2020

The singular value expansion for compact and non-compact operators

Daniel Crane
Michigan Technological University, dkcrane@mtu.edu

THE SINGULAR VALUE EXPANSION FOR COMPACT AND NON-COMPACT OPERATORS

By
Daniel K. Crane

A DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
In Mathematical Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY
2020

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Mathematical Sciences.

Department of Mathematical Sciences

Dissertation Advisor:	*Dr. Mark Gockenbach*
Committee Member:	*Dr. Allan A. Struthers*
Committee Member:	*Dr. Cécile Piret*
Committee Member:	*Dr. Brian Fick*
Department Chair:	*Dr. Mark Gockenbach*

Contents

List of Figures

List of Tables

Acknowledgments

I would first like to thank my parents for bringing me into existence and raising me. The time and money you've devoted to my education were much appreciated and I wouldn't be here without your support.

I want to thank my advisor Dr. Mark Gockenbach who has been an invaluable source of guidance throughout my time here.

I'd like to thank Samuel Judge as well as the other math majors at Taylor University. You all were a constant source of motivation for me.

Thanks to all the friends I've made while here at Michigan Tech. You have all made this experience that much more enjoyable.

Thanks to all the faculty and staff at MTU who have assisted me during my time here. In particular, I'd like to thank the members of my committee for the encouragement and advice you've given me during this last year. I'd also like to thank Jeanne Meyers, Margaret Perander, and Kimberly Puuri for all of the mostly unnoticed jobs they perform that keep the math department running smoothly. I've greatly appreciated all the help I've received from you.

Abstract

Given any bounded linear operator $T : X \to Y$ between separable Hilbert spaces X and Y, there exists a measure space $(M, \mathcal{A}, \mu)$ and isometries $V : L^2(M) \to X$, $U : L^2(M) \to Y$ and a nonnegative, bounded, measurable function $\sigma : M \to [0, \infty)$ such that

$$T = Um_\sigma V^\dagger,$$

with $m_\sigma : L^2(M) \to L^2(M)$ defined by $m_\sigma(f) = \sigma f$ for all $f \in L^2(M)$. The expansion $T = Um_\sigma V^\dagger$ is called the singular value expansion (SVE) of T.

The SVE is a useful tool for analyzing a number of problems such as the computation of the generalized inverse $T^\dagger$ of T, understanding the inverse problem $Tx = y$ and, regularizing $Tx = y$ using methods such as Tikhonov regularization. In fact, many standard Tikhonov regularization results can be derived by making use of the SVE.

The expansion $T = Um_\sigma V^\dagger$ can also be compared to the SVE of a compact operator $T : X \to Y$ which has the form

$$T = \sum_n \sigma_n u_n \otimes v_n$$

where the above sum may be finite or infinite depending on the rank of T. The set $\{\sigma_n\}$ is a sequence of positive real numbers that converge to zero if T has infinite rank. Such σ_n are the *singular values* of T. The sets $\{v_n\} \subset X$ and $\{u_n\} \subset Y$ are orthonormal sets of vectors that satisfy $Tv_n = \sigma_n u_n$ for all n. The vectors v_n and u_n are the *right and left singular vectors* of T, respectively. If the *essential range*, denoted $\mathcal{R}_{ess}(\sigma)$, forms a sequence of positive real numbers converging to zero (or is merely a finite set of nonnegative real numbers) and for each nonzero $s \in \mathcal{R}_{ess}(\sigma)$, the *essential preimage* of the singleton set $\{s\}$, denoted $\sigma_{ess}^{-1}(\{s\})$, is finite, then the bounded operator $T = Um_\sigma V^\dagger$ is in fact compact. The converse of this statement is also true.

If the operator T is compact, the singular values and vectors of T may be approximated by discretizing the operator and finding the singular value decomposition of a scaled Galerkin matrix. In general, the approximated singular values and vectors converge at the same rate, which is governed by the error (in the operator norm) in approximating T by the discretized operator. However, when the discretization is accomplished by projection (variational approximation), the computed singular values converge at an increased rate; the typical case is that the errors in the singular values are asymptotically equal to the square of the errors in the singular vectors (this statement must be modified if the approximations to the left and right singular vectors converge at different rates). Moreover, in the case of variational approximation, the error in the singular vectors can be compared with the optimal approximation error, with the two being asymptotically equal in the typical case.

Chapter 1

The singular value expansion

1.1 Introduction

In this chapter, we derive the singular value expansion (SVE) for a bounded operator between two separable Hilbert spaces. Before we do this, we will discuss two special cases of the SVE that should be familiar to the reader: the singular value decomposition (SVD) of a rectangular matrix and also the SVE of a compact operator from one separable Hilbert space to another. We will then derive the SVE for an operator that is not necessarily compact. The principal aim in discussing these three cases, is to demonstrate the relationship between a version of the spectral theorem and the SVE in all three contexts. Although a derivation of the SVE of a bounded operator can be found in the literature, this derivation relies on the polar decomposition theorem. We seek to derive the SVE in a way that mimics the logic of the derivations for the SVE in the matrix and compact operator cases and doesn't rely on the polar decomposition theorem.

It is assumed the reader is already familiar with most topics in undergraduate linear algebra, including the spectral theorem for a symmetric matrix. The reader is also expected to be familiar with the basics of Hilbert space theory.

1.1.1 The matrix case

We start this section by stating the spectral theorem for a real symmetric matrix, a result that can be found in most introductory linear algebra courses.

Theorem 1.1 (Spectral theorem for a real symmetric matrix) *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then each of the n eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ of A (counted according to multiplicity) are real numbers, there are n corresponding eigenvectors $v_1, v_2, ..., v_n$ of A that form an orthonormal basis for $\mathbb{R}^n$, and A may be written in the form $A = VDV^T$, where V is the orthogonal matrix $V = [v_1|v_2|...|v_n]$*

and $D = diag(\lambda_1, \lambda_2, ..., \lambda_n)$.

The proof of the spectral theorem may be found in many linear algebra textbooks, such as [19] or [9], and will be omitted here. This theorem is then used to derive the singular value expansion (SVD) for a rectangular matrix $A \in \mathbb{R}^{m\times n}$. A sketch of the proof is below.

Theorem 1.2 (SVD of a rectangular matrix) *Let $A \in \mathbb{R}^{m\times n}$. Then A may be written as $A = U\Sigma V^T$ where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal rectangular matrix with nonnegative diagonal entries.*

Proof: We start by assuming $m \geq n$. If $n > m$ we can apply the below proof to A^T and then take the transpose of the resulting SVD.

Consider the symmetric matrix A^TA. By the spectral theorem, we know there exist n orthonormal eigenvectors $v_1, v_2, ..., v_n$ of A^TA, corresponding to eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$, such that

$$A^TA = VDV^T,$$

where $V = [v_1|v_2|...|v_n]$ and $D = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$. Note that

$$\lambda_i = \lambda_i\langle v_i, v_i\rangle = \langle \lambda_i v_i, v_i\rangle = \langle A^TAv_i, v_i\rangle = |Av_i|^2 \geq 0.$$

This allows us to define the singular values $\sigma_1, \sigma_2, ..., \sigma_n$ by $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, 2, ..., n$. We also define vectors $u_1, u_2, ..., u_r$, where $u_i = \sigma_i^{-1}Av_i$ for $i = 1, 2, ..., r$ and r is defined by the property $\sigma_r > \sigma_{r+1} = ... = \sigma_n = 0$. Note that $\{u_i\}_{i=1}^r$ forms an orthonormal set because

$$\langle u_i, u_j\rangle = \langle \sigma_i^{-1}Av_i, \sigma_j^{-1}Av_j\rangle = \sigma_i^{-1}\sigma_j^{-1}\langle v_i, A^TAv_j\rangle = \sigma_i^{-1}\sigma_j^{-1}\sigma_j^2\langle v_i, v_j\rangle = \delta_{ij},$$

where δ_{ij} denotes the Kronecker delta. Extend $\{u_i\}_{i=1}^r$ to an orthonormal basis for all of $\mathbb{R}^m$ to form the set $\{u_i\}_{i=1}^m$, and define the orthogonal matrix $U = [u_1|u_2|...|u_m] \in \mathbb{R}^{m\times m}$. Then

$$U^TAV = U^T[Av_1|Av_2|...|Av_n] = U^T[\sigma_1u_1|\sigma_2u_2|...|\sigma_nu_n] = U^TU\Sigma = \Sigma,$$

where we have used the fact that $Av_i = \sigma_iu_i$ for each i. Applying U to the left and V^T to the right yields $A = U\Sigma V^T$. ■

The vectors $u_1, u_2, ..., u_m$ and $v_1, v_2, ..., v_n$ are referred to as the left and right singular vectors, respectively, for A. The values $\sigma_1, \sigma_2, ..., \sigma_r$ are called the singular values of A.

Using matrix multiplication, one can also write the singular value decomposition of the matrix A as a sum of rank one matrices:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T. \tag{1.1}$$

The above expansion for A is similar in form to the singular value expansion (SVE) of a compact operator, which we will discuss in the next section.

1.1.2 The compact operator case

We now consider the infinite-dimensional case, where a pair of theorems similar to Theorems 1.1 and 1.2 hold for compact operators. A compact operator is defined below

Definition 1.3 (Compact operator) *The linear operator $T : X \to Y$, where X, Y are Hilbert spaces, is compact if the image under T of any bounded set in X has compact closure.*

A compact operator is necessarily bounded. There are a number of equivalent formulations of a compact operator; in particular, T is compact if for any bounded sequence $\{x_n\} \subset X$, the sequence $\{Tx_n\} \subset Y$ has a convergence subsequence.

Every finite rank operator is compact. The most important class of compact operators is formed by integrals operators that are defined as follows. For M, a subset of $\mathbb{R}^n$, and with kernel $k : M \times M \to \mathbb{R}$ that is square integrable over $M \times M$, an integral operator $T : L^2(M) \to L^2(M)$ is defined by the equation

$$(Tx)(s) = \int_M k(s,t)x(t)\ dt.$$

Such operators come up in a number of applied math problems.

In deriving the SVE of a compact operator, we will start by stating the spectral theorem for a compact, self-adjoint operator T on a separable Hilbert space X. We will then use this theorem to derive the SVE of a compact operator $T : X \to Y$. Just as we did for the matrix case, we will omit the proof of the spectral theorem and provide a sketch of the derivation of the SVE.

Before stating the spectral theorem, we define the spectrum of a self-adjoint linear operator on a Hilbert space X. The spectrum is a generalization of the set of eigenvalues.

Definition 1.4 (Spectrum) *Given a self-adjoint linear operator $T : X \to X$ the spectrum of T is the set of all real numbers λ such that $T - \lambda I$ is not invertible, where I denotes the identity operator on X.*

Often the spectrum of an operator T is denoted by $\sigma(T)$. To avoid this notation clashing with future uses of the letter σ, we will use $\Lambda(T)$ to denote the spectrum of the operator T.

Theorem 1.5 (Spectral theorem for compact self-adjoint operators) *Let $T : X \to X$ be a compact, self-adjoint operator. Then every nonzero element $\lambda \in \Lambda(T)$ is an eigenvalue of T whose eigenspace is finite-dimensional. There also exists an orthonormal sequence $\{v_n\}$ of eigenvectors and a corresponding sequence of eigenvalues $\{\lambda_n\}$ such that*

$$T = \sum_n \lambda_n v_n \otimes v_n$$

where $\otimes$ denotes the outer product defined by $(v_n \otimes v_n)x = \langle x, v_n\rangle_X v_n$ for $x \in X$. In particular,

$$Tx = \sum_n \lambda_n \langle x, v_n\rangle_X v_n \quad \textit{for all } x \in X$$

These sequences $\{\lambda_n\}$ and $\{v_n\}$ may be finite or infinite. If they are infinite, then $\lambda_n \to 0$ as $n \to \infty$ and the series converges to T in the operator norm.

For a proof of this theorem, the reader may consult [10].

Just as we did in the matrix case, we use the spectral theorem for compact self-adjoint operators to derive the singular value expansion (SVE) for a compact operator $T : X \to Y$. The proof below can be also be found in [10].

Theorem 1.6 (SVE of a compact operator) *Let $T : X \to Y$ be a compact operator. Then there exist (finite or infinite) orthonormal sequences $\{v_n\} \subset X$ and $\{u_n\} \subset Y$ and positive numbers $\sigma_1 \geq \sigma_2 \geq ...$ such that*

$$T = \sum_n \sigma_n u_n \otimes v_n. \tag{1.2}$$

If the series is infinite, then it converges in the operator norm to T and $\sigma_n \to 0$ as $n \to \infty$. Also,

$$Tv_n = \sigma_n u_n \quad \textit{for all } n$$

and

$$Tx = \sum_n \sigma_n \langle v_n, x\rangle_X u_n \quad \textit{for all } x \in X.$$

Proof: We use an argument similar to the one used in the Theorem 1.2. We note that T^*T is compact and self-adjoint. So the spectral theorem for compact operators gives us

$$T^*T = \sum_n \lambda_n v_n \otimes v_n,$$

where $\{v_n\}$ is an orthonormal sequence in X, $|\lambda_1| \geq |\lambda_2| \geq ... > 0$ and, if the above sum is infinite, $\lambda_n \to 0$ as $n \to \infty$. Note that

$$\lambda_n = \lambda_n \langle v_n, v_n \rangle_X = \langle \lambda_n v_n, v_n \rangle_X = \langle T^* T v_n, v_n \rangle_X = \langle T v_n, T v_n \rangle_X \geq 0,$$

which means we may define $\sigma_n = \sqrt{\lambda_n}$ and $u_n = \sigma_n^{-1} T v_n$. Next, note that

$$\begin{aligned} \langle u_n, u_m \rangle_Y = \langle \sigma_n^{-1} T v_n, \sigma_m^{-1} T v_m \rangle_Y = \sigma_n^{-1} \sigma_m^{-1} \langle v_n, T^* T v_m \rangle_X &= \sigma_n^{-1} \sigma_m^{-1} \langle v_n, \sigma_m^2 v_m \rangle_X \\ &= \sigma_n^{-1} \sigma_m^{-1} \sigma_m^2 \langle v_n, v_m \rangle_X \\ &= \delta_{nm}, \end{aligned}$$

which shows $\{u_n\}$ is an orthonormal sequence. It is easily shown that $\mathcal{N}(T) = \mathcal{N}(T^*T) = \text{sp}\{v_1, v_2, ...\}^\perp$, which implies

$$\begin{aligned} Tx = T\left(\text{proj}_{\mathcal{N}(T)^\perp} x\right) = T\left(\sum_n \langle x, v_n \rangle_X v_n\right) = \sum_n \langle x, v_n \rangle_X T v_n &= \sum_n \sigma_n \langle x, v_n \rangle_X u_n \\ &= \sum_n \sigma_n (u_n \otimes v_n) x \\ &= \left(\sum_n \sigma_n u_n \otimes v_n\right) x. \end{aligned}$$

This shows that

$$T = \sum_n \sigma_n u_n \otimes v_n$$

in the pointwise sense. If the above sum is finite, the proof is complete. Otherwise, we finish the proof by proving that this sum converges to T in the operator norm. Let $x \in X$ with $\|x\|_X = 1$. Then

$$Tx - \left(\sum_{n=1}^N \sigma_n u_n \otimes v_n\right) x = \sum_{n=N+1}^\infty \sigma_n \langle x, v_n \rangle_X u_n.$$

Thus,

$$\begin{aligned} \left\| Tx - \sum_{n=1}^N \sigma_n \langle x, v_n \rangle_X u_n \right\|_Y^2 = \left\| \sum_{n=N+1}^\infty \sigma_n \langle x, v_n \rangle_X u_n \right\|_Y^2 &= \sum_{n=N+1}^\infty \sigma_n^2 |\langle x, v_n \rangle_X|^2 \\ &\leq \sigma_{N+1}^2 \sum_{n=N+1}^\infty |\langle x, v_n \rangle_X|^2 \\ &\leq \sigma_{N+1}^2 \|x\|_X^2 = \sigma_{N+1}^2. \end{aligned}$$

The facts that $\sigma_{N+1} \to 0$ as $N \to \infty$ and that this result holds for all unit vectors

$x \in X$ implies that the sum converges to the operator T in the operator norm. ■

1.2 The spectral theorem

In reading through the proofs in the previous section, one should detect a pattern in the derivations. We start with the spectral theorem for a self-adjoint operator and then use that to produce the SVD or SVE of a non-self-adjoint operator. This same principle may be applied to a bounded (perhaps non-compact) linear operator $T : X \to Y$, where X and Y are real, separable Hilbert spaces. Much of the derivation in this section comes from results discussed in [16] starting on page 49. Because this version of the SVE is both less well known and less accessible than the previous theorems, we will devote more time going through it. To get started, we first need several preliminary results.

Given a bounded, self-adjoint linear operator A on the separable Hilbert space X, and a continuous function f defined on a compact subset of $\mathbb{R}$, we want to define the operator $f(A) : X \to X$ in a sensible manner. For $f(x) = x$, it makes sense to define $f(A) = A$, for $f(x) = x^2$, $f(A) = A^2$, and so forth. Using this kind of reasoning, if $p(x) = \sum_{k=0}^{n} \alpha_k x^k$ is a polynomial, then $p(A) = \sum_{k=0}^{n} \alpha_k A^k$ is the obvious definition for $p(A)$. We then extend this idea to defining $f(A)$ for continuous functions f on a compact subset K of $\mathbb{R}$. To do this, we consider $C(K)$, the space of continuous functions defined on K. For our purposes, the spectrum $\Lambda(A)$ of A (which is always compact and because A is self-adjoint, is a subset of $\mathbb{R}$), will act as our compact set K. By the Stone-Weierstass theorem, $\mathbb{P}(K)$, the set of polynomials defined on K, is dense in $C(K)$.

Thus, we may define the operator $\Phi_A : \mathbb{P}(K) \to \mathcal{L}(X)$ where $\mathcal{L}(X)$ denotes the collection of all bounded, self-adjoint linear operators on X, by

$$\Phi_A(p) = p(A).$$

The space $\mathbb{P}(K)$ is dense in $C(K)$ and it can be proven that Φ_A is bounded, which follows from the fact that

$$\|p(A)\|_{\mathcal{L}(X)} = \|p\|_{C(K)}. \tag{1.3}$$

A proof for (1.3) can be found in [16]. Thus, if $\{p_n\}$ is a Cauchy sequence in $\mathbb{P}(K)$, then $\{p_n(A)\}$ is a Cauchy sequence in $\mathcal{L}(X)$, which implies that Φ_A may be continuously extended to all of $C(K)$.

Thus, we may define $f(A)$ by

$$f(A) = \overline{\Phi}_A(f). \tag{1.4}$$

where $\overline{\Phi}_A$ denotes the extension of Φ_A to all of $C(K)$. For a more rigorous development on this idea, the reader may consult [16].

The next result we require is the existence of the spectral measure. In deriving this, we must first review several definitions.

Definition 1.7 (σ-algebra) *Given a set M, a σ-algebra on M is a collection Σ of subsets of M that contains M and is closed under complements and countable unions.*

Definition 1.8 (Borel sets) *Given a set M with a known topology, the Borel subsets of M form the smallest σ-algebra of M that contains all the open sets of M.*

Definition 1.9 (Measure) *Given a set M and $\mathcal{A}$, a σ-algebra on M, a measure μ is a function from $\mathcal{A}$ to the interval $[0,\infty]$ that satisfies $\mu(\emptyset) = 0$, and for any countable collection $\{E_i\}_{i=1}^{\infty}$ with $E_i \in \mathcal{A}$ for all i and $E_i \cap E_j = \emptyset$ for $i \neq j$, we have $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$.*

Definition 1.10 (Borel measure) *A Borel measure μ is a measure defined on the Borel subsets of M.*

Definition 1.11 (Non-negative functional) *Consider a bounded linear functional ℓ on the space of continuous, real-valued functions on a compact subset K of $\mathbb{R}^n$ (denoted $C(K)$). We say ℓ is non-negative if, for all non-negative functions $f \in C(K)$, $\ell(f) \geq 0$.*

Throughout this chapter, we will be using the standard topology on $\mathbb{R}$. We now state the Riesz representation theorem for compact subsets of $\mathbb{R}$. See [23, *p.* 310] for a proof.

Theorem 1.12 (Riesz representation) *Let K be a compact subset of $\mathbb{R}$ and let ℓ be a nonnegative, bounded linear functional on $C(K)$. Then there exists a unique, nonnegative Borel measure μ on K such that*

$$\ell(f) = \int_K f(x)\ d\mu(x) \ \ \forall f \in C(K).$$

For $\psi \in X$, consider the bounded linear functional $\ell_\psi : C(K) \to \mathbb{R}$ defined by $\ell_\psi(f) = \langle \psi, f(A)\psi \rangle_X$. If f is non-negative, we may define $\sqrt{f}$ and $\sqrt{f(A)}$ by (1.4). Thus, $\langle \psi, f(A)\psi \rangle_X = \| \sqrt{f(A)}\psi \|_X^2 \geq 0$ so ℓ is non-negative. Therefore, we can apply the Riesz representation theorem to ℓ_ψ to produce a measure μ_ψ which satisfies

$$\langle \psi, f(A)\psi \rangle_X = \int_K f(x)\ d\mu_\psi(x)\ \forall f \in C(K).$$

This measure μ_ψ is called the *spectral measure* associated with ψ. This measure is then used to define the space $L^2(K, \mu_\psi)$ of square integrable functions on K with the measure μ_ψ. For convenience of presentation, we will denote this space $L^2(K)$ rather than $L^2(K, \mu_\psi)$ and the measure will either be explicitly given or intuited from context.

We next define a cyclic vector for the operator A, which we will use to prove a version of the spectral theorem for certain bounded self-adjoint linear operators.

Definition 1.13 (Cyclic vector) *Given a separable Hilbert space X and a bounded, self-adjoint linear operator $A : X \to X$, we say a vector ψ is a cyclic vector for A if the subspace*

$$X_\psi = span\{p(A)\psi : p \in \mathbb{P}\} \tag{1.5}$$

is dense in X.

In general, a self-adjoint operator A need not have a cyclic vector. However, in the case that it does, we may prove the spectral theorem below.

Theorem 1.14 *Let X be a separable Hilbert space and let $A : X \to X$ be a bounded, self-adjoint linear operator. If A has a cyclic vector $\psi \in X$, then there exists a unitary operator $V : L^2(\Lambda(A)) \to X$ (under the spectral measure μ_ψ) such that for every $f \in L^2(\Lambda(A))$ we have*

$$\left(V^{-1}AVf\right)(\lambda) = \lambda f(\lambda) \text{ for almost every } \lambda \in \Lambda(A).$$

Proof: For $f \in C(\Lambda(A))$, we define $f(A)$ by (1.4). Let μ_ψ be the spectral measure on $\Lambda(A)$ satisfying the equation

$$\langle \psi, f(A)\psi \rangle_X = \int_{\Lambda(A)} f \, d\mu_\psi \; \forall f \in C(\Lambda(A)).$$

We then define $V : C(\Lambda(A)) \to X$ by $Vf = f(A)\psi$ for each $f \in C(\Lambda(A))$. Note that

$$\|Vf\|_X^2 = \langle f(A)\psi, f(A)\psi \rangle_X = \langle \psi, f(A)^2\psi \rangle_X = \int_{\Lambda(A)} f^2 \, d\mu_\psi = \|f\|^2_{L^2(\Lambda(A))},$$

which implies that V is an isometry. We know that $C(\Lambda(A))$ is dense in $L^2(\Lambda(A))$, which implies there exists a unique continuous extension $\overline{V}$ of V from $L^2(\Lambda(A))$ to X that satisfies

$$\|\overline{V}f\|_X = \|f\|_{L^2(\Lambda(A))} \text{ for each } f \in L^2(\Lambda(A)).$$

Because ψ is cyclic, we know $\overline{X}_\psi = X$. Thus, for $x \in X$, there is a sequence $\{f_n\}$ of functions such that $f_n(A)\psi \to x$ as $n \to \infty$ or, equivalently, $\overline{V}f_n \to x$ as $n \to \infty$. Because $\overline{V}$ is an isometry, $f_n \to f$ for some $f \in L^2(\Lambda(A))$ and then by continuity of $\overline{V}$, we know $\overline{V}f = x$. Thus, $\overline{V}$ (which we will denote by V for convenience) maps $L^2(\Lambda(A))$ onto all of X, which means that V is an isometric isomorphism from $L^2(\Lambda(A))$ to X.

To finish the proof, let p be a polynomial defined on $\Lambda(A)$ and let q denote the polynomial defined by $q(t) = tp(t)$. Then

$$V^{-1}AVp = V^{-1}Ap(A)\psi = V^{-1}q(A)\psi = q. \tag{1.6}$$

Now note that each $f \in L^2(\Lambda(A))$ may be written as the limit of a sequence of functions in $C(\Lambda(A))$ and each function in $C(\Lambda(A))$ may be written as the limit of a sequence of polynomials defined on $\Lambda(A)$. Thus, the polynomials defined on $\Lambda(A)$ are dense in $L^2(\Lambda(A))$, which means (1.6) holds for all $f \in L^2(\Lambda(A))$ as well. This completes the proof. ■

The above theorem is inadequate in that not every self-adjoint operator A has a single cyclic vector. Below is a simple example of this phenomenon.

Example 1.15 *Let $A : X \to X$ (with X having dimension larger than one) be defined by $Ax = \lambda x$ for some real number λ. For any $\psi \in X$, the space X_ψ defined by equation* (1.5) *is one-dimensional. Thus, there is no cyclic vector for A.*

The remainder of this section will be devoted to proving the spectral theorem for an operator that does not have a cyclic vector. To do this, we require more definitions.

Definition 1.16 (Partially ordered set) *Given a set P, a partial ordering on the set, denoted by $\leq$, is a relation satisfying reflexivity ($a \leq a$ for all $a \in P$), antisymmetry ($a \leq b$ and $b \leq a$ implies $a = b$) and transitivity ($a \leq b$ and $b \leq c$ implies $a \leq c$).*

Definition 1.17 (Totally ordered set) *A totally ordered set $(P, \leq)$ is a partially ordered set where every two elements are related to each other.*

Definition 1.18 (Upper bound) *Given a subset S of a partially ordered set $(P, \leq)$, an upper bound on S is an element $p \in P$ such that $s \leq p$ for all $s \in S$.*

We now state Zorn's lemma.

Lemma 1.19 (Zorn's lemma) *Suppose a partially ordered set P has the property that every totally ordered subset has an upper bound. Then the set P contains a maximal element. That is, there is an element $p \in P$ such that $q \leq p$ for all $q \in P$ comparable to p.*

We can now prove the below two lemmas. The arguments presented here are restatements of proofs in [16].

Lemma 1.20 *Let X be a real Hilbert space and let A be a bounded self-adjoint linear operator on X. Let ϕ, $\psi \in X$. If $\psi \perp X_\phi$, then $X_\psi \perp X_\phi$.*

Proof: If $\psi \perp X_\phi$, then, for every $n \in \mathbb{Z}^+$, $\langle A^n\phi, \psi\rangle_X = 0$. Thus, for each $n, m \in \mathbb{Z}^+$, we have

$$\langle A^n\phi, A^m\psi\rangle_X = \langle A^{n+m}\phi, \psi\rangle_X = 0,$$

which completes the proof. ■

Before the next lemma, we require another definition.

Definition 1.21 *Let X be a separable Hilbert space and $\{X_n\}_{n\in\mathcal{N}}$ be a collection of mutually orthogonal subspaces of X. Then the orthogonal direct sum*

$$\bigoplus_{n\in\mathcal{N}} X_n$$

denotes the collection of all sums of the form $\sum_{k\in\mathcal{N}} x_k$ where each $x_k \in X_k$ and it is understood that only finitely many x_k are nonzero.

It is easy to see that this direct sum is a subspace of X. Also, if the set $\mathcal{N}$ is finite, the above direct sum is closed. If $\mathcal{N}$ is infinite, then it can be shown that the closure, denoted by

$$\overline{\bigoplus_{n\in\mathcal{N}} X_n}$$

is the collection of sums of the form $\sum_{k\in\mathcal{N}} x_k$ for $x_k \in X_k$ where we drop the assumption that only a finite number of the x_k's are nonzero and also assume that $\sum_{k\in\mathcal{N}} \|x_k\|_X^2 < \infty$. Note that the fact that X is a separable Hilbert space guarantees that the set $\mathcal{N}$ is at most countably infinite.

Lemma 1.22 *Let X be a real separable Hilbert space and let A be a bounded self-adjoint linear operator on X. Then there exists a collection $\{\psi_n\}_{n\in\mathcal{N}}$ of nonzero vectors in X such that for all $m, n \in \mathcal{N}$, with $m \neq n$, we have $X_{\psi_m} \perp X_{\psi_n}$ and*

$$\overline{\bigoplus_{n\in\mathcal{N}} X_{\psi_n}} = X$$

Proof: Let $S = \{x \in X : \|x\|_X = 1\}$ denote the unit sphere in X. Consider the collection of subsets

$$\mathcal{D} = \{D \subset S \,|\, \forall \phi, \psi \in D,\ \phi \neq \psi \implies X_\phi \perp X_\psi\}.$$

Note that $\mathcal{D}$ is a partially ordered set with respect to inclusion. Further, any totally ordered subset $\mathcal{D}_0$ of $\mathcal{D}$ has a upper bound given by the union $\bigcup_{D\subset\mathcal{D}_0} D$. Thus, by Zorn's lemma, there exists a maximal element $\mathcal{D}_{\max} \in \mathcal{D}$. Now let V be defined by

$$V = \bigoplus_{\phi\in\mathcal{D}_{\max}} X_\phi.$$

We finish the proof by showing that V is dense in X. Suppose it is not. Then there exists some $\psi \in S$ such that $\psi \perp V$. By Lemma 1.20, this implies that $X_\psi \perp X_\phi$ for each $\phi \in \mathcal{D}_{\max}$. But then $\mathcal{D}_{\max} \cup \{\psi\} \in \mathcal{D}$ and contains $\mathcal{D}_{\max}$, which contradicts the fact that $\mathcal{D}_{\max}$ is maximal. ■

It should be noted that the collection of cyclic vectors from the above lemma is by no means unique.

The utility of Lemma 1.22 is that while there exist operators that do not have a cyclic vector, we can still prove the result in Theorem 1.14 with a collection of cyclic vectors. To do this, we use Lemma 1.22 to produce a countable collection $\{\psi_n\}_{n\in\mathcal{N}}$ of vectors such that

$$\overline{\bigoplus_{n\in\mathcal{N}} \overline{X}_{\psi_n}} = X$$

and ψ_n is a cyclic vector for the operator $A_n = A|_{\overline{X}_{\psi_n}}$. Note we use $\overline{X}_{\psi_n}$ instead of X_{ψ_n} because A_n must be defined on a Hilbert space. We then apply Theorem 1.14 to each A_n. To produce an expansion for A, we must combine the expansions for each A_n. To do this, we will need to introduce the notion of the direct sum of measure spaces

$$(M, \mathcal{A}, \mu) = \bigoplus_{n\in\mathcal{N}} (M_n, \mathcal{A}_n, \mu_n)$$

as well as the direct sum of L^2 spaces

$$L^2(M) = \overline{\bigoplus_{n\in\mathcal{N}} L^2(M_n)}.$$

We will discuss these two in turn. If the collection of sets $\{M_n\}$ are each pairwise disjoint, then the new set M may be taken to be the union of each M_n. However, in many contexts (such as the case of repeated eigenvalues in $\Lambda(A)$) the sets M_n will not be disjoint. Thus, we define M to be the *disjoint union* of the sets M_n:

$$M = \bigsqcup_{n\in\mathcal{N}} M_n = \bigcup_{n\in\mathcal{N}} \{(x, n) \,|\, x \in M_n\}.$$

In our context, each M_n is a subset of $\mathbb{R}$, and we impose the standard topology on $\mathbb{R}$ and the subspace topology on M_n. That is, the open sets of M_n are all sets $M_n \cap U$, where U is open in $\mathbb{R}$.

We then impose the disjoint union topology on M. That is, if $S \subset M$, then S will have the form

$$S = \bigcup_{n\in\mathcal{N}} \{(s, n) \,|\, s \in S_n\}$$

where $S_n \subset M_n$, and S is open if and only if each S_n is open under the topology on M_n.

The set of measurable sets $\mathcal{A}$ is defined in the following way: a subset $S \subset M$ with the form $S = \bigcup\limits_{n \in \mathcal{N}} \{(s, n) \mid s \in S_n\}$ is in $\mathcal{A}$ as long as each $S_n \in \mathcal{A}_n$. Further the measure μ is defined in terms of the measures μ_n. That is, for $S = \bigcup\limits_{n \in \mathcal{N}} \{(s, n) \mid s \in S_n\}$,

$$\mu(S) = \sum_{n \in \mathcal{N}} \mu_n(S_n).$$

When taking the direct sum of the $L^2(M_n)$ spaces, we regard each $L^2(M_n)$ as a subspace of $L^2(M)$ by extending every $f_n \in L^2(M_n)$ to $\overline{f}_n \in L^2(M)$ by the equation

$$\overline{f}_n(x, k) = \begin{cases} f_n(x) & k = n \\ 0 & k \neq n. \end{cases}$$

This implies that $\{L^2(M_n)\}$ is a set of orthogonal subspaces. Thus, any function f in $\bigoplus_{n \in \mathcal{N}} L^2(M_n)$ can be written as the sum $f = \sum_{n \in \mathcal{N}} \overline{f}_n$. For convenience, we will henceforth use f_n to denote each extension $\overline{f}_n$. It is also understood that only a finite number of these functions are nonzero. Every element in the set $\overline{\bigoplus_{n \in \mathcal{N}} L^2(M_n)}$, however, may be written as $\sum_{n \in \mathcal{N}} f_n$ where potentially infinitely many of the f_n functions are nonzero. The norm of $f \in L^2(M)$ is given by

$$\|f\|^2_{L^2(M)} = \sum_{n \in \mathcal{N}} \|f_n\|^2_{L^2(M_n)}.$$

The last thing we must discuss is the topology that is placed on the set M. We place the standard topology on the subset M_n of $\mathbb{R}$ for each $n \in \mathcal{N}$ and the disjoint union topology on M. We note that the standard topology on M_n has a countable base, (that is, a countable collection of open sets $\mathcal{U} = \{U_i\}_{i=1}^{\infty}$ such that any open set E in M_n may be written as a union of elements in $\mathcal{U}$). Then because there is a countable number of sets M_n, M itself must also have a countable base. When M has this property, it is said to be second-countable.

The topological space M will also be Hausdorff. That is, for any pair of points $(\lambda_1, m_1), (\lambda_2, m_2) \in M$, there exist open sets $\mathcal{O}_1$ and $\mathcal{O}_2$ such that $(\lambda_1, m_1) \in \mathcal{O}_1$, $(\lambda_2, m_2) \in \mathcal{O}_2$ and $\mathcal{O}_1 \cap \mathcal{O}_2 = \emptyset$.

Using all of this, we may now prove the spectral theorem.

Theorem 1.23 (The spectral theorem for bounded self-adjoint operators) *Let X be a real, separable Hilbert space and let $A : X \to X$ be a bounded, self-adjoint linear operator. Then there exists a second countable, Hausdorff measure space $(M, \mathcal{A}, \mu)$,an essentially bounded function $\lambda : M \to \mathbb{R}$, and an isometric isomorphism*

$V : L^2(M) \to X$ such that

$$V^{-1}AV = m_\lambda$$

where $m_\lambda : L^2(M) \to L^2(M)$ is the multiplication operator on $L^2(M)$ defined by $m_\lambda f = \lambda f$ for all $f \in L^2(M)$.

Proof: From Lemma 1.22, there is a countable subset $\mathcal{N} \subset \mathbb{Z}^+$ and a collection of vectors $\{\psi_n\}_{n\in\mathcal{N}}$ such that

$$\overline{\bigoplus_{n\in\mathcal{N}} \overline{X}_{\psi_n}} = X,$$

where $\overline{X}_{\psi_n}$ is invariant under A and each ψ_n is a cyclic vector for the operator $A_n = A|_{\overline{X}_{\psi_n}}$. We apply Theorem 1.14 to each A_n to produce the measure spaces $(\Lambda(A_n), \mathcal{A}_n, \mu_n)$ where μ_n is the spectral measure given in Lemma 1.12. Theorem 1.14 also gives us the isometric isomorphism $V_n : L^2(\Lambda(A_n)) \to \overline{X}_{\psi_n}$, defined by $V_n f = f(A_n)\psi_n$ for each $f \in L^2(\Lambda(A_n))$, such that

$$V_n^{-1} A_n V_n f(t) = t f(t)$$

for almost every $t \in \Lambda(A_n)$.
We then define the measure space $(M, \mathcal{A}, \mu)$ as

$$(M, \mathcal{A}, \mu) = \bigoplus_{n\in\mathcal{N}} (\Lambda(A_n), \mathcal{A}_n, \mu_n)$$

where M and μ are defined per the discussion preceding the theorem. We also let

$$L^2(M) = \overline{\bigoplus_{n\in\mathcal{N}} L^2(\Lambda(A_n))}.$$

Now define the operator $V : \bigoplus_{n\in\mathcal{N}} L^2(\Lambda(A_n)) \to X$ by

$$Vf = \sum_{n\in\mathcal{N}} V_n f_n.$$

where $f \in L^2(M)$ has the form $f = \sum_{n\in\mathcal{N}} f_n$, where $f_n \in L^2(\Lambda(A_n))$. Using the fact that $V_n f_n \in \overline{X}_{\psi_n}$ for all $n \in \mathcal{N}$ and $\overline{X}_{\psi_n} \perp \overline{X}_{\psi_m}$ for $m \neq n$, we have

$$\|Vf\|_X^2 = \sum_{n\in\mathcal{N}} \|V_n f_n\|_X^2 = \sum_{n\in\mathcal{N}} \|f_n\|_{L^2(M_n)}^2 = \|f\|_{L^2(M)}^2,$$

which shows that V is an isometry. Now continuously extend V to the operator $\overline{V} : L^2(M) \to X$ that satisfies $\|\overline{V}f\|_X^2 = \|f\|_{L^2(M)}^2$ for all $f \in L^2(M)$. For convenience, we will use V instead of $\overline{V}$ to denote this operator.

