1/m$, which follows from Butzer & Westphal (1971) who proved that $\|T_m f\| = o(1/m)$ implies $f = 0$. Moreover, it will never be uniform[4] in the sense that $\|T_m\| \to 0$ as $m \to \infty$ (see, e.g., Kakutani & Petersen (1981)). We may therefore *not* assume that $\lambda = 1$ is an isolated simple eigenvalue of K, since otherwise $T_m = \frac{1}{m}(I - K_0)^{-1}(I - K_0^m) \to 0$ as $m \to \infty$, which is a contradiction.

We therefore have to impose assumptions on the functions that T_m is applied to. For a function $f \in L^2(\mu)$ we let

$$\mu_f(\Delta) := \|E(\Delta)f\|^2, \quad \Delta \in \mathfrak{B}(\mathbb{T}),$$

where E denotes the spectral measure of the unitary operator K, cf. Appendix D. Then μ_f is a finite measure on $\mathbb{T}$ describing the spectral distribution of f. Define the finite set of functions

$$\mathcal{F} := \{Q\psi_{ij} : i,j \in [1:N]\} \cup \{Qg_{ij} : i,j \in [1:N]\},$$

the arcs $S_\theta := \{e^{it} : -\theta \le t \le \theta\}$, $\theta \in (0,\pi)$, and the constant

$$M := \frac{8(1 + \|C^{-1}\|_F^2 \|C_+\|_F^2)^2}{\|C_+\|_F^2} \cdot \max\{\mathbb{E}_0, \mathbb{E}_+\}.$$

Theorem 4.5. *Assume that K is unitary and suppose that there exist $\alpha \in (1,2)$ and $\kappa \ge 0$ such that for some $\theta \in (0,\pi)$,*

$$\mu_f(S_\gamma) \le \kappa \cdot \mu_f(S_\theta) \cdot (\gamma/\pi)^\alpha \tag{7}$$

for all $f \in \mathcal{F}$ and all $\gamma \in (0,\theta]$. Then for $\varepsilon \in (0,2)$ we have

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le \frac{C(\alpha,\kappa,\theta)M}{m^\alpha \varepsilon^2},$$

where

$$C(\alpha,\kappa,\theta) := 2\max\left\{\frac{2}{1-\cos\theta}, \frac{3\kappa}{(\alpha-1)(2-\alpha)}\right\}.$$

If $\kappa = 0$, then

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le \frac{M}{(1-\cos\theta)\cdot m^2\varepsilon^2}.$$

Let us briefly discuss the condition (7). For this, assume that the spectrum of K_0 consists of eigenvalues $\lambda_n = e^{it_n}$ ($t_n \searrow 0$) with corresponding eigenfunctions φ_n. Then (7) is equivalent to $\sum_{n:t_n \le \gamma} |\langle f, \varphi_n\rangle|^2 \lesssim \gamma^\alpha$. As $\gamma \mapsto \gamma^\alpha$ is convex, this allows for relatively large gaps between the eigenvalues in relation to the coefficients $|\langle f, \varphi_n\rangle|^2$ (in fact, $|\langle f, \varphi_n\rangle|^2/(t_n - t_{n+1}) \le t_n^{\alpha-1}$ is sufficient), while a behavior like $\sum_{n:t_n<\gamma} |\langle f, \varphi_n\rangle|^2 \sim \gamma^{1/2}$ forces the eigenvalues to be dense at $\lambda = 1$. Hence, (7) relates the coefficient decay with the position of the eigenvalues.

[4] at least if μ is non-atomic

4.3. Case (S2): i.i.d. sampling. Bounds on the EDMD estimation error

In this subsection, we let $\mu = \nu$ and draw i.i.d. data samples (x_k, y_k) from $d\mu_0 := \rho(\cdot, dy)\, d\mu$, according to case **(S2)**. Our main result in this case is the following theorem.

Theorem 4.6. *Assume that $C_+ \neq 0$. Let $\varepsilon > 0$ and set $\sigma = 2\|C^{-1}\|_F \|C_+\|_F + \varepsilon$. Then*

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le \frac{\sigma^2}{m\varepsilon^2}\left[\big(L\|C_+\|_F^{-2} + \|C^{-1}\|_F^2\big)\|\varphi\|^2 - 2\right].$$

If, in addition, $\varphi \in L^\infty(\mu)$, then

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le 2\exp\left(-\frac{m\varepsilon^2\|C_+\|_F^2}{2\tau^2(1+L)^2}\right) + 2\exp\left(-\frac{m\varepsilon^2}{8\tau^2\|C^{-1}\|_F^2}\right),$$

where $\tau = \sigma\|\varphi\|_\infty$.

5. Numerical Examples

In this part, we illustrate the deduced error bounds by means of two highly complex examples. The first is a protein folding simulation of a 35-amino acid chain, and the second is a nonlinear chaotic partial differential equation modeling flame propagation.

5.1. Molecular Dynamics Simulation

We apply the Koopman approach to analyze the folding kinetics of the Fip35 WW-domain. This 35-amino acid protein has been used as a benchmarking system in many previous studies due to its small size and fast folding time scale of about ten micro-seconds. We use the 1.1-milli-second molecular dynamics (MD) simulation data set published by D.E. Shaw research in Lindorff-Larsen et al. (2011). The underlying dynamics are a modified version of Hamiltonian dynamics, defined by a many-body potential energy function V plus a thermostat, which ensures that the position space dynamics are effectively stochastic and reversibly sample the Boltzmann distribution $\exp(-\beta V(x))\,\mathrm{d}x$, enabling estimation of statistical averages via ergodic sampling. The complete data set comprises about $5.6 \cdot 10^6$ data points.

We build our Koopman model on a 528-dimensional space of inter-atomic distances (closest heavy-atom inter-residue distances), which is a standard featurization for molecular simulation data sets. We employ a basis set of random Fourier features (RFFs) (Rahimi & Recht, 2007), that is, complex plane waves of the forms

$$\psi_j(x) = \exp(i\omega_j^\top x),$$

where ω_j are frequencies drawn from the spectral measure associated to a Gaussian radial basis function kernel. These features were shown to provide a fairly automatic way of generating an expressive dictionary in Nüske & Klus (2023). The Gaussian bandwidth parameter σ, the number of random features p, and the lag time t for Koopman learning are tuned using the VAMP score metric (Wu & Noé, 2020), see once again Nüske & Klus (2023) for details. As shown in Figure 2, $\sigma = 30$, $p = 500$ and $t = 1\mu$s emerge as suitable parameters to accurately estimate the folding time scale. The leading eigenvectors of the Koopman model can be used to identify the folded and unfolded states from the simulation data, as illustrated in Figure 3. The two leading eigenvectors of the Koopman model are transformed into membership functions $\chi_{0,1}$ by the PCCA method (Deuflhard & Weber, 2005). A value close to one of these membership functions indicates that the system is in either the folded or unfolded state, as illustrated by representative structures on the left.

As the MD simulation is reversible, we can apply the error bounds in Eq. (6). Computing the exact asymptotic variances $\sigma_{m,0}^2$ and $\sigma_{m,+}^2$ via Theorem 4.1 and Eq. (6) requires access to the true eigenvalues and Galerkin matrices C, C_+. Since these are not available, we estimate them by learning a reference model on all available data points. We then compute the error relative to this reference model if the Galerkin matrices are estimated using fewer data points, with m ranging from $m = 2000$ to $m = 2 \cdot 10^6$.

In Figure 1 (left), we show the root mean square errors $[\mathbb{E}[\|\widehat{C} - C\|_F^2]]^{1/2}$ (red) and the 90% error quantile (blue), estimated by our theoretical bounds in Eq. (6) (squares) vs. data-based estimates relative to the reference model (circles) using 50 different sub-samples of the data set, each of size m. Throughout, solid lines refer to C, while dashed lines refer to C_+, but the results are almost indistinguishable. The black line indicates a qualitative decay of the form $cm^{-1/2}$, where the pre-factor c

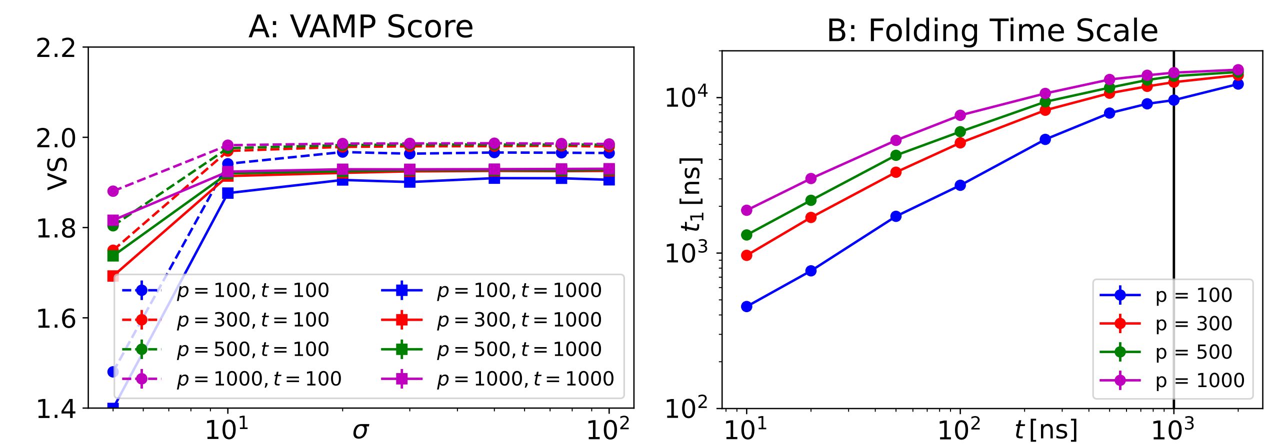

Figure 2. A: VAMP score metric as a function of kernel bandwidth σ for different feature sizes p and lag times t (in [ns]). Regardless of the lag time, $p = 500$ and $\sigma = 30$ emerge as optimal. B: Estimated folding timescale for $\sigma = 30$ and different feature sizes p as a function of lag time. It can be seen that convergence is achieved for $t = 1\,\mu$s and $p \geq 500$.

