COMPUTING SPECTRAL PROPERTIES OF INFINITE-DIMENSIONAL OPERATORS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Andrew James Horning

December 2021

COMPUTING SPECTRAL PROPERTIES OF INFINITE-DIMENSIONAL OPERATORS

Andrew James Horning, Ph.D.

Cornell University 2021

This dissertation introduces a cohesive framework for numerically computing spectral properties related to the discrete and continuous spectrum of infinite-dimensional operators. Approximations to eigenvalues and eigenvectors, spectral measures, and generalized eigenvectors are constructed by sampling the range of the resolvent operator at strategic points in the complex plane. These algorithms are developed and analyzed directly in the abstract infinite-dimensional Hilbert space setting. They require only two essential computational ingredients: (1) solving linear equations with complex shifts and (2) taking inner products in the Hilbert space. Numerical implementations for a broad class of differential and integral operators, leveraging state-of-the-art adaptive spectral methods, are provided in an accompanying MATLAB package called `SpecSolve`, which is demonstrated through a collection of examples.

BIOGRAPHICAL SKETCH

Andrew developed an interest in science and engineering while working as a technician in a small biomedical metallurgy facility in Glens Falls, New York, between 2010 and 2012. He earned an associate's degree in engineering science at Hudson Valley Community College in 2014 and a bachelor's degree in physics and mathematics at Rensselaer Polytechnic Institute in 2016. He conducted his dissertation research at the Center for Applied Mathematics at Cornell University between 2016 and 2021. He will begin postdoctoral studies as an Applied Math Instructor at the Massachusetts Institute of Technology in September 2021.

For my parents, Richard and Cheryl Horning.

ACKNOWLEDGEMENTS

When I arrived at Cornell five years ago, I was required to take a course in numerical analysis. I began with little interest in the subject, but the lecturer injected such curiosity and passion into each class! That's how I met Professor Alex Townsend, my adviser. I want to thank Alex for guiding me with the same boundless enthusiasm, keen insight, and genuine concern that characterize his classroom. He has modeled the highest standards in research, teaching, and mentorship, and given me a strong foothold in the world of computational math research. I will be unpacking what I have learned from him for years to come.

It has been a real pleasure to participate in applied math at Cornell. I will miss my fellow CAMsters in the Center for Applied Math and that rich, collaborative atmosphere that they cultivate year after year on the 6th floor of Rhodes Hall. I thank Professors Tim Healey, John Hubbard, Steve Strogatz, Bob Strichartz, and Alex Vladimirsky for bringing humor, wonder, and an inquisitive spirit to my graduate courses. And I am indebted to my committee members Professors David Bindel and Anil Damle for the nutrient-dense "food-for-thought" they distributed generously among their students. I am especially grateful to Erika Fowler-Decatur for the tireless advocacy and administrative excellence that kept CAM running so smoothly. My graduate years could not have been the same without Marc Aurèle Gilles, Heather Wilber, Tianyi Shi, Dan Fortunato, and Nicolas Boullé. It is wonderful to be part of a vibrant young research group and I'm eager to see where these five will shine next.

A thrilling aspect of graduate research is the opportunity to work with colleagues around the globe. It has been a privilege to work with Professor Yuji Nakatsukasa during the last year of my Ph.D. (see Chapter 3). His enthusiasm for numerical linear algebra is infectious! I have also been fortunate to collabo-

rate with Dr. Matt Colbrook at Cambridge University (see Chapter 4). The hours spent laboring over coffee and code have paid off a hundredfold in camaraderie and mathematical insight. I am grateful to Dr. Marcus Webb at the University of Manchester for his advocacy and words of advice. I thank Dr. Rich Lehoucq and Professors Mark Embree, Lin Lin, Vanni Noferini, Sheehan Olver, Mikaël Slevinsky, Yousef Saad, and Nick Trefethen for stimulating conversations and correspondence that have shaped this dissertation for the better.

In the summer of 2018, Dr. Rhonda Morgan invited me to spend 10 weeks at the Jet Propulsion Laboratory in Pasadena, California. It was a summer of 'firsts' for me, and I deeply appreciate Rhonda's mentorship through it all. Alongside Rhonda, I thank Drs. Mike Turmon, Jeff Jewell, Stuart Shaklan, and Eric Nielson for their encouragement, constructive criticism, and the best 10 week crash course in real-world computational science that one could hope for.

This work could not have succeeded without the support and love of my parents, Richard and Cheryl, my brother, Sean, and my wife, Ana. I also want to thank my Durkeetown church family and my brothers-by-faith – Robert Arnold, Chris and Nick Miller, and Zachary Prater – for providing a haven when I needed rest from research.

David and Sharon Covington once told me that a good graduate education can feel like trying to drink from a fire hose. Without the steadfast instruction of kind-hearted mentors, colleagues, and friends, I might have washed away in the torrent. I am grateful that, instead, they taught me how to drink deeply.

CONTENTS

CHAPTER 1
INTRODUCTION

In this thesis, we introduce new algorithms to compute spectral properties of infinite-dimensional operators, including eigenvalues and eigenfunctions associated with the discrete spectrum, generalized eigenfunctions related to the continuous spectrum, and spectral measures. The algorithms are formulated and analyzed in a Hilbert space setting, providing broadly applicable templates for spectral computations in infinite-dimensional spaces. A collection of practical MATLAB implementations for differential and integral operators, which leverage high-precision spectral methods and fast linear algebra, are available in the public repository `SpecSolve` on GitHub [33].

Chapter 2 is focused on the classical problem of computing eigenvalues and eigenfunctions. We propose an operator analogue of the FEAST matrix eigensolver [89, 116, 144] to compute discrete portions of the spectrum of a closed operator in a target region in the complex plane. If the target eigenvalues are well-conditioned, the operator analogue respects this and efficiently computes them to near machine precision accuracy. The algorithm is particularly adept at computing high-frequency modes of differential operators that possess self-adjoint structure with respect to weighted Hilbert spaces.

Chapter 3 examines the stability of the numerical linear algebra routines that underpin the algorithm of Chapter 2, i.e., FEAST and other contour integral eigensolvers that target interior eigenvalues with rational filters. Remarkably, subspace iteration with a rational filter is shown to be robust even when an eigenvalue is near a filter's pole. These "dangerous eigenvalues" contribute to large round-off errors in the first iteration but are self-correcting in later itera-

tions. For normal matrices, two iterations are enough to reduce round-off errors to the order of machine precision. In contrast, Krylov methods accelerated by rational filters with fixed poles typically fail to converge when an eigenvalue is close to a pole. We conclude with a simple restart strategy that recovers full precision in the target eigenpairs for Arnoldi with shift-and-invert enhancement.

