Freie Universität Berlin
Department of Mathematics
and Computer Science

Master thesis

Singular Value Decomposition of Operators on Reproducing Kernel Hilbert Spaces

Mattes Mollenhauer

June 7, 2018
Berlin

First supervisor:
Prof. Dr. Christof Schütte
Second supervisor:
Dr. Ingmar Schuster

Contents

1. Introduction

The *singular value decomposition* (SVD) is one of the most prominent mathematical results which transitioned from linear algebra and matrix analysis into a vast variety of numerical methods in computational science. It is used for instance in numerical solvers for systems of linear equations and optimization problems (Golub and Van Loan 2013; Boyd and Vandenberghe 2009), signal and image processing (Sadek 2012) and in a variety of related methods in statistics and machine learning such as *principal component analysis* (PCA; Pearson 1901; Hotelling 1933), *canonic correlation analysis* (CCA; Hotelling 1936), *latent semantic analysis* (LSA; Deerwester et al. 1990) and the estimation of *hidden Markov models* (Hsu, Kakade and Zhang 2012).

The SVD of a matrix can be used to obtain a majority of the characterising properties of the associated linear map such as range, nullspace, condition and different operator norms. Furthermore, the SVD can be used to optimally approximate matrices under rank constraints, solve least square problems or to directly compute the *Moore–Penrose pseudoinverse* (Mirsky 1960; Golub, Hoffman and Stewart 1987; Penrose 1955). Although the matrix SVD naturally extends to compact operators on Hilbert spaces (see for example Reed and Simon 1980; Rudin 1991), this infinite-dimensional generalization is not as multifaceted as the finite-dimensional case in terms of numerical applications. This is mainly due to the complicated numerical representation of infinite-dimensional operators and the resulting problems concerning the computation of their SVD. As a remedy, one usually considers finite-rank operators based on finite-dimensional subspaces given by a set of fixed basis elements. The SVD of compact operators, particularly of finite-rank operators will be the main focus of this thesis.

We will combine the theory of the SVD of finite-rank operators with the concept of the *reproducing kernel Hilbert space* (RKHS), a special class of Hilbert space allowing for a high-dimensional representation of the abstract mathematical notion of "data" in a feature space. A big part of the theory of RKHSs was originally developed in a functional analytic setting (Aronszajn 1950) and made its way into pattern recognition (Aizerman, Braverman and Rozoner 1964). RKHSs gained popularity in statistics (Berlinet and Thomas-Agnan 2004) and machine learning. Examples of the most prominent kernel-based methods are *support vector machines* (SVMs; Boser, Guyon and Vapnik 1992; Cortes and Vapnik 1995) and *kernel PCA* (Schölkopf, Smola and Müller 1998). RKHSs are mainly used in the aforementioned fields to implicitly give a nonlinear extension of linear methods by embedding observations into a high-dimensional RKHS and rewriting the method in terms of the inner product of the RKHS. This strategy is known as the *kernel trick* (Boser, Guyon and Vapnik 1992) and has been successfully applied to a variety of techniques and algorithms (see Schölkopf and Smola 2002 for an overview).

The approach of embedding a countable number of observations into the RKHS can be generalized to the embedding of probability distributions associated with random variables on the observation space into the RKHS (Smola et al. 2007). As a result, the theory of so-called *kernel mean embeddings* (see Muandet et al. 2017 for a comprehensive review) spawned more abstract probabilistic approaches to problems in statistics and machine learning. Recent advancements show that data-driven RKHS methods in various fields such as transfer operator theory, time series analysis and image and text processing can be formulated in terms of kernel mean embeddings and

naturally give rise to finite-rank RKHS operators (Song, Huang, et al. 2009; Klus, Schuster and Muandet 2017).

In this thesis, we will combine the functional analytic background of the Hilbert space operator SVD and the theory of RKHSs to

(I) develop a self-contained and rigorous mathematical framework for the SVD of finite-rank operators acting on RKHSs,

(II) show that the SVD of finite-rank RKHS operators can be computed numerically by solving an auxiliary eigenvalue problem of a real matrix,

(III) provide an overview of finite-rank operators on RKHSs which occur in practical computational applications and

(IV) illustrate strategies to apply the theory of the RKHS operator SVD in practice by giving numerical examples based on the theory of kernel transfer operators.

In Section 2, we give a concise overview of Hilbert space theory and different classes of Hilbert space operators as well as their eigendecompositions and singular value decompositions. Section 3 serves as an introduction to the theory of RKHSs and kernel methods. In Section 4, we unite the concepts of Section 2 and Section 3 by developing a spectral theory of finite-rank RKHS operators. We use a recent result by Klus, Schuster and Muandet (2017) to derive a method to numerically compute the SVD of a finite-rank RKHS operator by solving an eigenvalue problem of a real matrix. In Section 6, we prove important properties of the SVD of compact operators in terms of low-rank approximation optimality and generalize the notion of an inverse operator. Additionally, we provide results of experimental RKHS operator SVDs in the context of transfer operator theory to illustrate the capabilities of the developed theory. We conclude our results and give a general outlook on possible future research directions in Section 7. Appendix A contains a compilation of the most important properties and theorems related to the SVD of matrices as a general reference.

2. Preliminaries

This section serves as a collection of functional analytic tools to study linear operators on Hilbert spaces. Most of the presented statements are classical results in operator theory and spectral theory and can be found in the literature on the mentioned topics (see for example Reed and Simon 1980; Rudin 1991; Werner 2007). For most of the results in this section we therefore do not provide proofs. For more context-specific results and assertions, we refer to the related proofs in the literature.

In what follows, let H be a Hilbert space, $\langle \cdot, \cdot \rangle_H$ its inner product and $\|\cdot\|_H$ the norm induced by the inner product. If it is clear to which inner product and norm we refer, we will drop the subscripts. Although a large part of the theory presented in the following sections can be generalized to complex Hilbert spaces and even Banach spaces in some cases, we choose the practically oriented approach and only work with real Hilbert spaces. Whenever we introduce a Hilbert space, it is therefore considered to be real. We will use the term *operator* synonymously for linear map between Hilbert spaces throughout the entire text.

2.1. Hilbert Spaces

For a Hilbert space H, we call a set $\{h_i\}_{i \in I} \subseteq H$ with an index set I an *orthonormal system* (ONS), if $\langle h_i, h_j \rangle = \delta_{ij}$ for all $i, j \in I$. If additionally $\operatorname{span}\{h_i\}_{i \in I}$ is dense in H, then we call $\{h_i\}_{i \in I}$ a *complete orthonormal system* (CONS). We can find a CONS for every Hilbert space H. Furthermore, every ONS of a Hilbert space can be extended to a CONS. If H is separable, then the index set I of every CONS of H is countable.

Theorem 2.1 (Werner 2007, Theorem V.4.9)**.** *Let $\{h_i\}_{i \in I} \subseteq H$ be a CONS. Then the following statements hold:*

(i) We have the series expansion

$$x = \sum_{i \in I} \langle h_i, x \rangle \, h_i$$

*for all $x \in H$ (*generalized Fourier series*).*

(ii) We can generalize the Pythagorean theorem by means of property (i) as

$$\|x\|^2 = \sum_{i \in I} \langle h_i, x \rangle^2$$

*for every $x \in H$ (*Parseval's identity*).*

If M is a subset of H, the set $M^\perp := \{x \in H \mid \langle x, m \rangle = 0 \text{ for all } m \in M\} \subseteq H$ is called the *orthogonal complement* of M. Note that orthogonal complements are always closed, furthermore we have $(\overline{M})^\perp = M^\perp$. If M is a closed subspace of H, then we can express H through the internal direct sum of M and its orthogonal complement as

$$H = M \oplus M^\perp.$$

For a linear operator $A : H \to F$ between two Hilbert spaces H and F we set

$$\begin{aligned} \operatorname{range}(A) &:= \{Ax \in F \,|\, x \in H,\ Ax \text{ is defined}\}, \\ \operatorname{dom}(A) &:= A^{-1}(\operatorname{range}(A)), \\ \operatorname{rank}(A) &:= \dim \operatorname{range}(A), \\ \ker(A) &:= A^{-1}(\{0\}). \end{aligned}$$

Here, $A^{-1}(U) \subseteq H$ for a set $U \subseteq F$ is the preimage of U under A.

Definition 2.2. *If $A : H \to F$ is a linear operator between the Hilbert spaces H and F, we call*

$$\|A\| := \sup_{\|x\|_H = 1} \|Ax\|_F$$

the operator norm *of A induced by $\|\cdot\|_H$ and $\|\cdot\|_F$. If $\|A\| < \infty$, we say that A is a* bounded operator.

The operator norm is indeed a norm on the Banach space of all bounded linear operators from H to F which we denote by $\mathfrak{B}(H, F)$. Whenever an operator is an element of $\mathfrak{B}(H, F)$, we assume, without loss of generality, that its domain is the whole space H. In addition to the classical norm properties, the operator norm is compatible with the norm on H in the sense that $\|Ax\|_F \leq \|A\| \, \|x\|_H$ holds for all $x \in H$. For Hilbert spaces H, F, G and operators $A_1 \in \mathfrak{B}(H, F)$ and $A_2 \in \mathfrak{B}(F, G)$, it is also compatible with the composition of linear operators in the sense that the inequality $\|A_2 A_1\| \leq \|A_2\| \, \|A_1\|$ holds true. If $H = F$, we write $\mathfrak{B}(H) := \mathfrak{B}(H, H)$ for simplicity. A linear operator between two Hilbert spaces is continuous if and only if it is bounded. We say that $A \in \mathfrak{B}(H)$ is *positive*, if $\langle Ax, x \rangle \geq 0$ for all $x \in H$. The *dual space* of H is given as $H' := \mathfrak{B}(H, \mathbb{R})$, we call elements of the dual space *bounded functionals*.

Theorem 2.3 (Fréchet–Riesz)**.** *Let H be a Hilbert space. The map*

$$\begin{aligned} \nu : H &\to H' \\ x &\mapsto \langle \cdot, x \rangle \end{aligned}$$

is an isometric isomorphism.

Theorem 2.3 is also called *Riesz representation theorem*. This fundamental result gives the insight that the dual of every Hilbert space H can isomorphically be identified with H itself. For every functional $l \in H'$, there exists a unique element $x_l \in H$ such that $l = \langle \cdot, x_l \rangle$ and $\|l\| = \|x_l\|_H$ holds true. For $A \in \mathfrak{B}(H, F)$, we call the unique operator $A^* \in \mathfrak{B}(F, H)$ satisfying $\langle Ax, y \rangle_F = \langle x, A^* y \rangle_H$ for all $x \in H$ and $y \in F$ the *adjoint* of A. If $H = F$ and $A = A^*$, we call A *self-adjoint*.

Definition 2.4. *If $x \in H$ and $y \in F$ are both nonzero, we define the* tensor product operator

$$\begin{aligned} y \otimes x : H &\to F, \\ h &\mapsto \langle x, h \rangle_H \, y. \end{aligned}$$

Note that tensor product operators are bounded with $\|y \otimes x\| = \|x\|_H \|y\|_F$ since by the Cauchy–Schwarz inequality on H, we have

$$\sup_{\|h\|_H=1} \|\langle x, h\rangle_H \, y\|_F \leq \sup_{\|h\|_H=1} \|h\|_H \|x\|_H \|y\|_F = \|x\|_H \|y\|_F \tag{2.1}$$

with equivalence being attained for the choice of $h = x \|x\|_H^{-1}$. We define the set $\mathcal{E} := \operatorname{span}\{y \otimes x \,|\, x \in H, y \in F\}$ and call the completion of $\mathcal{E}$ with respect to the inner product

$$\langle y_1 \otimes x_1, y_2 \otimes x_2\rangle := \langle y_1, y_2\rangle_F \langle x_1, x_2\rangle_H \tag{2.2}$$

the *tensor product* of the spaces F and H, which we will write as $F \otimes H$. The tensor product space $F \otimes H$ is a Hilbert space with the inner product given in (2.2). If $\{h_i\}_{i\in I} \subseteq H$ and $\{f_j\}_{j\in J} \subseteq F$ are CONSs, then $\{f_j \otimes h_i\}_{i\in I, j\in J}$ is a CONS of $F \otimes H$. The importance of tensor product operators and tensor products of Hilbert spaces will become apparent in the context of finite-rank operators and Hilbert–Schmidt operators in the next sections.

2.2. Spectral Theory of Compact Operators

In this section, we will define the class of compact operators on Hilbert spaces and introduce classical results on their spectrum and their singular value decomposition.

Definition 2.5. *A linear operator $A : H \to F$ is called* compact*, if for every bounded sequence $(x_n) \subseteq H$, there is a subsequence (x_{n_j}) such that the sequence $(Ax_{n_j}) \subseteq F$ converges.*

The characterisation of compact operators in Definition 2.5 is equivalent to the property that $\overline{\{Ax \,|\, x \in H,\, \|x\| \leq 1\}}$ is compact in F. Every compact operator is bounded. In particular, the set of compact operators from H to F, denoted by $\mathfrak{K}(H, F)$, is a closed subspace of $\mathfrak{B}(H, F)$ and therefore a Banach space. If for Hilbert spaces H, F, G and two operators $A_1 \in \mathfrak{B}(H, F)$ and $A_2 \in \mathfrak{B}(F, G)$ either A_1 or A_2 is compact, then the composition $A_2 A_1$ is also compact.

Theorem 2.6 (Spectral theorem)**.** *If an operator $A \in \mathfrak{K}(H)$ is self-adjoint, then there exists an either finite or countably infinite ordered index set I, an ONS $\{e_i\}_{i\in I} \subseteq H$ and real numbers $(\lambda_i)_{i\in I} \subseteq \mathbb{R} \setminus \{0\}$ such that*

$$A = \sum_{i\in I} \lambda_i (e_i \otimes e_i). \tag{2.3}$$

If the index set I is not finite, then the resulting sequence $(\lambda_i)_{i\in I}$ is a zero sequence.

We call the expression in (2.3) the *eigendecomposition* or *spectral decomposition* of the operator A.

Remark 2.7. Some important properties of compact and self-adjoint operators can directly be obtained from Theorem 2.6:

(i) The numbers λ_i are exactly the nonzero eigenvalues of A since $Ae_i = \lambda_i e_i$. In particular, e_i is a unit-length eigenvector corresponding to the eigenvalue λ_i.

(ii) We may express the space H as an internal direct sum in terms of the ONS $\{e_i\}_{i\in I}$ as

$$H = \ker(A) \oplus \overline{\operatorname{span}\{e_i\}_{i\in I}}.$$

(iii) We have $\|A\| = \max_{i\in I} |\lambda_i|$.

(iv) The operator A is positive if and only if $\lambda_i > 0$ for all $i \in I$.

We now give the most important result of this section, the singular value decomposition of compact operators.

Theorem 2.8 (Singular value decomposition)**.** *An operator $A \in \mathfrak{B}(H, F)$ is compact if and only if there exists an either finite or countably infinite ordered index set I, two ONSs $\{v_i\}_{i\in I} \subseteq H$ and $\{u_i\}_{i\in I} \subseteq F$ and real numbers $(\sigma_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ such that*

$$A = \sum_{i\in I} \sigma_i (u_i \otimes v_i). \tag{2.4}$$

If the index set I is not finite, then the resulting sequence $(\sigma_i)_{i\in I}$ is a zero sequence.

Remark 2.9. If the SVD of $A \in \mathfrak{K}(H, F)$ is given by (2.4), its adjoint may be expressed through the SVD

$$A^* = \sum_{i\in I} \sigma_i (v_i \otimes u_i).$$

In particular, the adjoint of a compact operator is compact.

The following result shows the connection of the eigendecomposition and the SVD of compact operators. It will be of particular importance in the next sections, therefore we provide a proof.

Lemma 2.10. *Let $A \in \mathfrak{K}(H, F)$. If $(\lambda_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ denote the nonzero eigenvalues of A^*A counted with their multiplicities and $\{v_i\}_{i\in I} \subseteq H$ their corresponding unit-length eigenvectors, then the singular value decomposition of A is given by*

$$A = \sum_{i\in I} \lambda_i^{1/2} (u_i \otimes v_i), \tag{2.5}$$

where $u_i := \lambda_i^{-1/2} A v_i$.

Proof. The operator A^*A is compact, positive and self-adjoint. Theorem 2.6 immediately yields

$$A^*A = \sum_{i\in I} \lambda_i (v_i \otimes v_i)$$

for strictly positive eigenvalues $\lambda_i > 0$ and an ONS $\{v_i\}_{i\in I} \subseteq H$. This implies $A^*A\big|_{\operatorname{span}\{v_i\}^\perp} = 0$. For every $v \in \operatorname{span}\{v_i\}^\perp$ we then have

$$\langle A^*Av, h\rangle_H = \langle Av, Ah\rangle_F = 0 \quad \text{for all } h \in H.$$

By choosing $h = v$, we directly deduce $Av = 0$ and therefore

$$A\big|_{\operatorname{span}\{v_i\}^\perp} = 0.$$

Let V' be a CONS of H such that $\{v_i\}_{i\in I} \subseteq V'$. By expanding an element $h \in H$ in terms of V' as given in Theorem 2.1, we have the representation $h = \sum_{v'\in V'} \langle h, v'\rangle\, v'$ and therefore

$$Ah = \sum_{v'\in V'} \langle h, v'\rangle_H\, Av' = \sum_{i\in I} \langle h, v_i\rangle_H\, Av_i = \sum_{i\in I} \lambda_i^{1/2}\, \langle h, v_i\rangle_H\, \lambda_i^{-1/2} Av_i$$

holds for all $h \in H$. By defining $u_i := \lambda_i^{-1/2} Av_i$, we end up with the claimed characterization of A given in (2.5). It now remains to show that $\{u_i\}_{i\in I}$ is an ONS in F. By construction, we have

$$\begin{aligned}\langle u_i, u_j\rangle_F = (\lambda_i\lambda_j)^{-1/2}\, \langle Av_i, Av_j\rangle_F &= (\lambda_i\lambda_j)^{-1/2}\, \langle A^*Av_i, v_j\rangle_H \\ &= \lambda_i^{1/2}\lambda_j^{-1/2}\, \langle v_i, v_j\rangle_H = \delta_{ij}.\end{aligned}$$

Therefore, $\{u_i\}_{i\in I} \subseteq F$ is an ONS and the representation of A in (2.5) is its SVD. □

Remark 2.11. From Lemma 2.10 one may gain additional insight into the connection of the spectral decomposition and the SVD of compact operators.

(i) The singular values σ_i of the operator $A \in \mathfrak{K}(H, F)$ and the nonzero eigenvalues of $A^*A \in \mathfrak{K}(H)$ satisfy $\sigma_i^2 = \lambda_i$. The orthonormal system of eigenvectors of A^*A given by the nonzero eigenvalues is exactly the orthonormal system of right singular vectors of A.

(ii) If an operator $A \in \mathfrak{K}(H)$ is positive and self-adjoint, its SVD and its eigendecomposition are equivalent.

(iii) Analogously to the statement from Remark 2.7(iii) for self-adjoint operators, we can obtain $\|A\| = \max_{i\in I} \sigma_i$ from the SVD of A.

2.3. Finite-rank and Hilbert–Schmidt Operators

Definition 2.12. *Let $A \in \mathfrak{B}(H, F)$. We say that A is* r-dimensional*, if* $\operatorname{rank}(A) = r$*. If $r < \infty$, we say that A is* finite-dimensional *or* finite-rank*.*

We will refer to the set of all finite-rank operators from H to F as $\mathfrak{F}(H, F)$ as well as $\mathfrak{F}(H)$ in the case of $H = F$. Note that in Definition 2.12 we explicitly require finite-rank operators to be bounded, since there are also unbounded linear operators $A : H \to F$ such that $\operatorname{rank}(A) < \infty$. The next result provides a detailed characterisation of finite-rank operators.

Theorem 2.13 (Weidmann 1976, Theorem 6.1)**.** *Let $A : H \to F$ be a linear operator between two Hilbert spaces H and F such that* $\operatorname{dom}(A) = H$*. Then it holds that $A \in \mathfrak{F}(H, F)$ with* $\operatorname{rank}(A) = r$ *if and only if there exist linearly independent sets $\{h_i\}_{1\le i\le r} \subseteq H$ and $\{f_i\}_{1\le i\le r} \subseteq F$ such that*

$$A = \sum_{i=1}^{r} f_i \otimes h_i.$$

If A admits this representation, we additionally have

$$A^* = \sum_{i=1}^{r} h_i \otimes f_i$$

as well as $\|A\| \leq \sum_{i=1}^{r} \|h_i\|_H \|f_i\|_F$.