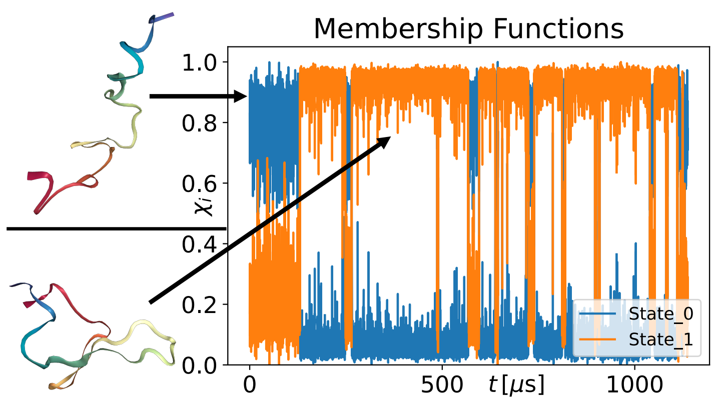

Figure 3. Identification of the folded state of Fip35 based on the Koopman model.

is the average ratio of the data-based RMSE over $m^{-1/2}$, for all values of m greater than the folding time scale. The values of m on the horizontal axis are normalized against the number of data points m_f required to reach the folding time scale $10\,\mu$s. Compared to the theoretical results, the actual error is about an order or magnitude smaller. We notice however, that after an initial period of about the length of the folding time scale, the asymptotic decay of the error is well-described by our estimates (see the black line in the left plot of Figure 1).

5.2. Nonlinear PDE

In our second example, we study the Kuramoto-Sivashinsky equation in two space dimensions, which is a widely-studied deterministic PDE modeling the dynamics of chaotic flame front propagation:

$$\partial_t x + \nabla^2 x + \nabla^4 x + |\nabla x|^2 = 0. \tag{8}$$

The system state $x : \Omega \times [0, \infty) \to \mathbb{R}$ (cf. Figure 4 left) depends on both space and time, and we consider a rectangular domain $\Omega = [-30\pi, 30\pi]^2$ with periodic boundary conditions.

The numerical simulation is realized using the open-source code *shenfun* (Mortensen, 2018). For the used domain size, the system exhibits chaotic dynamics, which means that we can use ergodic sampling from a single long trajectory that covers the entire chaotic attractor. In our experiments, we collect a total of $M = 100,000$ samples with a time step of $h = 0.01$. The spectral element discretization in space yields snapshots of dimension $128 \times 128 = 16,384$, which we reduce to $64 \times 64 = 4,096$ for computational reasons. As this is still prohibitively large for classical dictionaries, we rely

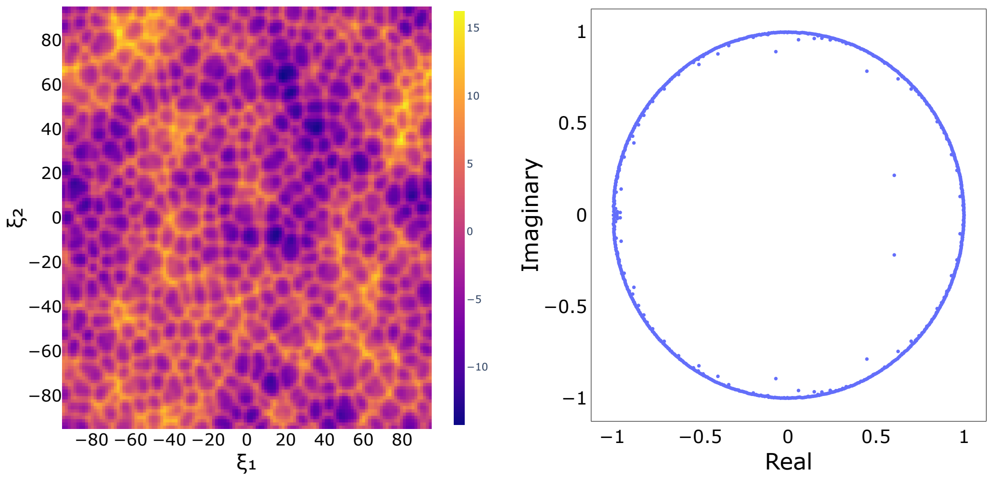

Figure 4. Left: Exemplary snapshot of a state $x(\xi, t)$ on the attractor governed by the Kuramoto-Sivashinsky equation (8). Right: Eigenvalues of the EDMD matrix $\widehat{K}_M$

on the kernel variant of EDMD (Williams et al., 2016). We choose a dictionary $\mathcal{D}$ consisting of $N = 1000$ radial basis functions with a twice differentiable Matérn kernel (of order $\nu = 2.5$), centered at N spatially equidistant grid points, which means that our Koopman approximation is of dimension $\widehat{K}_m \in \mathbb{R}^{N\times N}$. To estimate $\widehat{K}_m$ we choose a lag time of $\Delta t = 1$, and as we do not have access to the true Koopman compression $K_{\mathbb{V}}$, we study the behavior of the relative error between $\widehat{K}_m$ and $\widehat{K}_M$ for increasing m.

Figure 4 (right) shows the eigenvalue spectrum of the matrix $\widehat{K}_M$ for $N = 1000$ radial basis functions and all M samples. As usual for chaotic systems (cf. Rowley et al. (2009)) and proven in Subsection 4.2, the eigenvalues are distributed around the unit circle. Consequently, the eigenvalue $\lambda = 1$ is *not* isolated and thus, according to Theorem 4.5, the convergence rate with respect to m should range somewhere between one and two. When taking a closer look at Figure 1 (right), this is exactly the behavior that we observe.

6. Acknowlegdment

We thank *D.E. Shaw Research* for providing the Fip35 simulation data set.

7. Conclusion

We provided novel sharp error bounds on EDMD for both ergodic and i.i.d. sampling for discrete- and continuous-time nonlinear dynamical systems under rather mild assumptions. For i.i.d. sampling, we established convergence at an exponential rate. Contrary, the convergence of EDMD with ergodic sampling depends on whether the underlying system is stochastic or deterministic. While the error decays at a linear rate in the first case, we show that a broad class of deterministic systems exhibits superlinear convergence. Our theoretical results are underpinned by numerical simulations of highly-complex systems such as the folding kinetics of an acid protein in molecular dynamics and the chaotic nonlinear Kuramoto-Sivashinsky equation. In future research, we investigate suitable choices of the dictionary to speed up the convergence according to the newly identified conditions of Theorem 4.5 in the deterministic case.

References

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

Brunton, B. W., Johnson, L. A., Ojemann, J. G., and Kutz, J. N. Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of neuroscience methods*, 258:1–15, 2016.

Brunton, S. L., Budisic, M., Kaiser, E., and Kutz, J. N. Modern Koopman theory for dynamical systems. *SIAM Review*, 64 (2):229–340, 2022.

Butzer, P. and Westphal, U. The mean ergodic theorem and saturation. *Indiana University Mathematics Journal*, 20: 1163–1174, 1971.

Conway, J. B. *A Course in Functional Analysis*. Springer New York, NY, 2010.

Davies, E. Metastable states of symmetric Markov semigroups II. *Journal of the London Mathematical Society*, s2-26: 541–556, 12 1982.

Deuflhard, P. and Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005.

Dogra, A. S. and Redman, W. Optimizing neural networks via Koopman operator theory. *Advances in Neural Information Processing Systems*, 33:2087–2097, 2020.

Giannakis, D., Kolchinskaya, A., Krasnov, D., and Schumacher, J. Koopman analysis of the long-term evolution in a turbulent convection cell. *Journal of Fluid Mechanics*, 847:735–767, 2018.

Golany, T., Radinsky, K., Freedman, D., and Minha, S. 12-lead ecg reconstruction via Koopman operators. In *International Conference on Machine Learning*, pp. 3745–3754. PMLR, 2021.

Han, M., Euler-Rolle, J., and Katzschmann, R. K. Desko: Stability-assured robust control with a deep stochastic koopman operator. In *International Conference on Learning Representations*, 2021.

Kachurovskii, A. and Podvigin, I. Fejér sums for periodic measures and the von Neumann ergodic theorem. *Doklady Mathematics*, 98:344–347, 2018.

Kakutani, S. and Petersen, K. The speed of convergence in the ergodic theorem. *Monatshefte für Mathematik*, 91:11–18, 1981.

Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., and Noé, F. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28:985–1010, 6 2018.

Klus, S., Nüske, F., and Peitz, S. Koopman analysis of quantum systems. *Journal of Physics A: Mathematical and Theoretical*, 55(31):314002, 2022.

Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, 17(5):315, 1931.

Korda, M. and Mezić, I. On convergence of extended dynamic mode decomposition to the Koopman operator. *Journal of Nonlinear Science*, 28:687–710, 2018.

Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L., and Pontil, M. Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces. *Advances in Neural Information Processing Systems*, 35: 4017–4031, 2022.

Kostic, V. R., Lounici, K., Novelli, P., and Pontil, M. Sharp spectral rates for Koopman operator learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Krengel, U. *Ergodic Theorems*. De Gruyter, Berlin, New York, 1985.

Kutz, J. N., Brunton, S. L., Brunton, B. W., and Proctor, J. L. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

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

Mauroy, A., Susuki, Y., and Mezić, I. *Koopman operator in systems and control*. Springer, 2020.

Mezić, I. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual review of fluid mechanics*, 45: 357–378, 2013.

Mollenhauer, M. *On the statistical approximation of conditional expectation operators*. PhD thesis, Freie Universität Berlin, 2021.

Mortensen, M. Shenfun: High performance spectral Galerkin computing platform. *Journal of Open Source Software*, 3(31): 1071, 2018.