Chapters 4 and 5 develop rigorous computational tools for spectral measures and generalized eigenfunctions of self-adjoint operators. Inspired by density-of-states algorithms in physics and density estimation techniques in statistics, we construct smoothed approximations of spectral measures and develop a practical framework for sampling from them in Chapter 4. The resulting algorithms exploit regularity to accelerate convergence in the smoothing parameter. This feature allows us to compute spectral properties to high accuracy and avoid solving ill-conditioned linear systems, as demonstrated through careful numerical experiments. Chapter 5 extends these techniques to compute generalized eigenfunctions in a rigged Hilbert space setting. We introduce wave-packet approximations, analyze convergence in the natural dual topology, and apply the method to study scattering modes in a 2-dimensional quantum waveguide.

The unifying theme of our work is the central role of the resolvent operator: each algorithm we develop constructs spectral properties by sampling the range of the resolvent operator at carefully selected points in the complex plane. We believe this framework is powerful and timely for three reasons. First, the main computational task boils down to solving operator equations with complex shifts, for which we leverage cutting-edge technology in sparse spectral discretizations and adaptive infinite-dimensional linear algebra. Second, the number of linear equations needed for accurate results is controlled by care-

fully selecting complex shifts, a process naturally informed by the vibrant research field of rational approximation. Third, a flurry of recent developments in computational spectral theory reveal that the resolvent provides a robust way to access spectral properties of infinite-dimensional operators.

Spectral theory has developed a rich and powerful repertoire of techniques for analyzing infinite-dimensional operators over the past century. This thesis aims to encapsulate these techniques in publicly available software so that applied mathematicians and computational scientists can probe the spectrum of infinite-dimensional operators using robust, efficient, and user-friendly computational tools.

1.1 The eigenvalue problem

Given a linear operator $\mathcal{L}$ with domain $\mathcal{D}(\mathcal{L})$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{L}$ with associated eigenvector $u \in \mathcal{D}(\mathcal{L})$ ($u \neq 0$) if they satisfy [87, p. 172]

$$\mathcal{L}u = \lambda u, \qquad u \in \mathcal{D}(\mathcal{L}). \tag{1.1}$$

Linear operators may act on finite-dimensional spaces or infinite-dimensional spaces. For example, when $\mathcal{D}(\mathcal{L})$ is finite-dimensional, (1.1) reduces to the classical matrix eigenvalue problem. Throughout this thesis, we focus on differential and integral operators acting on infinite-dimensional spaces of functions.

The function spaces we are interested in are Hilbert spaces,[1] a natural setting for computational spectral theory. Given a Hilbert space $\mathcal{H}$, we denote the inner product by $(\cdot,\cdot)_{\mathcal{H}}$ and the norm by $\|\cdot\|_{\mathcal{H}} = \sqrt{(\cdot,\cdot)_{\mathcal{H}}}$. When the Hilbert space is

[1] The Hilbert spaces we encounter are separable, with a countable dense subset [122, p. 69].

clear from the context, we omit the subscript $\mathcal{H}$. The inner product and norm play key roles in both algorithms and analyses because they allow us to calculate orthogonal projections numerically and quantify approximation errors.

Linear operators in infinite dimensions may be unbounded: $\mathcal{D}(\mathcal{L})$ need not be all of $\mathcal{H}$. We consider operators whose domain $\mathcal{D}(\mathcal{L})$ is dense in $\mathcal{H}$ and whose graph is closed in $\mathcal{H} \times \mathcal{H}$.[2] We say that $\mathcal{L}$ is closed and densely defined in $\mathcal{H}$. These two technical conditions allow us to treat important classes of unbounded operators, such as differential operators and singular integral operators, while leveraging key analytical tools from spectral theory [87, Ch. III § 5-6].

Note that a bounded operator $\mathcal{B}$ on $\mathcal{H}$ is closed with operator norm [87, p. 164]

$$\|\mathcal{B}\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \|\mathcal{B}f\|_{\mathcal{H}} < \infty.$$

Conversely, a closed operator with $\mathcal{D}(\mathcal{B}) = \mathcal{H}$ is always bounded [87, Thm. 5.20]. In what follows, all operators encountered are linear, closed, and densely defined unless designated otherwise. We reserve $\mathcal{B}$ for bounded operators.

1.2 Why compute eigenvalues?

Trefethen and Embree [157] identify three aspects of spectral analysis that make eigenvalues and eigenvectors so extraordinarily useful:

(1) When eigenvectors diagonalize operators, they can reduce complex problems to a batch of much simpler scalar problems involving eigenvalues.

[2] The graph of $\mathcal{L}$ is a linear manifold composed of pairs $(u, \mathcal{L}u)$, for each $u \in \mathcal{D}(\mathcal{L})$ [87, p. 165].

(2) Eigenvalues can provide insight into resonance phenomena in linear systems, answering the question, "which inputs can produce a heightened response?"
(3) Eigenvalues often predict the asymptotic behavior of systems that evolve in time. They answer, "what is the dominant response to a general input?"

While these attributes are not exhaustive, the appearance of eigenvalues in dynamical systems, stochastic processes, optimization, statistics, and countless applications is often inextricably linked with (1)-(3).

Linear stability analysis. To illustrate how (1)-(3) often work together in the analysis of a physical model, consider the initial boundary value problem (IBVP) with periodic boundary conditions

$$u_t = \mathcal{L}u + \mathcal{N}(u), \qquad u_t(x,0) = g(x), \qquad u(-1,t) = u(1,t). \tag{1.2}$$

Here, $\mathcal{L}$ and $\mathcal{N}$ are linear and nonlinear ordinary differential operators (with respect to the variable x), respectively. In many instances, (1.2) supports steady-states, traveling wave states, or other phenomena whose stability is of critical importance in the physical problem under study [4, 95, 128]. When $\mathcal{L}$ is diagonalized by an orthogonal set of eigenvectors (see section 1.4 for a precise definition), the stability analysis often reduces to determining whether or not the eigenvalues of $\mathcal{L}$ are contained in one half-plane [4, 94, 128, 157].

In a typical example, liquid drops forming on a rigid substrate obey a particular form of (1.2) known as the Cahn–Hilliard equation [94]. Small environmental perturbations to these droplets can be decomposed into the eigenmodes of a fourth-order differential operator. Whether the drops persist after perturbation is determined by the associated eigenvalues [95]. Eigenvalues in the right half-plane correspond to unstable modes, i.e., perturbations which cause large