Now let $x \in X$. The vector x may be written as $x = \sum_{n \in \mathcal{N}} x_n$, where $x_n \in \overline{X}_{\psi_n}$ for all $n \in \mathcal{N}$ and $\sum_{n \in \mathcal{N}} \|x_n\|_X^2 < \infty$. Because $\mathcal{R}(V_n) = \overline{X}_{\psi_n}$, there is a function $f_n \in L^2(\Lambda(A_n))$ such that $x_n = V_n f_n$ for each $n \in \mathcal{N}$. Define $f = \sum_{n \in \mathcal{N}} f_n$ and note that $f \in L^2(M)$ because

$$\|f\|_{L^2(M)}^2 = \sum_{n \in \mathcal{N}} \|f_n\|_{L^2(M_n)}^2 = \sum_{n \in \mathcal{N}} \|x_n\|_X^2 < \infty.$$

Also $x = \sum_{n \in \mathcal{N}} V_n f_n = V \left(\sum_{n \in \mathcal{N}} f_n \right) = Vf$. So $x \in \mathcal{R}(V)$, which implies V is an isometric isomorphism from $L^2(M)$ to X.

Next, note that the operator $V^{-1} : X \to L^2(M)$ is such that if $x \in X$, then $x = \sum_{n \in \mathcal{N}} x_n$ for $x_n \in \overline{X}_{\psi_n}$ and V^{-1} sends x to the function $f = \sum_{n \in \mathcal{N}} f_n$ such that $V_n f_n = x_n$ for all $n \in \mathcal{N}$; that is $V^{-1} \left(\sum_{n \in \mathcal{N}} x_n \right) = \sum_{n \in \mathcal{N}} V_n^{-1} x_n$.

Now define the function $\lambda_n : M_n \to \mathbb{R}$ by the formula $\lambda_n(t) = t$ for $t \in \Lambda(A_n)$. Then define $\lambda : M \to \mathbb{R}$ by $\lambda = \sum_{n \in \mathcal{N}} \lambda_n$ where each λ_n has been extended to all of M by setting $\lambda_n(t, m) = t * \delta_{mn}$ for $(t, m) \in M$. This implies that for $f_n \in L^2(M_n)$, $V_n^{-1} A_n V_n f_n = \lambda_n f_n$.

Finally, for $f \in L^2(M)$, we have

$$\begin{aligned} V^{-1} A V f = V^{-1} A \left(\sum_{n \in \mathcal{N}} V_n f_n \right) = V^{-1} \left(\sum_{n \in \mathcal{N}} A_n V_n f_n \right) &= \sum_{n \in \mathcal{N}} V_n^{-1} A_n V_n f_n \\ &= \sum_{n \in \mathcal{N}} \lambda_n f_n \\ &= \lambda f = m_\lambda f. \end{aligned}$$

This completes the proof. ■

Before moving on, we present an example to illustrate the construction proved above.

Example 1.24 *Consider for $\lambda \in \mathbb{R}$, the operator $A : X \to X$ defined by $Ax = \lambda x$ for each $x \in X$. The operator A does not have a cyclic vector. However, supposing X is a separable Hilbert space, we can find an orthonormal basis for X given by $\{\phi_n\}_{n=1}^{\infty}$ and let each ϕ_n act as a cyclic vector for $A_n = A|_{sp\{\phi_n\}}$. We then construct the measure space $(M_n, \mathcal{A}_n, \mu_n)$ where $M_n = \{\lambda\}$, $\mathcal{A}_n = \{\emptyset, \{\lambda\}\}$ and μ_n satisfies $\mu_n(\{\lambda\}) = 1$. The isometric isomorphism $V_n : L^2(M_n) \to sp\{\phi_n\}$ is defined as $V_n f = f(\lambda)\phi_n$ and satisfies*

$$A_n = V_n m_\lambda V_n^{-1}$$

where m_λ denotes the multiplication by λ on $L^2(M_n)$. So for $x = \beta\phi_n \in sp\{\phi_n\}$, we have

$$V_n m_\lambda V_n^{-1} x = V_n m_\lambda V_n^{-1} \beta\phi_n = V_n m_\lambda \beta = V_n \lambda\beta = \lambda\beta\phi_n = \lambda x = A_n x.$$

We then construct the set $M = \{(\lambda, n) : n \in \mathbb{Z}^+\}$, which is the disjoint union of each M_n. The measure μ is defined so that if $S = \{(\lambda, n) : n \in \mathcal{N}\}$ where $\mathcal{N}$ is some subset of $\mathbb{Z}^+$, then $\mu(S) = \sum_{n\in\mathcal{N}} \mu_n(\{\lambda\})$, which reduces to $\mu(S) = |\mathcal{N}|$ in this case. We then define $V : L^2(M) \to X$ by $Vf = V\left(\sum_{n=1}^{\infty} f_n\right) = \sum_{n=1}^{\infty} f_n(\lambda)\phi_n$ where each $f_n : \{\lambda\} \to \mathbb{R}$ is a function in $L^2(M_n)$. This V defines an isometry. It is also clear that $\mathcal{R}(V) = X$ so V is an isometric isomorphism. We can write any $x \in X$ as $x = \sum_{n=1}^{\infty} \beta_n\phi_n$ and let f_β denote the function in $L^2(M)$ such that $f_\beta(\lambda, n) = \beta_n$. Then

$$\begin{aligned} Vm_\lambda V^{-1}x = Vm_\lambda V^{-1}\sum_{n=1}^{\infty}\beta_n\phi_n = Vm_\lambda f_\beta = V\lambda f_\beta = \sum_{n=1}^{\infty}\lambda_n\beta_n\phi_n &= \lambda\sum_{n=1}^{\infty}\beta_n\phi_n \\ &= \lambda x = Ax. \end{aligned}$$

Thus $A = Vm_\lambda V^{-1}$.

One last thing we should note about the construction of the SVE is the flexibility we have in defining the sets M_n for the measure spaces $(M_n, \mathcal{A}_n, \mu_n)$. The spectral measure μ_n is defined on the spectrum $\Lambda(A_n)$. Thus, the simplest way to define M_n is by $M_n = \Lambda(A_n)$. However, one could also allow M_n to be any subset of $\mathbb{R}$ as long as $\Lambda(A_n) \subset M_n$. One then must adjust the definition of the measure μ_n so that for $S_n \subset M_n$, $\mu_n(S_n) = \mu(S_n \cap \Lambda(A_n))$.

Now that we have constructed the SVE for a self-adjoint operator, we no longer need to concern ourselves with the exact details of how M and the measure space $(M, \mathcal{A}, \mu)$ are constructed. We may simply conclude that M is a set with a topology on it that is both second countable and Hausdorff and there is an isometric isomorphism $V : L^2(M) \to X$ such that $A = Vm_\lambda V^{-1}$ where $\lambda : M \to \mathbb{R}$ is an essentially bounded function and $m_\lambda : L^2(M) \to L^2(M)$ is defined by $m_\lambda f = \lambda f$ for all $f \in L^2(M)$.

1.3 The singular value expansion

We will now use the spectral theorem for a bounded self-adjoint linear operator to derive the SVE for a general bounded linear operator $T : X \to Y$ in the same way we did in the matrix and compact operator case. Before we do this however, we will need to prove the following two lemmas and theorem. The results below were proven by Gockenbach in [8].

Lemma 1.25 *Let $(M, \mathcal{A}, \mu)$ be a measure space and let $\theta : M \to [0, \infty)$ be a measurable function that is nonzero a.e. Let $\{\alpha_k\}$ be any sequence of positive numbers converging monotonically to zero and, for $k \in \mathbb{Z}^+$, let $E_k = \{x \in M : \theta(x) < \alpha_k\}$. If $\mu(E_n) < \infty$ for some $n \in \mathbb{Z}^+$, then $\mu(E_k) \to 0$ as $k \to \infty$.*

Proof: We first define $E = \bigcap_{k=1}^{\infty} E_k$; then $E = \{x \in M : \theta(x) = 0\}$. Then $\mu(E) = 0$ by hypothesis. But then, because $E_{k+1} \subset E_k$ for all k and $\mu(E_n) < \infty$, it follows from a standard theorem of measure theory (see Theorem 1.8 of [7]) that

$$0 = \mu(E) = \mu\left(\cap_{k=1}^{\infty} E_k\right) = \lim_{k\to\infty} \mu(E_k),$$

which proves the result. ■

Lemma 1.26 *Let $(M, \mathcal{A}, \mu)$ be a measure space and let $\theta : M \to [0, \infty)$ be a measurable function that is positive and finite a.e. Define*

$$S = \{f \in L^2(M) : \theta^{-1} f \in L^2(M)\}. \tag{1.7}$$

Then S is dense in $L^2(M)$.

Proof: Let $f \in L^2(M)$ be given. For every $\epsilon > 0$, since f^2 is integrable, there exists a measurable subset N_ϵ of M such that $\mu(N_\epsilon) < \infty$ and

$$\int_{M\setminus N_\epsilon} f^2 < \epsilon.$$

(To see this, note that Theorems 2.10, 2.14 of [7] imply that there is a simple function $g : M \to [0, \infty)$ such that $\left|\int g - \int f^2\right| < \epsilon$. Define N_ϵ to be the support of g. It is clear that $\mu(N_\epsilon) < \infty$, because otherwise g could not be integrable.) Now let $\epsilon > 0$ be given. For each $k \in \mathbb{Z}^+$, define $E_k = \{x \in M : \theta(x) < 1/k\} \cap N_{\epsilon/2}$, $F_k = (M\setminus E_k) \cap N_{\epsilon/2}$, and $f_k : M \to [0, \infty)$ by $f_k = f\chi_{F_k}$ (where χ_A is the characteristic function of $A \in \mathcal{A}$). We wish to show that $f_k \in S$ for all k and, for k sufficiently large, $\|f_k - f\|_{L^2(M)} < \epsilon$. We see that $\theta^{-1} f_k \in L^2(M)$ because

$$\int (\theta^{-1} f_k)^2 = \int \theta^{-2} f_{\chi_{F_k}}^2 = \int_{F_k} \theta^{-2} f^2 \leq k^2 \int_{F_k} f^2 \leq k^2 \int f^2 < \infty$$

since $\theta \geq 1/k$ on F_k. This shows that $\theta^{-1} f_k \in L^2(M)$, that is, $f_k \in S$.

Finally, $M = F_k \cup E_k \cup (M\setminus N_{\epsilon/2})$ and therefore

$$\begin{aligned}\int (f_k - f)^2 &= \int_{F_k} (f_k - f)^2 + \int_{E_k} (f_k - f)^2 + \int_{M\setminus N_{\epsilon/2}} (f_k - f)^2 \\ &= \int_{F_k} (f - f)^2 + \int_{E_k} (0 - f)^2 + \int_{M\setminus N_{\epsilon/2}} (0 - f)^2 \\ &= \int_{E_k} f^2 + \int_{M\setminus N_{\epsilon/2}} f^2.\end{aligned}$$

The second integral is less than $\epsilon/2$ by construction of $N_{\epsilon/2}$. Moreover, $A \to \int_A f^2$

defines a measure on $\mathcal{A}$ (a simple result to prove from the definition of measure). Since $E_{k+1} \subset E_k$ for all k and each E_k has finite measure, we know that

$$\int_{E_k} f^2 \to \int_E f^2 = 0,$$

where $E = \cap_{k=1}^{\infty} E_k = \{x \in M : \theta(x) = 0\}$ and $\mu(E) = 0$ by assumption. This implies that

$$\int_{E_k} f^2 < \frac{\epsilon}{2}$$

for all k sufficiently large and hence that

$$\int (f_k - f)^2 < \epsilon$$

for all k sufficiently large. Thus, $f_k \to f$ in $L^2(M)$. Since f was an arbitrary element of $L^2(M)$, this shows that S is dense in $L^2(M)$. ■

Next, we derive the SVE for a bounded linear operator in the special case that $\mathcal{N}(T) = \{0\}$. As mentioned in Section 1.1, the derivation of the SVE for a bounded operator follows from applying the spectral theorem for bounded operators to T^*T and following steps very similar to those found in Theorems 1.2 and 1.6.

Theorem 1.27 *Let X and Y be real separable Hilbert spaces and let $T : X \to Y$ be a bounded linear operator with $\mathcal{N}(T) = \{0\}$. Then there exist a measure space $(M, \mathcal{A}, \mu)$, isometric isomorphisms $V : L^2(M) \to X$, $U : L^2(M) \to \overline{\mathcal{R}(T)}$, and an essentially bounded measurable function $\sigma : M \to [0, \infty)$ such that*

$$T = U m_\sigma V^{-1}.$$

Moreover, $\sigma > 0$ a.e.

Proof: By the spectral theorem for bounded self-adjoint operators, there exist a measure space $(M, \mathcal{A}, \mu)$, an isometric isomorphism $V : L^2(M) \to X$, and a bounded measurable function $\theta : M \to \mathbb{R}$ such that

$$T^*T = V m_\theta V^{-1}.$$

We will first show that $\theta \geq 0$ a.e., which will follow if we prove that

$$\langle m_\theta f, f\rangle_{L^2(M)} \geq 0 \text{ for all } f \in L^2(M).$$

We will prove this by first noting that $m_\theta = V^{-1}T^*TV$ and hence, for all $f \in L^2(M)$,

$$\langle m_\theta f, f\rangle_{L^2(M)} = \langle V^{-1}T^*TVf, f\rangle_{L^2(M)} = \langle TVf, TVf\rangle_Y \geq 0$$

where we've used the fact that $V^{-1} = V^*$ because V is an isometric isomorphism. So we conclude $\theta \geq 0$ a.e., as desired.

Now we define $E = \{x \in M : \theta(x) = 0\}$. If $\mu(E) > 0$, then $\chi_E \neq 0$ in $L^2(M)$, which implies that $V\chi_E \neq 0$ in X and hence that $T^*TV\chi_E \neq 0$ (because $\mathcal{N}(T^*T) = \mathcal{N}(T) = \{0\}$). But

$$T^*TV\chi_E = Vm_\theta\chi_E = V(\theta\chi_E) = 0$$

because $\theta = 0$ on E and $\chi_E = 0$ on $M\backslash E$. This contradiction shows that $\mu(E)$ must be zero, that is, $\theta > 0$ a.e. in M. Now we define $\sigma = \sqrt{\theta}$ and

$$S = \{f \in L^2(M) : \sigma^{-1}f \in L^2(M)\}. \tag{1.8}$$

By Lemma 1.26, we see that S is dense in $L^2(M)$. We then define $U : S \to Y$ by $U = TVm_{\sigma^{-1}}$. Since $\sigma^{-1}f \in L^2(M)$ for all $f \in S$, U is well-defined. We also see that it is linear and densely defined. Next, we have

$$\begin{aligned}\|Uf\|_Y^2 &= \langle Uf, Uf\rangle_Y = \langle TVm_{\sigma^{-1}}f, TVm_{\sigma^{-1}}f\rangle_Y \\ &= \langle f, m_{\sigma^{-1}}V^{-1}T^*TVm_{\sigma^{-1}}f\rangle_{L^2(M)} \\ &= \langle f, m_{\sigma^{-1}}m_{\sigma^2}m_{\sigma^{-1}}f\rangle_{L^2(M)} \\ &= \langle f, f\rangle_{L^2(M)} = \|f\|_{L^2(M)}^2.\end{aligned}$$

This shows that $\|Uf\|_Y = \|f\|_{L^2(M)}$ for all $f \in S$. Since U is bounded and densely defined, it can be extended to a bounded operator whose domain is all of $L^2(M)$. For convenience, we will use U to denote the extension as well (i.e. U satisfies $U|_S = TVm_{\sigma^{-1}}$). Then, by continuity, we have $\|Uf\|_Y = \|f\|_{L^2(M)}$ for all $f \in L^2(M)$. This shows that U is an isometry from $L^2(M)$ to $\mathcal{R}(U)$.

Next, we show that $T = Um_\sigma V^{-1}$. For each $x \in X$, $m_\sigma V^{-1}x \in S$ because $m_{\sigma^{-1}}m_\sigma V^{-1}x = V^{-1}x \in L^2(M)$. Therefore, for each $x \in X$,

$$Um_\sigma V^{-1}x = TVm_{\sigma^{-1}}m_\sigma V^{-1}x = TVV^{-1}x = Tx.$$

Therefore, $Um_\sigma V^{-1} = T$.

Lastly, we will show that $\mathcal{R}(U) = \overline{\mathcal{R}(T)}$. To do this, let $y \in \overline{\mathcal{R}(T)}$. Then there exists a sequence $\{x_n\} \subset X$ such that $Tx_n \to y$, that is, $Um_\sigma V^{-1}x_n \to y$. This means that $\{Um_\sigma V^{-1}x_n\}$ is a Cauchy sequence and hence, because U is an isometry, $\{m_\sigma V^{-1}x_n\}$ is Cauchy in $L^2(M)$. So suppose $m_\sigma V^{-1}x_n \to f \in L^2(M)$. Then

$$Uf = \lim_{n\to\infty} Um_\sigma V^{-1}x_n = y,$$

which shows that $y \in \mathcal{R}(U)$. Since $\mathcal{R}(U) \subset \overline{\mathcal{R}(T)}$ by definition of U, it follows that $\mathcal{R}(U) = \overline{\mathcal{R}(T)}$. This completes the proof. ■

We will now use the above theorem to derive the SVE of a bounded operator T :

$X \to Y$ without the assumption that $\mathcal{N}(T) = \{0\}$. Before we derive this however, we recall the definition of the Moore-Penrose generalized inverse of an operator.

Definition 1.28 *Given, a bounded linear operator $T : X \to Y$, the Moore-Penrose generalized inverse of T, denoted $T^\dagger$ is the unique linear extension of the operator $\hat{T}^{-1}$ to $\mathcal{D}(T^\dagger) = \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$ with $\mathcal{N}(T^\dagger) = \mathcal{R}(T)^\perp$, where $\hat{T} = T|_{\mathcal{N}(T)^\perp}$.*

Theorem 1.29 (SVE of a bounded operator) *Let X and Y be real separable Hilbert spaces and let $T : X \to Y$ be a bounded linear operator. Then there exist a measure space $(M, \mathcal{A}, \mu)$, isometries $V : L^2(M) \to X$ and $U : L^2(M) \to Y$, and an essentially bounded, measurable function $\sigma : M \to [0, \infty)$ that is positive a.e. such that*

$$T = U m_\sigma V^\dagger$$

with $m_\sigma : L^2(M) \to L^2(M)$ defined by $m_\sigma(f) = \sigma f$ for all $f \in L^2(M)$.

Proof: Apply Theorem 1.27 to the operator $\hat{T} = T|_{\mathcal{N}(T)^\perp}$. This gives us

$$\hat{T} = \hat{U} m_\sigma \hat{V}^{-1},$$

where $\hat{V} : L^2(M) \to \mathcal{N}(T)^\perp$, $\hat{U} : L^2(M) \to \overline{\mathcal{R}(\hat{T})} = \overline{\mathcal{R}(T)}$ are isometric isomorphisms and $m_\sigma : L^2(M) \to L^2(M)$ is a multiplication operator with $\sigma > 0$ almost everywhere. We can then define $U : L^2(M) \to Y$ and $V : L^2(M) \to X$ such that $Uf = \hat{U}f$ and $Vf = \hat{V}f$ for all $f \in L^2(M)$. In effect, we are just changing the codomains of $\hat{U}$ and $\hat{V}$. Then U and V will be isometries (not necessarily isometric isomorphisms). Next, let $x \in X$ and decompose x as $x = x_1 + x_2$ for $x_1 \in \mathcal{N}(T)^\perp = \mathcal{R}(V)$ and $x_2 \in \mathcal{N}(T) = \mathcal{R}(V)^\perp$. Then

$$\begin{aligned} Tx = T(x_1 + x_2) = Tx_1 = \hat{T}x_1 = \hat{U} m_\sigma \hat{V}^{-1} x_1 = U m_\sigma \hat{V}^{-1} x_1 &= U m_\sigma V^\dagger (x_1 + x_2) \\ &= U m_\sigma V^\dagger x, \end{aligned}$$

where $V^\dagger$ is the generalized inverse of V. Also note that $\mathcal{N}(V^\dagger) = \mathcal{R}(V)^\perp = \mathcal{N}(T)$ and $V^\dagger|_{\mathcal{N}(T)^\perp} = V_1^{-1}$. Thus,

$$T = U m_\sigma V^\dagger. \blacksquare \tag{1.9}$$

This expansion for the operator T allows us to find expansions for T^* and $T^\dagger$ as well. These expansions will be used prominently in future chapters. To derive expansions for these operators, we first prove the following lemma.

Lemma 1.30 *Let H_1, H_2 be Hilbert spaces and let $V : H_1 \to H_2$ be an isometry. Then $V^* = V^\dagger$.*

Proof: Let $x \in H_1$ and $y \in H_2$. Consider the inner product $\langle Vx, y \rangle_{H_2}$. We can decompose $y = y_1 + y_2$ with $y_1 \in \mathcal{R}(V)$ and $y_2 \in \mathcal{R}(V)^\perp$. Then $y_1 = Vz$ for some

$z \in H_1$ and

$$\begin{aligned}\langle Vx, y\rangle_{H_2} = \langle Vx, y_1 + y_2\rangle_{H_2} = \langle Vx, Vz\rangle_{H_2} = \langle x, z\rangle_{H_1} &= \langle x, V^\dagger Vz\rangle_{H_1} \\ &= \langle x, V^\dagger y_1\rangle_{H_1} \\ &= \langle x, V^\dagger (y_1 + y_2)\rangle_{H_1} \\ &= \langle x, V^\dagger y\rangle_{H_1}\end{aligned}$$

($V^\dagger V = \text{proj}_{\mathcal{N}(V)^\perp} = I$ because $\mathcal{N}(V)$ is trivial). This proves that $V^* = V^\dagger$. ■

We now note the following expansions:

$$\begin{aligned} T^* &= Vm_\sigma U^\dagger \\ T^*T &= Vm_{\sigma^2} V^\dagger \\ TT^* &= Um_{\sigma^2} U^\dagger \\ T^\dagger &= Vm_{\sigma^{-1}} U^\dagger \end{aligned} \tag{1.10}$$

The last equation holds by the following lemmas.

Lemma 1.31 *Let S be defined as in equation* (1.8) *and define $U(S) = \{Us : s \in S\}$. Then $U(S) = \mathcal{R}(T)$.*

Proof: If $s \in S$, then $Us = TVm_{\sigma^{-1}}s \in \mathcal{R}(T)$ by definition. Thus, $U(S) \subset \mathcal{R}(T)$. On the other hand, if $y = Tx$ for some $x \in X$, then $y = Um_\sigma V^\dagger x$ and $m_\sigma V^\dagger x \in S$. Therefore $\mathcal{R}(T) \subset U(S)$. This completes the proof. ■

Lemma 1.32 $T^\dagger = Vm_{\sigma^{-1}}U^\dagger$.

Proof: We will begin by proving that $\mathcal{D}(T^\dagger) = \mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$. First suppose $y \in \mathcal{D}(T^\dagger) = \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$. Write $y = y_1 + y_2$ for $y_1 \in \mathcal{R}(T)$ and $y_2 \in \mathcal{R}(T)^\perp$ and consider $U^\dagger y$. We can write $y_1 = Tx = Um_\sigma V^\dagger x$ for some $x \in X$ and note that $\mathcal{N}(U^\dagger) = \mathcal{R}(T)^\perp$ (which follows from the fact that $\mathcal{R}(U) = \overline{\mathcal{R}(T)}$, which we showed in Theorem 1.27). This implies

$$U^\dagger y = U^\dagger y_1 = U^\dagger U m_\sigma V^\dagger x = m_\sigma V^\dagger x \in S$$

where we have used the fact that $U^\dagger U = \text{proj}_{\mathcal{N}(U)^\perp} = I$ because $\mathcal{N}(U)$ is trivial. Thus, $U^\dagger y \in S$ which implies that $y \in \mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$ and thus, $\mathcal{D}(T^\dagger) \subset \mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$.

On the other hand, suppose $y \in \mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$. Here, we write $y = y_1 + y_2$ where $y_1 \in \overline{\mathcal{R}(T)}$ and $y_2 \in \mathcal{R}(T)^\perp$. Then y being in $\mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$ implies that $U^\dagger y = U^\dagger y_1 \in S$ Thus, $UU^\dagger y_1 \in U(S) = \mathcal{R}(T)$ by the previous lemma. The fact that $UU^\dagger = \text{proj}_{\mathcal{R}(U)} = \text{proj}_{\overline{\mathcal{R}(T)}}$ implies that $UU^\dagger y_1 = y_1$. Thus, $y_1 \in \mathcal{R}(T)$ which implies that $y \in \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$. So $\mathcal{D}(Vm_{\sigma^{-1}}U^\dagger) \subset \mathcal{D}(T^\dagger)$ and hence, $\mathcal{D}(T^\dagger) = \mathcal{D}(Vm_{\sigma^{-1}}U^\dagger)$.

Next, let $y \in \mathcal{D}(T^\dagger)$ and again decompose y as $y = y_1 + y_2$ with $y_1 \in \mathcal{R}(T) \subset \mathcal{R}(U)$ and $y_2 \in \mathcal{R}(T)^\perp = \mathcal{R}(U)^\perp$. Then $x = T^\dagger y$ is the unique element of $\mathcal{N}(T)^\perp$ such that $Tx = y_1$. We can check that $x = Vm_{\sigma^{-1}}U^\dagger y$ satisfies this condition. First, note that $x \in \mathcal{R}(V) = \mathcal{N}(T)^\perp$. In addition,

$$Tx = (Um_\sigma V^\dagger)(Vm_{\sigma^{-1}}U^\dagger)y = UU^\dagger y = \text{proj}_{\mathcal{R}(U)}y = \text{proj}_{\overline{\mathcal{R}(T)}}y = y_1,$$

where we have used the fact that $V^\dagger V = I = m_\sigma m_{\sigma^{-1}}$ on $L^2(M)$. Thus, for all $y \in \mathcal{D}(T^\dagger)$, $T^\dagger y = Vm_{\sigma^{-1}}U^\dagger y$. ■

As mentioned earlier in the chapter, the SVE is not unique. If T has two different SVE's given by $T = U_1 m_{\sigma_1} V_1^\dagger = U_2 m_{\sigma_2} V_2^\dagger$, we can prove the below theorem regarding the spaces $L^2(M_1)$ and $L^2(M_2)$.

Theorem 1.33 *Let $T : X \to Y$ be a bounded linear operator with two distinct SVE's given by $T = U_1 m_{\sigma_1} V_1^\dagger$ and $T = U_2 m_{\sigma_2} V_2^\dagger$. Then the associated Hilbert spaces $L^2(M_1)$ and $L^2(M_2)$ are isometrically isomorphic.*

Proof: Consider the operator $V_1 : L^2(M_1) \to X$. This operator is an isometry. The operator $\hat{V}_1 : L^2(M_1) \to \mathcal{R}(V_1) = \mathcal{N}(T)^\perp$ with $\hat{V}_1 f = V_1 f \ \forall f \in L^2(M_1)$, however, is an isometric isomorphism. In addition, $\hat{V}_1^{-1} : \mathcal{N}(T)^\perp \to L^2(M_1)$, which is equivalent to $V_1^\dagger\big|_{\mathcal{N}(T)^\perp}$, is an isometric isomorphism. Using similar reasoning, we know that $\hat{V}_2 : L^2(M_2) \to \mathcal{R}(V_2) = \mathcal{N}(T)^\perp$ is an isometric isomorphism and so is $\hat{V}_2^{-1} : \mathcal{N}(T)^\perp \to L^2(M_2)$. Thus, $\hat{V}_2^{-1}\hat{V}_1$ is an isometric isomorphism from $L^2(M_1)$ to $L^2(M_2)$. In addition, $\hat{V}_1^{-1}\hat{V}_2$ is an isometric isomorphism from $L^2(M_2)$ to $L^2(M_1)$.

Likewise, we could construct isometric isomorphisms $\hat{U}_1^{-1}\hat{U}_2 : L^2(M_2) \to L^2(M_1)$ and $\hat{U}_2^{-1}\hat{U}_1 : L^2(M_1) \to L^2(M_2)$. ■

Chapter 2

Basic properties of the SVE of a bounded linear operator

2.1 Introduction

In the previous chapter, we derived the singular value expansion (SVE) $T = Um_\sigma V^\dagger$ of a bounded operator T. Here, we will define and discuss the essential range of a measurable function and the essential preimage of a set in $\mathbb{R}$ under the same function. This chapter is motivated by the paucity of references in the literature regarding the essential range as well as the complete lack of references in the literature on the essential pre-image. In [24], Rudin discusses the essential range only in a few of the end-of-chapter problems. Because of this, we will prove several basic properties of the essential range that are difficult to find in the literature. We will also define the essential pre-image of a set $S \subset \mathbb{R}$ and derive several of its properties. Later on, these results will be used to derive conditions on the multiplication operator m_σ that guarantee T is compact.

We start by discussing assumptions and notation. Throughout this chapter, we will be working with a measure space $(M, \mathcal{A}, \mu)$. We assume there is a topology $\mathcal{T}$ on the set M and that the collection $\mathcal{A}$ is the associated Borel σ-algebra defined on M. We also assume that $\mathcal{T}$ has a countable base $\mathcal{B}$ (which means the topology is *second countable*). Finally, we assume $\mathcal{T}$ is Hausdorff. These assumptions will be used many times in the proofs of this chapter, but we will avoid explicitly stating them in every theorem or lemma.

The motivation for defining the essential range of a measurable function f is that the range of f is not well defined. If f, g are measurable functions and $f = g$ except on a set of measure zero, then $f = g$, but $\mathcal{R}(f)$ may not be equal to $\mathcal{R}(g)$.

The essential range is defined below and does not change when the function is altered on a measure zero set. The reader can also consult [24] for a definition of the essential range.

Definition 2.1 (Essential range) *Let $f : M \to \mathbb{R}$ be a measurable function. The essential range of f is the set*

$$\begin{aligned}\mathcal{R}_{ess}(f) &= \{s \in \mathbb{R} : \forall \epsilon > 0,\ \mu(\{t \in M : |f(t) - s| < \epsilon\}) > 0\} \\ &= \{s \in \mathbb{R} : \forall \epsilon > 0,\ \mu(f^{-1}(s - \epsilon, s + \epsilon)) > 0\}.\end{aligned} \tag{2.1}$$

We also note that for a measurable function f, the set $f^{-1}(s - \epsilon, s + \epsilon)$ is not well defined because we may change f on a set of measure zero and thus change the set $f^{-1}(s - \epsilon, s + \epsilon)$. However, the *measure* of this set *is* well-defined.

Because $f^{-1}(S)$ for some subset $S \subset \mathbb{R}$ is also not well defined, we next define the essential preimage of a set.

Definition 2.2 (Essential preimage) *Let $f : M \to \mathbb{R}$ be a measurable function. Given a measurable subset $S \subset \mathbb{R}$, the essential preimage of S under f is defined by*

$$f_{ess}^{-1}(S) = M \setminus \bigcup \{E \in \mathcal{T} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\} \tag{2.2}$$

Before exploring the properties of these sets in the next section, we derive two alternate definitions for $f_{ess}^{-1}(S)$ that we will use in future proofs.

Lemma 2.3 *Let $f : M \to \mathbb{R}$ be a measurable function. If S is a measurable subset of $\mathbb{R}$, then*

$$f_{ess}^{-1}(S) = \{t \in M : \forall E \in \mathcal{T}, t \in E \implies S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset\}. \tag{2.3}$$

Proof: Define $U = \{t \in M : \forall E \in \mathcal{T}, t \in E \implies S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset\}$. Let $t \in U$ and $E \subset M$ be an open neighborhood of t. Then $S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset$, which implies that

$$t \notin \bigcup \{E \in \mathcal{T} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\} \tag{2.4}$$

and thus, $t \in f_{ess}^{-1}(S)$.

Conversely, suppose $t \in f_{ess}^{-1}(S)$. Then (2.4) holds and for any open $E \subset M$, we know $S \cap \mathcal{R}_{ess}(f|_E) = \emptyset$ implies $t \notin E$. This is equivalent to saying if $E \subset M$ is open, then $t \in E$ implies $S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset$, which implies $t \in U$. This concludes the proof. ■

We next wish to prove a simplification of the definition of essential preimage which makes use of the fact that $\mathcal{T}$ is second countable. We will use this result several times in future sections. Before we can prove this lemma however, we must prove a basic property about the essential range.

Lemma 2.4 *Let $f : M \to \mathbb{R}$ be a measurable function and let E be a measurable subset of M. Then $\mathcal{R}_{ess}(f|_E) \subset \mathcal{R}_{ess}(f)$.*

Proof: Because $\{t \in E : |f(t) - s| < \epsilon\} \subset \{t \in M : |f(t) - s| < \epsilon\}$, then

$$\begin{aligned} s \in \mathcal{R}_{ess}(f|_E) &\implies \forall \epsilon > 0,\ \mu(\{t \in E : |f(t) - s| < \epsilon\}) > 0 \\ &\implies \forall \epsilon > 0,\ \mu(\{t \in M : |f(t) - s| < \epsilon\}) > 0 \\ &\implies s \in \mathcal{R}_{ess}(f). \end{aligned}$$

This completes the proof. ■

Lemma 2.5 *Let $f : M \to \mathbb{R}$ be a measurable function. If S is a measurable subset of $\mathbb{R}$, then*

$$f_{ess}^{-1}(S) = M \setminus \bigcup \{E \in \mathcal{B} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\} \tag{2.5}$$

and

$$f_{ess}^{-1}(S) = \{t \in M : \forall E \in \mathcal{B}, t \in E \implies S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset\} \tag{2.6}$$

where $\mathcal{B}$ denotes the countable base of $\mathcal{T}$.

Proof: In this proof, we will make use of the assumptions that $\mathcal{T}$, the topology on M, is Hausdorff and second countable. To prove the result in (2.5), it suffices to show that

$$\bigcup \{E \in \mathcal{B} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\} = \bigcup \{E \in \mathcal{T} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\}.$$

If $t \in \bigcup \{E \in \mathcal{B} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\}$ then of course $t \in \bigcup \{E \in \mathcal{T} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\}$ because $\mathcal{B} \subset \mathcal{T}$. To prove the converse, if $t \in \bigcup \{E \in \mathcal{T} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\}$, then there exists $E \in \mathcal{T}$ such that $t \in E$ and $S \cap \mathcal{R}_{ess}(f|_E) = \emptyset$. But then there exists some $E' \in \mathcal{B}$ such that $t \in E' \subset E$ and $S \cap \mathcal{R}(f|_{E'}) = \emptyset$ (since $\mathcal{R}(f|_{E'}) \subset \mathcal{R}(f|_E)$). Thus, $t \in \bigcup \{E \in \mathcal{B} : S \cap \mathcal{R}_{ess}(f|_E) = \emptyset\}$. This proves (2.5).

To prove (2.6), we show that the two sets

$$U_1 = \{t \in M : \forall E \in \mathcal{T}, t \in E \implies S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset\}$$

and

$$U_2 = \{t \in M : \forall E \in \mathcal{B}, t \in E \implies S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset\}$$

are equal. Let $t \in U_1$ and $E \in \mathcal{B}$ satisfy $t \in E$. Because $\mathcal{B} \subset \mathcal{T}$, $E \in \mathcal{T}$ as well and t being in U_1 implies $S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset$. Thus $t \in U_2$.

To prove the converse, let $t \in U_2$ and let $E \in \mathcal{T}$ such that $t \in E$. Then there exists $E' \in \mathcal{B}$ such that $t \in E' \subset E$. Moreover, $t \in U_2$ and $t \in E'$ imply that $S \cap \mathcal{R}_{ess}(f|_{E'}) \neq \emptyset$. Because $\mathcal{R}_{ess}(f|_{E'}) \subset \mathcal{R}_{ess}(f|_E)$ by Lemma 2.4, this implies that $S \cap \mathcal{R}_{ess}(f|_E) \neq \emptyset$, and we have shown that $t \in U_1$. This completes the proof. ■

In Section 2.2, we will prove some more basic properties of the essential range and preimage. Section 2.3 will be devoted to proving results regarding isolated points in the essential range. We will then use all of these results to relate properties of the essential range of m_σ to the compactness of the operator T in Section 2.4. We will present several concluding remarks in Section 2.5.