Nüske, F. and Klus, S. Efficient approximation of molecular kinetics using random Fourier features. *The Journal of Chemical Physics*, 159:074105, 2023.

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

Philipp, F., Schaller, M., Worthmann, K., Peitz, S., and Nüske, F. Error bounds for kernel-based approximations of the Koopman operator. *ArXiv preprint, arXiv:2301.08637*, 2023a.

Philipp, F., Schaller, M., Worthmann, K., Peitz, S., and Nüske, F. Error analysis of kernel EDMD for prediction and control in the Koopman framework. *Arxiv preprint, arXiv:2312.10460*, 2023b.

Pinelis, I. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Retchin, M., Amos, B., Brunton, S., and Song, S. Koopman constrained policy optimization: A Koopman operator theoretic method for differentiable optimal control in robotics. In *ICML 2023 Workshop on Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, 2023.

Rowley, C., Mezić, I., Bagheri, S., Schlatter, P., and Henningson, D. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

Rozwood, P., Mehrez, E., Paehler, L., Sun, W., and Brunton, S. Koopman-assisted reinforcement learning. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Schütte, C., Koltai, P., and Klus, S. On the numerical approximation of the Perron-Frobenius and Koopman operator. *Journal of Computational Dynamics*, 3:1–12, 9 2016.

Weissenbacher, M., Sinha, S., Garg, A., and Yoshinobu, K. Koopman Q-learning: Offline reinforcement learning via symmetries of dynamics. In *International Conference on Machine Learning*, pp. 23645–23667. PMLR, 2022.

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

Williams, M. O., Rowley, C. W., and Kevrekidis, I. G. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265, 2016.

Wu, H. and Noé, F. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science*, 30:33–66, 2020.

Wu, H., Nüske, F., Paul, F., Klus, S., Koltai, P., and Noé, F. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *The Journal of chemical physics*, 146(15), 2017.

Zhang, C. and Zuazua, E. A quantitative analysis of Koopman operator methods for system identification and predictions. *Comptes Rendus Mécanique*, 351(S1):1–31, 2023.

A. Proofs of the main results

In this section, we provide proofs for our main theorems. For the convenience of the reader, we repeat the particular results.

A.1. Proof of Theorem 4.2 in Subsection 4.1

In this subsection, we let $\mu = \pi$ be an ergodic and invariant measure for the process X_n, i.e., case **(S1)**. Furthermore, we assume that for each $(N-1)$-dimensional subspace $M \subset \mathbb{R}^N$ and $x \in \Psi^{-1}(M)$ we have $\rho(x, \Psi^{-1}(M)) = 0$ in order to guarantee the a.s. invertibility of $\widehat{C}$, see Lemma C.4.

Theorem 4.2. *Assume that $\lambda = 1$ is an isolated simple eigenvalue of K. For $\varepsilon > 0$, define the constant*

$$\alpha := \big[1 + 4\|(I - K_0)^{-1}\|\big] \cdot \big[2\|C^{-1}\|_F \|C_+\|_F + \varepsilon\big]^2 \cdot \big[\big(\|C_+\|_F^{-2} + \|C^{-1}\|_F^2\big)\|\varphi\|^2 - 2\big].$$

Then we have

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le \frac{\alpha}{m\varepsilon^2}.$$

In particular, if $\delta \in (0,1)$, then for $m \ge \frac{\alpha}{\delta\varepsilon^2}$ ergodic samples, with probability at least $1-\delta$ we have that $\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F \le \varepsilon$.

Proof. In a first step, we provide exact formulae for the variances of $\widehat{C}$ and $\widehat{C}_+$ in Theorem 4.1. Next, we find bounds on these variances in Proposition A.1 and apply Markov's inequality to deduce a probabilistic bound on the errors $\|C_+ - \widehat{C}_+\|_F^2$ and $\|C - \widehat{C}\|_F^2$. Then Lemma C.5 immediately yields the desired bound for $\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F^2$. □

For completeness, we repeat Theorem 4.1 here. To this end, recall the definitions of $\mathbb{E}_+$ and $\mathbb{E}_0$:

$$\mathbb{E}_+ := \langle K\varphi, \varphi\rangle - \|C_+\|_F^2 \qquad \text{and} \qquad \mathbb{E}_0 := \|\varphi\|^2 - \|C\|_F^2.$$

Theorem 4.1. *Define the quantities*

$$\sigma_{m,+}^2 := \mathbb{E}_+ + \sum_{i,j=1}^N \langle p_m(K_0)Qg_{ij}, Qg_{ji}^*\rangle \qquad \textit{and} \qquad \sigma_{m,0}^2 := \mathbb{E}_0 + \sum_{i,j=1}^N \langle K_0 p_m(K_0)Q\psi_{ij}, Q\psi_{ij}\rangle,$$

where $\psi_{ij} = \psi_i\psi_j$, $g_{ij} = \psi_i \cdot K\psi_j$, $g_{ji}^* = \psi_j \cdot K^*\psi_i$ $(i,j = 1,\dots,N)$, $Q = P_{L_0^2(\mu)}$, and p_m is the polynomial

$$p_m(z) = 2\sum_{k=1}^{m-1}(1 - \tfrac{k}{m})z^{k-1}.$$

Then the variances of $\widehat{C}_+$ and $\widehat{C}$ admit the following representations:

$$\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] = \frac{\sigma_{m,+}^2}{m}, \tag{9}$$

$$\mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] = \frac{\sigma_{m,0}^2}{m}. \tag{10}$$

Proof. With $\Phi_k := \Psi(x_k)\Psi(x_{k+1})^\top$, $k = 0,\dots,m-1$, as in the proof of Proposition A.4, we have

$$\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] = \frac{1}{m}\big[\langle K\varphi,\varphi\rangle - \|C_+\|_F^2\big] + \frac{2}{m}\sum_{k=1}^{m-1}(1 - \tfrac{k}{m})\mathbb{E}\big[\big\langle C_+ - \Phi_k, C_+ - \Phi_0\big\rangle_F\big].$$

In this case, the cross terms in the second summand do not vanish. First,

$$\mathbb{E}\big[\big\langle C_+ - \Phi_k, C_+ - \Phi_0\big\rangle_F\big] + \|C_+\|_F^2 = \mathbb{E}\big[\big\langle \Phi_k, \Phi_0\big\rangle_F\big]$$

and next,

$$\begin{aligned}
\mathbb{E}\big[\big\langle\Phi_k,\Phi_0\big\rangle_F\big] &= \sum_{i,j=1}^N \mathbb{E}\big[\psi_i(x_k)\psi_j(x_{k+1})\psi_i(x_0)\psi_j(x_1)\big] \\
&= \sum_{i,j=1}^N \int\!\!\int\!\!\int\!\!\int \psi_i(x')\psi_j(y')\psi_i(x)\psi_j(y)\,\rho(x',dy')\,\rho_{k-1}(y,dx')\,\rho(x,dy)\,d\mu(x) \\
&= \sum_{i,j=1}^N \int\!\!\int\!\!\int \psi_i(x')\psi_i(x)\psi_j(y)(K\psi_j)(x')\,\rho_{k-1}(y,dx')\,\rho(x,dy)\,d\mu(x) \\
&= \sum_{i,j=1}^N \int\!\!\int \psi_i(x)\psi_j(y)\big(K^{k-1}g_{ij}\big)(y)\,\rho(x,dy)\,d\mu(x) \\
&= \sum_{i,j=1}^N \int \psi_i(x)\big(K_1[\psi_j\cdot K^{k-1}g_{ij}]\big)(x)\,d\mu(x) \\
&\overset{(*)}{=} \sum_{i,j=1}^N \big\langle\psi_j\cdot K^*\psi_i, K^{k-1}g_{ij}\big\rangle = \sum_{i,j=1}^N \big\langle g_{ji}^*, K^{k-1}g_{ij}\big\rangle.
\end{aligned}$$

For a justification of $(*)$ see Lemma B.2 in the Appendix.

Let $P := P_{\mathbb{1}} = \langle\,\cdot\,,\mathbb{1}\rangle\mathbb{1}$. Then, since $\int g_{ji}^*\,d\mu = \int g_{ij}\,d\mu$,

$$\sum_{i,j=1}^N \big\langle Pg_{ji}^*, K^{k-1}Pg_{ij}\big\rangle = \sum_{i,j=1}^N \big\langle\langle g_{ji}^*,\mathbb{1}\rangle\mathbb{1},\langle g_{ij},\mathbb{1}\rangle\mathbb{1}\big\rangle = \sum_{i,j=1}^N \left|\int g_{ij}\,d\mu\right|^2 = \sum_{i,j=1}^N |\langle\psi_i,K\psi_j\rangle|^2 = \|C_+\|_F^2.$$

Similarly, we get

$$\sum_{i,j=1}^N \|Pg_{ji}^*\|^2 = \sum_{i,j=1}^N \|Pg_{ij}\|^2 = \|C_+\|_F^2. \tag{11}$$

Therefore,

$$\mathbb{E}\big[\big\langle C_+-\Phi_k, C_+-\Phi_0\big\rangle_F\big] = \sum_{i,j=1}^N \big\langle K_0^{k-1}Qg_{ij}, Qg_{ji}^*\big\rangle,$$

and (9) is proved.