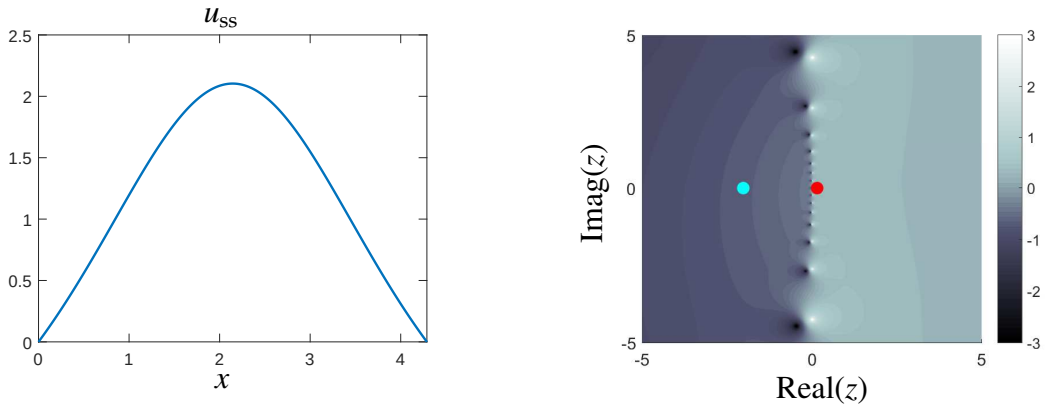

Figure 1.1: *Stability of liquid droplets.* A liquid droplet's radial profile is modeled by the curve $u_{ss}(x)$ (left panel), and the droplet is determined to be unstable due to the presence of an eigenvalue of a fourth-order differential operator in the right half-plane (right panel, red dot) [83, 95]. The color map in the right-hand panel corresponds to a rational filter used to target unstable modes in the right half-plane and filter out stable modes (right panel, blue dot) in the left half-plane (see section 2.6.2 of Chapter 2).

changes to the droplet profile over time. Figure 1.1 depicts the radial profile of a symmetric droplet (left panel) and the eigenvalues that determine its stability (right panel). We return to this problem in section 2.6.2 of Chapter 2.

1.3 The spectrum of a linear operator

The spectrum of $\mathcal{L}$ is defined through singular points of the resolvent [157, p. 28][3]

$$(\mathcal{L} - z)^{-1} : \mathcal{H} \to \mathcal{D}(\mathcal{L}), \qquad z \in \Theta(\mathcal{L}). \tag{1.3}$$

The resolvent function in (1.3) is a bounded inverse for all complex numbers in the resolvent set $\Theta(\mathcal{L}) = \{z \in \mathbb{C} : \|(\mathcal{L}-z)^{-1}\|_{\mathcal{H}} < \infty\}$ [87, p. 173]. The spectrum of $\mathcal{L}$ is the complement $\Lambda(\mathcal{L}) = \mathbb{C} \setminus \Theta(\mathcal{L})$ and it is convenient to adopt the convention

[3]We highlight the similarities between finite-dimensional and infinite-dimensional eigensolvers in Chapters 2 and 3 by adopting conventional linear algebra notation therein and referring to $(z\mathcal{I} - \mathcal{L})^{-1}$ as the resolvent, where $\mathcal{I}$ is the identity on $\mathcal{H}$.

that $\|(\mathcal{L} - z)^{-1}\|_{\mathcal{H}} = \infty$ when $z \in \Lambda(\mathcal{L})$ [157, p. 29].

While the spectrum of a matrix consists precisely of its eigenvalues, $\Lambda(\mathcal{L})$ can be empty or contain more exotic spectral types (e.g., see Figure 1.2) [41, p. 17]. This is because $(\mathcal{L}-z)^{-1}$ may fail to exist as a bounded operator in ways that have no finite-dimensional analogue. It is helpful to classify elements in the spectrum accordingly. Before proceeding, we note that the closed operator $\mathcal{L} - z$ has a bounded inverse if and only if it maps $\mathcal{D}(\mathcal{L})$ one-to-one and onto $\mathcal{H}$ [87, p. 167].

Point spectrum. A complex number $\lambda \in \Lambda(\mathcal{L})$ is in the *point spectrum*, denoted $\Lambda_{\mathrm{p}}(\mathcal{L})$, if $\mathcal{L} - \lambda$ is not one-to-one. In this case, $\mathcal{L} - \lambda$ has a nontrivial null space: λ is an eigenvalue and any vector in the null space is an associated eigenvector satisfying (1.1). The dimension of the null space is called the geometric multiplicity of λ [87, p. 173], which may be finite or infinite.

Continuous spectrum. A complex number $\lambda \in \Lambda(\mathcal{L})$ is in the *continuous spectrum*, denoted $\Lambda_{\mathrm{c}}(\mathcal{L})$, if $\mathcal{L} - \lambda$ is one-to-one while its range is dense in, rather than onto, $\mathcal{H}$. In this case, there is no eigenfunction $u \in \mathcal{H}$ associated with λ. However, $\|\mathcal{L} - \lambda\|_{\mathcal{H}}$ is unbounded below and there is always a sequence of functions $f_1, f_2, f_3, \ldots \in \mathcal{H}$ such that $\|(\mathcal{L} - \lambda) f_n\|_{\mathcal{H}} \to 0$ as $n \to \infty$ [41, p. 17].[4]

Residual spectrum. An operator may also have *residual spectrum*, consisting of points $\lambda \in \Lambda(\mathcal{L})$ where $\mathcal{L} - z$ is one-to-one, but does not have a dense range in $\mathcal{H}$ (and, therefore, is not onto). This phenomenon may arise, for instance, when the domain of a differential operator is specified inappropriately by imposing too many boundary conditions on the eigenvalue problem [157, p. 29]. We do not consider operators with residual spectrum in the remainder of this thesis.

[4] In general, this sequence does not converge in $\mathcal{H}$, but it may converge in a weaker topology. This idea is intimately connected with generalized eigenvectors (see section 1.4.2).

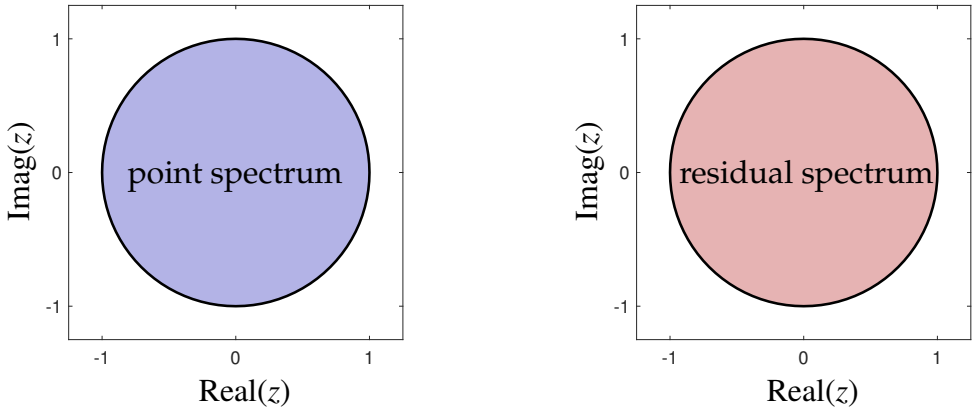

Figure 1.2: *Spectra of shift operators.* The left shift operator $\mathcal{B} : (a_1, a_2, a_3, \ldots) \to (a_2, a_3, a_4, \ldots)$ is bounded on $l^2(\mathbb{N})$: its point spectrum (blue) fills the interior of the unit disk, i.e., eigenvalues λ with eigenvectors $(1, \lambda, \lambda^2, \ldots)$. The right shift operator $\mathcal{B} : (a_1, a_2, a_3, \ldots) \to (0, a_1, a_2, \ldots)$ is also bounded on $l^2(\mathbb{N})$: it has no eigenvectors but residual spectrum (red) fills the interior of the unit disk. The continuous spectrum of both operators fills the unit circle, see [157, p. 29].

