Kernel Autocovariance Operators of Stationary Processes: Estimation and Convergence

Mattes Mollenhauer[1]   MATTES.MOLLENHAUER@FU-BERLIN.DE
Stefan Klus[3]   S.KLUS@HW.AC.UK
Christof Schütte[1,2]   SCHUETTE@ZIB.DE
Péter Koltai[1]   PETER.KOLTAI@FU-BERLIN.DE

[1]*Institute of Mathematics, Freie Universität Berlin*
Arnimallee 6, D-14195 Berlin
[2]*Zuse Intitute Berlin*
Takustraße 7, D-14195 Berlin
[3]*School of Mathematical & Computer Sciences, Heriot–Watt University*
Edinburgh, EH14 4AS, UK

Editor: Ingo Steinwart

Abstract

We consider autocovariance operators of a stationary stochastic process on a Polish space that is embedded into a reproducing kernel Hilbert space. We investigate how empirical estimates of these operators converge along realizations of the process under various conditions. In particular, we examine ergodic and strongly mixing processes and obtain several asymptotic results as well as finite sample error bounds. We provide applications of our theory in terms of consistency results for *kernel PCA* with dependent data and the *conditional mean embedding* of transition probabilities. Finally, we use our approach to examine the nonparametric estimation of Markov transition operators and highlight how our theory can give a consistency analysis for a large family of spectral analysis methods including kernel-based *dynamic mode decomposition*.

Keywords: stationary time series, autocovariance operator, kernel mean embedding, mixing, ergodic process

1. Introduction

The *kernel mean embedding*, i.e., the embedding of a probability distribution into a *reproducing kernel Hilbert space* (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007), and the closely related theory of *kernel covariance operators* have spawned a vast variety of nonparametric models and statistical tests over the last years. For an overview of the kernel mean embedding theory, we refer the reader to the survey by Muandet et al. (2017) and the references therein. Kernel covariance operators serve as the theoretical foundation of several spectral analysis and component decomposition techniques including *kernel principal component analysis*, *kernel independent component analysis* and *kernel canonical correlation analysis*. Consistency results and the statistical analysis of these methods can therefore be directly based on the estimation of kernel covariance operators (Blanchard et al., 2007;

Fukumizu et al., 2007; Rosasco et al., 2010). Moreover, kernel covariance operators and their connection to L^p-space integral operators and random matrices are a fundamental concept used to formalize statistical learning (see for example Smale and Zhou, 2007; Rosasco et al., 2010).

In this paper, we extend the statistical theory of kernel covariance operators from the independent scenario to *kernel autocovariance operators* of a stationary stochastic process (that is, kernel cross-covariance operators with respect to a time-lagged version of the process). Recently, several nonparametric models for dependent data, sequence modeling, and time series analysis based on kernel mean embeddings have emerged. Popular approaches include filtering (Song et al., 2009; Fukumizu et al., 2013; Gebhardt et al., 2019), transition models (Sun et al., 2019; Grünewälder et al., 2012b) and reinforcement learning (van Hoof et al., 2015; Lever et al., 2016; van Hoof et al., 2017; Stafford and Shawe-Taylor, 2018; Gebhardt et al., 2018), to only name a few. A theoretical tool to understand these concepts is the kernel autocovariance operator, as its plays an important role in the nonparametric approximation of transition probabilities. This concept has been introduced in an operator-theoretic sense under the name *conditional mean embedding* by Song et al. (2009) under strong technical requirements (see also Klebanov et al., 2020). These requirements have later been relaxed by developing the theory in a vector-valued regression scenario (Grünewälder et al., 2012a; Park and Muandet, 2020; Mollenhauer and Koltai, 2020; Li et al., 2022). Although time series are one of the primary fields of application, consistency results for the empirical conditional mean embedding have been limited to the case of independent data until now. As an application of the results of this paper, we prove standard consistency statements for dependent data.

Recent results (Klus et al., 2018a, 2020; Mollenhauer and Koltai, 2020) show that eigenfunctions of *Markov transition operators* can be approximated with the conditional mean embedding. In particular, it was discovered that a large family of kernel-based spectral analysis and model order reduction techniques for stochastic processes and dynamical systems (see Kutz et al., 2016; Klus et al., 2018b and Wu and Noé, 2020 for an overview) implicitly approximate the spectral decomposition of a transition operator defined on an RKHS. This operator can be expressed in terms of kernel autocovariance operators. Different versions of these methods are popular in fluid dynamics (Schmid, 2010; Tu et al., 2014; Williams et al., 2015a,b), signal processing (Molgedey and Schuster, 1994), machine learning (Harmeling et al., 2003; Kawahara, 2016; Hua et al., 2017), and molecular dynamics (Pérez-Hernández et al., 2013; Schwantes and Pande, 2015) under the names *dynamic mode decomposition* and *time-lagged independent component analysis*. Until now, a full statistical convergence analysis of these techniques has not been conducted to the best of our knowledge. A theoretical examination of kernel autocovariance operators contributes significantly to the understanding of kernel-based versions of the aforementioned approaches.

The theory of weakly dependent random processes taking values in infinite-dimensional Banach spaces or Hilbert spaces has become increasingly important especially due to applications in the field of *functional data analysis* (Hörmann and Kokoszka, 2010; Horváth and Kokoszka, 2012). In infinite-dimensional statistics, the estimation of covariance and cross-covariance operators (Baker, 1970, 1973) is a fundamental concept. Under parametric model

assumptions about the process, the estimation of covariance and autocovariance operators has been examined in various scenarios. For autoregressive (AR) processes in Banach spaces and Hilbert spaces, weak convergence and asymptotic normality has been established (Bosq, 2000, 2002; Mas, 2002; Dehling and Sharipov, 2005; Mas, 2006). Soltani and Hashemi (2011) add the assumption of periodic correlation for AR processes in Hilbert spaces. Allam and Mourid (2014, 2019) provide rates for almost sure convergence of covariance operators in Hilbert–Schmidt norm for an AR process with random coefficients. For processes in an L^2 function space, the weak convergence of covariance operators has been examined by Kokoszka and Reimherr (2013) under the assumption of L^4-*m approximability* (a concept generalizing *m-dependence*, which includes certain autoregressive and nonlinear models, see Hörmann and Kokoszka, 2010) in the context of functional principal component analysis.

In this paper, we consider a stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ taking values in a Polish space E. Let $\mathscr{H}$ be a reproducing kernel Hilbert space (RKHS) with the *canonical feature map* $\varphi : E \to \mathscr{H}$. In contrast to the previously mentioned work, we investigate autocovariance operators of the corresponding embedded RKHS-valued process

$$\big(\varphi(X_t)\big)_{t\in\mathbb{Z}}.$$

We face the challenge that properties of the E-valued process $(X_t)_{t\in\mathbb{Z}}$ which quantify the convergence speed of any empirical statistic must transfer accordingly to the embedded version of the process in the Hilbert space $\mathscr{H}$. However, we can not require restrictive assumptions about the feature map φ in order to ensure applicability of our results for various RKHSs in practice. Hence, in contrast to the previously mentioned literature, we consider a more general setting that does not require either $(X_t)_{t\in\mathbb{Z}}$ or $(\varphi(X_t))_{t\in\mathbb{Z}}$ to obey any specific parametric time series model.

Recently, Blanchard and Zadorozhnyi (2019) derived a Bernstein-type inequality for Hilbert space processes for a class of mixing properties called $\mathcal{C}$-mixing (Maume-Deschamps, 2006). As a special case, the authors show that under fairly restrictive Lipschitz conditions on the feature map φ, this mixing property is preserved under the RKHS embedding of a so-called τ-mixing process. The derived inequality is then used to obtain concentration bounds for the context of RKHS learning theory, including kernel covariance operator estimation *without a time lag*. As described for example by Hang and Steinwart (2017), the class of $\mathcal{C}$-mixing coefficients is only partly related to the classical strong mixing coefficients found in the literature (Doukhan, 1994; Bradley, 2005), which we will consider in this paper, in particular the concept of α-*mixing*.

The contributions of this paper are:

(i) A mathematical framework for kernel autocovariance operators of a stationary discrete-time process $(X_t)_{t\in\mathbb{Z}}$ taking values in a Polish space. This framework allows the investigation of ergodicity and strong mixing in the context of the embedded process $(\varphi(X_t))_{t\in\mathbb{Z}}$ under minimal requirements on the feature map φ. In particular, our assumptions are easily justifiable for further application in practical work on RKHS-based time series models.

(ii) A collection of various asymptotic as well as nonasymptotic results about the estimation error of empirical kernel autocovariance operators based on single trajectories of the process $(X_t)_{t \in \mathbb{Z}}$. These results are presented in a form that is directly accessible for work on related topics.

(iii) Applications of our results to

    (a) the consistency of kernel PCA with dependent data;

    (b) the consistency of the conditional mean embedding of transition probabilities under the typical technical assumptions; and

    (c) the estimation of Markov transition operators and their role in a family of spectral analysis methods for dynamical systems.

This paper is structured as follows. In Section 2, we recall the required preliminaries from spectral theory, Bochner integration, and reproducing kernel Hilbert spaces and formulate our working assumptions. Section 3 addresses the strong law of large numbers of empirical kernel autocovariance operators under the hypothesis of ergodicity. We introduce the concept of strong mixing and derive standard probabilistic results including the central limit theorem in Section 4. A general concentration bound for the estimation error can be found in Section 5. Based on these results, we highlight applications to kernel PCA from dependent data (Section 6), the conditional mean embedding (Section 7) and the approximation of Markov transition operators (Section 8). We conclude our work in Section 9.

2. Preliminaries

2.1 General Notation

We give an overview of our notation and collect well-known facts from operator theory and probability theory. For details, we refer the reader to Reed and Simon (1980) and Kallenberg (2002). In what follows, we write B for a separable real Banach space with norm $\|\cdot\|_B$, and H for a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. $\mathrm{L}(B)$ stands for the Banach space of bounded linear operators on B equipped with the operator norm $\|\cdot\|$. The expression $H \otimes H$ denotes the tensor product space: $H \otimes H$ is the Hilbert space completion of the algebraic tensor product with respect to the inner product $\langle x_2 \otimes x_1, x_2' \otimes x_1' \rangle_{H \otimes H} = \langle x_1, x_1' \rangle_H \langle x_2, x_2' \rangle_H$. For $x_1, x_2 \in H$, we interpret the element $x_2 \otimes x_1 \in H \otimes H$ as the linear rank-one operator $x_2 \otimes x_1 \colon H \to H$ defined by $x \mapsto \langle x, x_1 \rangle_H x_2$. Whenever $(e_i)_{i \in I}$ is a complete orthonormal system (CONS) in H, $(e_i \otimes e_j)_{i,j \in I}$ is a CONS in $H \otimes H$. Thus, when H is separable, $H \otimes H$ is separable.

Every compact operator A on H admits a *singular value decomposition*, that is, there exist orthonormal systems $\{u_i\}_{i \in J}$ and $\{v_i\}_{i \in J}$ in H such that

$$A = \sum_{i \in J} \sigma_i(A) u_i \otimes v_i, \tag{1}$$

where $(\sigma_i(A))_{i \in J}$ are the strictly positive and nonincreasingly ordered (including multiplicities) singular values of A with an (either countably infinite or finite) index set J. The

convergence in (1) is meant with respect to the operator norm. The *rank* of A is defined as the cardinality of J.

For $1 \leq p \leq \infty$, the *p-Schatten class* $S_p(H)$ consists of all compact operators A on H such that the norm $\|A\|_{S_p(H)} := \|(\sigma_i(A))_{i \in J}\|_{\ell_p}$ is finite. Here $\|(\sigma_i(A))_{i \in J}\|_{\ell_p}$ denotes the ℓ_p sequence space norm of the sequence of singular values. The spaces $S_p(H)$ are two-sided ideals in $L(H)$. Moreover $\|A\| \leq \|A\|_{S_q(H)} \leq \|A\|_{S_p(H)}$ holds for $p \leq q$, i.e., $S_p(H) \subseteq S_q(H)$. For $p = 2$, we obtain the Hilbert space of *Hilbert–Schmidt operators* on H equipped with the inner product $\langle A_1, A_2 \rangle_{S_2(H)} = \mathrm{Tr}\,(A_1^* A_2)$. For $p = 1$, we obtain the Banach algebra of *trace class operators*. For $p = \infty$, we obtain the Banach algebra of compact operators equipped with the operator norm $\|A\| = \|A\|_{S_\infty(H)}$. The Schatten classes are the completion of finite-rank operators (i.e., operators in $\mathrm{span}\{x \otimes x' \mid x, x' \in H\}$) with respect to the corresponding norm.

We will make frequent use of the fact that the tensor product space $H \otimes H$ can be isometrically identified with the space of Hilbert–Schmidt operators on H, i.e., we have $S_2(H) \simeq H \otimes H$. For elements $x_1, x_1', x_2, x_2' \in H$, we have the relation $\langle x_2 \otimes x_1, x_2' \otimes x_1' \rangle_{H \otimes H} = \langle x_2 \otimes x_1, x_2' \otimes x_1' \rangle_{S_2(H)}$, where the tensors are interpreted as rank-one operators as described above. This property extends to $\mathrm{span}\{x \otimes x' \mid x, x' \in H\}$ by linearity and defines a linear isometric isomorphism between $H \otimes H$ and $S_2(H)$, which can be seen by considering Hilbert–Schmidt operators in terms of their singular value decompositions.