Comparing Theorem 2.13 with the SVD (2.4) of Theorem 2.8, we immediately see that $\mathfrak{F}(H,F) \subseteq \mathfrak{K}(H,F)$ and that the compact operators of the form

$$A = \sum_{i=1}^{r} \sigma_i (u_i \otimes v_i), \tag{2.6}$$

namely with a finite index set $I = \{1, \ldots, r\}$ are r-dimensional. Additionally, we see that the tensor product operators given in Definition 2.4 are exactly the class of one-dimensional operators and the class of r-dimensional operators is exactly the subset of operators in $\mathfrak{K}(H,F)$ which admit the SVD given in (2.6). The next result describes the connection of finite-rank operators and compact operators in more detail.

Lemma 2.14. *Let $A \in \mathfrak{K}(H,F)$. Then there exists a sequence of operators $(A_n)_{n\in\mathbb{N}}$ in $\mathfrak{F}(H,F)$ such that*

$$\lim_{n\to\infty} \|A - A_n\| = 0.$$

In particular, we have $\overline{\mathfrak{F}(H,F)} = \mathfrak{K}(H,F)$ with respect to the operator norm.

Note that finite-rank operators may naturally be expressed as elements in

$$\operatorname{span}\{y \otimes x \,|\, x \in H, y \in F\}.$$

We will now introduce the class of Hilbert–Schmidt operators, which extends the concept of tensor product operators from Definition 2.4.

Definition 2.15. *Let H and F be Hilbert spaces and $\{h_i\}_{i\in I} \subseteq H$ be a CONS in H. We call an operator $A : H \to F$ a* Hilbert–Schmidt operator*, if it satisfies*

$$\sum_{i\in I} \|Ah_i\|_F^2 < \infty.$$

It is important to mention that the class of Hilbert–Schmidt operators is well defined in the sense that Definition 2.15 does not depend on the choice of the particular CONS. We will refer to $HS(H,F)$ as the set of Hilbert–Schmidt operators and for $A_1, A_2 \in HS(H,F)$ define an inner product as

$$\langle A_1, A_2 \rangle_{HS} := \sum_{i\in I} \langle A_1 h_i, A_2 h_i \rangle_F . \tag{2.7}$$

Hence, $HS(H,F)$ is itself a Hilbert space. Note that $\langle \cdot, \cdot \rangle_{HS}$ induces the norm

$$\|A\|_{HS} = \left(\sum_{i\in I} \|Ah_i\|_F^2 \right)^{1/2}, \tag{2.8}$$

which we initially required to be finite in Definition 2.15. We will now characterize the space of Hilbert–Schmidt operators in detail and establish connections to the classes of compact operators and finite-rank operators.

Lemma 2.16. *The following statements hold true for the space of Hilbert–Schmidt operators $HS(H,F)$:*

(i) If $A \in HS(H,F)$, then A is compact.

(ii) Let $A \in HS(H,F)$. Then there exists a sequence of operators $(A_n)_{n\in\mathbb{N}}$ in $\mathfrak{F}(H,F)$ such that

$$\lim_{n\to\infty} \|A - A_n\|_{HS} = 0.$$

In particular, we have $\overline{\mathfrak{F}(H,F)} = HS(H,F)$ with respect to the Hilbert–Schmidt norm $\|\cdot\|_{HS}$.

(iii) $HS(H,F) \simeq F \otimes H$, where $F \otimes H := \overline{\operatorname{span}\{y \otimes x \,|\, y \in F, x \in H\}}$ with respect to the inner product given by $\langle y_1 \otimes x_1, y_2 \otimes x_2\rangle_{F\otimes H} = \langle y_1, y_2\rangle_F \langle x_1, x_2\rangle_H$.

Example 2.17. For the special case $H = \mathbb{R}^m$ and $F = \mathbb{R}^n$ the classes $\mathfrak{F}(H,F)$, $\mathfrak{K}(H,F)$ and $\mathfrak{B}(H,F)$ coincide and we can represent linear operators from H to F as matrices $A = (a_{ij}) \in \mathbb{R}^{n\times m}$. The Hilbert–Schmidt norm is now also called the *Frobenius norm* given by

$$\|A\|_{HS} = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2\right)^{1/2}.$$

Remark 2.18. Let $A \in HS(H,F)$. By Lemma 2.16(i), we may express A in terms of its SVD as

$$A = \sum_{i\in I} \sigma_i (u_i \otimes v_i)$$

with ONSs $\{v_i\}_{i\in I} \subseteq H$ and $\{u_i\}_{i\in I} \subseteq F$. If we choose $\{h_j\}_{j\in J}$ to be any CONS extension of $\{v_i\}_{i\in I}$, we can use (2.8) and the SVD of A to deduce

$$\|A\|_{HS}^2 = \sum_{j\in J} \|Ah_j\|_F^2 = \sum_{i\in I} \|Av_i\|_F^2 = \sum_{i\in I} \|\sigma_i u_i\|_F^2 = \sum_{i\in I} \sigma_i^2.$$

3. Reproducing Kernel Hilbert Spaces

In this section, we introduce the concept of reproducing kernel Hilbert spaces and give an overview of their most important mathematical properties. We follow the derivations of Steinwart and Christmann (2008) and Schölkopf and Smola (2002).

Definition 3.1. *A Hilbert space H consisting of functions mapping from a nonempty set $\mathbb{X}$ to $\mathbb{R}$ is called a* reproducing kernel Hilbert space (RKHS) *if the evaluation functional*

$$l_x(f) := f(x), \quad f \in H \tag{3.1}$$

is a bounded functional on H for all $x \in \mathbb{X}$.

If H is an RKHS, we will write $\mathscr{H} := H$ and always refer to the set $\mathbb{X}$ as the corresponding *observation space* as described in Definition 3.1 unless stated otherwise. For an RKHS $\mathscr{H}$, Theorem 2.3 assures the existence of a unique function $k_x \in \mathscr{H}$, such that

$$l_x(f) = \langle f, k_x \rangle \quad \text{for all } f \in \mathscr{H} \tag{3.2}$$

for every $x \in \mathbb{X}$. We may therefore construct a function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that

$$k(x, y) := k_x(y).$$

By construction, k satisfies

$$f(x) = \langle f, k(x, \cdot) \rangle$$

for all $f \in \mathscr{H}$ and $x \in \mathbb{X}$. From this, we also obtain the frequently used characterization

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle . \tag{3.3}$$

The formalization of the previous observations leads to the theoretical framework for the characterization of reproducing kernel Hilbert spaces and their properties.

Definition 3.2. *Let $\mathscr{H}$ be an RKHS. A function $k(x, y) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a* reproducing kernel*, if we have $k(x, \cdot) \in \mathscr{H}$ for every $x \in \mathbb{X}$ and*

$$f(x) = \langle f, k(x, \cdot) \rangle \tag{3.4}$$

for all $f \in \mathscr{H}$ and $x \in \mathbb{X}$. Property (3.4) *is called the* reproducing property*.*

We will now introduce a second class of kernel functions and then establish a connection to the class of reproducing kernels.

Definition 3.3. *Let $\mathbb{X}$ be a nonempty set. A function $k : \mathbb{X} \times \mathbb{X} :\to \mathbb{R}$ is called a* symmetric and positive semi-definite kernel *if $k(x, y) = k(y, x)$ for all $x, y \in \mathbb{X}$ and*

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

for every positive integer m and all choices $c_1, \ldots, c_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in \mathbb{X}$.

Proposition 3.4. *Let $\mathscr{H}$ be an RKHS. Then $\mathscr{H}$ has a unique reproducing kernel k which is symmetric and positive semi-definite.*

Remark 3.5. The theory of kernel functions encompasses several other classes of kernels. However, we want to emphasize that we only consider symmetric and positive semi-definite kernel functions in this thesis. Whenever we use the term "kernel", we refer to the type of symmetric and positive semi-definite kernel from Definition 3.3. We additionally want to mention that the terminology associated with the theory of kernel functions is not consistent in the literature. In some references, kernel functions with the property as given by Definition 3.3 are called *symmetric and positive definite* (Steinwart and Christmann 2008).

The uniqueness of k allows to use the above construction from Riesz representation theorem to examine the properties of k. Symmetry and positive semi-definiteness of k can then also directly be obtained by using the properties of the inner product of $\mathscr{H}$ in combination with the characterization (3.3). In combination with the next classical result, we then get a one-to-one correspondence between symmetric and positive semi-definite kernels and RKHSs.

Theorem 3.6 (Aronszajn 1950)**.** *Let k be a symmetric and positive semi-definite kernel on a nonempty set $\mathbb{X}$. Then there exists a unique RKHS $\mathscr{H}$ with k as its reproducing kernel. Furthermore, $\mathscr{H}$ admits the characterization*

$$\mathscr{H} = \overline{\operatorname{span}\{k(x,\cdot) \mid x \in \mathbb{X}\}}.$$

We may define the map $\varphi : \mathbb{X} \to \mathscr{H}$ given by

$$\varphi(x) := k(x,\cdot)$$

and refer to it as the *canonical feature map* induced by k. Note that $\varphi(x)$ can be seen as the representative of x in the Hilbert space $\mathscr{H}$. It is clear that we can express the kernel k in terms of the feature map by rewriting (3.3) as

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle .$$

The following result shows that a tensor product RKHS can be constructed in a very intuitive way from two RKHSs by simply multiplying their kernels. The importance of this statement will later become apparent when examining Hilbert–Schmidt operators on RKHSs.

Lemma 3.7 (Berlinet and Thomas-Agnan 2004, Theorem 13)**.** *If $k : \mathbb{X}_1 \times \mathbb{X}_1 \to \mathbb{R}$ is the kernel of the RKHS $\mathscr{H}$ with associated feature map φ and $l : \mathbb{X}_2 \times \mathbb{X}_2$ is the kernel of the RKHS $\mathscr{F}$ with the associated feature map ψ, then $\mathscr{H} \otimes \mathscr{F}$ is an RKHS with the kernel $k \cdot l : (\mathbb{X}_1 \times \mathbb{X}_1) \times (\mathbb{X}_2 \times \mathbb{X}_2) \to \mathbb{R}$ given by*

$$\big((x,y),(x',y')\big) \mapsto k(x,y)\, l(x',y')$$

and admits the feature map $\varphi \otimes \psi : \mathbb{X}_1 \times \mathbb{X}_2 \to \mathscr{H} \otimes \mathscr{F}$ given by

$$(x,x') \mapsto \varphi(x) \otimes \psi(x').$$

At this point, one may note that the correspondence of tensor product feature map and product kernel is given by the relation

$$\begin{aligned} k(x,y)\,l(x',y') &= \langle \varphi(x), \varphi(y) \rangle_{\mathscr{H}} \, \langle \psi(x'), \psi(y') \rangle_{\mathscr{F}} \\ &= \big\langle \varphi(x) \otimes \psi(x'), \varphi(y) \otimes \psi(y') \big\rangle_{\mathscr{H} \otimes \mathscr{F}} \end{aligned}$$

by definition of the inner product on the tensor product space $\mathscr{H} \otimes \mathscr{F}$ as given by (2.2).

The feature map φ embeds observations from $\mathbb{X}$ into the RKHS $\mathscr{H}$. In statistics and machine learning, most problems are formulated with respect to a finite set of data in the observation space $\mathbb{X}$.

Definition 3.8. *If $D = \{x_1, \ldots, x_n\} \subseteq \mathbb{X}$ is a set of observations, then we call*

$$G := \big(k(x_i, x_j)\big)_{1 \le i,j \le n} = \big(\langle \varphi(x_i), \varphi(x_j) \rangle \big)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$$

the Gram matrix *associated with D.*

Remark 3.9. We note that the Gram matrix is always symmetric and positive semi-definite based on Definition 3.3.

Theorem 3.10 (Horn and Johnson 2013, Theorem 7.2.10)**.** *Let $D = \{x_1, \ldots, x_n\} \subseteq \mathbb{X}$ be a set of observations and G its Gram matrix given by the kernel associated with the feature map $\varphi : \mathbb{X} \to \mathscr{H}$. Then* $\operatorname{rank}(G) = \dim \operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\}$*. In particular, G is symmetric and strictly positive definite if and only if $\{\varphi(x_1), \ldots, \varphi(x_n)\}$ is a linearly independent system in $\mathscr{H}$.*

Remark 3.11. Based on Theorem 3.10, the Gram matrix G is invertible if and only if $\{\varphi(x_1), \ldots, \varphi(x_n)\}$ is a linearly independent system in $\mathscr{H}$.

By linearity of the inner product in $\mathscr{H}$ in each argument, we may numerically evaluate the inner product of arbitrary elements in the subspace $\operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\}$ by means of the Gram matrix as

$$\left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{i=1}^{n} \beta_i \varphi(x_i) \right\rangle_{\mathscr{H}} = \mathbf{a}^T G \mathbf{b},$$

where $\mathbf{a} = [\alpha_1, \ldots, \alpha_n]^T$ and $\mathbf{b} = [\beta_1, \ldots, \beta_n]^T$ are coefficient column vectors in $\mathbb{R}^n$. We note that we can therefore compute the $\mathscr{H}$-norm on $\operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\}$ by

$$\left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|_{\mathscr{H}} = \left(\mathbf{a}^T G \mathbf{a}\right)^{1/2}.$$

The Cauchy–Schwarz inequality on $\mathscr{H}$ can be rewritten in terms of the kernel as

$$k(x,y)^2 \le k(x,x) k(y,y)$$

for $x, y \in \mathbb{X}$. Since the geometry of $\operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\}$ is fully given by the Gram matrix G in terms of norms and inner products, Gram matrices are a powerful concept and a recurring theme in most algorithms based on RKHS theory. It is important to notice that one can describe $\operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\}$ without any knowledge of the feature map φ and the structure of $\mathscr{H}$ itself, since we only need kernel evaluations of the form $k(x_i, x_j)$ to construct the Gram matrix G.

Remark 3.12 (Kernel trick; Schölkopf and Smola 2002, Remark 2.8)**.** If an algorithm based on the RKHS $\mathscr{H}$ is given exlusively in terms of a symmetric and positive semi-definite kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, we may exchange k with another symmetric and positive semi-definite kernel $l : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ to transform the algorithm into an alternative version based on the RKHS $\mathscr{F}$ associated with the kernel l. In particular, if $\mathbb{X}$ is itself a Hilbert space, we see that the $k(x, x') := \langle x, x' \rangle_{\mathbb{X}}$ is a symmetric and positive semi-definite kernel on $\mathbb{X}$, the so-called *linear kernel*. In this case, we can identify the resulting RKHS $\mathscr{H}$ with the dual space X'. As a result, we can *kernelize* algorithms formulated in terms of $\langle \cdot, \cdot \rangle_{\mathbb{X}}$ by replacing the inner product on $\mathbb{X}$ with a nonlinear kernel l. The paragon for this kernelization of a linear method with this trick is the kernel PCA procedure (Schölkopf, Smola and Müller 1998).

Classical nonlinear kernels on the finite-dimensional observation space $\mathbb{X} \subseteq \mathbb{R}^n$ are for example the *polynomial kernel* $k(x, y) := (\langle x, y \rangle_{\mathbb{X}} + c)^d$ with degree $d \in \mathbb{N}$ for $c \leq 0$ and the *Gaussian radial basis function kernel* $k(x, y) := \exp\left(-\frac{1}{2\rho^2} \|x - y\|_2^2\right)$ for a bandwidth $\rho > 0$. Kernels have also been defined and successfully used on text data domains (Lodhi et al. 2002) and graph domains (Vishwanathan et al. 2010). As a more abstract and technical example, kernels can also be defined on the σ-algebra associated with a probability space on an arbitrary nonempty sample space (Schölkopf and Smola 2002). The operating ranges of kernel methods are as diverse as the mathematical interpretations and examples of kernel functions. If we for example consider an abstract observation space $\mathbb{X}$ without a norm or even a metric, we open up a vast library of RKHS-based algorithms to work on elements of $\mathbb{X}$ by introducing an appropriate kernel on $\mathbb{X}$.

4. RKHS Operators

In this section, we examine finite-rank operators between RKHSs. These operators can be defined through a finite number of fixed subspace basis elements in the corresponding RKHSs. In practice, finite-rank RKHS operators are usually estimates of infinite-dimensional operators based on a set of empirical observations (see Section 5). We may therefore later refer to this special type of finite-rank operator as *empirical RKHS operator*, although the concepts in this section are more general and do not need the assumption of the data in the observation space being given from random events. In order to provide efficient tools to describe empirical RKHS operators, we introduce the so-called *feature matrix notation*. We then combine the properties of empirical RKHS operators with the results from classical Hilbert space spectral theory and show that eigendecompositions and singular value decompositions of empirical RKHS operators can be computed by solving auxiliary eigenvalue problems of real matrices.

4.1. Feature Matrix Notation

The notation used in this section was previously employed by Song, Huang, et al. (2009), Klus, Schuster and Muandet (2017) and Muandet et al. (2017) to represent operators on RKHSs. We introduce it in detail and give a comprehensive overview of its properties.

Let $\mathscr{H}$ and $\mathscr{F}$ denote RKHSs based on the observation spaces $\mathbb{X}$ and $\mathbb{Y}$, respectively (see Table 1 for the explanation of related objects). Furthermore, let $x_1, \ldots, x_m \in \mathbb{X}$ and $y_1, \ldots, y_n \in \mathbb{Y}$ denote fixed elements in the observation spaces. Then we call

$$\begin{aligned}\Phi &:= [\varphi(x_1), \ldots, \varphi(x_m)] \text{ and} \\ \Psi &:= [\psi(y_1), \ldots, \psi(y_n)]\end{aligned}$$

their associated *feature matrices*. Note that technically, feature matrices are not matrices, but row vectors in $\mathscr{H}^m$ and $\mathscr{F}^n$. Since the embedded observations in the form of $\varphi(x_i) \in \mathscr{H}$ and $\psi(y_j) \in \mathscr{F}$ can themselves be interpreted as (possibly infinite-dimensional) vectors, the term feature "matrix" is used. We may assume that the elements in a feature matrix form a linearly independent set in the corresponding RKHS. This is for example given in the situation of $k(\cdot,\cdot)$ being a radial basis kernel and the observations $x_1, \ldots, x_m \in \mathbb{X}$ consisting of pairwise distinct elements. The assumption of the entries in the feature matrices being linearly independent will later on simplify the reasoning on the operators expressed in terms of the feature matrices. We will interpret the components of Φ and Ψ as the basis elements of finite-dimensional subspaces of $\mathscr{H}$ and $\mathscr{F}$.

As an analogue to matrix products and matrix-vector multiplication, we may extend the feature matrix notation to represent typical operations with feature matrices. We will now define these operations. In the following, let Υ denote a feature matrix always based on the very same RKHS as the feature matrix Φ, while elements of Ψ and Φ can potentially be in distinct RKHSs $\mathscr{H}$ and $\mathscr{F}$.