An important subset of the point spectrum is the *discrete spectrum*, the collection of isolated eigenvalues of $\mathcal{L}$ with finite multiplicity in the following sense [87, p. 181]. If Γ is a rectifiable, simple closed curve enclosing one isolated eigenvalue $\lambda \in \Lambda(\mathcal{L})$, then the spectral projector associated with λ is [87, p. 178]

$$\mathcal{P}_\lambda = -\frac{1}{2\pi i} \int_\Gamma (\mathcal{L} - z)^{-1} \, dz. \tag{1.4}$$

The *multiplicity* of λ is the rank of $\mathcal{P}_\lambda$, i.e., the dimension of its range in $\mathcal{H}$. The structure of the discrete spectrum and its sensitivity to bounded perturbations is comparable to the spectrum of a matrix [87, Ch. III § 6.5]. Two fundamental classes of infinite-dimensional operators have a purely discrete spectrum.

Compact operators. An operator $\mathcal{K}$ is called compact if the closure of the set $\{\mathcal{K}f : \|f\|_\mathcal{H} = 1\}$ is compact in $\mathcal{H}$ [122, p. 103]. Many integral operators encountered in application are compact. For example, given a domain $\Omega \subset \mathbb{R}^d$ and a kernel $K(x, y)$ that is square-integrable on $\Omega \times \Omega$, the integral operator

$$[\mathcal{K}f](x) = \int_\Omega K(x, y) \, f(y) \, dy,$$

is compact on $L^2(\Omega)$, the space of square-integrable functions on Ω [135, Ch. 4.6].

Theorem 1.3.1. *Given a Hilbert space $\mathcal{H}$, let $\mathcal{K} : \mathcal{H} \to \mathcal{H}$ be compact. Then, $\Lambda(\mathcal{K})$ is a countable set with no limit point other than zero and every nonzero $\lambda \in \Lambda(\mathcal{K})$ is an eigenvalue of $\mathcal{K}$ with finite multiplicity.*

Proof. See the proof of [87, Ch. III Thm. 6.26]. □

Operators with compact resolvent. An operator $\mathcal{L}$ has a compact resolvent if $(\mathcal{L} - z)^{-1}$ is compact for at least one $z \in \Theta(\mathcal{L})$ [87, Ch. III § 6.8]. Most differential operators in classical boundary value problems have a compact resolvent [87, p. 187]. For example, the resolvent of a second-order, uniformly elliptic partial differential operator on a bounded domain $\Omega \subset \mathbb{R}^3$ is a compact integral operator on $L^2(\Omega)$, the space of square-integrable functions, whose kernel is called the Green's function [27, Ch. 6].[5]

Theorem 1.3.2. *Given a Hilbert space $\mathcal{H}$, let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ have compact resolvent. Then, $\Lambda(\mathcal{K})$ comprises a countable set of isolated eigenvalues with finite multiplicity, having no finite limit point, and $(\mathcal{L} - z)^{-1}$ is compact for every $z \in \Theta(\mathcal{L})$.*

Proof. See the proof of [87, Ch. III Thm. 6.29]. □

Although many differential and integral operators have purely discrete spectra, operators with continuous spectrum also occupy an important place in applied mathematics. Operators with non-empty continuous spectra include self-adjoint Toeplitz operators on $\ell^2(\mathbb{N})$ (square summable sequences, where $\mathbb{N} = \{1, 2, ...\}$) [20]; differential operators on unbounded domains [150,

[5] Strictly speaking, the domain and any variable coefficients of the operator must also satisfy mild regularity requirements, see [27, Ch. 6] or [54, Ch. 6].

Ch. V] [48, Ch. XIII,Ch. XIV] or on bounded domains with singular variable coefficients [63, 97]; and integral perturbations of multiplication operators and Cauchy-type integral operators [56, 92]. In physical systems that scatter or radiate energy, the associated operator typically has a mix of continuous and discrete spectra; see the RAGE theorem for a mathematical formulation [5, 53, 123].

1.4 Diagonalization in infinite dimensions

In a finite-dimensional space, eigenvalue analysis is particularly effective and natural for matrices diagonalized by an orthogonal basis of eigenvectors [157, § 1.1]. These are the normal matrices, which include real symmetric, complex Hermitian, skew symmetric, and unitary matrices. A simple characterization of normal matrices is that they commute with their adjoint.[6] Real symmetric and complex Hermitian matrices are special among normal matrices because they are also their own adjoint. The rest of this section addresses their infinite-dimensional analogue:

What does it mean to diagonalize a self-adjoint operator?[7]

It is helpful to start with the diagonalization of an $n \times n$ Hermitian matrix A. In this case, there is an orthonormal basis of eigenvectors $v_1, \ldots, v_n$ for $\mathbb{C}^n$ such that

$$v = \sum_{k=1}^{n} (v_k^* v) v_k, \quad v \in \mathbb{C}^n \qquad \text{and} \qquad Av = \sum_{k=1}^{n} \lambda_k (v_k^* v) v_k, \quad v \in \mathbb{C}^n, \tag{1.5}$$

[6] The adjoint of a finite-dimensional matrix is its conjugate transpose [87, Ch. I § 3.6].

[7] Most results in this section have straightforward extensions to normal operators [122, Ch. 13].

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of A, i.e., $Av_k = \lambda_k v_k$ for $1 \leq k \leq n$. The expansions in (1.5) demonstrate two key properties of the eigenvectors of A; they form a complete orthonormal basis for $\mathbb{C}^n$ and they diagonalize A. The expansions are equivalent to the more familiar matrix equations $V^*V = I$ and $A = V\Lambda V$, where V and Λ are the usual matrices of eigenvector and eigenvalues.

We now turn to the infinite-dimensional case. Here, the adjoint of $\mathcal{L}$, denoted $\mathcal{L}^*$, is defined by the relation $(g, \mathcal{L}u)_{\mathcal{H}} = (\mathcal{L}^* g, u)_{\mathcal{H}}$, for all $u \in \mathcal{D}(\mathcal{L})$ and $g \in \mathcal{D}(\mathcal{L}^*)$. The adjoint is closed, densely defined in $\mathcal{H}$, and uniquely determined when $\mathcal{L}$ is closed and densely defined [87, Ch. V § 3.1]. We say that $\mathcal{L}$ is *self-adjoint* if $\mathcal{L} = \mathcal{L}^*$ [87, Ch. V § 3.3]. A countable set of vectors $v_1, v_2, v_3, \ldots \in \mathcal{H}$ is called an orthonormal basis for $\mathcal{H}$ if the infinite series converges in the norm $\|\cdot\|_{\mathcal{H}}$ [27, § 4.9],

$$f = \sum_{k=1}^{\infty} (v_k, f)_{\mathcal{H}} v_k, \qquad \text{for all} \qquad f \in \mathcal{H}.$$

A self-adjoint operator $\mathcal{L}$ with purely discrete spectrum is diagonalized by an orthonormal basis of eigenvectors in the sense of (1.5).