For any topological space E, we will write $\mathcal{F}_E = \mathcal{B}(E)$ for its associated Borel field. For any collection of sets $\mathcal{M}$, $\sigma(\mathcal{M})$ denotes the intersection of all σ-fields containing $\mathcal{M}$. For any σ-field $\mathcal{F}$ and countable index set I, we write $\mathcal{F}^{\otimes I}$ as the product σ-field (i.e., the smallest σ-field with respect to which all coordinate projections on E^I are measurable). Note that when E is Polish (i.e., separable and completely metrizable), we have $\mathcal{B}(E^I) = \mathcal{B}(E)^{\otimes I}$, i.e. the Borel field on the product space generated by the product topology and the product of the individual Borel fields are equal. Put differently, the Borel field operator and the product field operator are compatible with respect to product spaces (Dudley, 2002, Proposition 4.1.17). Moreover, E^I equipped with the product topology is Polish.

In this paper, we will consider a stochastic process $(X_t)_{t \in \mathbb{Z}}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the observation space $(E, \mathcal{F}_E)$, where we assume E to be Polish. For a finite number of random variables $\xi_1, \ldots, \xi_n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in E, we write $\mathcal{L}(\xi_1, \ldots, \xi_n)$ for the *finite-dimensional law*, i.e., the *pushforward measure* on $(E^n, \mathcal{B}(E^n))$ induced by $\xi_1, \ldots, \xi_n$.

Assumption 1 (Stationarity) *We assume that the process $(X_t)_{t \in \mathbb{Z}}$ is stationary in the sense that all finite-dimensional laws are identical, that is*

$$\mathcal{L}(X_{t_1}, \ldots, X_{t_r}) = \mathcal{L}(X_{t_1 + \eta}, \ldots, X_{t_r + \eta})$$

for all $t_1, \ldots, t_r \in \mathbb{Z}$, $r \in \mathbb{N}$, and time lags $\eta \in \mathbb{N}$.

For any separable Banach space B, let $L^p(\Omega, \mathcal{F}, \mathbb{P}; B)$ denote the space of strongly $\mathcal{F} - \mathcal{F}_B$ measurable and Bochner p-integrable functions $f \colon \Omega \to B$ for $1 \leq p \leq \infty$ (see for example Diestel and Uhl, 1977). In the case of $B = \mathbb{R}$, we simply write $L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}) = L^p(\mathbb{P})$ for the standard space of real-valued Lebesgue p-integrable functions.

2.2 Reproducing Kernel Hilbert Spaces

We will briefly introduce the concept of reproducing kernel Hilbert spaces. For a detailed discussion of this topic, we refer the reader to Berlinet and Thomas-Agnan (2004), Steinwart and Christmann (2008) and Saitoh and Sawano (2016). To distinguish standard Hilbert spaces from reproducing kernel Hilbert spaces, we will use the script letter $\mathscr{H}$ for the latter.

Definition 2 (Reproducing kernel Hilbert space) *Let E be a set and $\mathscr{H}$ a space of functions from E to $\mathbb{R}$. Then $\mathscr{H}$ is called a* reproducing kernel Hilbert space (RKHS) *with corresponding inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ if there exists function $k \colon E \times E \to \mathbb{R}$ such that*

(i) $\langle f, k(x, \cdot) \rangle_{\mathscr{H}} = f(x)$ for all $f \in \mathscr{H}$ (reproducing property), and

(ii) $\mathscr{H} = \overline{\operatorname{span}\{k(x, \cdot) \mid x \in E\}}$, where the completion is with respect to the RKHS norm.

We call k the reproducing kernel *of $\mathscr{H}$.*

It follows in particular that $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathscr{H}}$. The *canonical feature map* $\varphi \colon E \to \mathscr{H}$ is given by $\varphi(x) := k(x, \cdot)$. Thus, we obtain $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}}$. Every RKHS has a unique symmetric and positive semi-definite kernel k with the reproducing property. Conversely, every symmetric positive semi-definite kernel k induces a unique RKHS with k as its reproducing kernel. In what follows, we will use the term *kernel* synonymously for reproducing kernel/symmetric positive semi-definite kernel for brevity.

We now impose a few restrictions on the considered RKHS, which we assume to be fulfilled for the remainder of this paper.

Assumption 3 (Separability) *The RKHS $\mathscr{H}$ is separable. Note that for a Polish space E, the RKHS induced by a continuous kernel $k \colon E \times E \to \mathbb{R}$ is always separable (see Steinwart and Christmann, 2008, Lemma 4.33). For a more general treatment of conditions implying separability, see Owhadi and Scovel (2017).*

Assumption 4 (Measurability) *The canonical feature map $\varphi \colon E \to \mathscr{H}$ is $\mathcal{F}_E - \mathcal{F}_{\mathscr{H}}$ measurable. This is the case when $k(x, \cdot) \colon E \to \mathbb{R}$ is $\mathcal{F}_E - \mathcal{F}_{\mathbb{R}}$ measurable for all $x \in E$. If this condition holds, then additionally all functions $f \in \mathscr{H}$ are $\mathcal{F}_E - \mathcal{F}_{\mathbb{R}}$ measurable and $k \colon E \times E \to \mathbb{R}$ is $\mathcal{F}_E^{\otimes 2} - \mathcal{F}_{\mathbb{R}}$ measurable (see Steinwart and Christmann, 2008, Lemmas 4.24 and 4.25).*

Assumption 5 (Existence of second moments) *We have $\varphi(X_0) \in L^2(\Omega, \mathcal{F}, \mathbb{P}, \mathscr{H})$. Note that this is trivially the case whenever $\sup_{x \in E} k(x, x) < \infty$.*

2.3 Kernel Mean Embeddings and Kernel Covariance Operators

We now introduce kernel mean embeddings and kernel covariance operators, which are simply the Bochner expectations and covariance operators of RKHS-embedded random variables.

For a random variable X on E satisfying $\varphi(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P}; \mathscr{H})$, we call

$$\mu_X := \mu_{\mathcal{L}(X)} := \mathbb{E}[\varphi(X)] \in \mathscr{H}$$

the *kernel mean embedding* or simply *mean embedding* (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017) of X. For every $f \in \mathscr{H}$, the mean embedding satisfies $\mathbb{E}[f(X)] = \langle f, \mu_X \rangle_{\mathscr{H}}$.

Definition 6 (Kernel (cross-)covariance operator) *For two random variables X, Y on E satisfying $\varphi(X), \varphi(Y) \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathscr{H})$, we call the trace class operator $C_{YX} \colon \mathscr{H} \to \mathscr{H}$ defined by*

$$C_{YX} = \mathbb{E}[(\varphi(Y) - \mu_Y) \otimes (\varphi(X) - \mu_X)]$$

the kernel cross-covariance operator *of X and Y. We call C_{XX} the* kernel covariance operator *of X.*

For all $f, g \in \mathscr{H}$, we have $\mathrm{Cov}[f(X), g(Y)] = \langle g, C_{YX} f \rangle_{\mathscr{H}}$ as well as $C_{YX} = C_{XY}^*$. As a consequence, C_{XX} is self-adjoint, positive semi-definite and trace class. For additional information about (cross-)covariance operators of Hilbertian random variables, see for example Parthasarathy (1967), Baker (1970) and Baker (1973).

In the literature, the covariance operator is sometimes used as a generalization of the uncentered second moment and therefore defined without centering of the random variables $\varphi(X)$ and $\varphi(Y)$ (Prokhorov, 1956; Parthasarathy, 1967; Bharucha-Reid, 1972; Fukumizu et al., 2013).

Definition 7 (Kernel autocovariance operator) *Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary stochastic process with values on E such that $\varphi(X_0) \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathscr{H})$. Let $\eta \in \mathbb{N}$. We call*

$$C(\eta) := C_{X_\eta X_0} = C_{X_{t+\eta} X_t}$$

the kernel autocovariance operator *of the process $(X_t)_{t \in \mathbb{Z}}$ with respect to the time lag η.*

2.4 Product Kernels and Hilbert–Schmidt Operators

The tensor product space $\mathscr{H} \otimes \mathscr{H} \simeq S_2(\mathscr{H})$ is itself an RKHS with the corresponding canonical feature map $\varphi \otimes \varphi \colon E \times E \to \mathscr{H} \otimes \mathscr{H}$ given by

$$\varphi \otimes \varphi\,(x_1, x_2) := \varphi(x_1) \otimes \varphi(x_2).$$

For more details, we refer the reader to Steinwart and Christmann (2008, Lemma 4.6). The corresponding kernel of $\mathscr{H} \otimes \mathscr{H}$ is the product kernel $k \cdot k \colon E^2 \times E^2 \to \mathbb{R}$.

The estimation of the uncentered kernel autocovariance operator can therefore be interpreted as the estimation of a kernel mean on the product RKHS $\mathscr{H} \otimes \mathscr{H}$. In particular, we can write $C(\eta)$ as the kernel mean embedding of the joint distribution $\mathcal{L}(X_\eta, X_0)$ on the measurable space $(E \times E, \mathcal{F}_E \otimes \mathcal{F}_E)$ using the product feature map $\varphi \otimes \varphi$. That is, we have

$$\mu_{\mathcal{L}(X_\eta, X_0)} = \mathbb{E}[\varphi(X_\eta) \otimes \varphi(X_0)],$$

which is exactly the uncentered kernel autocovariance operator of X.

Thus, the analysis of the estimation of the uncentered autocovariance operator covers the problem of estimating the kernel mean $\mu_{\mathcal{L}(X_0)}$ of the marginal $\mathcal{L}(X_0)$. In fact, by considering kernels on the product space $E \times E$ instead of E, we need to account for the challenge

that appropriate statistical properties of the process $(X_t)_{t\in\mathbb{Z}}$ (such as ergodicity, mixing, or decay of correlations) have to transfer to the product process $(X_{t+\eta}, X_t)_{t\in\mathbb{Z}}$ on $E \times E$ in order to provide results. We may therefore concentrate directly on the estimation of uncentered autocovariance operators based on $E \times E$ instead on the estimation of kernel mean embeddings on E. We emphasize that all of our results directly transfer to the much simpler case of estimating the kernel mean $\mu_{\mathcal{L}(X_0)}$ from dependent data by simply replacing the product space $E \times E$ with E and the product feature map $\varphi \otimes \varphi$ with φ.

Assumption 8 (Centered process) *Without loss of generality, we assume that the embedded process is centered, i.e., $\mu_{\mathcal{L}(X_0)} = \mathbb{E}[\varphi(X_0)] = 0$. In this case, the centered and the uncentered autocovariance operator coincide:*

$$C(\eta) = \mathbb{E}[\varphi(X_\eta) \otimes \varphi(X_0)].$$

Whenever Assumption 8 is not satisfied, the centered autocovariance operator can be computed by replacing the random variables $\varphi(X_0)$ and $\varphi(X_\eta)$ by their centered counterparts $\varphi(X_0) - \mu_{\mathcal{L}(X_0)}$ and $\varphi(X_\eta) - \mu_{\mathcal{L}(X_0)}$, respectively. It is important to note that unlike $\varphi(\cdot)$, the expression $\varphi(\cdot) - \mu_{\mathcal{L}(X_0)}$ is technically not the canonical feature map of $\mathscr{H}$ and should be treated with caution. In particular, it does not satisfy the reproducing property. It does however satisfy all conditions required to formally define its mean embedding and the corresponding covariance operators. Furthermore, in practical applications, the mean embedding $\mu_{\mathcal{L}(X_0)}$ is typically not known and replaced by an empirical estimate. This centering and its consequences regarding the empirical estimation of kernel covariance operators are investigated in detail by Blanchard et al. (2007).

In what follows, we will repeatedly use the shorthand

$$C_n(\eta) := \frac{1}{n} \sum_{t=1}^{n} \varphi(X_{t+\eta}) \otimes \varphi(X_t)$$

for the empirical estimate of $C(\eta)$ based on $n + \eta$ consecutive time steps of the process $(X_t)_{t\in\mathbb{Z}}$.

3. Strong Law of Large Numbers

We now address the strong law of large numbers for the estimator $C_n(\eta)$. To this end, we briefly introduce the concept of measure-preserving dynamical systems and ergodicity. For details, the reader may refer for example to Petersen (1983). It is well-known that every stationary process can be expressed in terms of a measure-preserving dynamical system when the underlying probability space is chosen accordingly (see for example Doob, 1953, Chapter X). This allows to study stationary processes with tools from ergodic theory. We will briefly describe this viewpoint below.