Definition 4.1 (Feature matrix notation)**.** *Let $\mathbf{w} \in \mathbb{R}^m$, $B \in \mathbb{R}^{n\times m}$, $x \in \mathbb{X}$ and $v \in \mathscr{H}$. Let furthermore $\Phi := [k(x_1,\cdot), \ldots, k(x_m,\cdot)]$, $\Upsilon := [k(x'_1,\cdot), \ldots, k(x'_n,\cdot)]$ be feature matrices. We define the following operations in terms of the feature matrices Φ and Υ:*

Random variable	X	Y
Observation space	$\mathbb{X}$	$\mathbb{Y}$
Observation	x	y
Kernel	$k(\cdot,\cdot)$	$l(\cdot,\cdot)$
RKHS	$\mathscr{H}$	$\mathscr{F}$
Feature map	φ	ψ
Feature matrix	Φ, Υ	Ψ

Table 1: Notation adapted from Klus, Schuster and Muandet (2017)

Kernel evaluation:

$$\Phi^T v := \begin{bmatrix} \langle \varphi(x_1), v \rangle_{\mathscr{H}} \\ \vdots \\ \langle \varphi(x_m), v \rangle_{\mathscr{H}} \end{bmatrix}, \quad \Phi^T k(x,\cdot) := \begin{bmatrix} k(x_1,x) \\ \vdots \\ k(x_m,x) \end{bmatrix} \in \mathbb{R}^m. \tag{4.1}$$

Linear feature combination:

$$\Phi \mathbf{w} := \sum_{j=1}^{m} w_j k(x_j,\cdot), \tag{4.2}$$

$$B\Phi^T := \left[\sum_{j=1}^{m} b_{1j} k(x_j,\cdot), \quad \dots \quad , \sum_{j=1}^{m} b_{nj} k(x_j,\cdot) \right]^T \in \mathscr{H}^n, \tag{4.3}$$

$$\Upsilon B := \left[\sum_{i=1}^{n} b_{i1} k(x'_i,\cdot), \quad \dots \quad , \sum_{i=1}^{n} b_{im} k(x'_i,\cdot) \right] \in \mathscr{H}^m. \tag{4.4}$$

Kernel matrix evaluation:

$$G_{\Upsilon\Phi} := \Upsilon^T \Phi := \Big(k(x'_i, x_j) \Big)_{\substack{0 \le i \le n \\ 0 \le j \le m}} \in \mathbb{R}^{n \times m}. \tag{4.5}$$

Remark 4.2. (i) The feature matrix notation will be frequently used in combination with matrix multiplication and matrix-vector multiplication. It is important to note that the feature matrix notation is compatible with the classical notation for matrix multiplication in the sense that it preserves associativity: for all $\mathbf{w} \in \mathbb{R}^n$, we have

$$(\Upsilon^T \Phi)\mathbf{w} = \begin{bmatrix} \sum_{j=1}^{m} w_j k(x'_1, x_j) \\ \vdots \\ \sum_{j=1}^{m} w_j k(x'_n, x_j) \end{bmatrix} = \Upsilon^T(\Phi \mathbf{w}) \in \mathbb{R}^n.$$

(ii) If the rows of B are linearly independent in $\mathbb{R}^m$, then the elements of $B\Phi^T$ are linearly independent in $\mathscr{H}$. The analogue statement holds for linearly independent columns of B and elements of the feature matrix ΥB. This can be

seen by the following argument. Let $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^m$ be linearly independent vectors. Let $\alpha_1, \ldots \alpha_n \in \mathbb{R}$. Then from

$$0 = \alpha_1 \Phi \mathbf{w}_1 + \cdots + \alpha_n \Phi \mathbf{w}_n = \Phi(\alpha_1 \mathbf{w}_1 + \cdots + \alpha_n \mathbf{w}_n)$$

we directly obtain $\alpha_1 \mathbf{w}_1 + \cdots + \alpha_n \mathbf{w}_n = 0 \in \mathbb{R}^m$ since the elements in Φ are linearly independent. Therefore we see $\alpha_1 = \cdots = \alpha_n = 0$. Hence, $\Phi\mathbf{w}_1, \ldots, \Phi\mathbf{w}_n$ are also linearly independent in $\mathscr{H}$.

(iii) We may write the Gram matrix associated with Φ as $G_{\Phi\Phi} := \Phi^T \Phi \in \mathbb{R}^{m \times m}$.

4.2. Empirical RKHS Operators

Based on the feature matrix notation, the expression $S := \Psi B \Phi^T$ defines a map from $\mathscr{H}$ to $\mathscr{F}$ given by

$$Sv = \Psi B \Phi^T v = \sum_{i=1}^{n} \psi(y_i) \sum_{j=1}^{m} b_{ij} \langle \varphi(x_i), v \rangle_{\mathscr{H}}$$

for all $v \in \mathscr{H}$. The next result guarantees that S is a bounded finite-rank operator and gives a first intuition about the structure and properties of S.

Proposition 4.3. *Let $\Phi := [\varphi(x_1), \ldots, \varphi(x_m)]$ and $\Psi := [\psi(y_1), \ldots, \psi(y_n)]$ be feature matrices and let $B \in \mathbb{R}^{n \times m}$, then the map given by*

$$S : \mathscr{H} \to \mathscr{F}$$
$$S = \Psi B \Phi^T$$

satisfies the properties

(i) $S \in \mathfrak{F}(\mathscr{H}, \mathscr{F})$,

(ii) $\operatorname{rank}(S) = \operatorname{rank}(B)$,

(iii) $S^* = \Phi B^T \Psi^T$ *and*

(iv) $\|S\| \leq \sum_{i=1}^{r} \sigma_i \|\Psi \mathbf{w}_i\|_{\mathscr{F}} \|\Phi \mathbf{z}_i\|_{\mathscr{H}}$, *where*

$$B = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) [\mathbf{z}_1, \ldots, \mathbf{z}_m]^T$$

is the singular value decomposition of the matrix B.

Proof. The linearity of S directly follows from the fact that the kernel evaluation as given in (4.1) is linear by definition of the inner product in $\mathscr{H}$. Having verified the linearity, we now show that properties (i)–(iv) can directly be obtained from Theorem 2.13. Let $B = W \Sigma Z^T$ be the singular value decomposition of the matrix B (see Theorem A.1). We can rewrite $S = (\Psi W)\Sigma(Z^T \Phi^T)$. If $\mathbf{w}_i$ and $\mathbf{z}_i$ denote the column vectors in W and Z, then S can be written as

$$Sv = \sum_{i=1}^{r} \sigma_i \Psi \mathbf{w}_i \langle \Phi \mathbf{z}_i, v \rangle_{\mathscr{H}} \text{ for all } v \in \mathscr{H}, \tag{4.6}$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of B. Since the elements in Φ and Ψ are linearly independent, we see that ΦZ and ΨW are also feature matrices containing the linearly independent elements $\Phi \mathbf{z}_i \in \mathscr{H}$ and $\Psi \mathbf{w}_i \in \mathscr{F}$ as shown in Remark 4.2(ii). Therefore, (4.6) satisfies the assumptions in Theorem 2.13, if we choose $\{\Phi \mathbf{z}_i\}_{1 \leq i \leq r} \subseteq \mathscr{H}$ and $\{\sigma_i \Psi \mathbf{w}_i\}_{1 \leq i \leq r} \subseteq \mathscr{F}$ to be the required linearly independent sets. Theorem 2.13 directly yields all the desired statements. □

Remark 4.4. Note that the characterization (4.6) is in general not a singular value decomposition of S, since the given basis elements in ΦZ and ΨW are not necessarily orthonormal systems in $\mathscr{H}$ and $\mathscr{F}$, respectively.

Proposition 4.3 justifies the definition of maps of the form $S = \Psi B \Phi^T$ as a class of finite-rank operators.

Definition 4.5. *A map $S \in \mathfrak{F}(\mathscr{H}, \mathscr{F})$ of the form $S = \Psi B \Phi^T$ with feature matrices Φ and Ψ and a real matrix B of compatible size is called* empirical RKHS operator.

4.3. Eigendecomposition of Empirical RKHS Operators

For this subsection, we set $\Upsilon := [k(x'_1, \cdot), \ldots, k(x'_n, \cdot)]$ and $\Phi := [k(x_1, \cdot), \ldots, k(x_n, \cdot)]$ and emphasize that both feature matrices carry the same number of components from the same RKHS $\mathscr{H}$. We now give a central result which was first formulated by Klus, Schuster and Muandet (2017). It leads to the important realization that eigenfunctions related to nonzero eigenvalues of an empirical RKHS operator $S = \Upsilon B \Phi^T$ can be computed by solving an auxiliary real $n \times n$ eigenvalue problem.

Proposition 4.6 (Klus, Schuster and Muandet 2017, Proposition 4.11.)**.** *Let*

$$S : \mathscr{H} \to \mathscr{H}$$
$$S = \Upsilon B \Phi^T$$

be an empirical RKHS operator given by the feature matrices Φ and Υ and the square matrix $B \in \mathbb{R}^{n \times n}$. Then the following statements hold:

(i) If λ is an eigenvalue of $B\Phi^T\Upsilon \in \mathbb{R}^{n \times n}$ with corresponding eigenvector $\mathbf{w} \in \mathbb{R}^n$, then

$$\Upsilon \mathbf{w} \in \mathscr{H}$$

is an eigenfunction of S corresponding to the eigenvalue λ.

(ii) Conversely, if $\lambda \neq 0$ is an eigenvalue of S corresponding to the eigenfunction $v \in \mathscr{H}$, then

$$B\Phi^T v \in \mathbb{R}^n$$

is an eigenvector of $B\Phi^T\Upsilon \in \mathbb{R}^{n \times n}$ corresponding to the eigenvalue λ.

In particular, the operator S and the matrix $B\Phi^T\Upsilon$ share the same nonzero eigenvalues.

Proof. (i) Let $\mathbf{w} \in \mathbb{R}^n$ be an eigenvector of the matrix $B\Phi^T\Upsilon$ corresponding to the eigenvalue λ. Then by the associativity of feature matrix multiplication and kernel evaluation, we have

$$S(\Upsilon \mathbf{w}) = \Upsilon(B\Phi^T\Upsilon \mathbf{w}) = \lambda \Upsilon \mathbf{w}.$$

Furthermore, since $\mathbf{w} \neq 0 \in \mathbb{R}^n$ and the elements in Υ are linearly independent, we have $\Upsilon\mathbf{w} \neq 0 \in \mathscr{H}$. Therefore, $\Upsilon\mathbf{w}$ is an eigenfunction of S corresponding to the eigenvalue λ.

(ii) Let v be an eigenfunction of S associated with the eigenvalue $\lambda \neq 0$. By assumption, we then have

$$\Upsilon B\Phi^T v = \lambda v.$$

By "multiplying" both sides from the left with $B\Phi^T \in \mathscr{H}$ in the sense of (4.1) and using the associativity of the feature matrix notation, we obtain

$$(B\Phi^T\Upsilon)B\Phi^T v = \lambda B\Phi^T v.$$

Furthermore, $B\Phi^T v$ can not be the zero vector in $\mathbb{R}^n$, because otherwise we would have $\Upsilon(B\Phi^T v) = Sv = 0 \neq \lambda v$, since λ was assumed to be a nonzero eigenvalue. Therefore, $B\Phi^T v$ is an eigenvector of the matrix $B\Phi^T\Upsilon$.

The fact that S and $B\Phi^T\Upsilon$ share the same nonzero eigenvalues follows directly. □

Remark 4.7. Proposition 4.6 yields the fundamental insight that eigenfunctions of empirical RKHS operators may be expressed as a linear combination of elements contained in the feature matrices. However, there exist modified formulations of this result (Klus, Schuster and Muandet 2017). As an example, we can define the alternative auxiliary problem

$$\Phi^T\Upsilon B\mathbf{w} = \lambda\mathbf{w}. \tag{4.7}$$

For eigenvalues $\lambda \in \mathbb{R}$ and eigenvectors $\mathbf{w} \in \mathbb{R}^n$ satisfying (4.7), we see that $\Upsilon Bv \in \mathscr{H}$ is an eigenfunction of S. Conversely, for eigenvalues $\lambda \neq 0$ and eigenfunctions $v \in \mathscr{H}$ of S, the auxiliary matrix has the eigenvector $\Phi^T v \in \mathbb{R}^n$. The proof works analogously to the proof of Proposition 4.6. For a second example, we make use of the inverted Gram matrix $(\Phi^T\Phi)^{-1} \in \mathbb{R}^{n\times n}$. Note that by assumption, Φ contains linearly independent elements in $\mathscr{H}$, therefore $\Phi^T\Phi$ is invertible by Theorem 3.10. If an eigenvalue $\lambda \neq 0$ and an eigenvector $\mathbf{w} \in \mathbb{R}^n$ solves (4.7), then also $\Phi(\Phi^T\Phi)^{-1}\mathbf{w}$ is an eigenfunction of S as shown by Klus, Schuster and Muandet (2017).

Remark 4.8. While we need the assumption that the eigenvalue λ of S is nonzero to infer the eigenvector of the auxiliary matrix from the eigenfunction from S, this assumption is not needed the other way around. This has the simple explanation that a rank deficiency of B always introduces a rank deficiency to $S = \Upsilon B\Phi^T$. On the other hand, if $\mathscr{H}$ is infinite dimensional, S as a finite-rank operator *always* has a natural rank deficiency, even when B has full rank. In this case, S has the eigenvalue 0, while B does not.

In order to use Proposition 4.6 as a consistent tool to compute eigenfunctions of RKHS operators, we must ensure that all eigenfunctions corresponding to nonzero eigenvalues of empirical RKHS operators can be computed. In particular, we have to be certain that eigenvalues with a higher geometric multiplicity allow to capture a full set of linearly independent basis eigenfunctions in the associated eigenspace. The next result gives the desired statement.

Lemma 4.9. *Let $S = \Upsilon B\Phi^T \in \mathfrak{F}(\mathscr{H})$ be an empirical RKHS operator. Then it holds:*

(i) If $\mathbf{w}_1 \in \mathbb{R}^n$ and $\mathbf{w}_2 \in \mathbb{R}^n$ are linearly independent eigenvectors of $B\Phi^T\Upsilon$, then $\Upsilon\mathbf{w}_1 \in \mathscr{H}$ and $\Upsilon\mathbf{w}_2 \in \mathscr{H}$ are linearly independent eigenfunctions of S.

(ii) If v_1 and v_2 are linearly independent eigenfunctions belonging to the eigenvalue $\lambda \neq 0$ of S, then $B\Phi^T v_1 \in \mathbb{R}^n$ and $B\Phi^T v_2 \in \mathbb{R}^n$ are linearly independent eigenvectors of $B\Phi^T\Upsilon$.

In particular: if $\lambda \neq 0$, then we have

$$\dim\ker(B\Phi^T\Upsilon - \lambda I_n) = \dim\ker(S - \lambda\mathcal{I}_{\mathscr{H}}).$$

Proof. The eigenvalue-eigenfunction correspondence is covered in Proposition 4.6, it therefore remains to check the linear independence in statements (i) and (ii).

(i) This has already been proven in Remark 4.2(ii).

(ii) Proof by contradiction: let v_1 and v_2 be linearly independent eigenfunctions associated with the eigenvalue $\lambda \neq 0$ of S. Then assume for some $\alpha \neq 0 \in \mathbb{R}$, we have $B\Phi^T v_1 = \alpha B\Phi^T v_2$. Applying Υ from the left to both sides, we get

$$\Upsilon B\Phi^T v_1 = Sv_1 = \lambda v_1 = \alpha\lambda v_2 = \alpha S v_2 = \Upsilon\alpha B\Phi^T v_2,$$

which contradicts the linear independency of v_1 and v_2. Therefore, $B\Phi^T v_1$ and $B\Phi^T v_2$ have to be linearly independent in $\mathbb{R}^n$.

From (i) and (ii), we can directly infer

$$\dim\ker(B\Phi^T\Upsilon - \lambda I_n) = \dim\ker(S - \lambda\mathcal{I}_{\mathscr{H}})$$

by contradiction: let $\lambda \neq 0$ be an eigenvalue of S and $B\Phi^T\Upsilon$. We assume that $\dim\ker(B\Phi^T\Upsilon - \lambda I_n) > \dim\ker(S - \lambda\mathcal{I}_{\mathscr{H}})$. This implies that there exist two eigenvectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$ of $B\Phi^T\Upsilon$, which generate two linearly dependent eigenfunctions $\Upsilon\mathbf{w}_1, \Upsilon\mathbf{w}_2 \in \mathscr{H}$, contradicting statement (i). Hence, we must have $\dim\ker(B\Phi^T\Upsilon - \lambda I_n) \leq \dim\ker(S - \lambda\mathcal{I}_{\mathscr{H}})$. Analogously applying the same logic to statement (ii), we get $\dim\ker(B\Phi^T\Upsilon - \lambda I_n) \geq \dim\ker(S - \lambda\mathcal{I}_{\mathscr{H}})$, which concludes the proof. □

We now formally conclude that Proposition 4.6 is a consistent way to compute the eigenfunctions of empirical RKHS operators by solving an auxiliary matrix eigenvalue problem. This is the first central result of this section.

Corollary 4.10. *If $S = \Upsilon B\Phi^T$ is an empirical RKHS operator and $\lambda \in \mathbb{R}$ is nonzero, we have*

$$\{\Upsilon\mathbf{w} \,|\, B\Phi^T\Upsilon\mathbf{w} = \lambda\mathbf{w}\} = \ker(S - \lambda\mathcal{I}_{\mathscr{H}}).$$

Corollary 4.10 justifies to refer to the eigenvalue problems

$$Sv = \lambda v, \quad v \in \mathscr{H} \tag{4.8}$$

as *primal problem* and

$$B\Phi^T\Upsilon\mathbf{w} = \lambda\mathbf{w}, \quad \mathbf{w} \in \mathbb{R}^n \tag{4.9}$$

as *auxiliary problem*, respectively.

4.4. Singular Value Decomposition of Empirical RKHS Operators

We have seen that we can compute eigenfunctions corresponding to nonzero eigenvalues of an empirical RKHS operator S. By applying this technique to the self-adjoint operator S^*S, we obtain a closed form for the singular value decomposition of S.

Proposition 4.11. *Let $S = \Psi B \Phi^T \in \mathfrak{F}(\mathscr{H}, \mathscr{F})$ be an empirical RKHS operator defined by feature matrices $\Phi = [\varphi(x_1), \ldots, \varphi(x_m)]$ and $\Psi = [\psi(y_1), \ldots, \psi(y_n)]$ and a matrix $B \in \mathbb{R}^{n \times m}$. Then the singular value decomposition of S is given by*

$$S = \sum_{i=1}^{r} \lambda_i^{1/2} (u_i \otimes v_i),$$

where

$$\begin{aligned} v_i &:= (\mathbf{w}_i^T \Phi^T \Phi \mathbf{w}_i)^{-1/2} \Phi \mathbf{w}_i, \\ u_i &:= \lambda_i^{-1/2} S v_i, \end{aligned}$$

with the nonzero eigenvalues $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ of the matrix

$$M\Phi^T\Phi \in \mathbb{R}^{m \times m} \quad \textit{with} \quad M := B^T \Psi^T \Psi B \in \mathbb{R}^{m \times m}$$

counted with their multiplicities and corresponding eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_r \in \mathbb{R}^m$.

Proof. By Remark 4.2(i) and Proposition 4.3, the operator

$$S^*S = \Phi(B^T \Psi^T \Psi B)\Phi^T = \Phi M \Phi^T$$

is an empirical RKHS operator on $\mathscr{H}$. Naturally, S^*S is also positive and self-adjoint. We apply Proposition 4.10 to calculate the unit-length eigenfunctions

$$v_i := \|\Phi \mathbf{w}_i\|_{\mathscr{H}}^{-1} \Phi \mathbf{w}_i = (\mathbf{w}_i^T \Phi^T \Phi \mathbf{w}_i)^{-1/2} \Phi \mathbf{w}_i$$

of S^*S by means of the auxiliary problem

$$M\Phi^T\Phi \mathbf{w}_i = \lambda_i \mathbf{w}_i, \quad \mathbf{w}_i \in \mathbb{R}^m$$

for nonzero eigenvalues λ_i. We use Lemma 2.10 to establish the connection between the eigenfunctions of S^*S and singular functions of S and obtain the desired form for the singular value decomposition of S. □

Remark 4.12. (i) As explained in Remark 4.7, several different auxiliary problems to compute the eigendecomposition of S^*S can be derived. As a result, we can reformulate the calculation of the SVD of S for every possible auxiliary problem.

(ii) We want to emphasize that Proposition 4.11 gives a numerically computable form of the SVD of the empirical RKHS operator S. However, since the auxiliary problem of the eigendecomposition of $S^*S = \Phi M \Phi^T$ always involves several matrix multiplications, better conditioned approaches for practical applications may be investigated. We will adress this again in the outlook on future research in Section 7.

