Dissertation
zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

On the Statistical Approximation of Conditional Expectation Operators

vorgelegt von
Mattes Mollenhauer

eingereicht am Fachbereich für Mathematik und Informatik
der Freien Universität Berlin

Berlin, Oktober 2021

Erstgutachter:	Prof. Dr. Christof Schütte
Zweitgutachter:	Dr. Tim Sullivan
Tag der Disputation:	11. März 2022

„... denn in allem Chaos ist Kosmos und in aller Unordnung geheime Ordnung“

Carl Gustav Jung

Acknowledgements

First of all, I would like to thank Jette for always being there and keeping me sane during my journey of mathematical research.

I am greatly indebted to my parents as well as Maren, Ralf, my grandparents and all my family for their unconditional support and understanding. Thank you. I owe special thanks to Ralf for the careful proofreading of my manuscript.

Furthermore, I would like to express my gratitude and appreciation to Christof Schütte, Péter Koltai and Stefan Klus for giving me the chance and the freedom to explore various fascinating topics within our mathematical field of work. Without their trust in my abilities and careful guidance, this work would not have been possible.

I am thankful to have met great colleagues along the way: Ingmar Schuster, Andreas Bittracher, Ilja Klebanov and Tim Sullivan, who provided important advice in various situations. The discussions with you helped to shape my perception of several mathematical concepts occurring in this thesis.

Last but certainly not least: to Paul K., Mike, Hümmet, Jo, Alex, Filip, Jacek, Timm, Paul M., Max and Ahmed—thank you for staying friends over all those years and having my back.

Preface

This thesis consists of five chapters. The first three chapters serve an introductory purpose and describe the author's perspective on a well-known numerical problem. To the best of the author's knowledge, some viewpoints and results in the third chapter are new in this very specific context. The last two chapters contain results of mathematical research which were obtained in the following work:

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

(ii) M. Mollenhauer, S. Klus, C. Schütte, and P. Koltai. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020a.

The author of this thesis is responsible for the basic concept, the mathematical research and the writing of the manuscripts listed above. At the time of writing this thesis, both papers are undergoing a scientific peer review process.

Some parts of this thesis are identical to the content of the papers (i) and (ii). We indicate the connection between these papers and this thesis at the beginning of each chapter. All the passages taken from the above papers and all presented results are ultimately due to the author of this thesis.

The perspectives and ideas in this thesis are strongly influenced by some of the results from the author's earlier collaborations which are not presented in this thesis:

(i) I. Schuster, M. Mollenhauer, S. Klus, and K. Muandet. Kernel conditional density operators. In S. Chiappa and R. Calandra, editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 993–1004. PMLR, 2020,

(ii) M. Mollenhauer, I. Schuster, S. Klus, and C. Schütte. Singular value decomposition of operators on reproducing kernel Hilbert spaces. In O. Junge, O. Schütze, G. Froyland, S. Ober-Blobaum, and K. Padberg-Gehle, editors, *Advances on Dynamics, Optimization and Computation. Series: Studies in Systems, Decision and Control*, volume 304, pages 109–131. Springer, 2020b and

(iii) S. Klus, B. E. Husic, M. Mollenhauer, and F. Noé. Kernel methods for detecting coherent structures in dynamical data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):123112, 2019.

In the main text, we refer to all of the above preprints and publications by their individual entry in the bibliography of this thesis.

Zusammenfassung

Diese Dissertation erörtert die datengetriebene Approximation des sogenannten *conditional expectation operators*, welcher den Erwartungswert einer reellwertigen Transformation einer Zufallsvariablen bedingt auf eine zweite Zufallsvariable beschreibt. Sie stellt dieses klassische numerische Problem in einem neuen theoretischen Zusammenhang dar und beleuchtet es mit verschiedenen ausgewählten Methoden der modernen statistischen Lerntheorie. Es werden sowohl ein bekannter parametrischer Projektionsansatz aus dem numerischen Bereich als auch ein nichtparametrisches Modell auf Basis eines *reproducing kernel Hilbert space* untersucht.

Die Untersuchungen dieser Arbeit werden motiviert duch den speziellen Fall, in dem der *conditional expectation operator* die Übergangswahrscheinlichkeiten eines Markovprozesses beschreibt. In diesem Kontext sind die Spektraleigenschaften des resultierenden *Markov transition operators* von großem praktischen Interesse zur datenbasierten Untersuchung von komplexer Dynamik. Die oben genannten vorgestellten Schätzer werden in diesem Szenario in der Praxis verwendet.

Es werden diverse neue Konvergenz- und Approximationsresultate sowohl für stochastisch unabhängige als auch abhängige Daten gezeigt. Als Werkzeuge für diese Ergebnisse dienen Konzepte aus den Theorien inverser Probleme, schwach abhängiger stochastischer Prozesse, der Störung von Spektraleigenschaften und der Konzentration von Wahrscheinlichkeitsmaßen. Zur theoretischen Rechtfertigung des nichtparametrischen Modells wird das Schätzen von *kernel autocovariance operators* von stationären Zeitreihen untersucht. Diese Betrachtung kann zusätzlich vielfältig in anderen Zusammenhängen genutzt werden, was anhand von neuen Ergebnissen zur Konsistenz von kernelbasierter Hauptkomponentenanalyse mit schwach abhängigen Daten demonstriert wird.

Diese Dissertation ist theoretischer Natur und dient nicht zur unmittelbaren Umsetzung von neuen numerischen Methoden. Sie stellt jedoch den direkten Zusammenhang von bekannten Ansätzen in diesem Feld zu relevanten statistischen Arbeiten der letzten Jahre her, welche sowohl stärkere theoretische Ergebnisse als auch effizientere praktische Schätzer für dieses Problem in der Zukunft möglich machen.

Selbstständigkeitserklärung

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Alle Ausführungen, die wörtlich oder inhaltlich aus Schriften anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Mattes Mollenhauer

Contents

1. Introduction

This chapter contains passages taken from Mollenhauer et al. (2020a).

Given the joint distribution of two random variables X and Y taking values in some standard Borel space $(E, \mathcal{F}_E)$ with marginal distributions $X \sim \pi$ and $Y \sim \nu$, we investigate the statistical approximation of the linear *conditional expectation operator* $P : L^2(\nu) \to L^2(\pi)$ defined by

$$[Pf](x) := \mathbb{E}[f(Y) \mid X = x].$$

The conditional expectation operator can be understood as the functional-analytic analogue of the *Markov transition kernel* $p : E \times \mathcal{F}_E \to [0, 1]$ given by

$$p(x, \mathcal{A}) = \mathbb{P}[Y \in \mathcal{A} \mid X = x] \quad \text{for all } x \in E \text{ and } \mathcal{A} \in \mathcal{F}_E,$$

which describes the stochastic connection between X and Y. In contrast to the somewhat intangible measure-theoretic properties of the Markov transition kernel however, the operator P offers an accessible perspective in terms of the comprehensive functional analytic theory of linear operators.

For example, when P describes the transition probabilities of a Markov process, its spectral properties allow to characterize a wide variety of important features of the underlying dynamics. Due to this fact, the data-driven approximation of P based on empirical observations drawn from the joint distribution of X and Y has become ubiquitous in the numerical theory of dynamical systems and Markov processes. Interestingly, although the approximation of P is of a statistical nature in its essence, even well-known numerical approximation schemes have not yet been rigorously investigated from a statistical perspective.

Consequently, our primary goal in this thesis is to bridge the gap between the classical numerical approximation of P and developments in modern high-dimensional and nonparametric statistics. In particular, we derive results for the estimation of P and its spectral properties by accessing tools from important fields such as the concentration of measure, random matrices, spectral perturbation theory, kernel-based nonparametric inference and regularization theory. Our main motivation for this analysis are practical applications in which P describes the transitions of a Markov process or random

dynamical system. Therefore, our consistency results explicitly cover the case that the underlying data consists of subsequent observations from an empirical realization of a Markov process. *We neither claim to present an exhaustive theory of statistically optimal results nor derive new numerical methods for immediate practical applications in this thesis. Instead, we develop a mathematical perspective which allows the application of more sophisticated estimators and sampling techniques from the aforementioned statistical fields to the data-driven approximation of P in the future.*

This thesis is structured as follows. In *Chapter 2*, we introduce the necessary mathematical background and review the most important concepts from measure theory, integration and linear functional analysis.

In *Chapter 3*, we introduce the conditional expectation operator P and its properties. We investigate its singular value decomposition in the context of the so-called *canonical components* and address the empirical estimation of the singular functions of P in terms of a spectral perturbation result. We connect this general framework to the special case when P describes the transition probabilities of a Markov process. In particular, we discuss a parametric projection method for the empirical analysis of random dynamical systems which enjoys widespread practical use. We establish asymptotic as well as non-asymptotic convergence results for both independent and dependent observations and elaborate on theoretical limitations of these results from a statistical perspective.

In *Chapter 4*, we develop a framework for the nonparametric estimation of P based on a *reproducing kernel Hilbert space*. We establish a connection to the theory of *regularized least squares regression* with vector-valued kernels and related concepts from nonparametric inference such as the *kernel mean embedding* and the *maximum mean discrepancy*. We show that the nonparametric approximation of P admits a closed form solution for the special case of *Tikhonov–Phillips regularization*. We conclude by highlighting that this closed form solution and its empirical estimate are connected to a class of popular kernel-based spectral analysis techniques for dynamical systems.

In *Chapter 5*, we address the estimation of *kernel autocovariance operators*, which are strongly connected to the nonparametric approximation of P and several other statistical models for stationary time series. We prove various convergence results under the assumptions of ergodicity and mixing. As an application, we use these results to show consistency of *kernel principal component analysis* when the underlying data is dependent.

Each of the last three chapters concludes with an individual summary and provides an outlook on open problems and possible topics for future research.

2. Mathematical preliminaries

This chapter contains passages taken from the preliminary sections of Mollenhauer et al. (2020a) and Mollenhauer and Koltai (2020).

We briefly introduce the needed concepts from measure theory and integration (Diestel and Uhl, 1977), compact linear operators and spectral theory in Hilbert spaces (Dunford and Schwartz, 1988a,b; Reed and Simon, 1980) and basic probability theory (Dudley, 2002; Kallenberg, 2002). Although we aim to provide a mathematical exposition in the most general form possible, we may occasionally choose readability and clarity over an extreme degree of abstractness. Therefore, some details in our presentation can certainly be transferred to a more general scenario without significant technical effort.

2.1. Topology, measure and integration

For any topological space (E, τ), we will write $\mathcal{F}_E := \mathcal{B}(E)$ for its associated Borel field. For any collection $\mathcal{M}$ of subsets of E, $\sigma(\mathcal{M})$ denotes the intersection of all σ-fields containing $\mathcal{M}$. For any σ-field $\mathcal{F}$ on E and countable index set I, we write $\mathcal{F}^{\otimes I}$ as the product σ-field on the product space E^I (i.e., the smallest σ-field with respect to which all coordinate projections on E^I are measurable). If (E, τ) is separable and completely metrizable, it is called a *Polish space*. If (E, τ) is a Polish space, 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 (Dudley, 2002, Proposition 4.1.17). A Polish space equipped with its Borel field is called a *standard Borel space*. If $(E, \mathcal{F}_E)$ is a standard Borel space, then the space $(E^I, \mathcal{F}_E^{\otimes I})$ is again a standard Borel space.

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 $\langle \cdot, \cdot \rangle_H$. Let $(\Omega, \mathcal{F}, \pi)$ be a measure space. For $1 \leq q \leq \infty$, let $L^q(\Omega, \mathcal{F}, \pi; B)$ denote the Banach space of equivalence classes of $\mathcal{F} - \mathcal{F}_B$ measurable and Bochner q-integrable functions $f \colon \Omega \to B$ equipped with the norms

$$\|f\|^q_{L^q(\Omega,\mathcal{F},\pi;B)} := \int_\Omega \|f(\omega)\|_B^q \, \mathrm{d}\pi(\omega) \quad 1 \leq q < \infty,$$
$$\|f\|_{L^\infty(\Omega,\mathcal{F},\pi;B)} := \operatorname*{ess\,sup}_{\omega \in \Omega} \|f(\omega)\|_B \,.$$

In the case of $B = \mathbb{R}$, we simply write $L^q(\pi) := L^q(\Omega, \mathcal{F}, \pi; \mathbb{R})$ for the standard space of real-valued Lebesgue q-integrable functions when the choice of σ-field on Ω is clear. For $\mathcal{F} - \mathcal{F}_B$ measurable functions $f : \Omega \to B$, the map $\omega \to \|f(\omega)\|_B$ is $\mathcal{F} - \mathcal{B}(\mathbb{R})$ measurable and the conditions $f \in L^q(\Omega, \mathcal{F}, \pi; B)$ and $\|f\|_B \in L^q(\pi)$ are equivalent.

If $A : B \to B'$ is a bounded linear operator between Banach spaces and $f \in L^1(\Omega, \mathcal{F}, \pi; B)$, then $Af(\cdot) \in L^1(\Omega, \mathcal{F}, \pi; B')$ and we have

$$A\left(\int_E f(\omega)\,\mathrm{d}\pi(\omega)\right) = \int_E Af(\omega)\,\mathrm{d}\pi(\omega).$$

When $B = H$ is a Hilbert space, the choice of $q = 2$ defines the Hilbert space $L^2(\Omega, \mathcal{F}, \pi; H)$, where the above norm is induced by the inner product

$$\langle f,\, g\rangle_{L^2(\Omega,\mathcal{F},\pi;H)} := \int_\Omega \langle f(\omega),\, g(\omega)\rangle_H \,\mathrm{d}\pi(\omega).$$

2.2. Linear operators and spectral theory

The expression $\mathfrak{B}(B, B')$ denotes the Banach space of bounded linear operators from B to another Banach space B' and is equipped with the operator norm $\|\cdot\|$. For the case $B = B'$, we abbreviate $\mathfrak{B}(B, B') = \mathfrak{B}(B)$. We will also write $\|\cdot\| = \|\cdot\|_{B\to B'}$, if the choice of norms on the underlying spaces B, B' needs to be emphasized.

Tensor product spaces. The expression $H' \otimes H$ denotes the tensor product of Hilbert spaces H, H'. The Hilbert space $H' \otimes H$ is the completion of the algebraic tensor product with respect to the inner product $\langle x_1' \otimes x_1,\, x_2' \otimes x_2\rangle_{H'\otimes H} = \langle x_1',\, x_2'\rangle_{H'} \langle x_1,\, x_2\rangle_H$ for $x_1, x_2 \in H$ and $x_1', x_2' \in H'$. We interpret the element $x' \otimes x \in H' \otimes H$ as the linear rank-one operator $x' \otimes x \colon H \to H'$ defined by $\tilde{x} \mapsto \langle \tilde{x},\, x\rangle_H\, x'$ for all $\tilde{x} \in H$. Whenever $(e_i)_{i\in I}$, $(e_j')_{j\in J}$ are complete orthonormal systems (CONSs) in H and H', $(e_j' \otimes e_i)_{i\in I, j\in J}$ is a CONS in $H' \otimes H$. Thus, when H and H' are separable, $H' \otimes H$ is separable.

Spectral theorem for compact self-adjoint operators. An operator $A : H \to H'$ is called *compact* if it maps bounded sets in H to relatively compact sets in H'. That is, for every bounded sequence $(x_n)_{n\in\mathbb{N}}$ in H, the sequence $(Ax_n)_{n\in\mathbb{N}}$ in H' contains a converging subsequence. For every compact self-adjoint operator $A : H \to H$, there exists an either finite or countably infinite index set $I = \{1, 2, \dots\}$ and an orthonormal system of *eigenvectors* $\{v_i\}_{i\in I} \subset H$ with a corresponding sequence of real *eigenvalues* $(\mu_i(A))_{i\in I}$ such that A admits the *spectral decomposition*

$$A = \sum_{i\in I} \mu_i(A)\, v_i \otimes v_i, \tag{2.1}$$

where the above sum converges in the operator norm. If the index set I is infinite, then the only possible accumulation point of $(\mu_i)_{i\in I}$ is 0. For simplicity, we may assume in some situations that the eigenvalues are ordered nonincreasingly in absolute value, i.e., we have $|\mu_i(A)| \geq |\mu_{i+1}(A)|$ such that $(\mu_i)_{i\in I}$ converges to 0 if I is infinite. This convention is particularly convenient if A is *positive*, i.e. $\mu_i(A) \geq 0$ for all $i \in I$.

Singular value decomposition. Every compact operator $A : H \to H'$ on Hilbert spaces admits a *singular value decomposition*. That is, there exist an either finite or countably infinite index set $I = \{1, 2, \dots\}$ and orthonormal systems of *left singular vectors* $(u_i)_{i\in I} \subset H'$ and *right singular vectors* $(v_i)_{i\in I} \subset H$ such that

$$A = \sum_{i\in I} \rho_i(A)\, u_i \otimes v_i, \tag{2.2}$$

where $(\rho_i(A))_{i\in I} \subset \mathbb{R}_+$ are the strictly positive and nonincreasingly ordered *singular values* of A. The convergence in (2.2) is again meant with respect to the operator norm. The *rank* of A is defined as the cardinality of I and written as $\operatorname{rank}(A)$. We additionally set $\operatorname{rank}(A) = \infty$ if A is noncompact. The singular values of A can be obtained from the nonincreasingly ordered nonzero eigenvalues of the positive self-adjoint operator A^*A, i.e., we have $A^*Av_i = \rho_i(A)^2 v_i$ and therefore $\mu_i(A^*A) = \rho_i(A)^2$.

Schatten classes. For integers $1 \leq p < \infty$, the *p-Schatten class* $S_p(H, H')$ consists of all compact operators A from H to H' such that the norm $\|A\|_{S_p(H,H')} := \|(\rho_i(A))_{i\in I}\|_{\ell_p}$ is finite. Here $\|(\rho_i(A))_{i\in I}\|_{\ell_p}$ denotes the ℓ_p sequence space norm of the sequence of the singular values of A. We set $S_\infty(H, H')$ to be the class of compact operators from H to H' equipped with the operator norm and write $S_p(H) := S_p(H, H)$ for all $1 \leq p \leq \infty$. It is clear that $\|A\|_{S_q(H,H')} \leq \|A\|_{S_p(H,H')}$ holds for $1 \leq p \leq q \leq \infty$, i.e., $S_p(H, H') \subseteq S_q(H, H')$. For integer numbers $1 \leq p, q, r \leq \infty$ satisfying $1/p + 1/p = 1/r$ and operators $A \in S_p(H, H')$ and $B \in S_q(H', H'')$, we have $BA \in S_r(H, H'')$. Furthermore, a composition of a p-Schatten operator with any bounded operator yields a p-Schatten operator again.

For $p = 1$, we obtain the Banach space of *trace class operators* $S_1(H, H')$. For a CONS $(e_i)_{i\in I}$ in H, we define the *trace* of an operator $A \in S_1(H)$ as

$$\operatorname{Tr}(A) := \sum_{i\in I} \langle Ae_i,\, e_i\rangle_H < \infty,$$

which induces the inner product

$$\langle A_1,\, A_2\rangle_{S_2(H,H')} := \operatorname{Tr}(A_1^* A_2) = \operatorname{Tr}(A_2^* A_1)$$

on $S_2(H, H')$. We call the resulting Hilbert space equipped with this inner product the space of *Hilbert–Schmidt operators*. The trace and the Hilbert–Schmidt inner product are independent of the chosen CONS $(e_i)_{i\in I}$. For $A, B \in S_2(H)$, we additionally have $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$.

The p-Schatten classes are the completion of *finite-rank operators* (i.e., operators in $\operatorname{span}\{x' \otimes x \mid x \in H, x' \in H'\}$) 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 from H to H', i.e., we have $S_2(H, H') \simeq H' \otimes H$. For elements $x_1, x_2 \in H$, $x_1', x_2' \in H'$, we have the relation $\langle x_1' \otimes x_1,\, x_2' \otimes x_2\rangle_{H'\otimes H} = \langle x_1' \otimes x_1,\, x_2' \otimes x_2\rangle_{S_2(H,H')}$, where the tensors are interpreted as rank-one operators as described above. This identification of tensors as rank-one operators extends to $\operatorname{span}\{x' \otimes x \mid x \in H, x' \in H'\}$ by linearity and defines a linear isometric isomorphism between $H' \otimes H$ and $S_2(H, H')$, which can also be seen by considering Hilbert–Schmidt operators in terms of their singular value decompositions. We will frequently switch between these two viewpoints when considering Hilbert–Schmidt operators.

2.3. Conditional expectations and Markov kernels

We consider random variables $X, Y : \Omega \to E$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in a measurable space $(E, \mathcal{F}_E)$. Throughout this thesis, unless explicitly stated otherwise, the space $(E, \mathcal{F}_E)$ is a standard Borel space, i.e., a Polish space equipped with its Borel field.

We assume without loss of generality that $(\Omega, \mathcal{F}, \mathbb{P})$ is rich enough to support all performed operations. For a finite number of random variables $X_1, \dots, X_n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in E, we write $\mathcal{L}(X_1, \dots, X_n)$ for the joint *law* of $(X_1, \dots, X_n)$ on $(E^n, \mathcal{B}(E^n))$, which is the pushforward measure $\mathcal{L}(X_1, \dots, X_n) := \mathbb{P} \circ \gamma^{-1}$ of the coordinate map $\gamma : \Omega \to E^n$ given by

$$\gamma(\omega) = (X_1(\omega), \dots, X_n(\omega)).$$

We write $X \overset{d}{=} Y$, if the random variables X and Y are equal in distribution, i.e., their laws are equal. If $\mathcal{L}(X) = \pi$ for some probability measure π on $(E, \mathcal{F}_E)$, we use the shorthand $X \sim \pi$.

Whenever $X \sim \pi$, we have $f \in L^1(E, \mathcal{F}_E, \pi; B)$ if and only if $f \circ X \in L^1(\Omega, \mathcal{F}, \mathbb{P}; B)$. In this case, the *change of variables* formula

$$\int_{\mathcal{A}} f(x)\, \mathrm{d}\pi(x) = \int_{X^{-1}(\mathcal{A})} f(X)\, \mathrm{d}\mathbb{P}$$

holds for all $\mathcal{A} \in \mathcal{F}_E$. The *expectation* of $f \in L^1(E, \mathcal{F}_E, \pi; B)$ is defined as

$$\mathbb{E}[f(X)] := \int_E f(x) \, \mathrm{d}\pi(x) = \int_\Omega f(X) \, \mathrm{d}\mathbb{P}.$$

Conditional expectation. Let $X \sim \pi$ and $Y \sim \nu$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(E, \mathcal{F}_E)$. For every σ-field $\mathcal{H} \subseteq \mathcal{F}$ and $f \in L^1(E, \mathcal{F}_E, \nu; B)$, we obtain the *conditional expectation* of $f(Y)$ with respect to $\mathcal{H}$ as the $\mathbb{P}$-a.e. uniquely defined function $\mathbb{E}[f(Y) \,|\, \mathcal{H}] \in L^1(\Omega, \mathcal{H}, \mathbb{P}; B)$ which satisfies

$$\int_\mathcal{A} \mathbb{E}[f(Y) \,|\, \mathcal{H}] \, \mathrm{d}\mathbb{P} = \int_\mathcal{A} f(Y) \, \mathrm{d}\mathbb{P} =: \mathbb{E}[f(Y) \,|\, \mathcal{A}]$$

for all events $\mathcal{A} \in \mathcal{H}$. As a special case, we obtain the *conditional probability* of events $\mathcal{A} \in \mathcal{F}_E$ as

$$\mathbb{P}[Y \in \mathcal{A} \,|\, \mathcal{H}] := \mathbb{E}[\mathbb{1}_\mathcal{A}(Y) \,|\, \mathcal{H}].$$

We set

$$\mathbb{E}[f(Y) \,|\, X] := \mathbb{E}[f(Y) \,|\, \sigma(X)],$$

where $\sigma(X) \subseteq \mathcal{F}$ is the σ-field generated by X. For every $f \in L^1(E, \mathcal{F}_E, \nu; B)$, there exists a $\sigma(X) - \mathcal{F}_B$ measurable function $\xi_f : E \to B$ such that the so-called *Doob–Dynkin representation*

$$\mathbb{E}[f(Y) \,|\, X] = \xi_f(X)$$

holds $\mathbb{P}$-a.e. (see for example Kallenberg, 2002, Lemma 1.13). In this representation, the map ξ_f is uniquely determined π-a.e. on E.

Given $x \in E$, the above construction allows us to evaluate pointwise expressions of the form $\mathbb{E}[f(Y) \,|\, X = x] := \xi_f(x)$ as well as $\mathbb{P}[Y \in \mathcal{A} \,|\, X = x] := \mathbb{E}[\mathbb{1}_\mathcal{A}(Y) \,|\, X = x] = \xi_{\mathbb{1}_\mathcal{A}}(x)$ for events $\mathcal{A} \in \mathcal{F}_E$. The next section shows that it is possible to disintegrate joint probability distributions with respect to this fiberwise representation.

Regular conditional distribution and Markov kernel. Let $p : E \times \mathcal{F}_E \to \mathbb{R}$ be a *Markov kernel*[1] of the conditional distribution of Y given X, i.e.,

(i) $p(x, \cdot)$ is a probability measure on $(E, \mathcal{F}_E)$ for every $x \in E$ and

(ii) for every $\mathcal{A} \in \mathcal{F}_E$, the map $x \mapsto p(x, \mathcal{A})$ is an $\mathcal{F}_E - \mathcal{B}(\mathbb{R})$ measurable function

[1] We distinguish different notions of *kernels* in this thesis. In what follows, we will often refer to reproducing kernels/symmetric positive semidefinite kernels simply as *kernel*, while the kernel p defining a conditional distribution will always be called a *Markov kernel*.

such that

$$\mathbb{P}[Y \in \mathcal{A} \mid X = x] = \mathbb{E}[\mathbb{1}_{\mathcal{A}}(Y)|X = x] = \int_{\mathcal{A}} p(x, \mathrm{d}y) = p(x, \mathcal{A})$$

for all $x \in E$ and events $\mathcal{A} \in \mathcal{F}_E$. The Markov kernel p defines a so-called *regular version* of the above conditional distribution which allows to consider the fiberwise disintegration

$$\mathbb{P}[X \in \mathcal{A}_1, Y \in \mathcal{A}_2] = \mathbb{E}[\mathbb{1}_{\mathcal{A}_1 \times \mathcal{A}_2}(X, Y)] = \int_{\mathcal{A}_1} p(x, \mathcal{A}_2)\, \mathrm{d}\pi(x)$$

for all $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{F}_E$, see Dudley (2002, Theorem 10.2.1). Such a Markov kernel p exists always in our scenario, since the space E is Polish (Dudley, 2002, Theorem 10.2.2). Additionally, two regular versions of the same conditional distribution with corresponding Markov kernels p, p' naturally coincide almost everywhere, i.e., we have $p(x, \cdot) = p'(x, \cdot)$ for π-a.e. $x \in E$.

3. The structure of bivariate distributions

Section 3.5 contains passages taken from Mollenhauer and Koltai (2020). Section 3.6 contains passages taken from Mollenhauer et al. (2020a).

This chapter is named after the paper of Lancaster (1958) about the phenomenon that joint distributions of two random variables may be uniquely described in terms of two sets of maximally correlated transformations—the so-called *canonical variables* or *canonical components*. Two decades earlier, a simplified version of this idea was used by Hotelling (1936), who described the optimal *linear* transformations of two Euclidean random variables leading to a maximal correlation. This approach is now widely known under the name *canonical correlation analysis* (CCA). Various extensions of the classical linear CCA have been introduced in statistics and machine learning in the context of regression (Breiman and Friedman, 1985), functional data analysis (Leurgans et al., 1993), reproducing kernels (Akaho, 2001; Melzer et al., 2001; Bach and Jordan, 2002) and neural networks (Andrew et al., 2013), to name only a few important milestones in this now rapidly evolving field.

The connection between the theory of the canonical components and the underlying measure-theoretic concept of joint probability distributions can be established in terms of the conditional expectation operator and its spectral properties, which we will introduce in this chapter.

Independently of its importance in the context of the canonical components, the conditional expectation operator also plays a crucial role in the context of Markov processes and random dynamical systems. Conditional expectation operators which describe the transition probabilites of Markov processes (so-called *Markov transition operators*) are linked to a multitude of important dynamical features of the underlying process (Lasota and Mackey, 1994; Meyn and Tweedie, 2009; Bovier and Den Hollander, 2016; Douc et al., 2018). This fact has lead to a vast variety of methods for the data-driven approximation of Markov transition operators and their adjoints (see Li 1976; Ding and Li 1991; Dellnitz and Junge 1999; Huisinga 2001; Junge and Koltai 2009; Schmid 2010; Pérez-Hernández et al. 2013; Noé and Nüske 2013; Schütte and Sarich 2013; Williams et al. 2015a,b; Klus et al. 2016, 2018, 2020; Wu and Noé 2020; Tian and Wu 2020 and the references therein).

Although these methods have been gravitating at least partly towards concepts from machine learning over the last years, the mathematical formalism under which they are derived is often influenced by classical numerical theory rather than a statistical viewpoint. However, connections between the canonical components and the dynamical theory have recently been used in applied scenarios (Koltai et al., 2018; Klus et al., 2019; Wu and Noé, 2020).

In this chapter, we give a general overview of the functional-analytic theory of the conditional expectation operator P, its empirical estimation as well as the special case that P is a Markov transition operator. We investigate a basic parametric projection scheme for the approximation of P which is widely used in practice to analyze stochastic dynamical systems and is strongly related to the theory of the canonical components. To the best of our knowledge, a statistical analysis of this approach has not been conducted yet. *Our endeavour to derive bounds for the estimation error illustrates that projection methods may exhibit typical theoretical limitations of parametric statistical models in high dimensions. This fact eventually motivates the investigation of tools from nonparametric inference for the remainder of this work.*

3.1. Overview

We define the conditional expectation operator in Section 3.2 and connect it to the theory of the canonical components in Section 3.3. In Section 3.4, we discuss a well-known parametric projection method for the estimation of P and its spectral properties and derive basic convergence results based on independent observations. Section 3.5 establishes the connection between the conditional expectation operator and the transition probabilies of Markov processes. Additionally, we give a brief overview of practical applications of the aforementioned projection method in the context of stochastic processes. In Section 3.6, we investigate the statistical details of the estimation of P when the underlying data is dependent by introducing the concepts of ergodicity and mixing.

3.2. Conditional expectation operators

As previously mentioned, we consider two random variables $X \sim \pi$ and $Y \sim \nu$ defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the standard Borel space $(E, \mathcal{F}_E)$. The *conditional expectation operator* $P\colon L^2(\nu) \to L^2(\pi)$ is defined by

$$[Pf](x) := \mathbb{E}[f(Y) \mid X = x] = \int_E f(y)\, p(x, \mathrm{d}y).$$

The operator P is a contractive linear map, which can easily be seen by making use of Jensen's inequality for conditional expectations and considering

$$\|Pf\|^2_{L^2(\pi)} = \mathbb{E}\left[\mathbb{E}[f(Y) \mid X]^2\right] \leq \mathbb{E}\left[\mathbb{E}[(f(Y))^2 \mid X]\right] = \mathbb{E}[f(Y)^2] = \|f\|^2_{L^2(\nu)}\,. \tag{3.1}$$

Note that by the law of total expectation, we have the relation

$$\langle f,\, Pg\rangle_{L^2(\pi)} = \mathbb{E}[f(X)g(Y)]$$

for all $f \in L^2(\pi)$, $g \in L^2(\nu)$. Considering the adjoint of the operator P and making use of the law of total expectation with respect to conditioning on Y, we have

$$\langle P^* f,\, g\rangle_{L^2(\nu)} = \mathbb{E}[f(X)g(Y)] = \mathbb{E}[\,\mathbb{E}[f(X)|Y]\,g(Y)]$$

for all $f \in L^2(\pi)$, $g \in L^2(\nu)$. This lets us identify the adjoint of P as the operator $P^* : L^2(\pi) \to L^2(\nu)$ satisfying

$$[P^* f](y) = \mathbb{E}[f(X) \mid Y = y] = \int_E f(x)\, p'(y, \mathrm{d}x)$$

where $p' : E \times \mathcal{F}_E \to \mathbb{R}$ is some Markov kernel associated with the conditional distribution of X given Y.

3.3. Nonlinear canonical components

The understanding of the relationship between two random variables X and Y is a fundamental problem in probability theory and statistics. Since the joint distribution of X and Y can potentially be very complicated, an interpretable simplification of their connection is of great interest. Such a simplification is reflected in the dominant spectral properties of P, which leads to the theory of canonical components.

3.3.1. Maximizing the sum of correlations

In order to understand the relationship between X and Y, we may aim to express it in terms of a finite number of characteristics of maximal statistical importance. However, the term *importance* is vague in this context and needs to be interpreted properly.

A way to solve this problem is to search for a finite number of pairwise decorrelated transformations which maximize the sum of cross-correlations when applied to X and Y. In other words, for some finite number r, we try to find two sets of functions

$$F := \{f_1, \ldots, f_r\} \subset L^2(\pi) \text{ and } G := \{g_1, \ldots, g_r\} \subset L^2(\nu)$$

satisfying $\mathbb{E}[f_i(X)f_j(X)] = \delta_{ij}$ and $\mathbb{E}[g_i(Y)g_j(Y)] = \delta_{ij}$ such that the sum of correlations

$$\sum_{i=1}^{r} \mathbb{E}[f_i(X)g_i(Y)]$$

between the transformations $f_i(X)$ and $g_i(Y)$ is maximal. We reformulate this idea as the optimization problem

$$\begin{aligned} &\max_{F,G} \sum_{i=1}^{r} \langle f_i,\, Pg_i \rangle_{L^2(\pi)} \\ &\qquad \text{subject to} \\ &\langle f_i,\, f_j \rangle_{L^2(\pi)} = \delta_{ij} \text{ and } \langle g_i,\, g_j \rangle_{L^2(\nu)} = \delta_{ij}. \end{aligned} \tag{3.2}$$

For convenience, we always consider the number $r \leq \operatorname{rank}(P)$ fixed here. If the operator P is compact, it is known that the analytical maximum of (3.2) is attained for the left and right singular functions associated with the r largest singular values of P.