It is possible to assume without loss of generality that the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ describing $(X_t)_{t\in\mathbb{Z}}$ is the *canonical probability space*, i.e., $\Omega := E^{\mathbb{Z}}$ and $\mathcal{F} := \mathcal{F}_E^{\otimes \mathbb{Z}}$. In this case, we can simply express the process $(X_t)_{t\in\mathbb{Z}}$ as the family of *coordinate projections*

on Ω: for $\omega = (\omega_t)_{t \in \mathbb{Z}} \in \Omega$, we can write

$$X_t(\omega) := \omega_t = X_0(T^t \omega) \quad \text{for all } t \in \mathbb{Z},$$

where T is the *left-shift operator* on Ω defined by $(T\omega)_i = \omega_{i+1}$ for all $i \in \mathbb{Z}$. Note that by stationarity of $(X_t)_{t \in \mathbb{Z}}$, the shift T is measure preserving in the sense that $\mathbb{P}[T^{-1}M] = \mathbb{P}[M]$ for all $M \in \mathcal{F}_E^{\otimes \mathbb{Z}}$. We call $(X_t)_{\in \mathbb{Z}}$ *ergodic* whenever T is *ergodic in the measure theoretical sense*, i.e., for all sets $M \in \mathcal{F}_E^{\otimes \mathbb{Z}}$, we have that the condition $T^{-1}M = M$ implies either $\mathbb{P}[M] = 0$ or $\mathbb{P}[M] = 1$.

We show that for any fixed time lag $\eta \in \mathbb{N}$, the kernel autocovariance operator $C(\eta)$ can be estimated almost surely from realizations of $(X_t)_{t \in \mathbb{Z}}$ whenever the process is ergodic. This comes as a natural consequence of the following generalized version of Birkhoff's ergodic theorem.

Theorem 9 (Beck and Schwartz, 1957, Theorem 6) *Let B be a reflexive Banach space and T an ergodic measure-preserving transformation on $(\Omega, \mathcal{F}, \mathbb{P})$. Then for each $f \in L^1(\Omega, \mathcal{F}, \mathbb{P}; B)$,*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(T^i \omega) = \mathbb{E}[f],$$

where the convergence holds $\mathbb{P}$-a.e. with respect to $\|\cdot\|_B$.

We can directly apply this result to obtain almost sure convergence of the empirical estimate of $C(\eta)$ in the case that $(X_t)_{t \in \mathbb{Z}}$ is ergodic.

Corollary 10 (Strong consistency) *Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary and ergodic process defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Polish space E. Then*

$$\lim_{n \to \infty} C_n(\eta) = C(\eta),$$

where the convergence is $\mathbb{P}$-a.e. with respect to $\|\cdot\|_{S_2(\mathscr{H})}$.

Proof The time-lagged product process $(X_t, X_{t+\eta})_{t \in \mathbb{Z}}$ on $E \times E$ can be expressed via the projection tuple $(X_t, X_{t+\eta})(\omega) = (X_0, X_\eta)(T^t \omega)$. By construction, (X_0, X_η) is $\mathbb{P} - \mathcal{F}_E \otimes \mathcal{F}_E$ measurable. Note that because of Assumption 4 and Assumption 5 the product feature map $\varphi \otimes \varphi$ given by $(x, y) \mapsto \varphi(y) \otimes \varphi(x)$ is an element of $L^1(E \times E, \mathcal{F}_E \otimes \mathcal{F}_E, \mathcal{L}(X_0, X_\eta); S_2(\mathscr{H}))$, where $S_2(\mathscr{H})$ is clearly reflexive. Therefore, it holds that the composition

$$\varphi \otimes \varphi \circ (X_0, X_\eta) \colon \Omega \to S_2(\mathscr{H})$$

given by $\omega \mapsto (X_0, X_\eta)(\omega) \mapsto \varphi(X_\eta) \otimes \varphi(X_0)(\omega)$ is an element of $L^1(\Omega, \mathcal{F}, \mathbb{P}; S_2(\mathscr{H}))$.

The statement follows immediately from the fact that we choose $\varphi \otimes \varphi \circ (X_0, X_\eta)$ as the observable f in Theorem 9 and obtain

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \varphi \otimes \varphi \circ (X_0, X_\eta) \circ T^t = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \varphi(X_{t+\eta}) \otimes \varphi(X_t) = C(\eta),$$

where the convergence is $\mathbb{P}$-a.e. in $S_2(\mathscr{H})$. ∎

Remark 11 (Convergence in Schatten norms) *Corollary 10 also yields $\mathbb{P}$-a.e. convergence $C_n(\eta) \to C(\eta)$ in $S_p(\mathscr{H})$ for all $p \geq 2$, as it holds that $\|\cdot\|_{S_p(\mathscr{H})} \leq \|\cdot\|_{S_2(\mathscr{H})}$. Note that $S_1(\mathscr{H})$ is reflexive if and only if $\mathscr{H}$ is finite-dimensional (see for example Simon, 2005, Theorem 3.2). However, in the finite-dimensional case, all Schatten classes coincide and the question for convergence in Schatten norms becomes trivial. In the general case, it is not clear whether the reflexivity assumption in Theorem 9 is not only sufficient but also necessary for a convergence to hold. To the best of our knowledge, no stronger generalization results of Birkhoff's ergodic theorem for Banach-valued random variables exist.*

4. Asymptotic Error Behavior

In this section, we apply limit theorems for weakly dependent Hilbertian random variables to the estimation error $C_n(\eta) - C(\eta)$ and investigate its asymptotic behaviour.

4.1 α-Mixing

We first recall the basic notions of α-mixing in statistics (see for example Bradley, 2005).

Definition 12 (α-mixing coefficient) *For σ-fields $\mathcal{F}_1$, $\mathcal{F}_2 \subseteq \mathcal{F}$, we define*

$$\alpha(\mathcal{F}_1, \mathcal{F}_2) := \sup_{A \in \mathcal{F}_1, B \in \mathcal{F}_2} |\mathbb{P}[A \cap B] - \mathbb{P}[A]\mathbb{P}[B]|.$$

For a process $(X_t)_{t \in \mathbb{Z}}$, we furthermore define

$$\alpha(n) := \alpha((X_t)_{t \in \mathbb{Z}}, n) := \begin{cases} \sup_{j \in \mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^j, \mathcal{F}_{j+n}^\infty), & n \geq 1, \\ 1/4, & n < 1, \end{cases}$$

where $\mathcal{F}_l^m := \sigma(X_t, l \leq t \leq m)$ denotes the σ-field generated by the process $(X_t)_{t \in \mathbb{Z}}$ for time horizons $-\infty \leq l \leq m \leq \infty$.

The process is called *α-mixing* or *strongly mixing*, when $\alpha(n) \to 0$ as $n \to \infty$. In this case, the convergence rate of $\alpha(n)$ is called the *mixing rate* of the associated process. In this paper, we will not focus on the various alternative strong mixing coefficients which are frequently used in statistics (Doukhan, 1994; Bradley, 2005; Rio, 2017), since α-mixing is the weakest concept among the strong mixing coefficients and covers a wide range of processes in practice. The coefficient $\alpha(n)$ is typically defined only for $n \geq 1$ in the literature. However, since it always holds that $0 \leq \alpha(\mathcal{F}_1, \mathcal{F}_2) \leq 1/4$, our definition above extends $\alpha(n)$ trivially to $n \in \mathbb{Z}$. This allows us to formulate several results in a more convenient way in what follows.

Remark 13 (Terminology) *The concept of strong mixing coefficients is typically much stronger than the strong mixing considered in ergodic theory (Petersen, 1983). Also note that we define the α-mixing coefficient for stationary processes. It can also be defined for nonstationary processes, while mixing in the ergodic theoretical sense typically arises from dynamical systems induced by measure-preserving transformations and is therefore primarily used in the context of stationary stochastic processes.*

Example 1 (Mixing processes) *A wide range of mixing processes are given by Doukhan (1994) and Bradley (2005). We list some important examples here.*

1. *Irreducible and aperiodic stationary Markov processes on $E \subseteq \mathbb{R}$ are α-mixing (in fact, a stronger mixing property called β-mixing or absolute regularity holds, see for example Bradley, 2005, Corollary 3.6).*

2. *Stationary Markov processes satisfying geometric ergodicity (for details see Meyn and Tweedie, 2009, Chapter 15) are α-mixing with $\alpha(n) = O(\exp(-cn))$ for some $c > 0$ (Bradley, 2005, Theorem 3.7).*

3. *We consider a stochastic dynamical system $(X_t)_{t \in \mathbb{N}_0}$ also known as the nonlinear state space model (Meyn and Tweedie, 2009, Chapter 7) given by the recursion*

$$X_t = h_t(X_{t-1}, Z_t), \quad t \geq 1,$$

   where $h_t \colon E \to E$ are measurable functions and Z_t are i.i.d. random variables on E which are independent of X_0. The process $(X_t)_{t \in \mathbb{N}_0}$ is a Markov process (see for example Kallenberg, 2002, Proposition 8.6). Therefore 1. and 2. apply in this case. General conditions under which this system is geometrically ergodic (i.e., geometrically α-mixing in the sense of 2.) are presented by Doukhan (1994, Section 2.4). As a very basic example, the so-called simple linear model on $E \subseteq \mathbb{R}$ given by

$$X_t = aX_{t-1} + Z_t, \quad a \in \mathbb{R}, t \geq 1,$$

   is geometrically ergodic if $|a| < 1$ and the random variables Z_t are integrable and admit an everywhere positive density on E (Meyn and Tweedie, 2009, Section 15.5.2).

4. *Under some requirements, commonly used linear and nonlinear process models on finite-dimensional vector spaces including AR, ARMA, ARCH, and GARCH are α-mixing with $\alpha(n) = O(\exp(-cn))$ for some $c > 0$, see Doukhan (1994, Section 2.4) and Fan and Yao (2003, Section 2.6.1).*

We make use of the following classical lemma which we prove for completeness. It ensures that the mixing rates of measurable transformations of a stationary process process are at least as fast as the mixing rates of the original process.

Lemma 14 (Mixing coefficients of transformed processes) *Let $(X_t)_{t \in \mathbb{Z}}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ be a process with values in the Polish space $(E, \mathcal{F}_E)$ equipped with its Borel field. Let $(F, \mathcal{F}_F)$ be another Polish space equipped with its Borel field. Let $\eta \in \mathbb{N}$ and $h \colon E^{\eta+1} \to F$ be a $\mathcal{F}_E^{\otimes \eta+1} - \mathcal{F}_F$ measurable transformation. Then for the F-valued process $(H_t)_{t \in \mathbb{Z}}$ given by*

$$H_t := h(X_t, \ldots, X_{t+\eta}), \quad t \in \mathbb{Z},$$

we have

$$\alpha((H_t)_{t \in \mathbb{Z}}, n) \leq \alpha((X_t)_{t \in \mathbb{Z}}, n - \eta) \tag{2}$$

for all $n \in \mathbb{Z}$. In particular, if $(X_t)_{t \in \mathbb{Z}}$ is α-mixing, then $(H_t)_{t \in \mathbb{Z}}$ is α-mixing with the same mixing rate as $(X_t)_{t \in \mathbb{Z}}$ or faster.

Proof It is sufficient to consider the case $n > \eta$, as (2) is trivial whenever $n \leq \eta$ by our definition of the α-mixing coefficients. Let $\mathcal{H}_l^m := \sigma(H_t, l \leq t \leq m) \subseteq \mathcal{F}$ be the σ-field

generated by $(H_t)_{t\in\mathbb{Z}}$ for time horizons l and m. By construction, for all $j \in \mathbb{Z}$ we have $\mathcal{H}^j_{-\infty} \subseteq \mathcal{F}^{j+\eta}_{-\infty}$ as well as $\mathcal{H}^\infty_j \subseteq \mathcal{F}^\infty_j$ for all $j \in \mathbb{Z}$. Therefore, we have

$$\alpha((H_t)_{t\in\mathbb{Z}}, n) = \sup_{j\in\mathbb{Z}} \alpha(\mathcal{H}^j_{-\infty}, \mathcal{H}^\infty_{j+n}) \leq \sup_{j\in\mathbb{Z}} \alpha(\mathcal{F}^{j+\eta}_{-\infty}, \mathcal{F}^\infty_{j+n}) = \alpha((X_t)_{t\in\mathbb{Z}}, n - \eta),$$

proving the claim. ∎

4.2 Limit Theorems

We now investigate the asymptotic statistical behavior of the estimator $C_n(\eta)$ under the assumption that $(X_t)_{t\in\mathbb{Z}}$ is strongly mixing. For brevity, we introduce the shorthand notation

$$\xi_t := \big(\varphi(X_{t+\eta}) \otimes \varphi(X_t)\big) - C(\eta) \tag{3}$$

for $t \in \mathbb{Z}$ and $\eta \in \mathbb{N}$. The process $(\xi_t)_{t\in\mathbb{Z}}$ is stationary and centered with values in $S_2(\mathscr{H})$. The estimation error can now be expressed as $C_n(\eta) - C(\eta) = \frac{1}{n}\sum_{t=1}^n \xi_t$.