5. Examples of Empirical RKHS Operators

In Section 3, we saw that we can represent a finite set of observations in $\mathbb{X}$ in terms of an RKHS $\mathscr{H}$ by means of the feature map $\varphi : \mathbb{X} \to \mathscr{H}$. We furthermore saw that the geometry of finite-dimensional subspaces of RKHSs can be sufficiently described by the empirical basis elements and the kernel function. Subsequently, in Section 4, we introduced empirical RKHS operators as an abstract concept of expressing bounded finite-rank operators between two RKHSs through a finite set of fixed basis elements in the observation spaces. This section extends the previous situation by two fundamental concepts:

(I) We generalize the discrete setting of observing a finite number of elements in $\mathbb{X}$ and $\mathbb{Y}$ to the continuous setting of marginal and conditional probability distributions on $\mathbb{X}$ and $\mathbb{Y}$. This approach will yield the theory of so-called *kernel mean embeddings* (Smola et al. 2007) and *conditional kernel mean embeddings* (Song, Huang, et al. 2009; Song, Fukumizu and Gretton 2013) as well as several classes of compact RKHS operators. Kernel mean embeddings are closely related to the theory of *kernel integral operators* (Bach 2017).

(II) We show that estimates of these RKHS operators based on a finite number of observations can naturally be expressed as empirical RKHS operators. Hereby, we connect the technical probabilistic scenario with the theory of finite-rank RKHS operators.

All concepts introduced in this section are closely related to each other and serve as the foundation for the numerical examples given in Section 6.2. The notation in this section is based on the notation given in Table 1. In addition to the previous section, we now assume that for the observation spaces $\mathbb{X}, \mathbb{Y}$ there exist σ-algebras $\mathfrak{A}_{\mathbb{X}}$,$\mathfrak{A}_{\mathbb{Y}}$ such that we can define random variables and their corresponding probability distributions on the measurable spaces $(\mathbb{X}, \mathfrak{A}_{\mathbb{X}})$ and $(\mathbb{Y}, \mathfrak{A}_{\mathbb{Y}})$. Furthermore, we will refer to $\mathbb{1}_n$ as the all-ones column vector in $\mathbb{R}^n$ and $\mathbb{1}_n\mathbb{1}_n^T$ the all-ones matrix in $\mathbb{R}^{n\times n}$. Let X and Y be random variables on the sample spaces $\mathbb{X}$ and $\mathbb{Y}$ and let $\mathbb{P}_X$ and $\mathbb{P}_Y$ be their probability distributions. Additionally, let $\mathbb{P}_{YX}$ denote the joint distribution of X and Y and $M_+^1(\mathbb{X})$ be the set of probability distributions on the measurable space $(\mathbb{X}, \mathfrak{A}_{\mathbb{X}})$.

5.1. Kernel Covariance Operators

Covariance operators can be interpreted as a generalization of covariance matrices for the case of infinite-dimensional spaces (Baker 1970; Baker 1973). By representing covariance operators in terms of RKHSs and their associated kernels, we obtain the class of *kernel covariance operators* (Fukumizu, Bach and Jordan 2004; Song, Fukumizu and Gretton 2013; Muandet et al. 2017). This section gives a general overview of kernel covariance operators and their empirical estimates. Additionally, we introduce the basic properties of kernel mean embeddings of marginal distributions. For a detailed discussion of kernel mean embeddings of marginal distributions, we refer the reader to Smola et al. (2007).

Lemma 5.1. *Let $\mathbb{P}_X \in M_+^1(\mathbb{X})$ and $\mathbb{E}_X[\|\varphi(X)\|_{\mathscr{H}}] < \infty$. Then we have*

$$\mu_{\mathbb{P}_X} := \int_{\mathbb{X}} \varphi(X) \, \mathrm{d}\mathbb{P}_X \in \mathscr{H}$$

as well as

$$\langle \mu_{\mathbb{P}_X}, f \rangle_{\mathscr{H}} = \mathbb{E}_X[f(X)] \quad \text{for all } f \in \mathscr{H}. \tag{5.1}$$

Definition 5.2. *If $\mathbb{P}_X \in M_+^1(\mathbb{X})$ and $\mathbb{E}_X[\|\varphi(X)\|_{\mathscr{H}}] < \infty$, then we call*

$$\mu_{\mathbb{P}_X} := \mathbb{E}_X[\varphi(X)] = \int_{\mathbb{X}} \varphi(X) \, \mathrm{d}\mathbb{P}_X$$

the kernel mean embedding *or simply* mean embedding *of $\mathbb{P}_X$.*

The property (5.1) can be interpreted as an extension of the kernel reproducing property (3.4) in the sense that the expectation operator $f \mapsto \mathbb{E}_X[f(X)]$ for $f \in \mathscr{H}$ can be represented by the element $\mu_{\mathbb{P}_X} \in \mathscr{H}$ via $\mathbb{E}_X[f(X)] = \langle \mu_{\mathbb{P}_X}, f \rangle_{\mathscr{H}}$ as an analogue to the point evaluation $f(x) = \langle k(x, \cdot), f \rangle_{\mathscr{H}}$ for $x \in \mathbb{X}$.

Lemma 5.3 (Smola et al. 2007, Theorem 2)**.** *Let $\{x_1, \ldots, x_m\} \subseteq \mathbb{X}$ be independently and identically distributed sampled data with respect to $\mathbb{P}_X$. As $m \to \infty$, we have*

$$\left\| \mu_{\mathbb{P}_X} - \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i) \right\|_{\mathscr{H}} \in \mathcal{O}(m^{-1/2}).$$

Based on Lemma 5.3, we obtain

$$\widehat{\mu}_{\mathbb{P}_X} := \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i)$$

as the *empirical estimate* for $\mu_{\mathbb{P}_X}$. For details and additional results about consistency and convergence of $\widehat{\mu}_{\mathbb{P}_X}$, we refer the reader to Altun and Smola (2006) and Song (2008). We are now equipped with the general background on kernel mean embeddings and introduce the concept of covariance operators and cross-covariance operators on RKHSs.

Lemma 5.4. *If $\mathbb{E}_X[\|\varphi(X)\|_{\mathscr{H}}^2] < \infty$ and $\mathbb{E}_Y[\|\psi(Y)\|_{\mathscr{F}}^2] < \infty$, then*

$$\mathbb{E}_{YX}[\psi(Y) \otimes \varphi(X)] = \int l(Y, \cdot) \otimes k(X, \cdot) \, \mathrm{d}\mathbb{P}_{YX} \in HS(\mathscr{H}, \mathscr{F}).$$

Proof. Based on Lemma 2.16(iii), the expression $l(y, \cdot) \otimes k(x, \cdot)$ defines an operator in $HS(\mathscr{H}, \mathscr{F}) \simeq \mathscr{F} \otimes \mathscr{H}$ for fixed $x \in \mathbb{X}$ and $y \in \mathbb{Y}$. We use the Hilbert–Schmidt norm

$$\|\psi(y) \otimes \varphi(x)\|_{HS} = \left(\langle \psi(y), \psi(y) \rangle_{\mathscr{F}} \langle \varphi(x), \varphi(x) \rangle_{\mathscr{H}} \right)^{1/2} = \|\psi(y)\|_{\mathscr{F}} \|\varphi(x)\|_{\mathscr{H}}$$

to show

$$\begin{aligned}\|\mathbb{E}_{YX}[\psi(Y)\otimes\varphi(X)]\|_{HS} &= \left\|\int l(Y,\cdot)\otimes k(X,\cdot)\,\mathrm{d}\mathbb{P}_{YX}\right\|_{HS}\\ &\leq \int \|l(Y,\cdot)\otimes k(X,\cdot)\|_{HS}\,\mathrm{d}\mathbb{P}_{YX}\\ &= \int \|l(Y,\cdot)\|_{\mathscr{F}}\,\|k(X,\cdot)\|_{\mathscr{H}}\,\mathrm{d}\mathbb{P}_{YX}\\ &= \mathbb{E}_{YX}[\|\psi(Y)\|_{\mathscr{F}}\,\|\varphi(X)\|_{\mathscr{H}}]\\ &\overset{(\star)}{\leq} \mathbb{E}_X[\|\varphi(X)\|_{\mathscr{H}}^2]^{1/2}\,\mathbb{E}_Y[\|\psi(Y)\|_{\mathscr{F}}^2]^{1/2} < \infty.\end{aligned}$$

Here, the inequality $(\star)$ is the Cauchy–Schwarz inequality for the expectation of the square integrable $\mathbb{R}$-valued random variables $\|\psi(Y)\|_{\mathscr{F}}$ and $\|\varphi(X)\|_{\mathscr{H}}$ with respect to $\mathbb{P}_{YX}$. □

Lemma 5.4 justifies the following definition of kernel covariance operators as Hilbert-Schmidt operators on RKHSs.

Definition 5.5. *If* $\mathbb{E}_X[\|\varphi(X)\|_{\mathscr{H}}^2] < \infty$ *and* $\mathbb{E}_Y[\|\psi(Y)\|_{\mathscr{F}}^2] < \infty$*, we call*

$$\mathbb{E}_{YX}[\psi(Y)\otimes\varphi(X)] \in HS(\mathscr{H},\mathscr{F})$$

the kernel cross-covariance operator *and*

$$\mathbb{E}_{YX}[\psi(Y)\otimes\varphi(X)] - \mu_{\mathbb{P}_Y}\otimes\mu_{\mathbb{P}_X} \in HS(\mathscr{H},\mathscr{F})$$

the centered kernel cross-covariance operator *of* Y *and* X*.*

We will use the shorthand $\mathcal{C}_{YX}$ for both the uncentered and the centered kernel cross-covariance operator. If $Y = X$, then $\mathcal{C}_{XX}$ is called *kernel covariance operator* or *centered kernel covariance operator*, respectively.

Remark 5.6. (i) Definition 5.5 is ambiguous in the sense that we use the notation $\mathcal{C}_{YX}$ for both the uncentered and the centered kernel cross-covariance operator. However, if we are referring to one of the two versions specifically in a general setting, we will refer to the used version of kernel cross-covariance operator explicitly in the context.

(ii) In the special case of $\mathbb{X}\subseteq\mathbb{R}^d$ and $\mathbb{Y}\subseteq\mathbb{R}^n$ both equipped with linear kernels, we obtain the covariance matrix $\mathcal{C}_{XX} = \big(\mathrm{Cov}[X_i,X_j]\big)_{ij} \in \mathbb{R}^{d\times d}$ and the cross-covariance matrix $\mathcal{C}_{YX} = \big(\mathrm{Cov}[Y_i,X_j]\big)_{ij} \in \mathbb{R}^{n\times d}$ as special cases for the centered cross-covariance operator.

(iii) From now on, we will use the general term (cross-)covariance operator synonymously with kernel (cross-)covariance operator as we are always working in the RKHS context.

Remark 5.7. The cross-covariance operator $\mathcal{C}_{YX}$ satisfies the following properties:

(i) $\mathcal{C}_{YX}^* = \mathcal{C}_{XY}$,

(ii) $\langle g, \mathcal{C}_{YX} f \rangle_{\mathscr{F}} = \mathrm{Cov}[g(Y), f(X)]$ for $f \in \mathscr{H}$ and $g \in \mathscr{F}$,

(iii) if $\mathbb{E}_{Y|X}[g(Y) \,|\, X = \cdot] \in \mathscr{H}$ for a $g \in \mathscr{F}$, then it holds that

$$\mathcal{C}_{XY} g = \mathcal{C}_{XX} \mathbb{E}_{YX}[g(Y) \,|\, X = \cdot].$$

For detailed proofs, we refer the reader to Fukumizu, Bach and Jordan (2004), Song, Fukumizu and Gretton (2013) and Muandet et al. (2017).

Let $\{(x_1, y_1), \dots (x_m, y_m)\} \subseteq \mathbb{X} \times \mathbb{Y}$ be a set of observation tuples sampled independently and identically distributed with respect to $\mathbb{P}_{YX}$ and let $\Phi := [\varphi(x_1), \dots, \varphi(x_m)]$ and $\Psi := [\psi(y_1), \dots, \psi(y_m)]$ be the corresponding feature matrices. Then the empirical estimates for $\mathcal{C}_{YX}$ are given by

$$\widehat{\mathcal{C}}_{YX} := \frac{1}{m} \sum_{i=1}^{m} \psi(y_i) \otimes \varphi(x_i) = \frac{1}{m} \Psi \Phi^T \tag{5.2}$$

for the uncentered cross-covariance operator and

$$\widehat{\mathcal{C}}_{YX} := \frac{1}{m} \sum_{i=1}^{m} (\psi(y_i) - \widehat{\mu}_{\mathbb{P}_Y}) \otimes (\varphi(x_i) - \widehat{\mu}_{\mathbb{P}_Y}) = \frac{1}{m} \Psi B \Phi^T \tag{5.3}$$

for the centered cross-covariance operator, where

$$B := I_m - \frac{1}{m} \mathbb{1}_m \mathbb{1}_m^T \in \mathbb{R}^{m \times m}$$

is the corresponding centering matrix.

Lemma 5.8. *As $m \to \infty$, we have*

$$\left\| \mathcal{C}_{YX} - \widehat{\mathcal{C}}_{YX} \right\|_{HS(\mathscr{H}, \mathscr{F})} \in \mathcal{O}(m^{-1/2}),$$

where $\mathcal{C}_{YX}$ is either the centered or uncentered cross-covariance operator and $\widehat{\mathcal{C}}_{YX}$ the corresponding uncentered estimate (5.2) or centered estimate (5.3), respectively.

Proof. We identify the space of Hilbert–Schmidt operators $HS(\mathscr{H}, \mathscr{F})$ with the product RKHS $\mathscr{F} \otimes \mathscr{H}$ as described in Lemma 2.16(iii). The operator $\mathcal{C}_{YX}$ may be reinterpreted as a kernel mean embedding of the distribution $\mathbb{P}_{YX}$ into the RKHS $\mathscr{F} \otimes \mathscr{H}$ based on the product feature map $\psi \otimes \varphi$ as given in Lemma 3.7. The estimate $\widehat{\mathcal{C}}_{YX}$ is just the estimate of the associated kernel mean embedding and its convergence is then guaranteed by Lemma 5.3. □

We emphasize the concluding result of this section: the estimates for the both the centered and the uncentered cross-covariance operator are empirical RKHS operators in the sense of Definition 4.5.

5.2. Conditional Mean Embedding Operators

In this section, we will extend the kernel mean embedding of a marginal distribution $\mathbb{P}_X$ to the embedding of a conditional distribution $\mathbb{P}_{Y|X} = \mathbb{P}[Y \mid X]$. We give general results about the existence of conditional mean embeddings and needed regularity assumptions. Furthermore, we will define a class of operators representing the conditional mean embeddings based on covariance and cross-covariance operators, the so-called *conditional mean embedding operators*. Concluding, we will show that empirical estimates of conditional mean embedding operators can be written as empirical RKHS operators. The background on conditional mean embeddings is given by Song, Huang, et al. (2009) and Song, Fukumizu and Gretton (2013). For a comprehensive review of the kernel mean embedding theory see Muandet et al. (2017).

Given a conditional probability distribution $\mathbb{P}_{Y|X}$ and the event $X = x \in \mathbb{X}$, we want to define the conditional mean embedding $\mu_{\mathbb{P}_{Y|X=x}}$ such that the two intuitive properties hold:

(I) $\mu_{\mathbb{P}_{Y|X=x}} := \mathbb{E}_{Y|X}[\psi(Y) \mid X = x] \in \mathscr{F}$ for all $x \in \mathbb{X}$ and

(II) $\mathbb{E}_{Y|X}[g(Y)|X = x] = \left\langle g, \mu_{\mathbb{P}_{Y|X=x}} \right\rangle_{\mathscr{F}}$ for all $g \in \mathscr{F}$.

Note that we distinguish the object $\mu_{\mathbb{P}_{Y|X=x}} \in \mathscr{F}$ for a fixed $x \in \mathbb{X}$ and the map

$$\begin{aligned} \mu_{\mathbb{P}_{Y|X}} : \mathbb{X} &\to \mathscr{F}, \\ x &\mapsto \mu_{\mathbb{P}_{Y|X=x}}. \end{aligned}$$

To emphasize this technical distinction, we want to rewrite the conditional mean embedding in terms of an operator $\mathcal{U}_{Y|X} : \mathscr{H} \to \mathscr{F}$ for which we require that

$$\mathcal{U}_{Y|X} k(x, \cdot) = \mu_{\mathbb{P}_{Y|X=x}} = \mathbb{E}_{Y|X}[\psi(Y)|X = x] \tag{5.4}$$

holds for all $x \in \mathbb{X}$. The following result yields insight into the existence and structure of such an operator $\mathcal{U}_{Y|X}$.

Lemma 5.9 (Song, Huang, et al. 2009, Theorem 4)**.** *Let $\mathcal{C}_{XX}^{-1} \in \mathfrak{B}(\mathscr{H})$ and let furthermore $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathscr{H}$. Then the operator*

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX} \mathcal{C}_{XX}^{-1}, \tag{5.5}$$

satisfies the properties (I) and (II) with the relation $\mu_{\mathbb{P}_{Y|X=x}} := \mathcal{U}_{Y|X} k(x, \cdot)$.

Definition 5.10. *Let $\mathcal{C}_{XX}^{-1} \in \mathfrak{B}(\mathscr{H})$ and $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathscr{H}$ for $g \in \mathscr{F}$, then the operator $\mathcal{U}_{Y|X}$ defined as above is called* conditional mean embedding operator *or simply* conditional mean embedding.

Although we are able to give an analytic expression of $\mathcal{U}_{Y|X}$ in terms of Lemma 5.9, the nonexistence or ill-posedness of the finding of the inverse covariance operator $\mathcal{C}_{XX}^{-1} \in \mathfrak{B}(\mathscr{H})$ may lead to complications in practice. By definition, $\mathcal{C}_{XX}$ is a compact and self-adjoint operator. Based on Theorem 2.6, it might possess a sequence of arbitrarily small eigenvalues. Therefore, one usually considers the regularized version $\mathcal{C}_{XX} + \alpha \mathcal{I}_{\mathscr{H}}$ for an appropriate regularization parameter $\alpha \in \mathbb{R}$ (Muandet et al. 2017)

to introduce a spectral shift to $\mathcal{C}_{XX}$ and produce a well-posed alternative inversion problem. The regularization theory of compact operators is beyond the scope of this thesis. We acknowledge the problematic nature of the given inversion problem and assume $\mathcal{C}_{XX}$ to be replaced with the accordingly regularized version wherever needed.

Lemma 5.11 (Song, Huang, et al. 2009, Theorem 6)**.** *Let* $\{(x_1, y_1), \dots (x_m, y_m)\} \subseteq \mathbb{X} \times \mathbb{Y}$ *be a set of observation tuples sampled independently and identically distributed with respect to* $\mathbb{P}_{YX}$ *and let* $\Phi := [\varphi(x_1), \dots, \varphi(x_m)]$ *and* $\Psi := [\psi(y_1), \dots, \psi(y_m)]$ *be the corresponding feature matrices. If* $\varphi(x) \in \operatorname{range}(\mathcal{C}_{XX})$*, then*

$$\left\| \mu_{\mathbb{P}_{Y|X=x}} - \widehat{\mu}_{\mathbb{P}_{Y|X=x}} \right\|_{\mathscr{H}} \in \mathcal{O}\left((m\alpha)^{-1/2} + \alpha^{1/2} \right),$$

where $\widehat{\mu}_{\mathbb{P}_{Y|X=x}} := \Psi(G_{\Phi\Phi} + \alpha\, m I_m)^{-1} \Phi^T \varphi(x)$.

We obtain the empirical estimate

$$\widehat{\mathcal{U}}_{Y|X} := \Psi(G_{\Phi\Phi} + \alpha\, m I_m)^{-1} \Phi^T$$

for the conditional mean embedding operator $\mathcal{U}_{Y|X}$. Note that $\widehat{\mathcal{U}}_{Y|X}$ is an empirical RKHS operator in the sense of Definition 4.5.

5.3. Kernel Transfer Operators

Transfer operators are used as a tool in various scientific communities to describe, analyze and model complex nonlinear dynamical systems and stochastic processes. Transfer operator-based methods are employed, for example, in fluid dynamics and the analysis of ocean currents (Froyland, Padberg-Gehle, et al. 2007), molecular dynamics (Pérez-Hernández et al. 2013) and meteorology and climate science (Tantet, Burgt and Dijkstra 2015) and can in theory be applied to a vast set of problems in other fields allowing for high dimensional and complex dynamical systems such as mechanical engineering or stock market modeling. When we use the general term "transfer operator" in this work, we usually refer to the *Perron–Frobenius operator* (Lasota and Mackey 1994) or the *Koopman operator* (Budiŝić, Mohr and Mezić 2012), which are closely related. This section is intended to be a short summary of the general transfer operator theory and serves as an overview of how transfer operators and RKHS theory can be combined, leading to the concept of *kernel transfer operators* (Klus, Schuster and Muandet 2017).