Theorem 3.3.1 (Canonical components)**.** *Let P be compact with the singular value decomposition*

$$P = \sum_{i \in I} \rho_i(P)\, u_i \otimes v_i.$$

Furthermore, let $r \leq \operatorname{rank}(P)$ be fixed and let

$$U := \{u_1, \ldots, u_r\} \subset L^2(\pi) \text{ and } V := \{v_1, \ldots, v_r\} \subset L^2(\nu) \tag{3.3}$$

denote the first r left and right singular functions of P. Then U and V solve the optimization problem (3.2), i.e., we have

$$\max_{F,G} \sum_{i=1}^{r} \langle f_i,\, Pg_i \rangle_{L^2(\pi)} = \sum_{i=1}^{r} \langle u_i,\, Pv_i \rangle_{L^2(\nu)} = \sum_{i=1}^{r} \rho_i(P) \tag{3.4}$$

where F and G range over all orthonormal sets in $L^2(\pi)$ and $L^2(\nu)$ of cardinality r in the maximum on the left-hand side.

A proof of Theorem 3.3.1 is given by Gohberg and Kreĭn (1969, Chapter II, Lemma 4.1) in a more general context. It can also be seen as a consequence of the well-known *Courant–Fischer minmax theorem* for the singular values of compact operators (Weidmann, 1980, Theorem 7.7).

Remark 3.3.2 (Unique solution and order of singular functions). Note that the sets of first r singular functions U and V may not be unique whenever we have $r < \operatorname{rank}(P)$ and $\rho_r(P) = \rho_{r+1}(P)$. In this case, one may change the order of singular functions associated with $\rho_r(P)$, leading to alternative choices for U and V. If however $\rho_r(P) \neq \rho_{r+1}(P)$, the sets U and V and hence also the solution of (3.2) are unique. For simplicity, we may therefore occasionally assume $r < \operatorname{rank}(P)$ and $\rho_r(P) \neq \rho_{r+1}(P)$ in what follows.

The functions in U and V are called the first r *canonical components* of X and Y.

Remark 3.3.3 (Compactness of P). Note that sufficient conditions for the compactness of P can easily be derived by applying the theory of integral operators (Halmos and Sunder, 1978) to the structure of the underlying Markov kernel. Several authors address this topic in our context (Dellnitz and Junge, 1999; Michaeli et al., 2016; Wu and Noé, 2020) and show that P is Hilbert–Schmidt under the assumption that the Markov kernel p admits a square integrable *transition density* with respect to some reference measure on the measurable space $(E^2, \mathcal{F}_E^{\otimes 2})$.

Remark 3.3.4 (Trivial first singular component). By (3.1), we have $\rho_1(P) = \|P\| \leq 1$. Note that for the ν-a.e. constant function $g \equiv 1$, we have $\|g\|_{L^2(\nu)} = 1$ and $Pg \equiv 1$ π-a.e., which shows

$$\|Pg\|_{L^2(\pi)} = 1 = \|P\|\,.$$

Hence we have $\rho_1(P) = 1$, where the two corresponding singular functions u_1 and v_1 are almost everywhere constant unit functions. The first singular component of P is therefore always trivial and does not carry relevant information. It is sufficient to consider only $r - 1$ remaining top singular components in practical applications.

3.3.2. Empirical estimation and spectral perturbation

We now assume that we have access to a generic compact empirical estimate of the compact operator P in terms of $\widehat{P} : L^2(\nu) \to L^2(\pi)$, i.e., $\widehat{P}$ is a compactly perturbed version of P. We assume further that we can compute the empirical singular value decomposition

$$\widehat{P} = \sum_{i \in \widehat{I}} \rho_i(\widehat{P})\, \widehat{u}_i \otimes \widehat{v}_i.$$

For $r \leq \min\{\operatorname{rank}(\widehat{P}), \operatorname{rank}(P)\}$ we define the sets of top r left and right empirical singular vectors associated with the empirical singular values $\rho_1(\widehat{P}), \ldots, \rho_r(\widehat{P})$ as $\widehat{U} := \{\widehat{u}_1, \ldots, \widehat{u}_r\}$ and $\widehat{V} := \{\widehat{v}_1, \ldots, \widehat{v}_r\}$, which we will use as estimates of the first r canonical components U and V as defined in (3.3).

One would expect that if $\widehat{P}$ is a good approximation of P, then $\widehat{U}$ and $\widehat{V}$ are good approximations of U and V. A theoretical justification of this approach can be given in terms of the perturbation of singular subspaces, which we will introduce now.

We measure the distances between the top r empirical singular functions $\widehat{U}$ and $\widehat{V}$ and the true canonical components U and V in terms of the orthogonal projectors associated with the corresponding finite-dimensional subspaces. This approach is standard practice in matrix analysis and operator perturbation theory (Stewart and Sun 1990, Chapter

I.5 and Chapter V as well as Bhatia 1997, Chapter VII) and is often used in high-dimensional statistics. We write $\Pi_U : L^2(\pi) \to L^2(\pi)$ for the orthogonal projector onto $\operatorname{span} U \subset L^2(\pi)$ given by

$$\Pi_U = \sum_{i=1}^{r} u_i \otimes u_i \tag{3.5}$$

and define $\Pi_V : L^2(\nu) \to L^2(\nu)$ analogously as well as $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ as the empirical counterparts of Π_U and Π_V.

The operator norm $\left\|\Pi_{\widehat{U}} - \Pi_U\right\|$ of the distance between the projectors is sometimes called the *aperture* (Akhiezer and Glazman, 1993) or *gap* (Stewart and Sun, 1990) between the corresponding subspaces $\operatorname{span} U$ and $\operatorname{span} \widehat{U}$. The Hilbert–Schmidt distance $\left\|\Pi_{\widehat{U}} - \Pi_U\right\|_{S_2(L^2(\pi))}$ allows for a similar natural interpretation, as both of these measures can be written in terms of the so-called *canonical angles* between the associated subspaces. We review the theory of distances between subspaces and spectral perturbation in more detail in Appendix A.4 and argue that the Hilbert–Schmidt distance between the orthogonal projectors is actually a fairly natural choice for this problem. For one-dimensional projections, one can easily see that the Hilbert–Schmidt distance bounds the distance between the corresponding singular vectors, i.e., for all $1 \leq i \leq r$ we have

$$\begin{aligned}
\|\Pi_{u_i} - \Pi_{\widehat{u}_i}\|^2_{S_2(L^2(\pi))} &= \left\|(u_i \otimes u_i) - (\widehat{u}_i \otimes \widehat{u}_i)\right\|^2_{S_2(L^2(\pi))} \\
&= \|u_i \otimes u_i\|^2_{S_2(L^2(\pi))} - 2\left\langle u_i \otimes u_i,\, \widehat{u}_i \otimes \widehat{u}_i \right\rangle_{S_2(L^2(\pi))} + \|\widehat{u}_i \otimes \widehat{u}_i\|^2_{S_2(L^2(\pi))} \\
&= \|u_i\|^4_{L^2(\pi)} - 2\left\langle u_i,\, \widehat{u}_i \right\rangle^2_{L^2(\pi)} + \|\widehat{u}_i\|^4_{L^2(\pi)} \\
&= 2 - 2\left\langle u_i,\, \widehat{u}_i \right\rangle^2_{L^2(\pi)} \\
&\geq 2 - 2\left\langle u_i,\, \widehat{u}_i \right\rangle_{L^2(\pi)} = \|u_i - \widehat{u}_i\|^2_{L^2(\pi)},
\end{aligned}$$

whenever we assume that the orientation of u_i and $\widehat{u}_i$ is given in a way such that $0 \leq \langle u_i, \widehat{u}_i \rangle_{L^2(\pi)}$ holds (see also Zwald and Blanchard 2006). Note that in the above derivation, we also make use of $\langle u_i, \widehat{u}_i \rangle_{L^2(\pi)} \leq 1$ by Cauchy–Schwarz.

Let $\Delta := P - \widehat{P}$ denote the perturbation introduced by the estimation procedure. We modify a well-known version of the *Davis–Kahan theorem* (Davis and Kahan, 1970; Wedin, 1972) due to Yu et al. (2015) in Appendix A.4 and obtain the following spectral stability bound which quantifies the maximal distance between the singular subspaces with respect to the global estimation error $\|\Delta\|$.

Theorem 3.3.5 (Singular subspace perturbation)**.** *Let P and $\widehat{P}$ be compact and choose $r < \min\{\operatorname{rank}(P), \operatorname{rank}(\widehat{P})\}$. Furthermore, assume $\rho_r(P) \neq \rho_{r+1}(P)$. Then we have*

$$\max\left\{\left\|\Pi_{\widehat{U}} - \Pi_U\right\|_{S_2(L^2(\pi))}, \left\|\Pi_{\widehat{V}} - \Pi_V\right\|_{S_2(L^2(\nu))}\right\} \leq \frac{2\sqrt{2r}(2\|\Delta\| + \|\Delta\|^2)}{\rho_r(P)^2 - \rho_{r+1}(P)^2}.$$

Proof. Theorem 3.3.5 is obtained as a combination of Theorem A.4.5 with Lemma A.4.3 in the appendix and the fact that we have $\rho_1(P) = 1$. ■

3.4. A parametric model: projection method

We will now review a standard type of parametric numerical projection method (see for example Klus et al., 2016, 2018; Wu and Noé, 2020) for the estimation of P. Different versions of this method are widely used for the approximation of spectral properties of P in the case of random dynamical systems and Markov processes as we will later describe in Section 3.5. To the best of our knowledge, no abstract error analysis from a statistical perspective has been conducted for these approaches.

In order to discretize P, we choose two orthonormal systems $\Phi := \{\phi_1, \dots, \phi_s\} \subseteq L^2(\pi)$ and $\Psi := \{\psi_1, \dots, \psi_s\} \subseteq L^2(\nu)$ and consider the finite-dimensional projection of P given by

$$P_{\Phi,\Psi} := \Pi_\Phi P \Pi_\Psi.$$

Here, Π_Φ and Π_Ψ are the orthogonal projections onto the ansatz spaces $\operatorname{span}\Phi \subseteq L^2(\pi)$ and $\operatorname{span}\Psi \subseteq L^2(\nu)$. We note that it is generally possible to choose sets Φ and Ψ of different cardinality. However, we restrict ourselves to the case that both sets contain the same number of basis functions for simplicity. We assume that we have access to independent and identically distributed (iid) sample pairs $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{L}(X, Y)$ from the joint distribution of X and Y. The finite-dimensional operator $P_{\Phi,\Psi}$ can now simply be estimated in terms of a matrix, which we will describe in this section. In the case of Markov transition operators, one typically works with data obtained from a realization of the underlying process and therefore faces non-iid samples, which we will address in more detail in Section 3.6.

3.4.1. Empirical basis orthonormalization

Since the true marginals π and ν of X and Y are not known in practical applications, it is generally not possible to choose the orthonormal ansatz functions $\Phi \subseteq L^2(\pi)$ and $\Psi \subseteq L^2(\nu)$ a priori. Instead, Φ and Ψ can be chosen as linearly independent sets of functions and orthonormalized empirically. This procedure is performed in various applications in the context of Markov processes (Williams et al., 2015a; Koltai et al., 2018; Wu and Noé, 2020). The empirical orthonormalization can be interpreted as a form of *whitening* from a statistical perspective (Kessy et al., 2018), which we illustrate for the ansatz functions Φ in what follows.

We define the *Gramian matrix* $\mathbf{G} = (\mathbf{G}_{ij})_{1\leq i,j\leq s} \in \mathbb{R}^{s\times s}$ associated with Φ as

$$\mathbf{G}_{ij} := \langle \phi_i, \phi_j \rangle_{L^2(\pi)}, \quad 1 \leq i, j \leq s.$$

Since we assume that we have access to iid samples $X_1, \ldots, X_n \sim \pi$, we can compute the *empirical Gramian matrix* $\widehat{\mathbf{G}} = (\widehat{\mathbf{G}}_{ij})_{1\leq i,j\leq s} \in \mathbb{R}^{s\times s}$ as an approximation of the true Gramian as

$$\widehat{\mathbf{G}}_{ij} := \frac{1}{n}\sum_{k=1}^{n} \phi_i(X_k)\phi_j(X_k), \quad 1 \leq i, j \leq s.$$

Both $\mathbf{G}$ and $\widehat{\mathbf{G}}$ are symmetric. Because Φ contains linearly independent functions in $L^2(\pi)$, the matrix $\mathbf{G}$ is strictly positive. We note that $\widehat{\mathbf{G}}$ is only positive semidefinite in general. However, we first introduce the whitening procedure assuming that $\widehat{\mathbf{G}}$ is $\mathbb{P}$-a.e. invertible. We relax this assumption later on.

Invertible empirical Gramian $\widehat{\mathbf{G}}$. Let the empirical matrix $\widehat{\mathbf{G}}$ be $\mathbb{P}$-a.e. invertible. We note that a necessary condition for the invertibility of $\widehat{\mathbf{G}}$ is $n \geq s$. Interpreting Φ as an s-dimensional column vector of functions, we define a set of empirically orthonormalized ansatz functions $\bar{\phi}_i$ as linear combinations of the ansatz functions ϕ_i in terms of the vector $\bar{\Phi}$ given by

$$\bar{\Phi} := \widehat{\mathbf{G}}^{-1/2}\Phi, \tag{3.6}$$

i.e., the i-th row of $\widehat{\mathbf{G}}^{-1/2}$ contains the coordinate vector expression of the function $\bar{\phi}_i$ in terms of the basis Φ.

Our newly constructed ansatz functions $\bar{\Phi}$ satisfy

$$\left\langle \bar{\phi}_i, \bar{\phi}_j \right\rangle_{L^2(\pi)} = (\widehat{\mathbf{G}}^{-1/2}\mathbf{G}\widehat{\mathbf{G}}^{-1/2})_{ij} \approx \delta_{ij}, \quad 1 \leq i, j \leq s$$

whenever $\widehat{\mathbf{G}}^{-1/2}$ is close to $\mathbf{G}^{-1/2}$ in operator norm. We now investigate how well $\widehat{\mathbf{G}}^{-1/2}$ approximates $\mathbf{G}^{-1/2}$. First, we begin with a bound of the estimation error $\widehat{\mathbf{G}} - \mathbf{G}$ under the assumption that the basis functions Φ are $\mathbb{P}$-a.e. uniformly bounded by some constant.

Lemma 3.4.1 (Estimation of Gramian)**.** *Let $|\phi_i(X)| \leq M$ $\mathbb{P}$-a.e. for all $1 \leq i \leq s$. Then for every $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|_F \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{8s^2M^4}\right), \tag{3.7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. We define the $\mathbb{R}^{s\times s}$-valued random variable $\xi(X) := (\phi_i(X)\phi_j(X))_{1\leq i,j\leq s}$. Note that $\mathbf{G} = \mathbb{E}[\xi(X)]$ as well as $\widehat{\mathbf{G}} = \frac{1}{n}\sum_{k=1}^{n}\xi(X_k)$ hold. Furthermore, we have

$$\|\xi(X)\|_F^2 = \sum_{1\leq i,j\leq s}\phi_i(X)^2\phi_j(X)^2 \leq s^2M^4 \quad \mathbb{P}\text{-a.e.} \tag{3.8}$$

We apply the vector-valued Hoeffding inequality due to Pinelis (1994) given in Corollary A.5.2 to the zero-mean iid random variables $\xi(X_1) - \mathbb{E}[\xi(X)], \ldots, \xi(X_n) - \mathbb{E}[\xi(X)]$ which are $\mathbb{P}$-a.e. uniformly bounded in Frobenius norm by $2sM^2$ and obtain the assertion. ■

The next result gives the desired statement.

Theorem 3.4.2 (Whitening)**.** *Assume that* $\operatorname{rank}(\mathbf{G}) = s$ *and* $|\phi_i(X)| \leq M$ $\mathbb{P}$*-a.e. for all* $1 \leq i \leq s$*. Furthermore assume that* $\operatorname{rank}(\widehat{\mathbf{G}}) = s$ *holds* $\mathbb{P}$*-a.e. Then for every* $\epsilon > 0$*, we have*

$$\mathbb{P}\left[\left\|\widehat{\mathbf{G}}^{-1/2} - \mathbf{G}^{-1/2}\right\| \geq \sqrt{\frac{\epsilon}{\mu_s(\mathbf{G})\mu_s(\widehat{\mathbf{G}})}}\right] \leq 2\exp\left(-\frac{n\epsilon^2}{8s^2M^4}\right),$$

where $\mu_s(\mathbf{G}) > 0$ *and* $\mu_s(\widehat{\mathbf{G}}) > 0$ *are the smallest eigenvalues of* $\mathbf{G}$ *and* $\widehat{\mathbf{G}}$*, respectively.*

The proof can be found in Appendix A.1. Before we discuss the interpretation of Theorem 3.4.2, we abandon the assumption that $\widehat{\mathbf{G}}$ is $\mathbb{P}$-a.e. invertible.

Noninvertible empirical Gramian $\widehat{\mathbf{G}}$. In practice, it may certainly occur with some probability that the empirical estimate $\widehat{\mathbf{G}}$ is not of full rank or the numerical computation of $\widehat{\mathbf{G}}^{-1/2}$ is numerically not feasible. In this case, the empirical transformation (3.6) results in an ill-posed inverse problem which requires a regularization strategy (see Appendix A.6 for a brief overview of the theory of inverse problems). In practical applications, the underlying empirical regularization problem is sometimes solved via pseudoinverse matrices (Williams et al., 2015a). Here we derive a basic error bound to sketch the difficulties which may arise for regularized analogues of (3.6) using the example of the well-known Tikhonov–Phillips regularization scheme (Tikhonov and Arsenin, 1977). The Tikhonov–Phillips regularization scheme replaces the potentially nonexistent empirical inverse $\widehat{\mathbf{G}}^{-1/2}$ with the always existing regularized estimate $(\widehat{\mathbf{G}} + \lambda\mathbf{I})^{-1/2}$ with some appropriately small regularization parameter $\lambda > 0$. It is possible to obtain operator norm bounds for the estimation error

$$(\widehat{\mathbf{G}} + \lambda\mathbf{I})^{-1/2} - \mathbf{G}^{-1/2}$$

and deduce the optimal choice of λ depending on the number of samples n such that convergence is guaranteed in the infinite sample limit. We provide such a bound in the next result.

Theorem 3.4.3 (Regularized whitening). *Assume that* $\operatorname{rank}(\mathbf{G}) = s$ *and* $|\phi_i(X)| \leq M$ $\mathbb{P}$*-a.e. for all* $1 \leq i \leq s$*. Then for every regularization parameter* $\lambda > 0$ *and every* $\epsilon > 0$*, we have*

$$\mathbb{P}\left[\left\|(\widehat{\mathbf{G}} + \lambda \mathbf{I})^{-1/2} - \mathbf{G}^{-1/2}\right\| \geq \frac{\sqrt{\epsilon}}{\lambda} + \frac{\sqrt{\lambda}}{\mu_s(\mathbf{G})}\right] \leq 2 \exp\left(-\frac{n\epsilon^2}{8s^2M^4}\right),$$

where $\mu_s(\mathbf{G}) > 0$ *is the smallest eigenvalue of* $\mathbf{G}$*.*

The proof can be found in Appendix A.1. In essence, Theorem 3.4.3 highlights the fact that an optimally regularized estimate $(\widehat{\mathbf{G}} + \lambda \mathbf{I})^{-1/2}$ requires a reasonable choice of the regularization parameter λ. In particular, λ balances the error introduced by the regularization itself (depending on the degree of ill-posedness of the analytical inversion problem expressed in terms of $\frac{\sqrt{\lambda}}{\mu_s(\mathbf{G})}$) and the smoothing effect on the noise of the corresponding estimator (given in terms of $\frac{\sqrt{\epsilon}}{\lambda}$). This problem is typical in regularization theory (see Engl et al. 1996 and Appendix A.6), as we will observe multiple times throughout this thesis.

Estimation in high dimensions. Without further considering the intricacies of individual regularization techniques at this point, we can already see that the estimation and inversion of high-dimensional matrices can make the empirical orthonormalization of the basis functions a challenging problem in practice. In general, the number of samples n must be scaled quadratically relative to the number of ansatz functions s to maintain the same level of error confidence when deriving typical error bounds as given in Lemma 3.4.1, Theorem 3.4.2 and Theorem 3.4.3. This is due to the fact that the norms of $\mathbf{G}$ and $\widehat{\mathbf{G}}$ can grow rapidly as the dimension $s \times s$ increases. Similar phenomena are investigated throughout the field of high-dimensional statistics (Vershynin, 2018; Wainwright, 2019). We also note that in addition to the infeasible invertibility assumption in Theorem 3.4.2, the error bound depends on the minimal eigenvalue of the empirical Gramian, which is a random quantity itself. Hence, Theorem 3.4.2 is not really useful in practice and only serves the purpose of illustrating the whitening procedure in theory. In practical scenarios, bounds for appropriate regularization schemes like Theorem 3.4.3 allow to circumvent this issue. It is likely that our bounds can be improved by employing deeper results from matrix analysis and estimators which are used in high-dimensional statistics. However, applying these techniques to our scenario may introduce some technical difficulties and requires some work, which we elaborate on below.

Remark 3.4.4 (Stronger error bounds). Our error bounds rely on the application of a Hoeffding-type bound, which requires fairly mild assumptions in terms of boundedness of the random matrices. It is possible to straightforwardly apply a vector-valued version of Bernstein's inequality instead (Pinelis and Sakhanenko 1986, see Appendix A.5.2). This

requires incorporating higher moment bounds for the random matrices, which may again depend on the dimension $s \times s$ without further structural assumptions about the basis functions and hence lead to the same challenges as shown here.
In a broader sense, the estimation of $\mathbf{G}$ is closely related to various problems in high-dimensional statistics and compressed sensing such as the estimation of covariance matrices and matrix completion (see for example Wainwright 2019, Chapters 6 and 10 and the literature reviews therein). Although there exists a large amount of work on the efficient estimation of high-dimensional matrices and related concentration results (Vershynin, 2012; Tropp, 2015), we face a situation which does not align with some of the typical assumptions imposed in these scenarios. In particular

(i) the estimate $\widehat{\mathbf{G}}$ consists of dependent entries and

(ii) the matrix $\mathbf{G}$ is of full rank s by assumption and we therefore expect the estimate $\widehat{\mathbf{G}}$ to have a large rank with high probability.

A reasonable way to improve the estimation of $\mathbf{G}$ might be the application of *thresholding estimators* which are commonly used for high-dimensional covariance matrices (Bickel and Levina, 2008; Cai and Liu, 2011; Cai and Yuan, 2012). However, these approaches require additional assumptions in terms of sparsity of $\mathbf{G}$.

3.4.2. Estimation of $P_{\Phi,\Psi}$

We now investigate the estimation of the finite-dimensional operator $P_{\Phi,\Psi}$ and the overall error of the corresponding projection method. Justified by the considerations in the previous section, we now work under the idealized assumption that Φ and Ψ are both orthonormal systems in our further analysis.

Given some generic estimate $\widehat{P}_{\Phi,\Psi}$ of $P_{\Phi,\Psi}$, we can decompose the overall estimation error $\Delta = \widehat{P}_{\Phi,\Psi} - P$ in terms of

$$\left\|\Delta\right\| \leq \left\|\widehat{P}_{\Phi,\Psi} - P_{\Phi,\Psi}\right\| + \left\|P_{\Phi,\Psi} - P\right\| \quad \mathbb{P}\text{-a.e.} \tag{3.9}$$

The first term on the right-hand side is the stochastic *sample error* which can alternatively be measured in any arbitrary matrix norm since both involved operators are finite-dimensional. The sample error will be our main object of interest. The second term is the deterministic *model error*, which is introduced by the projection of P onto the ansatz spaces $\operatorname{span}\Phi$ and $\operatorname{span}\Psi$. The model error is investigated in the classical numerical theory of projection methods for integral equations and deterministic operator approximation (Hackbusch, 1995) and will therefore not be in our focus in this work. At this point, it is sufficient for us to note that

(i) if P is noncompact, it is clear that for an increasing dimension s of the ansatz spaces, the model error can never converge to 0. Instead, the distance between the projection $P_{\Phi,\Psi}$ and the analytical operator P can only converge in some weaker topology. If however P is compact, then the model error converges to 0 as $s \to \infty$ (see Appendix A.2), assuming for simplicity that we have $\dim L^2(\pi) = \dim L^2(\nu) = \infty$.

(ii) The model error is generally not known in practical applications, as it is based on the unknown distribution $\mathcal{L}(X, Y)$ and is determined by how well the ansatz spaces $\operatorname{span}\Phi$ and $\operatorname{span}\Psi$ capture the dominant structure of P. Pragmatically, a larger choice of s leads to a smaller model error and we therefore expect to deal with large values of s in practice.

We now focus on the projection estimate $\widehat{P}_{\Phi,\Psi}$ and the corresponding sample error based on iid sample pairs $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{L}(X, Y)$. The finite-rank operator $P_{\Phi,\Psi}$ can be expressed in terms of a matrix with respect to the orthonormal bases Φ and Ψ in terms of $\mathbf{A} = (\mathbf{A}_{ij})_{1 \le i,j \le s} \in \mathbb{R}^{s \times s}$ given by

$$\mathbf{A}_{ij} := \mathbb{E}[\phi_i(X)\,\psi_j(Y)] = \langle \phi_i,\, P\psi_j \rangle_{L^2(\pi)}\,, \quad 1 \le i, j \le s.$$

The matrix $\mathbf{A}$ allows the straighforward construction of an estimate $\widehat{\mathbf{A}} \in \mathbb{R}^{s \times s}$ which represents the action of an estimated finite-rank operator $\widehat{P}_{\Phi,\Psi}$ defined on $\operatorname{span}\Phi$ and $\operatorname{span}\Psi$ in terms of the entrywise Monte Carlo sums

$$\widehat{\mathbf{A}}_{ij} := \frac{1}{n}\sum_{t=1}^{n} \phi_i(X_t)\,\psi_j(Y_t) =: \left\langle \phi_i,\, \widehat{P}_{\Phi,\Psi}\psi_j \right\rangle_{L^2(\pi)}\,, \quad 1 \le i, j \le s.$$

By the strong law of large numbers, we clearly have

$$\widehat{\mathbf{A}}_{ij} \to \mathbf{A}_{ij} \quad \mathbb{P}\text{-a.e.}, \quad 1 \le i, j \le s,$$

and therefore $\widehat{\mathbf{A}} \to \mathbf{A}$ $\mathbb{P}$-a.e. and $\widehat{P}_{\Phi,\Psi} \to P_{\Phi,\Phi}$ $\mathbb{P}$-a.e in any matrix norm and corresponding norm for finite-rank operators, respectively. Analogously to the sample error bound of the empirical Gramian in Lemma 3.4.1, we can derive a sample error bound for $\widehat{P}_{\Phi,\Psi}$.

Theorem 3.4.5 (Sample error)**.** *Let $|\phi_i(X)| \le M$ and $|\psi_i(Y)| \le M$ $\mathbb{P}$-a.e. for all $1 \le i \le s$. Then for every $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left\|\widehat{P}_{\Phi,\Psi} - P_{\Phi,\Psi}\right\|_{S_2(L^2(\nu),L^2(\pi))} \ge \epsilon\right] \le 2\exp\left(-\frac{n\epsilon^2}{8s^2M^4}\right).$$

Proof. We can apply technique of the proof of Lemma 3.4.1 to the matrices $\widehat{\mathbf{A}}$ and $\mathbf{A}$ equipped with the Frobenius norm and use the fact that

$$\left\|\widehat{\mathbf{A}} - \mathbf{A}\right\|_F = \left\|\widehat{P}_{\Phi,\Psi} - P_{\Phi,\Psi}\right\|_{S_2(L^2(\nu),L^2(\pi))}$$

since we assume that Φ and Ψ are orthonormal systems. ∎

Theorem 3.4.5 introduces the same theoretical difficulties which we already discussed for the estimation of the Gramian $\mathbf{G}$ in the previous section to the estimation of $P_{\Phi,\Psi}$. In particular, for the error bound above and comparable generic matrix concentration bounds, the sample size n needs to scale at least quadratically with respect to the dimension s of the ansatz spaces to maintain the level of confidence.

Remark 3.4.6 (Stronger error bounds). The question of stronger error bounds for the estimate $\widehat{P}_{\Phi,\Psi}$ may be answered in terms of the so-called *intrinsic dimension* (also called the *effective rank* or *effective dimension*). Whenever one assumes that the variance structure of $\widehat{P}_{\Phi,\Psi}$ is close to an operator of low rank, bounds which are independent of the ambient dimension $s \times s$ may be derived (see Tropp, 2015, Section 7 for an overview).

3.5. Markov transition operators

The conditional expectation operator P and its spectral properties are of particular interest in the context of Markov processes and dynamical systems. We now connect our setup for the estimation of P from the preceding sections to the theory of Markov transition operators.

3.5.1. General overview

We consider a Markov process $(X_t)_{t\in\mathbb{Z}}$ on the state space $(E, \mathcal{F}_E)$. If for some fixed time $t \in \mathbb{Z}$ and lag time $\tau \in \mathbb{N}$, we choose $Y := X_{t+\tau}$ and $X := X_t$ in the context of our previously derived formalism, then the conditional expectation operator

$$[Pf](x) = \mathbb{E}[f(X_{t+\tau}) \mid X_t = x]$$

is called the *Markov transition operator* describing the transition from X_t to $X_{t+\tau}$. If the process $(X_t)_{t\in\mathbb{Z}}$ is time-homogeneous, then P does not depend on t, but only on the time lag τ. If in addition $(X_t)_{t\in\mathbb{Z}}$ is stationary, we have $\pi = \nu$ and hence $P : L^2(\pi) \to L^2(\pi)$. We introduce the mathematical background of Markov processes and related definitions in more detail in Section 3.5.3.

We will not necessarily require the process $(X_t)_{t\in\mathbb{Z}}$ to be homogeneous and stationary in general, but restrict ourselves to a simplified scenario in which we investigate only the one-step time transition from X_t to X_{t+1} for some fixed $t \in \mathbb{Z}$. However, when we investigate estimators for P based on a single empirical realization of $(X_t)_{t\in\mathbb{Z}}$, homogeneity and stationarity of the process will be a requirement. We also note that in this case, all transitions with a larger lag time can be treated in terms of powers of P and iterates of the corresponding Markov transition kernel, which can be seen in the very basic construction of discrete-time homogeneous Markov processes (Meyn and Tweedie, 2009). It is worth

mentioning that the adjoint of the Markov transition operator P^* can be interpreted as a propagator of signed measures under the dynamics (see for example Rudolf 2012, Section 3). In the theory of Markov processes, the Markov transition operator P, its adjoint and the underlying Markov transition kernel p are often used synonymously. In the context of dynamical systems however, P is often called the *Koopman operator*, while P^* is often called the *Perron–Frobenius operator*. Both operators can also be defined on different L^p spaces (Baxter and Rosenthal, 1995).

The Markov transition operator is a fundamentally important tool for the analysis of various properties of Markov processes and dynamical systems. For instance, it is known that the spectrum of P 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*, Bovier and Den Hollander 2016), several contraction and concentration results and many more (see for example Davies, 1982a,b, 1983; Roberts et al., 1997; Roberts and Tweedie, 2001; Kontoyiannis and Meyn, 2003, 2005, 2017; Huisinga et al., 2004; Huisinga and Schmidt, 2006; Paulin, 2015).

In addition to the interpretation as canonical components, the left and right singular functions of P are known to be connected to so-called *coherent sets* of the dynamical system $(X_t)_{t\in\mathbb{Z}}$ (Froyland, 2013). Coherent sets can be interpreted as a generalization of almost invariant sets of a time-homogeneous dynamical system to the case when the underlying dynamics are time-inhomogeneous. In this case, one considers sets which are minimally dispersive under the evolution of the dynamics instead of almost invariant domains of the state space which are spatially fixed (Froyland et al., 2010).

We will not investigate this vast field and the theoretical connections between dynamical behaviour of $(X_t)_{t\in\mathbb{Z}}$ and the functional-analytic properties of P in more detail here. Instead, we acknowledge the fact that the data-driven discovery of spectral properties of P may allow the construction of various estimators for several important features of $(X_t)_{t\in\mathbb{Z}}$ of practical interest.

3.5.2. Practical applications

We briefly give a non-exhaustive overview of practical applications of the parametric projection method from Section 3.4 in the context of Markov transition operators. The eigenfunctions of empirical projections of Markov transition operators, their adjoints or closely related objects are computed in a wide variety of scenarios. Notable examples are the discretization schemes of Dellnitz and Junge (1999) and Schütte (1999), the so-called *dynamic mode decomposition* (DMD) and its extensions (Schmid, 2010; Rowley et al., 2009; Williams et al., 2015a; Tu et al., 2014; Kutz et al., 2016) as well as the

corresponding adjoint analogues of versions of DMD which are also known as *time-based independent component analysis* (TICA, Molgedey and Schuster 1994; Pérez-Hernández et al. 2013) and the *variational approach of conformation dynamics* (VAC, Noé and Nüske 2013). The latter are known to be connected to a solution of the so-called *blind source separation problem* (Hyvärinen and Oja, 2000), i.e., the separation of a mixture of superimposed signals. The detailed relations between these approaches and several generic projection schemes from numerical analysis are highlighted by Klus et al. (2016) and Klus et al. (2018).