Lemma 15 (Mixing process in product RKHS) *Let $\eta \in \mathbb{N}$ be the time lag used to define the $S_2(\mathscr{H})$-valued process $(\xi_t)_{t\in\mathbb{Z}}$ given in by (3). Then the α-mixing coefficients of of $(\xi_t)_{t\in\mathbb{Z}}$ satisfy*

$$\alpha((\xi_t)_{t\in\mathbb{Z}}, n) \leq \alpha((X_t)_{t\in\mathbb{Z}}, n - \eta) \quad \text{for all } n \in \mathbb{Z}. \tag{4}$$

In particular, if the process $(X_t)_{t\in\mathbb{Z}}$ is α-mixing, then $(\xi_t)_{t\in\mathbb{Z}}$ is α-mixing with the same rate as $(X_t)_{t\in\mathbb{Z}}$ or faster.

Proof It is sufficient to note that $(\xi_t)_{t\in\mathbb{Z}}$ is obtained from $(X_t,\ldots,X_{t+\eta})_{t\in\mathbb{Z}}$ via a measurable transformation. The assertion follows from Lemma 14. ∎

We will make use of the fact that whenever the kernel k is bounded, the process $(\xi_t)_{t\in\mathbb{Z}}$ is almost surely bounded. In particular, if $\sup_{x\in E} k(x,x) = c < \infty$, then for all $t \in \mathbb{Z}$, we have

$$\begin{aligned}
\|\xi_t\|_{L^\infty(\Omega,\mathcal{F},\mathbb{P};S_2(\mathscr{H}))} &\leq \operatorname*{ess\,sup}_{\omega\in\Omega} \|\varphi(X_{t+\eta}) \otimes \varphi(X_t)\|_{S_2(\mathscr{H})} + \|C(\eta)\|_{S_2(\mathscr{H})} \\
&\leq \sup_{x\in E} \|\varphi(x)\|^2_{\mathscr{H}} + \mathbb{E}\left[\|\varphi(X_{t+\eta}) \otimes \varphi(X_t)\|_{S_2(\mathscr{H})}\right] \\
&\leq 2\sup_{x\in E} \|\varphi(x)\|^2_{\mathscr{H}} = 2\sup_{x\in E} k(x,x) = 2c.
\end{aligned} \tag{5}$$

Several properties of the estimation error $C_n(\eta) - C(\eta)$ can be proven by applying results from the asymptotic theory of weakly dependent Hilbertian processes to $(\xi_t)_{t\in\mathbb{Z}}$. We begin with one of the strongest results of this type which is an approximation of the rescaled estimation error $n(C_n(\eta) - C(\eta))$ by a Gaussian process.

To this end, we briefly recall the definition of a Gaussian measure on a Hilbert space (see for example Bogachev, 1998), which generalizes the concept of the finite-dimensional

normal distribution. A probability measure ν on a Hilbert space H equipped with its Borel field is called a *Gaussian measure* if its characteristic function $\tilde{\nu}: H \to \mathbb{C}$ defined by $\tilde{\nu}(x) := \int_H \exp(i \langle x, y \rangle_H) \, \mathrm{d}\nu(y)$ is of the form

$$\tilde{\nu}(x) = \exp\left(\langle \mu, x \rangle_H - \frac{1}{2} \langle x, Tx \rangle_H \right) \tag{6}$$

for some $\mu \in H$ (the *mean* of ν) and positive-semidefinite self-adjoint operator $T \in S_1(H)$ (the *covariance operator* of ν, see Bogachev, 1998, Theorem 2.3.1). If ν satisfies (6), we also write $\nu = \mathcal{N}(\mu, T)$. We note that the covariance operator T should not be confused with the kernel autocovariance operator $C(\eta)$, as T describes a limiting covariance structure of the sequence of laws associated with the empirical estimates of $C(\eta)$ in what follows.

We now introduce the shorthand notation $L(n) := \max(\log n, 1)$ for $n \in \mathbb{N}$.

Theorem 16 (Almost sure invariance principle) *Let $(X_t)_{t \in \mathbb{Z}}$ be stationary and α-mixing with coefficients $(\alpha(t))_{t \in \mathbb{Z}}$ such that $\sum_{t=1}^\infty \alpha(t) < \infty$. Furthermore, let $\sup_{x \in E} k(x, x) < \infty$. Let $\eta \in \mathbb{N}$ be the time lag used to define the $S_2(\mathscr{H})$-valued process $(\xi_t)_{t \in \mathbb{Z}}$ given in by (3). Then the linear operator $T_\eta: S_2(\mathscr{H}) \to S_2(\mathscr{H})$ defined by*

$$T_\eta := \mathbb{E}[\xi_0 \otimes \xi_0] + \sum_{t=1}^\infty \mathbb{E}[\xi_0 \otimes \xi_t] + \sum_{t=1}^\infty \mathbb{E}[\xi_t \otimes \xi_0] \tag{7}$$

is trace class. Furthermore, there exists a Gaussian measure $\mathcal{N}(0, T_\eta)$ on $S_2(\mathscr{H})$ and a sequence of i.i.d. $S_2(\mathscr{H})$-valued Gaussian random variables $(Z_t)_{t \in \mathbb{Z}} \sim \mathcal{N}(0, T_\eta)$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that we have $\mathbb{P}$-a.e.

$$\left\| n(C_n(\eta) - C(\eta)) - \sum_{t=1}^n Z_t \right\|_{S_2(\mathscr{H})} = o\left(\sqrt{nL(L(n))} \right).$$

Proof The assumptions ensure that $(\xi_t)_{t \in \mathbb{Z}}$ is $\mathbb{P}$-a.e. bounded by (5) and has summable mixing coefficients by (4). We can directly apply the almost sure invariance principle from Dedecker and Merlevède (2010, Corollary 1) to $(\xi_t)_{t \in \mathbb{Z}}$, which yields the assertion. We emphasize that the $\mathbb{P}$-a.e. boundedness of $(\xi_t)_{t \in \mathbb{Z}}$ and the summability of the mixing coefficients ensure that this result can be applied, as explicitly mentioned by the authors (see also Merlèvede, 2008, Remark 3). ∎

A strongly related statement is a standard central limit theorem for weakly dependent sequences which ensures asymptotic normality in the space $S_2(\mathscr{H})$.

Theorem 17 (Central limit theorem) *Let $\eta \in \mathbb{N}$ be a fixed time lag. Under the assumptions of Theorem 16, the laws of the sequence $\sqrt{n}(C_n(\eta) - C(\eta))$ converge weakly to a Gaussian measure $\mathcal{N}(0, T_\eta)$ on $S_2(\mathscr{H})$ with covariance operator T_η defined by (7).*

Proof By our previous analysis and the argumentation in the proof of Theorem 16, the process $(\xi_t)_{t \in \mathbb{Z}}$ satisfies all assumptions of the central limit theorem by Merlevède et al. (1997, Corollary 1). The above assertions follow directly. ∎

The next result is a compact law of the iterated logarithm. It ensures that an appropriately rescaled version of the estimation error approximates a compact limiting set almost surely. Additionally, it characterizes this set as the accumulation points of the estimation error sequence and gives a norm bound in $S_2(\mathscr{H})$ depending on the mixing rate. We obtain our result by applying an infinite-dimensional generalization of the classical law of the iterated logarithm for real-valued random variables. The original law of the iterated logarithm plays an important role in sequential hypothesis testing (Robbins, 1970; Siegmund, 1985), leading to potential applications in reinforcement learning as noted by Kaufmann et al. (2016).

Let $\mathrm{acc}(x_n) \subseteq X$ denote the set of all *accumulation points* of a sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space (X, d) and $\mathrm{dist}(x, A) := \inf\{d(x, y) \,|\, y \in A\}$ be the *distance* between a point $x \in X$ and a set $A \subseteq X$.

Theorem 18 (Compact law of the iterated logarithm) *Let the process $(X_t)_{t\in\mathbb{Z}}$ be stationary and α-mixing with coefficients $(\alpha(t))_{t\in\mathbb{Z}}$ such that $\sum_{t=1}^{\infty} \alpha(t) < \infty$. Furthermore, let $\sup_{x\in E} k(x, x) = c < \infty$. Then for every time lag $\eta \in \mathbb{N}$ there exists a compact, convex and symmetric set $K_\eta \subseteq S_2(\mathscr{H})$, such that $\mathbb{P}$-a.e.*

$$\lim_{n\to\infty} \mathrm{dist}\left(\frac{\sqrt{n}(C_n(\eta) - C(\eta))}{\sqrt{2L(L(n))}}, K_\eta \right) = 0 \tag{8}$$

as well as $\mathbb{P}$-a.e.

$$\mathrm{acc}\left(\frac{\sqrt{n}(C_n(\eta) - C(\eta))}{\sqrt{2L(L(n))}} \right) = K_\eta. \tag{9}$$

Moreover, whenever $\sum_{t=1}^{\infty} \alpha(t - \eta) = M < \infty$, we have

$$\sup_{A\in K_\eta} \|A\|_{S_2(\mathscr{H})} = (4c^2 + 32\, c^2 M)^{1/2}. \tag{10}$$

For the proof of Theorem 18, see Appendix A. We note that from (9), we can deduce

$$\limsup_{n\to\infty} \frac{\sqrt{n}\, \|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})}}{\sqrt{2L(L(n))}} = \sup_{A\in K_\eta} \|A\|_{S_2(\mathscr{H})},$$

which gives us the rate of $\mathbb{P}$-a.e. convergence of the estimation error.

Corollary 19 (Rate of convergence) *Under the assumptions of Theorem 18, we have*

$$\|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})} = O\left(\frac{\sqrt{2L(L(n))}}{\sqrt{n}} \right) \quad \mathbb{P}\text{-a.e.}$$

for every time lag $\eta \in \mathbb{N}$.

In particular, Corollary 19 shows that boundedness of the kernel and summability of the mixing coefficients are sufficient to obtain the rate of convergence which is known to hold in the strong law of large numbers for independent random variables as given by the classical law of the iterated logarithm (see for example de Acosta, 1983).

14

Remark 20 (Optimality of mixing rate assumptions) *The results in this section require that the α-mixing coefficients of $(X_t)_{t \in \mathbb{Z}}$ are summable. We now briefly address the question whether similar asymptotic statements can be derived under less strict assumptions. Merlevède et al. (1997), Merlèvede (2008) and Dedecker and Merlèvede (2010) use more general and much more technical quantile conditions than the summability of the mixing coefficients in order to prove the asymptotic results which we apply here. These quantile conditions are known to be necessary for a central limit theorem to hold for real-valued processes. We refer the reader to Doukhan et al. (1994, Section 4) for additional information. For bounded random variables however, the summability of the mixing coefficients is equivalent to the mentioned quantile conditions. This is investigated by Rio (1995, Application 1). A similar argument can be used for the law of the iterated logarithm (see also Rio, 1995).*

5. Concentration Bounds

In addition to the previous asymptotic results, concentration properties for the estimation error can be derived by using concentration properties of mixing Hilbertian processes. We now introduce the covariance operator of the $S_2(\mathscr{H})$-valued random variable ξ_t defined in (3), which we will denote by $\Gamma_\eta \in S_1(S_2(\mathscr{H}))$. The eigendecomposition of Γ_η allows to quantify the second moment of the empirical kernel autocovariance operators $C_n(\eta)$, which is needed to derive concentration bounds. We emphasize that the operator Γ_η should not be confused with the operator T_η introduced in (7), which describes the *asymptotic covariance* of $\sqrt{n}(C_n(\eta) - C(\eta))$, while Γ_η describes the variance of the marginal $\mathcal{L}(\xi_0)$ of the stationary process $(\xi_t)_{t \in \mathbb{Z}}$.

Theorem 21 (Error bound) *Let $(X_t)_{t \in \mathbb{Z}}$ be stationary with α-mixing coefficients $(\alpha(t))_{t \in \mathbb{Z}}$ Let furthermore $\sup_{x \in E} k(x, x) = c < \infty$. Then for every time lag $\eta \in \mathbb{N}$, $\epsilon > 0$, $\nu \in \mathbb{N}$, $n \geq 2$ as well as $q \in \{1, \ldots, \lfloor n/2 \rfloor\}$ and $\delta \in (0, 1)$, we have*

$$
\mathbb{P}\Big[\|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})} > \epsilon\Big] \leq 4\nu \exp\left(-\frac{(1-\delta)\epsilon^2 q}{32\nu c^2}\right)
$$

$$
+ 22\nu q \left(1 + \frac{8c\sqrt{\nu}}{\epsilon(1-\delta)^{1/2}}\right)^{1/2} \alpha(\lfloor n/2q \rfloor - \eta) + \frac{1}{\delta\epsilon^2} \sum_{j > \nu} \lambda_j,
$$

where the nonnegative real numbers $(\lambda_j)_{j \geq 1}$ are the nonincreasingly ordered eigenvalues of the covariance operator

$$
\Gamma_\eta \colon S_2(\mathscr{H}) \to S_2(\mathscr{H})
$$

defined by

$$
\Gamma_\eta := \mathbb{E}\left[\Big((\varphi(X_\eta) \otimes \varphi(X_0)) - C(\eta)\Big) \otimes \Big((\varphi(X_\eta) \otimes \varphi(X_0)) - C(\eta)\Big)\right]. \tag{11}
$$

Proof As previously noted, the process $(\xi_t)_{t \in \mathbb{Z}}$ defined by (3) with the time lag $\eta \in \mathbb{N}$ is stationary, centered and almost surely bounded by $2c$ in the norm of $S_2(\mathscr{H})$. Moreover, its α-mixing coefficients satisfy the bound (4). We can therefore apply the concentration bound given by Bosq (2000, Theorem 2.12) to the process $(\xi_t)_{t \in \mathbb{Z}}$, which yields the assertion. ∎

The above bound requires an optimal trade-off between ν, q, and δ. In particular, knowledge about the decay of the eigenvalues $(\lambda_j)_{j\geq 1}$ of Γ_η allows to choose ν, q and δ such a way that the bound can be simplified as demonstrated by Bosq (2000, Corollary 2.4). We also note that $\sum_{j>\nu} \lambda_j < \infty$ for every $\nu \in \mathbb{N}$, since Γ_η is trace class.