For most of this section, we will omit detailed technicalities related to measure theory and stochastic processes. For a more thorough investigation of the measure theoretic background of transfer operator theory, see Lasota and Mackey (1994) and Klus, Koltai and Schütte (2016).

Let $(X_t)_{t\in T}$ be an $\mathbb{X}$-valued stochastic process where T is an appropriate index set and $(\mathbb{X}, \mathfrak{A}_{\mathbb{X}})$ is a measurable space. Although a more general treatment of transfer operators is possible, we will consider the typical setting $\mathbb{X} \subseteq \mathbb{R}^n$ and $T = \mathbb{R}_{\geq 0}$. We assume that $(X_t)_{t\in T}$ admits a function $p_\tau : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}[X_{t+\tau} \in \mathcal{A} \,|\, X_t = x] = \int_{\mathcal{A}} p_\tau(y, x)\, \mathrm{d}y$$

for $x, y \in \mathbb{X}$, $\mathcal{A} \in \mathfrak{A}_{\mathbb{X}}$ and $\tau > 0$. The function p_τ is to be understood as the conditional *transition density* describing the transfer from a state $x \in \mathbb{X}$ to a measurable set $\mathcal{A}$ with respect to a *lag time* τ under the dynamical laws of the system $(X_t)_{t\in T}$.

Definition 5.12. *Let $(X_t)_{t\in T}$ be an $\mathbb{X}$-valued stochastic process and $p_\tau(y, x)$ its transition density with respect to a lag time τ. We define the* Koopman operator *$\mathcal{K}_\tau : L^\infty(\mathbb{X}) \to L^\infty(\mathbb{X})$ by the mapping rule*

$$(\mathcal{K}_\tau f)(x) = \int_{\mathbb{X}} p_\tau(y, x) f(y) \,\mathrm{d}y$$

for all $f \in L^\infty(\mathbb{X})$ and the Perron–Frobenius operator *$\mathcal{P}_\tau : L^1(\mathbb{X}) \to L^1(\mathbb{X})$ by*

$$(\mathcal{P}_\tau p)(y) = \int_{\mathbb{X}} p_\tau(y, x) p(x) \,\mathrm{d}x$$

for all $p \in L^1(\mathbb{X})$.

The Koopman operator can be interpreted as the propagator of observable functions $f \in L^\infty(\mathbb{X})$ under the given dynamics since $(\mathcal{K}_\tau f)(x) = \mathbb{E}[f(X_{t+\tau}) \,|\, X_t = x]$. Similarly, the Perron Frobenius operator propagates system state densities $p_t \in L^1(\mathbb{X})$ associated with a fixed time $t \in T$ to evolved state densities $p_{t+\tau} = \mathcal{P}_\tau p_t$ at a time $t + \tau \in T$. Although both operators are linear on $L^\infty(\mathbb{X})$ and $L^1(\mathbb{X})$, respectively, they are able to capture nonlinear components of the underlying dynamics on $\mathbb{X}$ because they are based on infinite-dimensional spaces. If the dynamics of the system are *time-homogeneous* in the sense that the distribution of $X_{t+\tau}$ conditioned on $X_t = x \in \mathbb{X}$ depends only on x and τ for all $t \in T$, then we have $\mathcal{K}_{\tau_1+\tau_2} = \mathcal{K}_{\tau_1}\mathcal{K}_{\tau_2}$ as well as $\mathcal{P}_{\tau_1+\tau_2} = \mathcal{P}_{\tau_1}\mathcal{P}_{\tau_2}$.

If there exists an almost everywhere non-negative function $\pi \in L^1(\mathbb{X})$ of unit norm such that $\mathcal{P}_\tau \pi = \pi$, then we call π an *invariant* or *stationary density* of the system. If the stationary density exists and is unique, the system is called *ergodic* (Klus, Koltai and Schütte 2016) and we may define the space $L^1_\pi(\mathbb{X})$ of the Lebesgue-integrable functions with respect to the density π (Klus, Nüske, et al. 2018) through the weighted norm

$$\|f\|_{L^1_\pi(\mathbb{X})} := \int_{\mathbb{X}} |f(x)| \pi(x) \,\mathrm{d}x.$$

Definition 5.13. *Let $(X_t)_{t\in T}$ be an $\mathbb{X}$-valued ergodic stochastic process and $p_\tau(y, x)$ its transition density with respect to a lag time $\tau > 0$. Let furthermore $\pi \in L^1(\mathbb{X})$ be the stationary density of the system given by $(X_t)_{t\in T}$. Then we call $\mathcal{T}_\tau : L^1_\pi(\mathbb{X}) \to L^1_\pi(\mathbb{X})$ defined by*

$$(\mathcal{T}_\tau u)(y) = \int_{\mathbb{X}} \frac{\pi(x)}{\pi(y)} p_\tau(y, x) u(x) \,\mathrm{d}x$$

the Perron–Frobenius operator with respect to the invariant density *or simply* transfer operator.

Similarly to the operator $\mathcal{T}_\tau$, we can define the Perron–Frobenius operator with respect to arbitrary density functions, although $\mathcal{T}_\tau$ is the most commonly used version.

Note that we have the relation $\pi\mathcal{T}_\tau u = \mathcal{P}_\tau(u\pi)$ and therefore we get the equivalence $\mathcal{P}\pi = \pi \Leftrightarrow \mathcal{T}_\tau 1_{\mathbb{X}} = 1_{\mathbb{X}}$ where $1_{\mathbb{X}} \in L^1_\pi(\mathbb{X})$ is the constant one function on $\mathbb{X}$ (see Koltai, Wu, et al. 2018). Additionally, if the transition density is continuous in both arguments, the transfer operators are compact. Given $1 < p, q < \infty$ such that $p^{-1} + q^{-1} = 1$, the operators $\mathcal{K}_\tau, \mathcal{P}_\tau$ and $\mathcal{T}_\tau$ can alternatively be defined on the spaces $L^p(\mathbb{X})$, $L^q(\mathbb{X})$ and $L^q_\pi(\mathbb{X})$, respectively (for details see Baxter and Rosenthal 1995; Klus, Nüske, et al. 2018). The obvious choice $p = q = 2$ gives the possibility to work the Hilbert space $L^2(\mathbb{X})$ (Noé and Nüske 2013; Koltai, Wu, et al. 2018). By defining the π-weighted duality pairing for $f \in L^p(\mathbb{X})$ and $g \in L^q(\mathbb{X})$ as

$$\langle f, g\rangle_\pi := \int_{\mathbb{X}} f(x)g(x)\pi(x)\,\mathrm{d}x,$$

we see that

$$\langle \mathcal{K}_\tau f, g\rangle_\pi = \langle f, \mathcal{T}_\tau g\rangle_\pi . \tag{5.6}$$

If additionally the system admits the so-called *detailed balance* or *reversibility condition* as

$$\pi(x)\, p_\tau(y, x) = \pi(y)\, p_\tau(x, y) \tag{5.7}$$

for all $x, y \in \mathbb{X}$, then we have the important equality $\mathcal{K}_\tau = \mathcal{T}_\tau$ and therefore (5.6) yields self-adjointness of $\mathcal{K}_\tau$ and $\mathcal{T}_\tau$ with respect to $\langle\cdot,\cdot\rangle_\pi$. Intuitively, the forward and backward dynamics of a reversible system are indistinguishable. If the system is reversible, then $\mathcal{T}_\tau$ admits an eigendecomposition as given by Theorem 2.6. It can furthermore be shown that the eigenvalues can be ordered as

$$1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \dots,$$

which allows a powerful interpretation of the eigendecomposition in terms of the dynamics of $(X)_{t\in T}$. The eigenvalues correspond to a set of $\langle\cdot,\cdot\rangle_\pi$-orthogonal eigenfunctions $\phi_i \in L^1_\pi(\mathbb{X})$ such that $\mathcal{T}_\tau\phi_i = \lambda_i\phi_i$. The eigenvalue $\lambda_1 = 1$ is associated with the stationary density $\phi_1 = \pi$ and the other eigenfunctions can be interpreted as a representation of exponentially decaying modes in the dynamics of the system (see Noé and Nüske 2013; Nüske et al. 2014; Klus, Nüske, et al. 2018), ordered by magnitude through the eigenvalues. There exist similar approaches to model dynamical systems and stochastic processes with transfer operators in the case of time-inhomogeneous dynamics or systems that do not fulfill the detailed balance condition (Koltai, Wu, et al. 2018), these approaches rely mostly on the SVD of the transfer operators. We will discuss the interpretation of the SVD of transfer operators in more details in Section 6.2. We now extend the theory of transfer operators to their representation in RKHSs. We omit the technical background and only give an overview of the definitions, analytic forms and corresponding estimators. For detailed derivations and proofs, the reader may refer to Klus, Schuster and Muandet (2017).

Kernel Perron–Frobenius operator. The kernel Perron–Frobenius operator is the result of the assumption that the underlying densities p_t describing the system $(X_t)_{t\in T}$ are elements of the RKHS $\mathscr{H}$. For $t \in T$ and a fixed lag time $\tau > 0$, we

define the process $(Y_t)_{t\in T}$ through the relation $Y_t := X_{t+\tau}$. That is, we introduce a τ time-lagged version of the system. Note that both random variables are defined on the same sample space $\mathbb{X}$. We are now looking for the RKHS density propagation operator $\mathcal{A}_\tau^{(k)} : \mathscr{H} \to \mathscr{H}$ satisfying $p_t \mapsto p_{t+\tau}$. The following proposition gives the analytic form of $\mathcal{A}_\tau^{(k)}$ with respect to either the uniform density on $\mathbb{X}$ or the invariant density π of the system.

Proposition 5.14 (Klus, Schuster and Muandet 2017, Proposition 4.1. & Corollary 4.2.)**.** *The kernel Perron–Frobenius operator $\mathcal{A}_\tau^{(k)} : \mathscr{H} \to \mathscr{H}$ satisfying*

$$\left(\mathcal{A}_\tau^{(k)} p_t\right)(y) = \int_{\mathbb{X}} \frac{p_{\mathbb{X}}(x)}{p_{\mathbb{X}}(y)} p_\tau(y,x) p_t(x)\,\mathrm{d}x$$

is given by

$$\mathcal{A}_\tau^{(k)} := \mathcal{C}_{XX}^{-1}\mathcal{C}_{YX},$$

where we consider either the case of $p_{\mathbb{X}}$ being the uniform density on $\mathbb{X}$ and write $\mathcal{P}_\tau^{(k)} := \mathcal{A}_\tau^{(k)}$ or the system with respect to the invariant density $p_{\mathbb{X}} = \pi$ and write $\mathcal{T}_\tau^{(k)} := \mathcal{A}_\tau^{(k)}$.

As discussed in the previous section, the regularized version $(\mathcal{C}_{XX} + \alpha\mathcal{I}_{\mathscr{H}})^{-1}$ of the inverse kernel covariance operator $\mathcal{C}_{XX}^{-1}$ may need to be considered. By incorporating the estimate (5.2) for the uncentered covariance and cross-covariance operators as derived in Section 5.1, we obtain the empirical estimates for $\mathcal{P}_\tau^{(k)}$ and $\mathcal{T}_\tau^{(k)}$.

Corollary 5.15 (Klus, Schuster and Muandet 2017, Proposition 4.3.)**.** *Let $Y_t = X_{t+\tau}$. If the set of data tuples $\{(x_1,y_1),\dots,(x_m,y_m)\} \subset \mathbb{X}\times\mathbb{X}$ sampled independently and identically distributed from $\mathbb{P}_{YX}$ with respect to either the uniform density on $\mathbb{X}$ or the invariant density π of $(X_t)_{t\in T}$ and $\Phi = [\varphi(x_1),\dots,\varphi(x_m)]$ and $\Psi = [\varphi(y_1),\dots,\varphi(y_m)]$ are feature matrices, the empirical estimate for the associated kernel transfer operator is given by the empirical RKHS operator*

$$\widehat{\mathcal{A}}_\tau^{(k)} := \Psi A \Phi^T,$$

where $A \in \mathbb{R}^{m\times m}$ is given by either $A_1 = G_{\Phi\Psi}^{-1} G_{\Phi\Phi}^{-1} G_{\Phi\Psi}$ or $A_2 = G_{\Phi\Psi}^{-1} G_{\Psi\Phi}^{-1} G_{\Psi\Psi}$.

Kernel Koopman operator. Analogously to the derivation of the kernel Perron–Frobenius operator, we now define the kernel Koopman operator $\mathcal{K}_\tau^{(k)} : \mathscr{H} \to \mathscr{H}$. We reinterpret our definition of obervables f of the system and assume that $f_t \in \mathscr{H}$ and that the set of all observables is invariant under $\mathcal{K}_\tau^{(k)}$. By definition of the Koopman operator, the kernel Koopman operator shall satisfy

$$f_{t+\tau} := \left(\mathcal{K}_\tau^{(k)} f_t\right)(x) = \mathbb{E}[f_t(X_{t+\tau}) \,|\, X_t = x].$$

This is essentially a restatement of Remark 5.7(iii) with the form

$$\mathcal{K}_\tau^{(k)} := \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY}.$$

Again, we may consider an appropriately regularized operator $(\mathcal{C}_{XX} + \alpha\mathcal{I}_{\mathscr{H}})^{-1}$. Naturally, we obtain an estimate of the kernel Koopman operator as the empirical RKHS operator

$$\widehat{\mathcal{K}}_\tau^{(k)} := \widehat{\mathcal{C}}_{XX}^{-1}\widehat{\mathcal{C}}_{XY} = (\Phi\Phi^T)^{-1}(\Phi\Psi^T) = \Phi G_{\Phi\Phi}^{-1}\Psi^T.$$

Embedded kernel transfer operators. Note that the kernel Koopman operator $\mathcal{K}_\tau^{(k)}$ and the variants of the kernel Perron–Frobenius operator $\mathcal{A}_\tau^{(k)}$ are not self-adjoint with respect to $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. As shown by Klus, Schuster and Muandet (2017), one may extend the interpretation of kernel transfer operators to the so-called *embedded transfer operators* which do not directly act on observables and densities as elements of $\mathscr{H}$, but $\mathscr{H}$-embedded analogues of observables $f_t \in L^\infty(\mathbb{X})$ and $p_t \in L^1(\mathbb{X})$. We skip the technical details and directly introduce the *embedded Koopman operator*

$$\mathcal{K}_\tau^{(\mathcal{E})} := \mathcal{C}_{XY}\mathcal{C}_{XX}^{-1}$$

and the *embedded Perron–Frobenius operator*

$$\mathcal{P}_\tau^{(\mathcal{E})} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} = \mathcal{U}_{Y|X}.$$

We then directly see that we have the relations $(\mathcal{K}_\tau^{(k)})^* = \mathcal{P}_\tau^{(\mathcal{E})}$ and $(\mathcal{P}_\tau^{(k)})^* = \mathcal{K}_\tau^{(\mathcal{E})}$. By Proposition 4.3(iii), the empirical estimates of the embedded transfer operators can therefore be written as the algebraically transposed form of the estimates of associated adjoint kernel transfer operator. Hence, we obtain the corresponding estimates in terms of the empirical RKHS operators

$$\widehat{\mathcal{K}}_\tau^{(\mathcal{E})} := \Phi A^T \Psi^T$$

for A being one of the matrices given in Corollary 5.15 and

$$\widehat{\mathcal{P}}_\tau^{(\mathcal{E})} := \Psi G_{\Phi\Phi}^{-1} \Phi^T.$$

The detailed connection of embedded transfer operators and kernel transfer operators is beyond the scope of this overview. We refer the reader to Klus, Schuster and Muandet (2017) for a comprehensive discussion of this topic.

6. Applications

In this section, we give applications of the SVD of compact operators and study its approximative properties and optimality characteristics as well as its connection to pseudoinverse operators. These theoretical results directly generalize the SVD of matrices (see Appendix A for an overview of the SVD in matrix theory). Additionally, we provide numerical examples based on the theory of kernel transfer operators from Section 5.3 and provide a basic intuition of the interpretation of the results.

6.1. Low-Rank Operator Approximation

The following results generalize the approximation of a matrix by another matrix under rank constraints through a construction based on the SVD. The classical proof for the low-rank matrix approximation was first published by Eckart and Young (1936) for the Frobenius norm. It was later added by Mirsky (1960) that the original statements also hold true for every unitarily invariant norm on matrix spaces (that is, for matrix norms which are invariant under orthonormal transformations). Since then, similar results with additional constraints on the approximating matrix have been formulated (Golub, Hoffman and Stewart 1987). We give an infinite-dimensional generalization of the classical Eckart–Young theorem for the operator norm and then derive a proof for the Hilbert–Schmidt norm.

Without loss of generality, we will add additional assumptions on the form of the SVD of compact operators and slightly modify the previously used notation. If an operator $A \in \mathfrak{K}(H, F)$ admits the SVD

$$A = \sum_{i \in I} \sigma_i (u_i \otimes v_i), \tag{6.1}$$

we will assume that the index set has the form $I = \{1, 2, \dots\}$ and that $\sigma_i \geq \sigma_{i+1}$ holds for all $i, i+1 \in I$, not changing the either finite sum or absolutely convergent series in (6.1). Hence, the SVD components in the sum of (6.1) are uniquely determined up to permutation of rank-one operators of the form $(u_i \otimes v_i)$ and $(u_j \otimes v_j)$, if they correspond to a mutual singular value $\sigma_i = \sigma_j$. The expression $\sigma_j(A)$ denoting the jth singular value of an arbitrary compact operator $A \in \mathfrak{K}(H, F)$ is therefore well defined for $j \leq \operatorname{rank}(A)$. Also note that we have $\operatorname{rank}(A) = |I|$ by construction.

Theorem 6.1 (Generalized Eckart–Young Theorem, operator norm)**.** *Let $A \in \mathfrak{K}(H, F)$ admit the SVD given in (6.1). Then for all $k < \operatorname{rank}(A)$, the operator given by the truncated SVD*

$$A_k := \sum_{i=1}^{k} \sigma_i(A)(u_i \otimes v_i) \tag{6.2}$$

satisfies the optimality property

$$A_k = \operatorname*{arg\,min}_{\substack{B \in \mathfrak{F}(H,F) \\ \operatorname{rank}(B) = k}} \|A - B\|.$$

Proof. Let $A \in \mathfrak{K}(H, F)$ admit the SVD in (6.1). Note that by Remark 2.11(iii) and the form of the operator A_k given by (6.2), we have

$$\|A - A_k\| = \sigma_{k+1}(A).$$

It now remains to show that the inequality $\|A - B\| \geq \sigma_{k+1}(A)$ holds for all operators $B \in \mathfrak{F}(H, F)$ with $\operatorname{rank}(B) = k$. Since $H = \ker(B) \oplus \ker(B)^\perp$ and $\dim \ker(B)^\perp = k$, there exists a nonzero element $x \in \operatorname{span}\{v_i\}_{1 \leq i \leq k+1} \cap \ker(B)$. We use Theorem 2.1(ii) and without loss of generality, we assume

$$\|x\|_H^2 = \sum_{i=1}^{k+1} \langle x, v_i \rangle_H^2 = 1.$$

By the definition of the operator norm, we now formulate

$$\begin{aligned} \|A - B\|^2 \geq \|(A - B)x\|_H^2 \geq \|Ax\|_H^2 &= \sum_{i=1}^{k+1} (\sigma_i \langle x, v_i \rangle_H)^2 \\ &\geq \sigma_{k+1}^2(A) \|x\|_H^2 = \sigma_{k+1}^2(A), \end{aligned}$$

which proves the statement. □

Theorem 6.1 offers the possibility to obtain bounds on singular values of compact operators which are perturbed by finite-rank operators. For the case of finite-dimensional spaces, similar statements are classical results for the eigenvalues of Hermitian matrices. They are known as *Weyl's inequalities* in matrix analysis (see Weyl 1912 for the original formulation and Bhatia 1997 for an overview). As we will now see, these inequalities generalize naturally to singular values of compact operators.