Theorem 1.4.1. *Given a Hilbert space $\mathcal{H}$, let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be self-adjoint and densely defined. If the spectrum of $\mathcal{H}$ comprises a countable set of isolated eigenvalues with finite multiplicity, then there is an orthogonal basis of eigenvectors $u_1, u_2, u_3, \ldots \in \mathcal{H}$ with real eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots$ satisfying* (1.1) *and*

$$f = \sum_{k \in \mathcal{I}} (u_k, f)_{\mathcal{H}} u_k, \quad f \in \mathcal{H} \qquad \text{and} \qquad \mathcal{L}u = \sum_{k \in \mathcal{I}} \lambda_k (u_k, u)_{\mathcal{H}} u_k, \quad u \in \mathcal{D}(\mathcal{L}),$$

where $\mathcal{I}$ is the countable set indexing the eigenvalues of $\mathcal{L}$ (including multiplicities).

Proof. The proof follows from equations (3.18)-(3.22) in [87, Ch. V § 3.5]. □

The spectral theorems for compact self-adjoint operators and self-adjoint operators with compact resolvent are special cases of Theorem 1.4.1 [27,

§ 4.11]. Theorem 1.4.1 can also be written with the spectral projectors introduced in (1.4), as

$$f = \sum_{\lambda \in \Lambda} \mathcal{P}_\lambda f, \quad f \in \mathcal{H} \qquad \text{and} \qquad \mathcal{L}u = \sum_{\lambda \in \Lambda} \lambda_k \mathcal{P}_\lambda u, \quad u \in \mathcal{D}(\mathcal{L}). \tag{1.6}$$

1.4.1 The spectral theorem

If $\mathcal{L}$ has non-empty continuous spectrum, then eigenfunctions of $\mathcal{L}$ do not form a basis for $\mathcal{H}$ or diagonalize $\mathcal{L}$. However, the spectral theorem for self-adjoint operators states that the projections $\mathcal{P}_\lambda$ in (1.6) can be replaced by a projection-valued measure $\mathcal{E}$ [119, Thm. VIII.6]. The measure $\mathcal{E}$ assigns an orthogonal projector to each Borel-measurable set such that

$$f = \int_{\mathbb{R}} d\mathcal{E}(y) f, \quad f \in \mathcal{H} \qquad \text{and} \qquad \mathcal{L}u = \int_{\mathbb{R}} y \, d\mathcal{E}(y) u, \quad u \in \mathcal{D}(\mathcal{L}). \tag{1.7}$$

Analogous to (1.6), $\mathcal{E}$ decomposes $\mathcal{H}$ and diagonalizes the operator $\mathcal{L}$.

The spectral measure of $\mathcal{L}$ with respect to $f \in \mathcal{H}$ is a scalar measure defined as $\mu_f(\Omega) := (\mathcal{E}(\Omega) f, f)$, where $\Omega \subset \mathbb{R}$ is a Borel-measurable set [119]. It is useful to examine Lebesgue's decomposition of μ_f [135], i.e.,

$$d\mu_f(y) = \underbrace{\sum_{\lambda \in \Lambda^{\mathrm{p}}(\mathcal{L})} (\mathcal{P}_\lambda f, f)\, \delta(y - \lambda) dy}_{\text{discrete part}} + \underbrace{\rho_f(y)\, dy + d\mu_f^{(\mathrm{sc})}(y)}_{\text{continuous part}}. \tag{1.8}$$

The discrete part of μ_f is a sum of Dirac delta distributions, supported on the set of eigenvalues of $\mathcal{L}$, and the coefficient of each δ in the sum is $(\mathcal{P}_\lambda f, f) = \|\mathcal{P}_\lambda f\|^2$. The continuous part of μ_f consists of an absolutely continuous[8] part with Radon–Nikodym derivative $\rho_f \in L^1(\mathbb{R})$ and a singular continuous component

[8] We take "absolutely continuous" to be with respect to Lebesgue measure.

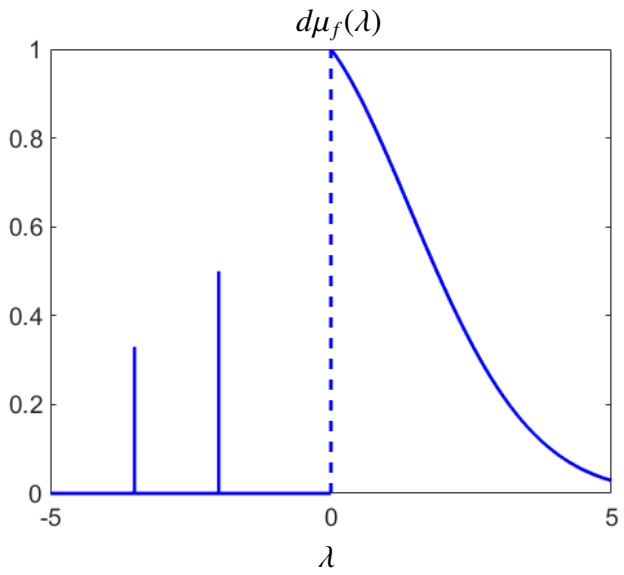

Figure 1.3: *Spectral measures.* The spectral measure μ_f of an operator with mixed spectral types can be decomposed into atoms (shown as spikes) located at eigenvalues, an integrable density (shown as smooth curve) on the absolutely continuous spectrum, and a singular continuous component (not depicted).

$\mu_f^{(\mathrm{sc})}$ (see Figure 1.3). When $\|f\| = 1$, the spectral measure μ_f is a probability measure.

Spectral measures of the form μ_f are related to correlation in stochastic processes and signal-processing [62, 86] [121, Ch. 7], scattering cross-sections in particle physics [50–52], the local density-of-states in crystalline materials [15, 75, 98], and many other quantities [39, 44, 91, 160, 167]. Furthermore, through spectral measures one can compute the functional calculus of $\mathcal{L}$, which is used to solve evolution equations such as the Schrödinger equation in quantum mechanics [79, 99]. We develop an algorithm to compute scalar spectral measures of $\mathcal{L}$ in Chapter 4.

1.4.2 The nuclear spectral theorem

Points in the continuous spectrum can sometimes be associated with generalized eigenfunctions. These objects are similar to eigenfunctions but are not in the Hilbert space. The nuclear spectral theorem describes when a self-adjoint operator has a complete system of generalized eigenfunctions. To make this precise, we need to equip (i.e., rig) a Hilbert space with a suitable dense subspace, containing test functions, and its topological dual, containing generalized eigenfunctions. It is helpful to begin with an illustrative example.