2.2 Properties of the essential range/preimage

In this section, we will prove several basic properties of the essential range and then one relation between the essential range and essential preimage. Many of these results are straightforward to prove, but are difficult to find in the literature.

The first two lemmas show that $\mathcal{R}_{ess}(f)$ is a closed set and is contained in $\overline{\mathcal{R}(f)}$.

Lemma 2.6 *If $f : M \to \mathbb{R}$ is measurable, then $\mathcal{R}_{ess}(f)$ is closed.*

Proof: It suffices to show that $\mathbb{R} \setminus \mathcal{R}_{ess}(f)$ is open. Suppose $s \in \mathbb{R} \setminus \mathcal{R}_{ess}(f)$. Then there exists $\epsilon > 0$ such that $\mu(f^{-1}(s-\epsilon, s+\epsilon)) = 0$. For any $s' \in (s-\epsilon/2, s+\epsilon/2)$, we know

$$f^{-1}(s'-\epsilon/2, s'+\epsilon/2) \subset f^{-1}(s-\epsilon, s+\epsilon)$$

and thus,

$$\mu(f^{-1}(s'-\epsilon/2, s'+\epsilon/2)) = 0.$$

Thus, if $s' \in (s-\epsilon/2, s+\epsilon/2)$, then $s' \notin \mathcal{R}_{ess}(f)$. Hence $\mathbb{R} \setminus \mathcal{R}_{ess}(f)$ is open. ■

Lemma 2.7 *Let $f : M \to \mathbb{R}$ be measurable. Then $\mathcal{R}_{ess}(f) \subset \overline{\mathcal{R}(f)}$.*

Proof: We prove the contrapositive. Suppose $s \notin \overline{\mathcal{R}(f)}$. Then there exists some $\epsilon > 0$ such that $(s-\epsilon, s+\epsilon) \cap \overline{\mathcal{R}(f)} = \emptyset$ and hence $f^{-1}(s-\epsilon, s+\epsilon) = \emptyset$. This implies that $s \notin \mathcal{R}_{ess}(f)$. ■

Our next lemma demonstrates that if the underlying set M has positive measure, then the essential range is always nonempty.

Lemma 2.8 *If $f : M \to \mathbb{R}$ is measurable and $\mu(M) > 0$, then $\mathcal{R}_{ess}(f) \neq \emptyset$.*

Proof: Note that $\mathbb{R} = \bigcup_{n\in\mathbb{Z}}(n-1, n]$ and thus

$$M = f^{-1}(\mathbb{R}) = \bigcup_{n\in\mathbb{Z}} f^{-1}((n-1, n])$$

$$\implies \mu(M) = \sum_{n\in\mathbb{Z}} \mu(f^{-1}((n-1, n])),$$

which implies that $\mu(f^{-1}((n-1, n])) > 0$ for at least one $n \in \mathbb{Z}$. Thus, there exists an interval $[r_1, r_2]$ such that $\mu(E) > 0$ where $E = f^{-1}([r_1, r_2])$. By Lemma 2.4, it suffices to show that $\mathcal{R}_{ess}(f|_E) \neq \emptyset$. We argue by contradiction. Let $F = f|_E$ and assume $\mathcal{R}_{ess}(F) = \emptyset$. Then for all $s \in [r_1, r_2]$, there exists some $\epsilon_s > 0$ such that

$$\mu(\{t \in E : |F(t) - s| < \epsilon_s\}) = 0.$$

Since $[r_1, r_2]$ is compact, there exists a finite sequence $s_1, s_2, ..., s_n$ such that

$$[r_1, r_2] \subset \bigcup_{j=1}^{n} (s_j - \epsilon_{s_j}, s_j + \epsilon_{s_j})$$

$$\implies F^{-1}([r_1, r_2]) \subset \bigcup_{j=1}^{n} \{t \in E : |F(t) - s_j| < \epsilon_{s_j}\}$$

$$\implies \mu(F^{-1}([r_1, r_2])) \leq \sum_{j=1}^{n} \mu(\{t \in E : |F(t) - s_j| < \epsilon_{s_j}\}) = 0.$$

But this contradicts the assumption that $\mu(f^{-1}([r_1, r_2])) > 0$. ■
Next, we prove that the intersection of $\mathcal{R}(f)$ and $\mathcal{R}_{ess}(f)$ cannot be empty.

Lemma 2.9 *Suppose $f : M \to \mathbb{R}$ is measurable and $\mu(M) > 0$. Then*

$$\mathcal{R}(f) \cap \mathcal{R}_{ess}(f) \neq \emptyset.$$

Proof: We will assume that $\mathcal{R}(f) \cap \mathcal{R}_{ess}(f) = \emptyset$ and prove that the measure of M must be zero. Given $s \in \mathcal{R}(f)$, the fact that $s \notin \mathcal{R}_{ess}(f)$ implies that there exists an open interval (a_s, b_s) containing s and such that $\mu(f^{-1}(a_s, b_s)) = 0$. Moreover, since the topological space $\mathbb{R}$ is second-countable, there exists a countable collection $\{(a_n, b_n) : n \in N\}$ of open intervals such that

$$\mathcal{R}(f) \subset \bigcup_{n \in N} (a_n, b_n)$$

where N denotes a countable index set and $\mu(f^{-1}(a_n, b_n)) = 0$ for all $n \in N$. But then

$$M = f^{-1}(\mathcal{R}(f)) \subset f^{-1}\left(\bigcup_{n \in N} (a_n, b_n)\right) = \bigcup_{n \in N} f^{-1}(a_n, b_n)$$

$$\Rightarrow \mu(M) \leq \sum_{n \in N} \mu(f^{-1}(a_n, b_n)) = 0.$$

This completes the proof. ■

Corollary 2.10 *Let $f : M \to \mathbb{R}$ be measurable. If $U \subset M$ and $f(U) \cap \mathcal{R}_{ess}(f) = \emptyset$, then $\mu(U) = 0$.*

Proof: If $\mu(U) > 0$, then Lemma 2.9 implies that

$$\mathcal{R}(f|_U) \cap \mathcal{R}_{ess}(f|_U) \neq \emptyset.$$

Since $\mathcal{R}(f|_U) = f(U)$ and $\mathcal{R}_{ess}(f|_U) \subset \mathcal{R}_{ess}(f)$, it follows that

$$\mu(U) > 0 \implies f(U) \cap \mathcal{R}_{ess}(f) \neq \emptyset,$$

which proves the contrapositive. ■

We end this section by proving that the essential preimage of the essential range is equal to all of M save for a set of measure zero.

Lemma 2.11 *If $f : M \to \mathbb{R}$ is essentially bounded, then*

$$\mu\left(M \setminus f_{ess}^{-1}(\mathcal{R}_{ess}(f))\right) = 0.$$

Proof: By Lemma 2.5,

$$\begin{aligned} M \setminus f_{ess}^{-1}(\mathcal{R}_{ess}(f)) &= \bigcup \{E \in \mathcal{B} : \mathcal{R}_{ess}(f) \cap \mathcal{R}_{ess}(f|_E) = \emptyset\} \\ &= \bigcup \{E \in \mathcal{B} : \mathcal{R}_{ess}(f|_E) = \emptyset\}. \end{aligned}$$

By Lemma 2.10, $\mathcal{R}_{ess}(f|_E) = \emptyset$ implies that $\mu(E) = 0$. Thus, $M \setminus f_{ess}^{-1}(\mathcal{R}_{ess}(f))$ is a countable union of open sets of measure zero and hence is itself a set of measure zero. ■

2.3 Isolated points

We now turn our attention to points $s \in \mathcal{R}_{ess}(f)$ that are isolated from all other points in $\mathcal{R}_{ess}(f)$. We say a point $s \in \mathcal{R}_{ess}(f)$ is an isolated point of $\mathcal{R}_{ess}(f)$ if there exists an open interval (a, b) such that $\mathcal{R}_{ess}(f) \cap (a, b) = \{s\}$. We will eventually prove that such points are eigenvalues of the multiplication operator m_f. We will also prove that in order to guarantee that $T = U m_\sigma V^\dagger$ is a compact operator, that the $\mathcal{R}_{ess}(\sigma)$ must consist of either a finite collection of isolated points, or a countably infinite collection of isolated points whose only limit point is zero.

In this section, we will derive a number of properties of the set $f^{-1}(\{s\})$, given that s in an isolated point of $\mathcal{R}_{ess}(f)$. These results will then be used in the next section.

Lemma 2.12 *Let s be an isolated point of $\mathcal{R}_{ess}(f)$. Then $t \in f_{ess}^{-1}(\{s\})$ if and only if for all open $E \subset M$ containing t, we have $\mu(E \cap f^{-1}(\{s\})) > 0$.*

Proof: Start by choosing an $\epsilon > 0$ such that $\mathcal{R}_{ess}(f) \cap (s-\epsilon, s+\epsilon) = \{s\}$. Let $t \in f_{ess}^{-1}(\{s\})$ and choose an $E \in \mathcal{T}$ containing t. Then from the definition of the essential preimage, we have

$$\begin{aligned} t \in f_{ess}^{-1}(\{s\}) &\implies \forall E \in \mathcal{T}, t \in E \implies s \in \mathcal{R}_{ess}(f|_E) \\ &\implies \forall E \in \mathcal{T}, t \in E \implies (\forall \epsilon' > 0, \mu(\{t' \in E : |f(t') - s| < \epsilon'\}) > 0) \end{aligned}$$

Thus, we have $\mu(\{x \in E \mid |f(x) - s| < \epsilon\}) > 0$. Then write

$$\{x \in E \mid |f(x)-s| < \epsilon\} = (E \cap f^{-1}(\{s\})) \bigcup (E \cap f^{-1}(s-\epsilon, s)) \bigcup (E \cap f^{-1}(s, s+\epsilon)).$$

Suppose the set $E \cap f^{-1}(s-\epsilon, s)$ has positive measure. Then for some $n \in \mathbb{Z}^+$, the set S_n defined by $S_n = E \cap f^{-1}[s-\epsilon+1/n, s-1/n]$ must have positive measure. Consider $f|_{S_n}$. Clearly, $\mathcal{R}(f|_{S_n}) \subset [s-\epsilon+1/n, s-1/n]$. Also, the assumption that $\mu(S_n) > 0$ along with Lemma 2.10 imply that $\mathcal{R}(f|_{S_n}) \cap \mathcal{R}_{ess}(f|_{S_n}) \neq \emptyset$. But this contradicts the fact that $\mathcal{R}_{ess}(f) \cap (s-\epsilon, s+\epsilon) = \{s\}$. Thus, $\mu(E \cap f^{-1}(s-\epsilon, s)) = 0$. A similar argument holds for the set $E \cap f^{-1}(s, s+\epsilon)$. Thus, the set $E \cap f^{-1}(\{s\})$ must have positive measure.

To prove the converse, suppose for some $t \in M$, every $E \in \mathcal{T}$ containing t satisfies $\mu(E \cap f^{-1}(\{s\})) > 0$. Then of course $0 < \mu(E \cap f^{-1}(\{s\})) \leq \mu(E \cap f^{-1}(s-\epsilon, s+\epsilon))$ for any $\epsilon > 0$. Thus, $t \in f_{ess}^{-1}(\{s\})$. ■

Lemma 2.13 *Let $f : M \to \mathbb{R}$ be a measurable function. If s is an isolated point of $\mathcal{R}_{ess}(f)$ with*

$$\mathcal{R}_{ess}(f) \cap (s-\epsilon, s+\epsilon) = \{s\}$$

for some $\epsilon > 0$, then $\mu(f^{-1}(\{s\})) > 0$ and $f(t) = s$ for almost every $t \in f^{-1}(s-\epsilon, s+\epsilon)$

Proof: Let $U = f^{-1}(s-\epsilon, s+\epsilon)$ and let $V = U \backslash f^{-1}(\{s\})$. Since $s \in \mathcal{R}_{ess}(f)$, $\mu(U) > 0$. Also, since $f(V) \subset (s-\epsilon, s+\epsilon)$ and $s \notin f(V)$, it follows that $f(V) \cap \mathcal{R}_{ess}(f) = \emptyset$ and hence $\mu(V) = 0$ by Lemma 2.10. But then

$$U = f^{-1}(\{s\}) \cup V \implies \mu(U) = \mu(f^{-1}(\{s\})) + \mu(V) = \mu(f^{-1}(\{s\})).$$

This shows that $\mu(f^{-1}(\{s\})) > 0$ and also that $f(t) = s$ for almost every $t \in U$ as desired. ■

This lemma immediately gives us the below theorem.

Theorem 2.14 *Let $f : M \to \mathbb{R}$ be a measurable function. If s is an isolated point of $\mathcal{R}_{ess}(f)$, then s is an eigenvalue of the multiplication operator m_f on $L^2(M)$ and the characteristic function $\chi_{f^{-1}(\{s\})}$ is a corresponding eigenvector.*

Proof: Follows immediately from Lemma 2.13. ■

Next, we prove that $f^{-1}(S) \setminus f_{ess}^{-1}(S)$ is a set of measure zero.

Lemma 2.15 *Let $S \subset M$ be measurable. Then $\mu(f^{-1}(S) \setminus f_{ess}^{-1}(S)) = 0$.*

Proof: Define $N = f^{-1}(S) \setminus f_{ess}^{-1}(S)$. For each $t \in N$, there exists an open set E_t containing t such that $S \cap \mathcal{R}_{ess}(f|_{E_t}) = \emptyset$. Now define the open set $E = \bigcup_{t \in N} E_t$. Clearly, $S \cap \mathcal{R}_{ess}(f|_E) = \emptyset$ and thus, because $\mathcal{R}_{ess}(f|_N) \subset \mathcal{R}_{ess}(f|_E)$, we know $S \cap \mathcal{R}_{ess}(f|_N) = \emptyset$ as well. But $\mathcal{R}(f|_N) \subset S$ and thus $\mathcal{R}(f|_N) \cap \mathcal{R}_{ess}(f|_N) = \emptyset$, which by Lemma 2.9 implies that $\mu(N) = 0$. ■

Corollary 2.16 *For any measurable set $S \subset M$, $\mu(f^{-1}(S)) = \mu(f^{-1}(S) \cap f_{ess}^{-1}(S))$.*

Proof: Note that $f^{-1}(S) = (f^{-1}(S) \cap f_{ess}^{-1}(S)) \cup (f^{-1}(S) \setminus f_{ess}^{-1}(S))$. The result then follows immediately from the Lemma 2.15. ■

Corollary 2.17 *Let $f : M \to \mathbb{R}$ be a measurable function and s be an isolated point of $\mathcal{R}_{ess}(f)$. Then $t \in f_{ess}^{-1}(\{s\})$ if and only if for all open sets $E \subset M$ containing t, we have $\mu(E \cap f^{-1}(\{s\}) \cap f_{ess}^{-1}(\{s\})) > 0$.*

Proof: We already know from Lemma 3.7 that $t \in f_{ess}^{-1}(\{s\})$ if and only if for all open sets $E \subset M$ containing t, we have $\mu(E \cap f^{-1}(\{s\})) > 0$. Now note that

$$f^{-1}(\{s\}) = (f^{-1}(\{s\}) \setminus f_{ess}^{-1}(\{s\})) \bigcup (f^{-1}(\{s\}) \cap f_{ess}^{-1}(\{s\}))$$

and $f^{-1}(\{s\}) \setminus f_{ess}^{-1}(\{s\})$ is a measure zero set by Lemma 2.15. Thus, $\mu(E \cap f^{-1}(\{s\})) = \mu(E \cap f^{-1}(\{s\}) \cap f_{ess}^{-1}(\{s\}))$. ■

Now we wish to prove a relation between isolated points $s \in \mathcal{R}_{ess}(f)$ and isolated points $\bar{t} \in f_{ess}^{-1}(\{s\})$. Before we do this, we prove the following lemma.

Lemma 2.18 *Let $f : M \to \mathbb{R}$ be a measurable function and let S be a subset of $\mathbb{R}$. Then for any open $E \subset M$,*

$$(f|_E)_{ess}^{-1}(S) = f_{ess}^{-1}(S) \cap E.$$

Proof: We have

$$t \in (f|_E)_{ess}^{-1}(S) \iff (\forall E' \subset E, E' \text{ open and } t \in E' \implies S \cap \mathcal{R}_{ess}(f|_{E'}) \neq \emptyset) \tag{2.7}$$

and

$$\begin{aligned} &t \in f_{ess}^{-1}(S) \cap E \iff \\ &t \in E \text{ and } \forall E'' \subset M, E'' \text{ open and } t \in E'' \implies S \cap \mathcal{R}_{ess}(f|_{E''}) \neq \emptyset. \end{aligned} \tag{2.8}$$

We will show that (2.7) and (2.8) are equivalent; it will then follow that

$$(f|_E)_{ess}^{-1}(S) = f_{ess}^{-1}(S) \cap E.$$

Suppose first that $t \in (f|_E)^{-1}_{ess}(S)$ and let E'' be an open neighborhood of t. Then $E' = E'' \cap E$ is an open subset of E such that $t \in E'$ and hence (2.7) implies that $S \cap \mathcal{R}_{ess}(f|_{E'}) \neq \emptyset$. Since $\mathcal{R}_{ess}(f|_{E'}) \subset \mathcal{R}_{ess}(f|_{E''})$ and t obviously belongs to E, it follows that the right-hand condition of (2.8) is satisfied and hence $t \in f^{-1}_{ess}(S) \cap E$. Conversely, suppose $t \in f^{-1}_{ess}(S) \cap E$ and let E' be an open subset of E containing t. Since E' is also an open subset of M, it follows from (2.8) that $S \cap \mathcal{R}_{ess}(f|_{E'}) \neq \emptyset$ and we see that the right-hand condition of (2.7) is satisfied. Therefore $t \in (f|_E)^{-1}_{ess}(S)$. This completes the proof. ■

Lemma 2.19 *Let $f : M \to \mathbb{R}$ be a measurable function. If $s \in \mathbb{R}$ is an isolated point of $\mathcal{R}_{ess}(f)$ and $\bar{t} \in M$ is an isolated point of $f^{-1}_{ess}(\{s\})$, then $\mu(\{\bar{t}\}) > 0$. Moreover, $f(\bar{t}) = s$.*

Proof: By assumption, there exists an open subset $E \subset M$ such that

$$f^{-1}_{ess}(\{s\}) \cap E = \{\bar{t}\}.$$

By definition of $f^{-1}_{ess}(\{s\})$, we know $s \in \mathcal{R}_{ess}(f|_E)$ and also that s is an isolated point of $\mathcal{R}_{ess}(f|_E)$. Hence, there is an $\epsilon > 0$ such that

$$\mathcal{R}_{ess}(f|_E) \cap (s - \epsilon, s + \epsilon) = \{s\}.$$

It follows from Lemma 2.13 and Corollary 2.16 that

$$\mu((f|_E)^{-1}_{ess}(\{s\})) \geq \mu((f|_E)^{-1}_{ess}(\{s\}) \cap (f|_E)^{-1}(\{s\})) = \mu((f|_E)^{-1}(\{s\})) > 0.$$

Next, Lemma 2.18 implies that

$$(f|_E)^{-1}_{ess}(\{s\}) = f^{-1}_{ess}(\{s\}) \cap E = \{\bar{t}\},$$

and hence, $\mu(\{\bar{t}\}) > 0$. We know from Lemma 2.13 that $f(t) = s$ for almost every $t \in (f|_E)^{-1}(s - \epsilon, s + \epsilon)$. But Lemma 2.15 implies that

$$(f|_E)^{-1}(s - \epsilon, s + \epsilon) = (f|_E)^{-1}_{ess}(s - \epsilon, s + \epsilon) \cup V = \{\bar{t}\} \cup V,$$

where V is a set of measure zero. It follows that $f(\bar{t}) = s$, and the proof is complete. ■

We have already shown that isolated points $s \in \mathcal{R}_{ess}(f)$ are eigenvalues of m_f. We now prove a result regarding the associated eigenvectors.

Lemma 2.20 *Let $f : M \to \mathbb{R}$ be a measurable function. If $s \in \mathbb{R}$ is an isolated point of $\mathcal{R}_{ess}(f)$, then s is an eigenvalue of m_f and any eigenvector $g : M \to \mathbb{R}$ of m_f corresponding to s is equal to χ_S for some subset $S \subset f^{-1}(\{s\}) \cap f^{-1}_{ess}(\{s\})$ of positive measure.*

Proof: Define $S_g = \{t \in M \mid g(t) \neq 0\}$ and $U = f^{-1}(\{s\}) \cap f^{-1}_{ess}(\{s\})$. Then

$$S_g = (S_g \cap (M \setminus f^{-1}(\{s\}))) \bigcup (S_g \cap (f^{-1}(\{s\}) \setminus f^{-1}_{ess}(\{s\}))) \bigcup (S_g \cap U)$$

and the set $S_g \cap M \setminus f^{-1}(\{s\})$ must have measure zero (otherwise g would not be an eigenvector). The set $S_g \cap (f^{-1}(\{s\}) \setminus f^{-1}_{ess}(\{s\}))$ also has measure zero by Lemma 2.15. This completes the proof. ■

2.4 Compact operators and the SVE

In this section, our goal is to derive a set of conditions on the multiplication operator m_σ that guarantees that the operator $T = Um_\sigma V^\dagger$ is compact. To do this, we start by proving several results about measurable sets that will be important for us later.

Lemma 2.21 *If $U \subset M$ has positive measure, then there exists $t \in U$ such that*

$$\forall E \in \mathcal{T},\ t \in E \implies \mu(E \cap U) > 0.$$

Proof: Because $\mathcal{T}$ is second countable with countable base $\mathcal{B}$, it suffices to prove that for all $E \in \mathcal{B}$, $t \in E$ implies $\mu(E \cap U) > 0$. By contradiction, assume for all $t \in U$, there exists some $E_t \in \mathcal{B}$ containing t such that $\mu(E_t \cap U) = 0$. Now notice $U = \bigcup\{E_t \cap U \mid t \in U\}$. Because $\mathcal{B}$ is countable, the set $\{E_t \cap U \mid t \in U\}$ will also be countable. Thus, $\mu(U) \leq \sum \mu(E_t \cap U) = 0$. This contradiction completes the proof. ■

Next, we consider a measurable set U such that U cannot be partitioned into two disjoint sets $U = U_1 \cap U_2$ with both U_1 and U_2 having positive measure.

Lemma 2.22 *Let $U \in \mathcal{A}$ be a set with positive measure such that*

$$\forall\ U_1, U_2 \in \mathcal{A}, ((U_1 \cap U_2 = \emptyset \text{ and } U = U_1 \cup U_2) \implies (\mu(U_1) = 0 \text{ or } \mu(U_2) = 0)).$$

Then there exists some $t \in M$ such that $U = \{t\} \cup U_0$, where $U_0 \in \mathcal{A}$, $\mu(\{t\}) = \mu(U)$, and $\mu(U_0) = 0$.

Proof: Let $t \in U$ be the element in U guaranteed by Lemma 2.21. Define $U_0 = U \setminus \{t\}$. Then of course $U = \{t\} \cup U_0$ and $\{t\} \cap U_0 = \emptyset$. For any open set $E \subset M$ containing t, we know that

$$U = (U \cap E) \bigcup (U \setminus E)$$

and $\mu(U \cap E) > 0$ by Lemma 2.21, and by assumption, $\mu(U \setminus E) = 0$. Then because $\mathcal{T}$ is second-countable, there exists a countable collection of open sets $\{F_j\}_{j=1}^{\infty}$ with

$t \in F_j\ \forall j \in \mathbb{Z}^+$ such that for all $t' \in U$ with $t' \neq t$, there exists at least one $F_{j'}$ such that $t' \notin F_{j'}$. Now define the sequence of sets $\{E_j\}_{j=1}^{\infty}$ by

$$E_j = \bigcap_{k=1}^{j} F_k$$

Then each E_j, being a finite intersection of open sets, is open and each E_j contains t. Also, $E_{n+1} \subset E_n$ for all $n \in \mathbb{Z}^+$ and it is clear that $\bigcap_{j=1}^{\infty} E_j = \{t\}$.

We thus, conclude that $\mu(U) = \mu(U \cap E_j)$ for all $j \in \mathbb{Z}^+$ and apply a well known theorem from measure theory to conclude

$$\mu(\{t\}) = \mu\left(\bigcap_{j=1}^{\infty}(U \cap E_j)\right) = \lim_{j\to\infty} \mu(U \cap E_j) = \lim_{j\to\infty} \mu(U) = \mu(U)$$

Thus, $\mu(U \setminus \{t\})$ must be equal to zero. This concludes the proof. ■

Our next result requires us to define a *positive partition* of a measurable set.

Definition 2.23 (Positive partition) *Let U be a measurable subset of M with $\mu(U) > 0$. The set $\{P_k\}_{k=1}^{n}$ is called a positive partition of U if $U = \bigcup_{k=1}^{n} P_k$, $P_k \cap P_m = \emptyset$ for $1 \leq k \neq m \leq n$ and $\mu(P_k) > 0$ for all k.*

We also define a partial ordering on the set of positive partitions of U.

Definition 2.24 (Partial ordering) *The partial ordering on the set of positive partitions of a measurable set $U \subset M$ is defined as*

$$\{P_k\}_{k=1}^{n} \leq \{Q_\ell\}_{\ell=1}^{m} \iff m \geq n \ \text{ and } \forall \ell \in \{1,2,...,m\},\ \exists k \in \{1,2,...,n\}, Q_\ell \subset P_k.$$

A chain of positive partitions is a collection of the form $\{P_k\}_{k=1}^{n} \leq \{Q_\ell\}_{\ell=1}^{m} \leq \{R_i\}_{i=1}^{p} \leq ....$ which may be infinite or finite.

Using these concepts, we prove the below lemma which demonstrates an important relationship between the cardinality of the set $f_{ess}^{-1}(\{s\})$ and the dimension of the eigenspace corresponding to the eigenvalue s.

Lemma 2.25 *Let $f : M \to \mathbb{R}$ be a measurable function. If $s \in \mathbb{R}$ is an isolated point of $\mathcal{R}_{ess}(f)$, then s is an eigenvalue of m_f with eigenspace $E_{m_f}(s)$, and*

$$\dim(E_{m_f}(s)) = n \iff f_{ess}^{-1}(\{s\}) \text{ contains exactly } n \text{ points.}$$

Proof: If s is an isolated point of $\mathcal{R}_{ess}(f)$, then by Corollary 2.16, s must be an eigenvalue of m_f. Suppose $f_{ess}^{-1}(\{s\})$ consists of exactly n elements. Then each $t \in f_{ess}^{-1}(\{s\})$ will be an isolated point of $f_{ess}^{-1}(\{s\})$ because M is a Hausdorff space. By

Lemma 2.19, this means $\mu(\{t\}) > 0$ and $f(t) = s$. Thus, if $f_{ess}^{-1}(\{s\}) = \{t_1, t_2, ..., t_n\}$, then $\{\chi_{\{t_1\}}, \chi_{\{t_2\}}, ..., \chi_{\{t_n\}}\}$ is an orthogonal collection of eigenvectors associated with s. Thus, the dimension of the eigenspace for s is at least n. Suppose $\dim(E_{m_f}(s)) > n$. Then there exists an eigenvector g associated with s that is orthogonal to $\chi_{\{t_i\}}$ for $i = 1, ..., n$. Also, by Lemma 2.20, g must be defined on $f_{ess}^{-1}(\{s\}) \cap f^{-1}(\{s\})$ except for a set of measure zero. Thus, $g = 0$ a.e. on the set M. This contradiction proves the eigenspace for s will have dimension exactly n.

To prove the converse, suppose the eigenspace of m_f associated with s has dimension n. Define the set $U = f_{ess}^{-1}(\{s\}) \cap f^{-1}(\{s\})$, and let $P_{pos}(U)$ be the collection of all positive partitions of U. Notice that for any positive partition $\{P_k\}_{k=1}^N$ of N subsets of U, the set of characteristic functions $\{\chi_{P_1}, \chi_{P_2}, ..., \chi_{P_N}\}$ is an orthogonal set of N eigenvectors of f corresponding to s.

Suppose there exists a chain of positive partitions in $P_{pos}(U)$ with no upper bound. Then the chain is an infinite sequence of positive partitions (otherwise the last element in the chain is an upper bound). If we assume, without loss of generality, that the elements in the chain are distinct, then the cardinalities of the elements in the chain form a strictly increasing sequence of positive integers that grow without bound. Then for any positive partition of the chain, $\{P_k\}_{k=1}^N$, the set of characteristic functions $\{\chi_{P_1}, \chi_{P_2}, ..., \chi_{P_N}\}$ is an orthogonal set of N eigenvectors of f corresponding to s. Thus, the dimension of the eigenspace cannot be bounded, a contradiction.

Thus, every chain in $P_{pos}(U)$ must have an upper bound. Zorn's lemma then implies there must exist some maximal positive partition $\{P_k\}_{k=1}^m$ such that each P_k cannot be partitioned into two subsets of positive measure. That is, for any $V_1, V_2 \subset P_k$ such that $V_1 \cap V_2 = \emptyset$ and $P_k = V_1 \cup V_2$, either $\mu(V_1) = 0$ or $\mu(V_2) = 0$. Thus, by Lemma 2.22, for each $k = 1, ..., m$ there is some $t_k \in P_k$ such that $P_k = \{t_k\} \cup P_{k,0}$ where $\mu(\{t_k\}) = \mu(P_k)$ and $\mu(P_{k,0}) = 0$.

We next prove that the sets $\{P_{k,0}\}$ are, in fact, empty. To do this, fix k and let $t \in P_{k,0} \subset U$. Then Corollary 2.17 implies that for every open set $E \subset M$ containing t, $\mu(E \cap U) > 0$. But because the topology on M is Hausdorff, we may find an open set E_t containing t such that $E_t \cap \{t_1, ..., t_m\} = \emptyset$. So then $\mu(E_t \cap U) = 0$ because $U \setminus \{t_1, ..., t_m\}$ is a set of measure zero. This contradiction shows that each $P_{k,0}$ must be empty.

This implies that $U = \{t_1, ..., t_m\}$ and each singleton set $\{t_j\}$ has positive measure. Thus $\{\chi_{\{t_1\}}, ..., \chi_{\{t_m\}}\}$ forms an orthogonal set of eigenvectors, which means $m \leq n$. Next, Lemma 2.20 implies that every eigenvector g associated with s must be equal to χ_S for some $S \subset U = \{t_1, ..., t_m\}$. This means that $n \leq m$ and therefore $m = n$.

Lastly, we must prove that the set $f_{ess}^{-1}(\{s\}) \setminus f^{-1}(\{s\})$ is empty. This will prove that $f_{ess}^{-1}(\{s\})$ consists of exactly n elements. To do this, suppose $t \in f_{ess}^{-1}(\{s\}) \setminus f^{-1}(\{s\})$. Then by definition of the essential preimage and range, for all $\epsilon > 0$, we have $\mu(f|_E^{-1}(s - \epsilon, s + \epsilon)) > 0$ for every open set E containing t. Again, select an open E_t such that $E_t \cap \{t_1, ..., t_n\} = \emptyset$. Lemma 2.13 then implies that the $f(r) = s$ for almost every $r \in (f|_{E_t})^{-1}(s-\epsilon, s+\epsilon)$. But that would imply that the characteristic function of the set $(f|_{E_t})^{-1}(s-\epsilon, s+\epsilon)$ is an eigenvector associated with s that is orthogonal to $\{t_1, ..., t_n\}$ which would contradict the fact that $\dim(E_{m_f}(s)) = n$. Thus, $f_{ess}^{-1}(\{s\})$ must consist of exactly n points. ■

Now that we have built up the necessary machinery, we can begin to prove several relationships between $T = Um_\sigma V^\dagger$ and the multiplication operator m_σ. We start with an easy result.

Theorem 2.26 *The range of $T = Um_\sigma V^\dagger$ fails to be closed if and only if the function σ is not bounded away from zero.*

Proof: By Corollary 2.17 of [10], $\mathcal{R}(T)$ will fail to be closed if and only if $T^\dagger = Vm_{\sigma^{-1}}U^\dagger$ is unbounded. $T^\dagger$ will fail to be bounded if and only if σ^{-1} fails to be essentially bounded. This holds if and only if σ is not bounded away from zero a.e. This completes the proof. ■

Next, we relate the spectrum of a self-adjoint operator $A : X \to X$ with SVE $A = Vm_\lambda V^{-1}$ to the essential range of the function λ. As in the previous chapter, to avoid ambiguity in our notation, we use $\Lambda(A)$ to denote the spectrum of A rather than the traditional $\sigma(A)$ notation.

Theorem 2.27 *If $A : X \to X$ is a bounded self-adjoint operator with spectral decomposition given by $A = Vm_\lambda V^{-1}$, then $\mathcal{R}_{ess}(\lambda) = \Lambda(A)$.*

Proof: Let $s \in \mathcal{R}_{ess}(\lambda)$. We will prove that s is an approximate eigenvalue of A and hence is contained in the spectrum of A. To show this, define

$$f_n = c_n \chi_{\lambda^{-1}\left(s-\frac{1}{n}, s+\frac{1}{n}\right)}$$

where χ_E denotes the characteristic function for the set E and c_n is a normalization constant such that $\|f_n\|_{L^2(M)} = 1$. It needs to be noted that $\lambda^{-1}\left(s-\frac{1}{n}, s+\frac{1}{n}\right)$ is not well defined because λ can be altered on a set of measure zero. However, $\chi_{\lambda^{-1}\left(s-\frac{1}{n}, s+\frac{1}{n}\right)}$ is well defined in $L^2(M)$. Note that it is possible to find a normalization c_n because $s \in \mathcal{R}_{ess}(\lambda)$ and hence $\chi_{\lambda^{-1}\left(s-\frac{1}{n}, s+\frac{1}{n}\right)} \neq 0$ in $L^2(M)$. We then define the sequence $\{x_n\} \subset X$ by $x_n = Vf_n$. Note that $\|x_n\|_X = 1$ for all n because V is an isometry. Then

$$\begin{aligned} \|Ax_n - sx_n\|_X^2 = \|Vm_\lambda V^{-1}x_n - sx_n\|_X^2 &= \|Vm_\lambda V^{-1}Vf_n - sVf_n\|_X^2 \\ &= \|m_\lambda f_n - sf_n\|_{L^2(M)}^2 \\ &= \int ((\lambda - s)f_n)^2 \, d\mu. \end{aligned}$$

We then note that the support of this integrand is $\lambda^{-1}\left(s-\frac{1}{n}, s+\frac{1}{n}\right)$, and hence $|\lambda - s| \leq \frac{1}{n}$. This allows us to conclude that

$$\int ((\lambda - s)f_n)^2 \, d\mu \leq \frac{1}{n^2}\int f_n^2 \, d\mu = \frac{1}{n^2} \to 0 \text{ as } n \to \infty.$$

Thus, there exists a sequence $\{x_n\} \subset X$ with $\|x_n\|_X = 1$ for all n such that $\|Ax_n - sx_n\|_X \to 0$, which means that s is an approximate eigenvalue of A and is hence contained in the spectrum of A. Thus, $\mathcal{R}_{ess}(\lambda) \subset \Lambda(A)$.

Conversely, suppose $s \notin \mathcal{R}_{ess}(\lambda)$. Then there exists an $\epsilon > 0$ such that

$$\mu(\lambda^{-1}(s-\epsilon, s+\epsilon)) = 0.$$

This means that λ is bounded away from s almost everywhere; that is, $|\lambda - s| \geq \epsilon$ a.e.. Thus, the operator $A - sI = Vm[\lambda - s]V^{-1}$ is a composition of invertible operators and hence, is invertible. It will have an inverse of the form

$$(A - sI)^{-1} = Vm\left[\frac{1}{\lambda - s}\right]V^{-1}.$$

It follows that s is not contained in the spectrum of A, and the proof is complete. ■

Corollary 2.28 *If $T : X \to Y$ is a bounded operator with singular value expansion given by $T = Um_\sigma V^\dagger$, then $\mathcal{R}_{ess}(\sigma) = \{s : s^2 \in \Lambda(T^*T)\}$.*

Proof: Apply Theorem 2.27 to the self-adjoint operator T^*T to obtain $\mathcal{R}_{ess}(\sigma^2) = \Lambda(T^*T)$. By the construction of the function σ, this means $\mathcal{R}_{ess}(\sigma) = \{s : s^2 \in \Lambda(T^*T)\}$. ■

We now prove the main result of this chapter, giving conditions on the multiplication operator m_σ that guarantee that $T = Um_\sigma V^\dagger$ is compact. We start with the self-adjoint case.