The proof of (10) is similar. Here, we set $\widetilde{\Phi}_k = \Psi(x_k)\Psi(x_k)^\top$ and observe that

$$\mathbb{E}\big[\|C-\widehat{C}\|_F^2\big] = \frac{1}{m}\big[\mathbb{E}[\|\widetilde{\Phi}_0\|_F^2] - \|C_+\|_F^2\big] + \frac{2}{m}\sum_{k=1}^{m-1}(1-\tfrac{k}{m})\big(\mathbb{E}\big[\big\langle\widetilde{\Phi}_k,\widetilde{\Phi}_0\big\rangle_F\big] - \|C\|_F^2\big).$$

Next,

$$\begin{aligned}
\mathbb{E}\big[\big\langle\widetilde{\Phi}_k,\widetilde{\Phi}_0\big\rangle_F\big] &= \sum_{i,j=1}^N \mathbb{E}\big[\psi_i(x_k)\psi_j(x_k)\psi_i(x_0)\psi_j(x_0)\big] = \sum_{i,j=1}^N \mathbb{E}\big[\psi_{ij}(x_k)\psi_{ij}(x_0)\big] \\
&= \sum_{i,j=1}^N \int\!\!\int \psi_{ij}(x)\psi_{ij}(y)\,\rho_k(x,dy)\,d\mu(x) \\
&= \sum_{i,j=1}^N \int \psi_{ij}(x)(K^k\psi_{ij})(x)\,d\mu(x) = \sum_{i,j=1}^N \big\langle K^k\psi_{ij},\psi_{ij}\big\rangle.
\end{aligned}$$

In particular, $\mathbb{E}\big[\|\widetilde{\Phi}_0\|_F^2\big] = \sum_{i,j=1}^N \|\psi_{ij}\|^2 = \|\varphi\|^2$. Finally, making use of

$$\sum_{i,j=1}^{N} \left\langle K^k P\psi_{ij}, P\psi_{ij}\right\rangle = \sum_{i,j=1}^{N} \|P\psi_{ij}\|^2 = \|C\|_F^2, \tag{12}$$

we obtain (10). □

Proposition A.1. *We have*

$$\begin{aligned}\sigma_{m,+}^2 &\le [1 + \|p_m(K_0)\|]\,\mathbb{E}_+\\ \sigma_{m,0}^2 &\le [1 + \|K_0 p_m(K_0)\|]\,\mathbb{E}_0.\end{aligned}$$

If, in addition, $\lambda = 1$ *is an isolated simple eigenvalue of* K*, then*

$$\begin{aligned}\sigma_{m,+}^2 &\le \big[1 + 4\|(I-K_0)^{-1}\|\big]\,\mathbb{E}_+,\\ \sigma_{m,0}^2 &\le \big[1 + 4\|K_0(I-K_0)^{-1}\|\big]\,\mathbb{E}_0.\end{aligned}$$

Proof. Let again $P := P_{\mathbb{1}} = \langle\,\cdot\,,\mathbb{1}\rangle\mathbb{1}$. Then

$$\begin{aligned}\sum_{i,j=1}^{N} \langle p_m(K_0)Qg_{ij}, Qg_{ji}^*\rangle &\le \left(\sum_{i,j=1}^{N} \|p_m(K_0)\|^2\|Qg_{ij}\|^2\right)^{1/2}\left(\sum_{i,j=1}^{N}\|Qg_{ji}^*\|^2\right)^{1/2}\\ &= \|p_m(K_0)\|\left(\sum_{i,j=1}^{N}\big[\|g_{ij}\|^2 - \|Pg_{ij}\|^2\big]\right)^{1/2}\left(\sum_{i,j=1}^{N}\big[\|g_{ji}^*\|^2 - \|Pg_{ji}^*\|^2\big]\right)^{1/2}.\end{aligned}$$

Now, as $(Kf)^2 \le Kf^2$ for $f \in L^4(\mu)$,

$$\sum_{i,j=1}^{N}\|g_{ij}\|^2 = \sum_{i,j=1}^{N}\int \psi_i^2(K\psi_j)^2\,d\mu \;\le\; \sum_{i,j=1}^{N}\int \psi_i^2\cdot K\psi_j^2\,d\mu = \langle K\varphi,\varphi\rangle.$$

Since we also have $(K^*f)^2 \le K^*f^2$ for $f \in L^4(\mu)$ (cf. Appendix B), we similarly conclude $\sum_{i,j=1}^N \|g_{ji}^*\|^2 \le \langle K\varphi,\varphi\rangle$ and thus (cf. (11))

$$\sum_{i,j=1}^{N}\langle p_m(K_0)Qg_{ij}, Qg_{ji}^*\rangle \;\le\; \|p_m(K_0)\|\big(\langle K\varphi,\varphi\rangle - \|C_+\|_F^2\big) = \|p_m(K_0)\|\cdot\mathbb{E}_+.$$

Similarly (cf. (12)),

$$\sum_{i,j=1}^{N}\langle K_0p_m(K_0)Q\psi_{ij}, Q\psi_{ij}\rangle \le \|K_0p_m(K_0)\|\sum_{i,j=1}^{N}\big[\|\psi_{ij}\|^2 - \|P\psi_{ij}\|^2\big] = \|K_0p_m(K_0)\|\cdot\mathbb{E}_0.$$

Assume now that $\lambda = 1$ is an isolated simple eigenvalue of K. Since

$$p_m(z) = \frac{2}{1-z}\left(1 - \frac{1}{m}\sum_{k=0}^{m-1} z^k\right), \qquad z \in \mathbb{C}\backslash\{0\},$$

making use of $\|K_0\| \le 1$, we observe that

$$\|p_m(K_0)\| = \left\|2(I-K_0)^{-1}\Big(I - \tfrac{1}{m}\sum_{k=0}^{m-1}K_0^k\Big)\right\| \;\le\; 4\|(I-K_0)^{-1}\|.$$

Similarly, $\|K_0p_m(K_0)\| \le 4\|K_0(I-K_0)^{-1}\|$. □

A.2. Proof of Theorem 4.5 in Subsection 4.2

Let us consider the deterministic subcase of ergodic sampling ,i.e., case **(S1)**, where K is a composition operator with a measure preserving map $T : \mathcal{X} \to \mathcal{X}$. Note that, in this case, we have $K_1(fg) = Kf \cdot Kg$ for $f, g \in L^2(\mu)$. Let us further assume that T is bijective, so that the corresponding Koopman operator K is unitary. Hence, there exists a bounded self-adjoint operator A on $L^2(\mu)$ with $\sigma(A) \subset [-\pi, \pi]$ such that $K = e^{iA}$. We set $A_0 := A|_{L^2_0(\mu)}$.

The following lemma is a slightly extended version of Lemma 4.4.

Lemma 4.4 (Extended version). *For the variances of $\widehat{C}_+$ and $\widehat{C}$, respectively, we have*

$$\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] = \frac{1}{m}\sum_{i,j=1}^{N}\big\langle F_m(A_0)Qg_{ij}, Qg_{ij}\big\rangle = \sum_{i,j=1}^{N}\bigg\|\frac{1}{m}\sum_{k=0}^{m-1}K_0^kQg_{ij}\bigg\|^2 \tag{13}$$

and

$$\mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] = \frac{1}{m}\sum_{i,j=1}^{N}\big\langle F_m(A_0)Q\psi_{ij}, Q\psi_{ij}\big\rangle = \sum_{i,j=1}^{N}\bigg\|\frac{1}{m}\sum_{k=0}^{m-1}K_0^kQ\psi_{ij}\bigg\|^2, \tag{14}$$

where F_m denotes the well known Fejér kernel

$$F_m(t) = \sum_{|k|\le m-1}\big(1 - \tfrac{|k|}{m}\big)e^{ikt} = \frac{1}{m}\left(\frac{1-\cos(mt)}{1-\cos t}\right)^2.$$

Proof. We set $q_m(z) := zp_m(z)$ and observe that

$$\operatorname{Re} q_m(e^{it}) = \sum_{k=1}^{m-1}\big(1 - \tfrac{k}{m}\big)e^{ikt} + \sum_{k=1}^{m-1}\big(1 - \tfrac{k}{m}\big)e^{-ikt} = -1 + \sum_{|k|\le m-1}\big(1 - \tfrac{|k|}{m}\big)e^{ikt} = F_m(t) - 1.$$

This proves the first equality in (14) as $\sum_{i,j=1}^N \|Q\psi_{ij}\|^2 = \mathbb{E}_0$. Let E denote the spectral measure of the unitary operator K_0. For $\Delta \in \mathfrak{B}(\mathbb{T})$ we compute

$$\begin{aligned}\langle E(\Delta)Qg_{ij}, Qg_{ji}^*\rangle &= \langle E(\Delta)Qg_{ij}, \psi_j \cdot K^*\psi_i\rangle = \langle \psi_j \cdot E(\Delta)Qg_{ij}, K^*\psi_i\rangle = \big\langle K(\psi_j \cdot E(\Delta)Qg_{ij}), \psi_i\big\rangle \\ &= \langle K\psi_j \cdot KE(\Delta)Qg_{ij}, \psi_i\rangle = \langle KE(\Delta)Qg_{ij}, \psi_i \cdot K\psi_j\rangle = \langle KE(\Delta)Qg_{ij}, Qg_{ij}\rangle,\end{aligned}$$

and hence

$$\begin{aligned}\operatorname{Re}\langle p_m(K_0)Qg_{ij}, Qg_{ji}^*\rangle &= \operatorname{Re}\int p_m(\lambda)\, d\langle E_\lambda Qg_{ij}, Qg_{ji}^*\rangle = \operatorname{Re}\int p_m(\lambda)\, d\langle K_0E_\lambda Qg_{ij}, Qg_{ij}\rangle \\ &= \operatorname{Re}\int \lambda p_m(\lambda)\, d\langle E_\lambda Qg_{ij}, Qg_{ij}\rangle = \big\langle \operatorname{Re} q_m(K_0)Qg_{ij}, Qg_{ij}\big\rangle \\ &= \|Qg_{ij}\|^2 - \langle F_m(A_0)Qg_{ij}, Qg_{ij}\rangle,\end{aligned}$$

which yields the first equality in (13). Now, for $f \in L_0^2(\mu)$ we compute

$$\bigg\|\frac{1}{m}\sum_{k=0}^{m-1}K_0^kf\bigg\|^2 = \frac{1}{m^2}\sum_{k,\ell=0}^{m-1}\big\langle K_0^{k-\ell}f, f\big\rangle = \frac{1}{m}\sum_{k=-(m-1)}^{m-1}\frac{m-|k|}{m}\big\langle K_0^kf, f\big\rangle = \langle F_m(A_0)f, f\rangle.$$

The lemma is proved. □

Remark A.2. The representation $\big\|\frac{1}{m}\sum_{k=0}^{m-1}K_0^kf\big\|^2 = \langle F_m(A_0)f, f\rangle$ of the square norm of the ergodic series by means of the Fejér kernel is well known, see (Kachurovskii & Podvigin, 2018).