Corollary 6.2 (Gohberg and Krein 1969, Corollary 2.1)**.** *Let the operator $A \in \mathfrak{K}(H, F)$ admit the SVD in (6.1). For any operator $B \in \mathfrak{F}(H, F)$ with $j := \operatorname{rank}(B)$, we have*

$$\sigma_{j+1}(A) \geq \sigma_{j+k+1}(A + B) \tag{6.3}$$

for all $0 \leq j$ and $0 \leq k < \operatorname{rank}(A)$ such that $j + k < \operatorname{rank}(A + B)$ as well as

$$\sigma_{j+1}(A + B) \geq \sigma_{j+k+1}(A) \tag{6.4}$$

for all $0 \leq j$ and $0 \leq k < \operatorname{rank}(A + B)$ such that $j + k < \operatorname{rank}(A)$.

Proof. Let $B \in \mathfrak{F}(H, F)$ with $0 \leq \operatorname{rank}(B) =: j$ and $0 \leq k < \operatorname{rank}(A)$ such that $j + k < \operatorname{rank}(A + B)$. Let furthermore $A_k \in \mathfrak{F}(H, F)$ be the rank k approximation of A given by (6.2). As in the proof of Theorem 6.1, we obtain

$$\sigma_{k+1}(A) = \|A - A_k\| = \|(A + B) - (B + A_k)\|. \tag{6.5}$$

We see that $B + A_k$ is a finite-rank operator with a rank of at most $j + k$. We now apply the reasoning from the proof of Theorem 6.1 to (6.5) and obtain

$$\|(A + B) - (B + A_k)\| \geq \min_{\substack{C \in \mathfrak{F}(H,F) \\ \operatorname{rank}(C) \leq j+k}} \|(A + B) - C\| = \sigma_{j+k+1}(A + B),$$

which proves (6.3). Since A and $A + B$ are both compact operators, we can swap their roles and prove (6.4) analogously. □

Corollary 6.3 (Generalized Eckart–Young Theorem, Hilbert–Schmidt norm)**.** *Let $A \in HS(H, F)$ admit the SVD given in (6.1). Then for all $0 \leq k < \operatorname{rank}(A)$, the operator given by the truncated SVD*

$$A_k = \sum_{i=1}^{k} \sigma_i(A)(u_i \otimes v_i)$$

satisfies the optimality property

$$A_k = \operatorname*{arg\,min}_{\substack{B \in \mathfrak{F}(H,F) \\ \operatorname{rank}(B)=k}} \|A - B\|_{HS}\,.$$

Proof. We note that by Remark 2.18, we have

$$\|A - A_k\|_{HS}^2 = \sum_{i \in I \setminus \{1,\dots,k\}} \sigma_i^2(A). \tag{6.6}$$

Let $B_k \in \mathfrak{F}(H, F)$ be any operator of rank k. We now apply Corollary 6.2. By setting $B := -B_k$ in (6.4), we get

$$\sigma_i(A - B_k)^2 \geq \sigma_{i+k}(A)^2 \tag{6.7}$$

for $1 \leq i \leq \operatorname{rank}(A - B_k)$ and $i + k \leq \operatorname{rank}(A)$. We remark that (6.7) is simply a version of (6.4) with a shifted index. We are now able to see that for an index set $J = \{1, 2, \dots\}$ corresponding to the SVD of $A - B_k$, we get

$$\|A - B_k\|_{HS}^2 = \sum_{j \in J} \sigma_j^2(A - B_k) \geq \sum_{j \in J} \sigma_{j+k}^2(A) = \sum_{i \in I \setminus \{1,\dots,k\}} \sigma_i^2(A),$$

which proves the statement in combination with (6.6). □

6.2. SVD of Kernel Transfer Operators

As briefly described in Section 5.3, the eigenfunctions of transfer operators contain information about characteristic structures in the dynamics of the associated system. Therefore, several approaches to numerically approximate transfer operators based on a finite set of observations of the dynamics have emerged. Methods like *dynamic mode decomposition* (DMD, see Schmid 2010; H. Tu et al. 2014) and *time-lagged independent component analysis* (TICA, also *time-structure based independent component analysis*, see Molgedey and G. Schuster 1994; Pérez-Hernández et al. 2013) rely on the eigendecomposition of a linear approximation of the Perron–Frobenius operator and the Koopman operator through estimates of covariance- and autocovariance matrices. Over time, DMD and TICA have been generalized into nonlinear frameworks based on function spaces called *extended dynamic mode decomposition* (EDMD, see Williams, Kevrekidis and Rowley 2014) and *variational approach of conformation dynamics* (VAC, see Noé and Nüske 2013; Nüske et al. 2014) as well as the RKHS-based methods *kernel EDMD* (Kevrekidis, Rowley and Williams 2016) and *kernel TICA* (Schwantes and Pande 2015). For a general overview of transfer operator approximation techniques and the relations between all of the previously mentioned methods, we refer the reader to Klus, Koltai and Schütte 2016 and Klus, Nüske, et al. 2018. As the Koopman

operator and the Perron–Frobenius operator are closely related, so are all of the aforementioned methods. They are mainly used to identify and interpret dominant structures in the dynamics. Furthermore, they are applied to reduce the dimensionality of the system by projecting the dynamics onto a subset of kinetic coordinates spanned by the eigenfunctions of the dominant eigenvalues. We now give a brief overview of how the SVD of transfer operators is connected to the eigendecomposition-based methods. We summarize a general experimental approach regarding the simulation of stochastic processes to obtain empirical data for the estimation of transfer operators. We exemplarily perform numerical experiments on simulation data of two stochastic processes: the double-well potential and the stochastic Lorenz system.

6.2.1. Interpretation of the Transfer Operator SVD

As described in Section 5.3, a major part of the general theory of transfer operator eigendecompositions is based on the assumptions that

(i) the system is time-homogeneous and admits a stationary density π such that we are able to describe the dynamics with respect to a reference density in terms of the transfer operator $\mathcal{T}_\tau$ as given in Definition 5.13,

(ii) the system is reversible in the sense of the detailed balance condition (5.7) such that the transfer operator $\mathcal{T}_\tau$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$ and

(iii) we have access to trajectory data that is sampled with respect to the invariant density such that we are able to estimate $\mathcal{T}_\tau$.

Assumption (i) ensures that $\mathcal{T}_\tau$ is well defined and assumption (ii) justifies to write

$$\mathcal{T}_\tau = \sum_{i=1}^{\infty} \lambda_i (\phi_i \otimes \phi_i)$$

for positive eigenvalues $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \dots$ and corresponding orthogonal eigenfunctions $\phi_i \in L^2_\pi(\mathbb{X})$. Here, $L^2_\pi(\mathbb{X})$ is the space of equivalence classes of square Lebesgue-integrable functions with respect to the norm induced by $\langle \cdot, \cdot \rangle_\pi$ as given in Section 5.3. We then heuristically identify a so-called spectral gap in the magnitude of the eigenvalues in the sense that $1 - \lambda_r \ll \lambda_r - \lambda_{r+1}$ for an $r > 1$. By the generalized Eckart-Young theorem from Section 6.1, we obtain the low-rank approximation

$$T_\tau := \sum_{i=1}^{r} \lambda_i (\phi_i \otimes \phi_i) = \underset{\substack{B \in \mathfrak{F}(L^2_\pi(\mathbb{X})) \\ \operatorname{rank}(B) = r}}{\arg\min} \|\mathcal{T}_\tau - B\| \tag{6.8}$$

as the eigendecomposition and the SVD of $\mathcal{T}_\tau$ coincide in this case. That is, the dynamics in the subspace $\operatorname{span}\{\phi_i\}_{1 \leq i \leq r}$ under the rank-reduced operator $T_\tau \approx \mathcal{T}_\tau$ are the most dominant r-dimensional characteristics of the system (Koltai, Wu, et al. 2018). The eigenfunctions have been used to identify so-called *metastable states* in the case of time-homogeneous dynamics (see for example Schütte and Sarich 2013). Metastable states of the system can be interpreted as the sets in the state space in which trajectories tend to stay for a longer period of time relative to the overall dynamics. Hence, the transitions between metastable states can be considered as rare

events or "slow" components of the dynamics. As a general intuition, the eigenfunctions can be approximately represented by a linear combination of characteristic functions over the metastable sets in the state space. For a more detailed discussion concerning the relation of eigenfunctions and metastable states, see Koltai, Wu, et al. 2018.

However, if one or more of the assumptions (i)–(iii) are not satisfied, for instance in the case of non-reversible dynamics or time-inhomogeneity, it is natural that the eigendecomposition of the estimated transfer operators loses its low-rank approximation optimality properties and (6.8) does in general not hold true. In this case, the singular value decomposition of the given transfer operator may be used and the singular functions replace the previously considered eigenfunctions (Wu and Noé 2017). Furthermore, if we estimate a transfer operator based on data sampled with respect to an arbitrary empirical density $\rho_0 \neq \pi$ at an initial time t_0, the empirical density ρ_1 at a time $t_1 > t_0$ might look completely different. Therefore, the information given by the eigenfunctions or singular functions may be biased based on ρ_0 and ρ_1. Generalizing the concept of the transfer operator $\mathcal{T}_\tau$ with respect to the invariant density, we can define the *adapted transfer operator* (Denner 2017; Koltai, Wu, et al. 2018) for $\tau := t_1 - t_0$ and set

$$\begin{aligned} \mathcal{V}_{\tau,t_0} &: L^2_{\rho_0}(\mathbb{X}) \to L^2_{\rho_1}(\mathbb{X}) \\ \mathcal{V}_{\tau,t_0} u &:= \frac{1}{\rho_1} \mathcal{P}_\tau(u\rho_0), \end{aligned}$$

where $\rho_1 := \mathcal{P}_{\tau,t_0}\,\rho_0$ is the propagated empirical density from the initial time t_0 to t_1 and $\mathcal{P}_{\tau,t_0}$ is the analytic Perron–Frobenius operator at time t_0. The adapted transfer operator $\mathcal{V}_{\tau,t_0}$ preserves some of the aforementioned properties of $\mathcal{T}_\tau$ in the sense that $\mathcal{V}_{\tau,t_0}\mathbb{1}_\mathbb{X} = \mathbb{1}_\mathbb{X}$ and its leading singular value is $\sigma_1 = 1$. As the adjoint of the adapted transfer operator, we obtain an appropriate reformulation of the Koopman operator $\mathcal{K}_{\tau,t_0} : L^2_{\rho_1}(\mathbb{X}) \to L^2_{\rho_0}(\mathbb{X})$ with $\mathcal{K}_{\tau,t_0} g(x) = \mathbb{E}[g(X_{t_1}) \,|\, X_{t_0} = x]$. We hence get a time-parametrized family of adapted transfer operators that describe the dynamics. If the dynamics are time-inhomogeneous, the notion of metastability loses its significance, since the previously considered rare transitions can happen in between sets that move and disperse over time. As a more general concept, the notion of *finite-time coherent sets* (Froyland, Santitissadeekorn and Monahan 2010; Froyland 2013; Koltai, Ciccotti and Schütte 2016) was introduced. Coherent sets can be interpreted as minimally dispersive sets under the dynamics with respect to a finite evolution in time. If the dynamics are time-inhomogeneous, coherent sets can be thought of as a generalization of the concept of metastability.

For a general overview of approaches to model nonreversible systems and time-inhomogeneous dynamics with transfer operators, see Koltai, Wu, et al. 2018. The technical background of metastable and coherent structures and their detailed connection to eigendecompositions and singular value decompositions of transfer operators as well as the optimal approximation of transfer operators is beyond the scope of this work. Here, we will give numerical examples to illustrate the estimation of the kernel transfer operators for time-homogeneous systems as defined in Section 5.3 and show how the eigendecompositions and singular value decompositions of the resulting empirical RKHS operators can be applied.

6.2.2. Experimental Problem Setting

We consider a time-homogeneous stochastic differential equation of the form

$$\mathrm{d}X_t = f(X_t)\mathrm{d}t + g(X_t)\,\mathrm{d}W_t \tag{6.9}$$

on the state space $\mathbb{X} \subseteq \mathbb{R}^d$. Here, W_t is standard Brownian motion. To obtain a discrete numerical solution of (6.9), we may define the Euler–Maruyama discretization step map

$$\xi_h x_t := x_{t+h} = x_t + hf(x_t) + g(x_t)(W_{t+h} - W_t), \tag{6.10}$$

for a step size $h > 0$ and a random variable $(W_{t+h} - W_t) \sim \mathcal{N}(0, h)$. We choose a kernel $k(\cdot, \cdot)$ corresponding to a feature map φ. By fixing a sample shift number $s \in \mathbb{N}$ and considering observations $\{x^{(1)}, \ldots, x^{(n)}\} \subseteq \mathbb{X}$, we can define the feature matrices

$$\begin{aligned}\Phi &:= [\varphi(x^{(1)}), \ldots, \varphi(x^{(n)})],\\ \Psi &:= [\varphi(\xi_h^s x^{(1)}), \ldots, \varphi(\xi_h^s x^{(n)})].\end{aligned}$$

Here, $\xi_h^s x^{(i)}$ denotes the propagated observation obtained by iteratively applying the discretization step map s times. In particular, the data in Ψ represents a timeshifted discrete system of the data in Φ for a lag time $\tau := hs$. In general, the sampling of the data $\{x^{(1)}, \ldots, x^{(n)}\}$ is not trivial and can have a major impact on the model and its interpretation. We can for example

1. obtain the data by simulating one or several long trajectories of the system by iteratively applying the discretization step (6.10) such that $x^{(i+1)} = \xi_h x^{(i)}$. We can then choose Φ and Ψ to be contiguous blocks of data in the trajectories each shifted by a window of s consecutive observations or

2. uniformly sample the data $\{x^{(1)}, \ldots, x^{(n)}\}$ from a relevant domain in the state space and manually propagate each observation independently by applying ξ_h^s.

When me make use of the first option, we try to sample from an approximation of the stationary density of the system (6.9), assuming it exists. The latter alternative gives the possibility to observe a short time period of the system with respect to the uniform distribution on the state space, if the number of discretization steps is small enough with respect to the amount of dispersion of the observations under the dynamics. To approximate $\mathcal{T}_\tau$ and $\mathcal{K}_\tau$, one often simulates long trajectories as in the first approach. To approximate $\mathcal{P}_\tau$, one typically tries to simulate many short trajectories using the second approach (Klus, Koltai and Schütte 2016). In general, the eigenfunctions and singular functions of the transfer operators have to be interpreted with respect to the empirical distribution according to which the data was sampled. For an overview of alternative approaches to model the dynamics, if the available data is not sampled in a non-stationay scenario, see Wu and Noé (2017) and Koltai, Wu, et al. (2018).

Having defined the feature matrices Φ and Ψ, we can estimate the kernel transfer operator

$$\widehat{\mathcal{A}}_\tau^{(k)} = \Psi A \Phi^T$$

with a choice of A as either $A_1 = G_{\Phi\Psi}^{-1} G_{\Phi\Phi}^{-1} G_{\Phi\Psi}$ or $A_2 = G_{\Phi\Psi}^{-1} G_{\Psi\Phi}^{-1} G_{\Psi\Psi}$ and the kernel Koopman operator

$$\widehat{\mathcal{K}}_\tau^{(k)} = \Phi G_{\Phi\Phi}^{-1} \Psi^T$$

as discussed in Section 5.3. Their eigenfunctions and singular functions can be represented as linear combinations of the elements in Φ and Ψ as shown in Proposition 4.6 and Proposition 4.11.

Possible Drawbacks. The eigenfunctions and singular functions of all kernel transfer operators are elements in $\operatorname{span}\{\varphi(x^{(i)})\}_{1\leq i\leq n}$ or $\operatorname{span}\{\varphi(\xi_h^s x^{(i)})\}_{1\leq i\leq n}$, respectively, depending on the choice of the formulation of the auxiliary problem of Proposition 4.6 and Remark 4.7. This fact leads to the dilemma that the sampling of the observations $x^{(i)}$ and the effect of the dynamics on the empirical data given by the propagated observations $\xi_h^s x^{(i)}$ potentially restrict our solution space and influence the interpretability of the experiments. As an example, we might use a radial basis kernel and decompose the kernel transfer operator $\widehat{\mathcal{T}}_\tau^{(k)}$ with respect to the invariant density into its singular components. We expect the dominant left and right singular functions to be the constant functions $\mathbb{1}_{\mathbb{X}}$. Of course, the propagated observations $\xi_h^s x^{(i)}$ do not necessarily cover the domain $\mathbb{X}$ in a sense that we can reliably approximate $\mathbb{1}_{\mathbb{X}}$ with elements from the solution space $\operatorname{span}\{\varphi(\xi_h^s x^{(i)})\}_{1\leq i\leq n}$. The effect of the discrete dynamics ξ_h^s on the data might lead to empirical density peaks in some regions of the state space and also to regions which are covered scarcely by observations. Additionally, the geometries of the RKHS $\mathscr{H}$ and the originally considered space $L_\rho^2(\mathbb{X})$ for a reference density ρ might clash in the sense that scaling elements to unit length with respect to $\|\cdot\|_{\mathscr{H}}$ differs from scaling with respect to $\|\cdot\|_{L_\rho^2(\mathbb{X})}$. As a result, the eigenfunctions or singular functions should not be compared by the magnitude of their exact function values but rather by the geometric shapes of their level sets. When we analyze estimated kernel transfer operators, different choices of the sampling of the observations $x^{(i)}$ not only affect the reference distribution of the given operator, but ultimately determine the resulting solution space.

6.2.3. Example: Double-Well Potential

In this example, we consider the time-homogeneous and reversible double-well problem on $\mathbb{R}^2$ given by the stochastic differential equation

$$\begin{cases} \mathrm{d}X_t &= -\nabla_X V(X_t, Y_t)\,\mathrm{d}t + \epsilon\,\mathrm{d}W_t^{(1)} \\ \mathrm{d}Y_t &= -\nabla_Y V(X_t, Y_t)\,\mathrm{d}t + \epsilon\,\mathrm{d}W_t^{(2)}, \end{cases} \tag{6.11}$$

for the potential $V(x, y) = (x^2 - 1)^2 + y^2$ and two independent standard Wiener processes $W_t^{(1)}$ and $W_t^{(2)}$. For our example, we choose $\epsilon = 0.7$. The solutions to the double-well problem exhibit metastable behaviour around the two wells of $V(x, y)$ at $(1, 0) \in \mathbb{R}^2$ and $(-1, 0) \in \mathbb{R}^2$. Trajectories usually stay close to the related metastable point while a transfer between the two states is observed as a rare event. For a visualization of a typical trajectory of the double-well system, see Figure 1.

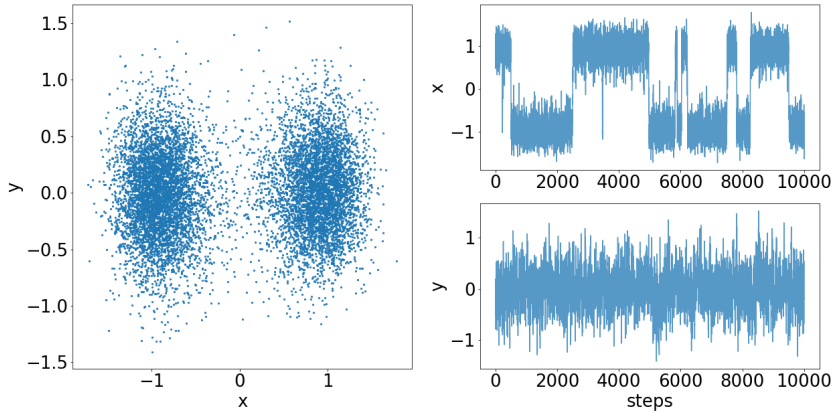

Figure 1: Example trajectory of the double-well problem given in (6.11). The x-coordinate component of the trajectory exhibits metastable behaviour and occasionally switches between the two wells. The y-coordinate shows Gaussian noise.