Theorem 2.29 *Let $A : X \to X$ be a bounded self-adjoint linear operator with spectral decomposition $A = Vm_\lambda V^{-1}$. Then A is compact if and only if one of the following is true.*

1. $\mathcal{R}_{ess}(\lambda)$ forms a sequence of nonzero numbers converging to zero (along with zero) and for each nonzero $s \in \mathcal{R}_{ess}(\lambda)$ the set $\lambda_{ess}^{-1}(\{s\})$ is finite. Or

2. $\mathcal{R}_{ess}(\lambda)$ is a finite set of nonnegative numbers and for each $s \in \mathcal{R}_{ess}(\lambda)$ the set $\lambda_{ess}^{-1}(\{s\})$ is finite.

Proof: Suppose A is compact. Then by Theorem 1.5, $\Lambda(A)$ consists of either a finite collection of nonnegative numbers or a sequence of positive real numbers that converge to zero (along with zero itself). Corollary 2.28 implies that $\mathcal{R}_{ess}(\lambda)$ must also have this property.

Now let s be a nonzero element in $\mathcal{R}_{ess}(\lambda)$. Because A is compact, s must be an eigenvalue of A with finite dimensional eigenspace. Lemma 2.25 then implies that $\lambda_{ess}^{-1}(\{s\})$ must be finite.

To prove the converse, suppose for every $s \in \mathcal{R}_{ess}(\lambda)$, the set $\lambda_{ess}^{-1}(\{s\})$ is finite. Then the set $\lambda_{ess}^{-1}(\mathcal{R}_{ess}(\lambda))$ is a countable collection $\{e_n\}$ of values in M, each e_n is an isolated point of M (because M is Hausdorff) and thus $\mu(\{e_n\}) > 0$, and $\chi_{\{e_n\}}$ is an eigenvector for m_λ with eigenvalue $\lambda_n = \lambda(e_n)$.

Also by Lemma 2.11, the set $M \setminus \lambda_{ess}^{-1}(\mathcal{R}_{ess}(\lambda))$ must have measure zero. This means we may define every function f in $L^2(M)$ on the set $\{e_n\}$. That is, we may write

$$f = \sum_n f(e_n)\chi_{\{e_n\}}$$

and also

$$m_\lambda f = \sum_n \lambda_n f(e_n)\chi_{\{e_n\}} = \sum_n \lambda_n \langle f, \chi_{\{e_n\}}\rangle_{L^2(M)}\chi_{\{e_n\}}.$$

Therefore,

$$m_\lambda = \sum_n \lambda_n \chi_{\{e_n\}} \otimes \chi_{\{e_n\}}$$

where by assumption, λ_n is either a finite set of nonnegative values or else a sequence of positive numbers satisfying $\lambda_n \to 0$. This implies that m_λ is compact and thus, that $A = Vm_\lambda V^{-1}$ is also compact. ■

We can now easily prove the non-self-adjoint version of this theorem.

Corollary 2.30 *Let $T : X \to Y$ be a bounded linear operator with singular value expansion $T = Um_\sigma V^\dagger$. Then T is compact if and only if one of the following is true.*
1. $\mathcal{R}_{ess}(\sigma)$ forms a sequence of nonzero numbers converging to zero (along with zero) and for each nonzero $s \in \mathcal{R}_{ess}(\sigma)$ the set $\sigma_{ess}^{-1}(\{s\})$ is finite. Or

2. $\mathcal{R}_{ess}(\sigma)$ is a finite set of nonnegative numbers and for each nonzero $s \in \mathcal{R}_{ess}(\sigma)$ the set $\sigma_{ess}^{-1}(\{s\})$ is finite.

Proof: Apply Theorem 2.29 to the self-adjoint operator $T^*T = Vm_{\sigma^2}V^\dagger$. Then the given properties hold for the function $\sigma^2 : M \to \mathbb{R}$ and thus must also hold for the positive square root $\sigma : M \to \mathbb{R}$. ■

2.5 Concluding remarks

We conclude this chapter by first discussing the ambiguity in defining the *singular values* of a bounded linear operator T that does not occur when the operator is a matrix, or is compact.

The singular values of a matrix $A \in \mathbb{R}^{m \times n}$ are familiar to the reader. There are two ways to conceptualize them. The first is to think of them as the square roots of the eigenvalues of $A^T A$. The second approach is to employ the min-max theorem given below.

Theorem 2.31 *Let $A \in \mathbb{R}^{m \times n}$ be a matrix with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$. Then*

$$\sigma_k = \min_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=n-k+1}} \max_{\substack{x \in S \\ \|x\|=1}} \|Ax\| \tag{2.9}$$

and

$$\sigma_k = \max_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=k}} \min_{\substack{x \in S \\ \|x\|=1}} \|Ax\|, \tag{2.10}$$

where $\| \cdot \|$ denotes the Euclidean norm.

Similarly, if $T : X \to Y$ is a compact operator, one can think of the singular values for T as the square roots of the elements in the spectrum of T^*T or one could use the following theorem, which is very similar to Theorem 2.31.

Theorem 2.32 *Let $T : X \to Y$ be a compact operator with singular values $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ....$ Then*

$$\sigma_k = \inf_{\substack{S \subset X \\ \dim(S)=k-1}} \sup_{\substack{x \in S^{\perp} \\ \|x\|_X=1}} \|Tx\|_Y \tag{2.11}$$

and

$$\sigma_k = \sup_{\substack{S \subset X \\ \dim(S)=k}} \inf_{\substack{x \in S \\ \|x\|_X=1}} \|Tx\|_Y \tag{2.12}$$

For the matrix and compact operator cases, these two approaches will each yield the same set of values we called the *singular values* of the operator. However, when we move to the more general, bounded operator case, the two approaches do *not* produce the same values. Consider the following example.

Example 2.33 *Let $M = [0, 1] \cup \{2, 3\}$ and define the measure space $(M, \mathcal{A}, \mu)$ with μ being Lebesgue measure on $[0, 1]$ and counting measure on $\{2, 3\}$. Let $T : L^2(M) \to L^2(M)$ be defined by $(Tf)(t) = tf(t)$ for each $f \in L^2(M)$.*

We will now determine the singular values of the operator T using the two approaches we discussed earlier. It is not hard to see that the spectrum of $T^*T = T^2$ is the set $[0, 1] \cup \{4, 9\}$. Taking the square roots of each value, we obtain the set $[0, 1] \cup \{2, 3\}$.

However, if we attempt to find the singular values using equations (2.11) *or* (2.12)*, we would find the values $\lambda_1 = 3$, $\lambda_2 = 2$, and then $\lambda_n = 1$ for $n \geq 3$. Thus, the two approaches for finding singular values of an operator do not produce the same set of* *values in the general setting where T is a bounded linear operator; in particular when the continuous or residual spectrum of the operator is nonempty.*

The example above illustrates that there is not a clear definition of singular values for a bounded linear operator. But if one wanted to choose one, we would argue that the essential range $\mathcal{R}_{ess}(\sigma)$, rather than the min-max theorems, would be the more useful definition. For one, the set $\mathcal{R}_{ess}(\sigma)$ is larger than the set one obtains from the min-max theorems. In addition, if $\mathcal{R}_{ess}(\sigma)$ is not bounded away from zero, then $\mathcal{R}(T)$ will fail to be closed and $Tx = y$ is an ill-posed problem. The min-max principle may not reveal that about T.

The last thing we will discuss in this chapter are some potential objections to our definition of the essential preimage. Our ultimate goal in this chapter was to derive results like the ones found in Theorem 2.29 and Corollary 2.30. Our definition of essential preimage was useful in that it allowed us to accomplish this. But whether or not our definition for the essential preimage is useful in all contexts is another question.

For instance, in Lemma 2.15 we proved that $\mu(f^{-1}(S) \setminus f_{ess}^{-1}(S)) = 0$, but we never proved that $\mu(f_{ess}^{-1}(S) \setminus f^{-1}(S)) = 0$. The reason for this is because $f_{ess}^{-1}(S) \setminus f^{-1}(S)$ *need not* be a measure zero set. Consider the following example.

Example 2.34 *Let $M = [0,1]$ and $(M, \mathcal{A}, \mu)$ be a measure space with μ being Lebesgue measure on $(0,1]$ and counting measure on $\{0\}$. Now define the function $f : M \to \mathbb{R}$ by*

$$f(x) = \begin{cases} 0 & x = 0 \\ 1 & x > 0. \end{cases}$$

Using our definition of the essential preimage, we find that $f_{ess}^{-1}(\{1\}) = [0,1]$. Then because $f^{-1}(\{1\}) = (0,1]$, we have

$$f_{ess}^{-1}(\{1\}) \setminus f^{-1}(\{1\}) = \{0\}$$

which is a set of positive measure. In addition, $\mu(f_{ess}^{-1}(\{1\})) = 2$ and $\mu(f^{-1}(\{1\})) = 1$.

This example illustrates a potentially undesirable property of our definition of the essential preimage. It also raises the following question: is it possible to produce a definition of essential preimage where $f_{ess}^{-1}(S)$ and $f^{-1}(S)$ are the same save for a set of measure zero? Producing a robust definition of the essential preimage (if there is one to be found) is a potential topic of future study.

Chapter 3

Tikhonov regularization

3.1 Introduction

Tikhonov regularization is a popular regularization method for ill-posed problems. In [14], Groetsch proves many of the standard results regarding Tikhonov regularization. Some of his proofs rely on the singular value expansion (SVE) of a compact operator. In [10], Gockenbach derives many of the usual Tikhonov regularization results without relying on the SVE at all. In this chapter, we demonstrate that the SVE of a bounded (not necessarily compact) operator may be used to derive many Tikhonov regularization results similar to those found in [14] or [10].

Before we begin, we recall the setting of Tikhonov regularization. We are interested in equations of the form $Tx = y$, where T is a bounded linear operator between two real, separable Hilbert spaces X and Y. In particular, we are interested in the case in which $\mathcal{R}(T)$ fails to be closed. When this holds, solving $Tx = y$ for x is an ill-posed problem because the solution x to $Tx = y$ (if it exists) does not depend continuously on the data y. To solve such an ill-posed problem, we employ a regularization method to compute a stable estimate of $T^\dagger y$, where $T^\dagger$ denotes the generalized inverse of T and is an unbounded operator when $\mathcal{R}(T)$ is not closed.

This chapter will be similar in structure to pages $15-52$ of [14]. In Section 3.2, we will derive a condition that guarantees the convergence of regularization methods to the solution of $Tx = y$. We will discuss convergence rates of these methods in Section 3.3. The case of inexact data will then be discussed in Section 3.4. We discuss Tikhonov regularization as a specific instance of these methods in Section 3.5. In Section 3.6 on converse results, we will prove that the convergence rates from Section 3.3 are, in fact, optimal. Finally, we will end the chapter by exploring the discrepancy principle in Section 3.7, which is based on the idea that if the error in the data y is bounded by some constant δ, then we will only try to make the residual $\|Tx - y\|_Y$ as small as δ.

3.2 Convergence

One way to approach an inverse problem of the form $Tx = y$ is to construct operators $S_\lambda : Y \to X$, $\lambda > 0$, that are bounded and approximate the generalized inverse $T^\dagger$ (which, when $\mathcal{R}(T)$ fails to be closed, is unbounded). These operators S_λ should satisfy

$$S_\lambda y \to T^\dagger y \text{ as } \lambda \to 0^+ \text{ for each } y \in \mathcal{D}(T^\dagger) = \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp.$$

For now, we will assume that y is known exactly. The vector $x_{0,y} = T^\dagger y$ satisfies $T^*Tx_{0,y} = T^*y$ and $x_{0,y} \in \mathcal{N}(T)^\perp$. Therefore, we attempt to approximate $x_{0,y}$ with

$$S_\lambda y = R_\lambda(T^*T)T^*y,$$

where R_λ is a family of continuous functions defined on $[0, \|T\|^2]$ (which contains the spectrum of T^*T) that approximate $f(t) = 1/t$. That is, $R_\lambda(t) \to 1/t$ as $\lambda \to 0^+$ for $t > 0$. Using the SVE, the expression $R_\lambda(T^*T)T^*$ can be written as

$$R_\lambda(T^*T)T^* = Vm\left[R_\lambda(\sigma^2)\sigma\right]U^\dagger.$$

Notice from equation (1.10), $T^\dagger = Vm_{\sigma^{-1}}U^\dagger$, which underscores the requirement for $R_\lambda(\sigma^2)$ to converge pointwise to $1/\sigma^2$. We now prove the following convergence result.

Theorem 3.1 *Suppose $\{R_\lambda\}_{\lambda>0}$ is a family of continuous real-valued functions defined on $[0, \|T\|^2]$ that satisfies*

$$R_\lambda(t) \to \frac{1}{t} \text{ as } \lambda \to 0^+ \text{ for each } t \in (0, \|T\|^2]. \tag{3.1}$$

In addition, suppose there exists $C > 0$ such that

$$|tR_\lambda(t)| \le C \text{ for each } t \in (0, \|T\|^2], \text{ and for each } \lambda \ge 0. \tag{3.2}$$

Then $R_\lambda(T^*T)T^*y \to T^\dagger y$ as $\lambda \to 0^+$ for each $y \in \mathcal{D}(T^\dagger)$.

Proof: Let $y \in \mathcal{D}(T^\dagger)$ be given and define $\overline{y} = \operatorname{proj}_{\overline{\mathcal{R}(T)}} y$. Note that $T^*y = T^*\overline{y}$ because $\mathcal{R}(T)^\perp = \mathcal{N}(T^*)$. We have $R_\lambda(T^*T)T^*y = Vm\left[R_\lambda(\sigma^2)\sigma\right]U^\dagger\overline{y}$ and $T^\dagger y = Vm\left[\sigma^{-1}\right]U^\dagger\overline{y}$. Therefore,

$$\begin{aligned}
\|R_\lambda(T^*T)T^*y - T^\dagger y\|_X^2 &= \left\|Vm\left[R_\lambda(\sigma^2)\sigma\right]U^\dagger\overline{y} - Vm\left[\sigma^{-1}\right]U^\dagger\overline{y}\right\|_X^2 \\
&= \left\|m\left[R_\lambda(\sigma^2)\sigma - \sigma^{-1}\right]U^\dagger\overline{y}\right\|_{L^2(M)}^2 \\
&= \int \left(R_\lambda(\sigma^2)\sigma^2 - 1\right)^2 \overline{f}^2 \, d\mu,
\end{aligned}$$

where we have defined $\overline{f} = m\left[\sigma^{-1}\right] U^{\dagger}\overline{y}$. We have $\overline{f} \in L^2(M)$ by the following reasoning. We know $y \in \mathcal{D}(T^{\dagger})$ which implies that $\overline{y} \in \mathcal{R}(T)$. Then by Lemma 1.31, we know $\overline{y} = Us$ for some $s \in S$ (where S is defined as in (1.7)) and hence, $U^{\dagger}\overline{y} = U^{\dagger}Us = s \in S$. Thus, $\overline{f} \in L^2(M)$.

Next, $R_\lambda(\sigma^2)\sigma^2$ is uniformly bounded by (3.2) and thus $\| \left(R_\lambda(\sigma^2)\sigma^2 - 1\right) \overline{f}\|^2_{L^2(M)} \leq (C+1)^2\|\overline{f}\|^2_{L^2(M)}$. Also, $\left(R_\lambda(\sigma^2)\sigma^2 - 1\right)^2 \overline{f}^2$ converges to zero pointwise. So then

$$\|R_\lambda(T^*T)T^*y - T^{\dagger}y\|^2_X = \int \left(R_\lambda(\sigma^2)\sigma^2 - 1\right)^2 \overline{f}^2 \, d\mu \to 0$$

by the dominated convergence theorem. ■

This theorem demonstrates that $R_\lambda(T^*T)T^*$ converges pointwise to $T^{\dagger}$ on $\mathcal{D}(T^{\dagger}) = \mathcal{R}(T) \oplus \mathcal{R}(T)^{\perp}$. When $\mathcal{R}(T)$ fails to be closed, $\mathcal{D}(T^{\dagger})$ is a proper dense subset of Y and $T^{\dagger}$ is unbounded. The next lemma and theorem demonstrate that if $y \notin \mathcal{D}(T^{\dagger})$, then $R_\lambda(T^*T)T^*y$ is not even weakly convergent.

Lemma 3.2 *Let $y \in \overline{\mathcal{R}(T)}$. Then $TR_\lambda(T^*T)T^*y \to y$ as $\lambda \to \infty$.*

Proof: We start by noting

$$\begin{aligned}
\| \left(TR_\lambda(T^*T)T^* - I\right) y\|_Y &= \|Um\left[R_\lambda(\sigma^2)\sigma^2 - 1\right] U^{\dagger}y\|_Y \\
&= \|m\left[R_\lambda(\sigma^2)\sigma^2 - 1\right] U^{\dagger}y\|_{L^2(M)} \\
&= \int \left(R_\lambda(\sigma^2)\sigma^2 - 1\right)^2 (U^{\dagger}y)^2 \, d\mu.
\end{aligned}$$

Then by equations (3.1) and (3.2), there exists a constant C such that

$$(R_\lambda(\sigma^2)\sigma^2 - 1) \leq (C+1).$$

By applying the dominated convergence theorem, the above norm goes to zero. ■

Theorem 3.3 *If $y \notin \mathcal{D}(T^{\dagger})$, then for any sequence $\lambda_n \to 0$, the sequence $\{R_{\lambda_n}(T^*T)T^*y\}$ is not weakly convergent.*

Proof: We will prove the contrapositive. We again use $\overline{y}$ to denote the orthogonal projection of y onto $\overline{\mathcal{R}(T)}$. Suppose there exists a sequence $\lambda_n \to 0$ such that

$$R_{\lambda_n}(T^*T)T^*y = R_{\lambda_n}(T^*T)T^*\overline{y} \to z \in X \text{ weakly.}$$

Then, because T is a bounded operator, we know

$$TR_{\lambda_n}(T^*T)T^*\overline{y} \to Tz \text{ weakly}$$

as well. However, $TR_{\lambda_n}(T^*T)T^*\overline{y} \to \overline{y}$ as $n \to \infty$ by Lemma 3.2. Thus, $TR_{\lambda_n}(T^*T)T^*\overline{y} \to Tz$ weakly and $TR_{\lambda_n}(T^*T)T^*\overline{y} \to \overline{y}$ strongly. This means that $\overline{y} = Tz$, which implies that $y \in \mathcal{D}(T^\dagger)$. ■

Corollary 3.4 *If $y \notin \mathcal{D}(T^\dagger)$, then $\lim\limits_{\lambda \to 0} \|R_\lambda(T^*T)T^*y\|_X = \infty$.*

Proof: If the limit does not go to infinity, then there exists $\{\lambda_n\} \subset (0, \infty)$ such that $\lambda_n \to 0$ and the sequence $\{R_{\lambda_n}(T^*T)T^*y\}$ is bounded. But that would imply that $\{R_{\lambda_n}(T^*T)T^*y\}$ has a weakly convergent subsequence. Then by the previous theorem, $y \in \mathcal{D}(T^\dagger)$. This proves the contrapositive. ■

3.3 Convergence rates

In this section, we prove convergence rates for the method discussed in the previous section given certain conditions. For simplicity, we will use the notation $x_{\lambda,y} = R_\lambda(T^*T)T^*y$ to denote the approximation of $T^\dagger y$ and $x_{0,y}$ to denote $T^\dagger y$ itself. In general, we can not say anything about the rate of convergence of $x_{\lambda,y} \to x_{0,y}$. However, if we make the assumption that $x_{0,y}$ is contained in $\mathcal{R}((T^*T)^\nu)$ for some $\nu > 0$, then we can derive a rate of convergence.

Before we prove this, we first note that the condition $x_{0,y} \in \mathcal{R}((T^*T)^\nu)$ is equivalent to $\overline{y} \in \mathcal{R}(T(T^*T)^\nu)$, where $\overline{y}$ denotes the projection of y onto $\overline{\mathcal{R}(T)}$. Groetsch (in [14]) proves the convergence rate we are about to prove assuming $\overline{y} \in \mathcal{R}(T(T^*T)^\nu)$. We will prove the same convergence rate assuming $x_{0,y} \in \mathcal{R}((T^*T)^\nu)$.

Lemma 3.5 *$x_{0,y} \in \mathcal{R}((T^*T)^\nu)$ if and only if $\overline{y} \in \mathcal{R}(T(T^*T)^\nu)$.*

Proof: If $x_{0,y} \in \mathcal{R}((T^*T)^\nu)$, then obviously $\overline{y} = Tx_{0,y} \in \mathcal{R}(T(T^*T)^\nu)$. To prove the converse, we first note that $\mathcal{R}((T^*T)^\nu) \subset \mathcal{N}(T)^\perp$ by the following reasoning:

$$\begin{aligned} x \in \mathcal{R}((T^*T)^\nu) &\implies x = (T^*T)^\nu w \text{ for some } w \in X \\ &\implies x = Vm\left[\sigma^{2\nu}\right]V^\dagger w \text{ for some } w \in X \\ &\implies x \in \mathcal{R}(V) = \mathcal{N}(T)^\perp. \end{aligned}$$

Now suppose that $\overline{y} \in \mathcal{R}(T(T^*T)^\nu)$. Then $Tx_{0,y} = \overline{y} = T(T^*T)^\nu w$ and hence $x_{0,y} - (T^*T)^\nu w \in \mathcal{N}(T)$. But we know $x_{0,y} \in \mathcal{N}(T)^\perp$ and $(T^*T)^\nu w \in \mathcal{N}(T)^\perp$ so we conclude that $x_{0,y} - (T^*T)^\nu w \in \mathcal{N}(T) \cap \mathcal{N}(T)^\perp = \{0\}$. Thus $x_{0,y} = (T^*T)^\nu w$, which completes the proof. ■

We will now prove a convergence rate assuming the condition

$$t^\nu |1 - tR_\lambda(t)| \leq \phi(\lambda, \nu) \text{ for all } t \in (0, \|T\|^2],\ \nu > 0 \tag{3.3}$$

where $\phi(\lambda, \nu)$ is called a *rate of convergence function* and has the property that $\phi(\lambda, \nu) \to 0$ as $\lambda \to 0$ for every $\nu > 0$.

Theorem 3.6 *If* (3.3) *holds and* $x_{0,y} = (T^*T)^\nu w$ *for* $\nu > 0$ *and some* $w \in X$*, then*

$$\|x_{\lambda,y} - x_{0,y}\|_X \leq \phi(\lambda, \nu)\|w\|_X.$$

Proof: Using the singular value expansion, we obtain

$$x_{\lambda,y} - x_{0,y} = Vm\left[R_\lambda(\sigma^2)\sigma\right]U^\dagger\overline{y} - Vm\left[\sigma^{-1}\right]U^\dagger\overline{y} = Vm\left[R_\lambda(\sigma^2)\sigma - \sigma^{-1}\right]U^\dagger\overline{y}.$$

Note that $\overline{y} = Tx_{0,y} = Um_\sigma V^\dagger x_{0,y}$ so the above equation becomes

$$Vm\left[R_\lambda(\sigma^2)\sigma - \sigma^{-1}\right]m_\sigma V^\dagger x_{0,y} = Vm\left[R_\lambda(\sigma^2)\sigma^2 - 1\right]V^\dagger x_{0,y}.$$

If $x_{0,y} = (T^*T)^\nu w = Vm[\sigma^{2\nu}]V^\dagger w$ for some $w \in \mathcal{N}(T)^\perp$, the above expression can be simplified further to

$$Vm\left[R_\lambda(\sigma^2)\sigma^2 - 1\right]m[\sigma^{2\nu}]V^\dagger w = Vm\left[\left(R_\lambda(\sigma^2)\sigma^2 - 1\right)\sigma^{2\nu}\right]V^\dagger w.$$

We conclude that

$$\begin{aligned}\|x_{\lambda,y} - x_{0,y}\|_X = \left\|m\left[\left(R_\lambda(\sigma^2)\sigma^2 - 1\right)\sigma^{2\nu}\right]V^\dagger w\right\|_{L^2(M)} &\leq \phi(\lambda, \nu)\|V^\dagger w\|_{L^2(M)} \\ &\leq \phi(\lambda, \nu)\|w\|_X. \blacksquare\end{aligned}$$

We conclude this section by mentioning an alternative way to conceptualize the condition $\overline{y} = T(T^*T)^\nu w$:

$$\begin{aligned}\overline{y} = T(T^*T)^\nu w \iff \overline{y} = Um\left[\sigma^{2\nu+1}\right]V^\dagger w \iff &m\left[\sigma^{-2\nu-1}\right]U^\dagger\overline{y} \in L^2(M) \\ \iff &\int \sigma^{-4\nu-2}(U^\dagger\overline{y})^2\, d\mu < \infty.\end{aligned}$$

That is, $\overline{y} \in \mathcal{R}(T(T^*T)^\nu)$ is equivalent to $\sigma^{-2\nu-1}U^\dagger\overline{y} \in L^2(M)$.

3.4 Inexact data

The results from the previous sections hold when y is known exactly. In a more realistic scenario, we do not know the true data, but rather some noisy approximation of the data. In this section, we will prove a convergence result when the data is not known exactly.

We will let y^* denote the exact data and y, some noisy measurement of y^*. We then assume that $\|y - y^*\|_Y < \delta$ for some $\delta > 0$. We compute $x_{\lambda,y} = R_\lambda(T^*T)T^*y$ (as opposed to $x_{\lambda,y^*} = R_\lambda(T^*T)T^*y^*)$ and hope for some choice of λ depending on δ, the solution $x_{\lambda,y}$ converges to $x_{0,y^*} = T^\dagger y^*$ as $\delta \to 0$. Thus, in addition to determining a suitable family of functions $\{R_\lambda\}_{\lambda>0}$, we also must choose a suitable regularization parameter λ. For now, we will assume λ is chosen based on the value of δ, i.e. $\lambda = \lambda(\delta)$. We say that the approximation $x_{\lambda,y}$ is *regular* if there is some choice of the regularization parameter λ in terms of δ such that $x_{\lambda,y} \to x_{0,y^*}$ as $\delta \to 0$.

In this section, we will make use of the equation

$$x_{\lambda,y} - x_{0,y^*} = (x_{\lambda,y} - x_{\lambda,y^*}) + (x_{\lambda,y^*} - x_{0,y^*}) \tag{3.4}$$

where the expression in the first set of parentheses is called the *perturbation error* and the expression in the second set of parentheses is called the *regularization error*. Note that the regularization error was analyzed in the previous section. In this section, our attention will be devoted to analyzing the perturbation error.

In order to prove our results regarding the perturbation error, we assume there exists some positive constant C that satisfies equation (3.2). We also define the function

$$r(\lambda) = \max\{|R_\lambda(t)| : t \in [0, \|T\|^2]\}. \tag{3.5}$$

These allow us to prove the following lemmas.

Lemma 3.7 $\|T(x_{\lambda,y} - x_{\lambda,y^*})\|_Y \leq \delta C$.

Proof: First note that

$$T(x_{\lambda,y} - x_{\lambda,y^*}) = TR_\lambda(T^*T)T^*(y - y^*) = Um\left[R_\lambda(\sigma^2)\sigma^2\right]U^\dagger(y - y^*)$$

which implies

$$\begin{aligned}\|T(x_{\lambda,y} - x_{\lambda,y^*})\|_Y &= \left\|Um\left[R_\lambda(\sigma^2)\sigma^2\right]U^\dagger(y - y^*)\right\|_Y \\ &= \left\|m\left[R_\lambda(\sigma^2)\sigma^2\right]U^\dagger(y - y^*)\right\|_{L^2(M)} \\ &\leq C\|U^\dagger(y - y^*)\|_{L^2(M)} \leq C\delta.\blacksquare\end{aligned}$$

Lemma 3.8 $\|x_{\lambda,y} - x_{\lambda,y^*}\| \leq \delta\sqrt{C}\sqrt{r(\lambda)}$.

Proof: We note that

$$\begin{aligned}x_{\lambda,y} - x_{\lambda,y^*} = Vm\left[R_\lambda(\sigma^2)\sigma\right]U^\dagger(y - y^*) &= (Vm_\sigma U^\dagger)Um\left[R_\lambda(\sigma^2)\right]U^\dagger(y - y^*) \\ &= T^*Um\left[R_\lambda(\sigma^2)\right]U^\dagger(y - y^*).\end{aligned}$$

So then

$$\begin{aligned}\|x_{\lambda,y} - x_{\lambda,y^*}\|_X^2 &= \left\langle Um\left[R_\lambda(\sigma^2)\right]U^\dagger(y-y^*), T(x_{\lambda,y} - x_{\lambda,y^*})\right\rangle_Y \\ &\le \|m\left[R_\lambda(\sigma^2)\right]U^\dagger(y-y^*)\|_{L^2(M)}\|T(x_{\lambda,y} - x_{\lambda,y^*})\|_Y \\ &\le r(\lambda)\|U^\dagger(y-y^*)\|_{L^2(M)}C\delta \le r(\lambda)C\delta^2,\end{aligned}$$

where Lemma 3.7 has been used in the last step. Taking a square root completes the proof. ■

With these lemmas, we can now prove sufficient conditions for $x_{\lambda,y}$ to converge to x_{0,y^*}. We have to assume that $\lambda : [0,\infty) \to [0,\infty)$ is a continuous, increasing, and nonnegative parameter choice depending on δ, with $\lambda(0) = 0$.

Theorem 3.9 *If $y^* \in \mathcal{D}(T^\dagger)$, $\lambda(\delta) \to 0$ and $\delta^2 r(\lambda(\delta)) \to 0$ as $\delta \to 0$, then $x_{\lambda,y} \to x_{0,y^*}$ as $\delta \to 0$.*

Proof: Note that

$$\|x_{\lambda,y} - x_{0,y^*}\|_X \le \|x_{\lambda,y} - x_{\lambda,y^*}\|_X + \|x_{\lambda,y^*} - x_{0,y^*}\|_X.$$

We know that $\|x_{\lambda,y^*} - x_{0,y^*}\|_X \to 0$ as $\delta \to 0$ by Theorem 3.1 and, applying Lemma 3.8,

$$\|x_{\lambda,y} - x_{\lambda,y^*}\|_X \le \delta\sqrt{C}\sqrt{r(\lambda(\delta))},$$

which goes to zero as $\delta \to 0$ provided $\delta^2 r(\lambda(\delta))$ goes to zero. ■

3.5 Tikhonov regularization

In this section, we will apply the results from previous sections to ordinary Tikhonov regularization. We start by choosing a regularization parameter $\lambda \in \mathbb{Z}^+$ and minimizing the functional

$$F_{\lambda,y}(x) = \|Tx - y\|_Y^2 + \lambda\|x\|_X^2$$

over $x \in X$. We first note that the first two derivatives of $F_{\lambda,y}$ with respect to x are given by

$$\begin{aligned}\nabla F_{\lambda,y}(x) &= 2\left(T^*Tx - T^*y + \lambda x\right), \\ \nabla^2 F_{\lambda,y}(x) &= 2(T^*T + \lambda I).\end{aligned}$$

For $\lambda > 0$, the operator $\nabla^2 F_{\lambda,y}(x) = 2(T^*T + \lambda I)$ is positive definite, which implies that $F_{\lambda,y}$ is strictly convex. Thus, $F_{\lambda,y}$ has a unique minimizer which can be found by solving the equation $\nabla F_{\lambda,y}(x) = 0$ for x. That is, the minimizer of $F_{\lambda,y}$ is given by $x_{\lambda,y} = (T^*T + \lambda I)^{-1}T^*y$. We can express $x_{\lambda,y}$ in the context of the previous section by defining

$$R_\lambda(t) = (t+\lambda)^{-1}.$$

Then the Tikhonov solution $x_{\lambda,y} = R_\lambda(T^*T)T^*y$ will have the form

$$x_{\lambda,y} = Vm\left[\frac{\sigma}{\sigma^2+\lambda}\right]U^\dagger y.$$

We can also derive the form of $\phi(\lambda,\nu)$ from equation (3.3) by finding an upper bound for $t^\nu|1-tR_\lambda(t)|$ for the given function $R_\lambda(t)$ where $t \in [0, \|T\|^2]$. We have

$$t^\nu\left(1-\frac{t}{\lambda+t}\right) = \frac{t^\nu\lambda}{\lambda+t} \tag{3.6}$$

and we treat this as a function of a single variable t. To find an upper bound, we first consider the case where $\nu > 1$. Here we can obtain the following simple upper bound

$$\frac{t^\nu\lambda}{\lambda+t} < \frac{t^\nu\lambda}{t} = \lambda t^{\nu-1} \leq C\lambda$$

where $C = \|T\|^{2\nu-2}$. On the other hand, if $\nu < 1$ we can find an upper bound by maximizing (3.6) using calculus. We find a critical number $t = \frac{\nu\lambda}{1-\nu}$ and (3.6) is equal to

$$\nu^\nu(1-\nu)^{1-\nu}\lambda^\nu$$

at this value of t. If ν is between 0 and 1, then $\nu^\nu(1-\nu)^{1-\nu} \leq 1$ so we let $\phi(\lambda,\nu) = \lambda^\nu$ for $\nu \in (0,1]$. In summary, $\phi(\lambda,\nu)$ may be defined by

$$\phi(\lambda,\nu) = \begin{cases} \lambda^\nu & 0 < \nu < 1 \\ C\lambda & \nu \geq 1. \end{cases}$$

The above derivation implies (when the data y is known) that first order convergence is the fastest possible convergence given by these results. In Section 3.6, we will see that first order convergence is the best possible rate for Tikhonov regularization.

In addition, the constant C from (3.2) is 1 and $r(\lambda)$ from (3.5) is λ^{-1}. With these values, we quickly obtain the following convergence results for Tikhonov regularization.

Lemma 3.10 *Let $y \in \mathcal{D}(T^\dagger)$. If $x_{0,y} = T^\dagger y \in \mathcal{R}(T^*T)^\nu)$ for some $\nu \in (0,1]$, then*

$$\|x_{\lambda,y} - x_{0,y}\|_X = O(\lambda^\nu).$$

Proof: This lemma is just a special case of Theorem 3.6. ■

The lemma below will allow us to give another convergence condition.

Lemma 3.11 $\mathcal{R}((T^*T)^{1/2}) = \mathcal{R}(T^*)$.

Proof: Let $x \in \mathcal{R}((T^*T)^{1/2})$. Then $x = Vm_\sigma V^\dagger w$ for some $w \in X$. Note that $U^\dagger U$ is the identity on $L^2(M)$ so $x = Vm_\sigma U^\dagger(UV^\dagger w) = T^*(UV^\dagger w) \in \mathcal{R}(T^*)$. Conversely, if $x \in \mathcal{R}(T^*)$, then $x = Vm_\sigma U^\dagger y$ for some $y \in Y$. Because $V^\dagger V = I$ on $L^2(M)$, we have $x = Vm_\sigma V^\dagger(VU^\dagger y) = (T^*T)^{1/2}(VU^\dagger y) \in \mathcal{R}((T^*T)^{1/2})$. ■

Corollary 3.12 *Let $y^* \in \mathcal{D}(T^\dagger)$. If $x_{0,y} \in \mathcal{R}(T^*)$, then $\|x_{\lambda,y} - x_{0,y}\|_X = O(\lambda^{1/2})$.*

Proof: The result follows trivially from the previous two lemmas. ■

For the case of inexact data, we again let y^* denote the true data and y a noisy approximation of y^* satisfying $\|y - y^*\|_Y \leq \delta$ for $\delta > 0$. We analyze the convergence of $x_{\lambda,y}$ to x_{0,y^*} using

$$\|x_{\lambda,y} - x_{0,y^*}\|_X \leq \|x_{\lambda,y} - x_{\lambda,y^*}\|_X + \|x_{\lambda,y^*} - x_{0,y^*}\|_X. \tag{3.7}$$

By Lemma 3.8, we know the perturbation error, $\|x_{\lambda,y} - x_{\lambda,y^*}\|_X$, is bounded by $\delta\lambda^{-1/2}$ and if we assume $x_{0,y^*} \in \mathcal{R}(T^*)$, Corollary 3.12, implies that the regularization error $\|x_{\lambda,y^*} - x_{0,y^*}\|_X$, is bounded by $C\lambda^{1/2}$ for some constant $C > 0$. Thus, a convergence rate for Tikhonov regularization can be obtained by choosing λ in such as way as to minimize the quantity $\delta\lambda^{-1/2} + C\lambda^{1/2}$. Using calculus, it can be shown that $\lambda \propto \delta$ will give you such a minimum.

Lemma 3.13 *If $x_{0,y^*} \in \mathcal{R}(T^*)$ and $\lambda = m\delta$ for some constant $m > 0$, then*

$$\|x_{\lambda,y} - x_{0.y^*}\|_X = O(\sqrt{\delta}).$$

Proof: The proof follows from (3.7), Lemma 3.8 and Corollary 3.12. ■

In a more general setting where we assume $x_{0,y^*} \in \mathcal{R}((T^*T)^\nu)$, Lemma 3.10 implies that the regularization error is $O(\lambda^\nu)$ and thus, we can obtain a rate of convergence by choosing λ to minimize $\delta\lambda^{-1/2} + C\lambda^\nu$ for some constant $C > 0$. Using calculus, choosing $\lambda \propto \delta^{2/(2\nu+1)}$ gives us a minimum.