The singular functions of P are computed via a projection method in order to obtain simplified transition models of time-inhomogeneous processes (Koltai et al., 2018; Wu and Noé, 2020) and to discover coherent sets (Froyland et al., 2010; Froyland, 2013). These approaches coincide exactly with our setup for the empirical estimation of P and its singular value decomposition presented in Section 3.3 and Section 3.4.

3.5.3. Markov processes: technical background

We review some important properties of discrete-time Markov processes and refer the reader to Meyn and Tweedie (2009) and Douc et al. (2018) for more details. A stochastic process $(X_t)_{t\in\mathbb{Z}}$ taking values in $(E, \mathcal{F}_E)$ is called *stationary*, if we have

$$(X_{t_1}, \dots, X_{t_n}) \stackrel{d}{=} (X_{t_1+\eta}, \dots, X_{t_n+\eta})$$

for all $t_1, \dots, t_n \in \mathbb{Z}$ and $n, \eta \in \mathbb{N}$. It is called a *Markov process*, if for all bounded $\mathcal{F}_E - \mathcal{B}(\mathbb{R})$ measurable functions $f : E \to \mathbb{R}$, we have

$$\mathbb{E}[f(X_{t+1}) \,|\, \mathcal{F}_{-\infty}^t] = \mathbb{E}[f(X_{t+1}) \,|\, X_t]$$

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

Definition 3.5.1 (Homogeneity)**.** *A Markov process $(X_t)_{t\in\mathbb{Z}}$ is called* (time-)homogeneous*, if there exists a Markov transition kernel $p : E \times \mathcal{F}_E \to \mathbb{R}$ such that we have*

$$\mathbb{P}[X_{t+1} \in \mathcal{A} \,|\, \mathcal{F}_{-\infty}^t] = p(X_t, \mathcal{A}) \quad \mathbb{P}\text{-a.e.}$$

for all $\mathcal{A} \in \mathcal{F}_E$ and $t \in \mathbb{Z}$.

Definition 3.5.2 (Invariant measure & reversibility)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be homogeneous with Markov transition kernel p.*

(i) A probability measure μ on $(E, \mathcal{F}_E)$ is called an invariant measure *of $(X_t)_{t\in\mathbb{Z}}$, if*

$$\int_E p(x, \mathcal{A}) \, \mathrm{d}\mu(x) = \mu(\mathcal{A})$$

for all $\mathcal{A} \in \mathcal{F}_E$.

(ii) The process $(X_t)_{t\in\mathbb{Z}}$ *is said to be* reversible *with respect to a probability measure* μ *on* $(E, \mathcal{F}_E)$*, if we have*

$$\int_{\mathcal{A}_1} p(x, \mathcal{A}_2)\,\mathrm{d}\mu(x) = \int_{\mathcal{A}_2} p(x, \mathcal{A}_1)\,\mathrm{d}\mu(x)$$

for all $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{F}_E$*.*

If $(X_t)_{t\in\mathbb{Z}}$ is reversible with respect to μ, then μ is an invariant measure of $(X_t)_{t\in\mathbb{Z}}$. Whenever we fix a particular starting distribution of a homogeneous Markov process (say for simplicity that we restrict the time index to the nonnegative integers $\mathbb{N}_0$), stationarity is trivially equivalent to its starting distribution being an invariant measure of the process.

Theorem 3.5.3 (Douc et al., 2018, Theorem 1.4.2)**.** *A homogeneous Markov process* $(X_t)_{t\in\mathbb{N}_0}$ *with starting distribution* $X_0 \sim \pi$ *is stationary if and only if* π *is an invariant measure of* $(X_t)_{t\in\mathbb{N}_0}$*.*

For a homogeneous process $(X_t)_{t\in\mathbb{Z}}$ with Markov transition kernel p, it is clear that the definition of the Markov transition operator at time $t \in \mathbb{Z}$ given by

$$[Pf](x) = \mathbb{E}[f(X_{t+1}) \mid X_t = x] = \int_E f(y)\, p(x, \mathrm{d}y)$$

is independent of $t \in \mathbb{Z}$. However, we note that its domain $L^2(\mathcal{L}(X_{t+1}))$ and target space $L^2(\mathcal{L}(X_t))$ may vary for different choices of $t \in \mathbb{Z}$ in the nonstationary case. Clearly, if $(X_t)_{t\in\mathbb{Z}}$ is homogeneous and stationary with $X_0 \sim \pi$, then we have $P : L^2(\pi) \to L^2(\pi)$ for all choices of $t \in \mathbb{Z}$ in the above definition of the Markov transition operator. If additionally the process $(X_t)_{t\in\mathbb{Z}}$ is reversible with respect to π, then P is self-adjoint (see for example Douc et al. 2018, Lemma 22.1.10).

In what follows, whenever we refer to a Markov transition operator in either the inhomogeneous or homogeneous but nonstationary case, we implicitly consider the transition from a fixed time $t \in \mathbb{Z}$ to time $t + 1$ with $X_t \sim \pi$ and $X_{t+1} \sim \nu$. This ensures that P as well as its domain and target space are well-defined. We may require P to be compact or a Hilbert–Schmidt operator, which we already discussed in Remark 3.3.3. Wo do not require additional properties such as self-adjointness.

3.5.4. Transition reconstruction error and canonical components

We now investigate an alternative interpretation of the canonical components which illustrates the well-known connection between the canonical components and the low-rank approximation of linear operators. This perspective is particularly relevant in the context of Markov processes (Wu and Noé, 2020) and given in terms of a minimization of a

Hilbert–Schmidt distance instead of the classical maximization problem (3.2). Let P be the Markov transition operator of a not necessarily homogeneous Markov process $(X_t)_{t\in\mathbb{Z}}$.

Given some complete orthonormal system $\{e_i\}_{i\in I} \subset L^2(\nu)$ we define the *transition reconstruction error* associated with two empirical rank r orthogonal projectors $\Pi_{\widehat{U}} : L^2(\pi) \to L^2(\pi)$ and $\Pi_{\widehat{V}} : L^2(\nu) \to L^2(\nu)$ by

$$\mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) := \sum_{i\in I} \mathbb{E}\left[\left|\mathbb{E}[\, e_i(X_{t+1}) \mid X_t] - \Pi_{\widehat{U}}\left\{\mathbb{E}[\, \Pi_{\widehat{V}}\, e_i(X_{t+1}) \mid X_t]\right\}\right|^2\right].$$

The reconstruction error is essentially the sum of squared errors between the expected transition of the observables $\{e_i\}_{i\in I} \subset L^2(\nu)$ and their projected analogues on the subspaces span $\widehat{V}$ and span $\widehat{U}$, respectively.

We now show that $\mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}})$ is well-defined and independent of the choice of the complete orthonormal system $\{e_i\}_{i\in I}$ if P is a Hilbert–Schmidt operator. Moreover, the minimum of $\mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}})$ is attained when $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ are exactly the orthogonal projectors onto the first r canonical components of P. For convenience, we use the shorthand notation $S_2 := S_2(L^2(\nu), L^2(\pi))$ in what follows.

Lemma 3.5.4 (Minimizing the reconstruction error)**.** *Let $P \in S_2$. Additionally, let $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ let the orthogonal projection operators associated with some orthonormal sets $\widehat{U} \subset L^2(\pi)$ and $\widehat{V} \subset L^2(\nu)$ of cardinality $r \leq \operatorname{rank}(P)$. Then we have*

$$\mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) = \left\|P - \Pi_{\widehat{U}} P \Pi_{\widehat{V}}\right\|_{S_2}^2. \tag{3.10}$$

Moreover, we have

$$\min_{\Pi_{\widehat{U}}, \Pi_{\widehat{V}}} \mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) = \mathcal{E}(\Pi_U, \Pi_V) = \sum_{i>r} \rho_i(P)^2, \tag{3.11}$$

where $U \subset L^2(\pi)$ and $V \subset L^2(\nu)$ are the sets containing the first r left and right singular functions of P.

Proof. We have

$$\begin{aligned}
\mathcal{E}_r(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) &= \sum_{i\in I} \mathbb{E}\left[\left|\mathbb{E}[e_i(X_{t+1}) \mid X_t] - \Pi_{\widehat{U}}\{\mathbb{E}[\Pi_{\widehat{V}} e_i(X_{t+1}) \mid X_t]\}\right|^2\right] \\
&= \sum_{i\in I} \int_E \left|Pe_i - \Pi_{\widehat{U}} P \Pi_{\widehat{V}} e_i\right|^2 \mathrm{d}\pi = \sum_{i\in I} \left\|(P - \Pi_{\widehat{U}} P \Pi_{\widehat{V}}) e_i\right\|_{L^2(\pi)}^2 \\
&= \left\|P - \Pi_{\widehat{U}} P \Pi_{\widehat{V}}\right\|_{S_2}^2,
\end{aligned}$$

which proves (3.10). The second assertion follows from the fact that $\Pi_{\widehat{U}} P \Pi_{\widehat{V}}$ is an operator of rank of at most r together with the Eckart–Young–Mirsky theorem (see Appendix A.2.2). ∎

In what follows, let U and V be the sets containing the first r canonical components of P. As an indicator for the performance of two empirical projection operators $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ with respect to the transition reconstruction error, we may naturally consider the *excess reconstruction error*

$$\mathcal{E}(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) - \mathcal{E}(\Pi_U, \Pi_V) \geq 0.$$

We show that he excess reconstruction error can be bounded by the Hilbert–Schmidt distance between the corresponding projection operators.

Lemma 3.5.5. *(Excess reconstruction error) Let $P \in S_2$ and $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ be the orthogonal projection operators associated with some orthonormal sets $\widehat{U} \subset L^2(\pi)$ and $\widehat{V} \subset L^2(\nu)$ of cardinality r.*

$$\mathcal{E}(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) - \mathcal{E}(\Pi_U, \Pi_V) \leq \|P\|_{S_2} \left(\left\| \Pi_{\widehat{V}} - \Pi_V \right\|_{S_2} + \left\| \Pi_{\widehat{U}} - \Pi_U \right\|_{S_2} \right)$$

Proof. By expressing the Hilbert–Schmidt norm in (3.10) in terms of the inner product and applying the binomial formula, we first note that for all orthogonal projection operators $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$, we have

$$\begin{aligned} \mathcal{E}(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) &= \|P\|_{S_2}^2 - 2 \left\langle P, \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\rangle_{S_2} + \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\|_{S_2}^2 \\ &= \|P\|_{S_2}^2 - \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\|_{S_2}^2, \end{aligned} \tag{3.12}$$

where the second equality follows from the fact that

$$\begin{aligned} \left\langle P, \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\rangle_{S_2} &= \operatorname{Tr}(P^* (\Pi_{\widehat{U}} P \Pi_{\widehat{V}})) = \operatorname{Tr}(P^* \Pi_{\widehat{U}}^* \Pi_{\widehat{U}} P \Pi_{\widehat{V}}) \\ &= \operatorname{Tr}(\Pi_{\widehat{V}}^* P^* \Pi_{\widehat{U}}^* \Pi_{\widehat{U}} P \Pi_{\widehat{V}}) = \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\|_{S_2}^2. \end{aligned} \tag{3.13}$$

This can be seen by using idempotence and self-adjointness of orthogonal projectors together with $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ for arbitrary Hilbert–Schmidt operators A and B acting on a common Hilbert space (Weidmann, 1980, Theorem 7.11). By making use of (3.10), (3.12) and (3.13), we have

$$\begin{aligned} \mathcal{E}(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) - \mathcal{E}(\Pi_U, \Pi_V) &= - \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\|_{S_2}^2 + \|\Pi_U P \Pi_V\|_{S_2}^2 \\ &= - \left\langle P, \Pi_{\widehat{U}} P \Pi_{\widehat{V}} \right\rangle_{S_2} + \langle P, \Pi_U P \Pi_V \rangle_{S_2} \\ &\leq \|P\|_{S_2} \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} - \Pi_U P \Pi_V \right\|_{S_2} \\ &\leq \|P\|_{S_2} \left\| \Pi_{\widehat{U}} P \Pi_{\widehat{V}} - \Pi_{\widehat{U}} P \Pi_V + \Pi_{\widehat{U}} P \Pi_V - \Pi_U P \Pi_V \right\|_{S_2} \\ &= \|P\|_{S_2} \left\| \Pi_{\widehat{U}} P (\Pi_{\widehat{V}} - \Pi_V) + (\Pi_{\widehat{U}} - \Pi_U) P \Pi_V \right\|_{S_2} \\ &\leq \|P\|_{S_2} \left(\left\| P (\Pi_{\widehat{V}} - \Pi_V) \right\|_{S_2} + \left\| (\Pi_{\widehat{U}} - \Pi_U) P \right\|_{S_2} \right), \end{aligned}$$

where we use $\left\| \Pi_{\widehat{U}} \right\| \leq 1$ and $\|\Pi_V\| \leq 1$ in the last step. The fact $\|P\| \leq 1$ yields the assertion. ∎

When $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ are the sets of first r canonical components of a compact empirical operator $\widehat{P}$ with the overall approximation error $\Delta = \widehat{P} - P$, then our previous considerations from Section 3.3.2 lead to the following bound for the transition reconstruction error.

Theorem 3.5.6 (Transition reconstruction error bound)**.** *Assume that $P \in S_2$ and $\widehat{P}$ is a compact estimate of P with estimation error $\Delta = \widehat{P} - P$. For $r < \min\{\operatorname{rank}(P), \operatorname{rank}(\widehat{P})\}$, let $\Pi_{\widehat{U}}$ and $\Pi_{\widehat{V}}$ be the sets of first r canonical components of $\widehat{P}$. Furthermore, assume $\rho_r(P) \neq \rho_{r+1}(P)$. Then we have*

$$\mathcal{E}(\Pi_{\widehat{U}}, \Pi_{\widehat{V}}) - \mathcal{E}(\Pi_U, \Pi_V) \leq 4\sqrt{2r}\, \|P\|_{S_2} \frac{(2\,\|\Delta\| + \|\Delta\|^2)}{\sigma_r(P)^2 - \sigma_{r+1}^2(P)}.$$

Proof. Theorem 3.5.6 is the direct consequence of Theorem 3.3.5 applied to the assertion of Lemma 3.5.5. ∎

3.6. Estimation from dependent data

In order to establish consistency for the estimators of P with dependent observations (such as a sequence of subsequent data obtained from a Markov process), we need appropriate mathematical tools to describe this setting. Two standard assumptions for estimators based on dependent data are ergodicity and mixing. Since we will later work with these concepts in a more general setting, we introduce them here for general stationary stochastic processes which do not necessarily need to exhibit the Markov property.

3.6.1. Ergodicity

We consider a stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(E, \mathcal{F}_E)$. Stationary stochastic processes can be conveniently expressed in terms of dynamical systems on measure spaces. We briefly introduce the concept of measure-preserving dynamical systems and ergodicity. For details, the reader may refer for example to (Petersen, 1983, Section 1.2) and (Kallenberg, 2002, Section 9).

We may assume without loss of generality that the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ describing the stationary process $(X_t)_{t\in\mathbb{Z}}$ 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 as the family of *coordinate projections* on Ω: for $\omega = (\omega_t)_{t\in\mathbb{Z}} \in \Omega$, we can write $X_t(\omega) = \omega_t = X_0(T^t\omega)$ 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* (Petersen, 1983), i.e., for all events $M \in \mathcal{F}_E^{\otimes\mathbb{Z}}$, we have that $T^{-1}M = M$ implies either $\mathbb{P}[M] = 0$ or $\mathbb{P}[M] = 1$.

Ergodic processes are of interest to us since they obey a *strong law of large numbers*, which we obtain as the following generalization of Birkhoff's ergodic theorem.

Theorem 3.6.1 (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 $f \in L^1(\Omega, \mathcal{F}, \mathbb{P}; B)$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(T^i\omega) = \int_\Omega f(\omega)\, \mathrm{d}\mathbb{P}(\omega),$$

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

3.6.2. Strong mixing coefficients

In order to derive more detailed results for dependent random variables, we need a measure of the degree of their dependence. Several different concepts have been developed in order to deal with this problem. Here we will focus on the *strong mixing coefficients* which are often considered in statistics (Doukhan, 1994; Bradley, 2005).

Definition 3.6.2 (Mixing coefficients)**.** *For σ-fields $\mathcal{F}_1$, $\mathcal{F}_2 \subseteq \mathcal{F}$, we define*

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

$$\beta(\mathcal{F}_1, \mathcal{F}_2) := \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left|\mathbb{P}[A_i \cap B_j] - \mathbb{P}[A_i]\mathbb{P}[B_j]\right|,$$

where the supremum in the last equation ranges over all finite subsets $\{A_1, \ldots, A_I\} \subseteq \mathcal{F}_2$ and $\{B_1, \ldots, B_J\} \subseteq \mathcal{F}_2$ which form a partition of Ω. For a stochastic process $(X_t)_{t\in\mathbb{Z}}$, we furthermore define

$$\alpha(n) := \alpha((X_t)_{t\in\mathbb{Z}}, n) := \sup_{j\in\mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{\infty}) \quad n \in \mathbb{Z},$$

$$\beta(n) := \beta((X_t)_{t\in\mathbb{Z}}, n) := \sup_{j\in\mathbb{Z}} \beta(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{\infty}) \quad n \in \mathbb{Z},$$

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

Whenever $(X_t)_{t\in\mathbb{Z}}$ is stationary, we have a simplified expression for the mixing coefficients in terms of $\alpha(n) = \alpha(\mathcal{F}_{-\infty}^{0}, \mathcal{F}_{n}^{\infty})$ and analogously $\beta(n) = \beta(\mathcal{F}_{-\infty}^{0}, \mathcal{F}_{n}^{\infty})$. The process is

called α-*mixing* or just *strongly mixing*, when $\alpha(n) \to 0$ as $n \to \infty$ and β-*mixing* or *absolutely regular*, when $\beta(n) \to 0$ as $n \to \infty$. In this case, the convergence rates of $\alpha(n)$ and $\beta(n)$ are called the *mixing rate* of the associated process. We have $2\alpha(n) \leq \beta(n)$ for all $n \in \mathbb{N}$ and hence β-mixing is a stronger condition than α-mixing. In this paper, we will not focus on the various alternative strong mixing coefficients which are frequently used in statistics (Bradley, 2005), since α-mixing and β-mixing are amongst the most commonly used mixing concepts and cover a wide range of processes in practice.

Remark 3.6.3 (Terminology). The concept of *strong mixing coefficients* is typically much stronger than the *strong mixing* considered in ergodic theory (Petersen, 1983). The strong mixing coefficients are 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 3.6.4 (Mixing processes). A wide range of mixing processes can be found in Doukhan (1994) and Bradley (2005). We list some important examples here.

(1) Irreducible and aperiodic stationary Markov processes on $E \subseteq \mathbb{R}$ are β-mixing (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 β-mixing with $\beta(n) = O(\exp(-cn))$ for some $c > 0$, see 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

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

where $h_t : E \to E$ are measurable functions and ξ_t are iid 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 7.6). Therefore (1) and (2) apply in this case. Conditions under which this system is geometrically ergodic (i.e., geometrically β-mixing in the sense of (2)) are given by Doukhan (1994, Section 2.4).

(4) One can show that a time-discretized version of a diffusion process expressed as a stochastic differential equation results in a geometrically ergodic Markov process under certain assumptions (Lacour, 2008).

(5) Under some requirements, commonly used linear and nonlinear time series 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 lemma which ensures that measurable transformations of a finite number of components of a mixing process preserve mixing rates.

Lemma 3.6.5 (Transformed processes are mixing)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ be a stationary process with values in the standard Borel space $(E, \mathcal{F}_E)$. Let $(F, \mathcal{F}_F)$ be another standard Borel space. 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*

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

we have

$$\alpha((H_t)_{t\in\mathbb{Z}}, n) \leq \alpha((X_t)_{t\in\mathbb{Z}}, n - \eta) \tag{3.14}$$

for all $n \in \mathbb{Z}$. The same result applies when the α-mixing coefficient is replaced with the β-mixing coefficient in the expression above.

In particular, if $(X_t)_{t\in\mathbb{Z}}$ is α/β-mixing, then $(H_t)_{t\in\mathbb{Z}}$ is α/β-mixing with at least the same mixing rate as $(X_t)_{t\in\mathbb{Z}}$.

Proof. 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 $\mathcal{H}_{-\infty}^j \subseteq \mathcal{F}_{-\infty}^{j+\eta}$ as well as $\mathcal{H}_j^\infty \subseteq \mathcal{F}_j^\infty$ for all $j \in \mathbb{Z}$. Therefore, we have

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

and analogously when the α-mixing coefficient is replaced with the β-mixing coefficient. ■

3.6.3. Convergence results for dependent observations

We now illustrate how ergodicity and the theory of strongly mixing random variables can be used to assess the convergence when the parametric projection method described in Section 3.4 is applied to a Markov transition operator.

Let $(X_t)_{t\in\mathbb{Z}}$ be a homogeneous stationary Markov process with marginal $X_t \sim \pi$ and Markov transition operator $P : L^2(\pi) \to L^2(\pi)$. Note again that we can now restrict ourselves to the single orthonormal system $\Phi \subseteq L^2(\pi)$, since P operates only on $L^2(\pi)$ due to the stationarity of $(X_t)_{t\in\mathbb{Z}}$.

We assume we have access to a finite sequence of observations $X_1, \dots, X_{n+1}$ from a realization of $(X_t)_{t\in\mathbb{Z}}$. The projection estimator $\widehat{P}_{\Phi,\Phi}$ for the operator $P_{\Phi,\Phi}$ from Section 3.4.2 given in terms of the matrix $\mathbf{A} \in \mathbb{R}^{s\times s}$ defined by

$$\mathbf{A}_{ij} := \mathbb{E}[\phi_i(X_t)\, \phi_j(X_{t+1})] = \langle \phi_i,\, P\phi_j \rangle_{L^2(\pi)}\,.$$

now takes the form

$$\widehat{\mathbf{A}}_{ij} := \frac{1}{n}\sum_{t=1}^{n} \phi_i(X_t)\,\phi_j(X_{t+1}) = \left\langle \phi_i,\, \widehat{P}_{\Phi,\Phi}\phi_j \right\rangle_{L^2(\pi)}$$

and consists of evaluations of the basis functions at the one-step transition pairs (X_t, X_{t+1}). We now outline basic convergence results for this estimator.

Theorem 3.6.6 (Strong law of large numbers)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be ergodic. Then*

$$\widehat{P}_{\Phi,\Phi} \to P_{\Phi,\Phi} \quad \mathbb{P}\text{-}a.e.$$

as $n \to \infty$ w.r.t. the norm of $S_2(L^2(\pi))$ and therefore similarly for every other norm on the space of linear operators on the finite-dimensional space $\operatorname{span}\Phi$*.*

Proof. The stationary time-lagged product process $(X_t, X_{t+1})_{t\in\mathbb{Z}}$ on $E \times E$ can be expressed via the projection tuple $(X_t, X_{t+1})(\omega) = (X_0, X_1)(T^t\omega)$, where T is the shift operator described in Section 3.6.1. Similarly to the proof of Lemma 3.4.1, we define the measurable transformation $\xi : E \times E \to \mathbb{R}^{s\times s}$ given by

$$\xi(X_t, X_{t+1}) := (\phi_i(X_t)\phi_j(X_{t+1}))_{1\le i,j\le s}, \tag{3.15}$$

where $\mathbb{R}^{s\times s}$ is equipped with some arbitrary matrix norm—we choose the Frobenius norm here. We have $\mathbf{A} = \mathbb{E}[\xi(X_0, X_1)]$ as well as $\widehat{\mathbf{A}} = \frac{1}{n}\sum_{t=1}^{n}\xi(X_t, X_{t+1})$. It is clear that $\xi(X_0, X_1) \in L^1(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^{s\times s})$ and we can therefore apply Theorem 3.6.1 with the observable $\xi(X_0, X_1)$ to obtain

$$\widehat{\mathbf{A}} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\xi(X_0, X_1)\circ T^t = \mathbb{E}[\xi(X_0, X_1)] = \mathbf{A} \quad \mathbb{P}\text{-a.e.}$$

with respect to the Frobenius norm and therefore any arbitrary matrix norm. ∎

An upper bound for the speed of convergence in Hilbert–Schmidt norm can be deduced by straight forwardly applying an asymptotic result for α-mixing Hilbertian random variables due to Bosq (2000, Corollary 2.4).

Theorem 3.6.7 (Convergence speed)**.** *Let $|\phi_i(X_t)| \le M$ $\mathbb{P}$-a.e. for all $1 \le i \le s$. Let additionally $(X_t)_{t\in\mathbb{Z}}$ be α-mixing with coefficients satisfying $\alpha(n) \le ar^n$ for some $r \in (0,1)$ and $a > 0$ for all $n \in \mathbb{N}$, then*

$$\left\|\widehat{P}_{\Phi,\Phi} - P_{\Phi,\Phi}\right\|_{S_2(L^2(\pi))} = O\left(\frac{(\log n)^{3/2}}{n^{1/2}}\right) \quad \mathbb{P}\text{-}a.e.$$

Proof. We again consider the transformation ξ defined in (3.15). Note that we have the bound $\|\xi(X_0, X_1) - \mathbb{E}[\xi(X_0, X_1)]\|_F \leq 2sM^2$ $\mathbb{P}$-a.e. as derived in (3.8). The α-mixing rate of the stationary process $(\xi(X_t, X_{t+1}) - \mathbb{E}[\xi(X_0, X_1)])_{t\in\mathbb{Z}}$ is at least as fast as the mixing rate of $(X_t)_{t\in\mathbb{Z}}$ due to Lemma 3.6.5. Since the space $\mathbb{R}^{s\times s}$ is finite-dimensional, the covariance operator eigenvalue decay condition of from Bosq (2000, Corollary 2.4) is satisfied (the covariance operator of a finite-dimensional random variable can again be expressed in terms of a matrix and hence exhibits only a finite number of eigenvalues). The assertion follows directly by applying Bosq (2000, Corollary 2.4) to the stationary centered process $(\xi(X_t, X_{t+1}) - \mathbb{E}[\xi(X_0, X_1)])_{t\in\mathbb{Z}}$. ■

As our last result, we present the application of a Bernstein-type error bound for β-mixing sequences due to Rhomari (2002, 2011) similarly as in the preceding proof of Theorem 3.6.7. We present the original bound in Appendix A.5.3 in a simplified form.

Theorem 3.6.8 (Error bound)**.** *Let $|\phi_i(X_t)| \leq M$ $\mathbb{P}$-a.e. for all $1 \leq i \leq s$. Let additionally $(X_t)_{t\in\mathbb{Z}}$ be β-mixing with coefficients $\beta(n)$. For every $\epsilon > 0$ and $1 \leq l \leq \lfloor n/2 \rfloor$, we have*

$$\mathbb{P}\left[\left\|\widehat{P}_{\Phi,\Phi} - P_{\Phi,\Phi}\right\|_{S_2(L^2(\pi))} \geq \epsilon\right] \leq 4\exp\left(-\frac{n\epsilon^2}{4(1+2l/n)\sigma^2 + 8lsM^2\epsilon/3}\right) + \left(\frac{n}{l} + 2\right)\beta(l-1),$$

where

$$\sigma^2 \leq 4s^2M^4\left(1 + 5\sum_{i=1}^{l-1}\beta(i-1)\right).$$

Proof. We apply the bound given in Appendix A.5.3 to the stationary and $\mathbb{P}$-a.e. bounded process $(\xi(X_t, X_{t+1}) - \mathbb{E}[\xi(X_0, X_1)])_{t\in\mathbb{Z}}$ defined in (3.15). ■

The general assertion of the bound in Theorem 3.6.8 is very similar to the well-known Bernstein bound for independent Hilbertian random variables (see Appendix A.5.2 for details). In essence, the decay of the β-mixing coefficients quantifies an upper bound for the variance proxy $\sigma^2 \geq 0$. Moreover, a balance between the total number of samples n and the number of consecutive steps l considered in the variance term is required to obtain an overall optimal bound. This phenomenon can also be seen in other concentration inequalities of this type.

Remark 3.6.9. (Alternative concentration inequality) Instead of applying the generalized Bernstein bound by Rhomari (2002) as shown here, we can alternatively apply a slightly more complex bound of similar form for α-mixing sequences as derived by Bosq (2000, Theorem 2.12). We introduce this alternative bound later in this thesis in Theorem 5.5.1.

3.7. Summary and outlook

In this chapter, we establish a basic statistical viewpoint for the approximation of conditional expectation operators. Addressing the generic estimation of their spectral properties, we apply a perturbation result which proves stability of the dominant singular subspaces with respect to the estimation error. By introducing a classical projection method, we highlight the limits of typical error bounds for the empirical basis normalization and estimation procedure in high dimensions. We investigate the projection approach and its applications in the context of Markov transition operators and connect the empirical minimization of a reconstruction error to the perturbation of the singular functions. Finally, we derive convergence results for the projected estimate based on dependent data under ergodicity and mixing assumptions.

This chapter leaves us with several open questions and possible directions for future research, which we will address separately in what follows.

1. The error bounds for the projected estimate of P which we derive are far from optimal and rely only on the boundedness of the underlying ansatz functions. A potential way to improve these theoretical results is to incorporate low-rank assumptions for P and apply matrix concentration results based on the effective rank of the projection of P (see Remark 3.4.6). In a related setting, a low-rank model for finite-state Markov chains has been formulated by Gerber and Horenko (2017) by assuming an underlying latent process.

2. In order to derive more efficient approximation schemes for the projection of P, we need to circumvent the problem of the estimation of large matrices. It occurs reasonable to apply specifically designed techniques for models with large parameter sizes and work with ideas from compressive sensing, high-dimensional matrix estimation and ℓ_1-regularization. An overview of recent approaches in these fields can be found in Foucart and Rauhut (2013) and Wainwright (2019) and the literature reviews therein.

3. Our framework to derive theoretical convergence results for the projected estimate of P with dependent data is based on the situation that we observe subsequent time steps in a single realization of a stationary homogeneous Markov process. However, the assumption of stationarity may not be satisfied in general. A potential direction of further research could more sophisticated sampling techniques under the assumptions that a homogeneous Markov process can be simulated with arbitrary initial distributions. Also, by modifying concentration bounds based on the spectral gap of a Markov transition operator (Paulin, 2015; Fan et al., 2021; Jiang et al., 2018) it may be possible to prove convergence results for situations which are closer to practical applications (i.e., the detection of metastable states and coherent sets).

4. Nonparametric approximation of conditional expectation operators

This chapter contains passages taken from Mollenhauer and Koltai (2020).

As we have seen in the example of a projection method in the preceding chapter, classically used parametric models for the estimation of P are theoretically limited in their performance by both the difficulties related to the corresponding model error and the estimation of high-dimensional matrices. A natural reason for these limitations is the choice of ansatz spaces, which has to be conducted a priori without knowledge of the joint distribution of X and Y and the resulting structure of P.

To circumvent these issues, various nonparametric methods have been derived. In essence, they rely on ansatz functions which are adaptively chosen based on the data during the sampling procedure. However, from a theoretical perspective, it is actually not clear that any of the aforementioned problems of parametric methods are alleviated by adapting the basis functions to the data. In this chapter, we will thus focus on the two main types of theoretical questions arising from such an idea:

1. *Is it possible to precisely specify the corresponding hypothesis space when the basis functions are adapted to the data? If so, what happens if the model is misspecified, i.e., the true operator P is not contained in the hypothesis space? Put differently, which object is approximated in the infinite data limit in this case?*

2. *Can we expect stronger overall approximation properties compared to parametric methods? Which quantities affect the convergence speed of such a nonparametric approach and is it possible to derive reasonable error bounds?*

Prominent examples of such nonparametric approaches based on reproducing kernels are kernel-based *extended dynamic mode decomposition* (EDMD) and various related techniques (Williams et al., 2015b; Schwantes and Pande, 2015; Kawahara, 2016; Klus et al., 2020, 2019; Mollenhauer et al., 2020b; Tian and Wu, 2020). Although these methods work well in practice, their approximation-theoretic background, convergence properties and statistical behaviour are not understood yet. By making use of the

functional-analytic theory of *vector-valued reproducing kernel Hilbert spaces*, we will be able to give some answers to the above questions for these methods.

The main idea behind the kernel-based approximation of P is to only evaluate P over functions contained in a reproducing kernel Hilbert space $\mathscr{H}$ which is embedded into its domain $L^2(\nu)$. That is, instead of $P : L^2(\nu) \to L^2(\pi)$, one essentially approximates the modified conditional expectation operator

$$\mathscr{H} \ni f \mapsto \mathbb{E}[f(Y) \,|\, X = \cdot] \in L^2(\pi),$$

where the geometric properties on the domain are given by the inner product of $\mathscr{H}$ instead of $L^2(\nu)$. We begin our theoretical investigation at this point and work our way to a nonparametric model step by step. Our investigation shows that the theory of the nonparametric approximation of P is naturally related to several well-known concepts in kernel-based inference such as the *kernel conditional mean embedding*, the *maximum mean discrepancy* and in particular regularized least squares regression with vector-valued kernels.

4.1. Overview

We introduce the functional-analytic theory of vector-valued reproducing kernel Hilbert spaces in Section 4.2 and discuss the assumptions which we impose in this chapter. In Section 4.3, we briefly review the theory of integral and inclusion operators associated with reproducing kernels. Section 4.4 highlights the embedding of probability measures into reproducing kernel Hilbert spaces, which is strongly connected to the theory which we derive in the following sections.