6. Consistency of Weakly Dependent Kernel PCA

By considering the kernel covariance operator $C := C(0)$, we can easily obtain consistency results for kernel PCA (Schölkopf et al., 1998) for the case that the data is dependent. It is well known that kernel PCA approximates the spectral decomposition of C (see for example Blanchard et al., 2007), as we will briefly explain in Section 6.2. Consistency results for kernel PCA from independent data have been obtained by considering the spectral perturbation of covariance operators of Hilbertian random variables (Mas and Menneteau, 2003; Blanchard et al., 2007; Mas and Ruymgaart, 2015; Koltchinskii and Lounici, 2016, 2017; Reiß and Wahl, 2020). Various approaches exist in this context and we do not aim to provide a full overview here. Instead, we will show how our previous results lead to some elementary consistency statements for dependent data. By applying techniques from the previously mentioned literature, these results may be refined and extended accordingly.

We note that convergence in measure and weak convergence of standard linear Hilbertian PCA for $L^2([0,1])$-valued stochastic processes was previously investigated by Kokoszka and Reimherr (2013) under the assumption of L^4-m approximability.

6.1 Notation

For a compact self-adjoint positive-semidefinite operator C on $\mathscr{H}$, let $(\lambda_i(C))_{i\in I}$ denote the nonzero eigenvalues of C ordered nonincreasingly repeated with their multiplicities for the index set $I = \{1, 2, \dots\}$. Then C admits the spectral decomposition

$$C = \sum_{i\in I} \lambda_i(C)\, v_i \otimes v_i \tag{12}$$

where the v_i are the orthonormal eigenfunctions of C. In addition, let $(\mu_j(C))_{j\in J}$ denote the *distinct* eigenvalues of C for $J = \{1, 2, \dots\}$ with $\triangle_j(C) := \{i \in I \mid \lambda_i(C) = \mu_j\}$ as well as the multiplicity $m_j(C) := |\triangle_j(C)|$. Note that C can also be written as

$$C = \sum_{j\in J} \mu_j(C) P_j(C),$$

were $P_j(C)$ is the orthogonal spectral projector onto the eigenspace corresponding to $\mu_j(C)$ and the convergence is with respect to the operator norm. We will additionally consider the *spectral gap*

$$g_j(C) := \begin{cases} \mu_1(C) - \mu_2(C), & j = 1, \\ \min\{\mu_{j-1}(C) - \mu_j(C),\, \mu_j(C) - \mu_{j+1}(C)\}, & j \geq 2. \end{cases}$$

Note that $g_j(C) \neq 0$ by construction.

6.2 Operator Interpretation of Kernel PCA

Kernel PCA approximates a finite-rank truncation of the *Karhunen–Loève transformation* of the embedded random variable $\varphi(X_0)$ by approximating the spectral decomposition of the kernel covariance operator $C = \mathbb{E}[\varphi(X_0) \otimes \varphi(X_0)]$ (see for example Blanchard et al., 2007).

Consider the spectral decomposition (12) of C. Let $\{\tilde{v}_i\}_{i \geq 1}$ be an extension of the eigenfunctions $\{v_i\}_{i \in I}$ of C to a complete orthonormal system in $\mathscr{H}$ (that is, the addition of an appropriate ONS spanning the null space of C) By expanding the random variable $\varphi(X_0)$ in terms of $\{\tilde{v}_i\}_{i \geq 1}$, we get the Karhunen–Loève transformation

$$\varphi(X_0) = \sum_{i \in I} \langle \varphi(X_0), \tilde{v}_i \rangle_{\mathscr{H}} \, \tilde{v}_i = \sum_{i \in I} Z_i \tilde{v}_i, \tag{13}$$

where $Z_i := \langle \varphi(X_0), \tilde{v}_i \rangle_{\mathscr{H}} = \tilde{v}_i(X_0)$ are real-valued random variables and convergence in (13) is with respect to the norm of $\mathscr{H}$. Note that we have

$$\mathrm{Cov}[Z_i, Z_j] = \mathrm{Cov}[(\tilde{v}_i(X_0), \tilde{v}_j(X_0)] = \langle \tilde{v}_i, C\tilde{v}_j \rangle_{\mathscr{H}} = \lambda_i(C)\delta_{ij},$$

where we extend the set of eigenvalues to the null space, i.e., we set $\lambda_i(C) := 0$ for $i \neq I$. In practice, the data is usually projected onto the first r dominant eigenfunctions in order to obtain an optimal low-dimensional approximation of $\varphi(X_0)$. In particular, for all $r \in I$, the projector $P_{\leq r} := \sum_{i=1}^{r} v_i \otimes v_i$ minimizes the reconstruction error

$$R(T) := \mathbb{E} \left[\|\varphi(X_0) - T\varphi(X_0)\|_{\mathscr{H}}^2 \right] \tag{14}$$

over all operators T in the set of r-dimensional orthogonal projectors on $\mathscr{H}$. By performing a spectral decomposition of the empirical kernel covariance operator

$$C_n = \frac{1}{n} \sum_{t=1}^{n} \varphi(X_t) \otimes \varphi(X_t),$$

kernel PCA aims to approximate (13) (or $P_{\leq r}$ respectively). We are therefore interested in how well the spectral decomposition of the empirical operator C_n approximates the spectral decomposition of C.

6.3 Consistency Results

We can now combine typical results from spectral perturbation theory with our previous error analysis for C_n to obtain consistency statements. Note that we do not aim to provide a full analysis but rather illustrate how our results can be used to assess the error of kernel PCA with weakly dependent data. In the independent case, stronger results have been obtained for example by Koltchinskii and Lounici (2016, 2017), Milbradt and Wahl (2020) and Reiß and Wahl (2020) by directly considering (14).

Remark 22 (Measurability of spectral properties) *As for example shown by Dauxois et al. (1982), the eigenvalues and corresponding eigenprojection operators of C and C_n are measurable and therefore random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

Theorem 23 (Spectral perturbation bounds) *With the notation of Section 6.1, it holds that*

$$\sup_{i \geq 1} |\lambda_i(C) - \lambda_i(C_n)| \leq \|C - C_n\| \quad \mathbb{P}\text{-}a.e.$$

as well as

$$\|P_j(C) - P_j(C_n)\| \leq \frac{4 \|C - C_n\|}{g_j(C)} \quad \mathbb{P}\text{-}a.e.$$

for all $j \in J$.

See Gohberg and Kreĭn (1969, Corollary 2.3) and Koltchinskii and Lounici (2016, Lemma 1) for proofs of these statements.

The above bounds combined with the strong law of large numbers from Corollary 10 for $\|C - C_n\|$ yield consistency results of kernel PCA with weakly dependent data.

Corollary 24 (Spectral consistency & convergence rate) *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary and ergodic. Then kernel PCA is strongly consistent in the sense that we have spectral convergence $\sup_{i\geq 1} |\lambda_i(C) - \lambda_i(C_n)| \to 0$ $\mathbb{P}$-a.e. as well as $\|P_j(C_n) - P_j(C)\| \to 0$ $\mathbb{P}$-a.e. for all $j \geq 1$. In both cases, convergence takes place with at least the same rate as the convergence $C_n \to C$ in operator norm.*

Remark 25 *The preservation of convergence rates in Corollary 24 is particularly relevant whenever the assumptions of Corollary 19 hold. In this situation, the spectral convergence rate is given by Corollary 24. We note that these results are by no means optimal, as they do not consider the full reconstruction error of a finite-rank truncation of* (13), *like for example Reiß and Wahl (2020) in the independent case. Stronger results can be obtained by accessing deeper perturbation results (see for example Yu et al., 2015; Jirak and Wahl, 2020) and are not in the scope of this work.*

Whenever the estimation error $\|C - C_n\|$ can be bounded in probability (for example by applying Theorem 21) corresponding statements hold for the eigenvalues and spectral projectors as a result of Theorem 23.

Corollary 26 (Spectral concentration) *Let $\mathbb{P}\left[\|C_n(0) - C(0)\| \geq \epsilon\right] \leq f(\epsilon, n)$ for some function $f : \mathbb{R}_{>0} \times \mathbb{N} \to \mathbb{R}_{\geq 0}$. Then we have*

1. $\mathbb{P}\left[\sup_{i\geq 1} |\lambda_i(C) - \lambda_i(C_n)| \geq \epsilon\right] \leq f(\epsilon, n)$ *and*

2. $\mathbb{P}[\|P_j(C) - P_j(C_n)\| \geq \epsilon] \leq f(\frac{g_j(C)}{4}\epsilon, n)$ *for all j.*

Remark 27 *We note that in the case of weakly dependent data, the representation* (13) *might not always be a desirable model since time-related information in the realization of the process $(\varphi(X_t)_{t\in\mathbb{Z}})$ is discarded. As such, kernel PCA decomposes the RKHS only with respect to the covariance of $\mathcal{L}(\varphi(X_0))$ instead of using autocovariance information from $\mathcal{L}(\varphi(X_{t_1}), \varphi(X_{t_2}), \varphi(X_{t_3}), \dots)$. If one is interested in performing a decomposition that captures the* dynamic behavior *instead of only the* asymptotic spatial behavior, *different approaches are needed. In the context of* functional data analysis, *the concept of* harmonic PCA *or* dynamic PCA *(Panaretos and Tavakoli, 2013; Hörmann et al., 2015) yields optimal*

filter functions to reduce the dimensionality of a stationary stochastic process. We will address alternative time-based decomposition approaches in Section 8.

7. Conditional Mean Embedding of Stationary Time Series

We will now show how the previous theoretical results can be used to obtain consistency results for a large family of nonparametric time series models. A wide variety of kernel techniques for sequential data rely on the RKHS embedding of the conditional η-time step transition probability

$$\mathbb{P}[X_{t+\eta} \in \mathcal{A} \mid X_t], \quad \mathcal{A} \in \mathcal{F}_E, \quad \eta \in \mathbb{N}_{>0}, \tag{15}$$

which is modeled in terms of the *conditional mean embedding* (Song et al., 2009). In what follows, we will briefly outline the different derivations of the conditional mean embedding.

Applications of the conditional mean embedding in the context of sequential data include, among others, state-space models and filtering (Song et al., 2009; Fukumizu et al., 2013; Gebhardt et al., 2019), the embedding of transition probability models (Song et al., 2010; Grünewälder et al., 2012b; Nishiyama et al., 2012; Sun et al., 2019), predictive state representations (Boots et al., 2013), and reinforcement learning models (van Hoof et al., 2015, 2017; Stafford and Shawe-Taylor, 2018; Gebhardt et al., 2018).

7.1 Operator-theoretic Conditional Mean Embedding

In order to express the transition probability (15) in terms of the RKHS $\mathscr{H}$, one is interested in a conditional mean operator $U_\eta \colon \mathscr{H} \supseteq \mathrm{dom}(U_\eta) \to \mathscr{H}$ which satisfies

$$\langle f, U_\eta \varphi(x) \rangle_{\mathscr{H}} = \mathbb{E}[f(X_{t+\eta}) \mid X_t = x], \quad f \in \mathscr{H}. \tag{16}$$

Note that the action of U_η on $\varphi(x) \in \mathscr{H}$ is interpreted as conditioning on the event $\{X_t = x\}$, while evaluations of functions $f \in \mathscr{H}$ with $U_\eta \varphi(x)$ under the inner product can be interpreted as a conditional expectation operator in a weak sense. It is important to note that such an operator U_η does not exist in general. By using properties of the kernel covariance operators, it can be shown that $U_\eta := C(\eta)C(0)^\dagger$ satisfies (16) under strong technical assumptions (see Klebanov et al., 2020 for details). We call U_η the *conditional mean operator* and $U_\eta \varphi(x)$ the *conditional mean embedding* of the transition probability $\mathbb{P}[X_{t+\eta} \in \mathcal{A} \mid X_t = x]$. Here, $C(0)^\dagger \colon \mathrm{range}(C(0)) \oplus \mathrm{range}(C(0))^\perp \to \mathscr{H}$ is the *Moore–Penrose pseudoinverse* of the operator $C(0)$ (see for example Engl et al., 1996). Note that U_η is in general not globally defined and bounded, i.e., $\mathrm{range}(C(0)) \oplus \mathrm{range}(C(0))^\perp \neq \mathscr{H}$, since $\mathrm{range}(C(0))$ is generally not closed. Song et al. (2009) propose the regularized conditional mean operator

$$U_\eta^{(\gamma)} := C(\eta)\left(C(0) + \gamma I_{\mathscr{H}}\right)^{-1}, \tag{17}$$

with the empirical estimate $U_\eta^{(\gamma,n)} := C_n(\eta)\left(C_n(0) + \gamma I_{\mathscr{H}}\right)^{-1}$. Here, $I_{\mathscr{H}}$ denotes the identity operator on $\mathscr{H}$ and $\gamma > 0$ is a regularization parameter. Note that $U_\eta^{(\gamma)}$ as well as $U_\eta^{(\gamma,n)}$ are always well-defined Hilbert–Schmidt operators on $\mathscr{H}$. Song et al. (2009), Fukumizu

et al. (2013), and Fukumizu (2017) examine convergence of this estimate for the case of independent data pairs from the joint distribution of X_t and $X_{t+\eta}$ and show weak consistency with different rates under various technical assumptions. We extend these results to the case of dependent data.