Lemma 3.14 *If $x_{0,y^*} \in \mathcal{R}((T^*T)^\nu)$ for some $\nu \in (0,1]$ and $\lambda = m\delta^{2/(2\nu+1)}$, then*

$$\|x_{\lambda,y} - x_{0.y^*}\|_X = O(\delta^{2\nu/(2\nu+1)}).$$

Proof: This follows from equation (3.7), Lemma 3.8 and Lemma 3.10. ■

To conclude this section, we note that when the data is not known exactly, the fastest possible convergence rate of $x_{\lambda,y}$ to x_{0,y^*} given by these lemmas is $\delta^{2/3}$, which occurs when $x_{0,y^*} \in \mathcal{R}(T^*T)$ and $\lambda \sim \delta^{2/3}$. In the next section we will show this rate is optimal.

3.6 Converse results

In the last section we derived a rate of convergence for ordinary Tikhonov regularization. The error $\|x_{\lambda,y} - x_{0,y^*}\|_X$ was proportional to, at best, $\delta^{2/3}$. This rate turns out to be the best possible rate of convergence for Tikhonov regularization, as we will show in this section.

We will first show that the rate at which the regularization error goes to zero, which we proved in Theorem 3.6, is in fact optimal.

Theorem 3.15 *Suppose $y \in \mathcal{D}(T^\dagger)$ and $\|x_{\lambda,y} - x_{0,y}\|_X = o(\lambda)$. Then $x_{0,y} = 0$ and $x_{\lambda,y} = 0$ for all λ.*

Proof: Consider

$$(T^*T + \lambda I)(x_{\lambda,y} - x_{0,y}).$$

We note that $VV^\dagger$ is the projection onto $\mathcal{R}(V)$, $x_{\lambda,y} = Vm\left[\frac{\sigma}{\sigma^2+\lambda}\right]U^\dagger y \in \mathcal{R}(V)$, and $x_{0,y} = Vm[\sigma^{-1}]U^\dagger y \in \mathcal{R}(V)$. Next note that $(T^*T + \lambda I)x_{\lambda,y} = T^*y$ and also that $T^*Tx_{0,y} = T^*y$ because $x_{0,y}$ is the least squares solution of $Tx = y$. Thus, we obtain

$$(T^*T + \lambda I)(x_{\lambda,y} - x_{0,y}) = T^*y - T^*y - \lambda x_{0,y} = -\lambda x_{0,y}.$$

This implies that

$$\lambda\|x_{0,y}\| \leq \|T^*T + \lambda I\|\|x_{\lambda,y} - x_{0,y}\|_X \leq (\|T\|^2 + \lambda)\|x_{\lambda,y} - x_{0,y}\|_X = o(\lambda).$$

But the above equation is only possible if $x_{0,y} = 0$. Now let $\overline{y}$ denote the projection of y onto $\overline{\mathcal{R}(T)}$. Then of course $0 = Tx_{0,y} = \overline{y}$. This implies

$$x_{\lambda,y} = (T^*T + \lambda I)^{-1}T^*y = (T^*T + \lambda I)^{-1}T^*\overline{y} = 0$$

which completes the proof.■

Theorem 3.16 *If $y \in \mathcal{D}(T^\dagger)$ and $\|x_{\lambda,y} - x_{0,y}\|_X = O(\lambda)$, then $x_{0,y} \in \mathcal{R}(T^*T)$.*

Proof: Again, let $\overline{y}$ denote $\mathrm{proj}_{\overline{\mathcal{R}(T)}}y$. Note that $x_{0,y} = Vm_{\sigma^{-1}}U^\dagger\overline{y}$ and $x_{\lambda,y} = Vm\left[\frac{\sigma}{\sigma^2+\lambda}\right]U^\dagger\overline{y}$. Therefore,

$$\|x_{\lambda,y} - x_{0,y}\|_X^2 = \|Vm\left[\frac{\sigma}{\sigma^2+\lambda} - \frac{1}{\sigma}\right]U^\dagger\overline{y}\|_X^2 = \lambda^2\int \sigma^{-2}(\sigma^2+\lambda)^{-2}(U^\dagger\overline{y})^2\, d\mu$$

But because $\|x_{\lambda,y} - x_{0,y}\|_X = O(\lambda)$ by assumption, we know

$$\int \sigma^{-2}(\sigma^2+\lambda)^{-2}(U^\dagger\overline{y})^2\, d\mu = \int \sigma^{-6}(1+\lambda\sigma^{-2})^{-2}(U^\dagger\overline{y})^2\, d\mu$$

is bounded as $\lambda \to 0$. Note that the expression $(1+\lambda\sigma^{-2})^{-2}$ increases as λ decreases and converges pointwise to 1 as $\lambda \to 0$, so it follows by the monotone convergence theorem.

$$\int \sigma^{-6}(U^\dagger \overline{y})^2 \, d\mu < \infty.$$

Thus, $m_{\sigma^{-3}}U^\dagger\overline{y} \in L^2(M)$. For convenience of notation, let $f = m_{\sigma^{-3}}U^\dagger\overline{y}$. Note that $V^\dagger V$ is the identity on $L^2(M)$, so $m_{\sigma^{-3}}U^\dagger\overline{y} = V^\dagger V f$. This means $\overline{y} = Um_{\sigma^3}V^\dagger(Vf) \in \mathcal{R}(TT^*T)$. Thus, $x_{0,y} = T^\dagger\overline{y} = Vm_{\sigma^{-1}}U^\dagger\left(Um_{\sigma^3}V^\dagger(Vf)\right) = Vm_{\sigma^2}V^\dagger(Vf) = T^*T(Vf) \in \mathcal{R}(T^*T)$. This completes the proof. ■

We can also prove converse results for inexact data. Let y^* and y be defined as before and assume $\lambda : [0,\infty) \to [0,\infty)$ is a continuous, strictly increasing function with $\lambda(0) = 0$.

Lemma 3.17 *If $x_{0,y^*} \neq 0$, then $\lambda(\delta) = O(\|x_{0,y^*} - x_{\lambda(\delta),y}\|_X) + O(\delta)$.*

Proof: For convenience, we write λ for $\lambda(\delta)$ throughout this proof. Note that

$$\begin{aligned}(T^*T+\lambda I)(x_{0,y^*} - x_{\lambda,y}) &= T^*Tx_{0,y^*} + \lambda x_{0,y^*} - (T^*T+\lambda I)(T^*T+\lambda I)^{-1}T^*y \\ &= \lambda x_{0,y^*} + T^*(Tx_{0,y^*} - y) \\ &= \lambda x_{0,y^*} + T^*(\overline{y^*} - y),\end{aligned}$$

where $\overline{y^*}$ denotes the projection of y^* onto $\overline{\mathcal{R}(T)}$. Thus,

$$\|\lambda x_{0,y^*} + T^*(\overline{y^*} - y)\|_X = \|(T^*T+\lambda I)(x_{0,y^*} - x_{\lambda,y})\|_X.$$

By the reverse triangle inequality, we have

$$\lambda\|x_{0,y^*}\|_X - \|T^*(\overline{y^*} - y)\|_X \leq \|(T^*T+\lambda I)(x_{0,y^*} - x_{\lambda,y})\|_X.$$

From this, we obtain,

$$\begin{aligned}\lambda\|x_{0,y^*}\|_X &\leq \|(T^*T+\lambda I)(x_{0,y^*} - x_{\lambda,y})\|_X + \|T^*(\overline{y^*} - y)\|_X \\ &\leq (\|T\|^2+\lambda)\|x_{0,y^*} - x_{\lambda,y}\|_X + \|T^*\|\|\overline{y^*} - y\|_Y \\ &\leq (\|T\|^2+\lambda)\|x_{0,y^*} - x_{\lambda,y}\|_X + \|T^*\|\delta.\end{aligned}$$

The result follows after division by $\|x_{0,y^*}\|_X$. ■

We will use this lemma to prove the following theorem.

Theorem 3.18 *Suppose $\mathcal{R}(T)$ fails to be closed. If $\|x_{0,y^*} - x_{\lambda,y_n}\|_X = o(\delta_n^{2/3})$ for any sequence $\{y_n\} \subset Y$ such that $y_n \to y^*$ and $\|y^* - y_n\|_Y \leq \delta_n$ for some sequence $\{\delta_n\}$ with $\delta_n \to 0$, then $x_{0,y^*} = 0$.*

Proof: By Theorem 2.26, the range of T failing to be closed implies there exists a sequence $\{s_n\} \subset \mathcal{R}_{ess}(\sigma)$ such that $s_n \to 0$. Define $\{\delta_n\}_{n=1}^\infty$ by $\delta_n = s_n^3$. We note that $\frac{\sigma}{\sigma^2+\lambda}$ is a continuous function in σ for each fixed λ. So for all $n \in \mathbb{Z}^+$, there exists some $\epsilon_n > 0$ such that if $|s_n - \sigma| \le \epsilon_n$, then $\left|\frac{\sigma}{\sigma^2+\lambda} - \frac{s_n}{s_n^2+\lambda}\right| < \frac{1}{n}$. This condition implies (by the reverse triangle inequality) that for all $\sigma \in (s_n - \epsilon_n, s_n + \epsilon_n)$,

$$\left|\frac{\sigma}{\sigma^2+\lambda}\right| \ge \left|\frac{s_n}{s_n^2+\lambda}\right| - \frac{1}{n} \quad \text{and} \quad \left|\frac{\sigma}{\sigma^2+\lambda}\right| \le \left|\frac{s_n}{s_n^2+\lambda}\right| + \frac{1}{n}. \tag{3.8}$$

Now define $f_n = c_n \chi_{\sigma^{-1}(s_n-\epsilon_n, s_n+\epsilon_n)}$, where c_n is a normalization constant chosen so that $\|f_n\|_{L^2(M)} = 1$. Such a constant is possible because $s_n \in \mathcal{R}_{ess}(\sigma)$ and thus, $\chi_{\sigma^{-1}(s_n-\epsilon_n, s_n+\epsilon_n)} \neq 0$ in $L^2(M)$. Then define the sequence $\{y_n\} \subset Y$ by

$$y_n = y^* - \delta_n U f_n.$$

It follows that $\|y^* - y_n\|_Y \le \delta_n$. Now let $\lambda_n = \lambda(\delta_n)$. Then

$$\begin{aligned} x_{0,y^*} - x_{\lambda_n,y_n} &= x_{0,y^*} - x_{\lambda_n,y^*} + x_{\lambda_n,y^*} - x_{\lambda_n,y_n} \\ &= (x_{0,y^*} - x_{\lambda_n,y^*}) + (T^*T + \lambda_n I)^{-1} T^*(y^* - y_n) \\ &= (x_{0,y^*} - x_{\lambda_n,y^*}) + Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] U^\dagger(\delta_n U f_n) \\ &= (x_{0,y^*} - x_{\lambda_n,y^*}) + \delta_n Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] f_n. \end{aligned}$$

Therefore,

$$\begin{aligned} \|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 = \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X^2 &+ 2\delta_n \left\langle x_{0,y^*} - x_{\lambda_n,y^*}, Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] f_n \right\rangle_X \\ &+ \delta_n^2 \int \left(\frac{\sigma}{\sigma^2+\lambda_n}\right)^2 f_n^2 \, d\mu \\ &\ge -2\delta_n \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left\| m\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] f_n \right\|_{L^2(M)} \\ &+ \delta_n^2 \int \left(\frac{\sigma}{\sigma^2+\lambda_n}\right)^2 f_n^2 \, d\mu, \end{aligned}$$

where we have used the Cauchy-Schwartz inequality and the fact that V is an isometry. We then use the fact that the two $L^2(M)$ integrals above are supported on the set $\sigma^{-1}(s_n - \epsilon_n, s_n + \epsilon_n)$, which implies

$$\|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 \ge -2\delta_n \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left(\frac{s_n}{s_n^2+\lambda_n} + \frac{1}{n}\right) + \delta_n^2 \left(\frac{s_n}{s_n^2+\lambda_n} - \frac{1}{n}\right)^2,$$

where we've used the fact that $\|f_n\|_{L^2(M)} = 1$. By definition $s_n = \delta_n^{1/3}$; thus

$$\|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 \geq -2\delta_n \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left(\frac{\delta_n^{1/3}}{\delta_n^{2/3} + \lambda_n} + \frac{1}{n} \right) + \delta_n^2 \left(\frac{\delta_n^{1/3}}{\delta_n^{2/3} + \lambda_n} - \frac{1}{n} \right)^2 .$$

Multiplication by $\delta_n^{-4/3}$ gives us

$$\begin{aligned}
\delta_n^{-4/3} \|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 &\geq -2\|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left(\frac{1}{\delta_n^{2/3} + \lambda_n} + \frac{\delta_n^{-1/3}}{n} \right) \\
&\quad + \left(\frac{\delta_n^{2/3}}{\delta_n^{2/3} + \lambda_n} - \frac{\delta_n^{1/3}}{n} \right)^2 \\
&= -2\|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left(\frac{\delta_n^{-2/3}}{1 + \lambda_n \delta_n^{-2/3}} + \frac{\delta_n^{-1/3}}{n} \right) \\
&\quad + \left(\frac{1}{1 + \lambda_n \delta_n^{-2/3}} - \frac{\delta_n^{1/3}}{n} \right)^2 \\
&\geq -2\delta_n^{-2/3} \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \left(\frac{1}{1 + \lambda_n \delta_n^{-2/3}} + \frac{1}{n} \right) \\
&\quad + \left(\frac{1}{1 + \lambda_n \delta_n^{-2/3}} - \frac{\delta_n^{1/3}}{n} \right)^2 \\
&\geq -4\delta_n^{-2/3} \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \\
&\quad + \left(\frac{1}{1 + \lambda_n \delta_n^{-2/3}} - \frac{\delta_n^{1/3}}{n} \right)^2 ,
\end{aligned}$$

where we have used the fact that $\delta_n^{-1/3} \leq \delta_n^{-2/3}$ for small δ_n, also $\frac{1}{1+\lambda_n \delta_n^{-2/3}} < 1$ and $\frac{1}{n} < 1$. Now assume that $x_{0,y^*} \neq 0$ and consider the equation

$$\delta_n^{-4/3} \|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 \geq -4\delta_n^{-2/3} \|x_{0,y^*} - x_{\lambda_n,y^*}\|_X + \left(\frac{1}{1 + \lambda_n \delta_n^{-2/3}} - \frac{\delta_n^{1/3}}{n} \right)^2 . \quad (3.9)$$

We will compute the limit supremum of both sides of the inequality in (3.9) and obtain a contradiction. If we consider the left side of (3.9), we obtain,

$$\delta_n^{-4/3} \|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 \to 0$$

by assumption. We now consider both terms on the right side of the inequality in (3.9).

For the first term, $-4\delta_n^{-2/3}\|x_{0,y^*} - x_{\lambda_n,y^*}\|_X$, we note that $\|x_{0,y^*} - x_{\lambda_n,y_n}\|_X = o(\delta_n^{2/3})$ for any y_n such that $\|y^* - y_n\|_Y \leq \delta_n$. So in particular, it holds for $y_n = y^*$. Thus, $\|x_{0,y^*} - x_{\lambda_n,y^*}\|_X = o(\delta_n^{2/3})$ and hence $-4\delta_n^{-2/3}\|x_{0,y^*} - x_{\lambda_n,y^*}\|_X \to 0$ as $\delta_n \to 0$. For the second term, we know

$$\left(\frac{1}{1+\lambda_n\delta_n^{-2/3}} - \frac{\delta_n^{1/3}}{n}\right)^2 \to 1$$

because $\lambda_n\delta_n^{-2/3} \to 0$ as $\delta_n \to 0$ (by Lemma 3.17, which uses the fact that $x_{0,y^*} \neq 0$). Thus, the right side of (3.9) converges to 1. This contradiction shows that $x_{0,y^*} = 0$. ■

This theorem shows that $\delta^{2/3}$ is in fact the optimal rate of convergence of Tikhonov regularization with inexact data. Our next theorem is the converse of Lemma 3.14 and shows that the rate $\delta^{2/3}$ will only occur when $x_{0,y^*} \in \mathcal{R}(T^*T)$.

Theorem 3.19 *Suppose $\lambda(\delta) = c\delta^{2/3}$ for some constant $c \neq 0$. If $\|x_{0,y^*} - x_{\lambda(\delta_n),y_n}\| = O(\delta_n^{2/3})$ for any sequence $\{y_n\} \subset Y$ such that $y_n \to y^*$ and $\|y^* - y_n\|_Y \leq \delta_n$ for some sequence $\{\delta_n\}$ with $\delta_n \to 0$, then $x_{0,y^*} \in \mathcal{R}(T^*T)$.*

Proof: Let $\{\delta_n\}$ be a sequence with $\delta_n \to 0$ and let $\lambda_n = c\delta_n^{2/3}$. Define $y_n = \left(1 + \frac{\delta_n}{\|y^*\|_Y}\right) y^*$; then $\|y^* - y_n\|_Y = \delta_n$ for each $n \in \mathbb{Z}^+$. We have

$$\begin{aligned} x_{0,y^*} - x_{\lambda_n,y_n} &= Vm\left[\frac{1}{\sigma}\right]U^\dagger y^* - Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right]U^\dagger(1+\delta_n/\|y^*\|_Y)y^* \\ &= Vm\left[\frac{1}{\sigma} - \frac{(1+\delta_n/\|y^*\|_Y)\sigma}{\sigma^2+\lambda_n}\right]U^\dagger y^* \\ &= Vm\left[\frac{\lambda - \delta_n\sigma^2/\|y^*\|_Y}{\sigma(\sigma^2+\lambda_n)}\right]U^\dagger y^* \end{aligned}$$

Thus,

$$\begin{aligned} \|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 &= \int \left(\frac{\lambda - \delta_n\sigma^2/\|y^*\|_Y}{\sigma(\sigma^2+\lambda_n)}\right)^2 (U^\dagger y^*)^2 \, d\mu \\ &= \int \left(\frac{\lambda - \delta_n\sigma^2/\|y^*\|_Y}{1+\lambda_n\sigma^{-2}}\right)^2 \sigma^{-6}(U^\dagger y^*)^2 \, d\mu. \end{aligned}$$

By assumption, there exists a constant $M > 0$ such that

$$M \geq \delta_n^{-4/3}\|x_{0,y^*} - x_{\lambda_n,y_n}\|_X^2 = \delta_n^{-4/3}\int \left(\frac{\lambda_n - \delta_n\sigma^2/\|y^*\|_Y}{1+\lambda_n\sigma^{-2}}\right)^2 \sigma^{-6}(U^\dagger y^*)^2 \, d\mu.$$

We then note that $\lambda_n = c\delta_n^{2/3}$, which gives us

$$M \geq \int \left(\frac{c - \delta_n^{1/3}\sigma^2/\|y^*\|_Y}{1 + c\delta_n^{2/3}\sigma^{-2}} \right)^2 \sigma^{-6}(U^\dagger y^*)^2 \, d\mu.$$

Now note that the expression $\dfrac{c - \delta_n^{1/3}\sigma^2/\|y^*\|_Y}{1 + c\delta_n^{2/3}\sigma^{-2}}$ is such that the numerator increases as $\delta_n \to 0$ and the denominator decreases as $\delta_n \to 0$. Thus, it must increase monotonically as $\delta_n \to 0$. So we can take the limit as $\delta_n \to 0$ of both sides of the above expression and apply the monotone convergence theorem to obtain

$$M \geq c^2 \int \sigma^{-6}(U^\dagger y^*)^2 \, d\mu.$$

Thus, $m[\sigma^{-3}]U^\dagger y^* = f \in L^2(M)$. This implies that $y^* = Um[\sigma^3]V^\dagger(Vf)$, where we've used the fact that $V^\dagger V = I$ in $L^2(M)$. Therefore $y^* \in \mathcal{R}(TT^*T) = \mathcal{R}(Um_{\sigma^3}V^\dagger)$ and hence,

$$x_{0,y^*} = T^\dagger y^* = Vm_{\sigma^{-1}}U^\dagger(Um[\sigma^3]V^\dagger Vf) = Vm[\sigma^2]V^\dagger(Vf) = T^*T(Vf).$$

Thus, $x_{0,y^*} \in \mathcal{R}(T^*T)$. ■

3.7 The discrepancy principle

In this section, we will discuss the discrepancy principle, a method of choosing the regularization parameter λ based on the noise level δ. We again let y^* denote the true data and assume $y^* \in \mathcal{R}(T)$ throughout this section. We consider the problem $Tx = y$, where y is a noisy approximation of y^* satisfying

$$\|y^* - y\|_Y \leq \delta < \|y\|_Y. \tag{3.10}$$

The relation $\|y\|_Y > \delta$ is reasonable in that if $\delta \geq \|y\|_Y$, the noise level would be so high relative to the data y that we couldn't hope to find a reasonable approximation to $T^\dagger y^*$ from y. We must assume the noise level is low enough to make some sort of analysis possible.

The discrepancy principle chooses the parameter λ to satisfy

$$\|Tx_{\lambda,y} - y\|_Y = \delta. \tag{3.11}$$

The main idea behind the discrepancy principle is that if the data satisfies equation (3.10), then there is no reason to attempt to make the residual $\|Tx_{\lambda,y} - y\|_Y$ any smaller than δ. That is, the quality of the result can be no better than the quality of the input data.

To prove that it is possible to choose δ in this way, we first define the function $d_{\lambda,y} : Y \times [0, \infty) \to \mathbb{R}^+$ by

$$d_{\lambda,y} = \|Tx_{\lambda,y} - y\|_Y \tag{3.12}$$

and then prove the following theorem.

Theorem 3.20 *Suppose y^* and y satisfy* (3.10)*. Then the function $\lambda \to d_{\lambda,y}$ is continuous, increasing and contains δ in its range.*

Proof: We begin by noting that

$$\begin{aligned} Tx_{\lambda,y} - y = T(T^*T + \lambda I)^{-1}T^*y - y &= Um\left[\frac{\sigma^2}{\sigma^2+\lambda}\right]U^\dagger y - y \\ &= Um\left[\frac{\sigma^2}{\sigma^2+\lambda}\right]U^\dagger y - UU^\dagger y - \hat{y}, \end{aligned}$$

where $\hat{y}$ denotes the projection of y onto $\mathcal{R}(T)^\perp$ and $UU^\dagger y$ is the projection onto $\mathcal{R}(U) = \overline{\mathcal{R}(T)}$. Thus

$$Tx_{\lambda,y} - y = Um\left[\frac{\sigma^2}{\sigma^2+\lambda} - 1\right]U^\dagger y - \hat{y} = Um\left[\frac{-\lambda}{\sigma^2+\lambda}\right]U^\dagger y - \hat{y}$$

and hence,

$$\begin{aligned} d^2_{\lambda,y} = \|Tx_{\lambda,y} - y\|^2_Y &= \left\|Um\left[\frac{-\lambda}{\sigma^2+\lambda}\right]U^\dagger y\right\|^2_Y + \|\hat{y}\|^2_Y \\ &= \int\left(\frac{\lambda}{\sigma^2+\lambda}\right)^2 (U^\dagger y)^2\, d\mu + \|\hat{y}\|^2_Y, \end{aligned}$$

where the Pythagorean theorem holds because $\hat{y} \in \mathcal{R}(T)^\perp = \mathcal{R}(U)^\perp$. The above expression for $d_{\lambda,y}$ shows that $\lambda \to d_{\lambda,y}$ is continuous in λ. Also, the derivative of $\lambda/(\sigma^2+\lambda)$ with respect to λ is

$$\frac{\sigma^2}{(\sigma^2+\lambda)^2} > 0$$

which implies that $d_{\lambda,y}$ also increases with λ. We next note that by the monotone convergence theorem,

$$\lim_{\lambda\to\infty} d^2_{\lambda,y} = \|U^\dagger y\|^2_{L^2(M)} + \|\hat{y}\|^2_Y = \|UU^\dagger y\|^2_Y + \|\hat{y}\|^2_Y = \|y\|^2_Y > \delta^2$$

where the last equality comes from the Pythagorean theorem. Now note that $y^* \in \mathcal{R}(T)$ and we can decompose y as $y = \overline{y} + \hat{y}$ with $\overline{y} \in \overline{\mathcal{R}(T)}$ and $\hat{y} \in \mathcal{R}(T)^\perp$. Thus,

$\|y^* - y\|_Y^2 = \|y^* - \overline{y} - \hat{y}\|_Y^2 = \|y^* - \overline{y}\|_Y^2 + \|\hat{y}\|_Y^2$. Thus, $\|\hat{y}\|_Y^2 \leq \|y^* - y\|_Y^2$. We now use the dominated convergence theorem to compute the limit

$$\lim_{\lambda \to 0} d_{\lambda,y}^2 = \|\hat{y}\|_Y^2 \leq \|y^* - y\|_Y^2 \leq \delta^2.$$

Thus, $\lambda \to d_{\lambda,y}$ defines a continuous, increasing function such that

$$\lim_{\lambda \to 0} d_{\lambda,y}^2 \leq \delta^2 \text{ and } \lim_{\lambda \to \infty} d_{\lambda,y}^2 > \delta^2.$$

Thus, δ must exist in the range of $d_{\lambda,y}$. ■

We obtain further results by defining the residual

$$r_{\lambda,y} = y - Tx_{\lambda,y}. \tag{3.13}$$

Note that $d_{\lambda,y} = \|r_{\lambda,y}\|_Y$. We also have the relation

$$\begin{aligned} T^* r_{\lambda,y} = T^* y - T^* T x_{\lambda,y} &= Vm[\sigma]U^\dagger y - Vm\left[\frac{\sigma^3}{\sigma^2 + \lambda}\right] U^\dagger y \\ &= Vm\left[\frac{\sigma\lambda}{\sigma^2 + \lambda}\right] U^\dagger y = \lambda x_{\lambda,y}. \end{aligned} \tag{3.14}$$

When we analyze the approximations $x_{\lambda,y}$, we are interested in the squared error between x_{0,y^*} and $x_{\lambda,y}$. Using the above definitions, we find

$$\begin{aligned} \|x_{0,y^*} - x_{\lambda,y}\|_X^2 &= \|x_{0,y^*}\|_X^2 - \frac{2}{\lambda}\langle T^* r_{\lambda,y}, x_{0,y^*}\rangle_X + \|x_{\lambda,y}\|_X^2 \\ &= \|x_{0,y^*}\|_X^2 - \frac{2}{\lambda}\langle r_{\lambda,y}, y^*\rangle_Y + \|x_{\lambda,y}\|_X^2 \\ &= \|x_{0,y^*}\|_X^2 - \frac{2}{\lambda}\langle r_{\lambda,y}, y\rangle_Y + \frac{2}{\lambda}\langle r_{\lambda,y}, y^* - y\rangle_Y + \|x_{\lambda,y}\|_X^2 \\ &\leq E_{\lambda,y} \end{aligned}$$

where we've used the fact that $y^* \in \mathcal{R}(T)$ and also defined $E_{\lambda,y}$ by

$$E_{\lambda,y} = \|x_{\lambda,y}\|_X^2 - \frac{2}{\lambda}\langle r_{\lambda,y}, y\rangle_Y + \frac{2\delta}{\lambda} d_{\lambda,y} + \|x_{0,y^*}\|_X^2. \tag{3.15}$$

The next theorem will show that the above estimate of the error is minimized when $\delta = d_{\lambda,y}$.

Theorem 3.21 *If y and y^* satisfy* (3.10)*, then $E_{\lambda,y}$ is a minimum if and only if $d_{\lambda,y} = \delta$.*

Proof: Note that $d_{\lambda,y}$ must be positive for all positive λ. To show this, suppose

$r_{\lambda,y} = y - Tx_{\lambda,y} = 0$ for some $\lambda > 0$. Then

$$x_{\lambda,y} = (T^*T + \lambda I)^{-1}T^*y = (T^*T + \lambda I)^{-1}T^*Tx_{\lambda,y};$$

applying $(T^*T + \lambda I)$ to both sides gives us $\lambda x_{\lambda,y} = 0$. But then $x_{\lambda,y} = 0$ and hence, $y = Tx_{\lambda,y} = 0$ also. This means (3.10) will fail to hold, a contradiction. Thus, $d_{\lambda,y}$ must be positive for all $\lambda > 0$. We now take the derivative of $E_{\lambda,y}$ defined in (3.15) with respect to λ:

$$\begin{aligned}\frac{d}{d\lambda}\left(E_{\lambda,y}\right) = 2\langle x_{\lambda,y}, \dot{x}_{\lambda,y}\rangle_X + \frac{2}{\lambda^2}\langle r_{\lambda,y}, y\rangle_Y + \frac{2}{\lambda}\langle T\dot{x}_{\lambda,y}, y\rangle_Y \\ - \frac{2\delta}{\lambda^2}d_{\lambda,y} - \frac{2\delta}{\lambda d_{\lambda,y}}\langle T\dot{x}_{\lambda,y}, r_{\lambda,y}\rangle_Y.\end{aligned} \tag{3.16}$$

We note that $\dot{x}_{\lambda,y} = -(T^*T + \lambda I)^{-2}T^*y$ and $\dot{r}_{\lambda,y} = -T\dot{x}_{\lambda,y}$. We analyze each of the above terms in (3.16) in sequence. The first term can be written as

$$\begin{aligned}2\langle x_{\lambda,y}, \dot{x}_{\lambda,y}\rangle_X &= -2\langle x_{\lambda,y}, (T^*T + \lambda I)^{-1}x_{\lambda,y}\rangle_X \\ &= -\frac{2}{\lambda}\langle x_{\lambda,y}, (T^*T + \lambda I)^{-1}T^*r_{\lambda,y}\rangle_X \\ &= -\frac{2}{\lambda}\langle x_{\lambda,y}, T^*(TT^* + \lambda I)^{-1}r_{\lambda,y}\rangle_X \\ &= -\frac{2}{\lambda}\langle Tx_{\lambda,y}, (TT^* + \lambda I)^{-1}r_{\lambda,y}\rangle_Y,\end{aligned}$$

where the third equality follows from the SVE of T^* and $(T^*T+\lambda I)^{-1}$. For the second term,

$$\begin{aligned}\frac{2}{\lambda^2}\langle r_{\lambda,y}, y\rangle_Y &= \frac{2}{\lambda^2}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, (TT^* + \lambda I)y\rangle_Y \\ &= \frac{2}{\lambda^2}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, TT^*y\rangle_Y + \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, y\rangle_Y \\ &= \frac{2}{\lambda^2}\langle T^*(TT^* + \lambda I)^{-1}r_{\lambda,y}, T^*y\rangle_X + \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, y\rangle_Y \\ &= \frac{2}{\lambda^2}\langle (T^*T + \lambda I)^{-1}T^*r_{\lambda,y}, T^*y\rangle_X + \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, y\rangle_Y \\ &= \frac{2}{\lambda}\langle (T^*T + \lambda I)^{-1}x_{\lambda,y}, T^*y\rangle_X + \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, y\rangle_Y \\ &= \frac{2}{\lambda}\|x_{\lambda,y}\|_X^2 + \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1}r_{\lambda,y}, y\rangle_Y,\end{aligned}$$

where we have used the fact that $T^*r_{\lambda,y} = \lambda x_{\lambda,y}$. For the third equation:

$$\frac{2}{\lambda}\langle T\dot{x}_{\lambda,y}, y\rangle_Y = -\frac{2}{\lambda}\langle (T^*T + \lambda I)^{-2}T^*y, T^*y\rangle_X = -\frac{2}{\lambda}\|x_{\lambda,y}\|_X^2.$$

Thus, if we add the expressions for the first three terms of (3.16), we obtain

$$\frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, y\rangle_Y - \frac{2}{\lambda}\langle T x_{\lambda,y}, (TT^* + \lambda I)^{-1} r_{\lambda,y}\rangle_Y = \frac{2}{\lambda}\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, r_{\lambda,y}\rangle_Y.$$

The fourth term in (3.16) can be written as

$$\begin{aligned}
-\frac{2\delta}{\lambda^2} d_{\lambda,y} &= -\frac{2\delta}{\lambda^2 d_{\lambda,y}}\langle r_{\lambda,y}, r_{\lambda,y}\rangle_Y \\
&= -\frac{2\delta}{\lambda^2 d_{\lambda,y}}\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, (TT^* + \lambda I) r_{\lambda,y}\rangle_Y \\
&= -\frac{2\delta}{\lambda^2 d_{\lambda,y}}\left(\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, TT^* r_{\lambda,y}\rangle_Y + \lambda\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, r_{\lambda,y}\rangle_Y\right) \\
&= -\frac{2\delta}{\lambda^2 d_{\lambda,y}}\left(\langle T^*(TT^* + \lambda I)^{-1} r_{\lambda,y}, T^* r_{\lambda,y}\rangle_X + \lambda\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, r_{\lambda,y}\rangle_Y\right) \\
&= -\frac{2\delta}{\lambda d_{\lambda,y}}\left(\langle T^*(TT^* + \lambda I)^{-1} r_{\lambda,y}, x_{\lambda,y}\rangle_X + \langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, r_{\lambda,y}\rangle_Y\right) \\
&= -\frac{2\delta}{\lambda d_{\lambda,y}}\left(\langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, T x_{\lambda,y}\rangle_X + \langle (TT^* + \lambda I)^{-1} r_{\lambda,y}, r_{\lambda,y}\rangle_Y\right).
\end{aligned}$$

Lastly, the fifth term in (3.16) is

$$\begin{aligned}
-\frac{2\delta}{\lambda d_{\lambda,y}}\langle T\dot{x}_{\lambda,y}, r_{\lambda,y}\rangle_Y &= -\frac{2\delta}{\lambda d_{\lambda,y}}\langle \dot{x}_{\lambda,y}, T^* r_{\lambda,y}\rangle_X \\
&= -\frac{2\delta}{d_{\lambda,y}}\langle \dot{x}_{\lambda,y}, x_{\lambda,y}\rangle_X \\
&= \frac{2\delta}{\lambda d_{\lambda,y}}\langle T x_{\lambda,y}, (TT^* + \lambda I)^{-1} r_{\lambda,y}\rangle_Y,
\end{aligned}$$

where the last step was done by copying our work on the first expression of (3.16). Thus, when we add the expressions for all five terms to gether, we obtain

$$\frac{d}{d\lambda}\left(E_{\lambda,y}\right) = \frac{2}{\lambda}\left(1 - \frac{\delta}{d_{\lambda,y}}\right)\left\|(TT^* + \lambda I)^{-1/2} r_{\lambda,y}\right\|_Y^2. \tag{3.17}$$

We note that the norm in the above expression must be positive for all $\lambda > 0$. If the norm was equal to zero, then $r_{\lambda,y} = 0$ for some $\lambda > 0$ and we would obtain a contradiction as we did earlier in the proof. Now note that in the proof of Theorem 3.20, we proved that $d_{\lambda,y}$ increases as λ increases. Thus, (3.17) tells us $\frac{d}{d\lambda}\left(E_{\lambda,y}\right) < 0$ for $d_{\lambda,y} < \delta$ and $\frac{d}{d\lambda}\left(E_{\lambda,y}\right) > 0$ for $d_{\lambda,y} > \delta$. Thus, $E_{\lambda,y}$ is minimized if and only if $\delta = d_{\lambda,y}$. ■

We may now prove the convergence of $x_{\lambda,y}$ to x_{0,y^*}, given the choice of λ satisfying the discrepancy principle.

Theorem 3.22 *If y and y^* satisfy* (3.10) *and $\lambda = \lambda(\delta)$ satisfies* (3.11)*, then $x_{\lambda,y} \to x_{0,y^*}$ as $\delta \to 0$.*

Proof: By definition, $x_{\lambda,y}$ minimizes the functional $F_{\lambda,y}(x) = \|Tx - y\|_Y^2 + \lambda\|x\|_X^2$, so then

$$\|r_{\lambda,y}\|_Y^2 + \lambda\|x_{\lambda,y}\|_X^2 = F_{\lambda,y}(x_{\lambda,y}) \leq F_{\lambda,y}(x_{0,y^*}) = \|y^* - y\|_Y^2 + \lambda\|x_{0,y^*}\|_X^2 \leq \delta^2 + \lambda\|x_{0,y^*}\|_X^2.$$

But $\|r_{\lambda,y}\|_Y = \delta$, so we have

$$\|x_{\lambda,y}\|_X \leq \|x_{0,y^*}\|_X$$

for all δ. Thus, for each sequence $\{\delta_n\}$ converging to 0, there is a subsequence, which we'll also denote $\{\delta_n\}$ and a vector $w \in X$, such that

$$x_{\lambda_n,y} \to w \text{ weakly,}$$

where λ_n is defined by $\lambda_n = \lambda(\delta_n)$. We will prove that $w = x_{0,y^*}$ and $x_{\lambda_n,y} \to w$ for all subsequences $\{\delta_n\}$.