We begin the exposition of the theoretical core results of this chapter with the derivation of a nonparametric model and its approximation properties from a population perspective in Section 4.5. In Section 4.6, we establish a typical connection to regularization of inverse problems and use this theory to construct an empirical estimate based on our previous derivation. We investigate the special case of Tikhonov–Phillips regularization and obtain closed form solutions for the population version as well as the empirical case. Section 4.7 highlights practical applications and connects our model to existing approaches while Section 4.8 discusses related work.

4.2. Vector-valued reproducing kernel Hilbert spaces

We will give an overview of the concept of a *vector-valued reproducing kernel Hilbert space* (vRKHS), i.e., a Hilbert space consisting of functions from a nonempty set E to a

Hilbert space H. Since the construction of such a space is quite technical, we will not cover all mathematical details here but rather introduce the most important properties. For a rigorous treatment of this topic, we refer the reader to Carmeli et al. (2006) as well as Carmeli et al. (2010). In this chapter, $(E, \mathcal{F}_E)$ is assumed to be a second countable locally compact Hausdorff space E equipped with its Borel field.

Definition 4.2.1 (Operator-valued psd kernel)**.** *Let E be a nonempty set and H be a real Hilbert space. A function $K : E \times E \to \mathfrak{B}(H)$ is called an* operator-valued positive-semidefinite (psd) kernel*, if $K(x, x') = K(x', x)^*$ for all $x, x' \in E$ and additionally for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in E$ as well as $\beta_1, \ldots, \beta_n \in \mathbb{R}$ and $h \in H$, we have*

$$\sum_{i,j=1}^{n} \beta_i \beta_j \langle h,\, K(x_i, x_j)h \rangle_H \geq 0. \tag{4.1}$$

Let $K : E \times E \to \mathfrak{B}(H)$ be an operator-valued psd kernel. For a fixed $x \in E$ and $h \in H$, we obtain a function from E to H via

$$[K_x h](\cdot) := K(\cdot, x)h.$$

We can now consider the set

$$\mathscr{G}_{\text{pre}} := \text{span}\{K_x h \mid x \in E, h \in H\} \tag{4.2}$$

and define an inner product on $\mathscr{G}_{\text{pre}}$ by linearly extending the expression

$$\left\langle K_x h,\, K_{x'} h' \right\rangle_{\mathscr{G}} := \left\langle h,\, K(x, x')h' \right\rangle_H . \tag{4.3}$$

Let $\mathscr{G}$ be the completion of $\mathscr{G}_{\text{pre}}$ with respect to this inner product. We call $\mathscr{G}$ the *H-valued reproducing kernel Hilbert space* or more generally the vRKHS induced by the kernel K.

The space $\mathscr{G}$ is a Hilbert space consisting of functions from E to H with the *reproducing property*

$$\langle F(x),\, h \rangle_H = \langle F,\, K_x h \rangle_{\mathscr{G}} \tag{4.4}$$

for all $F \in \mathscr{G}$, $h \in H$ and $x \in E$. Additionally, we have

$$\|F(x)\|_H \leq \|K(x, x)\|^{1/2} \, \|F\|_{\mathscr{G}}, \quad x \in E \tag{4.5}$$

for all $F \in \mathscr{G}$. When K_x is understood as a linear operator from H to $\mathscr{G}$ fixed $x \in E$, the inner product given by (4.3) implies that K_x is a bounded operator for all $x \in E$. As a result, we can rewrite the reproducing property (4.4) as

$$F(x) = K_x^* F \tag{4.6}$$

for all $F \in \mathscr{G}$ and $x \in E$. Therefore we have

$$K_x^* K_{x'} = K(x, x'), \quad x, x' \in E \tag{4.7}$$

and the linear operators $K_x \colon \mathscr{H} \to \mathscr{G}$ and $K_x^* \colon \mathscr{G} \to \mathscr{H}$ are bounded with

$$\|K_x\| = \|K_x^*\| = \|K(x,x)\|^{1/2}. \tag{4.8}$$

In this work, we will deal with two very specific examples of psd kernels, which we will introduce in what follows.

4.2.1. Real-valued RKHS

When we identify the space of linear operators on $\mathbb{R}$ with $\mathbb{R}$ itself and consider a real-valued psd kernel

$$k \colon E \times E \to \mathbb{R} \tag{4.9}$$

in the sense of Definition 4.2.1, we obtain the well-known setting of the real-valued reproducing kernel Hilbert space (RKHS; Aronszajn 1950). The kernel k satisfies $k(x, x') = k(x', x)$ for all $x, x' \in E$. The space $\mathscr{H}$ consists of functions from E to $\mathbb{R}$ with the properties

(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.

It follows in particular that $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathscr{H}}$. The so-called *canonical feature map* $\varphi \colon E \to \mathscr{H}$ is given by $\varphi(x) := k(x, \cdot)$.

The space $\mathscr{H}$ has been thoroughly examined over the last decades and has numerous applications in statistics, approximation theory and machine learning. For details, the reader may consult Berlinet and Thomas-Agnan (2004), Steinwart and Christmann (2008) and Saitoh and Sawano (2016).

Remark 4.2.2 (Notation). In what follows, $\mathscr{H}$ will always denote the $\mathbb{R}$-valued RKHS induced by the kernel $k \colon E \times E \to \mathbb{R}$ with corresponding canonical feature map $\varphi \colon E \to \mathscr{H}$ as described in this section. We will write small letters $f, g, h \in \mathscr{H}$ for $\mathbb{R}$-valued RKHS functions.

4.2.2. $\mathscr{H}$-valued vRKHS

Let $\mathscr{H}$ be the real-valued RKHS induced by the kernel $k : E \times E \to \mathbb{R}$ as decribed in Section 4.2.1. Let $\mathrm{Id}_{\mathscr{H}}$ be the identity operator on $\mathscr{H}$. We define the map

$$K \colon E \times E \to \mathfrak{B}(\mathscr{H}),$$
$$K(x, x') := k(x, x')\mathrm{Id}_{\mathscr{H}}$$

for all $x, x' \in E$. It is straightforward to show that K is a psd kernel and therefore induces an $\mathscr{H}$-valued vRKHS $\mathscr{G}$ (see also Carmeli et al., 2010, Example 3.3.(i)).

Remark 4.2.3 (Notation). In what follows, $\mathscr{G}$ will always denote the $\mathscr{H}$-valued vRKHS induced by the kernel $K \colon E \times E \to \mathfrak{B}(\mathscr{H})$ given by $K(x, x') = k(x, x')\mathrm{Id}_{\mathscr{H}}$ as described in this section. We will write capital letters $F, G, H \in \mathscr{G}$ for $\mathscr{H}$-valued functions in order to distinguish them from real-valued functions $f, g, h \in \mathscr{H}$.

4.2.3. Isomorphism between $\mathscr{G}$ and $S_2(\mathscr{H})$

The foundation of our approach is given by the fact that elements of the vRKHS $\mathscr{G}$ defined by the kernel $K(x, x') = k(x, x')\mathrm{Id}_{\mathscr{H}}$ can be interpreted as Hilbert–Schmidt operators on $\mathscr{H}$. We again recall that the space of Hilbert–Schmidt operators $S_2(\mathscr{H})$ is isometrically isomorphic to the tensor product space $\mathscr{H} \otimes \mathscr{H}$ via an identification of rank-one operators as elementary tensors. We will use the latter to state the result, since a formulation in this way is more natural.

Theorem 4.2.4 ($\mathscr{G}$ is isomorphic to $\mathscr{H} \otimes \mathscr{H}$)**.** *Let $\mathscr{H}$ be the real-valued RKHS with corresponding kernel k. Let $\mathscr{G}$ be the vector-valued RKHS induced by the kernel $K(x, x') := k(x, x')\mathrm{Id}_{\mathscr{H}}$. The map Θ defined on rank-one tensors in $\mathscr{H} \otimes \mathscr{H}$ defining an $\mathscr{H}$-valued function on E by the relation*

$$[\Theta(f \otimes h)](x) := h(x)f = (f \otimes h)\varphi(x) = \langle h, \varphi(x)\rangle_{\mathscr{H}} f \tag{4.10}$$

for all $x \in E$ and $f, h \in \mathscr{H}$ maps to $\mathscr{G}$. Furthermore, extending Θ to $\mathscr{H} \otimes \mathscr{H}$ via linearity and completion yields an isometric isomorphism between $\mathscr{H} \otimes \mathscr{H}$ and $\mathscr{G}$.

A proof of Theorem 4.2.4 can be found in Carmeli et al. (2010, Proposition 3.5 & Example 3.3(i)). The isometric isomorphism

$$\Theta : \mathscr{H} \otimes \mathscr{H} \to \mathscr{G}$$

defined by (4.10) seems technical but actually becomes quite intuitive when one examines how the inner products of both spaces are connected via the kernels k and K. We outline this connection briefly below.

Let $x, x' \in E$ and $h, h' \in \mathscr{H}$. We define $F := K_x h \in \mathscr{G}$ and $F' := K_{x'} h' \in \mathscr{G}$ and note that we can express the inner product in $\mathscr{G}$ as

$$\begin{aligned}\left\langle F,\, F' \right\rangle_{\mathscr{G}} &= \left\langle K_{x'}^* K_x h,\, h' \right\rangle_{\mathscr{H}} = \left\langle k(x', x)\mathrm{Id}_{\mathscr{H}} h,\, h \right\rangle_{\mathscr{H}} \\ &= \left\langle \varphi(x'),\, \varphi(x) \right\rangle_{\mathscr{H}} \left\langle h,\, h' \right\rangle_{\mathscr{H}} \\ &= \left\langle h \otimes \varphi(x),\, h' \otimes \varphi(x') \right\rangle_{\mathscr{H} \otimes \mathscr{H}}.\end{aligned}$$

This derivation can be extended straightforwardly to a correspondence of vector-valued functions $F, F' \in \operatorname{span}\{K_x h \,|\, x \in E, h \in \mathscr{H}\} \subseteq \mathscr{G}$ and linear combinations of tensors in $\{h \otimes \varphi(x) \,|\, x \in E, h \in \mathscr{H}\} \subseteq \mathscr{H} \otimes \mathscr{H}$ by using bilinearity of the respective inner products. Since both spans are dense in the associated spaces, this property can be extended to the full spaces via completion. We now restate Theorem 4.2.4 in a more accessible way for our scenario. The formulation below shows that pointwise evaluation of functions in $\mathscr{G}$ may be conducted by the action of the corresponding operator in $S_2(\mathscr{H})$ on the canonical feature map φ. We will refer to this property as the *operator reproducing property*. We visualize the relations between $\mathscr{H} \otimes \mathscr{H}$, $S_2(\mathscr{H})$ and $\mathscr{G}$ in Figure 4.1.

Corollary 4.2.5 (Operator reproducing property)**.** *For every function $F \in \mathscr{G}$ there exists an operator $A := \Theta^{-1}(F) \in S_2(\mathscr{H})$ such that*

$$F(x) = A\varphi(x) \in \mathscr{H} \tag{4.11}$$

for all $x \in E$ with $\|A\|_{S_2(\mathscr{H})} = \|F\|_{\mathscr{G}}$ and vice versa.

Conversely, for any pair $F \in \mathscr{G}$ and $A \in S_2(\mathscr{H})$ satisfying property (4.11) we have $A = \Theta^{-1}(F)$.

Proof. The first assertion directly follows from Theorem 4.2.4 and the construction of Θ. It remains to prove the second assertion. Let $F \in \mathscr{G}$ and define $A := \Theta^{-1}(F)$. By the first assertion, A satifies (4.11). Assume there exists $B \in S_2(\mathscr{H})$ satisfying (4.11). Then by linearity, A and B coincide on $\operatorname{span}\{\varphi(x) \mid x \in E\}$, which is dense in $\mathscr{H}$. By continuity, we therefore have $A = B$. The operator in $S_2(\mathscr{H})$ satisfying (4.11) is therefore uniquely given by $\Theta^{-1}(F)$. ■

Remark 4.2.6 (Operator reproducing property). Not only does Corollary 4.2.5 describe how functions in $\mathscr{G}$ can be evaluated in terms of their operator analogues in $S_2(\mathscr{H})$, it also shows the *implicit* construction of $\mathscr{G}$ via Hilbert–Schmidt operators acting on the RKHS $\mathscr{H}$. In particular, the above result shows that the space of Hilbert–Schmidt operators $S_2(\mathscr{H})$ generates the vRKHS $\mathscr{G}$ via

$$\mathscr{G} = \{F : E \to \mathscr{H} \mid F = A\varphi(\cdot),\, A \in S_2(\mathscr{H})\}.$$

Our previous considerations show that $\mathscr{G}$ is precisely the vRKHS associated with the vector-valued kernel $K := k\mathrm{Id}_{\mathscr{H}}$.

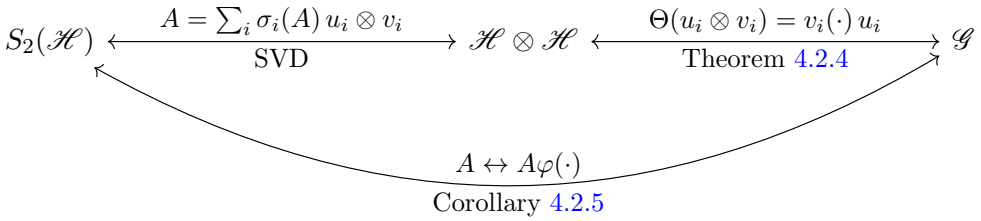

Figure 4.1.: Visualization of the isometric isomorphisms between $S_2(\mathscr{H})$, $\mathscr{H} \otimes \mathscr{H}$ and $\mathscr{G}$. Here, *SVD* refers to the singular value decomposition of compact operators.

Corollary 4.2.5 will be of central importance for our approach. The identification of an $\mathscr{H}$-valued vRKHS function in $\mathscr{G}$ with a corresponding Hilbert–Schmidt operator acting on $\mathscr{H}$ will be used to bridge the gap between vector-valued statistical learning theory and the nonparametric estimation of linear operators (Grünewälder et al., 2013).

4.2.4. Assumptions on $\mathscr{H}$

We impose some technical requirements on the RKHS $\mathscr{H}$ and the corresponding kernel k. Our first three assumptions allow that we can perform Bochner integration without being caught up in measurability and integrability issues later on (Diestel and Uhl, 1977). The fourth and the fifth assumption are needed to ensure that $\mathscr{H}$ supplies the typically used approximation qualities in a function space context.

Assumption 1 (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 2 (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 3 (Existence of second moments)**.** We have $\varphi \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ as well as $\varphi \in L^2(E, \mathcal{F}_E, \nu; \mathscr{H})$. Note that this is equivalent to $\mathbb{E}[\|\varphi(X)\|_{\mathscr{H}}^2] < \infty$ and $\mathbb{E}[\|\varphi(Y)\|_{\mathscr{H}}^2] < \infty$ which trivially holds for all probability measures π, ν on $(E, \mathcal{F}_E)$ case whenever $\sup_{x \in E} k(x, x) < \infty$.

Assumption 4 (C_0-kernel)**.** We assume that $\mathscr{H} \subseteq C_0(E)$, where $C_0(E)$ is the space of continuous functions $f\colon E \to \mathbb{R}$ such that for every $\epsilon > 0$, the set $\{x \mid |f(x)| \geq \epsilon\} \subseteq E$ is compact. In particular, this is the case if $x \mapsto k(x, x)$ is bounded on E and $k(x, \cdot) \in C_0(E)$

for all $x \in E$ (Carmeli et al., 2010, Proposition 2.2).

Assumption 5 (L^2-universal kernel, see Section 4.3)**.** We assume that $\mathscr{H}$ is dense in $L^2(\pi)$. In this case, the kernel k and the RKHS $\mathscr{H}$ are called L^2-*universal* (Carmeli et al., 2010; Sriperumbudur et al., 2011).

Remark 4.2.7. Since not all of our results will need all of the above assumptions, we collect some general implications of the different assumptions here.

1. Assumptions 1–3 ensure that $\mathscr{H}$ can be continuously embedded into both $L^2(\pi)$ and $L^2(\nu)$ (see Section 4.3).

2. The combination of Assumption 4 and Assumption 5 implies that $\mathscr{H}$ is even dense in $L^2(\nu)$ for all probability measures ν on $(E, \mathcal{F}_E)$ (Carmeli et al., 2010, Theorem 4.1 and Corollary 4.2).

3. Instead of Assumption 5, it is sometimes required in the literature that $\mathscr{H}$ is dense in $C_0(E)$ with respect to the supremum norm. This property is usually called C_0-*universality*. One can show that when Assumption 4 holds, C_0-universality is equivalent to L^2-universality (Sriperumbudur et al., 2011).

4. When Assumptions 1–5 are satisfied, then the vRKHS $\mathscr{G}$ induced by the kernel $K = k\mathrm{Id}_{\mathscr{H}}$ is dense in both $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ and $L^2(E, \mathcal{F}_E, \nu; \mathscr{H})$ (see Carmeli et al. 2010, Example 6.3 and Carmeli et al. 2010, Theorem 4.1). This is important for us, as we will make use of this fact later on.

Example 4.2.8. For $E \subseteq \mathbb{R}^d$, well-known translation invariant kernels such as the *Gaussian kernel* or *Laplacian kernel* satisfy all of the above assumptions for arbitrary probability measures π, ν on $(E, \mathcal{F}_E)$ (Sriperumbudur et al., 2011).

4.3. Integral and inclusion operators

The Assumptions 1–3 imply that $\mathscr{H}$ can be embedded into spaces of square integrable functions. This fact and its connections to integral operators defined by the corresponding kernels play a fundamental role in statistical learning theory.

4.3.1. Real-valued RKHS

We begin with general statements for the scalar kernel k (see for example Steinwart and Christmann, 2008, Chapter 4.3). Let the Assumptions 1–3 be satisfied. The *inclusion operator* $i_\pi : \mathscr{H} \to L^2(\pi)$ given by $f \mapsto [f]_\sim \in L^2(\pi)$ identifies $f \in \mathscr{H}$ with its equivalence class of π-a.e. defined functions in $L^2(\pi)$. It is bounded with $\|i_\pi\| \leq \|\varphi\|_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$

and Hilbert–Schmidt. The adjoint of i_π is the integral operator $i_\pi^*\colon L^2(\pi) \to \mathscr{H}$ given by

$$[i_\pi^* f](x) = \int_E k(x, x') f(x') \,\mathrm{d}\pi(x'), \quad f \in L^2(\pi).$$

The kernel k is L^2-universal if and only if i_π^* is injective.

The operator $C_{XX} := i_\pi^* i_\pi\colon \mathscr{H} \to \mathscr{H}$ is the *kernel covariance operator* associated with the measure π given by

$$C_{XX} = \int_E \varphi(x) \otimes \varphi(x) \,\mathrm{d}\pi(x) = \mathbb{E}[\varphi(X) \otimes \varphi(X)],$$

where the integral converges in trace norm. We define all of the above concepts analogously for the measure ν and the corresponding random variable Y. The *kernel cross-covariance operator* (Baker, 1973) of X and Y is the trace class operator given by

$$C_{YX} := \iint_{E \times E} \varphi(y) \otimes \varphi(x) \, p(x, \mathrm{d}y) \mathrm{d}\pi(x) = \mathbb{E}[\varphi(Y) \otimes \varphi(X)].$$

Both operators satisfy $\langle h,\, C_{XX} f \rangle_{\mathscr{H}} = \langle h,\, f \rangle_{L^2(\pi)} = \mathbb{E}[f(X)h(X)]$ as well as $\langle h,\, C_{YX} f \rangle_{\mathscr{H}} = \mathbb{E}[f(X)h(Y)]$ for all $f, h \in \mathscr{H}$.

Remark 4.3.1 (Scalar RKHSs and integral operators). Although the presented operators $i_\pi^* : L^2(\pi) \to \mathscr{H}$, $i_\pi i_\pi^* : L^2(\pi) \to L^2(\pi)$ and $C_{XX} : \mathscr{H} \to \mathscr{H}$ have the same analytical expression as integral operators, they are fundamentally different objects since they operate on different spaces. However, $i_\pi i_\pi^*$ and C_{XX} share the same nonzero eigenvalues and their eigenfunctions can be related (see for example Rosasco et al., 2010).

Remark 4.3.2 (Inclusion operators and notation). We will sometimes suppress the inclusion operators i_π and i_ν in our notation when the context is clear. In particular, for $f \in \mathscr{H}$ we will simply write $\|f\|_{L^2(\nu)}$ instead of $\|i_\nu f\|_{L^2(\nu)}$. Furthermore, under the above assumptions, we may understand the operator $P i_\nu : \mathscr{H} \to L^2(\pi)$ as a conditional expectation operator acting on functions of $\mathscr{H}$ via

$$[P i_\nu f](x) = \mathbb{E}[f(Y) \mid X = x] \in L^2(\pi) \quad \text{for } f \in \mathscr{H} \tag{4.12}$$

and use the norm of $\mathscr{H}$ on its domain. By abuse of notation, we may write $P : \mathscr{H} \to L^2(\pi)$ instead of $P i_\nu$ for the operator in (4.12). We will emphasize which version of P we refer to by simply distinguishing between $P : \mathscr{H} \to L^2(\pi)$ and $P : L^2(\nu) \to L^2(\pi)$. We write out the corresponding operator norms $\|P\|_{\mathscr{H} \to L^2(\pi)}$ and $\|P\|_{L^2(\nu) \to L^2(\pi)}$ to prevent confusion. Note that by boundedness of i_ν, we have $\|P\|_{\mathscr{H} \to L^2(\pi)} \leq \|i_\nu\| \, \|P\|_{L^2(\nu) \to L^2(\pi)}$. Similarly, for every bounded operator $A : \mathscr{H} \to \mathscr{H}$ we can consider the bounded operator $i_\pi A$ from $\mathscr{H}$ to $L^2(\pi)$, which we will also abbreviate as $A : \mathscr{H} \to L^2(\pi)$. At this point, it is worth mentioning that functions in $\mathscr{H}$ are generally defined pointwise, while elements of $L^2(\pi)$ are equivalence classes of π-a.e. equivalent functions.

4.3.2. $\mathscr{H}$-valued RKHS

Similarly to the above operators defined for the scalar kernel k, we can define the above concepts for the vector-valued kernel $K = k\mathrm{Id}_{\mathscr{H}}$ in the context of Bochner spaces (Carmeli et al., 2006, 2010).

When Assumptions 1–3 are satisfied, the space $\mathscr{G}$ is separable. The elements of $\mathscr{G}$ are $\mathcal{F}_E - \mathcal{F}_{\mathscr{H}}$ measurable functions. Additionally, they are Bochner square integrable w.r.t. π. The inclusion operator $\mathcal{I}_\pi \colon \mathscr{G} \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ given by $F \mapsto [F]_\sim$ is bounded with $\|\mathcal{I}_\pi\| \leq \|\varphi\|_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$.

The adjoint of $\mathcal{I}_\pi$ is the integral operator $\mathcal{I}_\pi^* \colon L^2(E, \mathcal{F}_E, \pi, \mathscr{H}) \to \mathscr{G}$ given by

$$[\mathcal{I}_\pi^* F](x) = \int_E K(x, x') F(x') \,\mathrm{d}\pi(x'), \quad F \in L^2(E, \mathcal{F}_E, \pi, \mathscr{H}).$$

The operator $T := \mathcal{I}_\pi^* \mathcal{I}_\pi \colon \mathscr{G} \to \mathscr{G}$ is the *generalized covariance operator* (also called *frame operator*, Carmeli et al. 2006) associated with the measure π given by

$$TF = \int_E K_x K_x^* F \,\mathrm{d}\pi(x) \tag{4.13}$$

for all $F \in \mathscr{G}$. T is bounded.

The following example shows that the generalized covariance operator T associated with $K(x, x') = k(x, x')\mathrm{Id}_{\mathscr{H}}$ is noncompact in general. This fact will be very important for us later on in the context of inverse problems.

Example 4.3.3 (Noncompact generalized covariance operator T). It is easy to see that for commonly used radial kernels k such as the Gaussian kernel on $E \subseteq \mathbb{R}^d$, the generalized covariance operator T is never compact.
Consider a measurable kernel $k : E \times E \to \mathbb{R}$ which induces an infinite-dimensional RKHS $\mathscr{H}$ satisfying Assumptions 1 and 2. Assume $k(x, y) > 0$ for all $x, y \in E$ and $k(x, x) = 1$ for all $x \in E$. Let $K = k\mathrm{Id}_{\mathscr{H}}$ and $(e_i)_{i\in\mathbb{N}} \subset \mathscr{H}$ be an ONS. We fix some $x' \in E$ and define $F_i := K_{x'} e_i \in \mathscr{G}$ for all $i \in \mathbb{N}$. Note that we have

$$\langle K_{x'} e_i,\, K_{x'} e_j \rangle_{\mathscr{G}} = \big\langle k(x', \cdot) e_i,\, k(x', \cdot) e_j \big\rangle_{\mathscr{G}} = k(x', x') \,\langle e_i,\, e_j \rangle_{\mathscr{H}} = \delta_{ij},$$

i.e., $(F_i)_{i\in\mathbb{N}}$ is an ONS in $\mathscr{G}$. Then it is possible to show that $(TF_i)_{i\in\mathbb{N}}$ consists of orthogonal elements of the same length:

$$\begin{aligned}
\langle TF_i,\, TF_j \rangle_{\mathscr{G}} &= \left\langle \int_E K_x F_i(x) \mathrm{d}\pi(x),\, \int_E K_x F_j(x) \mathrm{d}\pi(x) \right\rangle_{\mathscr{G}} \\
&= \left\langle \int_E k(x', x)\, K_x e_i \mathrm{d}\pi(x),\, \int_E k(x', x)\, K_x e_j \mathrm{d}\pi(x) \right\rangle_{\mathscr{G}} \\
&= \iint_{E^2} k(x', x) k(x', y) \left\langle K_y^* K_x e_i,\, e_j \right\rangle_{\mathscr{H}} \mathrm{d}[\pi \otimes \pi](x, y)
\end{aligned}$$

$$= \iint_{E^2} k(x', x)k(x', y)k(x, y) \langle e_i, e_j \rangle_{\mathscr{H}} \, \mathrm{d}[\pi \otimes \pi](x, y) = M\delta_{ij}$$

with the constant $M := \iint_{E^2} k(x', x)k(x', y)k(x, y) \, \mathrm{d}[\pi \otimes \pi](x, y) > 0$, which is independent of $i, j \in \mathbb{N}$. Consequently, we have $\|TF_i - TF_j\|_{\mathscr{G}}^2 = \|TF_i\|_{\mathscr{G}}^2 + \|TF_j\|_{\mathscr{G}}^2 = 2M$ for all $i \neq j$, i.e., no subsequence of $(TF_i)_{i\in\mathbb{N}}$ can be Cauchy. We therefore have constructed a bounded sequence $(F_i)_{i\in\mathbb{N}}$ in $\mathscr{G}$ such that $(TF_i)_{i\in\mathbb{N}}$ does not contain a convergent subsequence in $\mathscr{G}$, implying that T is not compact.

4.4. Kernel mean embedding and maximum mean discrepancy

Under Assumptions 1–3, the Bochner integrability of the feature map $\varphi : E \to \mathscr{H}$ can be elegantly used in combination with the reproducing property of $\mathscr{H}$ to express expectation operations via simple linear algebra.

In particular, the *kernel mean embedding* (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017) of the probability measure π defined by the Bochner expectation

$$\mu_\pi := \int_E \varphi(x) \, \mathrm{d}\pi(x) = \mathbb{E}[\varphi(X)] \in \mathscr{H} \tag{4.14}$$

naturally satisfies the expectation reproducing property

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\langle f, \varphi(X)\rangle_{\mathscr{H}}\right] = \langle f, \mu_\pi\rangle_{\mathscr{H}} \quad \text{for all } f \in \mathscr{H}. \tag{4.15}$$

We call the RKHS $\mathscr{H}$ (or equivalently the corresponding kernel k) *characteristic*, if the *mean embedding map*

$$\pi \mapsto \int_E \varphi(x) \, \mathrm{d}\pi(x) = \mu_\pi \in \mathscr{H}$$

defined on all probability measures on $(E, \mathcal{F}_E)$ for which the integral converges is injective.

Remark 4.4.1 (The RKHS $\mathscr{H}$ is characteristic). Our Assumptions 4 and 5 imply that $\mathscr{H}$ is characteristic (Carmeli et al., 2010; Sriperumbudur et al., 2010, 2011).

For two probability measures π, ν on $(E, \mathcal{F}_E)$, the *maximum mean discrepancy* (MMD) is defined by

$$d_k(\pi, \nu) := \sup_{\substack{f\in\mathscr{H} \\ \|f\|_{\mathscr{H}}\leq 1}} \left| \int_E f(x) \mathrm{d}\pi(x) - \int_E f(x) \mathrm{d}\nu(x) \right| = \|\mu_\pi - \mu_\nu\|_{\mathscr{H}} . \tag{4.16}$$

For characteristic kernels, the MMD constitutes a metric on the set of probability measures on $(E, \mathcal{F}_E)$. This fact has been used as a powerful tool in RKHS-based inference (Gretton

et al., 2012; Sejdinovic et al., 2013). The MMD can be interpreted as a so-called *integral probability metric* (Müller, 1997) and has been shown to metrize weak convergence of measures under some mild conditions (Simon-Gabriel et al., 2020).

Transferring (4.14) to a regular conditional distribution of Y given X, we define the $\mathscr{H}$-valued *conditional mean embedding* (CME) function (Park and Muandet, 2020a)

$$F_p(x) := \int_E \varphi(y)\, p(x, \mathrm{d}y) = \mathbb{E}[\varphi(Y) \mid X = x] \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H}) \tag{4.17}$$

and obtain a pointwise conditional version of the expectation reproducing property (4.15) as

$$\mathbb{E}[f(Y) \mid X = x] = \langle f,\, F_p(x) \rangle_{\mathscr{H}} \quad \text{for all } f \in \mathscr{H} \text{ and } x \in E. \tag{CME}$$

The fact that F_p (or analogously any other regular version of $\mathbb{E}[\varphi(Y) \mid X = \cdot]$) is a well-defined element in $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ can be seen by using Jensen's inequality for conditional Bochner expectations as

$$\begin{aligned} \|F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} &= \int_E \|F_p(x)\|^2_{\mathscr{H}} \,\mathrm{d}\pi(x) \\ &\leq \iint_{E\times E} \|\varphi(y)\|^2_{\mathscr{H}}\, p(x, \mathrm{d}y)\mathrm{d}\pi(x) = \mathbb{E}[\|\varphi(Y)\|^2_{\mathscr{H}}] < \infty \end{aligned}$$

together with Assumption 3.

The approximation of F_p is a key concept in a wide variety of models for kernel-based inference. If C_{XX} is injective, Song et al. (2009) and Fukumizu et al. (2013) show that under the assumption

$$\mathbb{E}[f(Y) \mid X = \cdot] = \langle f,\, F_p(\cdot) \rangle_{\mathscr{H}} \in \mathscr{H} \quad \text{for all } f \in \mathscr{H}, \tag{4.18}$$

we have a closed form expression of F_p via

$$F_p(x) = C_{YX} C^{\dagger}_{XX} \varphi(x) \tag{4.19}$$

for all $x \in E$ such that $\varphi(x) \in \operatorname{range}(C_{XX})$. Here, the (generally unbounded and not globally defined) operator $C^{\dagger}_{XX} \colon \operatorname{range}(C_{XX}) + \operatorname{range}(C_{XX})^{\perp} \to \mathscr{H}$ is the *Moore–Penrose pseudoinverse* of C_{XX} (see Appendix A.6). The assumption (4.18) is generally not satisfied (see Klebanov et al. 2020 for a detailed investigation of arising problems). Grünewälder et al. (2012a) and Park and Muandet (2020a) show that a Tikhonov–Phillips regularized version of the estimate of (4.19) can be understood as an empirical approximation of F_p with functions in $\mathscr{G}$ in a least squares regression context. However, no approximation qualities of the CME in the L^2-operator context are considered. We extend this theory in the next section and connect it to the CME regression model later on.

4.5. Nonparametric operator approximation

We now investigate the nonparametric estimation of P by means of reproducing kernels. We establish an approximation-theoretic viewpoint which connects our derivation to the theory of inverse problems and supervised learning.

The general idea is to estimate the operator $Pi_\nu : \mathscr{H} \to L^2(\pi)$, instead of P. By our previous considerations, Pi_ν is Hilbert–Schmidt under Assumptions 1–3 (see Steinwart and Christmann, 2008, Chapter 4.3), which already justifies the endeavor to approximate Pi_ν with operators of finite rank in the associated operator norm $\|\cdot\|_{\mathscr{H}\to L^2(\pi)}$. Our main problem is now given by the fact that we must restrict ourselves to a suitable hypothesis space which we can impose by using reproducing kernels. Furthermore, using the operator norm $\|\cdot\|_{\mathscr{H}\to L^2(\pi)}$ as an objective function introduces a supremum over the unit ball in $\mathscr{H}$, which we can not estimate consistently. The next result solves these issues. It shows that if we restrict ourselves to the class of Hilbert–Schmidt operators on $\mathscr{H}$ (which we examined in Section 4.2.3) and interpret the image space as $L^2(\pi)$, we obtain a surrogate problem in terms of an infinite-dimensional linear regression.

Note again that we may drop the inclusion operators i_ν and i_π from our notation for simplicity as described in Remark 4.3.2.