Since the assumptions for this operator-theoretic framework and especially the analytical existence of U_η are hard to verify, different interpretations have emerged. In settings where (16) does not have an analytical solution, the regularized estimate $U_\eta^{(\gamma,n)}$ minimizes an empirical risk functional, which we will briefly outline below.

7.2 Least-squares Conditional Mean Embedding.

In cases when U_η is not globally defined and bounded, it is natural to approximate a smooth solution to (16) by minimizing the regularized risk functional $\mathcal{E}_\eta^{(\gamma)} \colon S_2(\mathscr{H}) \to \mathbb{R}$ given by

$$\mathcal{E}_\eta^{(\gamma)}(A) := \sup_{\substack{f \in \mathscr{H} \\ \|f\|_{\mathscr{H}}=1}} \mathbb{E}\left[(\mathbb{E}[f(X_{t+\eta}) \mid X_t] - \langle f, A\varphi(X_t)\rangle_{\mathscr{H}})^2 \right] + \gamma \|A\|_{S_2(\mathscr{H})}^2, \qquad (18)$$

where $\gamma > 0$ is a regularization parameter. As first shown by Grünewälder et al. (2012a) and recently investigated by Park and Muandet (2020) and Mollenhauer and Koltai (2020), the risk $\mathcal{E}_\eta^{(\gamma)}(A)$ can be bounded from above by the *surrogate risk*

$$R_\eta^{(\gamma)}(A) := \mathbb{E}[\|\varphi(X_{t+\eta}) - A\varphi(X_t)\|_{\mathscr{H}}^2] + \gamma \|A\|_{S_2(\mathscr{H})}^2, \qquad (19)$$

which admits the minimizer $U_\eta^{(\gamma)}$ (Mollenhauer and Koltai, 2020). The corresponding *empirical surrogate risk*

$$R_\eta^{(\gamma,n)}(A) := \frac{1}{n}\sum_{t=1}^{n} \|\varphi(X_{t+\eta}) - A\varphi(X_t)\|_{\mathscr{H}}^2 + \gamma \|A\|_{S_2(\mathscr{H})}^2 \qquad (20)$$

attains its minimum at $U_\eta^{(\gamma,n)}$. We do not aim to cover all the mathematical intricacies of the CME and its connection to least squares regression here and refer the reader to Park and Muandet (2020) and Li et al. (2022) for more details.

7.3 Kernel Sum Rule

In the well-specified operator-theoretic setting of (16), Fukumizu et al. (2013, Theorem 2) show that the conditional mean operator U_η satisfies the more general so-called *kernel sum rule*, which is widely used in nonparametric Bayesian models, especially time series filtering. That is, for a *prior* measure z on $(E, \mathcal{F}_E)$ satisfying the integrability $\int_E \|\varphi(Z)\|_{\mathscr{H}} \,\mathrm{d}z(Z) < \infty$ with a kernel mean embedding $\mu_z = \int \varphi(Z)\mathrm{d}z(Z)$ such that $\mu_z \in \mathrm{dom}(C(0)^\dagger)$, we have

$$\langle f, U_\eta \mu_z \rangle_{\mathscr{H}} = \int_E \mathbb{E}[f(X_{t+\eta}) \mid X_t = x]\,\mathrm{d}z(x), \quad f \in \mathscr{H}. \qquad (21)$$

Note that the conditional mean property (16) is in fact a special case of the kernel sum rule when z is the Dirac measure at x, i.e., $\mu_z = \varphi(x)$. In applications, the embedded prior μ_z is usually estimated empirically by sampling from z. When $\widehat{\mu}_z$ is any kind of consistent estimate of μ_z, we obtain the plug-in estimator $U_\eta^{(\gamma,n)}\widehat{\mu}_z$ for $U_\eta \mu_z$.

7.4 Consistency Results

We now outline how our previous results allow to formulate consistency results for the kernel sum rule for dependent data. Prior consistency results for the operator-based setting of the conditional mean embedding and the kernel sum rule are limited to independent data pairs. Note again that a drawback of our approach is the typical assumption that the analytic expression $U_\eta \mu_z$ (and in particular $U_\eta \varphi(x)$) exists in $\mathscr{H}$ (see Klebanov et al., 2020), while a focus on the minimization properties allows to relax this assumption and consider convergence to a best approximation under the corresponding risk. We start by giving a generic error decomposition for the kernel sum rule in a form that admits the immediate application of our previous results.

Theorem 28 (Kernel sum rule error) *Let z be a prior finite measure on $(E, \mathcal{F}_E)$ and a kernel $k \colon E \times E \to \mathbb{R}$ with $\sup_{x \in E} k(x,x) = c < \infty$ such that the kernel sum rule (21) applies for a fixed time lag $\eta \in \mathbb{N}_{>0}$ (in particular, $\mu_z \in \operatorname{dom}(C(0)^\dagger)$). Then the empirical estimate $U_\eta^{(\gamma,n)} \widehat{\mu}_z$ admits the total error bound*

$$\left\| U_\eta^{(\gamma,n)} \widehat{\mu}_z - U_\eta \mu_z \right\|_{\mathscr{H}} \leq e_s(\mu_z, \widehat{\mu}_z, n, \gamma) + e_r(\mu_z, \gamma) \quad \mathbb{P}\text{-}a.e. \tag{22}$$

with the stochastic estimation error

$$e_s(\mu_z, \widehat{\mu}_z, n, \gamma) := \frac{c}{\gamma} \left\| \widehat{\mu}_z - \mu_z \right\|_{\mathscr{H}} + \frac{c^{3/2}}{\gamma^2} \left\| C_n(0) - C(0) \right\| + \frac{c^{1/2}}{\gamma} \left\| C_n(\eta) - C(\eta) \right\|$$

and the deterministic regularization error

$$e_r(\mu_z, \gamma) := c \left\| (C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1} \mu_z - C(0)^\dagger \mu_z \right\|_{\mathscr{H}}.$$

The proof for Theorem 28 can be found in Appendix A. In the context of inverse problems, the estimation error e_s is sometimes also called the *sample error*, while the regularization error is sometimes also called the *approximation error*.

Remark 29 (Kernel sum rule error) *The error decomposition (22) leads to the following insights:*

(i) *The deterministic regularization error $e_r(\mu_z, \gamma)$ captures the analytic nature of the inverse problem described by $C(0)u = \mu_z$ for $u \in \mathscr{H}$. As such, it is not affected by any estimation. Note that as $\gamma \to 0$, it holds that $e_r(\mu_z, \gamma) \to 0$ for every $\mu_z \in \operatorname{dom}(C(0)^\dagger)$. For details, we refer the reader to Engl et al. (1996). In practice, regularization is needed since for estimated right-hand sides $\widehat{\mu}_z$, the condition $\widehat{\mu}_z \in \operatorname{dom}(C(0)^\dagger)$ is in general not true—even if $\mu_z \in \operatorname{dom}(C(0)^\dagger)$. The convergence rate of $e_r(\mu_z, \gamma)$ depends on the eigendecomposition of $C(0)$ and can be assessed under additional assumptions about the decay rate of the eigenvalues. However, in general the convergence of the regularization error can be arbitrarily slow without additional assumptions, see Schock (1984).*

(ii) *For convergence of the total error (22), we need the two simultaneous conditions $e_s(\mu_z, \widehat{\mu}_z, n, \gamma) \to 0$ and $e_r(\mu_z, \gamma) \to 0$ as $n \to \infty$, $\gamma \to 0$ and $\widehat{\mu}_z \to \mu_z$. The typical*

trade-off between regularization error and estimation error is reflected in this fact. Our previous convergence results for the individual estimation errors of $C_n(0)$ and $C_n(\eta)$ allow to bound $e_s(\mu_z, \widehat{\mu}_z, n, \gamma)$ and derive regularization schemes $\gamma := \gamma(n, \mu_z, \widehat{\mu}_z)$ depending on the trajectory length n and the quality of the prior estimate $\widehat{\mu}_z$. Informally speaking, the individual estimation errors must tend to 0 faster than the regularization term, so γ should not be allowed to converge "too fast" with respect to the rate of increasing sample size n – this is the typical setting in the theory of inverse problems and regularization.

By incorporating additional knowledge about the convergence behavior of $C_n(0)$ and $C_n(\eta)$ from our previous results, Theorem 28 yields convergence rates of the estimation error e_s as well as admissible regularization schemes. We give an example below. For simplicity, we assume that $C_n(0)$ and $C_n(\eta)$ are estimated independently, which would of course require two realizations of length n of $(X_t)_{t \in \mathbb{Z}}$. To illustrate the idea, we require $\widehat{\mu}_z$ to converge with a $\mathbb{P}$-a.e. rate of $1/\sqrt{n}$, which we tie to the number of samples available for the estimation of $C_n(0)$ and $C_n(\eta)$ in order to avoid additional symbols for different samples. Our general approach presented here still applies when the convergence rate of $\widehat{\mu}_z$ is slower.

Example 2 (Kernel sum rule consistency) *Let $C_n(0)$ and $C_n(\eta)$ be estimated independently from samples $X_1, \ldots, X_n$. Assume that we have the prior convergence rate*

$$\|\mu_z - \widehat{\mu}_z\|_{\mathscr{H}} = O(1/\sqrt{n}) \quad \mathbb{P}\text{-a.e.}$$

Then under the conditions of Theorem 18, we have

$$e_s(\mu_z, \widehat{\mu}_z, n, \gamma) = O\left(\frac{\sqrt{2L(L(n))}}{\sqrt{n}\gamma^2}\right) \quad \mathbb{P}\text{-a.e.}$$

In particular, for every regularization scheme $\gamma = \gamma(n)$ such that

$$\gamma(n) \to 0 \text{ as well as } \frac{\sqrt{2L(L(n))}}{\sqrt{n}\gamma(n)^2} \to 0 \quad \mathbb{P}\text{-a.e.}$$

for $n \to \infty$, we have the overall convergence

$$U_\eta^{(\gamma(n),n)}\widehat{\mu}_z \to U_\eta\mu_z \quad \mathbb{P}\text{-a.e.}$$

in the norm of $\mathscr{H}$. Note that this follows since we have already covered the convergence of the regularization error $e_r(\mu_z, \gamma(n)) \to 0$ in Remark 29(i).

Remark 30 (Counterfactual mean embedding) *The results highlighted in this section require the fairly restrictive assumptions of the kernel sum rule as originally formulated by Fukumizu et al. (2013). As Muandet et al. (2021) have recently shown, the kernel sum rule estimator can be interpreted in the context of the so-called counterfactual mean embedding, where a convergence analysis can be based on the convergence of empirical kernel covariance operators under significantly weaker assumptions. Our results from the previous sections can be readily applied to the convergence analysis presented by Muandet et al. (2021, Section 4), leading to non-i.i.d. results in the framework of causal inference.*

8. Nonparametric Estimation of Markov Transition Operators

As the last application of our theory, we will briefly show how the risk functional (18) yields a nonparametric model for the estimation of *Markov transition operators*. Moreover, we elaborate on the recent discovery that this model is actually the theoretical foundation of a well-known family of several data-driven methods for the analysis of dynamical systems (Klus et al., 2020; Mollenhauer and Koltai, 2020). We only highlight immediate consequences of this approach and emphasize that several theoretical questions need to be answered separately in a vector-valued statistical learning context (Park and Muandet, 2020; Mollenhauer and Koltai, 2020). The aim of this section is to draw attention to the fact that statistical tools like strong mixing coefficients can be used to show consistency for a range of numerical methods used in other scientific disciplines. For the mathematical background of discrete-time Markov processes and Markov transition operators, we refer the reader to Douc et al. (2018).