Because T is a bounded operator,

$$Tx_{\lambda_n,y} \to Tw \text{ weakly.}$$

But $\|Tx_{\lambda_n,y} - y\|_Y = \delta_n \to 0$ and by (3.10), we know that $y \to y^*$. Thus,

$$\|Tx_{\lambda_n,y} - y^*\|_Y \leq \|Tx_{\lambda_n,y} - y\|_Y + \|y - y^*\|_Y \to 0,$$

and hence $Tx_{\lambda_n,y} \to y^*$. Therefore, $Tx_{\lambda_n,y} \to Tw$ weakly and $Tx_{\lambda_n,y} \to y^*$ strongly, which together imply $Tw = y^*$. Because $x_{\lambda,y} = (T^*T + \lambda I)^{-1}T^*y = T^*(TT^* + \lambda I)^{-1}y \in \mathcal{R}(T^*)$, it follows that $w \in \overline{\mathcal{R}(T^*)} = \mathcal{N}(T)^\perp$. Thus, $Tw = y^*$ and w is contained in $\mathcal{N}(T)^\perp$. It then follows that $w = x_{0,y^*}$ and thus

$$x_{\lambda_n,y} \to x_{0,y^*} \text{ weakly as } n \to \infty.$$

Then, because the norm function is weakly lower semi-continuous, as well as the fact that $\|x_{\lambda_n,y}\|_X \leq \|x_{0,y^*}\|_X$, we have

$$\|x_{0,y^*}\|_X \leq \liminf_{n\to\infty} \|x_{\lambda_n,y}\|_X \leq \limsup_{n\to\infty} \|x_{\lambda_n,y}\|_X \leq \|x_{0,y^*}\|_X.$$

This implies that $\|x_{\lambda,y}\|_X \to \|x_{0,y^*}\|_X$. We conclude the proof by noting that weak convergence along with convergence in norm implies strong convergence. Thus, $x_{\lambda_n,y} \to x_{0,y^*}$ as $\delta_n \to 0$. ■

We will now focus on proving convergence rates of $x_{\lambda,y} \to x_{0,y^*}$ for λ chosen according to the discrepancy principle.

Theorem 3.23 *If $\lambda = \lambda(\delta)$ satisfies $d_{\lambda,y} = \delta$, then $\lambda \leq \delta\frac{\|T\|^2}{\|y\|_Y - \delta}$.*

Proof: From (3.11), we have

$$\|y\|_Y - \delta = \|y\|_Y - \|Tx_{\lambda,y} - y\|_Y \leq \|y\|_Y + \|Tx_{\lambda,y}\| - \|y\|_Y = \|Tx_{\lambda,y}\|_Y$$

where we have used the reverse triangle inequality. Then (3.14) implies that

$$\|Tx_{\lambda,y}\|_Y \leq \|T\|\|x_{\lambda,y}\| = \|T\|\frac{\|T^* r_{\lambda,y}\|}{\lambda} \leq \|T\|^2 \frac{\|r_{\lambda,y}\|}{\lambda} = \|T\|^2 \frac{\delta}{\lambda}$$

and thus, we have

$$\|y\|_Y - \delta \leq \|T\|^2 \frac{\delta}{\lambda}$$

and the result follows directly. ■

We are now able to prove a convergence rate, assuming $x_{0,y^*} \in \mathcal{R}(T^*)$.

Theorem 3.24 *If $x_{0,y^*} \in \mathcal{R}(T^*)$ and λ is chosen by the discrepancy principle, then $\|x_{0,y^*} - x_{\lambda,y}\|_X = O(\sqrt{\delta})$.*

Proof: Let $x_{0,y^*} = T^* w$ for some $w \in Y$. Then, using our assumption that y^* is in $\mathcal{R}(T)$, we have $Tx_{0,y^*} = y^*$. Therefore,

$$\begin{aligned}
\|x_{\lambda,y} - x_{0,y^*}\|_X^2 &= \|x_{\lambda,y}\|_X^2 - 2\langle x_{\lambda,y}, x_{0,y^*}\rangle_X + \|x_{0,y^*}\|_X \\
&\leq 2\left(\|x_{0,y^*}\|_X^2 - \langle x_{\lambda,y}, x_{0,y^*}\rangle_X\right) \\
&= 2\langle x_{0,y^*} - x_{\lambda,y}, T^* w\rangle_X \\
&= 2\langle Tx_{0,y^*} - Tx_{\lambda,y}, w\rangle_Y \\
&= 2\langle y^* - y, w\rangle_Y + 2\langle y - Tx_{\lambda,y}, w\rangle_Y \\
&\leq 2\|y^* - y\|_Y\|w\|_Y + 2\|y - Tx_{\lambda,y}\|_Y\|w\|_Y \\
&\leq 4\delta\|w\|_Y,
\end{aligned}$$

where we used the fact that $\|y - Tx_{\lambda,y}\|_Y = \delta$ in the last step above. ■

Next, we prove that this rate of convergence cannot be improved upon except in the trivial case.

Theorem 3.25 *If λ is chosen by the discrepancy principle and $\|x_{\lambda,y} - x_{0,y^*}\|_X = o(\sqrt{\delta})$ for all y and y^* satisfying* (3.10)*, then $\mathcal{R}(T)$ is closed.*

Proof: Suppose $\mathcal{R}(T)$ fails to be closed but $\|x_{\lambda,y} - x_{0,y^*}\|_X = o(\sqrt{\delta})$. Then by Theorem 2.26, there exists a strictly decreasing sequence $\{s_n\} \subset \mathcal{R}_{ess}(\sigma)$ such that $s_n \to 0$. Let $\delta_n = s_n^2$ and λ_n denote the parameter chosen by the discrepancy principle. We then choose a sequence $\{\epsilon_n\}$ that satisfies the following properties. Each ϵ_n for $n > 1$, must be small enough so that $(s_1 - \epsilon_1, s_1 + \epsilon_1) \cap (s_n - \epsilon_n, s_n + \epsilon_n) = \emptyset$. Note also that the function $\sigma \to \frac{\sigma}{\sigma^2 + \lambda_n}$ is continuous in σ for each fixed λ_n. So we also

choose ϵ_n for each $n \in \mathbb{Z}^+$ so that if $|s_n - \sigma| < \epsilon_n$, then $\left|\frac{\sigma}{\sigma^2+\lambda_n} - \frac{s_n}{s_n^2+\lambda_n}\right| < \frac{1}{n}$ or, by the triangle inequality, $\frac{\sigma}{\sigma^2+\lambda_n} > \frac{s_n}{s_n^2+\lambda_n} - \frac{1}{n}$.
We now define

$$f_n = c_n \chi_{\sigma^{-1}(s_n - \epsilon_n, s_n + \epsilon_n)}$$

where c_n is a normalization constant such that $\|f_n\|_{L^2(M)} = 1$. Our requirement that $(s_1 - \epsilon_1, s_1 + \epsilon_1) \cap (s_n - \epsilon_n, s_n + \epsilon_n) = \emptyset$ implies that f_1 will be orthogonal to each f_n. We also define $y^* = Uf_1$ and $y_n = y^* + \delta_n U f_n$. Then $\|y^* - y_n\|_Y = \delta_n \leq \|y_n\|_Y$. It follows that

$$\begin{aligned}
\|x_{\lambda_n, y_n} - x_{0,y^*}\|_X^2 &= \left\| Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] f_1 - Vm[\sigma^{-1}]f_1 + \delta_n Vm\left[\frac{\sigma}{\sigma^2+\lambda_n}\right] f_n \right\|_X^2 \\
&= \int \left(\frac{\lambda_n}{\sigma(\sigma^2+\lambda_n)}\right)^2 f_1^2 \, d\mu + \delta_n^2 \int \left(\frac{\sigma}{\sigma^2+\lambda_n}\right)^2 f_n^2 \, d\mu \\
&\geq \delta_n^2 \left(\frac{s_n}{s_n^2+\lambda_n} - \frac{1}{n}\right)^2 \int f_n^2 \, d\mu \\
&\geq \left(\frac{\sqrt{\delta_n}}{1+\lambda_n\delta_n^{-1}} - \frac{\delta_n}{n}\right)^2,
\end{aligned}$$

where the second equality follows from the orthogonality of f_1 and f_n and the first inequality follows from the fact that the above integrals are supported only on the set $\sigma^{-1}(s_n - \epsilon_n, s_n + \epsilon_n)$. The last inequality follows from the equation $\delta_n = s_n^2$. By hypothesis, we know that

$$\frac{\sqrt{\delta_n}}{1+\lambda_n\delta_n^{-1}} - \frac{\delta_n}{n} = o(\sqrt{\delta_n}).$$

But this will only hold if $\lambda_n\delta_n^{-1} \to \infty$ as $n \to \infty$. But by Theorem 3.23, we know that

$$\lambda_n\delta_n^{-1} \leq \frac{\|T\|^2}{\|y_n\|_Y - \delta_n} \leq \frac{\|T\|^2}{\sqrt{1+\delta_n^2} - \delta_n} \to \|T\|^2 \text{ as } n \to \infty$$

where we used the fact that

$$\|y_n\|_Y^2 = \|f_1 + \delta_n f_n\|_{L^2(M)}^2 = 1 + \delta_n^2 + 2\delta_n\langle f_1, f_n\rangle_{L^2(M)} \geq 1 + \delta_n^2.$$

This contradiction completes the proof. ■

Chapter 4

Approximating the SVE of a compact operator

This chapter is an expanded version of our paper [6] that has been accepted for publication in the SIAM Journal on Numerical Analysis (SINUM). Several proofs that were cut out of the paper for economy of space are included here.

4.1 Introduction

The *singular value expansion* (SVE) of a compact linear operator $T : X \to Y$, where X and Y are separable Hilbert spaces, enables a straightforward analysis of several related problems: computing the generalized inverse $T^\dagger$ of T, understanding the inverse problem $Tx = y$ (given $y \in Y$, estimate $x \in X$), regularizing the inverse problem using Tikhonov regularization or another scheme, and so forth. Although the SVE is used mostly for analysis, it is employed in certain computational schemes, most notably truncated singular value expansion (TSVE) regularization. In this chapter, we analyze general schemes for approximating the singular values and vectors of a compact operator.

The SVE and its analysis have a long history; in the context of integral equations, this history dates from 1907 [25]. (For even earlier work, the reader can consult Stewart's brief history of the SVD [27].) In addition to deriving the basic properties of the SVE, much of the work focused on clarifying the sense in which the kernel of the integral operator could be represented in terms of the singular values and vectors (for example, [21], [26]) and characterizing the singular values, including the rate at which they converge to zero (for example, [29], [17], [26]). As we discuss below, there is little work in the literature about computing numerical estimates of the singular values and vectors of a compact operator.

Throughout this chapter, X and Y denote separable Hilbert spaces and $T : X \to Y$

denotes a compact linear operator with singular value expansion

$$T = \sum_{k=1}^{\infty} \sigma_k \psi_k \otimes \phi_k.$$

Thus, $\{\phi_k\}$ and $\{\psi_k\}$ are orthonormal sequences in X and Y, respectively, and $\{\sigma_k\}$ is a sequence of positive numbers decreasing monotonically to zero. (If T has finite rank, then the SVE contains only finitely many terms. For simplicity of exposition, we will assume the typical case that T has infinite rank.)

Let $\{T_h : X \to Y \,|\, h > 0\}$ be a family of compact linear operators with the property that $T_h \to T$ in the operator norm as $h \to 0$; more specifically, we assume that

$$\|T_h - T\|_{\mathcal{L}(X,Y)} \le \epsilon_h \to 0 \text{ as } h \to 0.$$

Each operator T_h has an SVE

$$T_h = \sum_k \sigma_{h,k} \psi_{h,k} \otimes \phi_{h,k},$$

where the sum contains finitely many terms if T_h has finite rank and infinitely many terms otherwise. This notation for T_h, its SVE, and ϵ_h will be used throughout the chapter. We wish to analyze how the singular values and singular vectors of T_h converge to those of T.

We will frequently use the fact that the singular values and singular vectors of T are related to the nonzero eigenvalues and corresponding eigenvectors of T^*T and TT^*. Specifically, for each $k \in \mathbb{Z}^+$, we have

$$\begin{aligned} T\phi_k &= \sigma_k \psi_k, \\ T^*T\phi_k &= \sigma_k^2 \phi_k, \\ TT^*\psi_k &= \sigma_k^2 \psi_k. \end{aligned}$$

It follows that the subspace of right singular vectors associated with the singular value σ_k can be defined as

$$E_k = \{\phi \in X \,:\, T^*T\phi = \sigma_k^2 \phi\}. \tag{4.1}$$

Similarly,

$$F_k = \{\psi \in Y \,:\, TT^*\psi = \sigma_k^2 \psi\} \tag{4.2}$$

is the subspace of left singular vectors associated with the singular value σ_k. We assume that the singular values are enumerated according to multiplicity. That is, if $\dim(E_k) = d > 1$, then the value σ_k appears d times in the list $\sigma_1, \sigma_2, \sigma_3, \ldots$.

By definition, the singular values of T and T_h satisfy

$$\sigma_1 \ge \sigma_2 \ge \sigma_3 \ge \cdots$$

and

$$\sigma_{h,1} \geq \sigma_{h,2} \geq \sigma_{h,3} \geq \cdots .$$

There is therefore no ambiguity in asserting that the singular values of T_h converge to those of T; it simply means that, for each $k \in \mathbb{Z}^+$, $\sigma_{h,k} \to \sigma_k$ as $h \to 0$, where it is understood that, for a given k, $\sigma_{h,k}$ is defined for all h sufficiently small.

The convergence of the singular vectors is a more subtle question. If the singular space E_k has dimension $d > 1$, then convergence of the singular values implies that there will be d singular values of T_h converging to σ_k:

$$\sigma_{h,k_i} \to \sigma_k \text{ as } h \to 0 \text{ for } i = 1, 2, \ldots, d$$

(where $\sigma_{k_i} = \sigma_k$ for all $i = 1, 2, \ldots, d$). However, there is no reason to expect that σ_{h,k_i}, $i = 1, 2, \ldots, d$, all have the same value; more typically, the singular space associated with each σ_{h,k_i} is one-dimensional. Since the right singular vectors associated with σ_k need only form an orthonormal basis for E_k, it need not be the case that ϕ_{h,k_i} converge to any particular ϕ_{k_j}. For this reason, we define

$$E_{h,k} = \mathrm{sp}\{\phi \in X \, : \, \exists\, \ell \in \mathbb{Z}^+, \, \sigma_{h,\ell} \to \sigma_k \text{ as } h \to 0 \text{ and } T_h^* T_h \phi = \sigma_{h,\ell}^2 \phi\} \tag{4.3}$$

and

$$F_{h,k} = \mathrm{sp}\{\psi \in Y \, : \, \exists\, \ell \in \mathbb{Z}^+, \, \sigma_{h,\ell} \to \sigma_k \text{ as } h \to 0 \text{ and } T_h T_h^* \psi = \sigma_{h,\ell}^2 \psi\}. \tag{4.4}$$

We then require that $E_{h,k}$ converge to E_k in the sense that the *gap* between $E_{h,k}$ and E_k converges to zero, and similarly for $F_{h,k}$ and F_k.

Definition 4.1 *Given subspaces U and V of a Hilbert space H, we define*

$$\delta(U, V) = \sup_{\substack{u \in U \\ \|u\|_H = 1}} \inf_{v \in V} \|v - u\|_H .$$

We then define the gap *between U and V by*

$$\hat{\delta}(U, V) = \max\{\delta(U, V), \delta(V, U)\}.$$

In the case that V is closed (the only case discussed in this chapter), we could equivalently define $\delta(U, V)$ as

$$\delta(U, V) = \sup_{\substack{u \in U \\ \|u\|_H = 1}} \|P_V u - u\|_H ,$$

where P_V denotes the (orthogonal) projection onto V. It is clear that $0 \leq \delta(U, V) \leq 1$. In general, $\delta(U, V)$ and $\delta(V, U)$ may differ, but it is known (see, for example, [**?** , §2]) that the two are equal when both are strictly less than 1. Moreover, this also holds if U and V are finite-dimensional subspaces with the same dimension.

Lemma 4.2 *Let U and V be k-dimensional subspaces of a Hilbert space H, where k is a positive integer. Then $\delta(U,V) = \delta(V,U)$.*

Proof: Without loss of generality, let us assume that $\delta(U,V) \le \delta(V,U)$. As noted above, $\delta(V,U) < 1$ implies that $\delta(U,V) = \delta(V,U)$. It suffices, therefore, to show that the assumption $\delta(U,V) < \delta(V,U) = 1$ produces a contradiction.

Since V is finite-dimensional,

$$\delta(V,U) = \max_{\substack{v \in V \\ \|v\|_H = 1}} \|P_U v - v\|_H.$$

Therefore, the assumption that $\delta(V,U) = 1$ implies that there exists $\hat{v} \in V$ such that $\|\hat{v}\|_H = 1$ and $\|P_U \hat{v} - \hat{v}\|_H = 1$. This is possible only if $P_U \hat{v} = 0$, that is, if $\hat{v} \in U^\perp$. On the other hand, the assumption that $\delta(U,V) < 1$ implies that $\|P_V u - u\|_H < 1$ for all $u \in U$ and hence that the null space of $P = P_V|_U$ (P_V restricted to U) is trivial. Since $\dim(U) = \dim(V) = k$, the fundamental theorem of linear algebra implies that P maps U onto V; thus, there exists $\hat{u} \in U$ such that $P\hat{u} = \hat{v}$. But then

$$\begin{aligned} \left\| P\left(\frac{\hat{u}}{\|\hat{u}\|_H}\right) - \frac{\hat{u}}{\|\hat{u}\|_H} \right\|_H < 1 &\Rightarrow \|P\hat{u} - \hat{u}\|_H < \|\hat{u}\|_H \\ &\Rightarrow \|\hat{v} - \hat{u}\|_H < \|\hat{u}\|_H \\ &\Rightarrow \|\hat{v}\|_H^2 + \|\hat{u}\|_H^2 < \|\hat{u}\|_H^2 \end{aligned}$$

(where we used $\delta(U,V) < 1$ in the first step and $\hat{v} \in U^\perp$ in the last step). Since $\|\hat{v}\|_H = 1$, the last inequality is impossible, and the proof is complete. ■

Convergence of the singular values and vectors of T_h to those of T follows from the theory of Babuška and Osborn [3]; the following result is Theorem 9.1 of [4], specialized to the self-adjoint case.

Theorem 4.3 *Let X be a Hilbert space, let $A : X \to X$ and $A_h : X \to X$, $h > 0$, be compact self-adjoint linear operators, and assume that $A_h \to A$ in the operator norm as $h \to 0$. Then, for any compact subset K of $\rho(A)$ (the resolvent of A), there exists $h_0 > 0$ such that for all $h \in (0, h_0)$, $K \subset \rho(A_h)$. If λ is a nonzero eigenvalue of A with multiplicity equal to m, then there are m eigenvalues $\lambda_{h,1}, \lambda_{h,2}, \ldots, \lambda_{h,m}$ of A_h, repeated according to their multiplicities, such that each $\lambda_{h,i} \to \lambda$ as $h \to 0$. Moreover, the gap between the direct sum of the eigenspaces of A_h corresponding to $\lambda_{h,1}, \lambda_{h,2}, \ldots, \lambda_{h,m}$ and the eigenspace of A corresponding to λ tends to zero as $h \to 0$.*

We can use Theorem 4.3 to demonstrate the convergence of the singular values and singular vectors of T_h to those of T by applying it to T^*T and TT^*.

Lemma 4.4 *There exists a constant $C > 0$ such that*

$$\|T_h^*T_h - T^*T\|_{\mathcal{L}(X,X)} \leq C\epsilon_h \text{ and } \|T_hT_h^* - TT^*\|_{\mathcal{L}(Y,Y)} \leq C\epsilon_h \; \forall \, h > 0.$$

Proof: We have

$$\begin{aligned}\|T_h^*T_h - T^*T\|_{\mathcal{L}(X,X)} &\leq \|T_h^*T_h - T^*T_h\|_{\mathcal{L}(X,X)} + \|T^*T_h - T^*T\|_{\mathcal{L}(X,X)} \\ &\leq (\|T_h\|_{\mathcal{L}(X,Y)} + \|T\|_{\mathcal{L}(X,Y)})\|T_h - T\|_{\mathcal{L}(X,Y)},\end{aligned}$$

and the first result follows (note that $\{\|T_h\|_X\}$ is bounded because $\{T_h\}$ converges as $h \to 0$; also, $\|T^*\|_{\mathcal{L}(Y,X)} = \|T\|_{\mathcal{L}(X,Y)}$). The proof of the second is similar. ■

Theorem 4.5 *For each $k \in \mathbb{Z}^+$,*

$$\begin{aligned}\sigma_{h,k} &\to \sigma_k \text{ as } h \to 0, \\ \hat{\delta}(E_{h,k}, E_k) &\to 0 \text{ as } h \to 0, \\ \hat{\delta}(F_{h,k}, F_k) &\to 0 \text{ as } h \to 0,\end{aligned}$$

where E_k, F_k, $E_{h,k}$, and $F_{h,k}$ are defined by (4.1–4.4).

Proof: The convergence of $\sigma_{h,k}$ to σ_k and $\hat{\delta}(E_{h,k}, E_k)$ to zero follows from applying Theorem 4.3 to $\{T_h^*T_h\}$ and T^*T, while the convergence of $\hat{\delta}(F_{h,k}, F_k)$ to zero follows from applying Theorem 4.3 to $\{T_hT_h^*\}$ and TT^*. ■

We will not use Theorems 4.3 and 4.5 in the rest of the chapter, preferring to give a direct analysis that also provides rates of convergence.

We will use the following well-known max-min characterizations of the singular values:

$$\begin{aligned}\sigma_k &= \max_{\substack{S\subset X \\ \dim(S)=k}} \min_{\substack{x\in S \\ x\neq 0}} \frac{\|Tx\|_Y}{\|x\|_X}, \\ \sigma_{h,k} &= \max_{\substack{S\subset X \\ \dim(S)=k}} \min_{\substack{x\in S \\ x\neq 0}} \frac{\|T_hx\|_Y}{\|x\|_X}\end{aligned} \tag{4.5}$$

(see [13, Theorem 8.6.1] or [18, Chapter 28, Theorem 4]).

A particular discretization (discretization by projection or variational approximation) is of special interest: We choose families $\{X_h\}$ and $\{Y_h\}$ of finite-dimensional subspaces of X and Y, respectively, having the property that $P_{X_h} \to I_X$ and $P_{Y_h} \to I_Y$ pointwise as $h \to 0$ (where I_X and I_Y denote the identity operators on X and Y, respectively), and define $T_h = P_{Y_h}TP_{X_h}$.

It is straightforward to use the max-min formulas to prove the following results:

Theorem 4.6

1. *For each $k \in \mathbb{Z}^+$, $|\sigma_{h,k} - \sigma_k| \leq \|T_h - T\|_{\mathcal{L}(X,Y)}$ for all $h > 0$. (This holds regardless of the form of T_h.)*

2. *If $T_h = P_{Y_h} T P_{X_h}$, then, for each $k \in \mathbb{Z}^+$, $\sigma_{h,k} \leq \sigma_k$ for all $h > 0$.*

Proof: The first statement has been proved in [12, Chapter IV, Cor. 1.6], but we will provide our own proof below. Using (4.5), we obtain

$$\begin{aligned}
\sigma_{h,k} &= \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \frac{\|T_h x\|_Y}{\|x\|_X} \\
&\leq \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|Tx\|_Y}{\|x\|_X} + \frac{\|(T_h - T)x\|_Y}{\|x\|_X} \right) \\
&\leq \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|Tx\|_Y}{\|x\|_X} + \|T_h - T\|_{\mathcal{L}(X,Y)} \right) \\
&= \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|Tx\|_Y}{\|x\|_X} \right) + \|T_h - T\|_{\mathcal{L}(X,Y)} = \sigma_k + \|T_h - T\|_{\mathcal{L}(X,Y)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sigma_k &= \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \frac{\|Tx\|_Y}{\|x\|_X} \\
&\leq \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|T_h x\|_Y}{\|x\|_X} + \frac{\|(T - T_h)x\|_Y}{\|x\|_X} \right) \\
&\leq \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|T_h x\|_Y}{\|x\|_X} + \|T - T_h\|_{\mathcal{L}(X,Y)} \right) \\
&= \max_{\substack{S \subset X \\ \dim(S)=k}} \min_{\substack{x \in S \\ x \neq 0}} \left(\frac{\|T_h x\|_Y}{\|x\|_X} \right) + \|T - T_h\|_{\mathcal{L}(X,Y)} = \sigma_{h,k} + \|T - T_h\|_{\mathcal{L}(X,Y)}.
\end{aligned}$$

Thus, $\sigma_k - \|T - T_h\|_{\mathcal{L}(X,Y)} \leq \sigma_{h,k} \leq \sigma_k + \|T - T_h\|_{\mathcal{L}(X,Y)}$.

To prove the second statement, note that if $\sigma_{h,k} = 0$, then the statement holds trivially. Now consider the operator $\overline{T_h} : X_h \to Y_h$ defined by $\overline{T_h} = P_{Y_h} T|_{X_h}$. We will prove that any nonzero singular value of $T_h = P_{Y_h} T P_{X_h}$ is also a singular value of $\overline{T_h}$. To show this, note that nonzero any singular value $\sigma_{h,k}$ of T_h satisfies $T_h^* T_h \phi = \sigma_{h,k}^2 \phi$ for some $\phi \in X$; that is,

$$(P_{X_h} T^* P_{Y_h} T P_{X_h}) \phi = \sigma_{h,k}^2 \phi.$$

In fact, the above equation implies that $\phi \in X_h$ and hence, the above equation can be written as

$$(P_{X_h}T^*P_{Y_h}T)\phi = \sigma_{h,k}^2\phi.$$

The adjoint of $\overline{T_h} : Y_h \to X_h$ is equal to $P_{X_h}T^*|_{P_{Y_h}}$ and thus, $\overline{T_h}^*\overline{T_h} = P_{X_h}T^*P_{Y_h}T|_{X_h}$. This implies that $\sigma_{h,k}$ and ϕ will also satisfy $\overline{T_h}^*\overline{T_h}\phi = \sigma_{h,k}^2\phi$. Thus, $\sigma_{h,k}$ and ϕ are a singular value and singular vector respectively, for $\overline{T_h}$.

We finish the proof by noting that for any singular value $\sigma_{h,k}$ of $\overline{T_h}$, we have

$$\begin{aligned} \sigma_{h,k} = \max_{\substack{S\subset X_h\\ \dim(S)=k}} \min_{\substack{x\in S\\ x\neq 0}} \frac{\|\overline{T_h}x\|_Y}{\|x\|_X} = \max_{\substack{S\subset X_h\\ \dim(S)=k}} \min_{\substack{x\in S\\ x\neq 0}} \frac{\|P_{Y_h}Tx\|_Y}{\|x\|_X} &\leq \max_{\substack{S\subset X_h\\ \dim(S)=k}} \min_{\substack{x\in S\\ x\neq 0}} \frac{\|Tx\|_Y}{\|x\|_X} \\ &\leq \max_{\substack{S\subset X\\ \dim(S)=k}} \min_{\substack{x\in S\\ x\neq 0}} \frac{\|Tx\|_Y}{\|x\|_X} = \sigma_k. \blacksquare \end{aligned}$$

Note that, in Theorem 4.6, we define $\sigma_{h,k} = 0$ for $k > \dim(X_h)$.

Relatively little work has been done on approximating the SVE of a compact operator. Hansen [15] analyzed the rate of convergence of the *method of moments* (equivalent to variational approximation) for the case that T is an integral operator of the first kind. Specifically, assume that

$$(Tx)(s) = \int_{I_t} k(s,t)x(t)\,dt,\ (T_hx)(s) = \int_{I_t} k_h(s,t)x(t)\,dt,\ s \in I_s,$$

where k_h is the kernel produced by variational approximation (namely, the projection of k onto the tensor product space $Y_h \otimes X_h$ in $L^2(I_s \times I_t)$). Hansen's analysis is based on

$$\tilde{\epsilon}_h = \|k_h - k\|_{L^2(I_s\times I_t)}$$

(an upper bound for $\|T_h - T\|_{\mathcal{L}(X,Y)}$), and his Theorem 4 implies that

$$\sigma_k - \sigma_{h,k} \leq \frac{\tilde{\epsilon}_h^2}{\sigma_{h,k}}\,\forall\, k = 1, 2, \ldots, N_h = \min\{\dim(X_h), \dim(Y_h)\}.$$

As we show below (Theorem 4.16), a more precise estimate can be formulated in terms of the optimal approximation errors for ϕ_k and ψ_k, which are $\|(I - P_{X_h})\phi_k\|_X$ and $\|(I - P_{Y_h})\psi_k\|_Y$, respectively, and ϵ_h (rather than $\tilde{\epsilon}_h$). (This statement assumes that the singular spaces corresponding to σ_k are one-dimensional; see Theorem 4.16 for the general statement.)

Regarding the singular vectors, Hansen's Theorem 5 implies (under the assumption that the singular spaces are one-dimensional) that

$$\max\{\|\phi_{h,k} - \phi_k\|_X, \|\psi_{h,k} - \psi_k\|_Y\} \leq \frac{\sqrt{2}\sqrt{\tilde{\epsilon}_h}}{\sqrt{\sigma_k - \sigma_{k+1}}} \, \forall \, k = 1, 2, \ldots, N_h.$$

Our Theorem 4.8 below improves this to an $O(\epsilon_h)$ upper bound for an arbitrary approximation T_h of T, while Theorem 4.11 improves the estimate for the case of variational approximation and shows when the error in the singular vectors is asymptotically optimal.

Given the paucity of results about approximation of the singular value expansion, it is natural to look to the literature on eigenvalues and eigenvectors, which is extensive (see, for example, [3] or [5]). We could aproximate the singular values and right singular vectors of T by approximating the eigenvalues and eigenvectors of T^*T as, for example, described in Babuška and Osborn [2]. They analyze Galerkin methods for solving self-adjoint eigenvalue problems posed in variational form; the eigenvalue problem $T^*T\phi = \sigma^2\phi$ would be posed in variational form as

$$\text{find } \phi \in X, \ \lambda \in \mathbb{R}, \text{ such that } \langle \phi, v \rangle_X = \lambda \langle T\phi, Tv \rangle_Y \ \forall \, v \in X \tag{4.6}$$

(with $\sigma^2 = \lambda^{-1}$). The Galerkin method discretizes this variational problem as

$$\text{find } \phi \in X_h, \ \lambda \in \mathbb{R}, \text{ such that } \langle \phi, v \rangle_X = \lambda \langle T\phi, Tv \rangle_Y \ \forall \, v \in X_h. \tag{4.7}$$

It is straightforward to show that (4.7) is equivalent to $T_h^*T_h\phi = \sigma^2\phi$ with $T_h = TP_{X_h}$ (again, with $\sigma^2 = \lambda^{-1}$).

Given any $u \in E_k$ (that is, any right singular vector corresponding to the singular value σ_k) with $\|u\|_X = 1$, and not assuming that E_k is one-dimensional, the optimal approximation error for u is $\|(I - P_{X_h})u\|_X$, and we obviously have

$$\|(I - P_{X_h})u\|_X \leq \|(I - P_{E_{h,k}})u\|_X.$$

Theorem 4.4 of [2] depends on

$$\hat{\epsilon}_h = \|T_h^*T_h - T^*T\|_{\mathcal{L}(X,X)}$$

and implies that

$$\|(I - P_{E_{h,k}})u\|_X \leq (1 + d_k\hat{\epsilon}_h)\|(I - P_{X_h})u\|_X, \tag{4.8}$$

where d_k is a constant proportional to the reciprocal of the gap between σ_k^2 and the closest distinct eigenvalue of T^*T. (Babuška and Osborn also prove a version of (4.8) with $\hat{\epsilon}_h$ replaced by $\|T_h^*T_h - T^*T\|_{\mathcal{L}(X_1,X_1)}^2$, where X_1 is the completion of $\mathcal{N}(T)^\perp$ under the inner product defined by $\langle u, v \rangle_{X_1} = \langle Tu, Tv \rangle_Y$. For some problems, this provides a more precise bound, because $\|T_h^*T_h - T^*T\|_{\mathcal{L}(X_1,X_1)}^2$ can be asymptotically smaller than $\hat{\epsilon}_h$ for some problems. We will not discuss this and similar refined bounds any further.) The bound (4.8) is comparable to the result in our Theorem 4.11:

$$\|(I - P_{E_{h,k}})u\|_X \le (1 + \sqrt{2}q_k\epsilon_h)\|(I - P_{X_h})u\|_X + \sqrt{2}q_k\epsilon_h\|(I - P_{Y_h})v\|_Y. \tag{4.9}$$

Our constant $\sqrt{2}q_k$ is proportional to the reciprocal of the gap between σ_k and the closest distinct singular value of T, which implies that $\sqrt{2}q_k$ is asymptotically smaller than d_k as $k \to \infty$. On the other hand, while $\hat{\epsilon}_h = O(\epsilon_h)$ as $h \to 0$, it could be the case that $\hat{\epsilon}_h$ is asymptotically smaller than ϵ_h. Therefore, the two constants, $d_k\hat{\epsilon}_h$ and $\sqrt{2}q_k\epsilon_h$, are not directly comparable. Nevertheless, (4.8) implies that

$$\|(I - P_{E_{h,k}})u\|_X \sim \|(I - P_{X_h})u\|_X \text{ as } h \to 0$$

and (4.9) yields the same result *if* $\epsilon_h\|(I - P_{Y_h})v\|_Y = o(\|(I - P_{X_h})u\|_X)$. We will see by example that this is not always the case and thus that optimal approximability of the right singular vectors can be lost due to poor approximability of the corresponding left singular vectors.

With respect to the singular values, Theorem 4.2 of [2] implies that there exists a right singular vector u corresponding to σ_k such that

$$\frac{\sigma_k^2 - \sigma_{h,k}^2}{\sigma_{h,k}^2} \le (1 + d_k\hat{\epsilon}_h)\|(I - P_{X_h})u\|_X^2.$$

Since

$$\frac{\sigma_k^2 - \sigma_{h,k}^2}{\sigma_{h,k}^2} = \frac{\sigma_k - \sigma_{h,k}}{\sigma_{h,k}} \cdot \frac{\sigma_k + \sigma_{h,k}}{\sigma_{h,k}}$$

and (since $\sigma_{h,k} \le \sigma_k$)

$$\frac{\sigma_{h,k}}{\sigma_k + \sigma_{h,k}} \le \frac{\sigma_{h,k}}{\sigma_{h,k} + \sigma_{h,k}} = \frac{1}{2},$$

we obtain

$$\frac{\sigma_k - \sigma_{h,k}}{\sigma_{h,k}} \le \frac{1}{2}(1 + d_k\hat{\epsilon}_h)\|(I - P_{X_h})u\|_X^2. \tag{4.10}$$

Our Theorem 4.16 implies that if we estimate the singular values by computing the SVE of $T_h = P_{Y_h}TP_{X_h}$ directly (rather than solving an eigenvalue problem), we obtain

$$\begin{aligned}\frac{\sigma_k - \sigma_{h,\ell}}{\sigma_k} \le &\frac{1}{2}\left(\|(I - P_{X_h})u\|_X^2 + \|(I - P_{Y_h})v\|_Y^2\right) + \\ &C_k\epsilon_h\left(\|(I - P_{X_h})u\|_X + \|(I - P_{Y_h})v\|_Y\right)^2 \ \forall\, h > 0 \text{ sufficiently small,}\end{aligned}$$

where $u \in X$ and $v \in Y$ satisfy $\|u\|_X = 1$, $\|v\|_Y = 1$, and $Tu = \sigma_k v$. If $\|(I - P_{X_h})u\|_X$ and $\|(I - P_{Y_h})v\|_Y$ go to zero at the same rate, this suggests that the error in the singular values (computed directly) is twice the error in the same quantities computed by the eigenvalue approach. However, this ignores the limitations of finite precision arithmetic, as discussed below.

It is natural to compute the singular values and singular vectors of T directly, rather than compute the eigenvalues and eigenvectors of T^*T, for several reasons. First, given that $\sigma_k \to 0$ as $k \to \infty$, we can expect to compute σ_k directly as long as it is larger than machine epsilon ϵ_{mach}. By the eigenvalue approach, we expect to be able to compute σ_k^2 as long as it remains above ϵ_{mach}, that is, as long as σ_k is larger than $\sqrt{\epsilon_{mach}}$. Thus, the eigenvalue approach reduces the range of the singular values than can be estimated. Second, both approaches require the computation of a Galerkin matrix. The eigenvalue approach requires $\hat{A}$ defined by

$$\hat{A}_{ij} = \langle x_i, T^*Tx_j \rangle_X = \langle Tx_i, Tx_j \rangle_X \,,$$

while the direct approach requires A defined by

$$A_{ij} = \langle y_i, Tx_j \rangle_X$$

(where $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ are the bases for X_h and Y_h, respectively). In the common case that T is an integral operator, $\hat{A}_{ij}$ is defined by a triple integral, while A_{ij} is defined by a double integral. Therefore, the direct approach may be more efficient. Finally, the eigenvalue approach gives no direct estimate of the left singular vectors.

In the remainder of the chapter, we will analyze the errors in both the singular values and the singular vectors in both the generic case (T_h is assumed only to converge to T in the operator norm) and the case of variational approximation ($T_h = P_{Y_h}TP_{X_h}$ and the related cases of $T_h = TP_{X_h}$ and $T_h = P_{Y_h}T$). Specifically, Section 4.2 contains error estimates for the computed singular vectors in the generic case. In Section 4.3, we present an analysis of the convergence of both the singular vectors and singular values in the case of variational approximation. Numerical examples are presented in Section 4.4. In Section 4.5, we show how to compute the singular values and singular vectors of $T_h = P_{Y_h}TP_{X_h}$ from the singular value decomposition (SVD) of a scaled Galerkin matrix. We present some conclusions in Section 4.6.