The main result of this section is the following extended version of Theorem 4.5 which is valid for exponents $\alpha \in (0,2)$. For its formulation, we define

$$C(\alpha) = \begin{cases} \frac{4-3\alpha}{1-\alpha} & \text{for } \alpha \in (0,1) \\ 3 & \text{for } \alpha = 1 \\ \frac{3}{(\alpha-1)(2-\alpha)} & \text{for } \alpha \in (1,2). \end{cases}$$

and

$$C(\alpha,\kappa,\theta) := 2\max\left\{\frac{2}{1-\cos\theta}, \kappa C(\alpha)\right\}.$$

Moreover, recall the constant

$$M := \frac{8(1+\|C^{-1}\|_F^2\|C_+\|_F^2)^2}{\|C_+\|_F^2}\cdot\max\{\mathbb{E}_0,\mathbb{E}_+\}.$$

Theorem 4.5. (Extended version) *Assume that K is unitary and suppose that there exist $\alpha \in (0,2)$, $\theta \in (0,\pi)$, and $\kappa \geq 0$ such that*

$$\mu_f(S_\gamma) \leq \kappa \cdot \mu_f(S_\theta) \cdot (\gamma/\pi)^\alpha \tag{15}$$

for all $f \in \mathcal{F}$ and all $\gamma \in (0,\theta]$. Then for $\varepsilon \in (0,2)$ we have

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \;\leq\; \frac{C(\alpha,\kappa,\theta)M}{m^\alpha\varepsilon^2}.$$

If $\kappa = 0$, then

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \;\leq\; \frac{M}{(1-\cos\theta)\cdot m^2\varepsilon^2}.$$

Proof. Theorem 4.5 is proved by combining Lemma C.5 with a result on probabilistic bounds on the errors $\|C_+ - \widehat{C}_+\|_F$ and $\|C - \widehat{C}\|_F$. These bounds are provided by the Proposition A.3 below. □

We set $C_0 = C$ and $\widehat{C}_0 = \widehat{C}$.

Proposition A.3. *Let the conditions of the extended version of Theorem 4.5 be satisfied. Then for $\varepsilon > 0$ we have*

$$\mathbb{P}\big(\|C_i - \widehat{C}_i\|_F > \varepsilon\big) \;\leq\; \frac{C(\alpha,\kappa,\theta)\cdot\mathbb{E}_i}{m^\alpha\varepsilon^2}, \qquad i \in \{+,0\}. \tag{16}$$

If $\mu_f(S_\theta) = 0$ for some $\theta \in (0,\pi)$ and all $f \in \mathcal{F}$, then

$$\mathbb{P}\big(\|C_i - \widehat{C}_i\|_F > \varepsilon\big) \;\leq\; \frac{2\mathbb{E}_i}{(1-\cos\theta)m^2\varepsilon^2}, \qquad i \in \{+,0\}. \tag{17}$$

Proof. We prove the statements for $i = 0$. Similar reasonings apply to the case $i = +$, respectively.

To show the second statement, assume that there exists $\theta \in (0,\pi)$ such that $\mu_f(S_\theta) = 0$ for all $f \in \mathcal{F}$, and let $r(z) := \sum_{k=0}^{m-1} z^k$. Then

$$\begin{aligned}\mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] &= \sum_{i,j=1}^N \left\|\frac{1}{m}\sum_{k=0}^{m-1} K_0^k Q\psi_{ij}\right\|^2 = \frac{1}{m^2}\sum_{i,j=1}^N \|r(K_0)Q\psi_{ij}\|^2 = \frac{1}{m^2}\sum_{i,j=1}^N \int_{\mathbb{T}\setminus S_\theta} |r(z)|^2\,d\mu_{Q\psi_{ij}}(z) \\ &= \frac{1}{m^2}\sum_{i,j=1}^N \int_{\mathbb{T}\setminus S_\theta}\left|\frac{1-z^m}{1-z}\right|^2 d\mu_{Q\psi_{ij}}(z) \;\leq\; \frac{1}{m^2}\sum_{i,j=1}^N \frac{4}{|1-e^{i\theta}|^2}\|Q\psi_{ij}\|^2 = \frac{2\mathbb{E}_0}{(1-\cos\theta)m^2}.\end{aligned}$$

Hence, (17) follows from Markov's inequality.

Let us now assume that (15) holds for all $f \in \mathcal{F}$ and all $\gamma \in (0, \theta]$. Let $f = Q\psi_{ij}$ for some fixed pair $i, j \in [1 : N]$. Then $f = f_1 + f_2$, where $f_1 = E(\mathbb{T} \backslash S_\theta) f$ and $f_2 = E(S_\theta) f$. Since $\mu_{f_1}(S_\theta) = 0$, we obtain as above that

$$\left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k f_1 \right\|^2 \le \frac{2\|f_1\|^2}{(1-\cos\theta)m^2}.$$

For f_2 and all $\gamma \in (0, \pi]$ we infer from (15) that

$$\begin{aligned}\mu_{f_2}(S_\gamma) &= \mu_f(S_\gamma \cap S_\theta) = \mu_f(S_{\min\{\gamma,\theta\}}) \le \kappa \mu_f(S_\theta) \min\{\gamma, \theta\}^\alpha \pi^{-\alpha} \\ &\le \kappa \mu_f(S_\theta)(\gamma/\pi)^\alpha = \kappa \|E(S_\theta) f\|^2 (\gamma/\pi)^\alpha = \kappa \|f_2\|^2 (\gamma/\pi)^\alpha.\end{aligned}$$

Hence, Theorem 2 in (Kachurovskii & Podvigin, 2018) implies

$$\left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k f_2 \right\|^2 \le \frac{\kappa \|f_2\|^2 C(\alpha)}{m^\alpha}.$$

We conclude

$$\begin{aligned}\left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k f \right\|^2 &\le 2\left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k f_1 \right\|^2 + 2\left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k f_2 \right\|^2 \le \frac{4\|f_1\|^2}{(1-\cos\theta)m^2} + \frac{2\kappa \|f_2\|^2 C(\alpha)}{m^\alpha} \\ &\le 2\max\left\{ \frac{2}{1-\cos\theta}, \kappa C(\alpha) \right\} \frac{\|f\|^2}{m^\alpha} = C(\alpha, \kappa, \theta) \frac{\|f\|^2}{m^\alpha},\end{aligned}$$

and therefore

$$\mathbb{E}\big[\|C - \widehat{C}\|_F^2\big] = \sum_{i,j=1}^N \left\| \frac{1}{m} \sum_{k=0}^{m-1} K_0^k Q\psi_{ij} \right\|^2 \le C(\alpha, \kappa, \theta) \sum_{i,j=1}^N \frac{\|Q\psi_{ij}\|^2}{m^\alpha} = \frac{C(\alpha, \kappa, \theta) \cdot \mathbb{E}_0}{m^\alpha}.$$

The proposition is proved. □

A.3. Proof of Theorem 4.6 in Subsection 4.3

In this subsection, we let $\mu = \nu$ satisfying (1), i.e., we consider the case **(S2)** of i.i.d. sampling and assume that the observables $\psi_1, \ldots, \psi_N$ are strongly μ-linearly independent, see Definition C.1. Let us recall the main result on the learning error in case **(S2)** as stated in Theorem 4.6.

Theorem 4.6. *Assume that $C_+ \neq 0$. Let $\varepsilon > 0$ and set $\sigma = 2\|C^{-1}\|_F \|C_+\|_F + \varepsilon$. Then*

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le \frac{\sigma^2}{m\varepsilon^2} \left[\big(L\|C_+\|_F^{-2} + \|C^{-1}\|_F^2\big)\|\varphi\|^2 - 2 \right].$$

If, in addition, $\varphi \in L^\infty(\mu)$, then

$$\mathbb{P}\big(\|C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+\|_F > \varepsilon\big) \le 2\exp\left(-\frac{m\varepsilon^2 \|C_+\|_F^2}{2\tau^2(1+L)^2} \right) + 2\exp\left(-\frac{m\varepsilon^2}{8\tau^2 \|C^{-1}\|_F^2} \right),$$

where $\tau = \sigma\|\varphi\|_\infty$.