In what follows, we assume that $(X_t)_{t \in \mathbb{Z}}$ is a stationary Markov process, i.e., it holds

$$\mathbb{E}\left[f(X_s) \mid \mathcal{F}^t_{-\infty}\right] = \mathbb{E}\left[f(X_s) \mid \sigma(X_t)\right]$$

for all bounded measurable functions $f \colon E \to \mathbb{R}$ and times $s \geq t$. For a fixed time lag $\eta \in \mathbb{N}_{>0}$, the *transition operator*, *(backward) transfer operator*[*] or *(stochastic) Koopman operator* $\mathcal{K}_\eta$ is defined by the relation

$$(\mathcal{K}_\eta f)(x) = \mathbb{E}[f(X_{t+\eta}) \mid X_t = x] \tag{23}$$

for all functions $f : E \to \mathbb{R}$ in some appropriately chosen subspace F of measurable functions. It describes the propagation of observable functions in F by the time step η under the dynamics given by $(X_t)_{t \in \mathbb{Z}}$.

The Markov transition operator is a fundamentally important tool for the analysis of various properties of Markov processes, Markov chain Monte Carlo methods, and dynamical systems. For instance, it is known that the spectrum of $\mathcal{K}_\eta$ and the associated eigenfunctions determine a crucial set of related properties of the underlying dynamics such as ergodicity, speed of convergence, the decomposition of the state space into almost invariant components (so-called *metastable states*), several contraction and concentration results and many more (see Meyn and Tweedie 2009; Rudolf 2012; Bovier and Den Hollander 2016; Douc et al. 2018 and the literature reviews therein).

By simply switching to the adjoint of A in the expression for $\mathcal{E}_\eta^{(\gamma)}(A)$ defined in (18) and using the reproducing property of $\mathcal{H}$, we have

$$\mathcal{E}_\eta^{(\gamma)}(A) := \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}}=1}} \mathbb{E}\left[(\mathbb{E}[f(X_{t+\eta}) \mid X_t] - (A^*f)(X_t))^2\right] + \gamma \|A^*\|^2_{S_2(\mathcal{H})}.$$

As a result, we can immediately interpret the adjoint of the conditional mean operator

$$U_\eta^{(\gamma)*} = (C(0) + \gamma I_{\mathcal{H}})^{-1} C(\eta)^*$$

---

[*]The name *backward transfer operator* is classically used in the context of continuous-time processes, where it is used to describe the solution to the *backwards Kolmogorov equation*. In the theory of dynamical systems, the term *Koopman operator* is commonly used.

as a smooth approximation of the transition operator $\mathcal{K}_\eta$ on the class of RKHS functions $F = \mathscr{H}$ with empirical estimate

$$U_\eta^{(\gamma,n)*} = (C_n(0) + \gamma I_{\mathscr{H}})^{-1} C_n(\eta)^*.$$

Note that all of our consistency results for kernel autocovariance operators and the conditional mean embedding transfer directly to this setting, as operator norm error bounds for the estimate of $U_\eta^{(\gamma)}$ are also valid for its adjoint. The approximation-theoretic details of this approach are investigated by Mollenhauer and Koltai (2020).

The idea of approximating the operator $\mathcal{K}_\eta$ via $U_\eta^{(\gamma)*}$ can be connected to data-driven spectral analysis techniques which are commonly used in fluid dynamics, molecular dynamics, and atmospheric sciences. One of the most widely used spectral analysis and methods is the so-called *extended dynamic mode decomposition* (EDMD,Williams et al., 2015a). EDMD computes Galerkin approximation of the eigendecomposition of $\mathcal{K}_\eta$ based on a finite set of basis functions in F. When F is chosen to be the RKHS $\mathscr{H}$, one obtains *kernel EDMD* (Williams et al., 2015b) as a special nonparametric version of EDMD. It was shown by Klus et al. (2020) that regularized kernel EDMD actually computes the eigendecomposition of $U_\eta^{(\gamma,n)*}$. In contrast to previous results that rely on ergodicity of the underlying system (Klus et al., 2016; Korda and Mezić, 2018), our results may lead to a refined convergence analysis by using mixing properties of the underlying system. To the best of our knowledge, prior consistency results for EDMD only cover convergence in strong operator topology (i.e., pointwise convergence) for parametric models, i.e. on fixed finite-dimensional subspaces spanned by a dictionary of basis functions. Furthermore, they mostly aim towards deterministic dynamical systems (Korda and Mezić, 2018). However, we note that the operator $U_\eta^{(\gamma)*}$ is in general not self-adjoint and a dedicated analysis of spectral properties and convergence is subject to future work.

Additionally, it is known that the operator $\mathcal{K}_\eta$ and its adjoint, the so-called *Perron–Frobenius* operator, can be connected to the solution of the so-called *blind source separation problem* (Klus et al., 2018b, 2020). In fact, eigenfunctions of compositions of empirical autocovariance operators (and their pseudoinverses) are used as projection coordinates in a kernel-based variant of *independent component analysis* (Harmeling et al., 2003; Schwantes and Pande, 2015). As such, consistency results formulated by us may be used to prove convergence for these approaches.

9. Conclusion

In this paper, we provided a mathematically rigorous analysis of kernel autocovariance operators and established classical limit theorems as well as nonasymptotic error bounds under classical ergodic and mixing assumptions. The results were mostly derived from theoretical work on discrete-time processes in Hilbert spaces and are presented in a form such that they can be easily applied in the context of RKHS-based time series models and frequency domain analysis. We highlighted high-level applications for kernel PCA, the conditional mean embedding, and the nonparametric estimation of Markov transition operators. The theory of vector-valued statistical learning from dependent data may be

connected to our considerations in future work. In the context of learning Markov transition operators, the kernel autocovariance operator may lead to an inverse problem that describes the analytical minimizer of an autoregression risk in an operator space.

Acknowledgements

Appendix A. Proofs

We report the proofs which were omitted in the main text.

A.1 Proof of Theorem 18

We apply Theorem 2 of Merlèvede (2008) to the process $(\xi)_{t \in \mathbb{Z}}$ defined by (3) for the fixed time lag $\eta \in \mathbb{N}$. This results verifies the existence of a compact set K_η with the desired properties such that both (8) and (9) hold. We note that Merlèvede (2008, Remark 3) ensures that our assumptions allow the application of this result. It now remains to show the norm bound (10) for K_η. The set K_η is the unit ball of the Hilbert space $\mathbb{H}_\eta \subseteq S_2(\mathscr{H})$, which is given by the completion of the range of $T_\eta^{1/2}$ (where T_η is given by (7) and $T_\eta^{1/2}$ denotes its operator square root) with respect to the inner product defined by

$$\left\langle T_\eta^{1/2} A, \, T_\eta^{1/2} B \right\rangle_{\mathbb{H}_\eta} := \langle A, \, B \rangle_{S_2(\mathscr{H})}, \quad A, B \in S_2(\mathscr{H}). \tag{24}$$

The space $\mathbb{H}_\eta$ is also called *Cameron–Martin space* or *abstract Wiener space* in the context of Gaussian measures (for details, we refer the reader to Bogachev, 1998, Chapter 2). It is noteworthy that $\mathbb{H}_\eta$ itself is sometimes also called *reproducing kernel Hilbert space* (associated with the so-called *covariance kernel* defined by T_η, see Merlèvede, 2008, Theorem 2), which may seem misleading in our context at first glance. In particular, note that the space $S_2(\mathscr{H})$ consists of Hilbert–Schmidt operators and not of real-valued functions as in our definition of an RKHS in this paper. The interpretation of $\mathbb{H}_\eta$ as a reproducing kernel Hilbert space requires an identification of $\mathbb{H}_\eta$ with its dual space as real-valued functions contained in the space $L^2(\mathcal{N}(0,T))$ as described by Bogachev (1998, Remark 2.2.3). The final connection between this setting and our definition of an RKHS is made clear by Mercer's theorem and the related theory of integral operators (Steinwart and Christmann, 2008, Chapter 4.5). We will not cover all the mathematical details here and refer the reader to more comprehensive treatments below.

For a technical construction of $\mathbb{H}_\eta$ and the limit set K_η in the law of the iterated logarithm in Banach spaces, we refer the reader to Kuelbs (1976, Section 2) as well as Goodman et al. (1981, Section 2). Note that these references elaborate on the i.i.d. case. However, for the

construction of $\mathbb{H}_\eta$ and K_η only an abstract limiting probability measure is needed, which is given by the Gaussian measure obtained from Theorem 17 and its covariance operator T_η defined by (7), just as shown in the proof of Merlèvede (2008, Theorem 2). We can therefore analyze properties of K_η by considering the Cameron–Martin space of the centered Gaussian measure induced by T_η, which is examined in the previously mentioned literature. The identity (24) can be verified by translating the abstract Banach space definition of (Kuelbs, 1976, Equation 2.3) to our scenario of the separable Hilbert space $S_2(\mathscr{H})$ as, for example, described by Bogachev (1998, Remark 2.3.3).

From (24), we obtain

$$\|A\|_{S_2(\mathscr{H})} \leq \left\|T_\eta^{1/2}\right\| \|A\|_{\mathbb{H}_\eta}, \quad A \in \mathbb{H}_\eta. \tag{25}$$

Since $K_\eta = \left\{A \in \mathbb{H}_\eta \mid \|A\|_{\mathbb{H}_\eta} \leq 1\right\}$, a bound for $\left\|T_\eta^{1/2}\right\| = \|T_\eta\|^{1/2}$ depending on the mixing rate of $(\xi_t)_{t \in \mathbb{Z}}$ is sufficient in order to provide a bound for elements of K in the norm of $S_2(\mathscr{H})$.

We now give a norm bound for $T_\eta = \mathbb{E}[\xi_0 \otimes \xi_0] + \sum_{t=1}^\infty \mathbb{E}[\xi_0 \otimes \xi_t] + \sum_{t=1}^\infty \mathbb{E}[\xi_t \otimes \xi_0]$. We clearly have

$$\|\mathbb{E}[\xi_0 \otimes \xi_0]\| \leq 4c^2,$$

since ξ_0 is almost surely bounded by $2c$ by (5).

Let $\alpha(n)$ be the mixing coefficients of $(X_t)_{t \in \mathbb{Z}}$. We now note that by (4), we have $\alpha((\xi_t)_{t \in \mathbb{Z}}, n) \leq \alpha(n - \eta)$ for all $n \in \mathbb{N}$. This allows to give a bound for the two remaining summands of T:

$$
\begin{aligned}
\left\|\sum_{t=1}^\infty \mathbb{E}[\xi_t \otimes \xi_0]\right\| &\leq \sum_{t=1}^\infty \|\mathbb{E}[\xi_t \otimes \xi_0]\| \\
&= \sum_{t=1}^\infty \sup_{\substack{\|A\|_{S_2(\mathscr{H})}=1 \\ \|B\|_{S_2(\mathscr{H})}=1}} |\mathbb{E}[\langle \xi_t, B\rangle \langle \xi_0, A\rangle]| \\
&\leq \sum_{t=1}^\infty \sup_{\substack{\|A\|_{S_2(\mathscr{H})}=1 \\ \|B\|_{S_2(\mathscr{H})}=1}} 4\,\alpha(\sigma(\xi_t), \sigma(\xi_0)) \|\langle \xi_t, B\rangle\|_{L^\infty(\mathbb{P})} \|\langle \xi_0, A\rangle\|_{L^\infty(\mathbb{P})} \\
&\leq \sum_{t=1}^\infty 16\,c^2 \alpha(t - \eta) = 16\,c^2 M,
\end{aligned}
$$

where we use Ibragimov's covariance inequality for strongly mixing and bounded random variables (Ibragimov, 1962, Lemma 1.2) in the third step (note that $\langle \xi_t, B\rangle$ and $\langle \xi_0, A\rangle$ are centered real-valued random variables which are $\mathbb{P}$-a.e. bounded by $2c$ because of Equation 5). By symmetry, we obtain the same bound for $\|\sum_{t=1}^\infty \mathbb{E}[\xi_0 \otimes \xi_t]\|$ and we end up with the total norm bound

$$\|T_\eta\| \leq 4c^2 + 32\,c^2 M,$$

which proves the claim in combination with (25). ∎

A.2 Proof of Theorem 28

Note that since $\sup_{x \in E} k(x,x) = c < \infty$, we have the $\mathbb{P}$-a.e. bounds $\|\mu_z\|_{\mathscr{H}} \leq c^{1/2}$ as well as $\|C(\eta)\| \leq c$ for all $\eta \in \mathbb{N}$. Additionally, the regularized inverse can be bounded as $\|(C(0) + \gamma I_{\mathscr{H}})^{-1}\| \leq \frac{1}{\gamma}$, which is easy to see from the corresponding spectral decomposition. These bounds hold analogously for the empirical versions of all above objects.[†] All following bounds below will be understood in the $\mathbb{P}$-a.e. sense for the remainder of this proof.