Theorem 4.5.1 (Regression and conditional mean approximation)**.** *Under the Assumptions 1–3, we have for every operator $A \in S_2(\mathscr{H})$ that*

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \leq \mathbb{E}\left[\|F_p(X) - A^*\varphi(X)\|^2_{\mathscr{H}}\right] = \|F_p - A^*\varphi(\cdot)\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}\,.$$

The given bound is sharp.

Proof. Let $A \in S_2(\mathscr{H})$. We have

$$\begin{aligned}
\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} &= \sup_{\|f\|_{\mathscr{H}}=1} \|Af - Pf\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \|[Af](\cdot) - \mathbb{E}[f(Y) \mid X = \cdot]\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \left\|\langle Af, \varphi(\cdot)\rangle_{\mathscr{H}} - \langle f, F_p(\cdot)\rangle_{\mathscr{H}}\right\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \left\|\langle f, A^*\varphi(\cdot) - F_p(\cdot)\rangle_{\mathscr{H}}\right\|^2_{L^2(\pi)} \\
&= \sup_{\|f\|_{\mathscr{H}}=1} \mathbb{E}\left[\langle f, A^*\varphi(X) - F_p(X)\rangle^2_{\mathscr{H}}\right] \\
&\leq \sup_{\|f\|_{\mathscr{H}}=1} \mathbb{E}\left[\|f\|^2_{\mathscr{H}}\,\|A^*\varphi(X) - F_p(X)\|^2_{\mathscr{H}}\right] \\
&= \mathbb{E}\left[\|A^*\varphi(X) - F_p(X)\|^2_{\mathscr{H}}\right] = \|A^*\varphi(\cdot) - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}\,,
\end{aligned}$$

where we use the reproducing property in $\mathscr{H}$ in the third equality and the Cauchy–Schwarz inequality. It is clear that the above bound is sharp by considering the case that we have $\mathbb{P}$-a.e. $A^*\varphi(X) - F_p(X) = h$ for some constant $h \in \mathscr{H}$. In this case the above bound is attained when we choose $f = h/\|h\|_{\mathscr{H}}$ in the supremum. ■

We note that this result can also be interpreted as an improvement of a surrogate risk bound derived by Grünewälder et al. (2012a, Section 3.1) and later on used by Park and Muandet (2020a) to approximate the CME. We will elaborate on this fact in more detail later on (see Section 4.5.2 and Remark 4.5.13 in particular). Our bound has a significant impact from an approximation viewpoint, which we will highlight in our following examination.

Theorem 4.5.2 (Approximation by Hilbert–Schmidt operators)**.** *Let Assumptions 1–5 be satisfied. Then for every $\delta > 0$, there exists a Hilbert–Schmidt operator $A\colon \mathscr{H} \to \mathscr{H}$, such that*

$$\|A - P\|_{\mathscr{H} \to L^2(\pi)} < \delta. \tag{4.20}$$

Proof. By Corollary 4.2.5, every operator $A^* \in S_2(\mathscr{H})$ corresponds to a function $F \in \mathscr{G}$ via $F(x) = A^*\varphi(x)$ for all $x \in E$ and vice versa. The space $\mathscr{G}$ is densely embedded into $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ by Remark 4.2.7(4). For every $\delta > 0$ we therefore have an operator $A^* \in S_2(\mathscr{H})$ such that the bound $\|A^*\varphi(\cdot) - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} = \|F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} < \delta$ holds. Together with the bound obtained in Theorem 4.5.1, this proves the assertion. ■

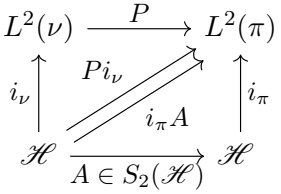

Figure 4.2.: Nonparametric approximation of P over functions in $\mathscr{H}$ by a Hilbert–Schmidt operator $A \in S_2(\mathscr{H})$. Theorem 4.5.2 shows that $Pi_\nu \approx i_\pi A$ to arbitrary accuracy in the associated operator norm. The operator A is approximated by finite-rank operators on $\mathscr{H}$ in Corollary 4.5.4.

Remark 4.5.3. Some remarks related to Theorem 4.5.2 are in order.

1. We do not require $P : L^2(\nu) \to L^2(\pi)$ to be a Hilbert–Schmidt operator or compact in order for the above statement to hold. Our result is not a contradiction to the known fact that operator norm limits of Hilbert–Schmidt operators are compact. The reason for that is that the compactness property is given with respect to the

norm $\|\cdot\|_{\mathscr{H}}$ on the domain, which is stronger than the norm $\|\cdot\|_{L^2(\nu)}$. Hence, the continuous extension to $A : L^2(\nu) \to \mathscr{H}$ via the known construction for bounded operators (Weidmann, 1980, Theorem 4.5) is generally not compact. This can equivalently be seen by the fact that i_ν does generally not admit a globally defined bounded inverse. We visualize Theorem 4.5.2 in Figure 4.2.

2. The assumptions on $\mathscr{H}$ are not restrictive, as they are well examined in statistical learning theory and often satisfied for particular RKHSs used in practice. It is actually sufficient to only require that $\mathscr{H}$ is dense in $L^2(\rho)$ for any probability measure ρ on $(E, \mathcal{F}_E)$, as this implies denseness in both $L^2(\pi)$ and $L^2(\nu)$.

3. We will later also investigate under which requirements there exists a Hilbert–Schmidt operator $A : \mathscr{H} \to \mathscr{H}$ such that $\|A - P\|_{\mathscr{H} \to L^2(\pi)} = 0$ (see Section 4.5.2).

We conclude that we can now approximate P with operators on $\mathscr{H}$ of finite rank.

Corollary 4.5.4. *Let Assumptions 1-5 be satisfied. Then there exists a sequence of finite-rank operators $(A_n)_{n\in\mathbb{N}}$ from $\mathscr{H}$ to $\mathscr{H}$ such that $\|A_n - P\|_{\mathscr{H} \to L^2(\pi)} \to 0$ as $n \to \infty$.*

Proof. Let $\delta > 0$. By the fact that the finite-rank operators on $\mathscr{H}$ are dense in $S_2(\mathscr{H})$ and Theorem 4.5.1, we can choose $A \in S_2(\mathscr{H})$ as well as a finite-rank operator A_n on $\mathscr{H}$ such that

$$\begin{aligned}\|A_n - P\|_{\mathscr{H} \to L^2(\pi)} &\leq \|A - P\|_{\mathscr{H} \to L^2(\pi)} + \|i_\pi\| \, \|A_n - A\|_{\mathscr{H} \to \mathscr{H}} \\ &\leq \|A - P\|_{\mathscr{H} \to L^2(\pi)} + \|i_\pi\| \, \|A_n - A\|_{S_2(\mathscr{H})} < \frac{\delta}{2} + \frac{\delta}{2}.\end{aligned}$$

■

As we will prove, such a sequence $(A_n)_{n\in\mathbb{N}}$ can be computed from data almost surely.

4.5.1. Measure-theoretic implications of the approximation of P

When $\mathscr{H}$ is characteristic, $P : \mathscr{H} \to L^2(\pi)$ uniquely determines the conditional distribution $p(x, \cdot)$ for π-a.e. $x \in E$ (that is, up to a choice of a regular version of the underlying conditional expectation). This underlines that the conditional expectation operator P interpreted as an operator with the domain $\mathscr{H}$ instead of $L^2(\nu)$ still captures sufficient information about the underlying joint distribution of X and Y. More generally, an approximation of P naturally yields a weighted approximation of the associated Markov kernel p in the MMD. This may provide a foundation for the adaptation of MMD-based hypothesis tests for Markov kernels.

Theorem 4.5.5 (Equivalence to approximation in MMD)**.** *Let Assumptions 1–3 be satisfied. Let $P, P' : \mathscr{H} \to L^2(\pi)$ be two well-defined bounded conditional expectation operators associated with the Markov kernels $p, p' : E \times \mathcal{F}_E \to \mathbb{R}$. Then we have*

$$\left\|P - P'\right\|^2_{\mathscr{H}\to L^2(\pi)} = \int_E d_k(p(x,\cdot), p'(x,\cdot))^2 \,\mathrm{d}\pi(x). \tag{4.21}$$

Proof. We have

$$\begin{aligned}
\left\|P - P'\right\|^2_{\mathscr{H}\to L^2(\pi)} &= \sup_{\substack{f\in\mathscr{H}\\ \|f\|_{\mathscr{H}}=1}} \left\|Pf - P'f\right\|^2_{L^2(\pi)} \\
&= \sup_{\substack{f\in\mathscr{H}\\ \|f\|_{\mathscr{H}}=1}} \left\| \int f(y)\, p(\cdot, \mathrm{d}y) - \int f(y)\, p'(\cdot, \mathrm{d}y) \right\|^2_{L^2(\pi)} \\
&= \sup_{\substack{f\in\mathscr{H}\\ \|f\|_{\mathscr{H}}=1}} \left\| \left\langle f, \int_E \varphi(y)\, p(\cdot, \mathrm{d}y) - \int_E \varphi(y)\, p'(\cdot, \mathrm{d}y) \right\rangle_{\mathscr{H}} \right\|^2_{L^2(\pi)} \\
&= \int_E \left\|\mu_{p(x,\cdot)} - \mu_{p'(x,\cdot)}\right\|^2_{\mathscr{H}} \mathrm{d}\pi(x) \\
&= \int_E d_k(p(x,\cdot), p'(x,\cdot))^2 \,\mathrm{d}\pi(x),
\end{aligned}$$

where we use the reproducing property in $\mathscr{H}$ in the third equality. ∎

Under more restrictive assumptions, the low-dimensional approximation of the adjoint of P by means of the MMD has been proposed in the context of random dynamical systems with a different estimation scheme (Tian and Wu, 2020).

Remark 4.5.6 (Assumptions of Theorem 4.5.5). For simplicity, we do not explicitly assume in Theorem 4.5.5 that the underlying random variables associated with P and P' are distributed with respect to the marginals π and ν. To show the above statement, it is sufficient that both operators are well-defined and bounded when the domain and image space are chosen to be $\mathscr{H}$ and $L^2(\pi)$. The proof of Theorem 4.5.5 shows that $\|P - P'\|^2_{\mathscr{H}\to L^2(\pi)}$ can equivalently be interpreted as the squared $L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ distance between the two conditional mean embeddings $F_p(x) = \int_E \varphi(y)\, p(x, \mathrm{d}y)$ and $F'_p(x) = \int_E \varphi(y)\, p'(x, \mathrm{d}y)$.

Remark 4.5.7 (Approximation of MMD in L^q-norm). Whenever the conditional expectation operators P, P' are well-defined and bounded operators from $\mathscr{H}$ to $L^q(\pi)$ for $1 \leq q \leq \infty$, we can analogously obtain versions of (4.21) for the respective L^q norm. In particular, in this case we have

$$\left\|P - P'\right\|^q_{\mathscr{H}\to L^q(\pi)} = \int_E d_k(p(x,\cdot), p'(x,\cdot))^q \,\mathrm{d}\pi(x). \tag{4.22}$$

When $\mathscr{H}$ is characteristic, we immediately obtain the following result. It shows that conditional expectation operators on $\mathscr{H}$ determine the conditional distribution of the associated random variables uniquely up to a choice of a regular version.

Corollary 4.5.8. *Let Assumptions 1–3 be satisfied and $\mathscr{H}$ be characteristic. With the notation of Theorem 4.5.5, we have $\|P - P'\|_{\mathscr{H}\to L^2(\pi)} = 0$ if and only if $p(x,\cdot) = p'(x,\cdot)$ for π-a.e. $x \in E$.*

Moreover, Corollary 4.5.8 implies that the joint distributions for the class of pairs of random variables X, Y with a fixed marginal $X \sim \pi$ are uniquely determined by $P : \mathscr{H} \to L^2(\pi)$.

Corollary 4.5.9. *Let X, X', Y, Y' be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(E, \mathcal{F}_E)$ such that $X \sim \pi$ and $X' \sim \pi$ and Assumptions 1–3 are satisfied for both pairs X, Y and X', Y'. Let $\mathscr{H}$ be characteristic and $P, P' : \mathscr{H} \to L^2(\pi)$ be bounded conditional expectation operators given by $Pf = \mathbb{E}[f(Y) \mid X = \cdot]$ and $P'f = \mathbb{E}[f(Y') \mid X' = \cdot]$ defined by some Markov kernels p and p' respectively. Then we have $\|P - P'\|_{\mathscr{H}\to L^2(\pi)} = 0$ if and only if $\mathcal{L}(X, Y) = \mathcal{L}(X', Y')$.*

Proof. Let $\|P - P'\|_{\mathscr{H}\to L^2(\pi)} = 0$. For any two events $\mathcal{A}, \mathcal{B} \in \mathcal{F}_E$, we perform the disintegration

$$\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \int_{\mathcal{A}} p(x, \mathcal{B}) \, \mathrm{d}\pi(x) \tag{4.23}$$

and analogously for the pair X', Y'. We apply Corollary 4.5.8, leading to the π-a.e. equivalence $p(\cdot, \mathcal{B}) = p'(\cdot, \mathcal{B})$. This gives $\mathbb{P}[X \in \mathcal{A}, Y \in \mathcal{B}] = \mathbb{P}[X' \in \mathcal{A}, Y' \in \mathcal{B}]$. The converse implication follows analogously. ∎

4.5.2. Least squares regression and connection to the CME

We now describe the theoretical foundation for the construction of an empirical estimate of P based on Theorem 4.5.1. In the process, we will see that our concept is closely related to the CME.

By the operator reproducing property from Corollary 4.2.5 we may rewrite the vRKHS least squares regression problem

$$\operatorname*{arg\,min}_{F \in \mathscr{G}} R(F) \text{ with } R(F) := \mathbb{E}[\|\varphi(Y) - F(X)\|^2_{\mathscr{H}}] \tag{4.24}$$

equivalently as

$$\operatorname*{arg\,min}_{A^* \in S_2(\mathscr{H})} \mathbb{E}[\|\varphi(Y) - A^*\varphi(X)\|^2_{\mathscr{H}}]. \tag{4.25}$$

As is well-known in statistical learning theory (see for example Cucker and Smale, 2002, Proposition 1), for all $F \in L^2(E, \mathcal{F}, \pi; \mathscr{H})$, the risk R allows for the decomposition

$$R(F) = \|F_p - F\|^2_{L^2(E,\mathcal{F},\pi;\mathscr{H})} + R(F_p), \tag{4.26}$$

where $R(F_p)$ represents the irreducible error term (see Theorem A.3.1 for a proof in the infinite-dimensional case). This reduces the regression problem (4.24) and equivalently problem (4.25) to an L^2-approximation of the conditional mean embedding F_p. In this context, F_p is often called *regression function*. Therefore, the so-called *excess risk* $R(F) - R(F_p) = \|F_p - F\|^2_{L^2(E,\mathcal{F},\pi;\mathscr{H})}$ of some estimate $F \in \mathscr{G}$ is typically investigated in nonparametric statistics.

The above formalism allows us to estimate the conditional mean operator P based on our previous results. By Theorem 4.5.1, we have

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \leq \|F_p - A^*\varphi(\cdot)\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} \tag{4.27}$$

for all $A^* \in S_2(\mathscr{H})$. We can now perform the vRKHS regression (4.25) and obtain an approximation of P in the norm $\|\cdot\|^2_{\mathscr{H}\to L^2(\pi)}$ in terms of $A \in S_2(\mathscr{H})$, which we implicitly interpret as an operator from $\mathscr{H}$ to $L^2(\pi)$. Theorem 4.5.2 and Corollary 4.5.4 show that this is possible up to an arbitrary degree of accuracy.

Along the lines of the known work on least squares regression of the form (4.24) or equivalently (4.25), we can distinguish the following two general cases (Szabó et al., 2016):

1. The *well-specified case*, i.e., there exists a regular version of the conditional distribution of Y given X such that $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$. For the well-specified case, we below obtain the known properties of the conditional mean embedding which were derived from the linear-algebraic perspective (Song et al., 2009; Klebanov et al., 2020, 2021).

2. The *misspecified case*, i.e., $F_p \in L^2(\pi) \setminus \mathscr{G}$. This is clearly the more interesting setting, as the well-specified case does typically not occur in practice. From the operator-theoretic perspective, this case has not been investigated yet.

Our previous results allow to reformulate the well-specified case and establish a connection to the CME.

Corollary 4.5.10 (Well-specified case)**.** *Let Assumption 1–3 be satisfied. Consider a fixed regular version of the distribution of Y conditioned on X given by some Markov kernel $p : E \times \mathcal{F}_E \to \mathbb{R}$. The following statements are equivalent:*

(i) We have $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$.

(ii) There exists an operator $A \in S_2(\mathscr{H})$ *such that*

$$[Af](x) = \langle Af,\, \varphi(x)\rangle_{\mathscr{H}} = \langle f,\, A^*\varphi(x)\rangle_{\mathscr{H}} = \mathbb{E}[f(Y) \mid X = x] \tag{4.28}$$

for all $x \in E$ *and* $f \in \mathscr{H}$.

Both (i) and (ii) imply (iii):

(iii) There exists an operator $A \in S_2(\mathscr{H})$ *which satisfies* $\|A - P\|_{\mathscr{H}\to L^2(\pi)} = 0$.

Proof. We show that (i) is equal to (ii). Let $F_p(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] \in \mathscr{G}$. Let $A^* \in S_2(\mathscr{H})$ be the unique operator such that $A^*\varphi(\cdot) = F_p(\cdot)$ by Corollary 4.2.5. By the reproducing property in $\mathscr{H}$, we can verify (4.28) immediately. For the converse implication, let (4.28) be satisfied for some operator $A^* \in S_2(\mathscr{H})$. Then by Corollary 4.2.5, we have the function $F \in \mathscr{G}$ with $F(\cdot) = A^*\varphi(\cdot)$ such that

$$\langle f,\, F(x)\rangle_{\mathscr{H}} = \mathbb{E}[f(Y) \mid X = x] = \mathbb{E}[\langle f,\, \varphi(Y)\rangle_{\mathscr{H}} \mid X = x] \tag{4.29}$$

for all $f \in \mathscr{H}$. The right-hand side of 4.29 is equal to $\langle f,\, \mathbb{E}[\varphi(Y)_{\mathscr{H}} \mid X = x]\rangle_{\mathscr{H}}$ for all $x \in E$ and $f \in \mathscr{H}$, we therefore have $F(\cdot) = \mathbb{E}[\varphi(Y) \mid X = \cdot] = F_p(\cdot) \in \mathscr{G}$ as claimed. The last statement follows from Theorem 4.5.1 by inserting A^* into the right-hand side of the bound, giving $\|A - P\|_{\mathscr{H}\to L^2(\pi)} = 0$. ■

Remark 4.5.11 (Connection to CME and well-specified case). By comparing (4.28) to the expectation reproducing property (CME), we see that in the well-specified case, the operator A^* satisfying (4.28) is exactly the operator which was introduced by Song et al. (2009) as the original conditional mean embedding. *That is, we obtain the approximation of* P *from* $\mathscr{H}$ *to* $L^2(\pi)$ *as the adjoint of the CME.* A similar connection was established by Klus et al. (2020) under the restrictive assumptions of Song et al. (2009) in the context of Markov transition operators.

Remark 4.5.12 (Well-specified case closed form solution). Klebanov et al. (2021, Theorem 5.8) prove in a slightly different context of tensor product spaces, without explicitly using vRKHSs, that in the well-specified case the operator A^* satisfying (4.28) can be expressed in terms of the covariance operators as $A^* = (C_{XX}^{\dagger} C_{XY})^*$. In fact, this proves that $(C_{XX}^{\dagger} C_{XY})^*$ is Hilbert–Schmidt in this case.

Remark 4.5.13 (Surrogate risk bound for the CME). In the well-specified case, Park and Muandet (2020a) investigate the estimation of the CME in terms of (4.24). Their results build upon the surrogate risk bound

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \leq R(A^*\varphi(\cdot)),$$

originally formulated by Grünewälder et al. (2012a). Our Theorem 4.5.1 improves this bound and eliminates the need for additional approximation results (see for example

Grünewälder et al., 2012a, Theorem 3.2) for the analysis of the misspecified case. By (4.26), our bound from Theorem 4.5.1 equals to

$$\|A - P\|^2_{\mathscr{H}\to L^2(\pi)} \leq R(A^*\varphi(\cdot)) - R(F_p),$$

which allows the approximation up to an arbitrary accuracy and removes the excess term $R(F_p)$.

We have seen that in the well-specified case, our results align with prior work on the CME. In the practically more relevant misspecified case however, the bound given by Theorem 4.5.1 significantly simplifies the theory of approximating the CME.

4.6. Regularization and empirical estimation

We now connect our previous results to the theory of supervised learning and derive empirical estimators of P. To this end, we will briefly review how the regression problem (4.24) can be formulated in terms of an inverse problem. The decomposition of R in (4.26) allows to obtain a solution by approximating F_p with functions in $\mathscr{G}$. This framework allows to derive the well-known formalism for supervised learning and regularization theory which will yield estimates of P. We refer to the seminal work for least squares regression with vRKHSs (De Vito and Caponnetto, 2005; Caponnetto and De Vito, 2007) for more details. This section contains the reformulation of our setting in terms of known results, making the theory of vRKHS regression applicable for the estimation of P. We use this framework to derive new results in Section 4.6.3.

4.6.1. Inverse problem

In the misspecified case, it is not necessarily clear that the minimizer of R over $\mathscr{G}$ exists. The analytical nature of this question can be naturally expressed in terms of an inverse problem. For the necessary background on inverse problems in Hilbert spaces and regularization theory, we refer to Engl et al. (1996) and Appendix A.6. We will formulate (4.24) a bit more verbosely in terms of the inclusion $\mathcal{I}_\pi : \mathscr{H} \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$, so that the connection to the inverse problem becomes clear.

If $F \in \mathscr{G}$, we have by (4.26) that

$$R(F) = \|\mathcal{I}_\pi F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})} + R(F_p).$$

Finding $F_{\mathscr{G}} := \arg\min_{F\in\mathscr{G}} R(F)$ is therefore equivalent to finding $F_{\mathscr{G}} \in \mathscr{G}$ such that

$$\|\mathcal{I}_\pi F - F_p\|^2_{L^2(E,\mathcal{F}_E,\pi;\mathscr{H})}$$

is minimal. As is well-known from the theory of inverse problems, this is equivalent to finding the optimal solution $F_{\mathscr{G}}$ of the potentially ill-posed inverse problem

$$\mathcal{I}_\pi F = F_p, \quad F \in \mathscr{G}. \tag{4.30}$$

The inverse problem (4.30) is again equivalent to finding the solution of the so-called *normal equation* (Engl et al., 1996, Theorem 2.6) given by

$$(\mathcal{I}_\pi^* \mathcal{I}_\pi) F = TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G}.$$

In particular, we obtain the following solution.

Theorem 4.6.1 (Regression solution)**.** *Let Assumptions 1–3 be satisfied. The optimal solution*

$$F_{\mathscr{G}} = \operatorname*{arg\,min}_{F \in \mathscr{G}} R(F) = \operatorname*{arg\,min}_{F \in \mathscr{G}} \|\mathcal{I}_\pi F - F_p\|^2_{L^2(E, \mathcal{F}_E, \pi; \mathscr{H})}$$

exists if and only if $\mathcal{I}_\pi^* F_p \in \operatorname{range}(T) + \operatorname{range}(T)^\perp =: \operatorname{dom}(T^\dagger)$,[1] *where the operator* $T^\dagger : \operatorname{range}(T) + \operatorname{range}(T)^\perp \to \mathscr{G}$ *is the pseudoinverse of* T*. In this case,* $F_{\mathscr{G}}$ *is given by the solution to the normal equation*

$$TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G} \tag{4.31}$$

by $F_{\mathscr{G}} = T^\dagger \mathcal{I}_\pi^* F_p$.

Remark 4.6.2 (Limitations of existing literature). Theorem 4.6.1 and the resulting normal equation (4.31) show that our surrogate problem is essentially a (potentially ill-posed) inverse problem with the following technical features:

(i) both the forward operator T and the right-hand side $\mathcal{I}_\pi^* F_p$ are unknown and must be discretized by sampling from $\mathcal{L}(Y, X)$ and

(ii) the forward operator T is in general not compact but only bounded, as Example 4.3.3 shows.

However, results on uniform upper and lower bounds for convergence rates in a vector-valued learning scenario are typically investigated in the case where the forward operator T of the problem (4.31) is trace class (Caponnetto and De Vito, 2007; Rastogi and Sampath, 2017; Rastogi et al., 2020, and references therein). In particular, the aforementioned authors assume $K_x K_x^* \in S_1(\mathscr{H})$ for all $x \in E$ and use the *effective dimension*

$$\mathcal{N}(\lambda) := \operatorname{Tr}\left((T + \lambda \operatorname{Id}_{\mathscr{G}})^{-1} T\right) \qquad \text{for } \lambda > 0$$

as the central tool in order to analyze the convergence of kernel-based regression problems (the reader may also refer to Blanchard and Mücke 2018; Lu et al. 2020; Lin et al. 2020

[1] An equivalent condition is $\Pi F_p \in \operatorname{range}(\mathcal{I}_\pi)$, where $\Pi \colon L^2(E, \mathcal{F}_E, \pi; \mathscr{H}) \to L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$ is the orthogonal projection onto the closure of $\operatorname{range}(\mathcal{I}_\pi)$.

for the scalar case). Example 4.3.3 shows that $\mathcal{N}(\lambda)$ is generally not finite in our setting. Moreover, most results on statistical inverse problems with noncompact forward operators seem to be derived under the assumption that the forward operator is known (see for example Cavalier, 2006; Bissantz et al., 2007) and do therefore not directly transfer to our scenario. Adapting these results in our setting would need a thorough perturbation analysis of the continuous spectrum of T. In addition, discretizing $\mathscr{G}$ in the noncompact case may introduce additional difficulties, see Remark 4.6.3.
To the best of our knowledge, Park and Muandet (2020a,b) are the only authors who address the estimation under assumptions which are satisfied in our case for Tikhonov regularization in the well-specified case (see Section 4.6.3). As these problems require a foundational new approach in the context of supervised learning problems, they are out of the scope of this work.

Remark 4.6.3 (Discretization of T). Note that due to the noncompactness of $\mathscr{G}$, a bit of caution is required when discussing its discretization. In particular, a naive estimate of T would be the Monte Carlo sum

$$T_n := \frac{1}{n}\sum_{i=1}^{n} K_{X_i} K_{X_i}^*$$

for iid $X_i \sim \pi$ and one would think that some strong law of large numbers and concentration results in Banach spaces would lead to the desired convergence results $T_n \to T$ in operator norm. Note that the Banach space of bounded operators on $\mathscr{G}$ is not separable, even if $\mathscr{G}$ itself is separable.[2] This fact may lead to measurability issues of the $\mathfrak{B}(\mathscr{G})$-valued object

$$\xi := K_X K_X^*.$$

Because of this fact, we defined the operator T in (4.13) pointwise as

$$TF = \int_E K_x K_x^* F \,\mathrm{d}\pi(x), \quad F \in \mathscr{G}$$

instead of an integral over the object ξ as defined above – which would need to converge in operator norm. As previously mentioned, available literature on vector-valued regression imposes the assumption $K_x K_x^* \in S_1(\mathscr{H})$, which is not satisfied in our scenario. In addition, versions of the strong law of large numbers in Banach spaces typically require additional properties (Ledoux and Talagrand, 1991, Section 7). For simplicity, we will therefore consider the strong operator topology formulation

$$TF = \int_E K_x K_x^* F \,\mathrm{d}\pi(x) = \mathbb{E}[\xi F] \in \mathscr{G} \tag{4.32}$$

for every $F \in \mathscr{G}$ instead of the norm topology on $\mathfrak{B}(\mathscr{G})$.

[2] This can be proven with the fact that the sequence space ℓ^∞ – which is not separable – can be isometrically embedded into $\mathfrak{B}(\mathscr{G})$.

4.6.2. Sampling operators and empirical estimation

For simplicity, we assume that the optimal solution $F_{\mathscr{G}} = \arg\min_{\mathscr{G}} R(F)$ exists, i.e., we have $\mathcal{I}_\pi^* F_p \in \mathrm{dom}(T^\dagger)$. We wish to compute a solution of the normal equation

$$TF = \mathcal{I}_\pi^* F_p, \quad F \in \mathscr{G} \tag{4.33}$$

in terms of $F_{\mathscr{G}} = T^\dagger \mathcal{I}_\pi^* F_p$ based on an empirical realization of $(X_t)_{t\in\mathbb{Z}}$.

In order to do this, we must discretize T as well as the right-hand side $\mathcal{I}_\pi^* F_p$. We now face the problem that (4.33) may be *ill-posed* in the sense that the solution does not continously depend on $\mathcal{I}_\pi^* F_p$ (and of course on T as well). To still be able to perform an estimation, a *regularization strategy* (Engl et al., 1996) is needed to ensure well-posedness in practice.

Let $\{g_\lambda(T) : \mathscr{G} \to \mathscr{G} \,|\, \lambda \in (0,\infty]\}$ be a regularization strategy.[3] For a fixed regularization parameter $\lambda > 0$, we define the regularized solution

$$F_\lambda := g_\lambda(T)\mathcal{I}_\pi^* F_p \in \mathscr{G}. \tag{4.34}$$

We now discretize the regularized problem (4.34) based on the iid data

$$\mathbf{z} := ((X_1, Y_1), \ldots, (X_n, Y_n))$$

with $(X_i, Y_i) \sim \mathcal{L}(X, Y)$. We generalize the *sampling operator approach* (Smale and Zhou, 2005) from the scalar setting to the vector-valued scenario and derive an empirical estimate of F_λ. Given the data above, we define the *sampling operator* $S_{\mathbf{x}} \colon \mathscr{G} \to \mathscr{H}^n$ given by $S_{\mathbf{x}} F := (F(X_t))_{t=1}^n = (K_{X_t}^* F)_{t=1}^n$. Here, we consider $\mathscr{H}^n$ as a Hilbert space equipped with the inner product

$$\langle \mathbf{f}, \mathbf{h} \rangle_{\mathscr{H}^n} := \frac{1}{n} \sum_{i=1}^n \langle f_i, h_i \rangle_{\mathscr{H}}$$

for $\mathbf{f} = (f_1, \ldots, f_n) \in \mathscr{H}^n$ and $\mathbf{h} = (h_1, \ldots, h_n) \in \mathscr{H}^n$. It is easy to see that the adjoint of $S_{\mathbf{x}}$ is the operator $S_{\mathbf{x}}^* \colon \mathscr{H}^n \to \mathscr{G}$ given by

$$S_{\mathbf{x}}^* \mathbf{h} = \frac{1}{n} \sum_{i=1}^n K_{X_i} h_i$$

for all $\mathbf{h} \in \mathscr{H}^n$ and the operator $T_{\mathbf{x}} := S_{\mathbf{x}}^* S_{\mathbf{x}} \colon \mathscr{G} \to \mathscr{G}$ satisfies

$$T_{\mathbf{x}} F = S_{\mathbf{x}}^* S_{\mathbf{x}} F = \frac{1}{n} \sum_{i=1}^n K_{X_i} K_{X_i}^* F$$

[3] We require $\{g_\lambda(T) : \mathscr{G} \to \mathscr{G} \,|\, \lambda \in (0,\infty]\}$ to be a parametrized family of *globally defined bounded operators* satisfying $g_\lambda(T)F \to T^\dagger F$ for all $F \in \mathrm{dom}(T^\dagger)$ as $\lambda \to 0$, see Appendix A.6.

for all $F \in \mathscr{G}$. Based on these considerations, we will use $S_{\mathbf{x}}^*$ and $T_{\mathbf{x}}$ as empirical estimates for $\mathcal{I}_\pi^*$ and T respectively based on the data $\mathbf{x}$. We define the *target data vector* $\Upsilon := (\varphi(Y_1), \ldots, \varphi(Y_n)) \in \mathscr{H}^n$ and obtain the empirical regularized solution

$$F_{\lambda,\mathbf{z}} := g_\lambda(T_{\mathbf{x}}) S_{\mathbf{x}}^* \Upsilon \in \mathscr{G} \tag{4.35}$$

as the discretized analogue of the analytical regularized solution (4.34).

Via the identification of F_λ and $F_{\lambda,\mathbf{z}}$ with operators through the isomorphism Θ in Corollary 4.2.5, we obtain the *analytical regularized operator solution*

$$A_\lambda := [\Theta^{-1}(F_\lambda)]^* \in S_2(\mathscr{H})$$

as well as the *empirical regularized operator solution*

$$A_{\lambda,\mathbf{z}} := [\Theta^{-1}(F_{\lambda,\mathbf{z}})]^* \in S_2(\mathscr{H}),$$

i.e., $F_\lambda(x) = A_\lambda \varphi(x)$ and $F_{\lambda,\mathbf{z}}(x) = A_{\lambda,\mathbf{z}} \varphi(x)$ for all $x \in E$.

4.6.3. Tikhonov-Phillips regularization

For the remainder of this chapter, we will restrict ourselves to the Tikhonov–Phillips regularization approach (Phillips, 1962; Tikhonov and Arsenin, 1977) to solve the (potentially ill-posed) inverse problem given by Theorem 4.6.1 in order to obtain the optimal solution $F_{\mathscr{G}}$ in $\mathscr{G}$ of the surrogate problem (assuming it exists).