Proof. Theorem 4.6 is an immediate consequence of the following Proposition A.4 below, which deduces a probabilistic error bound on $C_+ - \widehat{C}_+$ and $C - \widehat{C}$, and Lemma C.5 which allows to straightforwardly infer the claimed bound on $C^{-1}C_+ - \widehat{C}^{-1}\widehat{C}_+$. □

Proposition A.4. *The following probabilistic bounds on the estimation errors hold:*

$$\mathbb{P}\big(\|C_+ - \widehat{C}_+\|_F > \varepsilon\big) \le \frac{\langle K\varphi, \varphi\rangle - \|C_+\|_F^2}{m\varepsilon^2} \quad \text{and} \quad \mathbb{P}\big(\|C - \widehat{C}\|_F > \varepsilon\big) \le \frac{\|\varphi\|^2 - \|C\|_F^2}{m\varepsilon^2}. \tag{18}$$

If $\varphi \in L^\infty(\mu)$, then

$$\mathbb{P}\big(\|C_+ - \widehat{C}_+\|_F > \varepsilon\big) \le 2\, e^{-\frac{m\varepsilon^2}{2(1+L)^2\|\varphi\|_\infty^2}} \quad \text{and} \quad \mathbb{P}\big(\|C - \widehat{C}\|_F > \varepsilon\big) \le 2\, e^{-\frac{m\varepsilon^2}{8\|\varphi\|_\infty^2}}. \tag{19}$$

Proof. We set $\Phi_k := \Psi(x_k)\Psi(y_k)^\top$. Then we have

$$\begin{aligned}\mathbb{E}\big[\|C_+ - \widehat{C}_+\|_F^2\big] &= \mathbb{E}\bigg[\bigg\|\frac{1}{m}\sum_{k=0}^{m-1}\big[C_+ - \Phi_k\big]\bigg\|_F^2\bigg] = \frac{1}{m^2}\mathbb{E}\bigg[\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1}\big\langle C_+ - \Phi_k, C_+ - \Phi_\ell\big\rangle_F\bigg]\\ &= \frac{1}{m^2}\mathbb{E}\bigg[\sum_{k=0}^{m-1}\big\|C_+ - \Phi_k\big\|_F^2 + 2\sum_{k=1}^{m-1}\sum_{\ell=0}^{k-1}\big\langle C_+ - \Phi_k, C_+ - \Phi_\ell\big\rangle_F\bigg]\\ &= \frac{1}{m^2}\sum_{k=0}^{m-1}\mathbb{E}\big[\|C_+ - \Phi_k\|_F^2\big] + \frac{2}{m^2}\sum_{k=1}^{m-1}(m-k)\mathbb{E}\big[\big\langle C_+ - \Phi_k, C_+ - \Phi_0\big\rangle_F\big]\\ &= \frac{1}{m}\big(\mathbb{E}\big[\|\Phi_0\|_F^2\big] - \|C_+\|_F^2\big) + \frac{2}{m}\sum_{k=1}^{m-1}(1-\tfrac{k}{m})\mathbb{E}\big[\big\langle C_+ - \Phi_k, C_+ - \Phi_0\big\rangle_F\big].\end{aligned}$$

The sum $\frac{2}{m}\sum_{k=1}^{m-1}\ldots$ in the last expression vanishes as Φ_k and Φ_0 are stochastically independent. Concerning the first summand, we compute (for the definition of φ see Assumption **(A1)**)

$$\|\Phi_0\|_F^2 = \big\|\big(\psi_i(x_0)\psi_j(y_0)\big)_{i,j=1}^N\big\|_F^2 = \sum_{i,j=1}^N \psi_i^2(x_0)\psi_j^2(y_0) = \varphi(x_0)\varphi(y_0) \tag{20}$$

and hence

$$\mathbb{E}\big[\|\Phi_0\|_F^2\big] = \int\int \varphi(x)\varphi(y)\,\rho(x,dy)\,d\mu(x) = \int \varphi(x)(K\varphi)(x)\,d\mu(x) = \langle K\varphi,\varphi\rangle.$$

Now, apply Markov's inequality to the non-negative random variable $\|C_+ - \widehat{C}_+\|_F^2$ to obtain the first probabilistic bound. The second is proved similarly.

Now, assume that $\varphi \in L^\infty(\mu)$ and set $S := \|\varphi\|_{L^\infty(\mu)}$. Since $\|K\| \le L$, we observe that

$$\|C_+\|_F^2 = \sum_{i,j=1}^N |\langle\psi_i, K\psi_j\rangle|^2 \le \sum_{i,j=1}^N \|\psi_i\|^2 L^2 \|\psi_j\|^2 = L^2\|\varphi\|_1^2 \le L^2 S^2.$$

Next, let $(X,Y)\sim\mu_0$. Then $\|\Psi(X)\Psi(Y)^\top\|_F^2 = \varphi(X)\varphi(Y)$, see (20). We shall prove that $\varphi(X)\varphi(Y) \le S^2$ a.s.. To see this, we note that $X\sim\mu$ and thus $\varphi(X)\le S$ a.s., so that

$$\begin{aligned}\mathbb{P}\big(\varphi(X)\varphi(Y) > S^2\big) &\le \mathbb{P}(\varphi(Y) > S) = \mathbb{P}\big((X,Y)\in\mathcal{X}\times\varphi^{-1}((S,\infty))\big)\\ &= \int \rho(x,\varphi^{-1}((S,\infty)))\,d\mu(x) \le L^2\cdot\mu(\varphi^{-1}((S,\infty))) = 0.\end{aligned}$$

Thus, we obtain that $\|\Psi(x_k)\Psi(y_k)^\top\|_F \le S$ a.s. for all $k = 0,\ldots,m-1$.

Consider the random matrices $D_k := C_+ - \Psi(x_k)\Psi(y_k)^\top$, $k = 0,\ldots,m-1$. These are stochastically independent, $\mathbb{E}[D_k] = 0$, and $\|D_k\|_F \le (1+L)S$ a.s.. Hence, Hoeffding's inequality for bounded independent random variables in Hilbert spaces (see Corollary A.5.2 in (Mollenhauer, 2021) and Theorem 3.5 in (Pinelis, 1994)) implies that

$$\mathbb{P}\big(\|C_+ - \widehat{C}_+\|_F > \varepsilon\big) = \mathbb{P}\left(\left\|\tfrac{1}{m}\sum_{k=0}^{m-1} D_k\right\|_F > \varepsilon\right) \le 2\cdot e^{-\frac{m\varepsilon^2}{2(1+L)^2S^2}},$$

as claimed. The second bound is proved similarly. □

Remark A.5. The bound on $\|C - \widehat{C}\|_F$ in Proposition A.4 had been found already in Lemma 3.4.1 of (Mollenhauer, 2021) (where $\|\varphi\|_\infty$ is replaced by the slightly worse constant $N\cdot\max_{i=1,\ldots,N}\|\psi_i\|_\infty^2$).

B. Further aspects of the Koopman operator

The following considers both cases of i.i.d. and ergodic sampling, that is, we let $\mu\in\{\nu,\pi\}$ as in **(S1)** and **(S2)**.

B.1. Well-definedness of the Koopman operator

Note that (5) and (1) imply $\tau \ll \mu$, where $\tau(A) := \int \rho(x, A)\, d\mu(x)$, with a density $g \in L^\infty(\mu)$. In particular, $f \in L^p(\mu)$ implies $f \in L^p(\tau)$, since $\int |f|^p\, d\tau = \int |f|^p g\, d\mu \le \|g\|_\infty \int |f|^p\, d\mu$.

Note that for every simple function f on $\mathcal{X}$ we have

$$\int f\, d\tau = \int\int f(y)\, \rho(x, dy)\, d\mu(x).$$

Now, let $f \in L^1(\mu)$, $f \ge 0$. Then $f \in L^1(\tau)$, and there exists a sequence of simple functions $(f_n)_{n\in\mathbb{N}}$, $0 \le f_n \le f$, such that $f_n \uparrow f$ as $n \to \infty$. Hence,

$$\int f\, d\tau = \int (f - f_n)\, d\tau + \int\int f_n(y)\, \rho(x, dy)\, d\mu(x).$$

The first integral approaches zero as $n \to \infty$ by dominated convergence ($0 \le f - f_n \le 2|f|$), and the second integral tends to $\int\int f(y)\, \rho(x, dy)\, d\mu(x)$ as $n \to \infty$ by Beppo-Levi. In particular, $\int f(y)\, \rho(x, dy) < \infty$ for μ-a.e. $x \in \mathcal{X}$, which is the definition of the Koopman operator.

B.2. The adjoint of the Koopman operator

For $f \in L^1(\mu)$ define the signed measure

$$\mu_f(A) := \int \rho(x, A) f(x)\, d\mu(x), \qquad A \in \mathfrak{B}(\mathcal{X}).$$

If $\mu(A) = 0$, then $\int \rho(x, A)\, d\mu(x) \le L\mu(A) = 0$ (where $L = 1$ if $\mu = \pi$), hence $\rho(x, A) = 0$ for μ-a.e. $x \in \mathcal{X}$, and thus $\mu_f(A) = 0$. Therefore, there exists a unique $Pf \in L^1(\mu)$ such that $\mu_f(A) = \int_A Pf\, d\mu$ holds for all $A \in \mathfrak{B}(\mathcal{X})$. It is now easily seen that $P : L^1(\mu) \to L^1(\mu)$ is linear. Moreover, if $f \in L^1(\mu)$ and $\Delta_+ = \{f \ge 0\}$, $\Delta_- = \{f < 0\}$, we have $\mu_f = \mu_f^+ - \mu_f^-$, where $\mu_f^p m(A) = \pm \int_{\Delta_\pm} \rho(x, A)\, d\mu(x)$, thus

$$\int |Pf|\, d\mu = |\mu_f|(\mathcal{X}) = \mu_f^+(\mathcal{X}) + \mu_f^-(\mathcal{X}) = \int |f|\, d\mu.$$

Hence, P is an isometry on $L^1(\mu)$. It is moreover easy to see that P is a Markov operator, i.e., $P\mathbb{1} = \mathbb{1}$ and $Pf \ge 0$ if $f \ge 0$.

For a Borel set $A \in \mathfrak{B}(\mathcal{X})$ and $f \in L^q(\mu)$ (where $\frac{1}{p} + \frac{1}{q} = 1$) we have

$$\int_A K_p^* f\, d\mu = \langle K_p^* f, \mathbb{1}_A \rangle_{L^q, L^p} = \langle f, K_p \mathbb{1}_A \rangle_{L^q, L^p} = \int \rho(x, A) f(x)\, d\mu(x) = \mu_f(A) = \int_A Pf\, d\mu.$$

This implies $K_p^* f = Pf$. In particular, P maps $L^q(\mu)$ into $L^q(\mu)$ for all $q \in [1, \infty]$ and coicides there with K_p^* for $q \ne 1$.