We now successively insert appropriate zero-sum terms into the total error and apply the triangle inequality multiple times to obtain the worst-case estimation error. We have the overall decomposition

$$\left\| U_\eta^{(\gamma,n)} \widehat{\mu}_z - U_\eta \mu_z \right\|_{\mathscr{H}} \leq \underbrace{\left\| U_\eta^{(\gamma,n)} \widehat{\mu}_z - U_\eta^{(\gamma,n)} \mu_z \right\|_{\mathscr{H}}}_{(I)} + \underbrace{\left\| U_\eta^{(\gamma,n)} \mu_z - U_\eta \mu_z \right\|_{\mathscr{H}}}_{(II)}. \tag{26}$$

For these two error components, we get the individual bounds

$$(I) \leq \|C_n(\eta)\| \left\| (C_n(0) + \gamma I_{\mathscr{H}})^{-1}(\widehat{\mu}_z - \mu_z) \right\|_{\mathscr{H}} \leq \frac{c}{\gamma} \|\widehat{\mu}_z - \mu_z\|_{\mathscr{H}}$$

as well as

$$(II) \leq \underbrace{\left\| C_n(\eta)(C_n(0) + \gamma I_{\mathscr{H}})^{-1}\mu_z - C_n(\eta)(C(0) + \gamma I_{\mathscr{H}})^{-1}\mu_z \right\|_{\mathscr{H}}}_{(\star)}$$

$$+ \underbrace{\left\| C_n(\eta)(C(0) + \gamma I_{\mathscr{H}})^{-1}\mu_z - C(\eta)C(0)^{\dagger}\mu_z \right\|_{\mathscr{H}}}_{(\star\star)}.$$

For $(\star)$, we give a bound by

$$(\star) \leq \|C_n(\eta)\| \left\| (C_n(0) + \gamma I_{\mathscr{H}})^{-1} - (C(0) + \gamma I_{\mathscr{H}})^{-1} \right\| \|\mu_z\|_{\mathscr{H}}$$

$$\leq \frac{c^{3/2}}{\gamma^2} \|C_n(0) - C(0)\|,$$

where we use the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for invertible operators A and B. To obtain a bound for $(\star\star)$, we again insert a zero-sum term:

$$(\star\star) \leq \left\| C_n(\eta)(C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1}\mu_z - C(\eta)(C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1}\mu_z \right\|_{\mathscr{H}}$$

$$+ \left\| C(\eta)(C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1}\mu_z - C(\eta)C(0)^{\dagger}\mu_z \right\|_{\mathscr{H}}$$

$$\leq \|C_n(\eta) - C(\eta)\| \left\| (C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1} \right\| \|\mu_z\|_{\mathscr{H}}$$

$$+ \|C(\eta)\| \left\| (C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1}\mu_z - C(0)^{\dagger}\mu_z \right\|_{\mathscr{H}}$$

$$\leq \frac{c^{1/2}}{\gamma} \|C_n(\eta) - C(\eta)\| + c \left\| (C(0) + \gamma \mathcal{I}_{\mathscr{H}})^{-1}\mu_z - C(0)^{\dagger}\mu_z \right\|_{\mathscr{H}}.$$

The sum of the bounds (I), $(\star)$, and $(\star\star)$ yields the total bound as given in (22) after rearranging. $\blacksquare$

[†] For $\widehat{\mu}_z$, estimators of the form $\widehat{\mu}_z := \sum_i \beta_i \varphi(x_i)$ with coefficients $\sum_i |\beta_i| = 1$ naturally satisfy the bound.

References

A. Allam and T. Mourid. Covariance operator estimation of a functional autoregressive process with random coefficients. *Statistics & Probability Letters*, 84:1–8, 2014.

A. Allam and T. Mourid. Optimal rate for covariance operator estimators of functional autoregressive processes with random coefficients. *Journal of Multivariate Analysis*, 169: 130–137, 2019.

C. Baker. Mutual information for Gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

A. Beck and J. Schwartz. A vector-valued random ergodic theorem. *Proceedings of the American Mathematical Society*, 8:1049–1059, 1957.

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

A. T. Bharucha-Reid. *Random Integral Equations*. Academic Press, 1972.

G. Blanchard and O. Zadorozhnyi. Concentration of weakly dependent Banach-valued sums and applications to statistical learning methods. *Bernoulli*, 25(4B):3421–3458, 2019.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

V. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

B. Boots, A. Gretton, and G. J. Gordon. Hilbert space embeddings of predictive state representations. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 92–101. AUAI Press, 2013.

D. Bosq. *Linear Processes in Function Spaces*. Lecture Notes in Statistics. Springer, 2000.

D. Bosq. Estimation of mean and covariance operator of autoregressive processes in Banach spaces. *Statistical Inference for Stochastic Processes*, 5:287–306, 2002.

A. Bovier and F. Den Hollander. *Metastability: A Potential-Theoretic Approach*. Springer, 2016.

R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12:136–154, 1982.

A. de Acosta. A New Proof of the Hartman-Wintner Law of the Iterated Logarithm. *The Annals of Probability*, 11(2):270–276, 1983.

J. Dedecker and F. Merlèvede. On the almost sure invariance principle for stationary sequences of Hilbert-valued random variables. In I. Berkes, R. Bradley, H. Dehling, M. Peligrad, and R. Tichy, editors, *Dependence in Probability, Analysis and Number Theory*, pages 157–175. Kendrick Press, 2010.

H. Dehling and O. Sharipov. Estimation of mean and covariance operator for Banach space valued autoregressive processes with dependent innovations. *Statistical Inference for Stochastic Processes*, 8:137–149, 2005.

J. Diestel and J. Uhl. *Vector Measures.* American Mathematical Society, 1977.

J. Doob. *Stochastic Processes.* John Wiley & Sons, 1953.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains.* Springer, 2018.

P. Doukhan. *Mixing: Properties and Examples.* Springer-Verlag New York, 1994.

P. Doukhan, P. Massart, and E. Rio. The functional central limit theorem for strongly mixing processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 30(1): 63–82, 1994.

R. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, 1996.

J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, 2003.

K. Fukumizu. Nonparametric Bayesian inference with kernel mean embedding. In G. Peters and T. Matsui, editors, *Modern Methodology and Applications in Spatial-Temporal Modeling.* Springer, 2017.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

G. Gebhardt, K. Daun, M. Schnaubelt, and G. Neumann. Learning robust policies for object manipulation with robot swarms. In *2018 IEEE International Conference on Robotics and Automation*, pages 7688–7695, 2018.

G. Gebhardt, A. Kupcsik, and G. Neumann. The kernel Kalman rule. *Machine Learning*, 108(12):2113–2157, 2019.

I. C. Gohberg and M. G. Kreĭn. *Introduction To The Theory of Linear Nonselfadjoint Operators*, volume 18 of *Translations of Mathematical Monographs.* American Mathematical Society, 1969.

V. Goodman, J. Kuelbs, and J. Zinn. Some results on the LIL in Banach space with applications to weighted empirical processes. *Annals of Probability*, 9(5):713–752, 1981.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1823–1830, New York, NY, USA, 2012a. Omnipress.

S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *Proceedings of the 29th International Conference on Machine Learning*, pages 535–542, New York, NY, USA, 2012b. Omnipress.

H. Hang and I. Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2): 708–743, 2017.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, 2012.

J.-C. Hua, F. Noorian, D. Moss, P. H. W. Leong, and G. H. Gunaratne. High-dimensional time series prediction using kernel-based Koopman mode regression. *Nonlinear Dynamics*, 90(3):1785–1806, 2017.

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.

S. Hörmann, L. Kidzinski, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):319–348, 2015.

I. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability and its Applications*, 7:349–382, 1962.

M. Jirak and M. Wahl. Perturbation bounds for eigenspaces under a relative gap condition. *Proceedings of the American Mathematical Society*, 148:479–494, 2020.

O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2002.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Y. Kawahara. Dynamic mode decomposition with reproducing kernels for Koopman spectral analysis. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 911–919. Curran Associates, Inc., 2016.

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

S. Klus, P. Koltai, and C. Schütte. On the numerical approximation of the Perron–Frobenius and Koopman operator. *Journal of Computational Dynamics*, 3(1):51–79, 2016.

S. Klus, A. Bittracher, I. Schuster, and C. Schütte. A kernel-based approach to molecular conformation analysis. *The Journal of Chemical Physics*, 149:244109, 2018a.

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

S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020.

P. Kokoszka and M. Reimherr. Asymptotic normality of the principal components of functional time series. *Stochastic Processes and their Applications*, 123(5):1546 – 1562, 2013.

V. Koltchinskii and K. Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976–2013, 11 2016.

V. Koltchinskii and K. Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157, 02 2017.

M. Korda and I. Mezić. On convergence of Extended Dynamic Mode Decomposition to the Koopman operator. *Journal of Nonlinear Science*, 28:687–710, 2018.

J. Kuelbs. A strong convergence theorem for Banach space valued random variables. *The Annals of Probability*, 4(5):744–771, 1976.

J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

G. Lever, J. Shawe-Taylor, R. Stafford, and C. Szepesvári. Compressed conditional mean embeddings for model-based reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 1779–1787, 2016.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*, volume 36, 2022. To appear.

A. Mas. Weak convergence for the covariance operators of a Hilbertian linear process. *Stochastic Processes and their Applications*, 99(1):117–135, 2002.

A. Mas. A sufficient condition for the CLT in the space of nuclear operators—Application to covariance of random functions. *Statistics & Probability Letters*, 76(14):1503–1509, 2006.

A. Mas and L. Menneteau. Perturbation approach applied to the asymptotic study of random operators. In J. Hoffmann-Jørgensen, J. Wellner, and M. Marcus, editors, *High Dimensional Probability III. Progress in Probability*, volume 55, pages 127–134. Birkhäuser, 2003.

A. Mas and F. Ruymgaart. High-dimensional principal projections. *Complex Analysis and Operator Theory*, 9:35–63, 2015.

V. Maume-Deschamps. Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stochastic Dynamics*, 6:1049–1059, 2006.

F. Merlevède, M. Peligrad, and S. Utev. Sharp conditions for the CLT of linear processes in a Hilbert space. *Journal of Theoretical Probability*, 10:681–693, 1997.

F. Merlèvede. On a maximal inequality for strongly mixing random variables in Hilbert spaces. Application to the compact law of the iterated logarithm. *Annales de l'I.S.U.P.*, 52:47–60, 2008.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

C. Milbradt and M. Wahl. High-probability bounds for the reconstruction error of PCA. *Statistics & Probability Letters*, 161:108741, 2020.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3637, 1994.

M. Mollenhauer and P. Koltai. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10 (1–2):1–141, 2017.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert space embeddings of POMDPs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, UAI'12. AUAI Press, 2012.

H. Owhadi and C. Scovel. Separability of reproducing kernel spaces. *Proceedings of the American Mathematical Society*, 145:2131–2138, 2017.

V. M. Panaretos and S. Tavakoli. Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Processes and their Applications*, 123(7):2779 – 2807, 2013.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.

K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1), 2013.

K. Petersen. *Ergodic Theory*. Cambridge University Press, 1983.

Y. U. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability and its Applications*, 1(2):157–214, 1956.

M. Reed and B. Simon. *Methods of Mathematical Physics I: Functional Analysis*. Academic Press Inc., 2nd edition, 1980.

M. Reiß and M. Wahl. Nonasymptotic upper bounds for the reconstruction error of PCA. *The Annals of Statistics*, 48(2):1098–1123, 2020.

E. Rio. The functional law of the iterated logarithm for stationary strongly mixing sequences. *The Annals of Probability*, 23(3):1188–1203, 07 1995.

E. Rio. *Asymptotic Theory of Weakly dependent random processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer, 2017.

H. Robbins. Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

D. Rudolf. Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Math.*, 486, 2012.

S. Saitoh and Y. Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

E. Schock. On the asymptotic order of accuracy of Tikhonov regularization. *Journal of Optimization Theory and Applications*, 44(1):95–104, Sep 1984.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

C. R. Schwantes and V. S. Pande. Modeling molecular kinetics with tICA and the kernel trick. *Journal of Chemical Theory and Computation*, 11(2):600–608, 2015.

D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer-Verlag, New York, 1985.

B. Simon. *Trace Ideals and Their Applications*. Mathematical surveys and monographs 120. American Mathematical Society, 2nd edition, 2005.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

A. R. Soltani and M. Hashemi. Periodically correlated autoregressive Hilbertian processes. *Statistical Inference for Stochastic Processes*, 14(2):177–188, 2011.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. Association for Computing Machinery, 2009.

L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 991–998. Omnipress, 2010.

R. Stafford and J. Shawe-Taylor. ACCME: Actively compressed conditional mean embeddings for model-based reinforcement learning. In *European Workshop on Reinforcement Learning 14*, 2018.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

Y. Sun, Y. Duan, H. Gong, and M. Wang. Learning low-dimensional state embeddings and metastable clusters from time series data. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4563–4572. Curran Associates, Inc., 2019.

J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1 (2), 2014.

H. van Hoof, J. Peters, and G. Neumann. Learning of Non-Parametric Control Policies with High-Dimensional State Features. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 995–1003. PMLR, 2015.

H. van Hoof, G. Neumann, and J. Peters. Non-parametric policy search with limited information loss. *Journal of Machine Learning Research*, 18(73):1–46, 2017.

M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015a.

M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265, 2015b.

H. Wu and F. Noé. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.