General framework

Tikhonov–Phillips regularization is formally obtained by choosing the regularization strategy $g_\lambda(T) := (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} \in \mathfrak{B}(\mathscr{G})$. We replace the risk R with the *regularized risk*

$$R_\lambda(F) := R(F) + \lambda \|F\|_{\mathscr{G}}^2 \tag{4.36}$$

with a regularization parameter $\lambda > 0$. The unique minimizer of (4.36) exists for all λ and is exactly given by the regularized solution $F_\lambda = (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} \mathcal{I}_\pi^* F_p$, which is a standard result in inverse problems (Engl et al., 1996, Theorem 5.1). Based on the data $\mathbf{z}$, we define the *regularized empirical risk*

$$R_{\lambda,\mathbf{z}}(F) := \frac{1}{n} \sum_{i=1}^n \|\varphi(Y_i) - F(X_i)\|_{\mathscr{H}}^2 + \lambda \|F\|_{\mathscr{G}}^2 \tag{4.37}$$

for all $F \in \mathscr{G}$. We can reformulate (4.37) in terms of the sampling operator equivalently as $R_{\lambda,\mathbf{z}}(F) = \|S_{\mathbf{x}} F - \Upsilon\|_{\mathscr{H}^n}^2 + \lambda \|F\|_{\mathscr{G}}^2$ for all $F \in \mathscr{G}$. Therefore, $R_{\lambda,\mathbf{z}}$ admits a unique minimizer in $\mathscr{G}$ given by the regularized empirical solution $F_{\lambda,\mathbf{z}} = (T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} S_{\mathbf{x}}^* \Upsilon$, which we will consider from now on as the estimate of F_λ.

Uniform convergence rates

As mentioned previously in Remark 4.6.2, uniform convergence rates of $F_{\lambda,\mathbf{z}}$ can only be achieved under additional smoothness assumptions on F_p. Park and Muandet (2020a,b) investigate Tikhonov–Phillips regularization in the well specified case $F_p \in \mathscr{G}$ and show $R(F_{\lambda,\mathbf{z}}) - R(F_p) \in \mathcal{O}_p(n^{-1/4})$ for the regularization scheme $\lambda(n) \in \mathcal{O}(n^{-1/4})$ as $n \to \infty$ whenever the kernel k is bounded and the underlying data is independent.

Closed form Tikhonov–Phillips operator estimates

We show that for the Tikhonov–Phillips estimate, the adjoint of the regularized analytical operator solution, $A_\lambda^* = \Theta^{-1}(F_\lambda)$, which satisfies

$$A_\lambda^* = \operatorname*{arg\,min}_{A \in S_2(\mathscr{H})} \mathbb{E}[\|\varphi(Y) - A^*\varphi(X)\|_{\mathscr{H}}^2] + \lambda \|A\|_{S_2(\mathscr{H})}^2 ,$$

admits a closed form representation in terms of covariance operators associated with the kernel k. In fact, we prove that A_λ^* has the known form which Song et al. (2009) originally identified as the conditional mean embedding under the previously mentioned restrictive assumptions.

While this result does not come as a surprise at this point, we emphasize that this has not been proven before. Although Grünewälder et al. (2012a) establish a connection between the *empirical* regularized solution $F_{\lambda,\mathbf{z}}$ and a version of the *empirical* conditional mean embedding with a *rescaled regularization parameter*, a population analogue was never derived. A simple asymptotic argument via convergence in the infinite-data limit is hampered by the rescaling of the regularization parameter in this derivation. Interestingly, the population expression of A_λ which we derive here is sometimes taken for granted in the literature (see for example Fukumizu et al. 2013), even if it was never proven in the original work.

Our analysis offers a view on the beautiful duality between the generalized covariance operator T acting on $\mathscr{G}$, composition operators acting on $S_2(\mathscr{H})$ and the kernel covariance operator C_{XX}.

Remark 4.6.4. While our analysis is purely aimed at a theoretical understanding at this point, we expect that the following results will have a practical benefit, as they allow an asymptotic discussion of the spectral properties of the given estimates.

For an operator $B \in \mathfrak{B}(\mathscr{H})$, define the *right-composition operator*

$$\Xi_B \colon S_2(\mathscr{H}) \to S_2(\mathscr{H}), \tag{4.38}$$

$$A \mapsto AB. \tag{4.39}$$

It is easy to see that Ξ_B is a well-defined bounded operator since $S_2(\mathscr{H})$ is an ideal in $\mathfrak{B}(\mathscr{H})$ and we have $\|\Xi_B A\|_{S_2(\mathscr{H})} \leq \|A\|_{S_2(\mathscr{H})} \|B\|$. Furthermore, if B is invertible then Ξ_B is invertible and we have $\Xi_{B^{-1}} = \Xi_B^{-1}$.

The following result describes the connection between $\mathscr{G}$ and C_{XX} in terms of the composition operator $\Xi_{C_{XX}}$. In fact, it shows that $T\colon \mathscr{G} \to \mathscr{G}$ describes exactly the action of $\Xi_{C_{XX}}\colon S_2(\mathscr{H}) \to S_2(\mathscr{H})$ under the isomorphism $\Theta\colon S_2(\mathscr{H}) \to \mathscr{G}$.

$$\begin{array}{ccc} \mathscr{G} & \xrightarrow{\Theta^{-1}} & S_2(\mathscr{H}) \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle \Xi_{C_{XX}}} \\ \mathscr{G} & \xrightarrow[\Theta^{-1}]{} & S_2(\mathscr{H}) \end{array} \qquad \begin{array}{ccc} \mathscr{G} & \xrightarrow{\Theta^{-1}} & S_2(\mathscr{H}) \\ {\scriptstyle T+\lambda\mathrm{Id}_{\mathscr{G}}}\downarrow & & \downarrow{\scriptstyle \Xi_{C_{XX}+\lambda\mathrm{Id}_{\mathscr{H}}}} \\ \mathscr{G} & \xrightarrow[\Theta^{-1}]{} & S_2(\mathscr{H}) \end{array}$$

Figure 4.3.: Correspondence of T and $\Xi_{C_{XX}}$.

Theorem 4.6.5. *Let $F \in \mathscr{G}$ and $A := \Theta^{-1}(F) \in S_2(\mathscr{H})$. Then the diagrams in Figure 4.3 are both commutative diagrams, i.e., we have*

$$\Theta^{-1}(TF) = AC_{XX}$$

as well as

$$\Theta^{-1}[(T + \lambda\mathrm{Id}_{\mathscr{G}})F] = A(C_{XX} + \lambda\mathrm{Id}_{\mathscr{H}}).$$

Proof. Let $F \in \mathscr{G}$ and $A = \Theta^{-1}(F) \in S_2(\mathscr{H})$. We have $F(\cdot) = A\varphi(\cdot)$ by Corollary 4.2.5. From the definition of $\mathscr{G}$, we get

$$\begin{aligned} TF &= \int_E K_x F(x)\mathrm{d}\pi(x) = \int_E K_x[A\varphi(x)]\mathrm{d}\pi(x) \\ &= \int_E A[k(\cdot,x)\varphi(x)]\mathrm{d}\pi(x) = A\int_E k(\cdot,x)\varphi(x)\mathrm{d}\pi(x) \\ &= A\int_E [\varphi(x)\otimes\varphi(x)]\varphi(\cdot)\mathrm{d}\pi(x) = AC_{XX}\varphi(\cdot), \end{aligned}$$

where we use the fact that for every fixed $x' \in E$, the map $x \mapsto k(x',x)\varphi(x)$ is an element of $L^1(E, \mathcal{F}_E, \pi; \mathscr{H})$ due to Assumption 3 and Hölder's inequality. Because of this, the integration and the operator A commute (Diestel and Uhl, 1977, Chapter II.2, Theorem 6). The operator AC_{XX} is Hilbert–Schmidt and $TF = AC_{XX}\varphi(\cdot)$ confirms the operator reproducing property under Θ^{-1} from Corollary 4.2.5, hence we have $\Theta^{-1}(TF) = AC_{XX}$. Using this fact, we obtain

$$(T + \lambda\mathrm{Id}_{\mathscr{G}})F = AC_{XX}\varphi(\cdot) + \lambda A\varphi(\cdot) = A(C_{XX} + \lambda\mathrm{Id}_{\mathscr{H}})\varphi(\cdot),$$

confirming the same relation for the second assertion of the theorem. ∎

Theorem 4.6.5 allows us to easily derive the expression for the Tikhonov–Phillips estimate F_λ under Θ^{-1} in terms of its corresponding operator in $S_2(\mathscr{H})$ in terms of C_{XX} and C_{YX}.

Corollary 4.6.6 (Closed form analytical operator solution)**.** *We have*

$$\Theta^{-1}(F_\lambda) = A_\lambda^* = C_{YX}(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1},$$

i.e., the analytical regularized operator solution can be represented as

$$\Theta^{-1}(F_\lambda)^* = A_\lambda = (C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} C_{XY}. \tag{4.40}$$

Proof. By definition, we have $F_\lambda = g_\lambda(T)\mathcal{I}_\pi^* F_p = (T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1}\mathcal{I}_\pi^* F_p$. We can rearrange

$$\begin{aligned}
\mathcal{I}_\pi^* F_p &= \int_E K(\cdot, x) F_p(x) \mathrm{d}\pi(x) = \int_E k(\cdot, x) \int_E \varphi(y) p(x, \mathrm{d}y) \mathrm{d}\pi(x) \\
&= \iint_{E^2} \varphi(y) \left\langle \varphi(x), \varphi(\cdot) \right\rangle_{\mathscr{H}} p(x, \mathrm{d}y) \mathrm{d}\pi(x) \\
&= \left[\int_E \varphi(Y) \otimes \varphi(X) \mathrm{d}\mathbb{P} \right] \varphi(\cdot) = C_{YX} \varphi(\cdot).
\end{aligned}$$

We have thus shown that $C_{YX} = \Theta^{-1}(\mathcal{I}_\pi^* F_p)$ by the operator reproducing property from Corollary 4.2.5. Theorem 4.6.5 implies that the operator $(T + \lambda \mathrm{Id}_{\mathscr{G}})^{-1}$ acting on $\mathscr{G}$ may be represented under Θ^{-1} as by the right composition operator $\Xi_{(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1}}$ acting on $S_2(\mathscr{H})$, leading to

$$\Theta^{-1}(F_\lambda) = \Xi_{(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1}} C_{YX} = C_{YX}(C_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1}$$

as claimed. ∎

Analogously we obtain a closed form representation for the empirical regularized operator solution $A_{\lambda,\mathbf{z}}$, in terms of the empirical covariance operators

$$\widehat{C}_{XX} := \frac{1}{n} \sum_{i=1}^n \varphi(X_i) \otimes \varphi(X_i) \text{ and } \widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^n \varphi(Y_i) \otimes \varphi(X_i).$$

Theorem 4.6.7 (Closed form empirical operator solution)**.** *We have*

$$\Theta^{-1}(F_{\lambda,\mathbf{z}}) = A_{\lambda,\mathbf{z}}^* = \widehat{C}_{YX}(\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1},$$

i.e., the empirical regularized operator solution can be represented as

$$\Theta^{-1}(F_{\lambda,\mathbf{z}})^* = A_{\lambda,\mathbf{z}} = (\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} \widehat{C}_{XY}. \tag{4.41}$$

Theorem 4.6.7 can be proven by simply replacing T with the sample-based operator $T_{\mathbf{x}}$ in the proof of Theorem 4.6.5, leading to $\Theta^{-1}[(T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})F] = \Theta^{-1}(F)(\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})$ for all $F \in \mathscr{G}$. Furthermore replacing $\mathcal{I}_\pi^*$ with $S_{\mathbf{x}}^*$ in the proof of Corollary 4.6.6 yields $\Theta^{-1}(S_{\mathbf{x}}^* \Upsilon) = \widehat{C}_{YX}$, thereby confirming the claim when applying both results to $A_{\lambda,\mathbf{z}} = \Theta^{-1}(F_{\lambda,\mathbf{z}}) = \Theta^{-1}[(T_{\mathbf{x}} + \lambda \mathrm{Id}_{\mathscr{G}})^{-1} S_{\mathbf{x}}^* \Upsilon]$.

4.7. Practical applications

The derivation of the closed form for the regularized operator solution from the previous section allows to connect our theory to known spectral analysis techniques used in practice. We briefly sketch the connection to these approaches here and refer the reader to the mentioned literature for more details.

Klus et al. (2020) and Mollenhauer et al. (2020b) show that the eigenfunctions and singular functions of the regularized empirical estimate

$$A_{\lambda,\mathbf{z}} = (\widehat{C}_{XX} + \lambda \mathrm{Id}_{\mathscr{H}})^{-1} \widehat{C}_{XY} \tag{4.42}$$

can be computed by solving a matrix eigenproblem. In the case that P is the Markov transition operator as described in Section 3.5, it is furthermore shown by Klus et al. (2020) that this empirical eigenproblem coincides exactly with the regularized eigenproblem given by the well-known kernel-based version of EDMD (Williams et al., 2015b). Hence, the asymptotic viewpoint derived in our analysis ultimately proves that kernel EDMD essentially approximates the operator

$$P : \mathscr{H} \to L^2(\pi)$$

in the infinite-sample limit with a suitable regularization scheme, thereby providing a statistical model for kernel EDMD. A theory of the spectral convergence of kernel EDMD can now be developed by investigating the spectral perturbation under the convergence $\|A_{\lambda,\mathbf{z}} - P\|_{\mathscr{H} \to L^2(\pi)} \to 0$ for an admissible regularization scheme $\lambda(n)$ and $n \to \infty$ with suitable assumptions of the underlying process. In particular, our approximation results from Section 4.5 may be used to show that kernel EDMD may overcome the weak spectral convergence of standard EDMD which was proven by Korda and Mezić (2018). However, we note that this requires the eigenfunction of interest is contained in the RKHS $\mathscr{H}$. The details of this theory are not in the scope of this work and are subject to further research.

Additionally, the singular value decomposition of the estimate of P given by (4.42) coincides with the empirical spectral problem of kernel CCA, which is shown by Klus et al. (2020). As an application, the authors identify coherent sets of random dynamical systems by approximating the singular functions of the corresponding Markov transition operator.

4.8. Related work

This chapter is inspired by the recent development in RKHS-based statistical inference. Although our investigation is targeted at creating a more general mathematical perspective from an approximation viewpoint, we make use of the theoretical tools which were originally developed in statistical learning theory. We therefore highlight the most important work which impacted our analysis.

Over the last years, the theory of RKHS-based inference and the kernel mean embedding spawned a vast variety of methods in various statistical disciplines. In this context, a nonparametric approximation of the conditional mean operation

$$(x, f) \mapsto \mathbb{E}[f(Y) \mid X = x]$$

for functions f in some RKHS $\mathscr{H}$ over E was developed by Song et al. (2009) as a purely linear-algebraic concept under the name conditional mean embedding. This idea has since been used as the theoretical backbone for methods in Bayesian analysis, graphical models, time series analysis, spectral analysis and dimensionality reduction, filtering, reinforcement learning and many more (see for example Muandet et al. 2017 for a non-exhaustive selection of applications). Klebanov et al. (2021) recently noted that that the CME can be interpreted in the framework of *best linear unbiased estimation* (BLUE) in Hilbert spaces.

Although the CME as described by Song et al. (2009) performs successfully in applications, the mathematical assumptions imposed in the original work are typically violated; this has been thoroughly examined by Klebanov et al. (2020). The foundational problems in the theory of the CME led to an investigation of the approximation of RKHS-valued conditional Bochner expectations from a regression perspective. In particular, Grünewälder et al. (2012a) show that the empirical Tikhonov–Phillips solution of a regularized least squares regression problem in a vector-valued reproducing kernel Hilbert space coincides with the empirical estimate derived by Song et al. (2009). Additionally, Grünewälder et al. (2013) propose to use the same estimate for the approximation of linear operators in a very broad sense but do not offer an asymptotic perspective of this idea.

Park and Muandet (2020a) extend the asymptotic regression theory of the CME in the framework of least-squares regression in a *vector-valued reproducing kernel Hilbert space* (vRKHS) and regularization theory (see for example Caponnetto and De Vito 2007). In this context, uniform convergence rates are proven under the assumption that the true CME is contained in the hypothesis space. Klebanov et al. (2021) extend the operator-theoretic interpretation of the CME. In particular, they prove existence of an operator on an RKHS which expresses the conditional mean under the assumption that

the true conditional mean function is a member of a corresponding tensor product space. In fact, our analysis shows that this assumption is equivalent to the assumption under which Park and Muandet (2020a) derive convergence rates.

Comparison to this work

Concluding the overall picture of the aforementioned work: while the regression perspective of the CME (Grünewälder et al., 2012a; Park and Muandet, 2020a) allows to consider asymptotic interpretations and prove convergence results, it has the fundamental drawback that the algebraically interesting operator-theoretic perspective of P is not present. Even more so, the estimation of spectral properties of P (for example in the case of Markov transition operators) is impossible. Conversely, the operator-theoretic formulation of the CME (Song et al., 2009; Klebanov et al., 2020, 2021) lacks an asymptotic perspective and suffers from complex interdependencies of various assumptions (Klebanov et al., 2020), severely impeding a theoretical mathematical analysis. Additionally, the approximation viewpoint in the L^2-operator context has not been investigated yet. We show that this approximation admits a natural perspective in terms of the maximum mean discrepancy between the underlying Markov kernels.

Regarded in the context of the CME, our results can be interpreted as the missing link between the recent work of Klebanov et al. (2021) and Park and Muandet (2020a). In particular, we provide an asymptotic approximation perspective in the operator-theoretic context of conditional expectations. On our way, we moreover improve a surrogate risk bound used by Grünewälder et al. (2012a) and Park and Muandet (2020a) which serves as the theoretical foundation for the regression perspective of the CME. However, our results are formulated in a more general perspective in terms of the numerical approximation of linear operators and can certainly be regarded outside of the context of the previously mentioned work on the CME.

4.9. Summary and outlook

In this chapter, we investigate the approximation-theoretic details of the kernel-based discretization of P. When the domain of P is restricted to an RKHS, we show under which assumptions P can be approximated arbitrarily well by Hilbert–Schmidt operators acting on this RKHS. We connect our theory to the conditional mean embedding and its estimation. In particular, we exploit an isomorphism between these Hilbert–Schmidt operators and a vRKHS and prove that the kernel-based approximation of P is equivalent to the prototypical problem of least squares regression with a vector-valued kernel. Solving this problem with a Tikhonov–Phillips regularization scheme, we derive a closed form solution for both the analytical and the empirical setting. We confirm the empirical

solution to be connected to practical spectral decomposition problems which have been derived independently under much stronger assumptions in the context of the conditional mean embedding.

The results in this sections may be regarded as a first step in the direction of an asymptotic theory of the nonparametric approximation of P and its connection to the conditional mean embedding. However, they also leave us with a wide variety of open questions. We highlight some of these questions below.

1. We show that classical methods to derive uniform convergence rates for vector-valued kernel regression due to Caponnetto and De Vito (2007) do not apply in our scenario, as the involved generalized covariance operator is not compact in typical situations. This fact motivates a unified analysis of kernel-based least squares regression with infinite-dimensional output spaces.

2. We only investigate the special case that the least squares regression is solved via a Tikhonov–Phillips regularization scheme. However, several common alternative regularization techniques such as a spectral cutoff or iterative schemes (i.e., different versions of gradient descent) may be examined in this context.

3. The connection of the nonparametric approximation of P to the conditional mean embedding and the maximum mean discrepancy may allow to derive hypothesis tests based on P such as two-sample tests for Markov processes along the lines of Gretton et al. (2012) and Sejdinovic et al. (2013).

4. For the specific applications which we investigate in the context of Markov transition operators, it is of fundamental importance to understand the connections between the spectral properties of the conditional expectation operator $P : L^2(\nu) \to L^2(\pi)$ and the nonparametric model solution $Pi_\nu : \mathscr{H} \to L^2(\pi)$. It is clear that the inclusion operator i_ν must be injective in order to uniquely reconstruct singular functions of P (or eigenfunctions in the case $\nu = \pi$) from the empirical approximation of Pi_ν. Additional questions which address the existence of singular functions of P in $\operatorname{range}(i_\nu) \subseteq L^2(\nu)$ need to be discussed in a separate examination.

5. Kernel autocovariance operators of stationary processes

This chapter contains passages taken from Mollenhauer et al. (2020a).

In the previous chapter, we have seen that a regularized least squares approximation of the conditional expectation operator P over functions in an RKHS can be expressed in terms of kernel (cross-)covariance operators. An empirical estimate of this approximation is naturally obtained by replacing the involved operators with their empirical counterparts.

In a more general context, kernel covariance operators and kernel cross-covariance operators serve as the theoretical foundation of several spectral analysis and component decomposition techniques including *kernel principal component analysis*, *kernel canonical correlation analysis* and *kernel independent component analysis* (Schölkopf et al., 1998; Akaho, 2001; Bach and Jordan, 2002). Consistency results and the statistical analysis of these methods can therefore be directly based on the convergence of empirical kernel covariance operators (Blanchard et al., 2007; Fukumizu et al., 2007; Rosasco et al., 2010). Moreover, the estimation of kernel covariance operators and their connection to L^p integral operators are fundamental concepts in the formalization of statistical learning (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Rosasco et al., 2010).

A kernel cross-covariance operator which describes the relation between two random variables in a stochastic process is called *kernel autocovariance operator*. In this chapter, we extend convergence results for empirical kernel covariance operators based on independent data (see for example Rosasco et al., 2010) to the estimation of kernel autocovariance operators from subsequent observations of a stationary stochastic process. Our investigation is primarily motivated by the nonparametric approximation of the Markov transition operator P of a stationary Markov process as discussed in the previous chapter.

Our results transfer to a wide variety of statistical models for sequential data which rely on the estimation of kernel autocovariance operators. Popular approaches include state space models and filtering (Song et al., 2009; Fukumizu et al., 2013; Gebhardt et al., 2019), transition models (Sun et al., 2019; Grünewälder et al., 2012b), hypothesis testing

(Besserve et al., 2013; Chwialkowski and Gretton, 2014; Chwialkowski et al., 2014) 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). Although these techniques perform well in practice, the theoretical details of the estimation of kernel autocovariance operators from dependent data seem to be covered only sparsely in the literature.

We investigate the estimation of kernel autocovariance operators under the assumptions of ergodicity and mixing which we already imposed in Section 3.6 in the context of projection methods.

5.1. Overview

We revisit the real-valued RKHS which we introduced in the previous chapter and formally introduce kernel autocovariance operators of a stationary stochastic process in Section 5.2. In Section 5.3, we derive a strong law of large numbers for empirical kernel autocovariance operators based on the ergodicity of the underlying process. We investigate details of the asymptotic behaviour of the estimation error under mixing assumptions in Section 5.4 and derive analogues of classical statistical results such as the central limit theorem and the law of the iterated logarithm. Section 5.5 illustrates finite sample error bounds for mixing processes which are refined for the special case of the Gaussian kernel in Section 5.6. As an application, we show in Section 5.7 how our results can be used to prove concistency of kernel PCA when the underlying data is non-iid. Finally, we discuss related work in Section 5.8.

5.2. Kernel autocovariance operators

In this chapter, we consider a stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in the standard Borel space $(E, \mathcal{F}_E)$. As introduced in Section 4.2.1, let $\mathscr{H}$ denote the real-valued RKHS induced by the psd kernel $k : E \times E \to \mathbb{R}$ with corresponding canonical feature map $\varphi : E \to \mathscr{H}$.

We briefly recall the assumptions about the RKHS $\mathscr{H}$ which he have already discussed in Section 4.2.4. We assume $\mathscr{H}$ to be separable (Assumption 1). Furthermore, the feature map φ is required to satisfy measurability (Assumption 2) and the square integrability condition $\varphi(X_0) \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathscr{H})$ or equivalently $\mathbb{E}[k(X_0, X_0)] < \infty$ (Assumption 3). The remaining assumptions introduced in Section 4.2.4 will not be needed in this chapter.

For some fixed time lag $\eta \in \mathbb{N}$, we define the *kernel autocovariance operator* describing

the connection between the snapshots X_t and $X_{t+\eta}$ as

$$C(\eta) := C_{X_\eta X_0} = \mathbb{E}[\varphi(X_\eta) \otimes \varphi(X_0)] \in S_1(\mathscr{H}).$$

Our goal is to empirically approximate $C(\eta)$ in terms of the natural estimator

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

which requires the observation of $n+\eta$ consecutive time steps of the process $(X_t)_{t\in\mathbb{Z}}$.

5.3. Strong law of large numbers

We now address the strong law of large numbers for the estimator $C_n(\eta)$. 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 consequence of the generalized version of Birkhoff's ergodic theorem by Beck and Schwartz (1957) which we introduced in Theorem 3.6.1.

Corollary 5.3.1 (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 the standard Borel space $(E, \mathcal{F}_E)$. Then*

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

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

Proof. We recall the definition of ergodicity from Section 3.6.1 and the representation of the process $(X_t)_{t\in\mathbb{Z}}$ in terms of the left-shift operator T on the canonical probability space $\Omega = E^{\mathbb{Z}}$. The following argumentation is similar to the proof of Theorem 3.6.6. 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 2 and Assumption 3 , the product feature map

$$\varphi \otimes \varphi : E \times E \to S_2(\mathscr{H})$$

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, 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 assertion follows immediately from the fact that we choose $\varphi \otimes \varphi \circ (X_0, X_\eta)$ as the L^1-observable in Theorem 3.6.1 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 holds $\mathbb{P}$-a.e. in $S_2(\mathscr{H})$. ■

Remark 5.3.2 (Convergence in Schatten norms). Corollary 5.3.1 also yields $\mathbb{P}$-a.e. convergence $C_n(\eta) \to C(\eta)$ in $S_p(\mathscr{H})$ for all $p \geq 2$. 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 3.6.1 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.

5.4. Asymptotic error behavior

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

$$\xi_t := \big(\varphi(X_{t+\eta}) \otimes \varphi(X_t)\big) - C(\eta) \tag{5.1}$$

for $t \in \mathbb{Z}$ and fixed $\eta \in \mathbb{N}$. 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 $C_n(\eta) - C(\eta) = \frac{1}{n}\sum_{t=1}^{n} \xi_t$.

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

$$\begin{aligned}
\|\xi_t\|_{L^\infty(\Omega,\mathcal{F},\mathbb{P};S_2(\mathscr{H}))} &\leq \operatorname*{ess\,sup}_{\omega\in\Omega} \|\varphi(X_{t+\eta}) \otimes \varphi(X_t)\|_{S_2(\mathscr{H})} + \|C(\eta)\|_{S_2(\mathscr{H})} \\
&\leq \sup_{x\in E} \|\varphi(x)\|^2_{\mathscr{H}} + \mathbb{E}\left[\|\varphi(X_{t+\eta}) \otimes \varphi(X_t)\|_{S_2(\mathscr{H})}\right] \\
&\leq 2 \sup_{x\in E} \|\varphi(x)\|^2_{\mathscr{H}} \\
&= 2 \sup_{x\in E} k(x,x) = 2c.
\end{aligned} \tag{5.2}$$

We now recall the properties of the strong mixing coefficients introduced in Section 3.6.2. Lemma 3.6.5 clearly shows that for all $n \in \mathbb{Z}$, we have

$$\alpha((\xi_t)_{t\in\mathbb{Z}}, n) \leq \alpha((X_t)_{t\in\mathbb{Z}}, n - \eta). \tag{5.3}$$

In particular, if the process $(X_t)_{t\in\mathbb{Z}}$ is strongly mixing, then $(\xi_t)_{t\in\mathbb{Z}}$ is strongly mixing with at least the same rate as $(X_t)_{t\in\mathbb{Z}}$. Several properties of the estimation error $C_n(\eta) - C(\eta)$ can be proven by applying results from the asymptotic theory of weakly dependent random variables in Hilbert spaces to the process $(\xi_t)_{t\in\mathbb{Z}}$. 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, let $L(n) := \max(\log n, 1)$ for all $n \in \mathbb{N}$.

Theorem 5.4.1 (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 $\sum_{t=1}^{\infty} \alpha(t-\eta) < \infty$. Furthermore, let $\sup_{x\in E} k(x,x) < \infty$. Then the linear operator $T\colon S_2(\mathscr{H}) \to S_2(\mathscr{H})$ defined by*

$$T := \mathbb{E}[\xi_0 \otimes \xi_0] + \sum_{t=1}^{\infty} \mathbb{E}[\xi_0 \otimes \xi_t] + \sum_{t=1}^{\infty} \mathbb{E}[\xi_t \otimes \xi_0] \tag{5.4}$$

is trace class. Furthermore, there exists a Gaussian measure $\mathcal{N}(0,T)$ on $S_2(\mathscr{H})$ and a sequence of iid $S_2(\mathscr{H})$-valued Gaussian random variables $(Z_t)_{t\in\mathbb{Z}} \sim \mathcal{N}(0,T)$ 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 process $(\xi_t)_{t\in\mathbb{Z}}$ is $\mathbb{P}$-a.e. bounded by (5.2) and has summable mixing coefficients by (5.3). We can directly apply the almost sure invariance principle from Dedecker and Merlèvede (2010, Corollary 1) to $(\xi_t)_{t\in\mathbb{Z}}$, which yields the assertion. Note in particular that the authors emphasize that the almost sure boundedness of $(\xi_t)_{t\in\mathbb{Z}}$ and summability of the mixing coefficients imply that the assumptions of Dedecker and Merlèvede (2010, Corollary 1) are satisfied (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 5.4.2 (Central limit theorem)**.** *Under the assumptions of Theorem 5.4.1, the laws of $\sqrt{n}(C_n(\eta) - C(\eta))$ converge weakly to a Gaussian measure $\mathcal{N}(0,T)$ on $S_2(\mathscr{H})$ with covariance operator T defined by (5.4).*

Proof. By our previous analysis and the argumentation of the proof of Theorem 5.4.1, 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 a rescaled version of the estimation error approximates a compact limiting set almost surely. Additionally, it characterizes this set as the accumulation points of the sequence of estimation

errors and gives a norm bound in $S_2(\mathscr{H})$ depending on the mixing rate. We define the shorthand notation $a_n := \sqrt{2L(L(n))}$ for all $n \in \mathbb{N}$. Let furthermore $\operatorname{acc}(x_n)$ denote the set of all accumulation points of a sequence $(x_n)_{n\in\mathbb{N}}$ in a topological space.

Theorem 5.4.3 (Compact law of the iterated logarithm)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary and α-mixing with coefficients $(\alpha(t))_{t\in\mathbb{Z}}$ such that $\sum_{t=1}^{\infty} \alpha(t-\eta) < \infty$. Furthermore, let $\sup_{x\in E} k(x,x) = c < \infty$. Then there exists a compact, convex and symmetric set $K \subseteq S_2(\mathscr{H})$, such that $\mathbb{P}$-a.e.*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{\sqrt{n}(C_n(\eta) - C(\eta))}{a_n}, K\right) = 0 \tag{5.5}$$

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

$$\operatorname{acc}\left(\frac{\sqrt{n}(C_n(\eta) - C(\eta))}{a_n}\right) = K. \tag{5.6}$$

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

$$\sup_{A\in K} \|A\|_{S_2(\mathscr{H})} = (4c^2 + 32\, c^2 M)^{1/2}. \tag{5.7}$$

Theorem 5.4.3 is proven in Appendix A.1.

Remark 5.4.4 (Assumptions on the mixing rate). The results in this section require that the α-mixing coefficients of $(X_t)_{t\in\mathbb{Z}}$ are summable. We 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 argumentation can be derived for the law of the iterated logarithm (see also Rio, 1995).

5.5. Concentration bounds

In addition to the previous asymptotic results, nonasymptotic statements about the estimation error can be derived by applying concentration bounds for mixing Hilbertian random variables to the process $(\xi_t)_{t\in\mathbb{Z}}$.

Theorem 5.5.1 (Error bound)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary and α-mixing with coefficients $(\alpha(t))_{t\in\mathbb{Z}}$. Let $\sup_{x\in E} k(x,x) = c < \infty$. Then for every $\epsilon > 0$, $\nu \in \mathbb{N}$, $n \geq 2$ as well as*

$q \in \{1, \dots, \lfloor n/2 \rfloor\}$ 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

$$\Gamma \colon S_2(\mathscr{H}) \to S_2(\mathscr{H})$$

defined by

$$\Gamma := \mathbb{E}\left[\Big((\varphi(X_\eta) \otimes \varphi(X_0)) - C(\eta)\Big) \otimes \Big((\varphi(X_\eta) \otimes \varphi(X_0)) - C(\eta)\Big)\right]. \tag{5.8}$$

Proof. As previously noted, the process $(\xi_t)_{t\in\mathbb{Z}}$ defined by (5.1) 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 (5.3). 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. ∎

Remark 5.5.2. Alternatively to the bound by Bosq (2000), we can apply the bound for β-mixing sequences by Rhomari (2002), which is also given in Appendix A.5.3. We choose to apply the former result as presented above, since it allows for a convenient simplification when the kernel k is a Gaussian kernel (see Section 5.6).

The above bound requires an optimal tradeoff between the choices of ν, q, and δ. Note that we have $\sum_{j>\nu} \lambda_j < \infty$ for every $\nu \in \mathbb{N}$, since Γ is trace class. The knowledge of the mixing rate $(\alpha(t))_{t\geq 1}$ and the decay of the eigenvalues $(\lambda_j)_{j\geq 1}$ of Γ allows to drastically simplify the bound and derive $\mathbb{P}$-a.e. convergence rates. Whenever $(\alpha(t))_{t\geq 1}$ and $(\lambda_j)_{j\geq 1}$ decay exponentially, a straightforward application of Bosq (2000, Corollary 2.4) can be used to obtain a sharper bound and $\mathbb{P}$-a.e. convergence rates. We will state this result here for completeness and show how the decay of $(\lambda_j)_{j\geq 1}$ can be precisely bounded for the special case of the Gaussian kernel in the next section.