Lemma B.1. *For $f \in L^2(\mu)$ we have $(Pf)^2 \le Pf^2$ μ-a.e.*

Proof. Let f be a simple function, i.e., $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$, where the A_i are mutually disjoint and $\bigcup_{i=1}^n A_i = \mathcal{X}$. Then $\sum_{i=1}^n P\mathbb{1}_{A_i} = P\mathbb{1} = \mathbb{1}$, hence, by convexity of $z \mapsto z^2$,

$$(Pf)^2(x) = \Big(\sum_{i=1}^n a_i (P\mathbb{1}_{A_i})(x)\Big)^2 \le \sum_{i=1}^n a_i^2 (P\mathbb{1}_{A_i})(x) = (Pf^2)(x)$$

and therefore $(Pf)^2 \le Pf^2$. Similarly, $|Pf| \le P|f|$. If $f \in L^2(\mu)$, let (f_n) be a sequence of simple functions such that $\|f_n - f\|_{L^2(\mu)} \to 0$ as $n \to \infty$. Then, for every $A \in \mathfrak{B}(\mathcal{X})$,

$$\begin{aligned}
\int_A \left[(Pf)^2 - Pf^2\right] d\mu &\le \int_A ((Pf)^2 - (Pf_n)^2)\, d\mu + \int_A (Pf_n^2 - Pf^2)\, d\mu \\
&\le \int |P(f - f_n) P(f + f_n)|\, d\mu + \|P(f_n^2 - f^2)\|_{L^1(\mu)} \\
&\le \|P(f_n - f)\|_{L^2(\mu)} \|P(f_n + f)\|_{L^2(\mu)} + \|f_n^2 - f^2\|_{L^1(\mu)} \\
&\le 2\|f_n - f\|_{L^2(\mu)} \|f_n + f\|_{L^2(\mu)}.
\end{aligned}$$

This proves $(Pf)^2 \le Pf^2$ μ-a.e. for all $f \in L^2(\mu)$. □

Lemma B.2. *Let $f, g, h \in L^2(\mu)$ such that $g^2 Pf^2 \in L^1(\mu)$. Then $f \cdot K_1(gh) \in L^1(\mu)$ and*

$$\int f \cdot K_1(gh)\, d\mu = \int [g \cdot K^* f] \cdot h\, d\mu.$$

Proof. Let $(f_n) \subset L^\infty(\mu)$ be a sequence of simple functions such that $f_n \to f$ and $|f_n| \nearrow |f|$ pointwise μ-a.e. as $n \to \infty$. Then

$$\int |K^*(f_n - f)|\, d\mu \le \int K^*|f_n - f|\, d\mu = \int |f_n - f|\, d\mu,$$

which converges to zero as $n \to \infty$ by dominated convergence ($|f_n - f| \le 2|f|$). Hence, there exists a subsequence (f_{n_k}) such that $K^* f_{n_k} \to K^* f$ μ-a.e. as $k \to \infty$. WLOG, we may therefore assume that $K^* f_n \to K^* f$ μ-a.e. as $n \to \infty$. By monotone convergence,

$$\begin{aligned}\int |f||K_1(gh)|\, d\mu &= \lim_{n\to\infty} \int |f_n||K_1(gh)|\, d\mu \le \limsup_{n\to\infty} \int |f_n| \cdot K_1(|gh|)\, d\mu \\ &= \limsup_{n\to\infty} \int K_1^*|f_n| \cdot |gh|\, d\mu \le \int |g|P|f| \cdot |h|\, d\mu,\end{aligned}$$

which is a finite number. Hence, indeed, $f \cdot K_1(gh) \in L^1(\mu)$ and, by dominated convergence,

$$\int f \cdot K_1(gh)\, d\mu = \lim_{n\to\infty} \int f_n \cdot K_1(gh)\, d\mu = \lim_{n\to\infty} \int [g \cdot K_1^* f_n] \cdot h\, d\mu = \int [g \cdot K^* f] \cdot h\, d\mu,$$

as claimed. □

C. Auxiliary results

In this section, we provide auxiliary results on the invertibility of the considered matrices.

Definition C.1. Let μ be an arbitrary measure on $\mathcal{X}$, and let measurable functions $\psi_1, \dots, \psi_N : \mathcal{X} \to \mathbb{R}$ be given.

(a) We say that $\psi_1, \dots, \psi_N$ are linearly independent w.r.t. the measure μ (or simply μ-linearly independent) if $\sum_{j=1}^N \lambda_j \psi_j = 0$ μ-a.e. implies $\lambda_1 = \dots = \lambda_N = 0$.

(b) We say that $\psi_1, \dots, \Psi_N$ are strongly linearly independent w.r.t. the measure μ (or simply strongly μ-linearly independent) if $\mu\left(\sum_{j=1}^N \lambda_j \psi_j = 0\right) > 0$ implies $\lambda_1 = \dots = \lambda_N = 0$.

If $\psi_1, \dots, \Psi_N$ are strongly μ-linearly independent, it follows in particular that the sets of zeros $\psi_j^{-1}(\{0\})$ are null sets (w.r.t. μ). Furthermore, note that the following implications hold:

$$\text{strong } \mu\text{-linear independence} \implies \mu\text{-linear independence} \implies \text{linear independence}.$$

The following lemma holds for both cases **(S1)** and **(S2)**.

Lemma C.2. *Let $\mu \in \{\nu, \pi\}$ and let $\psi_1, \dots, \psi_N$ be μ-linearly independent. Then, the matrix C is invertible, and the matrix representation $\widetilde{K}_{\mathbb{V}}$ of the compression $P_{\mathbb{V}} K|_{\mathbb{V}}$ of K to $\mathbb{V}$ is given by $\widetilde{K}_{\mathbb{V}} = C^{-1} C_+$.*

Proof. Let $v \in \mathbb{R}^N$ such that $Cv = 0$. Then we have $\sum_{j=1}^N \langle \psi_i, \psi_j \rangle v_j = 0$ and hence $\langle \psi_i, \sum_{j=1}^N v_j \psi_j \rangle = 0$ for all $i \in [N]$. But this implies $\sum_{j=1}^N v_j \psi_j = 0$ μ-a.e. and thus $v_1 = \dots = v_N = 0$ as the ψ_j are μ-linearly independent. Hence, C is indeed invertible.

For any $j \in [N]$, we have $P_{\mathbb{V}} K \psi_j = \sum_{i=1}^N a_{ij} \psi_i$ with some $a_{ij} \in \mathbb{R}$ forming the matrix $\widetilde{K}_{\mathbb{V}} = (a_{ij})_{i,j=1}^N$. Next, for $i, j \in [N]$,

$$(C_+)_{ij} = \langle \psi_i, K\psi_j \rangle = \langle \psi_i, P_{\mathbb{V}} K \psi_j \rangle = \left\langle \psi_i, \sum_{\ell=1}^N a_{\ell j} \psi_\ell \right\rangle = \sum_{\ell=1}^N \langle \psi_i, \psi_\ell \rangle a_{\ell j} = (C\widetilde{K}_{\mathbb{V}})_{ij},$$

and the claim follows. □

Lemma C.3. *In case* **(S2)***, $\widehat{C}$ is invertible a.s. if and only if $m \geq N$ and $\psi_1, \dots, \psi_N$ are strongly μ-linearly independent.*

Proof. First of all, note that $\widehat{C}$ is invertible if and only if the rank of Ψ_X equals N.

Assume that the ψ_j are strongly linearly independent and $m \geq N$. Define the random variables $Y_j := \Psi(x_j) \in \mathbb{R}^N$, $j = 0, \dots, m-1$, and the pushforward measure $\tau := \mu \circ \Psi^{-1}$. Then the Y_j are τ-distributed and independent, and we have $\mathbb{P}(\lambda^\top Y_j = 0) = 0$ for all j and all $\lambda \in \mathbb{R}^N \backslash \{0\}$.

We shall show that $Y_0, \dots, Y_{N-1}$ are linearly independent a.s. For this, fix $j \in \{0, \dots, N-1\}$, let $J \subset \{0, \dots, N-1\} \backslash \{j\}$ with $1 \leq k \leq N-1$ elements, and define

$$V := \big\{(y_1, \dots, y_{k+1}) \in (\mathbb{R}^N)^{k+1} : y_{k+1} \in \operatorname{span}\{y_1, \dots, y_k\}\big\}.$$

Then we have

$$\begin{aligned}\mathbb{P}\big(Y_j \in \operatorname{span}\{Y_i : i \in J\}\big) = \tau^{k+1}(V) &= \int_{(\mathbb{R}^N)^k} \int_{\operatorname{span}\{y_1,\dots,y_k\}} d\tau(y)\, d\tau^k(y_1, \dots, y_k) \\ &= \int_{(\mathbb{R}^N)^k} \mathbb{P}\big(Y_j \in \operatorname{span}\{y_1, \dots, y_k\}\big)\, d\tau^k(y_1, \dots, y_k).\end{aligned}$$

Fix k vectors $y_1, \dots, y_k \in \mathbb{R}^N$. Then there exists $\lambda \in \mathbb{R}^N \backslash \{0\}$ such that $\lambda^\top [y_1, \dots, y_k] = 0$, hence

$$\mathbb{P}\big(Y_j \in \operatorname{span}\{y_1, \dots, y_k\}\big) \leq \mathbb{P}(\lambda^\top Y_j = 0) = 0.$$

This implies $\mathbb{P}\big(Y_j \in \operatorname{span}\{Y_i : i \in J\}\big) = 0$, which proves the claim.

Conversely, assume that Ψ_X has rank N a.s., let $\lambda \in \mathbb{R}^N \backslash \{0\}$ and set $Z := \{\lambda^\top \Psi = 0\} \subset \mathcal{X}$. Then

$$0 = \mathbb{P}(\lambda^\top \Psi_X = 0) = \mathbb{P}(x_k \in Z\ \forall k = 0, \dots, m-1) = \mu^m(Z^m) = [\mu(Z)]^m,$$

which proves that $\psi_1, \dots, \psi_N$ are strongly μ-linearly independent. □

Lemma C.4. *Let $m \geq N+1$. In case* **(S1)***, $\widehat{C}$ is invertible a.s. if for each $(N-1)$-dimensional subspace $M \subset \mathbb{R}^N$ and $x \in \Psi^{-1}(M)$ we have $\rho(x, \Psi^{-1}(M)) = 0$.*

Proof. If $\mathcal{L} \subset \mathbb{R}^N$ is a subspace with $\dim \mathcal{L} < N-1$, for any $x \in \mathcal{X}$ there exists an $(N-1)$-dimensional subspace $M \supset \mathcal{L}$ with $x \in \Psi^{-1}(M)$, and we obtain $\rho(x, \Psi^{-1}(\mathcal{L})) \leq \rho(x, \Psi^{-1}(M)) = 0$. Hence, the assumption holds for all subspaces $M \subset \mathbb{R}^N$.