4.2 Convergence of singular vectors: the general case

We have already seen that

$$|\sigma_{h,k} - \sigma_k| \le \|T_h - T\|_{L(X,Y)} \le \epsilon_h \text{ for all } h > 0.$$

We now show that $\hat{\delta}(E_{h,k}, E_k)$ and $\hat{\delta}(F_{h,k}, F_k)$ also converge to zero like $O(\epsilon_h)$. We begin by establishing more notation.

We have already defined E_k and F_k in (4.1) and (4.2); let I_k be the corresponding index set, so that

$$E_k = \mathrm{sp}\{\phi_\ell : \ell \in I_k\},\ F_k = \mathrm{sp}\{\psi_\ell : \ell \in I_k\}.$$

Let $\sigma_{h,\ell}$, $\phi_{h,\ell}$, $\psi_{h,\ell}$, $\ell \in \mathcal{I}_h$, be the singular values and singular vectors of T_h, where $\mathcal{I}_h = \{1, 2, \ldots, N_h\}$ or $I_h = \mathbb{Z}^+$, according as T_h has finite rank or not, and define

$$I_{h,k} = \{\ell \in \mathcal{I}_h : \sigma_{h,\ell} \to \sigma_k \text{ as } h \to 0\}.$$

(Note that $I_{h,k} = I_k$ for all $h > 0$ sufficiently small.) Then $E_{h,k} = \mathrm{sp}\{\phi_{h,\ell} : \ell \in I_{h,k}\}$. Extend $\{\phi_{h,\ell} : \ell \in \mathcal{I}_h\}$ to a complete orthonormal sequence $\{\phi_{h,\ell} : \ell \in \mathcal{J}_h\}$ for X ($\mathcal{I}_h \subset \mathcal{J}_h$), and define $\sigma_{h,\ell} = 0$ for $\ell \in \mathcal{J}_h \setminus \mathcal{I}_h$.

We will write gap_k for the gap between σ_k and the nearest distinct singular value of T. If $k > 1$ and $\sigma_{k-1} > \sigma_k = \sigma_{k+1} = \cdots = \sigma_{k+n_k-1} > \sigma_{k+n_k}$, then

$$\mathrm{gap}_k = \min\{\sigma_{k-1} - \sigma_k, \sigma_k - \sigma_{k+n_k}\},$$

while $\mathrm{gap}_1 = \sigma_1 - \sigma_t$, where σ_t is the largest singular value not equal to σ_1.

Lemma 4.7 *Assume that $\|T_h - T\|_{\mathcal{L}(X,Y)} \le \epsilon_h$ for $h > 0$. Then, for each $k \in \mathbb{Z}^+$,*

$$\max\left\{\frac{1}{|\sigma_k - \sigma_{h,\ell}|} : \ell \in \mathcal{J}_h \setminus I_{h,k}\right\} \le \frac{\sqrt{2}}{\mathrm{gap}_k}\ \forall\, h > 0 \textit{ sufficiently small.}$$

Proof: Let $k \in \mathbb{Z}^+$ be given and note that $\sigma_{h,\ell} \to \sigma_\ell$ as $h \to 0$ for all $\ell \in \mathbb{Z}^+$. It follows that, for $h > 0$ sufficiently small, $I_{h,k} = I_k$ and

$$|\sigma_{h,\ell} - \sigma_k| \ge \frac{\mathrm{gap}_k}{\sqrt{2}}\ \forall\, \ell \in \mathcal{J}_h \setminus I_{h,k}.$$

The result follows. ■

In Lemma 4.7, we could obviously replace $\sqrt{2}$ by any constant strictly greater than 1. We will write $q_k = \sqrt{2}/\mathrm{gap}_k$.

Theorem 4.8 *Assume that $\|T_h - T\|_{\mathcal{L}(X,Y)} \le \epsilon_h$ for $h > 0$. Let $k \in \mathbb{Z}^+$ and let E_k and $E_{h,k}$ be defined by (4.1) and (4.3), respectively. Then, for all $u \in E_k$ satisfying $\|u\|_X = 1$ and for all $h > 0$ sufficiently small, $\|(I - P_{E_{h,k}})u\|_X \le 2q_k\epsilon_h$.*

Proof: Given $u \in E_k$, we define $v = \sigma_k^{-1} T u$; then $T^* v = \sigma_k u$. We have

$$\begin{aligned} u &= \sum_{\ell \in \mathcal{J}_h} \langle \phi_{h,\ell}, u \rangle_X \, \phi_{h,\ell}, \\ P_{E_{h,k}} u &= \sum_{\ell \in I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X \, \phi_{h,\ell} \end{aligned}$$

and hence

$$u - P_{E_{h,k}} u = \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X \, \phi_{h,\ell}.$$

For each $\ell \in \mathcal{J}_h \setminus I_{h,k}$, we have

$$\begin{aligned} \sigma_k |\, \langle \phi_{h,\ell}, u \rangle_X \,| &= |\, \langle \phi_{h,\ell}, \sigma_k u \rangle_X \,| = |\, \langle \phi_{h,\ell}, T^* v \rangle_X \,| = |\, \langle T \phi_{h,\ell}, v \rangle_Y \,|, \\ \sigma_{h,\ell} |\, \langle \phi_{h,\ell}, u \rangle_X \,| &= |\, \langle \sigma_{h,\ell} \phi_{h,\ell}, u \rangle_X \,| = |\, \langle T_h^* \psi_{h,\ell}, u \rangle_X \,| = |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,|, \\ \sigma_k |\, \langle \psi_{h,\ell}, v \rangle_Y \,| &= |\, \langle \psi_{h,\ell}, \sigma_k v \rangle_Y \,| = |\, \langle \psi_{h,\ell}, T u \rangle_Y \,|, \\ \sigma_{h,\ell} |\, \langle \psi_{h,\ell}, v \rangle_Y \,| &= |\, \langle \sigma_{h,\ell} \psi_{h,\ell}, v \rangle_Y \,| = |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,|. \end{aligned}$$

Subtracting yields

$$\begin{aligned} (\sigma_k - \sigma_{h,\ell}) |\, \langle \phi_{h,\ell}, u \rangle_X \,| &= \ |\, \langle T \phi_{h,\ell}, v \rangle_Y \,| - |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,|, \\ (\sigma_k - \sigma_{h,\ell}) |\, \langle \psi_{h,\ell}, v \rangle_Y \,| &= |\, \langle \psi_{h,\ell}, T u \rangle_Y \,| - |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,|. \end{aligned}$$

We then add to obtain

$$\begin{aligned} &(\sigma_k - \sigma_{h,\ell}) \left(|\, \langle \phi_{h,\ell}, u \rangle_X \,| + |\, \langle \psi_{h,\ell}, v \rangle_Y \,| \right) \\ &= |\, \langle T \phi_{h,\ell}, v \rangle_Y \,| - |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,| + |\, \langle \psi_{h,\ell}, T u \rangle_Y \,| - |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,| \\ &= |\, \langle T \phi_{h,\ell}, v \rangle_Y \,| - |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,| + |\, \langle \psi_{h,\ell}, T u \rangle_Y \,| - |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,|, \end{aligned}$$

which yields

$$\begin{aligned} &|\sigma_k - \sigma_{h,\ell}| \left(|\, \langle \phi_{h,\ell}, u \rangle_X \,| + |\, \langle \psi_{h,\ell}, v \rangle_Y \,| \right) \\ = \ & \big|\, |\, \langle T \phi_{h,\ell}, v \rangle_Y \,| - |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,| + |\, \langle \psi_{h,\ell}, T u \rangle_Y \,| - |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,| \big| \\ \le \ & \big|\, |\, \langle T \phi_{h,\ell}, v \rangle_Y \,| - |\, \langle T_h \phi_{h,\ell}, v \rangle_Y \,| \big| + \big|\, |\, \langle \psi_{h,\ell}, T u \rangle_Y \,| - |\, \langle \psi_{h,\ell}, T_h u \rangle_Y \,| \big| \\ \le \ & \left| \langle (T - T_h) \phi_{h,\ell}, v \rangle_Y \right| + \left| \langle \psi_{h,\ell}, (T - T_h) u \rangle_Y \right|. \end{aligned}$$

This implies that

$$|\, \langle \phi_{h,\ell}, u \rangle_X \,| \le \frac{\left| \langle (T - T_h) \phi_{h,\ell}, v \rangle_Y \right| + \left| \langle \psi_{h,\ell}, (T - T_h) u \rangle_Y \right|}{|\sigma_k - \sigma_{h,\ell}|}. \tag{4.11}$$

Therefore,

$$\begin{aligned}
\|u - P_{E_{h,k}} u\|_X^2 &= \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X^2 \\
&\le \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \left(\frac{\left|\langle (T - T_h)\phi_{h,\ell}, v\rangle_Y\right| + \left|\langle \psi_{h,\ell}, (T - T_h)u\rangle_Y\right|}{|\sigma_k - \sigma_{h,\ell}|} \right)^2 \\
&\le 2q_k^2 \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \left(\left|\langle \phi_{h,\ell}, (T - T_h)^* v\rangle_X\right|^2 + \left|\langle \psi_{h,\ell}, (T - T_h)u\rangle_Y\right|^2 \right) \\
&\le 2q_k^2 \left(\|(T - T_h)^* v\|_X^2 + \|(T - T_h)u\|_Y^2 \right) \le 4q_k^2 \epsilon_h^2
\end{aligned}$$

(since $\|T_h - T\|_{\mathcal{L}(X,Y)} = \|T_h^* - T^*\|_{\mathcal{L}(Y,X)} \le \epsilon_h$ and $\|u\|_X = \|v\|_Y = 1$). Therefore, $\|u - P_{E_{h,k}} u\|_X \le 2q_k\epsilon_h$, as desired. ■

The previous theorem implies that $\delta(E_k, E_{h,k}) \le 2q_k\epsilon_h$ for all $h > 0$ sufficiently small. Since Lemma 4.2 shows that $\delta(E_{h,k}, E_k) = \delta(E_k, E_{h,k})$ for all h sufficiently small (h must be small enough that $\dim(E_{h,k}) = \dim(E_k)$), we obtain the desired bound on the gap between E_k and $E_{h,k}$.

Corollary 4.9 *For each $k \in \mathbb{Z}^+$,*

$$\hat{\delta}(E_{h,k}, E_k) \le 2q_k\epsilon_h \;\forall\, h > 0 \textit{ sufficiently small.}$$

The same analysis, applied to T^* and T_h^*, shows that

$$\hat{\delta}(F_{h,k}, F_k) \le 2q_k\epsilon_h \;\forall\, h > 0 \text{ sufficiently small,}$$

where F_k and $F_{h,k}$ are defined by (4.2) and (4.4).

4.3 Accelerated convergence

The case of variational approximation deserves special attention because it leads to increased rates of convergence. Given the families of finite-dimensional subspaces $\{X_h\}$ and $\{Y_h\}$ of X and Y, respectively, we define $T_h : X \to Y$ by $T_h = P_{Y_h} T P_{X_h}$ and

$$\epsilon_h = \|P_{Y_h} T P_{X_h} - T\|_{\mathcal{L}(X,Y)} \;\forall\, h > 0.$$

Under our assumptions on $\{X_h\}$ and $\{Y_h\}$ (namely, that $P_{X_h} \to I_X$ and $P_{Y_h} \to I_Y$ pointwise), it is guaranteed that $\epsilon_h \to 0$ as $h \to 0$.

Theorem 4.10 *The operator $P_{Y_h} T P_{X_h}$ converges in the operator norm to T as $h \to 0$.*

Proof: Consider

$$\begin{aligned}\|P_{Y_h}TP_{X_h}-T\|_{\mathcal{L}(X,Y)} &\le \|P_{Y_h}TP_{X_h}-P_{Y_h}T\|_{\mathcal{L}(X,Y)}+\|P_{Y_h}T-T\|_{\mathcal{L}(X,Y)}\\ &\le \|T(P_{X_h}-I_X)\|_{\mathcal{L}(X,Y)}+\|(P_{Y_h}-I_Y)T\|_{\mathcal{L}(X,Y)}.\end{aligned}$$

The second norm goes to zero by a standard result (see [1] or [11, Theorem 7]), while the first goes to zero by Theorem 16 of [11]. ■

The results of Section 4.2 apply and show that the singular values and corresponding singular spaces of T_h satisfy

$$\begin{aligned}|\sigma_{h,k}-\sigma_k| &\le \epsilon_h,\\ \hat{\delta}(E_{h,k},E_k) &\le 2q_k\epsilon_h,\\ \hat{\delta}(F_{h,k},F_k) &\le 2q_k\epsilon_h,\end{aligned}$$

where the above inequalities hold for all $h>0$ sufficiently small. We will now show that, for $T_h=P_{Y_h}TP_{X_h}$, better rates of convergence are obtained. Recall that $\{\phi_{h,\ell}:\ell\in I_h\}$ was extended to a complete orthonormal sequence $\{\phi_{h,\ell}:\ell\in\mathcal{J}_h\}$ for X. We now assume that this is done so that $\{\phi_{h,\ell}:\ell\in\tilde{I}_h\}$ ($I_h\subset\tilde{I}_h\subset\mathcal{J}_h$) is an orthonormal basis for X_h ($\tilde{I}_h=I_h$ if $\mathcal{N}(T_h)\cap X_h$ is trivial). Similarly, we extend $\{\psi_{h,\ell}:\ell\in I_h\}$ to a complete orthonormal sequence $\{\psi_{h,\ell}:\ell\in\mathcal{K}_h\}$ for Y and assume that $\{\psi_{h,\ell}:\ell\in\hat{I}_h\}$ is an orthonormal basis for Y_h. These definitions imply that $\{\phi_{h,\ell}:\ell\in\mathcal{J}_h\setminus\tilde{I}_h\}$ is a complete orthonormal sequence for $X_h^\perp$ and $\{\psi_{h,\ell}:\ell\in\mathcal{K}_h\setminus\hat{I}_h\}$ is a complete orthonormal sequence for $Y_h^\perp$.

4.3.1 Singular vectors

Theorem 4.11 *Suppose $u\in E_k$ and $v\in F_k$ satisfy $Tu=\sigma_k v$ and $T_h=P_{Y_h}TP_{X_h}$. Then, for all $h>0$ sufficiently small,*

$$\|(I-P_{E_{h,k}})u\|_X \le (1+\sqrt{2}q_k\epsilon_h)\|(I-P_{X_h})u\|_X+\sqrt{2}q_k\epsilon_h\|(I-P_{Y_h})v\|_Y \quad (4.12)$$

$$\|(I-P_{F_{h,k}})v\|_Y \le (1+\sqrt{2}q_k\epsilon_h)\|(I-P_{Y_h})v\|_Y+\sqrt{2}q_k\epsilon_h\|(I-P_{X_h})u\|_X. \quad (4.13)$$

Proof: As in the proof of Theorem 4.8,

$$(I-P_{E_{h,k}})u=\sum_{\ell\in\mathcal{J}_h\setminus I_{h,k}}\langle\phi_{h,\ell},u\rangle_X\,\phi_{h,\ell},$$

which yields

$$\|(I-P_{E_{h,k}})u\|_X^2=\sum_{\ell\in\mathcal{J}_h\setminus\tilde{I}_h}\langle\phi_{h,\ell},u\rangle_X^2+\sum_{\ell\in\tilde{I}_h\setminus I_{h,k}}\langle\phi_{h,\ell},u\rangle_X^2.$$

The first term on the right of the equals sign is exactly $\|(I - P_{X_h})u\|_X^2$. To estimate the second term, we use the following inequality from the proof of Theorem 4.8:

$$| \langle \phi_{h,\ell}, u \rangle_X | \leq \frac{1}{|\sigma_k - \sigma_{h,\ell}|} \left(\left| \langle (T - T_h)\phi_{h,\ell}, v \rangle_Y \right| + \left| \langle \psi_{h,\ell}, (T - T_h)u \rangle_Y \right| \right).$$

We argue as follows:

$$\begin{aligned}\sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X^2 &\leq \sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \left(\frac{\left| \langle (T - T_h)\phi_{h,\ell}, v \rangle_Y \right| + \left| \langle \psi_{h,\ell}, (T - T_h)u \rangle_Y \right|}{|\sigma_k - \sigma_{h,\ell}|} \right)^2 \\ &\leq 2q_k^2 \sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \left(\left| \langle (T - T_h)\phi_{h,\ell}, v \rangle_Y \right|^2 + \left| \langle \psi_{h,\ell}, (T - T_h)u \rangle_Y \right|^2 \right).\end{aligned}$$

For any $x \in X_h$,

$$(T - T_h)x = Tx - P_{Y_h} T P_{X_h} x = Tx - P_{Y_h} Tx = (I - P_{Y_h})Tx \in Y_h^\perp,$$

and therefore

$$\langle (T - T_h)\phi_{h,\ell}, v \rangle_Y = \langle (T - T_h)\phi_{h,\ell}, (I - P_{Y_h})v \rangle_Y$$

because $(T - T_h)\phi_{h,\ell} \in Y_h^\perp$ and $P_{Y_h} v \in Y_h$. Similarly, for $y \in Y_h$,

$$(T - T_h)^* y = T^* y - P_{X_h} T^* P_{Y_h} y = T^* y - P_{X_h} T^* y = (I - P_{X_h})T^* y \in X_h^\perp,$$

which yields

$$\langle \psi_{h,\ell}, (T - T_h)u \rangle_Y = \langle (T - T_h)^* \psi_{h,\ell}, u \rangle_X = \langle (T - T_h)^* \psi_{h,\ell}, (I - P_{X_h})u \rangle_X.$$

Therefore,

$$\begin{aligned}&\sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \left(\left| \langle (T - T_h)\phi_{h,\ell}, v \rangle_Y \right|^2 + \left| \langle \psi_{h,\ell}, (T - T_h)u \rangle_Y \right|^2 \right) \\ =& \sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \left(\left| \langle (T - T_h)\phi_{h,\ell}, (I - P_{Y_h})v \rangle_Y \right|^2 + \left| \langle (T - T_h)^* \psi_{h,\ell}, (I - P_{X_h})u \rangle_X \right|^2 \right) \\ =& \sum_{\ell \in \tilde{I}_h \setminus I_{h,k}} \left(\left| \langle \phi_{h,\ell}, (T - T_h)^*(I - P_{Y_h})v \rangle_X \right|^2 + \left| \langle \psi_{h,\ell}, (T - T_h)(I - P_{X_h})u \rangle_Y \right|^2 \right) \\ \leq& \left(\|(T - T_h)^*(I - P_{Y_h})v\|_X^2 + \|(T - T_h)(I - P_{X_h})u\|_Y^2 \right) \\ \leq& \epsilon_h^2 \left(\|(I - P_{Y_h})v\|_Y^2 + \|(I - P_{X_h})u\|_X^2 \right).\end{aligned}$$

The conclusion is

$$\|(I - P_{E_{h,k}})u\|_X^2 \leq \|(I - P_{X_h})u\|_X^2 + 2q_k^2 \epsilon_h^2 \left(\|(I - P_{Y_h})v\|_Y^2 + \|(I - P_{X_h})u\|_X^2 \right).$$

Inequality (4.12) follows immediately.

The proof of (4.13) is exactly analogous. ■

Inequality (4.12) suggests that the approximation error for a given right singular vector u is affected by the optimal approximation error for that singular vector and also the optimal approximation error for the corresponding left singular vector. As long as

$$\epsilon_h \|(I - P_{Y_h})v\|_Y = o\left(\|(I - P_{X_h})u\|_X\right) \text{ as } h \to 0,$$

it follows that

$$\|(I - P_{E_{h,k}})u\|_X \sim \|(I - P_{X_h})u\|_X \text{ as } h \to 0,$$

that is, the error in the approximation to u is asymptotically optimal. However, it is not difficult to construct an example in which

$$\|(I - P_{X_h})u\|_X = o\left(\epsilon_h \|(I - P_{Y_h})v\|_Y\right) \text{ as } h \to 0,$$

in which case the error in the approximation to u is suboptimal. Examples are presented in later sections.

The situation with respect to optimal approximation of the left singular vectors is exactly analogous.

The bound (4.12) is not directly comparable to the result (4.8) of Babuška and Osborn. When we approximate both the right and left singular vector simultaneously, it is not guaranteed that we obtain an optimal rate of convergence for both, although Theorem 4.11 shows that at least one of the left or right singular vectors will exhibit an optimal rate of convergence. The examples given below verify that suboptimal convergence is observed in some cases. If, for some reason, it were desired to approximate just the singular values and either the right singular vectors or the left singular vectors, we could (at least in principle) proceed with $T_h = TP_{X_h}$ or $T_h = P_{Y_h}T$.

Theorem 4.12

1. *Let the family $\{X_h\}$ of subspaces of X be given and, for each $h > 0$, define $Y_h = T(X_h)$, $T_h = TP_{X_h}$. Let $u \in E_k$ be given and define $v = \sigma_k^{-1}Tu$. Then, for all $h > 0$ sufficiently small,*

$$\|(I - P_{E_{h,k}})u\|_X \leq (1 + \sqrt{2}q_k\epsilon_h)\|(I - P_{X_h})u\|_X, \tag{4.14}$$

$$\|(I - P_{F_{h,k}})v\|_Y \leq \|(I - P_{Y_h})v\|_Y + \sqrt{2}q_k\epsilon_h\|(I - P_{X_h})u\|_X. \tag{4.15}$$

2. *Let the family $\{Y_h\}$ of subspaces of Y be given and, for each $h > 0$, define $X_h = T^*(Y_h)$, $T_h = P_{Y_h}T$. Let $v \in F_k$ be given and define $u = \sigma_k^{-1}T^*v$. Then,*

for all $h > 0$ sufficiently small,

$$\|(I - P_{E_{h,k}})u\|_X \leq \|(I - P_{X_h})u\|_X + \sqrt{2}q_k\epsilon_h\|(I - P_{Y_h})v\|_Y, \tag{4.16}$$
$$\|(I - P_{F_{h,k}})v\|_Y \leq (1 + \sqrt{2}q_k\epsilon_h)\|(I - P_{Y_h})v\|_Y. \tag{4.17}$$

Proof: The proofs of (4.14–4.17) are similar to the proofs of (4.12–4.13) and will not be given in detail. The difference between (4.12) and (4.14) is that, for $x \in X_h$ and $T = TP_{X_h}$, $(T - T_h)x = 0$ (rather than just $(T - T_h)x \in X_h^{\perp}$, as in the case of $T_h = P_{Y_h}TP_{X_h}$). The difference between (4.13) and (4.17) is similar. ■

Therefore, if we approximate T with $T_h = TP_{X_h}$, we are guaranteed that the error in the approximations of the right singular vectors is optimal and of the same quality as obtained in (4.8). Similarly. with $T_h = P_{Y_h}T$, we are guaranteed optimal approximation of the left singular vectors. In Section 4.5, we show how to compute the SVE of $P_{Y_h}TP_{X_h}$ by computing the SVD of a suitably scaled Galerkin matrix. The numerical computation of the SVE of $T_h = TP_{X_h}$ or $T_h = P_{Y_h}T$ is problematic; we also comment on this in Section 4.5.

4.3.2 Singular values

We continue to discuss the case of $T_h = P_{Y_h}TP_{X_h}$. To start, we require a bound on $\|T_h(I - P_{E_{h,k}})u\|_Y$ for $u \in E_k$. We will use the following technical results.

Lemma 4.13 *If $\{s_k\}$ is a strictly decreasing sequence of positive real numbers that converges to zero, then*

$$\frac{s_{k-1}}{s_{k-1} - s_k} < \frac{2s_k}{\min\{s_{k-1} - s_k, s_k - s_{k+1}\}} \;\forall\, k > 1.$$

Proof: Given $k > 1$, we consider two cases.

1. If $s_{k-1} - s_k \leq s_k - s_{k+1}$, then it follows that $s_{k-1} - s_k < s_k$ and hence that $s_{k-1} < 2s_k$. Therefore,

$$\frac{s_{k-1}}{s_{k-1} - s_k} = \frac{s_{k-1}}{\min\{s_{k-1} - s_k, s_k - s_{k+1}\}} < \frac{2s_k}{\min\{s_{k-1} - s_k, s_k - s_{k+1}\}}.$$

   Thus the result holds in this case.

2. If $s_k - s_{k+1} < s_{k-1} - s_k$, define

$$\theta_1 = \frac{s_k}{s_{k-1}},\; \theta_2 = \frac{s_{k+1}}{s_{k-1}}$$

and note that $\theta_1 \in (0,1)$, $\theta_2 \in (0,\theta_1)$. We must show that

$$\begin{aligned} &\frac{s_{k-1}}{s_{k-1}-s_k} < \frac{2s_k}{\min\{s_{k-1}-s_k, s_k-s_{k+1}\}} = \frac{2s_k}{s_k-s_{k+1}} \\ \Leftrightarrow\ &\frac{s_{k-1}(s_k-s_{k+1})}{s_k(s_{k-1}-s_k)} < 2 \ \Leftrightarrow\ \frac{\theta_1-\theta_2}{\theta_1(1-\theta_1)} < 2. \end{aligned}$$

Note that $s_k - s_{k+1} < s_{k-1} - s_k$ is equivalent to $\theta_1 - \theta_2 < 1 - \theta_1$. It is now an easy exercise to prove that $(\theta_1-\theta_2)/(\theta_1(1-\theta_1)) < 2$ for θ_1, θ_2 satisfying $0 < \theta_2 < \theta_1 < 1$ and $\theta_1 - \theta_2 < 1 - \theta_1$. This completes the proof. ■

Lemma 4.14 *For a given $k \in \mathbb{Z}^+$, there exists $h_0 > 0$ such that*

$$\frac{\sigma_{h,\ell}}{|\sigma_k - \sigma_{h,\ell}|} \le \frac{2\sigma_k}{\text{gap}_k} = \sqrt{2}q_k\sigma_k \ \forall\, \ell \notin I_{h,k} \ \forall\, h \in (0,h_0).$$

Proof: We will assume that $\sigma_k = \sigma_{k+1} = \cdots = \sigma_{k+n_k-1} > \sigma_{k+n_k}$ and either $k = 1$ or $\sigma_{k-1} > \sigma_k$. If we prove the result in this case, it obviously follows for any other value of k. Suppose first that $\ell \ge k + n_k$. Then $\sigma_{h,\ell} \le \sigma_\ell \le \sigma_{k+n_k} < \sigma_k$. Therefore,

$$\begin{aligned} \frac{\sigma_{h,\ell}(\sigma_k - \sigma_{k+n_k})}{\sigma_k(\sigma_k - \sigma_{h,\ell})} \le \frac{\sigma_{k+n_k}(\sigma_k - \sigma_{k+n_k})}{\sigma_k(\sigma_k - \sigma_{k+n_k})} < 1 &\Rightarrow \frac{\sigma_{h,\ell}\text{gap}_k}{\sigma_k(\sigma_k - \sigma_{h,\ell})} < 1 \\ &\Rightarrow \frac{\sigma_{h,\ell}}{\sigma_k - \sigma_{h,\ell}} < \frac{\sigma_k}{\text{gap}_k}. \end{aligned}$$

This proves the desired result in the case $\ell \ge k+n_k$. If $\ell < k$, then there exists $h_0' > 0$ such that $\sigma_{h,k-1} > \sigma_k$ for all $h \in (0,h_0')$. For such h,

$$\frac{\sigma_{h,\ell}}{|\sigma_k - \sigma_{h,\ell}|} = \frac{\sigma_{h,\ell}}{\sigma_{h,\ell} - \sigma_k} \le \frac{\sigma_{h,k-1}}{\sigma_{h,k-1} - \sigma_k}$$

(using the fact that $s/(s-\sigma_k)$ increases as s decreases toward σ_k). Since

$$\frac{\sigma_{h,k-1}}{\sigma_{h,k-1} - \sigma_k} \to \frac{\sigma_{k-1}}{\sigma_{k-1} - \sigma_k} \text{ as } h \to 0$$

and

$$\frac{\sigma_{k-1}}{\sigma_{k-1} - \sigma_k} < \frac{2\sigma_k}{\text{gap}_k}$$

by Lemma 4.13, it follows that there exists $h_0 \in (0,h_0')$ such that

$$\frac{\sigma_{h,k-1}}{\sigma_{h,k-1} - \sigma_k} \le \frac{2\sigma_k}{\text{gap}_k} \ \forall\, h \in (0,h_0).$$

Thus the desired results holds in the case that $\ell < k$, and the proof is complete. ■

We can now prove the desired bound on $\|T_h(I - P_{E_{h,k}})u\|_Y$ for $u \in E_k$.

Theorem 4.15 *Let $k \in \mathbb{Z}^+$ be given, suppose $T_h = P_{Y_h} T P_{X_h}$, and let $u \in E_k$, $v \in F_k$ satisfy $Tu = \sigma_k v$. Then, for all $h > 0$ sufficiently small,*

$$\|T_h(u - P_{E_{h,k}} u)\|_Y \leq 2\sigma_k q_k \epsilon_h \left(\|(I - P_{Y_h})v\|_Y + \|(I - P_{X_h})u\|_X\right).$$

Proof: We will assume that $\sigma_k = \sigma_{k+1} = \cdots = \sigma_{k+n_k-1} > \sigma_{k+n_k}$ and either $k = 1$ or $\sigma_{k-1} > \sigma_k$. We have

$$u = \sum_{\ell \in \mathcal{J}_h} \langle \phi_{h,\ell}, u \rangle_X \phi_{h,\ell},$$
$$P_{E_{h,k}} u = \sum_{\ell \in I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X \phi_{h,\ell},$$

and therefore

$$(I - P_{E_{h,k}})u = \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \langle \phi_{h,\ell}, u \rangle_X \phi_{h,\ell}.$$

This yields

$$T_h(u - P_{E_{h,k}} u) = \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \sigma_{h,\ell} \langle \phi_{h,\ell}, u \rangle_X \psi_{h,\ell}$$
$$\Rightarrow \|T_h(u - P_{E_{h,k}} u)\|_Y^2 = \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \sigma_{h,\ell}^2 \langle \phi_{h,\ell}, u \rangle_X^2.$$

Now we use the upper bound (4.11) and Lemma 4.14 to obtain

$$\begin{aligned}
&\|T_h(I - P_{E_{h,k}})u\|_Y^2 \\
\leq &\sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \frac{\sigma_{h,\ell}^2}{|\sigma_k - \sigma_{h,\ell}|^2} \left(\left|\langle (T - T_h)\phi_{h,\ell}, v \rangle_Y\right| + \left|\langle \psi_{h,\ell}, (T - T_h)u \rangle_Y\right|\right)^2 \\
\leq &\, 2\sigma_k^2 q_k^2 \sum_{\ell \in \mathcal{J}_h \setminus I_{h,k}} \left(\left|\langle (T - T_h)\phi_{h,\ell}, v \rangle_Y\right| + \left|\langle \psi_{h,\ell}, (T - T_h)u \rangle_Y\right|\right)^2,
\end{aligned}$$

where the last inequality holds for all $h > 0$ sufficiently small. Proceeding as in the proof of Theorem 4.11, we obtain

$$\|T_h(I - P_{E_{h,k}})u\|_Y^2 \leq 4\sigma_k^2 q_k^2 \epsilon_h^2 \left(\|(I - P_{Y_h})v\|_Y^2 + \|(I - P_{X_h})u\|_X^2\right).$$

The desired result follows. ■

We can now give our main result on the convergence of the singular values.

Theorem 4.16 *Let $k \in \mathbb{Z}^+$ be given and suppose $T_h = P_{Y_h} T P_{X_h}$. For each $\ell \in I_k$,* *there exist $u \in E_k$, $v \in F_k$, and a constant $C_k > 0$ such that*

$$0 \le \frac{\sigma_k - \sigma_{h,\ell}}{\sigma_k} \le \frac{e_X^2}{2} + \frac{e_Y^2}{2} + C_k \epsilon_h \left(e_X + e_Y\right)^2 \;\; \forall\, h > 0 \textit{ sufficiently small,}$$

where $e_X = \|(I - P_{X_h})u\|_X$ and $e_Y = \|(I - P_{Y_h})v\|_Y$.

Proof: Let $\ell \in I_k$ be given, let $h > 0$ be sufficiently small that $\ell \in I_{h,k}$, and choose $\phi_{h,\ell} \in E_{h,k}$, $\psi_{h,\ell} \in F_{h,k}$ such that

$$\|\phi_{h,\ell}\|_X = \|\psi_{h,\ell}\|_Y = 1 \text{ and } T_h \phi_{h,\ell} = \sigma_{h,\ell} \psi_{h,\ell}.$$

We note that $\sigma_{h,\ell} \le \sigma_k$ by Theorem 4.6. We must derive an upper bound on $\sigma_k - \sigma_{h,\ell}$. For $h > 0$ sufficiently small, $\delta(E_{h,k}, E_k) = \delta(E_k, E_{h,k}) < 1$, which implies that $P_{E_{h,k}}$ defines a bijection from E_k onto $E_{h,k}$ (see the proof of Lemma 4.2). Thus there exists $u \in E_k$ such that $\|u\|_X = 1$ and $P_{E_{h,k}} u = \alpha \phi_{h,\ell}$ for some $\alpha \in (0, 1]$. Define $v \in F_k$ by $Tu = \sigma_k v$ and note that $\|v\|_Y = 1$. As in the statement of the theorem, we will write

$$e_X = \|(I - P_{X_h})u\|_X, \;\; e_Y = \|(I - P_{Y_h})v\|_Y.$$

We have

$$\begin{aligned}
\sigma_k &= \langle v, Tu \rangle_Y \\
&= \left\langle P_{F_{h,k}} v + (I - P_{F_{h,k}})v, T(P_{E_{h,k}} u + (I - P_{E_{h,k}})u) \right\rangle_Y \\
&= \left\langle P_{F_{h,k}} v, T P_{E_{h,k}} u \right\rangle_Y + \left\langle (I - P_{F_{h,k}})v, T P_{E_{h,k}} u \right\rangle_Y + \\
&\quad \left\langle P_{F_{h,k}} v, T(I - P_{E_{h,k}})u \right\rangle_Y + \left\langle (I - P_{F_{h,k}})v, T(I - P_{E_{h,k}})u \right\rangle_Y .
\end{aligned}$$

We now consider each of these four inner products. For the first, we have

$$\left\langle P_{F_{h,k}} v, T P_{E_{h,k}} u \right\rangle_Y = \alpha \left\langle P_{F_{h,k}} v, T_h \phi_{h,\ell} \right\rangle_Y = \alpha \sigma_{h,\ell} \left\langle P_{F_{h,k}} v, \psi_{h,\ell} \right\rangle_Y .$$

By the Pythagorean theorem and Taylor's theorem,

$$\begin{aligned}
\alpha = \|P_{E_{h,k}} u\|_X &= \sqrt{1 - \|(I - P_{E_{h,k}})u\|_X^2} \\
&= 1 - \frac{\|(I - P_{E_{h,k}})u\|_X^2}{2} + O(\|(I - P_{E_{h,k}})u\|_X^4) \\
&\le 1 - \frac{e_X^2}{2} + O(\|(I - P_{E_{h,k}})u\|_X^4) \\
&= 1 - \frac{e_X^2}{2} + O(e_X^4)
\end{aligned}$$

where the bound from Theorem 4.11 is used in the last step. Similarly,

$$\left\langle P_{F_{h,k}} v, \psi_{h,\ell} \right\rangle_Y \le \|P_{F_{h,k}} v\|_Y \le 1 - \frac{e_Y^2}{2} + O(e_Y^4).$$

It follows that

$$\begin{aligned}
\left\langle P_{F_{h,k}} v, T P_{E_{h,k}} u \right\rangle_Y &\le \sigma_{h,\ell} \left(1 - \frac{e_X^2}{2} + O(e_X^4)\right) \left(1 - \frac{e_Y^2}{2} + O(e_Y^4)\right) \\
&= \sigma_{h,\ell} - \left(\frac{e_X^2}{2} + \frac{e_Y^2}{2} + O(e_X^4 + e_Y^4)\right) \sigma_{h,\ell} \\
&= \sigma_{h,\ell} - \left(\frac{e_X^2}{2} + \frac{e_Y^2}{2} + O(e_X^4 + e_Y^4)\right) \sigma_k + \\
&\quad \left(\frac{e_X^2}{2} + \frac{e_Y^2}{2} + O(e_X^4 + e_Y^4)\right) (\sigma_k - \sigma_{h,\ell}) \\
&= \sigma_{h,\ell} - \left(\frac{e_X^2}{2} + \frac{e_Y^2}{2}\right) \sigma_k + O(\sigma_k \epsilon_h (e_X + e_Y)^2)
\end{aligned}$$

(using the fact that $\sigma_k - \sigma_{h,\ell} \le \epsilon_h$ and $e_X, e_Y = O(\epsilon_h)$ by Theorems 4.6 and 4.8). To bound the second inner product, notice that $\left\langle (I - P_{F_{h,k}})v, T_h P_{E_{h,k}} u \right\rangle_Y = 0$ because $(I - P_{F_{h,k}})v \in F_{h,k}^{\perp}$ and $T_h P_{E_{h,k}} u \in F_{h,k}$. Thus

$$\begin{aligned}
\left\langle (I - P_{F_{h,k}})v, T P_{E_{h,k}} u \right\rangle_Y &= \left\langle (I - P_{F_{h,k}})v, T P_{E_{h,k}} u - T_h P_{E_{h,k}} u \right\rangle_Y \\
&= \left\langle (I - P_{F_{h,k}})v, (I - P_{Y_h}) T P_{E_{h,k}} u \right\rangle_Y \\
&= \left\langle (I - P_{Y_h})v, (I - P_{Y_h}) T P_{E_{h,k}} u \right\rangle_Y \\
&\le e_Y \|(I - P_{Y_h}) T P_{E_{h,k}} u\|_Y \\
&\le e_Y \left(\|(I - P_{Y_h}) T u\|_Y + \|(I - P_{Y_h}) T (I - P_{E_{h,k}}) u\|_Y \right) \\
&\le e_Y \left(\sigma_k e_Y + \epsilon_h \|(I - P_{E_{h,k}}) u\|_X \right) \\
&\le \sigma_k e_Y \left(e_Y + \frac{\epsilon_h}{\sigma_k} \|(I - P_{E_{h,k}}) u\|_X \right) \\
&\le \sigma_k e_Y \left(e_Y + \epsilon_h q_k \|(I - P_{E_{h,k}}) u\|_X \right) \\
&= \sigma_k e_Y^2 + O\left(\sigma_k \epsilon_h (e_X + e_Y)^2 \right)
\end{aligned}$$

(using Theorem 4.11 and the fact that $\|T - P_{Y_h} T\|_{\mathcal{L}(X,Y)} \le \|T - T_h\|_{\mathcal{L}(X,Y)} = \epsilon_h$). We use similar reasoning to bound the third inner product as follows:

$$\left\langle P_{F_{h,k}} v, T (I - P_{E_{h,k}}) u \right\rangle_Y \le \sigma_k e_X^2 + O\left(\sigma_k \epsilon_h (e_X + e_Y)^2 \right).$$

Finally, for the fourth inner product, we use Theorem 4.15 to obtain

$$\begin{aligned}
&\left\langle (I-P_{F_{h,k}})v, T(I-P_{E_{h,k}})u\right\rangle_Y \\
=\ &\left\langle (I-P_{F_{h,k}})v, T_h(I-P_{E_{h,k}})u\right\rangle_Y + \left\langle (I-P_{F_{h,k}})v, (T-T_h)(I-P_{E_{h,k}})u\right\rangle_Y \\
\le\ &\|(I-P_{F_{h,k}})v\|_Y \|T_h(I-P_{E_{h,k}})u\|_Y + \\
&\qquad \|(I-P_{F_{h,k}})v\|_Y \|T-T_h\|_{\mathcal{L}(X,Y)} \|(I-P_{E_{h,k}})u\|_X \\
\le\ &2\sigma_k q_k \epsilon_h \|(I-P_{F_{h,k}})v\|_Y \left(e_Y+e_X\right) + \sigma_k \left(\frac{\epsilon_h}{\sigma_k}\right) \|(I-P_{F_{h,k}})v\|_Y \|(I-P_{E_{h,k}})u\|_X \\
=\ &O\left(\sigma_k\epsilon_h(e_X+e_Y)^2\right).
\end{aligned}$$

We thus obtain

$$\begin{aligned}
\sigma_k &\le \sigma_{h,\ell} - \sigma_k\left(\frac{e_X^2}{2}+\frac{e_Y^2}{2}\right) + \sigma_k e_X^2 + \sigma_k e_Y^2 + O\left(\sigma_k\epsilon_h(e_X+e_Y)^2\right) \\
&= \sigma_{h,\ell} + \sigma_k\left(\frac{e_X^2}{2}+\frac{e_Y^2}{2}\right) + O\left(\sigma_k\epsilon_h(e_X+e_Y)^2\right)
\end{aligned}$$

and hence

$$0 \le \frac{\sigma_k - \sigma_{h,\ell}}{\sigma_k} \le \frac{e_X^2}{2} + \frac{e_Y^2}{2} + O\left(\epsilon_h(e_X+e_Y)^2\right),$$

as desired. ■

Since $\|(I-P_{X_h})u\|_X$ and $\|(I-P_{Y_h})v\|_Y$ are both $O(\epsilon_h)$ by Theorem 4.8, Theorem 4.16 implies that the relative error in each computed singular value is $O(\epsilon_h^2)$ when $T_h = P_{Y_h}TP_{X_h}$ (as opposed to $O(\epsilon_h)$ for a general approximation T_h of T).