Multiplication by x. Consider the "multiplication by x" operator on $L^2[-1,1]$, defined by $[\mathcal{L}u](x) = x\,u(x)$, for $x \in [-1,1]$. The operator $\mathcal{L}$ is self-adjoint in $L^2[-1,1]$, the Hilbert space of real square-integrable functions on the unit interval. The resolvent of $\mathcal{L}$, with action

$$(\mathcal{L} - z)^{-1} f = (x - z)^{-1} f, \qquad f \in L^2[-1,1],$$

is bounded if and only if $z \notin [-1,1]$. Consequently, its spectrum fills the unit interval. However, there is no square-integrable eigenfunction associated with $\lambda \in [-1,1]$, because $(x-\lambda)f = 0$ if and only if $f = 0$ almost everywhere.

Although $\mathcal{L}$ has no eigenfunctions in $L^2[-1,1]$, it does have a complete system of *generalized eigenfunctions* associated with the continuous spectrum. Given $\lambda \in [-1,1]$, the following distributional relationship holds

$$\int_{-1}^{1} \delta(y-\lambda)\mathcal{L}\phi(y)\,dy = \lambda \int_{-1}^{1} \delta(y-\lambda)\phi(y)\,dy, \qquad \phi \in C^\infty[-1,1], \tag{1.9}$$

since both sides of the equation are equal to $\lambda\,\phi(\lambda)$. Moreover, any $\phi \in C^\infty[-1,1]$ satisfies

$$\phi(x) = \int_{-1}^{1} \delta(x-\lambda)\phi(\lambda)\,d\lambda, \quad [\mathcal{L}\phi](x) = \int_{-1}^{1} \lambda\delta(x-\lambda)\phi(\lambda)\,d\lambda. \tag{1.10}$$

In (1.9), λ and $\delta(\lambda - y)$ are examples of generalized eigenvalues and generalized eigenfunctions, respectively, of $\mathcal{L}$, and the expansion in (1.10) expresses their completeness in $C^\infty[-1,1]$. The generalized eigenfunctions are not square-integrable functions, but belong to the dual of the test space $C^\infty[-1,1]$.

Rigged Hilbert space. Given a Hilbert space $\mathcal{H}$, a topological vector space Φ, and a continuous inclusion map $i : \Phi \to \mathcal{H}$ with a dense range, the dual inclusion map $i^* : \mathcal{H} \to \Phi^*$ is also continuous with dense range. These technical details make it sensible to write $\Phi \subset \mathcal{H} \subset \Phi^*$ and make the duality pairing between Φ and Φ^* compatible with the inner product on $\mathcal{H}$, in the sense that

$$(\phi, \psi)_{\Phi,\Phi^*} = (\phi, \psi)_{\mathcal{H}},$$

whenever $\phi \in \Phi \subset \mathcal{H}$ and $\psi \in \mathcal{H} \subset \Phi^*$. The embedding $\Phi \hookrightarrow \mathcal{H} \hookrightarrow \Phi^*$ is called a rigged Hilbert space, which is sometimes referred to as a Gel'fand triple.

Generalized eigenfunctions. Suppose that $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ is a self-adjoint operator on a rigged Hilbert space $\Phi \subset \mathcal{H} \subset \Phi^*$ such that $\Phi \subset \mathcal{D}(\mathcal{L})$ and $\mathcal{L}\Phi \subset \Phi$. Given $\lambda \in \lambda(\mathcal{L})$, $\psi_\lambda \in \Phi^*$ is called a generalized eigenfunction of $\mathcal{L}$ if it satisfies

$$(\mathcal{L}\phi, \psi_\lambda)_{\Phi,\Phi^*} = \lambda(\phi, \psi_\lambda)_{\Phi,\Phi^*}, \qquad \text{for all} \qquad \phi \in \Phi. \tag{1.11}$$

The scalar λ is called a generalized eigenvalue. Note that any classical (weak) eigenfunction $u_* \in \mathcal{D}(\mathcal{L})$ with eigenvalue λ_* is also a generalized eigenfunction. That is, we can use the compatibility of the duality pairing and the inner product on $\mathcal{H}$ to compute, for any $\phi \in \Phi$,

$$\begin{aligned}(\mathcal{L}\phi, u_*)_{\Phi,\Phi^*} &= (\mathcal{L}\phi, u_*)_{\mathcal{H}} = (\phi, \mathcal{L}u_*)_{\mathcal{H}} \\ &= \lambda_*(\phi, u_*)_{\mathcal{H}} = \lambda_*(\phi, u_*)_{\Phi,\Phi^*}.\end{aligned}$$

Now, a generalized eigenvalue $\lambda \in \lambda(\mathcal{L})$ may have multiple associated generalized eigenfunctions, denoted $\{\psi_{\lambda,k}\}_{k=1}^{m_\lambda}$ for some $m_\lambda \in \{0, 1, 2, \ldots, \infty\}$. Let

$M = \sup_{\lambda(\mathcal{L})} m_\lambda$ and consider a family of σ-finite Borel measures $\{\mu_k\}_{k=1}^M$ supported on $\lambda(A)$, such that $\operatorname{supp}(\mu_k) \subset \{\lambda \in \lambda(\mathcal{L}) : m_\lambda \geq k\}$. We say that $\mathcal{L}$ has a complete system of generalized eigenpairs (with respect to $\{\mu_k\}_{k=1}^M$) if any $\phi \in \Phi$ can be written as

$$\phi = \sum_{k=1}^{M} \int_{\lambda(\mathcal{L})} (\phi, \psi_{\lambda,k})_{\Phi,\Phi^*} \psi_{\lambda,k} \, d\mu_k(\lambda). \tag{1.12}$$

Note that by the definition of the generalized eigenfunctions, we have

$$\mathcal{L}\phi = \sum_{k=1}^{M} \int_{\lambda(\mathcal{L})} \lambda (\phi, \psi_{\lambda,k})_{\Phi,\Phi^*} \psi_{\lambda,k} \, d\mu_k(\lambda). \tag{1.13}$$

When these expressions hold, the right-hand sides are equivalent to the projection-valued measure (and direct-integral) formulation of the spectral theorem for general self-adjoint operators in a Hilbert space. Note that (1.10) in the "multiplication by x" example is a special case of (1.12) and (1.13).

Nuclear spectral theorem. To ensure that $\mathcal{L}$ has a complete set of generalized eigenpairs, Φ must satisfy an additional condition: it must be a *nuclear space*. Nuclear spaces are topological vector spaces with many useful features of finite-dimensional spaces, e.g., the unit ball is precompact in a nuclear space.

A precise characterization of nuclear spaces is beyond the scope of this thesis; we defer the details to [59, Ch. 3]. For our purpose, it is enough to note that smooth function spaces such as (a) $C^\infty(\mathcal{M})$, where $\mathcal{M}$ is a compact manifold, and (b) the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ of functions with rapidly decaying derivatives of all orders are nuclear spaces (when equipped with the correct topology). Their topological duals are, respectively, the space of compactly supported distributions on $\mathcal{M}$ and the space of tempered distributions on $\mathbb{R}^n$.