We choose the Gaussian radial basis kernel $k(x, y) = \exp(-\gamma \|x - y\|_2)$ with $\gamma = 0.3$ and independently sample $n = 10000$ observations $z^{(i)} = (x^{(i)}, y^{(i)}) \in \mathbb{R}^2$ from the uniform distribution on $[-2, 2] \times [-2, 2] \subseteq \mathbb{R}^2$. We construct $\Phi = [\varphi(z^{(1)}), \ldots, \varphi(z^{(n)})]$ from the initial data and set $\Psi = [\varphi(\xi_h^s z^{(1)}), \ldots, \varphi(\xi_h^s z^{(n)})]$, where ξ_h^s is the Euler–Maruyama discretization step for the system (6.11) with step size $h = 0.1$ and progagation step number $s = 20$. The estimate

$$\widehat{\mathcal{P}}_\tau^{(k)} = \Psi A \Phi^T$$

for $A = G_{\Phi\Psi}^{-1} G_{\Phi\Phi}^{-1} G_{\Phi\Psi}$ yields the empirical kernel Perron–Frobenius operator and by Proposition 4.6 we obtain the eigenfunctions $u_i = \Psi \mathbf{w}_i$ of $\widehat{\mathcal{P}}_\tau^{(k)}$ by solving the auxiliary problem

$$A \Phi^T \Psi \mathbf{w}_i = \lambda_i \mathbf{w}_i$$

for eigenvectors $\mathbf{w}_i$. We obtain the dominant eigenvalue $\lambda_1 \approx 1$ corresponding to the $\mathscr{H}$-representation of the approximated stationary density u_i of the system. For the second eigenvalue $\lambda_2 \approx 0.694$, the eigenfunction u_2 exhibits the characteristic separation of the two metastable states of the system. All other eigenvalues can be numerically considered as zero. For visualizations of the eigenfunctions u_1 and u_2, see Figure 2. These results are in agreement with the classical results for transfer operator eigendecompositions of the double well system (Klus, Koltai and Schütte 2016).

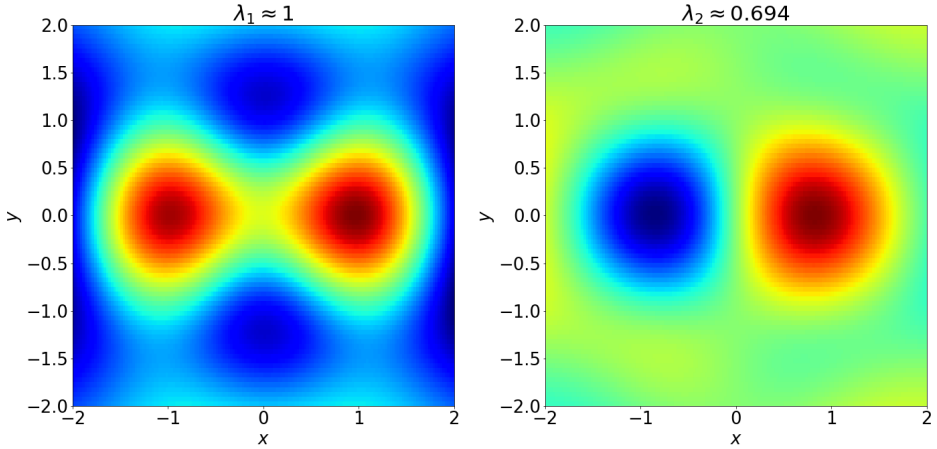

Figure 2: Eigenfunctions of the empirical kernel Perron–Frobenius operator $\widehat{\mathcal{P}}_\tau^{(k)}$ of the double well system. The left eigenfunction u_1 corresponds to the empirical stationary density represented in the RKHS $\mathscr{H}$, the right eigenfunction u_2 characterises the dynamics of the state transfer in between the two metastable sets.

6.2.4. Example: Stochastic Lorenz System

In this example, we consider the stochastic Lorenz system (Lorenz 1963; Chekroun, Simonnet and Ghil 2013) given by the stochastic differential equation

$$\begin{cases} \mathrm{d}X_t &= s(Y_t - X_t)\,\mathrm{d}t + \epsilon X_t\,\mathrm{d}W_t^{(1)} \\ \mathrm{d}Y_t &= (rX_t - Y_t - X_t Z_t)\,\mathrm{d}t + \epsilon Y_t\,\mathrm{d}W_t^{(2)} \\ \mathrm{d}Z_t &= (-bZ_t + X_t Y_t)\,\mathrm{d}t + \epsilon Z_t\,\mathrm{d}W_t^{(3)} \end{cases} \tag{6.12}$$

for the parameter configuration $s = 10$, $r = 28$ and $b = 8/3$ and independent standard Wiener processes $W_t^{(i)}$. The deterministic system for $\epsilon = 0$ has been studied in various scientific fields as a classic example of a time-autonomous system exhibiting chaotic behaviour (Tucker 1999). It has previously been considered in the context of transfer operators (Froyland and Padberg 2009; Brunton et al. 2017; Wu and Noé 2017). The trajectories of the deterministic system are described by the switching in between two lobes, which are connected by a transition domain close to the saddle point at $(0, 0, 0) \in \mathbb{R}^3$. Its nondeterministic counterpart shows the same overall characteristics, only affected by random perturbation in each possible state. For example trajectories for the deterministic as well as the nondeterministic case, see Figure 3.

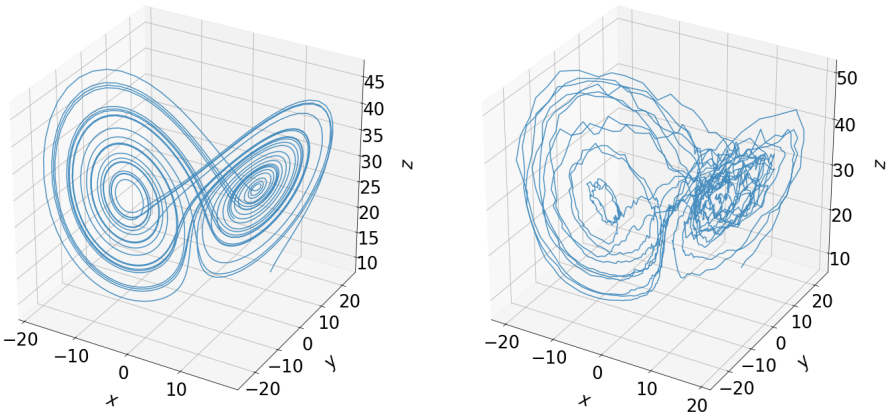

Figure 3: Example trajectories of the Lorenz system for the deterministic case with $\epsilon = 0$ (left) and the nondeterministic case with $\epsilon = 0.3$ (right).

We set $\epsilon = 0.3$ and simulate 10 trajectories of length 15 with $\tau = 0.75$ based on the Euler–Maruyama discretization step ξ_h from (6.10) for $h = 0.01$. We choose the Gaussian radial basis kernel $k(x, y) = \exp(-\gamma \|x - y\|_2)$ with $\gamma = 0.0025$ and compute the SVD of the estimated kernel Koopman operator $\widehat{\mathcal{K}}_\tau^{(k)} = \Phi B \Psi^T$ with $B = G_{\Phi\Phi}^{-1}$ as given in Section 5.3 by applying Proposition 4.11 and solving the auxiliary problem

$$B^T \Phi^T \Phi B \Psi^T \Psi \mathbf{w}_i = B^T \Psi^T \Psi \mathbf{w}_i = \sigma_i^2 \mathbf{w}_i.$$

We obtain the right singular functions

$$v_i := \|\Psi \mathbf{w}_i\|_{\mathscr{H}}^{-1} \, \Psi \mathbf{w}_i = (\mathbf{w}_i^T \Psi^T \Psi \mathbf{w}_i)^{-1/2} \, \Psi \mathbf{w}_i$$

and left singular functions

$$u_i := \sigma_i^{-1} \, \widehat{\mathcal{K}}_\tau^{(k)} v_i = \sigma_i^{-1} \, (\mathbf{w}_i^T \Psi^T \Psi \mathbf{w}_i)^{-1/2} \, \Phi B \Psi^T \Psi \, \mathbf{w}_i.$$

Note again that the singular functions are normalized with respect to the norm induced by the inner product of $\mathscr{H}$. For $\sigma_1 \approx 1$, we get an approximation of the constant function evaluated on the observations of used trajectories as right singular function. For $\sigma_2 \approx 0.823$ and $\sigma_3 \approx 0.721$, we obtain singular functions with complementary level sets (see Figure 4). They are in good agreement with the results on the almost-invariant manifolds and metastable sets given by Froyland and Padberg (2009), Brunton et al. (2017) and Wu and Noé (2017). We clearly observe a separation of two metastable structures, each containing an inner part of a lobe and the outer part of its opposing lobe. However, the left singular functions lack interpretability in all cases. This is presumably due to the bias of the eigenfunctions with respect to the τ-propagated empirical distribution of the initial data which we did not account for in the computation (Wu and Noé 2017).

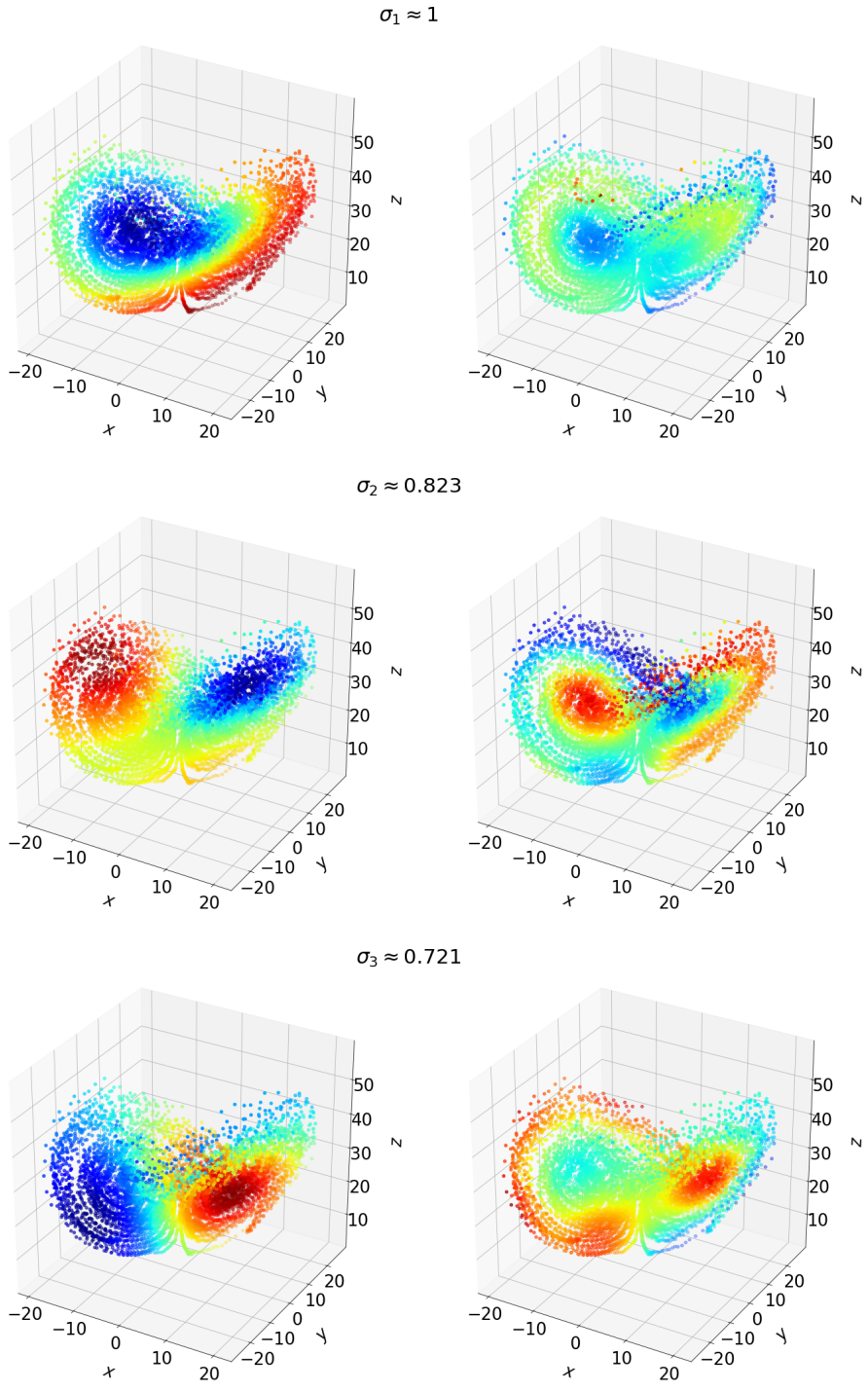

Figure 4: Dominant left singular functions (left) and right singular functions (right) of the empirical kernel Koopman operator of the stochastic Lorenz system evaluated on independently generated trajectory observations. The first right singular function v_1 can be interpreted as constant. The level sets of the right singular functions v_2 and v_3 can be used to distinguish between the two metastable components. All of the left singular functions lack interpretability.

6.3. Generalized Pseudoinverses

The general algebraic theory of generalized inverses of linear maps spawned a rich set of tools to tackle problems in various scientific fields. Results based on generalized inverses are for example used in numerical methods for regression problems, dynamical systems, Markov models and nonlinear equations. For a comprehensive review of the theory and practical applications of generalized inverses, we refer to Ben-Israel and Greville (2003). The *Moore–Penrose pseudoinverse* of a matrix (see Moore 1920; Penrose 1955) can analytically be expressed via the SVD of the given matrix. For a concise overview of the Moore–Penrose pseudoinverse of matrices and its properties, the reader may refer to Appendix A. In this section, we will generalize the theory of Moore–Penrose pseudoinverses of matrices to compact operators and show that we must impose certain regularity assumptions onto the considered class of operators to obtain a globally defined pseudoinverse operator.

The general problem one usually has in mind when considering generalized inverse operators is the following: Given a linear operator $A : H \to F$ and an element $y \in F$, we want to approximate a hypothetical optimal solution $x \in H$ of the not necessarily well-posed problem

$$Ax = y \tag{6.13}$$

if the operator A is not bijective. We will follow the general argumentation of Eubank and Hsing (2015) in this section to partly solve this problem for compact operators in terms of their SVD. Additionally, we will derive a natural interpretation of the term "optimality" of a solution of problem (6.13).

Lemma 6.4. *If $A \in \mathfrak{B}(H, F)$, then the restricted linear operator $\tilde{A}$ given by*

$$\begin{aligned} \tilde{A} : \ker(A)^\perp &\to \operatorname{range}(A), \\ x &\mapsto Ax \end{aligned}$$

is bijective. Furthermore, for its inverse we have $\tilde{A}^{-1} \in \mathfrak{B}(\operatorname{range}(A), \ker(A)^\perp)$.

Proof. The linearity of $\tilde{A}$ is clear. Surjectivity of $\tilde{A}$ onto $\operatorname{range}(A)$ directly follows by the fact that we do not exclude any element from the range of A by restricting A onto the orthogonal complement of its nullspace. Injectivity follows from the fact that we have $\ker(\tilde{A}) = \{0\}$ by construction. Since $\tilde{A}$ is bijective, its inverse exists. Boundedness of the inverse can be obtained as a consequence of the open mapping theorem as shown by Werner (2007, Corollary IV.3.4). □

Definition 6.5. *Let $A \in \mathfrak{B}(H, F)$. Then the linear operator*

$$A^+ : \operatorname{range}(A) + \operatorname{range}(A)^\perp \to H$$

given by

$$A^+ y := \begin{cases} \tilde{A}^{-1} y, & y \in \operatorname{range}(A) \\ 0, & y \in \operatorname{range}(A)^\perp \end{cases}$$

is called the (generalized) Moore–Penrose pseudoinverse *of A. Here $\tilde{A}$ is the restricted operator from Lemma 6.4.*

The Moore–Penrose pseudoinverse is defined on the set

$$\operatorname{range}(A) + \operatorname{range}(A)^{\perp} = \{w + v \,|\, w \in \operatorname{range}(A), v \in \operatorname{range}(A)^{\perp}\} \subseteq F.$$

If $\operatorname{range}(A)$ is closed, then we have the representation

$$\operatorname{range}(A) + \operatorname{range}(A)^{\perp} = \operatorname{range}(A) \oplus \operatorname{range}(A)^{\perp} = F, \tag{6.14}$$

so in this case A^{+} is defined on the whole space F. Note that Definition 6.5 only imposes the assumption of boundedness on A. The theory of pseudoinverse operators on Hilbert spaces can therefore be proceeded in this very general setting (see Ben-Israel and Greville 2003). However, we want to investigate the connection between the operator SVD and the Moore–Penrose pseudoinverse and hence add the assumption of compactness. We now examine the structure of $\operatorname{range}(A)$ in dependency of the SVD of $A \in \mathfrak{K}(H, F)$ and give sufficient conditions under which $\operatorname{range}(A)$ is closed such that (6.14) holds.

Proposition 6.6 (Picard condition, Eubank and Hsing 2015)**.** *Let $A \in \mathfrak{K}(H, F)$ have the SVD representation*

$$A = \sum_{i \in I} \sigma_i (u_i \otimes v_i).$$

Then every $y \in \operatorname{range}(A)$ satisfies the condition

$$\sum_{i \in I} \frac{\langle y, u_i \rangle_F^2}{\sigma_i^2} < \infty.$$

Proof. Let $y \in \operatorname{range}(A)$, hence for an element $x \in \overline{\operatorname{span}\{v_i\}_{i \in I}}$, we have $Ax = y$. By the SVD of A, we have $y \in \overline{\operatorname{span}\{u_i\}_{i \in I}}$. By writing $Ax = y$ in terms of the SVD of A, we have

$$y = \sum_{i \in I} \sigma_i \langle x, v_i \rangle_H u_i. \tag{6.15}$$

By expanding y in terms of Theorem 2.1(i), we have

$$y = \sum_{i \in I} \langle y, u_i \rangle_F u_i. \tag{6.16}$$

Therefore, we see that by comparison of (6.15) and (6.16) and the uniqueness of the ONS expansion coefficients, it must hold that $\langle y, u_i \rangle_F = \sigma_i \langle x, v_i \rangle$ for all $i \in I$. Since x as an element of H must satisfy $\|x\|_H^2 < \infty$, we conclude

$$\|x\|_F^2 = \sum_{i \in I} \langle x, v_i \rangle_H^2 = \sum_{i \in I} \frac{\langle y, u_i \rangle_F^2}{\sigma_i^2} < \infty$$

by using Theorem 2.1(ii). □

Corollary 6.7. *The range of a compact operator A is closed if and only if A is a finite-rank operator.*

Proof. Let $A \in \mathfrak{K}(H, F)$. If the rank of A is finite, then its range is a finite-dimensional subspace of F and hence closed. If the rank of A is infinite, then A admits a zero sequence of singular values which we denote by $(\sigma_i)_{i\in\mathbb{N}}$. We may therefore find a subsequence $(\sigma_{i_k})_{k\in\mathbb{N}}$ that satisfies

$$\sigma_{i_k} < \frac{1}{k^2}$$

for all $k \in \mathbb{N}$. We now construct an element $y \in F$ as

$$y = \sum_{k=1}^{\infty} \sigma_{i_k} u_{i_k},$$
$$\|y\|_F^2 = \sum_{k=1}^{\infty} \sigma_{i_k}^2 < \infty.$$

Since $u_{i_k} \in \operatorname{range}(A)$ for every $k \in \mathbb{N}$, we have $y \in \overline{\operatorname{range}(A)}$. Since y violates the Picard condition in the sense that the series

$$\sum_{i=1}^{\infty} \frac{\langle y, u_i\rangle_F^2}{\sigma_i^2} = \sum_{k=1}^{\infty} \frac{\sigma_{i_k}^2}{\sigma_{i_k}^2}$$

diverges, we see that y can not be an element of $\operatorname{range}(A)$ by Proposition 6.6, proving that $\operatorname{range}(A) \neq \overline{\operatorname{range}(A)}$. Therefore, $\operatorname{range}(A)$ is not closed. □

Corollary 6.7 leads to the important observation that the Moore–Penrose pseudoinverse A^+ of a bounded operator $A : H \to F$ is globally defined on the space F if and only if A is finite-rank. Based on Definition 6.5, we immediately see that for a compact operator A with an SVD as given in Theorem 6.6 and $y \in \operatorname{range}(A)$, the element $A^+y \in H$ takes the form

$$A^+y := \sum_{i\in I} \frac{1}{\sigma_i} \langle y, u_i\rangle_F \, v_i. \tag{6.17}$$

Note that Proposition 6.6 ensures that A^+y as given in (6.17) is well defined. If A is finite-rank with the SVD

$$A = \sum_{i=1}^{r} \sigma_i (u_i \otimes v_i),$$

then the Moore–Penrose pseudoinverse is a finite-rank operator defined on F and is naturally given by the SVD

$$A^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} (v_i \otimes u_i). \tag{6.18}$$

We are equipped with the general results on the existence of the Moore–Penrose pseudoinverse and its connection to the SVD of compact operators. We now examine the initial inverse problem given in (6.13) and its connection to the Moore–Penrose pseudoinverse.