4.4 Numerical experiments

The first example demonstrates the convergence guaranteed by Theorems 4.6 and 4.8.

Example 4.17 *Consider the first-kind integral operator $T : L^2(0,1) \to L^2(0,1)$ defined by*

$$(Tx)(s) = \int_0^1 k(s,t)x(t)\,dt,\ 0<s<1,$$

where $k(s,t) = se^{st}$. We will use the techniques described in this chapter, with X_h and Y_h chosen to be finite-element spaces, to estimate the SVE of T. Since the kernel is smooth, Chebyshev approximation allows for a sequence of approximations that converge exponentially quickly (see [20] and [28]). The finite-element approximations described here do not lead to a competitive algorithm, but they serve to illustrate the convergence theorems of this chapter.

To apply the above results, we define $X_h = Y_h$ to be the space of continuous piecewise polynomial functions relative to the uniform mesh on $[0,1]$ with $1/h$ elements. We denote the nodes by $t_0, t_1, \ldots, t_n$ (where $t_j = j/n$ for each j) and use the standard nodal basis $\{x_0, x_1, \ldots, x_n\}$ (defined by $x_i(t_j) = \delta_{ij}$). Note that $n = d/h$ for piecewise polynomials of degree d. We discretize the integral operator by interpolating the kernel k onto the tensor-product finite element space

$$Y_h \otimes X_h = \mathrm{sp}\{z_{k\ell} \,:\, 0 \le k, \ell \le n\},$$

where $z_{k\ell}$ is defined by $z_{k\ell}(s,t) = x_k(s)x_\ell(t)$. We then define $T_h : X_h \to Y_h$ by

$$(T_h x)(s) = \int_0^1 k_h(s,t)x(t)\,dt, \ 0 < s < 1,$$

where k_h is the interpolated kernel:

$$k_h(s,t) = \sum_{k=0}^{n}\sum_{\ell=0}^{n} k(t_k, t_\ell)x_k(s)x_\ell(t)$$

Standard finite element approximation results can be used to show that, for piecewise polynomials of degree d,

$$\|k_h - k\|_{L^2((0,1)\times(0,1))} \le Ch^{d+1}.$$

It follows immediately that $\|T_h - T\|_{\mathcal{L}(L^2(0,1),L^2(0,1))} \le Ch^{d+1}$ and therefore we expect $\sigma_{h,k}$, $E_{h,k}$, and $F_{h,k}$ to converge with an error of $O(h^{d+1})$.

It is easy to show that the Galerkin matrix A is given by $A = MCM$, where M is the Gram matrix for the basis $\{x_0, x_1, \ldots, x_n\}$ and C is defined by $C_{k\ell} = k(t_k, t_\ell)$. It follows that it is simple to implement this particular discretization.

Table 4.1 shows the computed values of $\sigma_{h,1}$, $\sigma_{h,2}$, $\sigma_{h,3}$ for $h = 2^{-3}, 2^{-4}, \ldots, 2^{-10}$, using piecewise linear functions ($d = 1$). The exact values are unknown, but we use Richardson extrapolation to estimate an exponent p such that the error appears to converge to zero like $O(h^p)$. As expected, the results suggest that the errors are $O(h^2)$. Although we do not show the results, the same method suggests that the singular functions $\{\phi_{h,k}\}$ and $\{\psi_{h,k}\}$ converge at the same rate. (The singular spaces all appear to be one-dimensional.) Figure 4.1 shows the first three right and left singular functions.

The next experiment illustrates the accelerated convergence guaranteed by Theorem 4.16.

h	$\sigma_{h,1}(p_{\mathrm{est}})$	$\sigma_{h,2}(p_{\mathrm{est}})$	$\sigma_{h,3}(p_{\mathrm{est}})$
$1/8$	$8.95937 \cdot 10^{-1}$	$4.25077 \cdot 10^{-2}$	$1.19714 \cdot 10^{-3}$
$1/16$	$8.93464 \cdot 10^{-1}$	$4.26331 \cdot 10^{-2}$	$1.22087 \cdot 10^{-3}$
$1/32$	$8.92847 \cdot 10^{-1}(2.00)$	$4.26627 \cdot 10^{-2}(2.09)$	$1.22608 \cdot 10^{-3}(2.19)$
$1/64$	$8.92693 \cdot 10^{-1}(2.00)$	$4.26700 \cdot 10^{-2}(2.02)$	$1.22734 \cdot 10^{-3}(2.05)$
$1/128$	$8.92655 \cdot 10^{-1}(2.00)$	$4.26718 \cdot 10^{-2}(2.00)$	$1.22765 \cdot 10^{-3}(2.01)$
$1/256$	$8.92645 \cdot 10^{-1}(2.00)$	$4.26722 \cdot 10^{-2}(2.00)$	$1.22773 \cdot 10^{-3}(2.00)$
$1/512$	$8.92643 \cdot 10^{-1}(2.00)$	$4.26723 \cdot 10^{-2}(2.00)$	$1.22775 \cdot 10^{-3}(2.00)$
$1/1024$	$8.92642 \cdot 10^{-1}(2.00)$	$4.26724 \cdot 10^{-2}(2.00)$	$1.22775 \cdot 10^{-3}(2.00)$

Table 4.1

The computed singular values $\sigma_{h,1}, \sigma_{h,2}, \sigma_{h,3}$ for Example 4.17. Richardson extrapolation is used to estimate p such that $|\sigma_{h,k} - \sigma_k| = O(h^p)$ appears to hold; the estimated exponents p_{est} appear in parentheses.

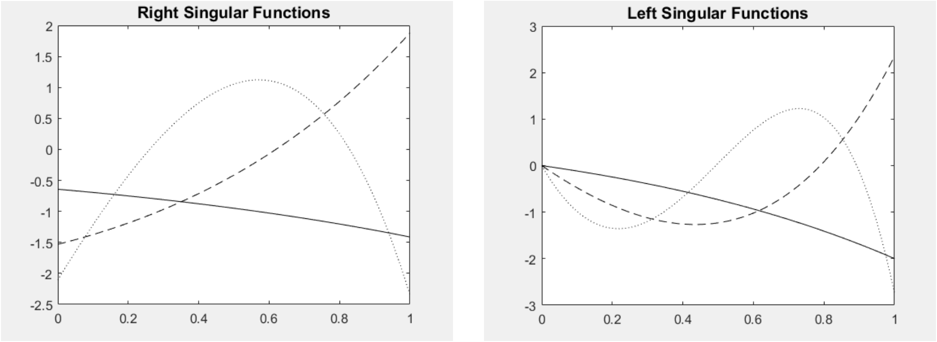

Figure 4.1: The computed singular functions $\phi_{h,1}, \phi_{h,2}, \phi_{h,3}$ (left) and $\psi_{h,1}, \psi_{h,2}, \psi_{h,3}$ (right). The solid curves represent $\phi_{h,1}$ and $\psi_{h,1}$, the dashed curves $\phi_{h,2}$ and $\psi_{h,2}$, and the dotted curves $\phi_{h,3}$ and $\psi_{h,3}$.

Example 4.18 *We now repeat Example 4.17 but using $T_h = P_{Y_h} T P_{X_h}$ to approximate T. We are required to compute the Galerkin matrix A defined by*

$$\begin{aligned} A_{k\ell} = \langle y_k, T_h x_\ell \rangle_{L^2(0,1)} &= \langle y_k, T x_\ell \rangle_{L^2(0,1)} \\ &= \int_0^1 \int_0^1 k(s,t) x_\ell(t) y_k(s) \, dt \, ds. \end{aligned}$$

We use a tensor-product Gauss quadrature rule to compute the entries of A to high accuracy. For an integral operator such as T, it is straightforward to show that $T_h = P_{Y_h} T P_{X_h}$ is the integral operator defined by the kernel $\hat{k}_h$, where $\hat{k}_h$ is the projection (in the L^2 inner product) of the true kernel k onto the tensor-product space $Y_h \otimes X_h$ (see [15]). It follows that $\|\hat{k}_h - k\|_{L^2((0,1)\times(0,1))}$ is no greater than $\|k_h - k\|_{L^2((0,1)\times(0,1))}$ (where k_h is the interpolated kernel used in Example 4.17), and numerical evidence *suggests that the asymptotic rate is the same:* $\|\hat{k}_h - k\|_{L^2((0,1)\times(0,1))} = O(h^{d+1})$ *if piecewise polynomials of degree d are used. By Theorem 4.16, then, we expect the computed singular values to converge at a rate of* $O((h^{d+1})^2) = O(h^{2d+2})$.

Table 4.2 shows the computed values of $\sigma_{h,1}, \sigma_{h,2}, \sigma_{h,3}$ *for* $h = 2^{-3}, 2^{-4}, \ldots, 2^{-10}$, *using piecewise linear functions* ($d = 1$). *As expected, the errors are consistent with* $O(h^4)$ *convergence. Table 4.3 shows the analogous results for piecewise quadratic polynomials* ($d = 2$). *In this case, the results are consistent with* $O(h^6)$ *convergence, as predicted by Theorem 4.16. (With piecewise quadratic functions, the singular values converge quickly enough that the errors reach the level of round-off error in our computations. This is reflected in the fact that the estimated exponent is not as close to* $2d + 2$ *as in the linear case, and even becomes negative in some cases.)*

h	$\sigma_{h,1}(p_{\text{est}})$	$\sigma_{h,2}(p_{\text{est}})$	$\sigma_{h,3}(p_{\text{est}})$
1/8	$8.92640 \cdot 10^{-1}$	$4.26673 \cdot 10^{-2}$	$1.22572 \cdot 10^{-3}$
1/16	$8.92642 \cdot 10^{-1}$	$4.26720 \cdot 10^{-2}$	$1.22763 \cdot 10^{-3}$
1/32	$8.92642 \cdot 10^{-1}(4.00)$	$4.26723 \cdot 10^{-2}(4.00)$	$1.22775 \cdot 10^{-3}(3.99)$
1/64	$8.92642 \cdot 10^{-1}(4.00)$	$4.26724 \cdot 10^{-2}(4.00)$	$1.22776 \cdot 10^{-3}(4.00)$
1/128	$8.92642 \cdot 10^{-1}(4.00)$	$4.26724 \cdot 10^{-2}(4.00)$	$1.22776 \cdot 10^{-3}(4.00)$
1/256	$8.92642 \cdot 10^{-1}(4.00)$	$4.26724 \cdot 10^{-2}(4.00)$	$1.22776 \cdot 10^{-3}(4.00)$
1/512	$8.92642 \cdot 10^{-1}(4.00)$	$4.26724 \cdot 10^{-2}(4.00)$	$1.22776 \cdot 10^{-3}(4.00)$
1/1024	$8.92642 \cdot 10^{-1}(4.00)$	$4.26724 \cdot 10^{-2}(4.00)$	$1.22776 \cdot 10^{-3}(4.00)$

Table 4.2

The computed singular values $\sigma_{h,1}, \sigma_{h,2}, \sigma_{h,3}$ for Example 4.18 (piecewise linear functions). Richardson extrapolation is used to estimate p such that $|\sigma_{h,k} - \sigma_k| = O(h^p)$ appears to hold; the estimated exponents p_{est} appear in parentheses.

With both piecewise linear and piecewise quadratic functions, the computed singular functions are consistent with $O(h^{d+1})$ *convergence; the increased rate of convergence applies only to the singular values.*

We also note that the first nine (that is, the nine largest) singular values of T *are approximately*

$$8.926 \cdot 10^{-1},\ 4.267 \cdot 10^{-2},\ 1.228 \cdot 10^{-3},\ 2.434 \cdot 10^{-5},\ 3.685 \cdot 10^{-7},$$
$$4.507 \cdot 10^{-9},\ 4.621 \cdot 10^{-11},\ 4.076 \cdot 10^{-13},\ 3.15 \cdot 10^{-15}.$$

Using piecewise linear functions and a uniform mesh with 256 elements, we are able to estimate these values accurately (error approximately 1% in σ_9 *and much less than 1% for* $\sigma_1, \ldots, \sigma_8$*) by computing the SVE of* $T_h = P_{Y_h} T P_{X_h}$. *Using the eigenvalue approach and the same mesh, it is possible only to estimate the first five singular values, with the error in* $\sigma_5 \approx 3.685 \cdot 10^{-7}$ *already about* 0.6%.

h	$\sigma_{h,1}(p_{\text{est}})$	$\sigma_{h,2}(p_{\text{est}})$	$\sigma_{h,3}(p_{\text{est}})$
$1/8$	$8.92642 \cdot 10^{-1}$	$4.26724 \cdot 10^{-2}$	$1.22775 \cdot 10^{-3}$
$1/16$	$8.92642 \cdot 10^{-1}$	$4.26724 \cdot 10^{-2}$	$1.22776 \cdot 10^{-3}$
$1/32$	$8.92642 \cdot 10^{-1}(5.93)$	$4.26724 \cdot 10^{-2}(5.91)$	$1.22776 \cdot 10^{-3}(5.86)$
$1/64$	$8.92642 \cdot 10^{-1}(5.98)$	$4.26724 \cdot 10^{-2}(5.95)$	$1.22776 \cdot 10^{-3}(5.93)$
$1/128$	$8.92642 \cdot 10^{-1}(6.27)$	$4.26724 \cdot 10^{-2}(5.98)$	$1.22776 \cdot 10^{-3}(5.96)$
$1/256$	$8.92642 \cdot 10^{-1}(1.91)$	$4.26724 \cdot 10^{-2}(7.38)$	$1.22776 \cdot 10^{-3}(6.00)$
$1/512$	$8.92642 \cdot 10^{-1}(-4.15)$	$4.26724 \cdot 10^{-2}(-2.39)$	$1.22776 \cdot 10^{-3}(4.79)$

Table 4.3

The computed singular values $\sigma_{h,1}, \sigma_{h,2}, \sigma_{h,3}$ for Example 4.18 (piecewise quadratic functions). Richardson extrapolation is used to estimate p such that $|\sigma_{h,k} - \sigma_k| = O(h^p)$ appears to hold; the estimated exponents p_{est} appear in parentheses. With piecewise quadratic approximations, the errors quickly reach the level of machine epsilon and so the estimates of p deteriorate as the mesh is refined.

Finally, as we have noted, Theorem 4.11 does not guarantee an optimal rate of convergence for the estimates of the singular vectors, at least not in all scenarios. The following example, in which we use different finite element spaces for X_h and Y_h, suggests that Theorem 4.11 correctly predicts the observed rate of convergence (optimal or suboptimal).

Example 4.19 *Let $T : L^2(0,1) \to L^2(0,1)$ be defined by*

$$(Tx)(s) = \int_0^1 k(s,t)x(t)\,dt,\ \ 0 < s < 1,$$

where k is the discontinuous kernel defined as follows:

$$k(s,t) = \begin{cases} s^2 - t, & s \le t, \\ se^{st}, & t < s. \end{cases}$$

We use continuous piecewise polynomials of degree $p-1$ and $q-1$ for X_h and Y_h, respectively, yielding the following rates of convergence:

$$\text{Right singular vectors: } \|(I - P_{X_h})u\|_X = O(h^p);$$
$$\text{Left singular vectors: } \|(I - P_{Y_h})v\|_Y = O(h^q).$$

Since k is discontinuous, $\epsilon_h = O(h)$. Theorems 4.11 and 4.16 suggest that we should *observe*

$$\begin{aligned}
\|(I-P_{E_{h,k}})u\|_X &= O\left(h^r\right),\ r=\min\{p,q+1\},\\
\|(I-P_{F_{h,k}})v\|_Y &= O\left(h^s\right),\ s=\min\{q,p+1\},\\
\frac{\sigma_k-\sigma_{h,k}}{\sigma_k} &= O\left(h^t\right),\ t=2\min\{r,s\}=2\min\{p,q\}.
\end{aligned}$$

Table 4.4 presents the results of our numerical experiments for different values of p *and* q*; the results are fully consistent with the predictions of Theorems 4.11 and 4.16. For illustration, we use* $k=1$*; specifically, we estimate* $u=\phi_1$ *and* $v=\psi_1$ *on uniform meshes with 16, 32, and 64 elements. We then use Richardson extrapolation to estimate* r*,* s*, and* t *so that*

$$\|(I-P_{E_{h,1}})u\|_X = O(h^r),\ \|(I-P_{F_{h,1}})v\|_Y = O(h^s),\ \textit{and}\ \frac{\sigma_k-\sigma_{h,k}}{\sigma_k} = O(h^t).$$

Although we do not show the results here, we observe the same behavior for various values of k*, except that, as* k *increases, finer meshes are needed to observe predicted rates of convergence.*

		$\|(I-P_{E_{h,1}})u\|_X$	$\|(I-P_{F_{h,1}})v\|_Y$	$\frac{\sigma_1-\sigma_{h,1}}{\sigma_1}$
p	q	observed r	observed s	observed t
2	2	2.0213	2.0057	4.0217
2	3	2.0201	2.9772	4.0450
3	2	2.8997	2.0061	4.0117
2	4	2.0201	**3.1194**	4.0395
4	2	**3.0198**	2.0061	4.0113
4	3	3.9531	2.9552	5.9103
5	2	**3.0194**	2.0061	4.0113
3	5	2.8427	**3.7315**	5.6847
5	3	**3.8940**	2.9552	5.9096
6	2	**3.0195**	2.0061	4.0113

Table 4.4

Observed rates of convergence of the first right and left singular vectors for different discretizations. In each case, the observed rate of convergence agrees with the prediction of Theorem 4.11. Suboptimal rates of convergence are indicated in boldface. The last column shows the rate of convergence for the first singular value; the results agree with Theorem 4.16.

4.5 SVE Computation

We now show how to compute the SVE of $T_h = P_{Y_h} T P_{X_h}$. It is clear that the right singular vectors of T_h belong to X_h and the left singular vectors to Y_h. Therefore, it suffices to show how to compute the SVE of an operator of the form $\hat{T} : \hat{X} \to \hat{Y}$, where $\hat{X}$ and $\hat{Y}$ are finite-dimensional subspaces of X and Y, respectively.

The following lemma will be useful.

Lemma 4.20 *Let $\{x_1, x_2, \ldots, x_n\}$ be a basis for a finite-dimensional inner product space $\hat{X}$, and let $M \in \mathbb{R}^{n \times n}$ be the Gram matrix for this basis. Then $\{\phi_1, \phi_2, \ldots, \phi_n\}$ is an orthonormal basis for $\hat{X}$ if and only if*

$$\phi_j = \sum_{i=1}^{n} U_{ij} x_i, \ \ j = 1, 2, \ldots, n, \tag{4.18}$$

where $M^{1/2}U$ is an orthogonal matrix.

Proof: Since $\{x_1, x_2, ..., x_n\}$ is a basis for X_h, any other basis $\{\phi_1, \phi_2, ..., \phi_n\}$ can be written in the form of (4.18) for some matrix $U \in \mathbb{R}^{n \times n}$. We then have

$$\begin{aligned} \langle \phi_i, \phi_j \rangle_X = \left\langle \sum_{k=1}^{n} U_{ki} x_k, \sum_{\ell=1}^{n} U_{\ell j} x_\ell \right\rangle_X &= \sum_{k=1}^{n} \sum_{\ell=1}^{n} U_{ki} U_{\ell j} \langle x_k, x_\ell \rangle_X \\ &= \sum_{k=1}^{n} \sum_{\ell=1}^{n} U_{ki} U_{\ell j} M_{k\ell} \\ &= \left(U^T M U \right)_{ij} = \left(\left(M^{1/2} U \right)^T \left(M^{1/2} U \right) \right)_{ij} \end{aligned}$$

It follows that $\{\phi_1, \phi_2, ..., \phi_n\}$ is orthonormal if and only if $M^{1/2}U$ is an orthogonal matrix. ■

We can now state the desired theorem.

Theorem 4.21 *Let $\hat{X}$ and $\hat{Y}$ be finite-dimensional subspaces of X and Y, respectively, and let $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ be bases for $\hat{X}$ and $\hat{Y}$, respectively. Suppose $\hat{T} : \hat{X} \to \hat{Y}$ is linear, and let $A \in \mathbb{R}^{m \times n}$ be the Galerkin matrix defined by*

$$A_{k\ell} = \left\langle y_k, \hat{T} x_\ell \right\rangle_Y .$$

Let $H^{-1/2} A M^{-1/2} = U \Sigma V^T$ be an SVD of the matrix $\hat{A} = H^{-1/2} A M^{-1/2}$, where M and H are the Gram matrices for $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$, respectively, *and define $\hat{V} = M^{-1/2}V$ and $\hat{U} = H^{-1/2}U$. Then an SVE of $\hat{T}$ is given by*

$$\hat{T} = \sum_{k=1}^{r} \hat{\sigma}_k \hat{\psi}_k \otimes \hat{\phi}_k, \tag{4.19}$$

where $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_r$ are the nonzero singular values of $\hat{A}$ and

$$\begin{aligned} \hat{\phi}_\ell &= \sum_{k=1}^{n} \hat{V}_{k\ell} x_k, \ \ell = 1, 2, \ldots, n, \\ \hat{\psi}_\ell &= \sum_{k=1}^{m} \hat{U}_{k\ell} y_k, \ \ell = 1, 2, \ldots, m. \end{aligned} \tag{4.20}$$

Proof: By Lemma 4.20, (4.20) defines orthonormal bases

$$\{\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_n\}, \ \{\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_m\}$$

of $\hat{X}$, $\hat{Y}$, respectively. It suffices to prove that (4.19) holds. We see that

$$\begin{aligned} & \hat{T}\left(\sum_{\ell=1}^{n} \alpha_\ell x_\ell\right) = \sum_{\ell=1}^{m} \beta_\ell y_\ell \\ \iff & \left\langle y_k, \hat{T}\left(\sum_{\ell=1}^{n} \alpha_\ell x_\ell\right)\right\rangle_Y = \left\langle y_k, \sum_{\ell=1}^{m} \beta_\ell y_\ell \right\rangle_Y, \ k = 1, 2, ..., m \\ \iff & \sum_{\ell=1}^{n} \langle y_k, \hat{T} x_\ell \rangle_X \alpha_\ell = \sum_{\ell=1}^{m} \langle y_k, y_\ell \rangle_Y \beta_\ell, \ k = 1, 2, ..., m \\ \iff & A\alpha = H\beta \\ \iff & \beta = H^{-1} A\alpha. \end{aligned}$$

We must show that

$$\left(\sum_{k=1}^{r} \hat{\sigma}_k \hat{\psi}_k \otimes \hat{\phi}_k\right)\left(\sum_{\ell=1}^{n} \alpha_\ell x_\ell\right) = \sum_{\ell=1}^{m} \beta_\ell y_\ell,$$

where $\beta = H^{-1}A\alpha$. This can be shown directly:

$$\begin{aligned}
\left(\sum_{k=1}^{r} \hat{\sigma}_k \hat{\psi}_k \otimes \hat{\phi}_k\right)\left(\sum_{\ell=1}^{n} \alpha_\ell x_\ell\right) &= \sum_{k=1}^{r}\sum_{\ell=1}^{n} \hat{\sigma}_k \langle \hat{\phi}_k, x_\ell\rangle_X \alpha_\ell \hat{\psi}_k \\
&= \sum_{k=1}^{r}\sum_{\ell=1}^{n} \left(\hat{\sigma}_k \left\langle \sum_{i=1}^{n} \hat{V}_{ik} x_i, x_\ell \right\rangle_X \alpha_\ell \sum_{q=1}^{m} \hat{U}_{qk} y_q\right) \\
&= \sum_{k=1}^{r}\sum_{\ell=1}^{n}\sum_{i=1}^{n}\sum_{q=1}^{m} \left(\hat{U}_{qk}\hat{V}_{ik}\hat{\sigma}_k \langle x_i, x_\ell\rangle_X \alpha_\ell y_q\right) \\
&\sum_{q=1}^{m}\left(\sum_{k=1}^{r}\sum_{\ell=1}^{n}\sum_{i=1}^{n} \hat{U}_{qk}\hat{V}_{ik}\hat{\sigma}_k M_{i\ell}\alpha_\ell\right) y_q \\
&\sum_{q=1}^{m}\left(\sum_{k=1}^{r}\sum_{i=1}^{n} \hat{U}_{qk}\hat{V}_{ik}\hat{\sigma}_k (M\alpha)_i\right) y_q \\
&\sum_{q=1}^{m}\left(\sum_{k=1}^{r} \hat{U}_{qk}\hat{\sigma}_k (\hat{V}^T M\alpha)_k\right) y_q \\
&\sum_{q=1}^{m}\left(\hat{U}\Sigma\hat{V}^T M\alpha\right)_q y_q.
\end{aligned}$$

This gives the desired result, because

$$\hat{U}\Sigma\hat{V}^T M = H^{-1/2}U\Sigma V^T M^{-1/2}M = H^{-1/2}H^{-1/2}AM^{-1/2}M^{-1/2}M = H^{-1}A.\blacksquare$$

It might be desirable to compute the SVE of $\hat{T} : \hat{X} \to \hat{Y}$, where $\hat{X}$ is a given finite-dimensional subspace of X, $\hat{Y} = T(\hat{X})$, and $\hat{T}x = Tx$ for all $x \in \hat{X}$. We suppose first that $\{x_1, x_2, \ldots, x_n\}$ is a basis for $\hat{X}$ and that $\{y_1, y_2, \ldots, y_n\}$, where $y_i = Tx_i$ for $i = 1, 2, \ldots, n$, is linearly independent and hence is a basis for $\hat{Y} = T(X_h)$. In this case, the Galerkin matrix A is defined by

$$A_{ij} = \langle y_i, Tx_j\rangle_Y = \langle Tx_i, Tx_j\rangle_Y$$

and coincides with the Gram matrix for the basis $\{y_1, y_2, \ldots, y_n\}$. Since A is likely to be dense (even if X_h is a finite-element space), the square root $A^{1/2}$ required by Theorem 4.21 is expensive to compute when n is large. For this reason, we use the Cholesky factorization instead (as, indeed, we could have done in Theorem 4.21). Let $A = LL^T$ and $M = NN^T$ be Cholesky factorizations of A and M, respectively (so that L and N are lower triangular). We then define $\hat{A} = L^{-1}AN^{-T} = L^TN^{-T}$ and compute an SVD of $\hat{A}$: $L^TN^{-T} = U\Sigma V^T$. Defining $\hat{V} = N^{-T}V$ and $\hat{U} = L^{-T}U$, it is easy to show that

$$\hat{T} = \sum_{k=1}^{r} \hat{\sigma}_k \hat{\psi}_k \otimes \hat{\phi}_k,$$

where $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_r$ are the nonzero singular values of $\hat{A}$ and

$$\hat{\phi}_\ell = \sum_{k=1}^{n} \hat{V}_{k\ell} x_k, \ \ell = 1, 2, \ldots, n,$$

$$\hat{\psi}_\ell = \sum_{k=1}^{n} \hat{U}_{k\ell} y_k, \ \ell = 1, 2, \ldots, n.$$

The above scheme is likely to be effective if $\mathcal{N}(T) \cap \hat{X}$ is trivial and the singular vectors of T go to zero slowly enough to allow a sufficiently fine discretization of X for accurate approximation of the left singular vectors while maintaining the positive definiteness of A. For many operators T, though, the Galerkin matrix A will be singular for a reasonable discretization $\hat{X}$ of X. In this case, a more careful algorithm is needed.

4.6 Conclusions

The singular values and singular vectors of a compact operator can be estimated using a variety of discretizations. In the generic case, where the operator T is approximated by a family $\{T_h : h > 0\}$ of discretized operators, the errors in both singular values and singular vectors go to zero at least as fast as does $\|T_h - T\|_{\mathcal{L}(X,Y)}$. With the special choice of $T_h = P_{Y_h} T P_{X_h}$, the error in the computed singular vectors can be expressed in terms of the optimal approximation errors, but the approximability of both left and right singular vectors affects the error in either the computed left singular vectors or the computed right singular vectors. In many cases, the error in the computed singular vector is asymptotically optimal, but a poor degree of approximability in the left singular vectors, for example, can cause the error in the computed right singular vectors to be suboptimal. This is suggested by the bounds in Theorem 4.11 and confirmed by the numerical examples in Section 4.4. Still considering the case of $T_h = P_{Y_h} T P_{X_h}$, Theorem 4.16 shows that the computed singular values converge at an increased rate. The typical case is that the error goes to zero like the square of the optimal approximation error for the singular vectors, but once again the approximability of both left and right singular vectors must be taken into account.

Although the approximation of the singular values and singular vectors of T is related to the approximation of the eigenvalues and eigenvectors of T^*T, the fact that both left and right singular vectors (which may have different degrees of approximability) are involved means that the situation cannot be understood by simply transferring the results of the eigenvalue theory (due to Chatelin, Babuška and Osborn, and others) to the singular value problem. Moreover, as discussed in Section 4.1, there are advantages to computing the singular values and singular vectors directly, rather than by formulating a related eigenvalue problem, particularly the fact that we can approximate smaller singular values accurately.

References

[1] Philip Anselone. *Collectively Compact Operator Approximation Theory and Applications to Integral Equations.* Prentice-Hall, Upper Saddle River, 1971.

[2] I. Babuška and J. Osborn. Finite element Galerkin approximation of the eigenvalues and eigenvectors of self-adjoint problems. *Mathematics of Computation*, 52(186):275–297, 1989.

[3] I. Babuška and J. Osborn. Eigenvalue problems. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis. Volume 2: Finite element methods: part 1*, pages 641–787. Elsevier, Amsterdam, 1991.

[4] Daniele Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010.

[5] Françoise Chatelin. *Spectral Approximation of Linear Operators.* Society for Industrial and Applied Mathematics, Philadelphia, 2011.

[6] Daniel Crane, Mark Gockenbach, and Matthew Roberts. Approximating the Singular Value Expansion of a Compact Operator. *SIAM Journal on Numerical Analysis*, In Press, 2020.

[7] Gerald Folland. *Real Analysis: Modern Techniques and their Applications.* John Wiley & Sons, Hoboken, 1984.

[8] Mark Gockenbach. A singular value expansion for arbitrary bounded linear operators. *Unplublished.*

[9] Mark Gockenbach. *Finite Dimensional Linear Algebra.* CRC Press, Boca Raton, 2010.

[10] Mark Gockenbach. *Linear Inverse Problems and Tikhonov Regularization.* MAA Press, Washington D.C., 2016.

[11] Mark Gockenbach and Matthew Roberts. Approximating the generalized singular value expansion. *SIAM Journal on Numerical Analysis*, 56(5):2776–2795, 2018.

[12] Israel Gohberg, Seymor Goldberg, and Marinus Kaashoek. *Classes of Linear Operators*, volume 1. Springer, New York City, 1990.

[13] Gene Golub and Charles Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 4th edition, 2013.

[14] C. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*. Pitman Advanced Publishing Program, Boston, 1984.

[15] P.C. Hansen. Computation of the singular value expansion. *Computing*, 40:185–199, 1988.

[16] Frederic Helein. Spectral Theory. *Paris Diderot University*, 2014.

[17] Einar Hille and Jacob Tamarkin. On the characteristic values of linear integral equations. *Acta Math.*, 57:1–76, 1931.

[18] Peter D. Lax. *Functional Analysis*. John Wiley & Sons, Hoboken, 2002.

[19] David Lay. *Linear Algebra and its Applications*. Pearson, London, 4th edition, 2012.

[20] G. Little and J. B. Reade. Eigenvalues of analytic kernels. *SIAM Journal on Mathematical Analysis*, 15(1):133–136, 1984.

[21] James Mercer. XVI. Functions of positive and negative type, and their connection the theory of integral equations. *Royal Society*, 209.441-458:415–446, 1909.

[22] Albrecht Pietch. *History of Banach Spaces and Linear Operators*. Birkhäuser, Boston, 2007.

[23] H. Royden. *Real Analysis*. MacMillan Publishing CO., Inc., New York City, 2nd edition, 1968.

[24] Walter Rudin. *Real and Complex Analysis*. Tata McGraw-Hill education, New York City, 2006.

[25] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. i teil. entwicklung willkiirlichen funktionen nach system vorgeschriebener. *Math. Ann.*, 63:433–476, 1907.

[26] Frank Smithies. The eigen-values and singular values of integral equations. *Proc. London Math. Soc.*, 43:255–279, 1937.

[27] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566, 1993.

[28] Alex Townsend and Lloyd N. Trefethen. Continuous analoques of matrix factorizations. *Proc. R. Soc. A*, 471.2173:20140585, 2015.

[29] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Math. Annalem*, 71:441–479, 1912.