Theorem 5.5.3 (Bosq, 2000, Corollary 2.4)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary and α-mixing with coefficients $(\alpha(t))_{t\in\mathbb{Z}}$. Let $\sup_{x\in E} k(x,x) = c < \infty$. Additionally, let $(\lambda_j)_{j\geq 1}$ be the nonincreasingly ordered eigenvalues of the covariance operator $\Gamma \colon S_2(\mathscr{H}) \to S_2(\mathscr{H})$ defined by (5.8).*

If there exist constants $r \in (0,1)$ and $a > 0$ such that

$$\alpha(t) \leq a r^t \text{ and } \lambda_j < a r^j$$

for all $t, j \in \mathbb{N}$, then for every $\epsilon > 0$, there exist positive constants k_1 and k_2 only depending on ϵ and the law $\mathcal{L}(X_0, X_\eta)$ such that for all $n \in \mathbb{N}$, we have

$$\mathbb{P}\Big[\|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})} > \epsilon\Big] \leq k_1 \exp(-k_2 n^{1/3}).$$

In addition, we have the convergence rate

$$\|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})} = O\left(\frac{(\log n)^{3/2}}{n^{1/2}}\right) \quad \mathbb{P}\text{-a.e.}$$

5.6. Example: Gaussian kernel

Theorem 5.5.1 and Theorem 5.5.3 show that the two main quantities of interest for a bound of the estimation error $\|C_n(\eta) - C(\eta)\|_{S_2(\mathscr{H})}$ are the mixing rate of $(X_t)_{t\in\mathbb{Z}}$ as well as the covariance of the law $\mathcal{L}(\varphi(X_\eta) \otimes \varphi(X_0))$. The latter is given in terms of the eigenvalues of the covariance operator Γ acting on $S_2(\mathscr{H})$ defined by (5.8).

While mixing rates can be assessed by imposing structural assumptions on the process $(X_t)_{t\in\mathbb{Z}}$, the analysis of the eigenvalues of Γ seems to be more intricate. We will now show that for the case that $\mathscr{H}$ is induced by a Gaussian kernel, decay rates of the eigenvalues of Γ can be obtained.

Theorem 5.6.1. *(Eigenvalue decay of Γ) Let $E \subseteq \mathbb{R}^d$ and $\mathscr{H}$ be the RKHS induced by the Gaussian kernel*

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2_{\mathbb{R}^d}}{2\sigma^2}\right)$$

for some width $\sigma > 0$. Let $(\lambda_j)_{j\geq 1}$ be the nonincreasingly ordered eigenvalues of the covariance operator Γ on $S_2(\mathscr{H})$ defined by (5.8). Then the following decay rates hold.

1. *When E is compact, then $(\lambda_j)_{j\geq 1} = O(\exp(-cj \log j))$ for some constant $c > 0$.*
2. *For arbitrary E, if $\mathcal{L}(X_0, X_\eta)$ is absolutely continuous with respect to the Lebesgue measure on $E \times E$ with joint density $p(x, y) \colon E \times E \to \mathbb{R}$ satisfying*

   $$p(x, y) < B \exp(-\|(x, y)\|^2_{\mathbb{R}^{2d}})$$

   for some constant $B > 0$, then $(\lambda_j)_{j\geq 1} = O(\exp(-cj))$ for some constant $c > 0$.
3. *In any case, without additional assumptions about E and $\mathcal{L}(X_0, X_\eta)$, we have $(\lambda_j)_{j\geq 1} = O(\exp(-cj^{1/(2d)}))$ for some constant $c > 0$.*

Theorem 5.6.1 is proven in Appendix A.1. We can now obtain $\mathbb{P}$-a.e. convergence rates and error bounds for the estimator $C_n(\eta)$ when the mixing properties of $(X_t)_{t\in\mathbb{Z}}$ are known. In combination with results for mixing rates of typical classes of processes (see Example 3.6.4), statements like the following are the immediate consequence.

Example 5.6.2 (Markov process on compact domain in $\mathbb{R}^d$). Let $E \subseteq \mathbb{R}^d$ be compact and $\mathscr{H}$ be the RKHS induced by the Gaussian kernel on E for some width $\sigma > 0$. If $(X_t)_{t\in\mathbb{Z}}$ is a stationary, geometrically ergodic Markov process on E, then the conclusions of Theorem 5.5.3 hold.

5.7. Application: statistical consistency of kernel PCA with dependent data

We now use our preceding results to prove the consistency of kernel PCA when the underlying data is dependent. To this end, we adopt the modern functional-analytic formalism to describe the statistical model of kernel PCA in Hilbert spaces (see for example Mas and Menneteau, 2003; Zwald and Blanchard, 2006; Blanchard et al., 2007; Mas and Ruymgaart, 2015; Koltchinskii and Lounici, 2016, 2017; Reiß and Wahl, 2020; Jirak and Wahl, 2020) and apply basic results from spectral perturbation theory. We do not aim to provide a full overview of the related work here. Instead, we will highlight how our work leads to some elementary statistical results for kernel PCA with dependent data. To the best of our knowledge, a detailed analysis of this setting is not yet available in the literature.

5.7.1. Minimizing the reconstruction error

We introduce the functional-analytic interpretation of kernel PCA and incorporate the scenario of dependent data. For simplicity, we may assume that the stationary embedded process $(\varphi(X_t))_{t\in\mathbb{Z}}$ is centered, i.e.,

$$\mathbb{E}[\varphi(X_0)] = 0,$$

which is commonly required in the theoretical analysis of kernel PCA and replaced by an empirical centering in practice (see for example Blanchard et al., 2007).

Given some fixed integer $0 < r \leq \dim \mathscr{H}$, an orthonormal system $F := \{f_1, \dots, f_r\} \subset \mathscr{H}$ and the corresponding orthogonal projection $\Pi_F : \mathscr{H} \to \mathscr{H}$ onto $\operatorname{span} F \subset \mathscr{H}$, we define the *reconstruction error* of the embedded random variable $\varphi(X_0)$ as

$$R(\Pi_F) := \mathbb{E}\left[\|\varphi(X_0) - \Pi_F\varphi(X_0)\|_{\mathscr{H}}^2\right]. \tag{5.9}$$

Our goal is now to choose the orthonormal set F and the corresponding orthogonal projection Π_F in such a way that the resulting reconstruction error $R(\Pi_F)$ is minimal over the set of all orthogonal projections on $\mathscr{H}$ of rank r. The main idea of kernel PCA is to then express the r-dimensional projection $\Pi_F\varphi(X_t)$ in terms of the Euclidean process of coordinate vectors given by

$$\tilde{X}_t := \begin{bmatrix} \langle f_1, \varphi(X_t)\rangle_{\mathscr{H}} \\ \vdots \\ \langle f_r, \varphi(X_t)\rangle_{\mathscr{H}} \end{bmatrix} = \begin{bmatrix} f_1(X_t) \\ \vdots \\ f_r(X_t) \end{bmatrix} \in \mathbb{R}^r.$$

The projected process $\tilde{X}_t$ can be seen as a low-dimensional approximation of $\varphi(X_t)$ with respect to (5.9).

We now consider the zero-lag kernel autocovariance operator of $(X_t)_{t\in\mathbb{Z}}$ given by

$$C := C(0) = \mathbb{E}[\varphi(X_0) \otimes \varphi(X_0)].$$

The reconstruction error $R(\Pi_F)$ can be conveniently expressed in an alternative way in terms of C.

Lemma 5.7.1 (Reconstruction error)**.** *For every r-dimensional orthogonal projection $\Pi_F : \mathscr{H} \to \mathscr{H}$, we have*

$$R(\Pi_F) = \mathrm{Tr}(C(\mathrm{Id}_{\mathscr{H}} - \Pi_F)).$$

Proof. Let $\Pi_F : \mathscr{H} \to \mathscr{H}$ be an orthogonal projection with corresponding orthonormal system $F = \{f_1, \dots, f_r\} \subset \mathscr{H}$. We extend F to a complete orthonormal system $F' := \{f_1, \dots, f_r, \dots\}$ of $\mathscr{H}$. We have

$$\begin{aligned} \mathrm{Tr}(C(\mathrm{Id}_{\mathscr{H}} - \Pi_F)) &= \sum_{f_j \in F'} \langle Cf_j, (\mathrm{Id}_{\mathscr{H}} - \Pi_F) f_j\rangle_{\mathscr{H}} = \sum_{j>r} \mathbb{E}[f_j(X_0)\, f_j(X_0)] \\ &= \sum_{j>r} \mathbb{E}\left[\langle f_j, \varphi(X_0)\rangle_{\mathscr{H}}^2\right] = \mathbb{E}\left[\|(\mathrm{Id}_{\mathscr{H}} - \Pi_F)\varphi(X_0)\|_{\mathscr{H}}^2\right] = R(\Pi_F), \end{aligned}$$

where we make use of $\langle Cf, h\rangle_{\mathscr{H}} = \mathbb{E}[f(X)h(X)]$ for all $f, h \in \mathscr{H}$ as well as the reproducing property in $\mathscr{H}$ and Parseval's identity. ∎

Let $C = \sum_{i\in I} \mu_i(C)\, v_i \otimes v_i$ denote the eigendecomposition of the zero-lag kernel autocovariance operator. We now assume $r \leq \mathrm{rank}(C)$. Let $V := \{v_1, \dots, v_r\}$ denote the set of the first r eigenvectors. Since we have

$$R(\Pi_F) = \mathrm{Tr}(C(\mathrm{Id}_{\mathscr{H}} - \Pi_F)) = \mathrm{Tr}(C) - \mathrm{Tr}(C\Pi_F)$$

by Lemma 5.7.1, a projection operator Π_F minimizing $R(\Pi_F)$ must maximize $\mathrm{Tr}(C\Pi_F)$.

Gohberg and Kreĭn (1969, Chapter II, Lemma 4.1) show that the maximum of $\operatorname{Tr}(C\Pi_F)$ is attained for the projection operator onto the span of the first r eigenfunctions of C. Consequently, we have

$$\arg\min_{\Pi_F} R(\Pi_F) = \Pi_V, \tag{5.10}$$

where the minimum on the left-hand side ranges over all orthogonal projection operators on $\mathscr{H}$ of rank r. Furthermore, from (5.10) we can clearly deduce $R(\Pi_V) = \sum_{i>r} \mu_i(C)$.

The above examination motivates to define the *empirical reconstruction error*

$$R_n(\Pi_F) := \sum_{t=1}^{n} \|\varphi(X_t) - \Pi_F \varphi(X_t)\|_{\mathscr{H}}^2$$

based on n subsequent observations $X_1, \ldots, X_n$ of the process $(X_t)_{t\in\mathbb{Z}}$. Analogously to Lemma 5.7.1, we can write

$$R_n(\Pi_F) = \operatorname{Tr}(C_n(\mathrm{Id}_{\mathscr{H}} - \Pi_F)),$$

where $C_n = \sum_{t=1}^{n} \varphi(X_t) \otimes \varphi(X_t)$ is the empirical zero-lag autocovariance operator with the empirical spectral decomposition

$$C_n = \sum_{i\in\widehat{I}} \mu_i(C_n)\, \widehat{v}_i \otimes \widehat{v}_i.$$

Again assuming $r \leq \operatorname{rank}(C_n)$, the *empirical kernel PCA solution* minimizing R_n is naturally given by the projection $\Pi_{\widehat{V}}$ onto the span of the first r empirical eigenvectors $\widehat{V} := \{\widehat{v}_1, \ldots, \widehat{v}_r\}$.

5.7.2. Convergence of the excess reconstruction error

The consistency of kernel PCA can be investigated by focusing on the convergence of the nonnegative *excess reconstruction error* $R(\Pi_{\widehat{V}}) - R(\Pi_V) \to 0$ with high probability as the number of samples n increases. We highlight a very simple approach relying on the eigenspace perturbation based on the Davis–Kahan theorem (see Appendix A.4). We write $\Delta = C - C_n$ for the estimation error of the kernel covariance operator.

Theorem 5.7.2 (Excess reconstruction error)**.** *Let* $r < \min\{\operatorname{rank}(C), \operatorname{rank}(C_n)\}$ *and* $\mu_r(C) \neq \mu_{r+1}(C)$*. Then we have*

$$R(\Pi_{\widehat{V}}) - R(\Pi_V) \leq 2^{3/2} \|C\|_{S_2(\mathscr{H})} \frac{\|\Delta\|_{S_2(\mathscr{H})}}{\mu_r(C) - \mu_{r+1}(C)}. \tag{5.11}$$

Proof. By Lemma 5.7.1, we have

$$R(\Pi_{\widehat{V}}) - R(\Pi_V) = \operatorname{Tr}(C(\Pi_{\widehat{V}} - \Pi_V)) = \left\langle C, \Pi_{\widehat{V}} - \Pi_V \right\rangle_{S_2(\mathscr{H})}$$

$$\leq \|C\|_{S_2(\mathscr{H})} \left\|\Pi_{\widehat{V}} - \Pi_V\right\|_{S_2(\mathscr{H})}.$$

We can now simply apply Lemma A.4.3 and Theorem A.4.4 to the term $\left\|\Pi_{\widehat{V}} - \Pi_V\right\|_{S_2(\mathscr{H})}$, yielding the assertion. ■

In combination with the convergence results for the estimation error Δ which we derive in the previous sections of this chapter, the bound from Theorem 5.7.2 can be used to obtain convergence and overall probability bounds for $R(\Pi_{\widehat{V}}) - R(\Pi_V)$ when the data is weakly dependent. In particular, the above result shows that the rate of convergence of the excess reconstruction error of kernel PCA is at least the rate of convergence of $\|\Delta\|_{S_2(\mathscr{H})}$. A very simple example of this theory is highlighted below. For more sophisticated error analyses and stronger results in the case of independent sampling, the reader may refer to the recent work by Reiß and Wahl (2020), Jirak and Wahl (2020) and Milbradt and Wahl (2020) and the references therein.

Example 5.7.3 (Geometrically ergodic Markov process on $E \subseteq \mathbb{R}^d$). Let $(X_t)_{t\in\mathbb{Z}}$ be the Markov process on the compact domain $E \subseteq \mathbb{R}^d$ with the properties as given in Example 5.6.2. Then the excess reconstruction error of kernel PCA with the Gaussian kernel with some arbitrary width $\sigma > 0$ converges as

$$R(\Pi_{\widehat{V}}) - R(\Pi_V) = O\left(\frac{(\log n)^{3/2}}{n^{1/2}}\right) \quad \mathbb{P}\text{-a.e.}$$

5.8. Related work

The theory of weakly dependent random processes taking values in infinite-dimensional 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; Hsing and Eubank, 2015). 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. The importance of autocovariance operators of stationary processes in the L^2 space context is underlined by the concepts of *spectral density operators* (Panaretos and Tavakoli, 2013b,a) as well as *dynamic functional principal components analysis* (Hörmann et al., 2015).

In contrast to the previously mentioned work on general processes taking values in Banach spaces and Hilbert spaces, we consider autocovariance operators of the embedded Hilbert space process

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

explicitly incorporating properties of the RKHS into our analysis. This scenario directly falls in line with the classic setting in learning theory, which has led to celebrated results and numerous applications in case of independent and identically distributed data. 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 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 covariance operator estimation *without a time lag*. To the best of our knowledge, this is the first time that RKHS covariance operator estimation is addressed in the context of weakly dependent data. 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 which are more commonly found in the literature (see Doukhan 1994, Bradley 2005 and Rio 2017 and the references therein for an overview). Additionally, our results cover the estimation of autocovariance operators with a time lag.

We note that convergence in measure of 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. To the best of our knowledge, results for kernel PCA with dependent data have not explicitly been stated in the literature yet.

5.9. Summary and outlook

In this chapter, we investigate the estimation of kernel autocovariance operators from a realization of a stationary stochastic process under the assumptions of ergodicity and mixing. We prove several convergence statements by applying standard results from probability theory and statistics with weakly dependent Hilbertian random variables. Our analysis is primarily motivated by the fact that the nonparametric estimation of

Markov transition operators requires the estimation of kernel covariance operators as discussed in the preceding chapter.

Although our results are generally not sufficient to prove the overall convergence of the Tikhonov–Phillips estimate of the Markov transition operator, they cover a substantial part of this problem. In particular, they can be used to show convergence of the regularized empirical estimate to the regularized analytic solution for a *fixed regularization parameter*. Overall convergence and optimal regularization schemes need to be derived in the general setting of inverse problems, which we briefly introduce in Appendix A.6.

Apart from the estimation of Markov transition operators, we demonstrate the versatility of our results by providing new consistency results for kernel PCA with weakly dependent data. Similarly, our results may be used further to examine theoretical properties of kernel-based hypothesis tests for time series which rely on covariance operators (Besserve et al., 2013; Chwialkowski and Gretton, 2014; Chwialkowski et al., 2014).

Bibliography

S. Akaho. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*. Springer-Verlag, 2001.

N. I. Akhiezer and I. M. Glazman. *Theory of linear operators in Hilbert space*. Dover Publications, Inc., 1993.

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.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, number 3 in Proceedings of Machine Learning Research, pages 1247–1255, Atlanta, Georgia, USA, 2013.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

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.

J. R. Baxter and J. S. Rosenthal. Rates of convergence for everywhere-positive Markov chains. *Statistics & Probability Letters*, 22(4):333–338, 1995.

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

M. Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1348–1361. PMLR, 2018.

A. Ben-Isreal and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, 2003.

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

M. Besserve, N. K. Logothetis, and B. Schölkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2535–2543. Curran Associates, Inc., 2013.

R. Bhatia. *Matrix analysis*. Springer Verlag, New York, 1997.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18:971–1013, 2018.

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.

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.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391): 580–598, 1985.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042, 2012.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 04(04):377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, 2010.

L. Cavalier. Inverse problems with non-compact operators. *Journal of Statistical Planning and Inference*, 136(2):390–400, 2006.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1422–1430, Bejing, China, 2014. PMLR.

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3608–3616. Curran Associates, Inc., 2014.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

E. B. Davies. Metastable states of symmetric Markov semigroups I. *Proceedings of the London Mathematical Society*, 3(1):133–150, 1982a.

E. B. Davies. Metastable states of symmetric Markov semigroups II. *Journal of the London Mathematical Society*, 2(3):541–556, 1982b.

E. B. Davies. Spectral properties of metastable Markov semigroups. *Journal of Functional Analysis*, 52(3):315–329, 1983.

C. Davis and W. M. Kahan. The Rotation of Eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

E. De Vito and A. Caponnetto. Risk bounds for the regularized least-squares algorithm with operator-valued kernels. Technical report, Massachusetts Institute of Technology, Cambridge, May 2005. CBCL Paper #249/AI Memo #2005-015.

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 W. Philipp. Almost sure invariance principles for weakly dependent vector-valued random variables. *The Annals of Probability*, 10(3):689–701, 08 1982.

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.

M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM Journal on Numerical Analysis*, 36(2):491–515, 1999.

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

J. Ding and T.-Y. Li. Markov finite approximation of the Frobenius–Perron operator. *Nonlinear Analysis: Theory, Methods & Applications*, 17:759–772, 1991.

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.

N. Dunford and J. T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988a.

N. Dunford and J. T. Schwartz. *Linear Operators. Part II.* Wiley Classics Library. John Wiley & Sons, Inc., New York, 1988b.

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.

J. Fan, B. Jiang, and Q. Sun. Hoeffding's inequality for general Markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139): 1–35, 2021.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

G. Froyland. An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. *Physica D: Nonlinear Phenomena*, 250:1–19, 2013.

G. Froyland, N. Santitissadeekorn, and A. Monahan. Transport in time-dependent dynamical systems: Finite-time coherent sets. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):043116, 2010.

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.

S. Gerber and I. Horenko. Toward a direct and scalable identification of reduced models for categorical processes. *Proceedings of the National Academy of Sciences*, 114(19): 4863–4868, 2017.

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.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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.

S. Grünewälder, G. Arthur, and J. Shawe-Taylor. Smooth operators. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1184–1192, Atlanta, Georgia, USA, 2013. PMLR.

W. Hackbusch. *Integral Equations: Theory and Numerical Treatment.* Birkhäuser, 1995.

P. R. Halmos and V. S. Sunder. *Bounded integral operators on L^2 spaces*, volume 96 of *Ergebnisse der Mathematik und ihrer Grenzgebiete.* Springer-Verlag, Berlin-Heidelberg, 1978.

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.

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

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* Wiley Series in Probability and Statistics. Wiley, 2015.

W. Huisinga. *Metastability of Markovian systems: A transfer operator based approach in application to molecular dynamics.* PhD thesis, Freie Universität Berlin, 2001.

W. Huisinga and B. Schmidt. Metastability and dominant eigenvalues of transfer operators. In *New algorithms for macromolecular simulation*, pages 167–182. Springer, 2006.

W. Huisinga, S. Meyn, and C. Schütte. Phase transitions and metastability in Markovian and molecular systems. *Annals of Applied Probability*, 14(1):419–458, 02 2004.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Network*, 13(4-5):411–430, 2000.

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.

B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

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. Junge and P. Koltai. Discretization of the Frobenius–Perron operator using a sparse Haar tensor basis: The Sparse Ulam method. *SIAM Journal on Numerical Analysis*, 47:3464–3485, 2009.

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

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.

A. Kessy, A. Lewin, and K. Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

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.

I. Klebanov, B. Sprungk, and T. J. Sullivan. The linear conditional expectation in Hilbert space. *Bernoulli*, 27(4):2267–2299, 2021.

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, 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, 2018.

S. Klus, B. E. Husic, M. Mollenhauer, and F. Noé. Kernel methods for detecting coherent structures in dynamical data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):123112, 2019.

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.

P. Koltai, H. Wu, F. Noé, and C. Schütte. Optimal data-driven estimation of generalized markov state models for non-equilibrium dynamics. *Computation*, 6(1):22, 2018.

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.

I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Annals of Applied Probability*, 13(1):304–362, 2003.

I. Kontoyiannis and S. P. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electronic Journal of Probability*, 10:61–123, 2005. doi: 10.1214/EJP.v10-231.

I. Kontoyiannis and S. P. Meyn. Approximating a diffusion by a finite-state hidden Markov model. *Stochastic Processes and their Applications*, 127(8):2482–2507, 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.

C. Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications*, 118(2):232 – 260, 2008.

H. Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29(3):719–736, 1958.

A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, volume 97 of *Applied Mathematical Sciences*. Springer, 2nd edition, 1994.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, 1991.

S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 55(3):725–740, 1993.

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.

T.-Y. Li. Finite approximation for the Frobenius–Perron operator. A solution to Ulam's conjecture. *J. Approx. Theory*, 17:177–186, 1976.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

S. Lu, P. Mathé, and S. V. Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis*, 48(1): 123–148, 2020.

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.

T. Melzer, M. Reiter, and H. Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *International Conference on Artificial Neural Networks*, pages 353–360. Springer, 2001.

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.

T. Michaeli, W. Wang, and K. Livescu. Nonparametric canonical correlation analysis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1967–1976, New York, New York, USA, 20–22 Jun 2016. PMLR.

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.

M. Mollenhauer, S. Klus, C. Schütte, and P. Koltai. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020a.

M. Mollenhauer, I. Schuster, S. Klus, and C. Schütte. Singular value decomposition of operators on reproducing kernel Hilbert spaces. In O. Junge, O. Schütze, G. Froyland, S. Ober-Blobaum, and K. Padberg-Gehle, editors, *Advances on Dynamics, Optimization and Computation. Series: Studies in Systems, Decision and Control*, volume 304, pages 109–131. Springer, 2020b.

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.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

F. Noé and F. Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation*, 11(2):635–655, 2013.

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, 2013a.

V. M. Panaretos and S. Tavakoli. Fourier analysis of stationary time series in function space. *The Annals of Statistics*, 41(2):568–603, 04 2013b.

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., 2020a.

J. Park and K. Muandet. Regularised least-squares regression with infinite-dimensional output space. *arXiv preprint arXiv:2010.10973*, 2020b.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:32 pp., 2015.

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.

D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM*, 9(1):84–97, 1962.

I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. *Probability in Banach Spaces: Proceedings of the Eighth International Conference*, 8:128–134, 01 1992.

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

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

A. Rastogi and S. Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3, 2017. ISSN 2297–4687.

A. Rastogi, G. Blanchard, and P. Mathé. Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems. *Electronic Journal of Statistics*, 14(2):2798–2841, 2020.

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.

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

N. Rhomari. On Bernstein type and maximal inequalities for dependent Banach-valued random vectors and applications. In F. Ferraty and Y. Romain, editors, *Dependence in Probability, Analysis and Number Theory*. Oxford University Press, 2011.

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.

G. Roberts, J. Rosenthal, et al. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

G. O. Roberts and R. L. Tweedie. Geometric L2 and L1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, pages 37–41, 2001.

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

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

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.

I. Schuster, M. Mollenhauer, S. Klus, and K. Muandet. Kernel conditional density operators. In S. Chiappa and R. Calandra, editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 993–1004. PMLR, 2020.

C. Schütte. Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules, 1999. Habilitation Thesis.

C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. Number 24 in Courant Lecture Notes. American Mathematical Society, 2013.

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. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41 (5):2263–2291, 2013.

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

C.-J. Simon-Gabriel, A. Barp, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

S. Smale and D.-X. Zhou. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285 – 302, 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.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In Y. W. Teh and M. Titterington, editors, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9, pages 773–780, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

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.

G. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

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.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

W. Tian and H. Wu. Kernel embedding based variational approach for low-dimensional approximation of dynamical systems. *arXiv preprint arXiv:2008.02962*, 2020.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, 1977.

J. A. Tropp. An Introduction to Matrix Concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

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.

R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

J. Weidmann. *Linear Operators in Hilbert Spaces*. Springer, 1980.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109:278–295, 1963.

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.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press, 2006.

A. Appendix

Section A.1 contains passages taken from Mollenhauer et al. (2020a).
Section A.3 contains passages taken from Mollenhauer and Koltai (2020).

A.1. Proofs

We provide proofs which are omitted in the main text.

Proof of Theorem 3.4.2. We first recall some basic facts from matrix analysis. For strictly positive symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, we have the identity

$$\mathbf{A}^{-1/2} - \mathbf{B}^{-1/2} = \mathbf{A}^{-1/2}(\mathbf{B}^{1/2} - \mathbf{A}^{1/2})\mathbf{B}^{-1/2} \tag{A.1}$$

as well as the bound

$$\left\|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\right\| \leq \left\|\mathbf{A} - \mathbf{B}\right\|^{1/2}, \tag{A.2}$$

which is a special case of Bhatia (1997, Theorem X.1.1.). This yields

$$\left\|\widehat{\mathbf{G}}^{-1/2} - \mathbf{G}^{-1/2}\right\| \leq \left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|^{1/2} \left\|\widehat{\mathbf{G}}^{-1/2}\right\| \left\|\mathbf{G}^{-1/2}\right\| = \sqrt{\frac{\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|}{\mu_s(\widehat{\mathbf{G}})\mu_s(\mathbf{G})}} \quad \mathbb{P}\text{-a.e.},$$

hence the assertion follows from Lemma 3.4.1. ■

Proof of Theorem 3.4.3. Note that we have the bound

$$\left\|(\mathbf{A} + \lambda\mathbf{I})^{-1/2}\right\| = \frac{1}{\sqrt{\mu_s(\mathbf{A} + \lambda\mathbf{I})}} = \frac{1}{\sqrt{\mu_s(\mathbf{A}) + \lambda}} \leq \frac{1}{\sqrt{\lambda}} \tag{A.3}$$

for any symmetric positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{s\times s}$.

All bounds involving $\widehat{\mathbf{G}}$ are to be understood in the $\mathbb{P}$-a.e. sense. We decompose the overall error as

$$\left\|(\widehat{\mathbf{G}} + \lambda\mathbf{I})^{-1/2} - \mathbf{G}^{-1/2}\right\| \leq \underbrace{\left\|(\widehat{\mathbf{G}} + \lambda\mathbf{I})^{-1/2} - (\mathbf{G} + \lambda\mathbf{I})^{-1/2}\right\|}_{(I)} + \underbrace{\left\|(\mathbf{G} + \lambda\mathbf{I})^{-1/2} - \mathbf{G}^{-1/2}\right\|}_{(II)}$$

and address (I) and (II) individually by applying (A.1), (A.2) and (A.3) in this order to both terms. For (I), we have

$$\begin{aligned}(I) &\leq \left\|(\widehat{\mathbf{G}} + \lambda \mathbf{I})^{1/2} - (\mathbf{G} + \lambda \mathbf{I})^{1/2}\right\| \left\|(\widehat{\mathbf{G}} + \lambda \mathbf{I})^{-1/2}\right\| \left\|(\mathbf{G} + \lambda \mathbf{I})^{-1/2}\right\| \\ &\leq \frac{1}{\lambda} \left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|^{1/2}.\end{aligned}$$

Analogously for (II), we obtain

$$\begin{aligned}(II) &\leq \left\|(\mathbf{G} + \lambda \mathbf{I})^{1/2} - \mathbf{G}^{1/2}\right\| \left\|(\mathbf{G} + \lambda \mathbf{I})^{-1/2}\right\| \left\|\mathbf{G}^{-1/2}\right\| \\ &\leq \frac{\|\lambda \mathbf{I}\|^{1/2}}{\sqrt{\mu_s(\mathbf{G}) + \lambda}\,\sqrt{\mu_s(\mathbf{G})}} \leq \frac{\sqrt{\lambda}}{\mu_s(\mathbf{G})}.\end{aligned}$$

In total, whenever $\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\| \leq \epsilon$ holds for some $\epsilon > 0$, then we have

$$\left\|(\widehat{\mathbf{G}} + \lambda \mathbf{I})^{-1/2} - \mathbf{G}^{-1/2}\right\| \leq \frac{\sqrt{\epsilon}}{\lambda} + \frac{\sqrt{\lambda}}{\mu_s(\mathbf{G})}$$

and hence the assertion follows from Lemma 3.4.1. ■

Proof of Theorem 5.4.3. We apply Merlèvede (2008, Theorem 2) to the process $(\xi)_{t\in\mathbb{Z}}$ defined in (5.1) and immediately obtain the existence of a compact set K with the desired properties such that both (5.5) and (5.6) 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 (5.7) for K. The set K is the unit ball of the Hilbert space $\mathbb{H}$, which is given by the completion of the range of $T^{1/2}$ (where T is given by (5.4) and $T^{1/2}$ denotes its operator square root) with respect to the inner product defined by

$$\left\langle T^{1/2}A,\, T^{1/2}B \right\rangle_{\mathbb{H}} := \langle A,\, B\rangle_{S_2(\mathscr{H})}, \quad A, B \in S_2(\mathscr{H}). \tag{A.4}$$

The space $\mathbb{H}$ is also called *Cameron–Martin space* or *abstract Wiener space* (for details, we refer the reader to Bogachev, 1998, Chapter 2). For a technical construction of $\mathbb{H}$ and the limit set K 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 iid case. However, for the construction of $\mathbb{H}$ and K only an abstract limiting probability measure is needed, which is given by the Gaussian measure obtained from Theorem 5.4.2 and its covariance operator T defined by (5.4), just as shown in the proof of Merlèvede (2008, Theorem 2). We can therefore analyze properties of K by considering the Cameron–Martin space of the centered Gaussian measure induced by T, which is examined in the previously mentioned literature. The identity (A.4) 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 (A.4), we obtain

$$\|A\|_{S_2(\mathscr{H})} \leq \left\|T^{1/2}\right\| \|A\|_{\mathbb{H}}, \quad A \in \mathbb{H}. \tag{A.5}$$

Since $K = \{A \in \mathbb{H} \mid \|A\|_{\mathbb{H}} \leq 1\}$, a bound for $\left\|T^{1/2}\right\| = \|T\|^{1/2}$ 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 $T = \mathbb{E}[\xi_0 \otimes \xi_0] + \sum_{t=1}^{\infty} \mathbb{E}[\xi_0 \otimes \xi_t] + \sum_{t=1}^{\infty} \mathbb{E}[\xi_t \otimes \xi_0]$. We clearly have

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

since ξ_0 is almost surely bounded by $2c$ by (5.2).

Let $\alpha(n)$ be the mixing coefficients of $(X_t)_{t\in\mathbb{Z}}$. We now note that by (3.14), 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 (5.2)). 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

$$\|T\| \leq 4c^2 + 32\, c^2 M, \tag{A.6}$$

which proves the claim in combination with (A.5). ■

Proof of Theorem 5.6.1. The key idea for this proof is to make use of the fact that the product RKHS $S_2(\mathscr{H}) \simeq \mathscr{H} \otimes \mathscr{H}$ is isometrically isomorphic to a Gaussian RKHS. This

allows to interpret Γ as a (centered) convolution operator in order to apply classical results from the theory of integral equations.

Let $\mathscr{G}$ denote the RKHS induced by the Gaussian kernel $\ell\colon E^2 \times E^2 \to \mathbb{R}$,

$$\ell(z, z') = \exp\left(\frac{-\|z - z'\|_{\mathbb{R}^{2d}}^2}{2\sigma^2}\right),$$

where $\|\cdot\|_{\mathbb{R}^{2d}}$ is the Euclidean norm on E^2. Let $\psi\colon E^2 \to \mathscr{G}$ be the feature map corresponding to $\mathscr{G}$. Then the pointwise defined map

$$\begin{aligned}\nu\colon \mathscr{H} \otimes \mathscr{H} &\to \mathscr{G}\\ \varphi(x) \otimes \varphi(y) &\mapsto \psi((x, y))\end{aligned}$$

is an isometry, which can be seen by expressing the respective inner products in terms of the corresponding kernels and using the fact that $k(x, x')k(y, y') = \ell((x, y), (x', y'))$. Extending ν to linear combinations gives a bijective isometry from the dense subset $\operatorname{span}\{\varphi(x)\otimes\varphi(x') \mid x, x' \in E\}$ of $\mathscr{H}\otimes\mathscr{H}$ to the dense subset $\operatorname{span}\{\psi((x, x')) \mid (x, x') \in E^2\}$ of $\mathscr{G}$. Finally, extending ν continuously to the respective completions yields an isometric isomorphism from $\mathscr{H} \otimes \mathscr{H}$ to $\mathscr{G}$.