Theorem 6.8 (Eubank and Hsing 2015, Theorem 3.5.10)**.** *Let $A \in \mathfrak{B}(H, F)$ and $y \in \operatorname{range}(A) + \operatorname{range}(A)^{\perp}$. Then we have*

$$A^{+}y = \operatorname*{arg\,min}_{x \in H} \|Ax - y\|_F . \tag{6.19}$$

Furthermore, $A^{+}y$ is the unique element in H with minimal norm satisfying (6.19).

We emphasize that the proof of Theorem 6.8 as given by Eubank and Hsing is only based on the very general formulation of Definition 6.5 and hence also holds for the case of noncompact bounded operators. However, in the case of A being a compact operator, we may use (6.17) to solve regression problems of the form (6.19) for feasible $y \in \operatorname{dom}(A^{+})$. If A is finite-rank, we may directly use the analytical SVD form of A^{+} given by (6.18) to solve the regression problem for all $y \in F$.

7. Conclusion

In this thesis, we examined linear operators on reproducing kernel Hilbert spaces from a spectral theoretic point of view. By combining classical results from functional analysis on the eigendecomposition and the SVD of compact operators with the characteristic properties of RKHSs, we provided the general framework for the mathematical intepretation of the SVD of finite-rank RKHS operators. We showed that the SVD of finite-rank RKHS operators can be computed numerically by solving an auxiliary matrix eigenvalue problem. To underline the practical importance and versatility of RKHS operators, we gave concrete examples in terms of kernel covariance operators, conditional mean embedding operators and kernel transfer operators.

As shown, finite-rank estimates of these operators arise naturally in practical applications. As an exemplary scenario, we simulated two classical stochastic differential equations, the double-well potential and the stochastic Lorenz system. We numerically analyzed estimates of their associated kernel transfer operators by computing their dominant eigencomponents and singular components in view of the classical interpretation of transfer operator-based methods in fluid dynamics and molecular dynamics. We furthermore provided general results related to the SVD of compact operators in terms of the low-rank operator approximation and generalized Moore–Penrose pseudoinverse. These results transfer directly to the special case of the RKHS operator SVD and can therefore be used to extend the spectral theory of empirical RKHS operators in numerical applications.

Although we were able to provide the mathematical background of the RKHS operator theory, the practical realization of RKHS operator-based spectral decomposition methods still contains details of unknown or problematic nature and possible areas where further research is needed:

(I) The RKHS operator SVD ultimatively relies on the RKHS operator eigendecomposition. To compute the SVD of an empirical RKHS operator S, we computed the eigendecomposition of the empirical RKHS operator S^*S. In order to obtain the SVD of S directly, more effective methods may be derived by generalizing iterative computation schemes or block operator reformulations for the matrix SVD to the RKHS setting. In addition, a more direct approach may yield a more numerically robust scheme. Since the auxiliary problem related to the eigendecomposition of S^*S contains several matrix multiplications, its numerical condition may be ineligible and hardly assessable for practical use. When using a more direct method, the matrix multiplications in the auxiliary problem may be bypassed.

(II) By working with empirical RKHS operators, we naturally restricted ourselves to finite-dimensional subspaces of the form $\operatorname{span}\{\varphi(x_1), \ldots, \varphi(x_n)\} \subseteq \mathscr{H}$ as solution spaces for the resulting eigenfunctions and singular functions. When RKHS operators are used to extend mathematical frameworks such as transfer operators, this restricted solution space can have an impact on the interpretability of the method. Based on the desired interpretation of the resulting singular values and singular functions, a sampling bias of the available data might influence the approximative qualities of the solution significantly. In particular, we saw that the singular functions of the Koopman operator associated with the stochastic Lorenz system may be hard to interpret, since the geometry of

the finite-dimensional solution space can differ from the original solution space. Hence, consistency results about the numerical approximation of solutions for RKHS operator-based methods in terms of the finite-dimensional solution spaces have to be derived individually for specific application scenarios.

(III) By Theorem 6.8, the generalized Moore–Penrose pseudoinverse allows the computation of an optimal solution of the regression problem

$$\arg\min_{h \in \mathscr{H}} \|Sh - f\|_{\mathscr{F}}$$

for an empirical RKHS operator $S : \mathscr{H} \to \mathscr{F}$ and an element $f \in \mathscr{F}$. Combined with the numerical SVD of empirical RKHS operators, this could be a starting point for the development of RKHS operator regression approaches.

(IV) The combination of the low-rank operator approximation as given by Theorem 6.1 can be combined with the theory of the generalized Moore–Penrose pseudoinverses. The Moore–Penrose pseudoinverse of an operator $A \in \mathfrak{F}(H, F)$ of rank r may be approximated by the truncated SVD

$$A_k^+ = \sum_{i=1}^{k} \frac{1}{\sigma_i}(v_i \otimes u_i)$$

for $k < r$ with the singular values σ_i^{-1} ordered accordingly by their magnitude. In practice, the approximated pseudoinverse operator A_k^+ may be an alternative to classical regularization techniques in several applications. This approach transfers naturally to the case of empirical RKHS operators and their numerical SVD.

As a general focus, it remains to be examined how numerical methods in other scientific fields can be generalized in terms of the RKHS operator approach. Based on the work of Klus, Schuster and Muandet (2017), we gave a general overview of kernel transfer operators by translating the classical transfer operator theory into the RKHS setting. As a result, we can formulate kernel-based transfer operator methods as alternatives to the existing approaches. Similar connections may be discovered in other areas of research. We can for example use string-based and graph-based kernels to derive RKHS operator methods over more abstract and non-metric domains. The very general investigation of the mathematical background of RKHS operators in this thesis suggests that the RKHS operator theory may yield promising approaches in machine learning, fluid dynamics, signal processing, mechanical engineering and similar fields.

A. Appendix

This appendix serves as a brief overview of the singular value decomposition of real matrices and its properties as well as the Moore–Penrose pseudoinverse. For details and proofs, the reader may refer to Golub and Van Loan (2013) and Ben-Israel and Greville (2003).

Theorem A.1 (Singular value decomposition)**.** *Let $A \in \mathbb{R}^{n\times m}$. Then there exist orthogonal matrices $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{m\times m}$ and such that*

$$U^T A V = \Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} \in \mathbb{R}^{n\times m}, \tag{A.1}$$

where $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

Remark A.2. The σ_i are called the *singular values* of A. For $i \leq r$, we call the column $\mathbf{v}_i$ of V corresponding to the singular value σ_i the i-th *right singular vector* of A and the column $\mathbf{u}_i$ of U corresponding to σ_i the i-th *left singular vector* of A, respectively. Theorem A.1 is equivalent to the assertion that every matrix $A \in \mathbb{R}^{n\times m}$ can be represented by orthonormal systems of column vectors $\{\mathbf{u}_i\}_{1\leq i\leq r} \subseteq \mathbb{R}^n$ and $\{\mathbf{v}_i\}_{1\leq i\leq r} \subseteq \mathbb{R}^m$ as

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

for an $r \leq \min\{m, n\}$ and real numbers $\sigma_1 \geq \cdots \geq \sigma_r > 0$. From this representation one can also deduce the properties

$$\begin{aligned} \operatorname{rank}(A) &= r, \\ \operatorname{range}(A) &= \operatorname{span}\{\mathbf{u}_i\} \text{ and} \\ \ker(A) &= (\operatorname{span}\{\mathbf{v}_i\})^{\perp}, \end{aligned}$$

as well as the *spectral norm* $\|A\| = \sigma_1$ as a special case of the operator norm and the *Frobenius norm* as a special case of the Hilbert–Schmidt norm

$$\|A\|_{HS} = \big(\sum_{i=1}^{r} \sigma_i^2\big)^{1/2}.$$

Theorem A.3 (Eckart and Young 1936/Mirsky 1960)**.** *Let $A \in \mathbb{R}^{n\times m}$ admit the SVD given in (A.1) and $k \leq \operatorname{rank}(A)$. If $\|\cdot\|_*$ is either the Frobenius norm on $\mathbb{R}^{n\times m}$ or the spectral norm on $\mathbb{R}^{n\times m}$, then*

$$\sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \operatorname*{arg\,min}_{\substack{B\in\mathbb{R}^{n\times m} \\ \operatorname{rank}(B)=k}} \|A - B\|_* . \tag{A.2}$$

Definition A.4 (Moore 1920/Penrose 1955)**.** *If $A \in \mathbb{R}^{n\times m}$ admits the singular value decomposition given in (A.1), then we define the matrix $A^+ \in \mathbb{R}^{m\times n}$ by*

$$A^+ = V^T \Sigma^+ U,$$

where Σ^+ is the matrix obtained by filling the first r diagonal entries of the $m \times n$ zero matrix with the inverse singular values σ_i^{-1} of A. We call A^+ the Moore–Penrose pseudoinverse *or simply* pseudoinverse *of A.*

Remark A.5. The Moore–Penrose pseudoinverse generalizes the matrix inverse to non-square and non-invertible matrices in the sense that for $A \in \mathbb{R}^{n\times m}$ we have

(i) $AA^+A = A$,

(ii) $A^+AA^+ = A^+$,

(iii) AA^+ and A^+A are symmetric,

(iv) $(A^+)^+ = A$ and

(v) $A^+ = A^{-1}$ if $n = m$ and A is invertible.

Lemma A.6 (Least squares optimal solution, Golub and Van Loan 2013, 5.5.2)**.** *Let $A \in \mathbb{R}^{n\times m}$. Then $A^+ \in R^{m\times n}$ is the unique solution of minimal norm to the minimization problem*

$$\min_{B\in\mathbb{R}^{m\times n}} \|AB - I_n\|_F .$$

Additionally, for arbitrary $b \in \mathbb{R}^n$, the vector $A^+b \in \mathbb{R}^m$ is the unique solution of minimal norm to the minimization problem

$$\min_{x\in\mathbb{R}^m} \|Ax - b\|_2^2 .$$

References

Aizerman, A., E. Braverman and L. Rozoner (1964). "Theoretical foundations of the potential function method in pattern recognition learning". In: *Autom. Remote Control* 25, pp. 821–837.

Altun, Y. and A. Smola (2006). "Unifying Divergence Minimization and Statistical Inference via Convex Duality". In: *Proceedings of the International Conference on Computational Learning Theory (COLT)*, pp. 139–153.

Aronszajn, N. (1950). "Theory of reproducing Kernels". In: *Transactions of the American Mathematical Society* 68, pp. 337–404.

Bach, F. (2017). "On the Equivalence Between Kernel Quadrature Rules and Random Feature Expansions". In: *Journal of Machine Learning Research* 18.1, pp. 714–751.

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

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

Baxter, J. R. and J. S. Rosenthal (1995). "Rates of convergence for everywhere- positive Markov chains". In: *Statistics & probability letters* 22, pp. 333–338.

Ben-Israel, A. and T. Greville (2003). *Generalized Inverses: Theory and Applications*. 2nd ed. Springer.

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

Bhatia, R. (1997). *Matrix Analysis*. Springer.

Boser, B. E., I. M. Guyon and V. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92*, p. 144.

Boyd, S. and L. Vandenberghe (2009). *Convex Optimization*. 7th ed. Cambridge University Press.

Brunton, S. L. et al. (2017). "Chaos as an intermittently forced linear system". In: *Nature Communications* 8, p. 19.

Budiŝić, M., R. Mohr and I. Mezić (2012). "Applied Koopmanism". In: *Chaos: An interdisciplinary Journal of Nonlinear Science* 22.

Chekroun, M., E. Simonnet and M. Ghil (2013). "Stochastic climate dynamics: Random attractors and time-dependent invariant measures". In: *Physica D Nonlinear Phenomena* 240, pp. 1685–1700.

Cherkassy, V. and F. Mulier (1998). *Learning From Data: Concepts, Theory and Methods*. John Wiley and Sons.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks". In: *Machine Learning*, pp. 273–297.

Deerwester, S. et al. (1990). "Indexing by Latent Semantic Analysis". In: *Journal of the American society for information science* 41 (6), pp. 391–407.

Denner, A. (2017). *Coherent structures and transfer operators*. PhD thesis, Technische Universität München.

Eckart, C. and G. Young (1936). "The Approximation of One Matrix by Another of Lower Rank". In: *Psychometrika* 1, pp. 211–218.

Eubank, R. and T. Hsing (2015). *Theoretical Foundations of Functional Data Analysis with an Introduction to Linear Operators*. 1st ed. Wiley.

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

Froyland, G. and K. Padberg (2009). "Almost-invariant sets and invariant manifolds — Connecting probabilistic and geometric descriptions of coherent structures in flows". In: *Physica D* 238, pp. 1507–1523.

Froyland, G., K. Padberg-Gehle, et al. (2007). "Detection of Coherent Oceanic Structures via Transfer Operators". In: *Physical review letters* 98, p. 224503.

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

Fukumizu, K., F. R. Bach and M. I. Jordan (2004). "Dimensionality reduction for supervised learning with Reproducing Kernel Hilbert Spaces". In: *Journal of Machine Learning Research* 5, pp. 73–99.

Gohberg, I. and M. Krein (1969). *Introduction To The Theory of Linear Nonselfadjoint Operators*. American Mathematical Society.

Golub, G., A. Hoffman and G. Stewart (1987). "A generalization of the Eckart-Young-Mirsky matrix approximation theorem". In: *Linear Algebra and its Applications* 88/89, pp. 317–327.

Golub, G. and C. Van Loan (2013). *Matrix Computations*. 4th ed. John Hopkins University Press.

H. Tu, J. et al. (2014). "On Dynamic Mode Decomposition: Theory and Applications". In: *Journal of Computational Dynamics* 1, pp. 391–421.

Horn, R. and C. Johnson (2013). *Matrix Analysis*. 2nd ed. Cambridge University Press.

Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24, 417–441 and 498–520.

Hotelling, H. (1936). "Relations Between Two Sets of Variates". In: *Biometrika* 28.3/4, pp. 321–377.

Hsu, D., S. M. Kakade and T. Zhang (2012). "A spectral algorithm for learning Hidden Markov Models". In: *Journal of Computer and System Sciences* 78 (5), pp. 1460–1480.

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

Kloeden, P. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Springer.

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

Klus, S., F. Nüske, et al. (2018). "Data–driven Model Reduction and Transfer Operator Approximation". In: *Journal of Nonlinear Science* 28, pp. 985–1010.

Klus, S., I. Schuster and K. Muandet (2017). "Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces". In: *ArXiv e-prints*.

Koltai, P., G. Ciccotti and C. Schütte (2016). "On metastability and Markov state models for non-stationary molecular dynamics". In: *The Journal of Chemical Physics* 145.

Koltai, P., H. Wu, et al. (2018). "Optimal data–driven estimation of generalized Markov state models for non-equilibrium dynamics". In: *ArXiv e-prints*.

Lasota, A. and M. C. Mackey (1994). *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. 2nd ed. Springer.

Lodhi, H. et al. (2002). "Text classification using string kernels". In: *Journal of Machine Learning Research* 2, pp. 419–444.

Lorenz, E. (1963). " Deterministic nonperiodic flow". In: *Journal of Atmospheric Sciences* 20, pp. 130–141.

Mirsky, L. (1960). "Symmetric gauge functions and unitarily invariant norms". In: 11, pp. 50–59.

Molgedey, L. and H. G. Schuster (1994). "Separation of a mixture of independent signals using time delayed correlations". In: *Physical review letters* 72, pp. 3634–3637.

Moore, E. (1920). "On the Reciprocal of the General Algebraic Matrix". In: *Bulletin of American Mathematical Society* 26, pp. 394–395.

Muandet, K. et al. (2017). "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends in Machine Learning* 10.1-2, pp. 1–141.

Noé, F. and F. Nüske (2013). "A variational approach to modeling slow processes in stochastic dynamical systems". In: *Multiscale Modeling & Simulation* 11, pp. 635–655.

Nüske, F. et al. (2014). "Variational approach to molecular kinetics". In: *Journal of Chemical Theory and Computation* 10, pp. 1739–1752.

Pearson, K. (1901). "On Lines and planes of Closest Fit to Systems of Points in Space". In: *Philosophical Magazine* 11, pp. 559–572.

Penrose, R. (1955). "A generalized inverse for matrices". In: *Proceedings of the Cambridge Philosophical Society* 51, pp. 406–413.

Pérez-Hernández, G. et al. (2013). "Identification of slow molecular order parameters for Markov model construction". In: *The Journal of Chemical Physics* 139, p. 015102.

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

Rowley, C. et al. (2009). "Spectral analysis of nonlinear flows". In: *Journal of Fluid Mechanics* 641, pp. 115–127.

Rudin, W. (1991). *Functional Analysis*. 2nd ed. McGraw-Hill Inc.

Sadek, R. A. (2012). "SVD Based Image Processing Applications: State of The Art, Contributions and Research Challenges". In: *IJACSA* 3, pp. 26–34.

Schmid, P. (2010). "Dynamic Mode Decomposition of numerical and experimental data". In: *Journal of Fluid Mechanics* 65, pp. 5–28.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. The MIT Press.

Schölkopf, B., A. Smola and K.-R. Müller (1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10 (5), pp. 1299–1319.

Schütte, C. and M. Sarich (2013). "Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches". In: *AMS Courant Lecture Notes in Mathematics* 24.

Schwantes, C. and V. Pande (2015). "Modeling Molecular Kinetics with tICA and the Kernel Trick". In: *Journal of Chemical Theory and Computation* 11, pp. 600–608.

Smola, A. et al. (2007). "A Hilbert space embedding for distributions". In: *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pp. 13–31.

Song, L., K. Fukumizu and A. Gretton (2013). "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models". In: *IEEE Signal Processing Magazine* 30, pp. 98–111.

Song, L. (2008). *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, University of Sydney.

Song, L., J. Huang, et al. (2009). "Hilbert space embeddings of conditional distributions with applications to dynamical systems". In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968.

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

Tantet, A., F. van der Burgt and H. A. Dijkstra (2015). "An early warning indicator for atmospheric blocking events using transfer operators". In: *Chaos An Interdisciplinary Journal of Nonlinear Science* 25, p. 036406.

Tucker, W. (1999). "The Lorenz attractor exists". In: *Comptes Rendus de l Académie des Sciences - Series I - Mathematics* 328, pp. 1197–1202.

Vishwanathan, S. et al. (2010). "Graph Kernels". In: *Journal of Machine Learning Research* 11, pp. 1201–1242.

Weidmann, J. (1976). *Lineare Operatoren in Hilberträumen*. 3rd ed. Teubner.

Werner, D. (2007). *Funktionalanalysis*. 6th ed. Springer.

Weyl, H. (1912). "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)". In: *Mathematische Annalen* 71, pp. 441–479.

Williams, M., I. G. Kevrekidis and C. Rowley (2014). "A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition". In: *Journal of Nonlinear Science* 25, pp. 1307–1346.

Wu, H. and F. Noé (2017). "Variational approach for learning Markov processes from time series data". In: *ArXiv e-prints*.

Declaration of Authorship

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Alle direkten und indirekten Quellen und Hilfsmittel sind angebeben. Alle direkten und indirekten Zitate sind als solche gekennzeichnet. Diese Arbeit wurde bei keiner anderen Prüfungsbehörde eingereicht und auch nicht veröffentlicht.

I hereby declare that this thesis is my own work. I have authored it independently and all direct or indirect sources are listed and acknowledged as references. All directly or indirectly quoted passages in the text are marked as such. This thesis has not been submitted to any other examination authority or published.

Berlin, June 7, 2018.

Mattes Mollenhauer