In the following, let $A := \Psi^{-1}(\{0\})$. Assume that $\mu(A) > 0$. Note that $\mu(A) < 1$ by μ-linear independence of the ψ_j. We show that

$$\mathbb{P}([\Psi(X_1), \dots, \Psi(X_N)] \text{ invertible} \,|\, X_0 \in A) = 1 \tag{21}$$

$$\mathbb{P}([\Psi(X_0), \dots, \Psi(X_{N-1})] \text{ invertible} \,|\, X_0 \notin A) = 1. \tag{22}$$

Then the claim follows. By assumption, we have $\rho(x, A) = 0$ for $x \in A$, hence

$$\mathbb{P}(\Psi(X_1) = 0 \,|\, X_0 \in A) = \frac{1}{\mu(A)} \int_A \rho(x, A)\, d\mu(x) = 0$$

Next, let $k \in \{1, \dots, N-1\}$ and set $M(x_1, \dots, x_k) := \operatorname{span}\{\Psi(x_1), \dots, \Psi(x_k)\}$ for $x_1, \dots, x_k \in \mathcal{X}$. Then

$$\begin{aligned}&\mathbb{P}\big(\Psi(X_{k+1}) \in M(X_1, \dots, X_k) \,|\, X_0 \in A\big) \\ &\quad= \frac{1}{\mu(A)} \int_A \int_{x_1} \cdots \int_{x_k} \int_{\Psi^{-1}(M(x_1,\dots,x_k))} \rho(x_k, dx_{k+1}) \rho(x_{k-1}, dx_k) \dots \rho(x_0, dx_1)\, d\mu(x_0) \\ &\quad= \frac{1}{\mu(A)} \int_A \int_{x_1} \cdots \int_{x_k} \rho(x_k, \Psi^{-1}(M(x_1, \dots, x_k))) \rho(x_{k-1}, dx_k) \dots \rho(x_0, dx_1)\, d\mu(x_0).\end{aligned}$$

Now, note that for all $x_k \in \mathcal{X}$ we have $x_k \in \Psi^{-1}(M(x_1, \dots, x_k))$, hence $\rho(x_k, \Psi^{-1}(M(x_1, \dots, x_k))) = 0$. Therefore, we have $\mathbb{P}\big(\Psi(X_{k+1}) \in M(X_1, \dots, X_k) \,|\, X_0 \in A\big) = 0$ for all $k = 1, \dots, N-1$, which shows (21). The relation (22) can be shown similarly with all indices dropped by one and A replaced by $\mathcal{X} \backslash A$. The proof for (22) carries over to the case $\mu(A) = 0$ without conditioning and with $\mathcal{X} \setminus A$ replaced by $\mathcal{X}$. □

The following result provides an improved version of Theorem 12 in (Nüske et al., 2023).

Lemma C.5. *Let $C, D \in \mathbb{R}^{N\times N}$ be such that C is invertible and $D \neq 0$. Let $\widehat{C}, \widehat{D} \in \mathbb{R}^{N\times N}$ be random matrices such that $\widehat{C}$ is invertible a.s.. Then for any sub-multiplicative matrix norm $\|\cdot\|$ on $\mathbb{R}^{N\times N}$ and any $\varepsilon > 0$ we have*

$$\mathbb{P}\big(\|C^{-1}D - \widehat{C}^{-1}\widehat{D}\| > \varepsilon\big) \le \mathbb{P}\Big(\|D-\widehat{D}\| > \tfrac{\varepsilon}{\tau}\|D\|\Big) + \mathbb{P}\Big(\|C-\widehat{C}\| > \tfrac{\varepsilon}{\tau}\|C^{-1}\|^{-1}\Big),$$

where $\tau = 2\|C^{-1}\|\|D\| + \varepsilon$.

Proof. We have $C^{-1}D - \widehat{C}^{-1}\widehat{D} = \widehat{C}^{-1}(D-\widehat{D}) + (C^{-1} - \widehat{C}^{-1})D$ and thus

$$\begin{aligned}
\mathbb{P}\big(\|C^{-1}D - \widehat{C}^{-1}\widehat{D}\| > \varepsilon\big) &\le \mathbb{P}\big(\|\widehat{C}^{-1}\|\|D-\widehat{D}\| + \|D\|\|C^{-1}-\widehat{C}^{-1}\| > \varepsilon\big)\\
&\le \mathbb{P}\big(\|\widehat{C}^{-1}\|\|D-\widehat{D}\| > \tfrac{\varepsilon}{2} \ \vee\ \|D\|\|C^{-1}-\widehat{C}^{-1}\| > \tfrac{\varepsilon}{2}\big)\\
&= \mathbb{P}\Big(\|D-\widehat{D}\| > \tfrac{\varepsilon/2}{\|\widehat{C}^{-1}\|} \ \vee\ \|C^{-1}-\widehat{C}^{-1}\| > \tfrac{\varepsilon/2}{\|D\|}\Big)\\
&\le \mathbb{P}\Big(\|D-\widehat{D}\| > \frac{\varepsilon/2}{\|C^{-1}\|+\frac{\varepsilon/2}{\|D\|}} \ \vee\ \|C^{-1}-\widehat{C}^{-1}\| > \tfrac{\varepsilon/2}{\|D\|}\Big)\\
&\le \mathbb{P}\Big(\|D-\widehat{D}\| > \tfrac{\varepsilon/2}{\tau}\|D\|\Big) + \mathbb{P}\Big(\|C^{-1}-\widehat{C}^{-1}\| > \tfrac{\varepsilon/2}{\|D\|}\Big).
\end{aligned}$$

Next, we estimate

$$\|C^{-1}-\widehat{C}^{-1}\| = \|C^{-1}(\widehat{C}-C)\widehat{C}^{-1}\| \ \le\ \|C^{-1}\|\|\widehat{C}-C\|\big(\|\widehat{C}^{-1}-C^{-1}\| + \|C^{-1}\|\big).$$

Hence, if $\|C-\widehat{C}\| < \frac{1}{\|C^{-1}\|}$, then

$$\|C^{-1}-\widehat{C}^{-1}\| \ \le\ \frac{\|C^{-1}\|^2\|C-\widehat{C}\|}{1-\|C^{-1}\|\|C-\widehat{C}\|}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big(\|C^{-1}-\widehat{C}^{-1}\| > \tfrac{\varepsilon/2}{\|D\|}\Big) &\le \mathbb{P}\Big(\|C-\widehat{C}\| \ge \tfrac{1}{\|C^{-1}\|} \ \vee\ \tfrac{\|C^{-1}\|^2\|C-\widehat{C}\|}{1-\|C^{-1}\|\|C-\widehat{C}\|} > \tfrac{\varepsilon/2}{\|D\|}\Big)\\
&= \mathbb{P}\Big(\|C-\widehat{C}\| \ge \tfrac{1}{\|C^{-1}\|} \ \vee\ \|C-\widehat{C}\| > \tfrac{\varepsilon/2}{\|C^{-1}\|(\|D\|\|C^{-1}\|+\varepsilon/2)}\Big)\\
&= \mathbb{P}\Big(\|C-\widehat{C}\| > \tfrac{\varepsilon/2}{\|C^{-1}\|(\|D\|\|C^{-1}\|+\varepsilon/2)}\Big),
\end{aligned}$$

and the lemma is proved. □

D. Spectral measures of unitary operators

Let U be a unitary operator in a Hilbert space $\mathcal{H}$. By the spectral theorem for normal operators in Hilbert spaces (see, e.g., (Conway, 2010)), there exists an operator-valued measure E on the Borel sigma algebra $\mathcal{B}$ of $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$, which has the following properties:

- $E(\Delta)$ is an orthogonal projection for all $\Delta \in \mathfrak{B}$.
- $E(\mathbb{T}\backslash\sigma(U)) = 0$ and $E(\sigma(U)) = I$.
- $E(\Delta_1 \cap \Delta_2) = E(\Delta_1)E(\Delta_2) = E(\Delta_2)E(\Delta_1)$ for $\Delta_1, \Delta_2 \in \mathcal{B}$.
- $E(\Delta)\mathcal{H}$ is U-invariant for each $\Delta \in \mathcal{B}$.
- $\sigma(U|_{E(\Delta)\mathcal{H}}) \subset \sigma(U) \cap \Delta$ for closed $\Delta \in \mathcal{B}$.

The measure E is called the *spectral measure* of U. In the finite-dimensional case (i.e., when $U \in \mathbb{C}^{n\times n}$ is a unitary matrix), the projection $E(\Delta)$ is the orthogonal projection onto the sum of eigenspaces corresponding to the eigenvalues of U in Δ. This is also true in the infinite-dimensional case if U has only discrete spectrum in Δ (i.e., isolated eigenvalues).

Let $g : \mathbb{T} \to \mathbb{C}$ be a bounded measurable function. Then $g(U) := \int g \, dE$ defines a bounded normal operator, and for $f \in \mathcal{H}$ we have

$$\langle g(U)f, f\rangle = \int_{\mathbb{T}} g(z)\, d\mu_f(z) \qquad \text{and} \qquad \|g(U)f\|^2 = \int_{\mathbb{T}} |g(z)|^2\, d\mu_f(z),$$

where μ_f is the measure $\mu_f(\Delta) := \|E(\Delta)f\|^2$, $\Delta \in \mathcal{B}$.