We now decompose $\Gamma = \Gamma_1 - \Gamma_2$, where $\Gamma_1 := \mathbb{E}\left[(\varphi(X_\eta) \otimes \varphi(X_0)) \otimes (\varphi(X_\eta) \otimes \varphi(X_0))\right]$ and $\Gamma_2 := C(\eta) \otimes C(\eta)$. Since Γ_2 is a rank-one operator, Γ has the same asymptotic eigenvalue behavior as Γ_1 (Gohberg and Kreĭn, 1969, Chapter II, Corollary 2.1). It is therefore sufficient to only consider the eigenvalue decay of Γ_1. The isomorphism ν constructed above allows to interpret Γ_1 as an integral operator with respect to the Gaussian kernel ℓ, which makes the application of classical results from the theory of integral equations possible. For every operator $A \in S_2(\mathscr{H}) \simeq \mathscr{H} \otimes \mathscr{H}$ and all $x, x' \in E$, we write $A(x, x') = \langle A,\, \varphi(x) \otimes \varphi(x')\rangle_{\mathscr{H}\otimes\mathscr{H}}$ where we identify A with its representation in $\mathscr{H} \otimes \mathscr{H}$ via its singular decomposition. We get the representation

$$\begin{aligned}(\Gamma_1 A)(y, y') &= \left\langle \Gamma_1 A,\, \varphi(y) \otimes \varphi(y')\right\rangle_{\mathscr{H}\otimes\mathscr{H}}\\ &= \int \left\langle \varphi(X_\eta) \otimes \varphi(X_0),\, \varphi(y) \otimes \varphi(y')\right\rangle_{\mathscr{H}\otimes\mathscr{H}} A(X_\eta, X_0)\,\mathrm{d}\mathbb{P}\\ &= \int k(X_\eta, y)k(X_0, y')\, A(X_\eta, X_0)\,\mathrm{d}\mathbb{P}\\ &= \int \ell((X_\eta, X_0), (y, y'))\, A(X_\eta, X_0)\,\mathrm{d}\mathbb{P}\\ &= \int_E \ell(z, z')\, A(z)\,\mathrm{d}\mathcal{L}(X_\eta, X_0)(z)\end{aligned}$$

for all $y, y' \in E$, where $z := (x, x')$ and $z' := (y, y')$. We can therefore consider the eigenvalue problem

$$(\Gamma_1 A)(z') = \int_E \ell(z, z')\, A(z)\,\mathrm{d}\mathcal{L}(X_\eta, X_0)(z) = \lambda A(z'), \tag{A.7}$$

where A is interpreted as a real-valued function on E^2. The solution of integral equations of the form (A.7) for $A \in L^2(E^2, \mathcal{F}_E^{\otimes 2}, \mathcal{L}(X_\eta, X_0); \mathbb{R})$ is well examined. Let $(\lambda_j)_{j>0}$ denote the eigenvalues of Γ_1. When E is a compact domain, the eigenvalues have a super exponential decay of the form $O(\exp(-cj \log j))$ for some constant $c > 0$. When no assumptions about the domain E are made, exponential decay of the Lebesgue density of $\mathcal{L}(X_\eta, X_0)$ on E^2 leads to an exponential eigenvalue decay in terms of $O(\exp(-cj))$, which is a special case of results by Widom (1963) (see for example Bach and Jordan, 2002, Appendix C.2 for the specific cases considered in this context). Without any additional assumptions about the domain or the underlying distribution, a nearly exponential decay of eigenvalues of the form $O(\exp(-cj^{1/2d})$ is always guaranteed (Belkin, 2018, Theorem 5).

Note that when we interpret (A.7) as an operator on product RKHS functions in $\mathscr{H} \otimes \mathscr{H}$ instead of $L^2(E^2, \mathcal{F}_E^{\otimes 2}, \mathcal{L}(X_\eta, X_0); \mathbb{R})$, the resulting operator Γ_1 has the same eigenvalues as its L^2-analogue (Rosasco et al., 2010, Proposition 8), which proves all assertions of the theorem. ■

A.2. Functional analysis

We collect basic results from functional analysis.

A.2.1. Projection operators

Theorem A.2.1. *Let B and B' be Banach spaces and $K : B \to B'$ be a compact operator. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of bounded operators on B' converging to an operator $A : B' \to B'$ in strong operator topology. Then $A_n K \to AK$ in operator norm topology.*

A proof based on the uniform boundedness principle is given by Hackbusch (1995, Lemma 4.3.7. and Lemma 4.3.8.) for the case that K operates on a single space B and can easily be extended to the scenario of two distinct spaces B, B' given above.

Corollary A.2.2. *Let H and H' be Hilbert spaces and $K : H \to H'$ be a compact operator. Let furthermore $(\Pi_n)_{n\in\mathbb{N}}$ and $(\Pi'_n)_{n\in\mathbb{N}}$ be sequences of orthogonal projection operators on H and H' converging strongly to the identities on H and H' respectively. Then we have*

$$\left\| K - \Pi'_n K \Pi_n \right\| \to 0 \quad \text{as } n \to \infty.$$

Proof. Since $(\Pi_n)_{n\in\mathbb{N}}$ and $(\Pi'_n)_{n\in\mathbb{N}}$ converge strongly to the identity operators, the sequences $(\Pi_n^\perp)_{n\in\mathbb{N}}$ and $(\Pi'^\perp_n)_{n\in\mathbb{N}}$ of the complementary projection operators given by $\Pi_n^\perp := \mathrm{Id}_H - \Pi_n$ and $\Pi'^\perp_n := \mathrm{Id}_{H'} - \Pi'_n$ converge strongly to the zero operator on H and

H'. We have

$$\begin{aligned}\left\|K - \Pi'_n K \Pi_n\right\| &\leq \left\|K - \Pi'_n K\right\| + \left\|\Pi'_n K - \Pi'_n K \Pi_n\right\| \\ &= \left\|\Pi_n^{\perp} K\right\| + \left\|\Pi'_n K \Pi_n^{\perp}\right\| \leq \left\|\Pi_n^{\perp} K\right\| + \left\|K \Pi_n^{\perp}\right\| \\ &= \left\|\Pi_n^{\perp} K\right\| + \left\|\Pi_n^{\perp} K^*\right\|\end{aligned}$$

and hence the assertion follows from Theorem A.2.1. ■

A.2.2. Low-rank approximation

For every compact Hilbert space operator $A\colon H \to H'$ with SVD

$$A = \sum_{i \in I} \rho_i(A)\, u_i \otimes v_i,$$

we define the *rank-r truncation* of A as

$$A_r := \sum_{i=1}^{r} \rho_i(A)\, u_i \otimes v_i.$$

for every $r < \operatorname{rank}(A)$. The following well-known result characterizes the truncated operator A_r as the optimal approximation to A with rank constraint r.

Theorem A.2.3 (Eckart–Young–Mirsky theorem)**.** *Let $A\colon H \to H'$ be a compact operator and $r < \operatorname{rank}(A)$. Then we have*

$$\operatorname*{arg\,min}_{\substack{B \in \mathfrak{B}(H,H') \\ \operatorname{rank}(B) = r}} \|A - B\| = A_r$$

with $\|A - A_r\| = \rho_{r+1}(A)$. If additionally $A \in S_2(H, H')$, then we have analogously

$$\operatorname*{arg\,min}_{\substack{B \in \mathfrak{B}(H,H') \\ \operatorname{rank}(B) = r}} \|A - B\|_{S_2(H,H')} = A_r$$

with $\|A - A_r\|^2_{S_2(H,H')} = \sum_{i>r} \rho_i^2(A)$.

A proof for the approximation in operator norm is given by Gohberg and Kreĭn (1969, Chapter II, Theorem 2.1), the Hilbert–Schmidt case is shown by Hsing and Eubank (2015, Theorem 4.4.7.).

A.3. Statistical learning theory

It is well-known that for $F \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$, the standard least squares risk

$$R(F) := \mathbb{E}\left[\|\varphi(Y) - F(X)\|_{\mathscr{H}}^2\right],$$

can be rewritten in terms of the *regression function*

$$F_p(x) := \int_E \varphi(y)\, p(x, \mathrm{d}y) = \mathbb{E}[\varphi(Y) \mid X = x] \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H}).$$

We report the proof here for completeness (see also Cucker and Smale, 2002, Proposition 1 for a proof in the scalar case).

Theorem A.3.1 (Risk and regression function)**.** *Under Assumptions 1–3, the risk R can equivalently be rewritten as*

$$R(F) = \|F - F_p\|_{L^2(E, \mathcal{F}_E, \pi; \mathscr{H})}^2 + R(F_p) \tag{A.8}$$

for all $F \in L^2(E, \mathcal{F}_E, \pi; \mathscr{H})$.

Proof. We have

$$\begin{aligned} R(F) &= \mathbb{E}\left[\|\varphi(Y) - F(X)\|_{\mathscr{H}}^2\right] \\ &= \mathbb{E}\left[\|\varphi(Y) - F_p(X) + F_p(X) - F(X)\|_{\mathscr{H}}^2\right] \\ &= \mathbb{E}\left[\|\varphi(Y) - F_p(X)\|_{\mathscr{H}}^2\right] \\ &\quad + 2\mathbb{E}\left[\langle \varphi(Y) - F_p(X),\, F_p(X) - F(X)\rangle_{\mathscr{H}}\right] \\ &\quad + \mathbb{E}\left[\|F(X) - F_p(X)\|_{\mathscr{H}}^2\right], \end{aligned}$$

where we see that the first summand equals to $R(F_p)$. The second summand which contains the mixed terms vanishes since we have

$$\begin{aligned} &\mathbb{E}\left[\langle \varphi(Y) - F_p(X),\, F_p(X) - F(X)\rangle_{\mathscr{H}}\right] \\ &= \int_E \Big\langle \underbrace{\int_E \varphi(y)\, p(x, \mathrm{d}y)}_{=F_p(x)} - F_p(x),\, F_p(x) - F(x) \Big\rangle_{\mathscr{H}} \mathrm{d}\pi(x). \end{aligned}$$

The last summand can be rewritten as

$$\mathbb{E}\left[\|F(X) - F_p(X)\|_{\mathscr{H}}^2\right] = \|F - F_p\|_{L^2(E, \mathcal{F}_E, \pi; \mathscr{H})}^2$$

by change of measure, proving the assertion. ∎

A.4. Spectral subspace perturbation

The basic perturbation of the spectrum of compact operators can for example be described in terms of *Weyl's inequalities*, which are widely known in the context of matrix analysis (Bhatia, 1997, Section III.2). We state an infinite-dimensional version concerning the singular values of operators between two real Hilbert spaces.

Theorem A.4.1 (Gohberg and Kreĭn 1969, Chapter II, Corollary 2.3)**.** *Let $A : H \to H'$ and $\widehat{A} : H \to H'$ be compact operators between Hilbert spaces H and H'. Then we have*

$$\sup_{i \in \mathbb{N}} \left| \rho_i(A) - \rho_i(\widehat{A}) \right| \leq \left\| A - \widehat{A} \right\|,$$

where we set $\rho_i(A) := 0$ whenever $i > \operatorname{rank}(A)$ and analogously $\rho_i(\widehat{A}) := 0$ whenever $i > \operatorname{rank}(\widehat{A})$.

However, we are more interested in the perturbation of the subspaces associated with eigenvalues and singular values. In particular, we aim to understand the perturbation of the eigenspaces or singular components associated with the r dominant eigenvalues or singular values in a uniform fashion. Following the seminal work of Davis and Kahan (1970), we now give a very basic overview of the perturbation of finite-dimensional spectral subspaces of linear operators. At the heart of the approach lies the idea to measure the distance between two subspaces in terms of the so-called *canonical angles*. We refer the reader to (Stewart and Sun, 1990, Sections I.5 and II.4) and Bhatia (1997, Sections V.II.1 and V.II.3) for modern expositions of this theory.

We consider two orthonormal systems $U = \{u_1, \ldots, u_r\}$ and $\widehat{U} = \{\widehat{u}_1, \ldots, \widehat{u}_r\}$ in a real Hilbert space H. We define the r *canonical angles* between the subspaces $\operatorname{span} U$ and $\operatorname{span} \widehat{U}$ as the inverse cosines of the top r singular values of the operator $\Pi_U \Pi_{\widehat{U}}$ given by

$$\arccos \rho_1(\Pi_U \Pi_{\widehat{U}}), \ldots, \arccos \rho_r(\Pi_U \Pi_{\widehat{U}}).$$

Note that equivalently, the canonical angles are sometimes defined as the inverse sines of the singular values of the operator $\Pi_U^{\perp} \Pi_{\widehat{U}}$, (Bhatia, 1997, pp. 201–202). We define $\Theta(U, \widehat{U}) \in \mathbb{R}^{r \times r}$ as the matrix which contains the canonical angles on the diagonal. Additionally, let $\sin \Theta(U, \widehat{U}) \in \mathbb{R}^{r \times r}$ denote the matrix which contains the sines of the canonical angles on the diagonal.

Remark A.4.2 (Infinite dimensions). Although the typical theory of canonical angles is formulated for finite-dimensional Hilbert spaces H in terms of matrices, it can be extended straightforwardly to infinite-dimensional ambient spaces as long as U and $\widehat{U}$ are finite. The ambient dimension of H is not explicitly present in the results. To apply the finite-dimensional theory in infinite dimensions, it is therefore sufficient to just restrict

the setting to a finite-dimensional subspace of H that contains both $\operatorname{span} U$ and $\operatorname{span} \widehat{U}$. We directly formulate the results in an infinite-dimensional setting here.

The norms of the matrix $\sin \Theta(U, \widehat{U})$ containing the sines of the canonical angles can be written in terms of distances between the corresponding projectors. These distances again allow for a geometrical interpretation as the *gap* (Stewart and Sun, 1990) or *aperture* (Akhiezer and Glazman, 1993) between the subspaces.

Lemma A.4.3 (Distances between projectors)**.** *With the above notation, we have*

$$\left\|\sin \Theta(U, \widehat{U})\right\| = \left\|\Pi_{\widehat{U}} - \Pi_U\right\|$$

as well as

$$\left\|\sin \Theta(U, \widehat{U})\right\|_F = \frac{1}{\sqrt{2}} \left\|\Pi_{\widehat{U}} - \Pi_U\right\|_{S_2(H)},$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. Both assertions follow from Stewart and Sun (1990, Theorem I.5.5.) and the fact that $\left\|\Pi_{\widehat{U}} - \Pi_U\right\| = \rho_1(\Pi_{\widehat{U}} - \Pi_U)$ and $\left\|\Pi_{\widehat{U}} - \Pi_U\right\|_{S_2(H)}^2 = \sum_{i \in I} \rho_i(\Pi_{\widehat{U}} - \Pi_U)^2$. ∎

The importance of the canonical angles is reflected in the so-called *Davis–Kahan theorem* (Davis and Kahan, 1970), which gives a bound for the perturbation of the canonical angles of the eigenspaces of self-adjoint matrices. We present an infinite-dimensional extension of Yu et al. (2015, Theorem 2), which simplifies the original results by Davis and Kahan.

Theorem A.4.4 (Eigenspace perturbation)**.** *Let $A : H \to H$ and $\widehat{A} : H \to H$ be compact positive self-adjoint operators with nonincreasingly ordered eigenvalues $(\mu_i(A))_{i \in I}$ and $(\mu_i(\widehat{A}))_{i \in \widehat{I}}$. Let furthermore $r < \min\{\operatorname{rank}(A), \operatorname{rank}(\widehat{A})\}$ and U and $\widehat{U}$ be the sets containing the first r eigenvectors of A and $\widehat{A}$. Assume $\mu_r(A) \neq \mu_{r+1}(A)$. Then we have*

$$\left\|\sin \Theta(U, \widehat{U})\right\|_F \leq \frac{2 \min\left\{\sqrt{r} \left\|A - \widehat{A}\right\|, \left\|A - \widehat{A}\right\|_{S_2(H)}\right\}}{\mu_r(A) - \mu_{r+1}(A)}.$$

Proof. Let W and $\widehat{W}$ be the sets of first $r+1$ eigenvectors of A and $\widehat{A}$, respectively. We define the operators $A_0 := \Pi_{W \cup \widehat{W}} A \Pi_{W \cup \widehat{W}}$ and $\widehat{A}_0 := \Pi_{W \cup \widehat{W}} \widehat{A} \Pi_{W \cup \widehat{W}}$. By construction, the first $r+1$ eigenvalues and eigenvectors of A_0 and $\widehat{A}_0$ coincide with the first $r+1$ eigenvalues and eigenvectors of A and $\widehat{A}$, respectively. By choosing some orthonormal basis of the finite-dimensional space $\operatorname{span}(W \cup \widehat{W})$ and expressing A_0 and $\widehat{A}_0$ in terms of matrices, we can apply Yu et al. (2015, Theorem 2) to A_0 and $\widehat{A}_0$ and obtain the assertion. ∎

The above theorem can be directly transferred to the singular value decomposition by applying the proof provided by Yu et al. (2015, Theorem 3).

Corollary A.4.5 (Singular subspace perturbation)**.** *Let $A : H \to H'$ and $\widehat{A} : H \to H'$ be compact operators. Let furthermore $r < \min\{\operatorname{rank}(A), \operatorname{rank}(\widehat{A})\}$ and U and $\widehat{U}$ be the sets containing left singular functions corresponding to the top r dominant singular values of A and $\widehat{A}$, respectively. Assume $\rho_r(A) \neq \rho_{r+1}(A)$. Then we have*

$$\left\|\sin\Theta(U,\widehat{U})\right\|_F \leq \frac{2\left(2\rho_1(A) + \left\|A-\widehat{A}\right\|\right)\min\left\{\sqrt{r}\left\|A-\widehat{A}\right\|, \left\|A-\widehat{A}\right\|_{S_2(H)}\right\}}{\rho_r(A)^2 - \rho_{r+1}(A)^2}.$$

The above bound also holds when U and $\widehat{U}$ are replaced with the sets of top r right singular functions V and $\widehat{V}$ of A and $\widehat{A}$, respectively.

A.5. Concentration bounds in Hilbert spaces

We present concentration inequalities for both independent and weakly dependent vector-valued random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A.5.1. Generalized Hoeffding bound

We review a well-known vector-valued version of Hoeffding's inequality in Hilbert spaces due to Pinelis (1992, 1994). We call a sequence of random variables $(S_k)_{k\in\mathbb{N}}$ a *martingale sequence*, if $\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_1^n] = S_n$ for all $n \in \mathbb{N}$, where $\mathcal{F}_1^n$ is the σ-field induced by $(S_1, \ldots, S_n)$.

Theorem A.5.1 (Pinelis 1994, Theorem 3.5)**.** *Let $(S_k)_{k\in\mathbb{N}}$ be a martingale sequence with values in a separable real Hilbert space H such that $\sum_{k=1}^{\infty} \operatorname{ess\,sup} \|S_k - S_{k-1}\|_H^2 \leq C^2$ for some constant $C^2 > 0$. Then for every $\epsilon > 0$, we have*

$$\mathbb{P}\left[\sup_{k\in\mathbb{N}} \|S_k\|_H \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2C^2}\right).$$

The following vector-valued generalization of Hoeffding's inequality is an immediate consequence.

Corollary A.5.2 (Hoeffding's inequality in Hilbert spaces)**.** *Let $\xi_1, \ldots, \xi_n$ be independent random variables in a separable Hilbert space H such that $\mathbb{P}$-a.e. $\|\xi_i\|_H \leq M$ and $\mathbb{E}[\xi_i] = 0$ for all $1 \leq i \leq n$. Then for all $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\|_H \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2M^2}\right)$$

Proof. For a fixed $n \in \mathbb{N}$, define the martingale sequence $(S_k)_{k \in N}$ by

$$(S_k) := \mathbb{1}_{(1 \leq k \leq n)} \left(\sum_{i=1}^{k} \xi_i \right), \quad 0 \leq k.$$

Then S_k clearly satisfies

$$\sum_{k=1}^{\infty} \operatorname{ess\,sup} \|S_k - S_{k-1}\|_H^2 = \sum_{k=1}^{n} \operatorname{ess\,sup} \|\xi_k\|_H^2 \leq nM^2.$$

For every $\epsilon > 0$, applying Theorem A.5.1 to $(S_k)_{k \in \mathbb{N}}$ yields

$$\mathbb{P}\left[\left\| \sum_{i=1}^{n} \xi_i \right\|_H \geq \epsilon \right] \leq 2 \exp\left(-\frac{\epsilon^2}{2nM^2} \right),$$

which implies the assertion by multiplying ϵ with n. ∎

A.5.2. Generalized Bernstein bound

When the higher moments of the random variables $\xi_1, \ldots, \xi_n$ can be bounded, results like the following version of Bernstein's inequality due to Pinelis and Sakhanenko (1986, Corollary 1) can be proven and may lead to sharper inequalities.

Theorem A.5.3 (Bernstein's inequality in Hilbert spaces)**.** *Let $\xi_1, \ldots, \xi_n$ be independent random variables in a separable real Hilbert space H such that $\mathbb{E}[\xi_i] = 0$ for all $1 \leq i \leq n$ and there exist constants $\sigma^2 > 0$ and $L > 0$ such that*

$$\sum_{i=1}^{n} \mathbb{E}[\|\xi_i\|_H^p] \leq \frac{p!}{2} \sigma^2 L^{p-2} \text{ for all } p \geq 2. \tag{A.9}$$

Then for all $\epsilon > 0$, we have

$$\mathbb{P}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_H \geq \epsilon \right] \leq 2 \exp\left(-\frac{n^2 \epsilon^2}{2(\sigma^2 + nL\epsilon)} \right).$$

Note that whenever we have $\|\xi_i\|_H \leq L$ $\mathbb{P}$-a.e. and $\mathbb{E}[\|\xi_i\|_H^2] \leq \tilde{\sigma_i}^2$ for all $1 \leq i \leq n$, then

$$\sum_{i=1}^{n} \mathbb{E}[\|\xi_i\|_H^p] \leq \sum_{i=1}^{n} \mathbb{E}[\|\xi_i\|_H^2] L^{p-2} \leq n\tilde{\sigma}^2 L^{p-2}$$

such that (A.9) holds with $\sigma^2 := n\tilde{\sigma}^2$.

A.5.3. Generalized Bernstein bound for weakly dependent random vectors

We now briefly review a version of a tail bound for weakly dependent Hilbertian random variables due to Rhomari (2002).[1] This result can be understood as a generalized Bernstein bound for β-mixing sequences and incorporates variance proxies of the involved random variables in terms of second moments of their respective norms. The structure of these bounds is comparable to the bounds provided by Bosq (2000, Theorem 2.12 and Corollary 2.4) for α-mixing sequences which we apply in Section 5.5 and Section 5.6.

Both types of bounds can be used to derive convergence rates for corresponding strong laws of large numbers under additional assumptions on the mixing rates (Bosq 2000, Corollary 2.4; Rhomari 2011, Corollary 14.4).

We now state the vector-valued Bernstein bound for β-mixing sequences as given by (Rhomari 2002, Théorème 3.1). Note that the original reference states the result below for nonstationary sequences. We state it directly for stationary case and incorporate the resulting simplifications.

Theorem A.5.4 (Generalized Bernstein bound for β-mixing sequences)**.** *Let $(\xi_t)_{t\in\mathbb{Z}}$ be a stationary stochastic process taking values in a separable real Hilbert space H such that $\mathbb{E}[\xi_0] = 0$ with β-mixing coefficients $\beta(n)$. Fix some $n \in \mathbb{N}$. Suppose there exist constants $\sigma^2 > 0$ and $M > 0$ such that for $1 \leq l \leq \lfloor n/2 \rfloor$, we have*

$$\|\xi_1 + \cdots + \xi_l\|_H \leq lM \quad \mathbb{P}\text{-a.e.}, \tag{A.10}$$

$$\mathbb{E}[\|\xi_1 + \cdots + \xi_l\|_H^2] \leq l\sigma^2. \tag{A.11}$$

Then for all $\epsilon > 0$, the bound

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{t=1}^{n}\xi_t\right\|_H \geq \epsilon\right] \leq 4\exp\left(-\frac{n\epsilon^2}{4(1+2l/n)\sigma^2 + 4lM\epsilon/3}\right) + \left(\frac{n}{l}+2\right)\beta(l)$$

holds.

Remark A.5.5 (Simplification for bounded random variables). The conditions of Theorem A.5.4 are fulfilled if the ξ_t are almost surely bounded. In particular, if $\|\xi_0\|_H < M$ $\mathbb{P}$-a.e. then (A.10) holds trivially and (A.11) holds with

$$\sigma^2 \leq M^2\left(1 + 5\sum_{i=1}^{l-1}\beta(i)\right), \tag{A.12}$$

which is a slightly stronger version of the simplification provided (but not proven) by Rhomari (2002, 2011). We now prove (A.12). Let $\mathcal{F}_t \subseteq \mathcal{F}$ be the σ-field induced by

[1] see also Rhomari (2011) for related results in English

the random variable ξ_t. We recall the definition of the strong mixing coefficients from Section 3.6.2. Then by using the covariance inequality for α-mixing almost surely bounded sequences by Dehling and Philipp (1982, Lemma 2.2), we have

$$\begin{aligned}
\mathbb{E}[\|\xi_1 + \cdots + \xi_l\|_H^2] &= \mathbb{E}\big[\big\langle \sum_{i=1}^{l} \xi_i, \sum_{j=1}^{l} \xi_j \big\rangle_H\big] = \mathbb{E}\big[\sum_{i,j=1}^{l} \langle \xi_i, \xi_j \rangle_H\big] \\
&\leq \sum_{i=1}^{l} \mathbb{E}\big[\|\xi_i\|_H^2\big] + \sum_{i \neq j}^{l} \Big|\mathbb{E}\big[\langle \xi_i, \xi_j \rangle_H\big]\Big| \\
&\leq lM^2 + \sum_{i \neq j}^{l} 10M^2 \alpha(\mathcal{F}_i, \mathcal{F}_j) \\
&\leq lM^2 + l \sum_{i=1}^{l-1} 10M^2 \alpha(i) \\
&= lM^2 \left(1 + 10 \sum_{i=1}^{l-1} \alpha(i)\right).
\end{aligned}$$

The hence the assertion (A.12) follows from $\alpha(i) \leq \frac{1}{2}\beta(i)$ for all i.

A.6. Inverse problems and regularization

We give a concise overview of the most important concepts of ill-posed linear inverse problems in Hilbert spaces and their regularization. For details, we refer the reader to Engl et al. (1996).

In what follows, we consider two real Hilbert spaces H, F and a bounded operator $A : H \to F$. We are interested in finding a solution $h \in H$ to the operator equation

$$Ah = f \tag{A.13}$$

for a given right-hand side $f \in F$. The operator A is often called *forward operator* and f is often called *data* in this context. Problem (A.13) is called *well-posed*, if

(i) for all $f \in F$, a unique solution $h \in H$ exists and

(ii) the solution depends continuously on the data.

If these two conditions do not hold, then we call (A.13) *ill-posed*. Note that a necessary condition for (i) is range$(A) = F$, which already fails to hold in case A is compact and F is infinite-dimensional.

We now generalize the notion of a *solution* to problem (A.13) in order to deal with ill-posed problems in a meaningful way.

Definition A.6.1. *An element $u \in H$ is called* least squares *solution of* (A.13)*, if*

$$\|Au - f\|_F^2 = \inf_{h \in H} \|Ah - f\|_F^2 .$$

And element $u \in H$ is called best approximate solution *of* (A.13)*, if*

$$\|u\|_H = \inf\{\|h\|_H \mid h \in H \text{ is a least squares solution of (A.13)}\}.$$

Thus, the best approximate solution is a least squares solution of minimal norm. We characterize the existence of these solutions through the *Moore–Penrose pseudoinverse* of A. We first observe that the restriction of A to the orthogonal complement of its nullspace is always bijective: that is,

$$A\big|_{\ker(A)^\perp} : \ker(A)^\perp \to \operatorname{range}(A)$$

admits a bounded inverse. We now make use of this fact to define the Moore–Penrose pseudoinverse. Additionally, for two subsets $S \subseteq F$ and $S' \subseteq F$, we introduce the shorthand notation

$$S + S' := \{f + f' \mid f \in S, f' \in S'\} \subseteq F.$$

Definition A.6.2 (Pseudoinverse)**.** *Consider the (not necessarily closed) set*

$$M := \operatorname{range}(A) + \operatorname{range}(A)^\perp \subseteq F.$$

We call the operator $A^\dagger : M \to H$ given by

$$A^\dagger f := \begin{cases} 0 & f \in \operatorname{range}(A)^\perp, \\ (A\big|_{\ker(A)^\perp})^{-1} f & f \in \operatorname{range}(A) \end{cases}$$

the Moore–Penrose pseudoinverse *or simply* pseudoinverse *of A.*

Remark A.6.3. While $A^\dagger$ can always be defined on the domain M, it is generally not a bounded operator and is furthermore not globally defined as an operator from F to H. Note that $\operatorname{range}(A)$ is typically not a closed subspace of F, so in this case we have

$$M = \operatorname{range}(A) + \operatorname{range}(A)^\perp \neq F.$$

It can be shown that $A^\dagger$ is bounded if and only if $\operatorname{range}(A)$ is closed (Engl et al., 1996, Proposition 2.4).

For a more detailed analysis of the properties of generalized inverse operators, we refer the reader to Ben-Isreal and Greville (2003). We now describe the connection between $A^\dagger$ and the best approximate solutions of (A.13).

Theorem A.6.4 (Engl et al. 1996, Theorems 2.5 & 2.6)**.** *The problem* (A.13) *has a least squares solution if and only if* $f \in M = \mathrm{dom}(A^\dagger) \subset F$*. In this case, the best approximate solution is unique and is given in terms of*

$$u^\dagger := A^\dagger f \in H.$$

Equivalently, $u^\dagger$ *can be characterized as the best approximate solution to the so-called* normal equation

$$A^*Ah = A^*f, \quad h \in H$$

and therefore admits the representation

$$u^\dagger = (A^*A)^\dagger A^* f,$$

i.e., we have $A^\dagger = (A^*A)^\dagger A^*$*.*

In practical applications, we face the situation that we only have access to perturbed versions of the forward operator A or the data f (for example through discretization or empirical estimation). So even if the analytical best approximate solution $u^\dagger = A^\dagger f$ exists, we can not be certain that the perturbed problem can be solved and is stable, since $A^\dagger$ is generally not continuous. We therefore regularize the problem by introducing globally defined operators which approximate $A^\dagger$.

Definition A.6.5 (Regularization)**.** *A parametrized family of globally defined bounded operators*

$$\{g_\lambda(A) : F \to H \,|\, \lambda \in (0, \infty]\}$$

is called a regularization strategy *for* A*, if for every* $f \in M = \mathrm{dom}(A^\dagger)$*, there exists a sequence* $(\lambda_n)_{n\in N} \in (0, \infty]$*, such that*

$$g_{\lambda_n}(A)f \to u^\dagger = A^\dagger f, \quad n \to \infty.$$

In particular, we focus on the special case of Tikhonov–Phillips regularization (Phillips, 1962; Tikhonov and Arsenin, 1977) in this work.

Theorem A.6.6 (Tikhonov–Phillips regularization)**.** *For every* $\lambda > 0$*, the operator*

$$A^*A + \lambda \mathrm{Id}_H : H \to H$$

is bijective and therefore admits a bounded inverse. The family of operators

$$g_\lambda(A) := (A^*A + \lambda \mathrm{Id}_H)^{-1} A^*, \quad \lambda > 0$$

is a regularization strategy and for every $f \in M = \mathrm{dom}(A^\dagger)$*, we have*

$$u_\lambda := g_\lambda(A)f \to u^\dagger$$

as $\lambda \to 0$. *Furthermore, the regularized solution* $u_\lambda \in H$ *is the unique minimizer of the* Tikhonov error functional

$$h \mapsto \|Ah - f\|_F^2 + \lambda \|h\|_H^2, \quad h \in H.$$

For proofs of the above results, may consult Engl et al. (1996, Theorems 4.1 & 5.1). When $f \in M = \mathrm{dom}(A^\dagger)$, a natural question is to ask how fast the regularization error $\left\|u_\lambda - u^\dagger\right\|_H$ converges to 0 depending on the rate of $\lambda \to 0$. It is known that in general, the convergence rate can be arbitrarily slow and may be assessed under smoothness assumptions about the data f (Schock, 1984).

In practice, a perturbed problem given by an approximation $A_{\delta_1} : H \to F$ of the forward operator and an approximation $f_{\delta_2} \in F$ of the data is usually considered instead of problem (A.13). That is, we solve

$$A_{\delta_1} h = f_{\delta_2}$$

under the assumption that $\|A_{\delta_1} - A\| < \delta_1$ and $\|f_{\delta_2} - f\|_F \leq \delta_2$ hold as approximation errors of the forward operator and the data. One of the central goals of regularization theory is to assess the rate of convergence of the perturbed regularized solution

$$u_{\lambda,\delta_1,\delta_2} := g_\lambda(A_{\delta_1}) f_{\delta_2} \in H$$

to the true best approximate solution $u^\dagger$ depending on the regularization strategy g_λ and a suitable regularization parameter scheme $\lambda = \lambda(\delta_1, \delta_2)$ under the assumption that $\delta_1 \to 0$ and $\delta_2 \to 0$.