Theorem 1.4.2. *Let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be a self-adjoint operator on a rigged Hilbert space $\Phi \subset \mathcal{H} \subset \Phi^*$, where $\Phi \subset \mathcal{D}(\mathcal{L})$ and $\mathcal{L}\Phi \subset \Phi$. If Φ is nuclear, then there is an*

$M \in \{0, 1, 2, \ldots, \infty\}$ and a family of σ-finite Borel measures $\{\mu_k\}_{k=1}^{M}$ supported on $\lambda(\mathcal{L})$, such that $\mathcal{L}$ has a complete set of generalized eigenfunctions with respect to $\{\mu_k\}_{k=1}^{M}$. In particular, (1.12) *and* (1.13) *hold.*

1.5 Three paradigms for computing spectra

Computational methods that target spectra of linear operators incorporate a creative blend of techniques from functional analysis, spectral theory, approximation theory, linear algebra, and numerical analysis. We introduce and develop the key ideas relevant to our work in subsequent chapters of this thesis. Here, we outline a broader context for infinite-dimensional spectral problems by highlighting three computational paradigms under active development.

Discretize, then solve. Since the development of the QR algorithm in the 1960s, computing the spectrum of an infinite-dimensional operator $\mathcal{L}$ is usually done in two steps: first discretize $\mathcal{L}$ to obtain a finite dimensional matrix eigenvalue problem, and then solve the matrix eigenvalue problem with algorithms from numerical linear algebra [45, 58, 65, 106]. Ideally, these discretizations are data-sparse to admit fast linear algebra and have eigenvalues that converge rapidly to those of $\mathcal{L}$ (see Figure 1.4).

Infinite-dimensional linear algebra. Recently, algorithms of numerical linear algebra for solving linear systems and diagonalizing matrices have been extended to certain classes of structured infinite-dimensional matrices [32,103,104, 164]. Here, $\mathcal{L}$ is represented as a structured (e.g., block banded) matrix acting on square summable sequences, after choosing a basis for $\mathcal{H}$. The entries of the matrix are usually known explicitly or generated recursively [104]. The focus is

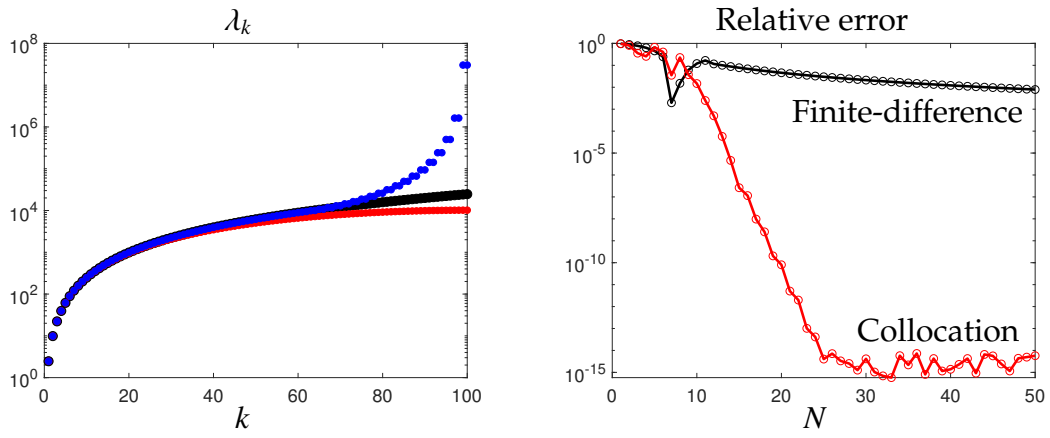

Figure 1.4: *Eigenvalues of discretizations.* The eigenvalues of $[\mathcal{L}u](x) = -u''(x)$ (black markers) are compared with those of 100×100 finite difference (red) and spectral collocation (blue) matrices (left panel). The former converge at an algebraic rate while the latter converge super-geometrically to those of $\mathcal{L}$ as $N \to \infty$, as seen in the relative error in the computed approximation to λ_5 (right panel).

on rigorously approximating spectral properties – such as eigenvalues, eigenvectors, and spectral measures – to a specified accuracy by "lazy evaluation," i.e., adaptively accessing entries from the infinite matrix.

Discretization-oblivious algorithms. The two previous paradigms both depend on a matrix representation of the operator, whether finite or infinite, and numerical approximations are constructed by manipulating entries of the matrix. An alternative approach is to design algorithms that directly manipulate $\mathcal{L}$ and functions in $\mathcal{H}$ [8, 61, 71]. For example, if $\mathcal{L}$ is bounded with $\mathcal{D}(\mathcal{L}) = \mathcal{H}$, one may design algorithms based on operator-function products $\mathcal{L}f$. Effective implementations naturally depend on discretization (typically adaptive) to run on a computer; however, the particular discretization used to compute $\mathcal{L}f$ is in some sense superfluous (provided that associated approximation errors do not compound).

Distinctions between these three paradigms are subtle at times, giving rise to slight differences in emphasis, analysis, and implementation [165]. Each

paradigm provides a helpful conceptual lens through which to view spectral computations. The algorithms introduced in the following chapters are best understood as "discretization-oblivious algorithms" for computing spectra of closed operators, built on two essential computational ingredients: (1) solving linear operator equations $(\mathcal{L} - z)u = f$ with complex shifts $z \in \mathbb{C}$ and (2) computing inner products $(f, g)_{\mathcal{H}}$ with functions $f, g \in \mathcal{H}$. For differential and integral operators on simple geometries, these two tasks are implemented with fast, adaptive spectral methods that automatically resolve dense subsets of smooth functions in $\mathcal{H}$. However, all of the algorithms that we propose may be adapted to more complicated settings, provided that one can execute these two tasks.

CHAPTER 2

COMPUTING ISOLATED EIGENVALUES AND EIGENFUNCTIONS

This chapter[1] is about computing the eigenvalues of $\mathcal{L}$ contained in a simply connected region $\Omega \subset \mathbb{C}$. Throughout, we assume that the boundary $\partial\Omega$ is a rectifiable, simple closed curve, that the spectrum $\Lambda(\mathcal{L})$ does not intersect $\partial\Omega$, and that Ω contains finitely many eigenvalues counting multiplicities. To simplify discussions about eigenfunctions, we assume that there are eigenfunctions of $\mathcal{L}$ that form a basis for the invariant subspace of $\mathcal{L}$ associated with Ω. Our framework and analyses are general, but the exposition is focused on a representative setting for clarity: $\mathcal{L}$ is a linear, ordinary differential operator of even order N. This simple context is sufficient to illustrate the attractive features of our approach.

Motivated by mathematical software for highly adaptive computations with functions [46], we propose the following strategy: an algorithm that solves (1.1) by directly manipulating $\mathcal{L}$ at the continuous level and only discretizes functions, not operators. By designing an eigensolver for $\mathcal{L}$ rather than intermediate discretizations, we are able to leverage spectrally accurate approximation schemes for functions while avoiding several pitfalls that plague spectral discretizations of (1.1) (see [158], [60, Ch. 2], and [157, Ch. 30]). For this reason, we view our proposed algorithms through the "discretization-oblivious" paradigm. This paradigm has been applied to Krylov methods [61], iterative eigensolvers [71], and contour integral projection eigensolvers [8] for differential operators. Related techniques for computing with operators on infinite di-

[1] This chapter is based on a paper with Alex Townsend [83]. I was the lead author and developed the theory, algorithms, and software with input from Alex in our weekly meetings.

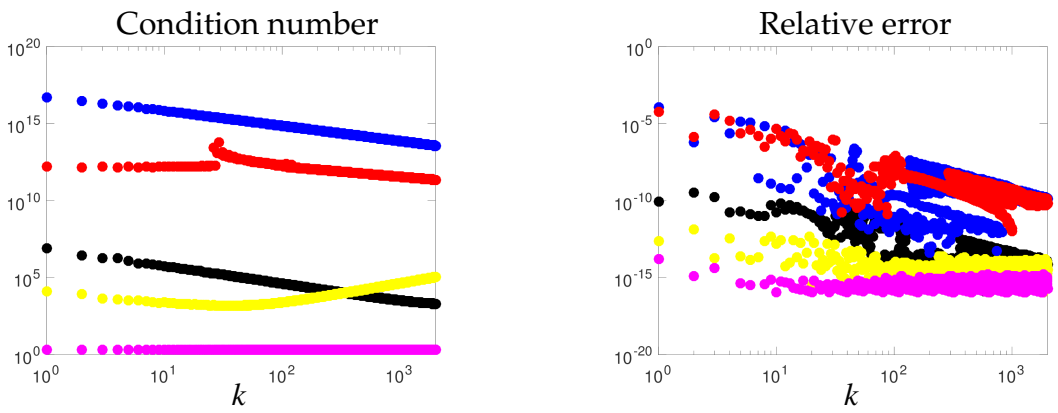

Figure 2.1: Left: The eigenvalue condition numbers [6] for 4000×4000 discretizations of (2.1) obtained by collocation (blue dots), tau (red dots), Chebyshev–Galerkin (black dots), and ultraspherical (yellow dots) spectral methods are compared to the eigenvalue condition numbers (magenta dots) of (2.1), which are preserved by the operator analogue of FEAST. Right: The relative errors in the first 2000 eigenvalues of each spectral discretization of (2.1), computed with a backward stable eigensolver [64, p. 385]. We observe fluctuations in the relative errors due to the ill-conditioning introduced by using nonsymmetric spectral discretizations of $\mathcal{L}$. In contrast, the relative errors (magenta dots) in the eigenvalues computed by `contFEAST`, a practical implementation of the operator analogue of FEAST (see section 2.3), are on the order of machine precision.

mensional spaces have been proposed and studied in [32, 104, 164].

As an example of the advantages of our methodology, consider the simplest possible differential eigenvalue problem given by

$$-u''(x) = \lambda u, \qquad u(\pm 1) = 0. \tag{2.1}$$

The eigenvalues of (2.1) are $\lambda_k = (k\pi/2)^2$, for $k \geq 1$, and are well-conditioned due to the fact that the eigenfunctions form a complete orthonormal set in the Hilbert space $L^2([-1,1])$ [87, p. 382]. However, spectral discretizations of (2.1) lead to highly non-normal matrices with eigenvalues that are far more ill-conditioned than expected. Due to this ill-conditioning, the accuracy in the computed eigenvalues can be extremely variable and difficult to predict, ranging from a few digits to nearly full precision (see Figure 2.1). It is possible to use structure-preserving spectral discretizations to solve (2.1) accurately [25, 129]. However, there is a lack of literature on designing spectral discretizations of (1.1) when $\mathcal{L}$

is self-adjoint or normal with respect to an inner product other than $L^2([-1,1])$. On the other hand, our solve-then-discretize methodology automatically preserves the normality or self-adjointness of $\mathcal{L}$ with respect to a relevant Hilbert space $\mathcal{H}$, provided that the inner product $(\cdot,\cdot)_{\mathcal{H}}$ can be evaluated.

At the heart of our approach is an operator analogue of the FEAST matrix eigensolver, which we briefly outline:

(1) We construct a basis for the eigenspace $\mathcal{V}$ corresponding to Ω by sampling the range of the associated spectral projector $\mathcal{P}_{\mathcal{V}}$.
(2) We extract an $\mathcal{H}$-orthonormal basis for $\mathcal{V}$ with a continuous analogue of the QR factorization [154].
(3) We perform a Rayleigh–Ritz projection [124, p. 98] of $\mathcal{L}$ onto $\mathcal{V}$ with the orthonormal basis in (2). We solve the resulting matrix eigenvalue problem to obtain approximations to the eigenvalues of $\mathcal{L}$ in Ω.

As with the FEAST matrix eigensolver, the spectral projector $\mathcal{P}_{\mathcal{V}}$ is applied approximately via a quadrature rule approximation. For matrices, this involves solving shifted linear systems, while for differential operators one needs to solve shifted linear differential equations. We solve these differential equations with the ultraspherical spectral method, which is a well-conditioned spectral method that is capable of resolving solutions that exhibit layers, rapid oscillations, and weak corner singularities [103].

Critically, we discretize basis functions for $\mathcal{V}$ as opposed to discretizing the differential operator $\mathcal{L}$ when solving (1.1). While discretizations of a normal operator $\mathcal{L}$ can lead to non-normal matrices, the Rayleigh–Ritz projection described in (3) always leads to a normal matrix eigenvalue problem when $\mathcal{L}$ is

normal (see Theorems 2.2.1 and 2.2.2). In fact, we prove that using a sufficiently good approximate basis for $\mathcal{V}$ does not significantly increase the sensitivity of the eigenvalues when $\mathcal{L}$ is normal (see section 2.4.2 for a precise statement). The result is a highly accurate eigensolver for normal differential operators $\mathcal{L}$, requiring $O(mMN\log(N)+m^2N+m^3)$ floating point operations, where $m = \dim(\mathcal{V})$ and M and N are the polynomial degrees used to resolve the variable coefficients in $\mathcal{L}$ and the eigenfunctions in $\mathcal{V}$, respectively.

The eigensolver we develop is competitive in the high-frequency regime because it efficiently resolves oscillatory basis functions in $\mathcal{V}$. Furthermore, it handles operators that are self-adjoint or normal with respect to non-standard Hilbert spaces. Finally, our algorithm is parallelizable like the FEAST matrix eigensolver [116]. This work is a step towards closing the gap between the frequency regimes that are accessible to computational techniques and asymptotic methods for differential eigenvalue problems posed on higher dimensional domains [13, 17].

The chapter is organized as follows. In section 2.2 we introduce an analogue of FEAST for closed operators and show that the operator analogue preserves eigenvalue sensitivity. In section 2.3 we discuss a practical implementation of the operator analogue for differential operators and provide two examples from Sturm–Liouville theory to illustrate its capabilities in the high-frequency regime. We analyze the convergence and stability of this implementation in section 2.4. Sections 2.5 and 2.6 develop further applications of the discretization-oblivious paradigm, including an operator analogue of the Rayleigh Quotient iteration and an extension of FEAST to unbounded search regions.

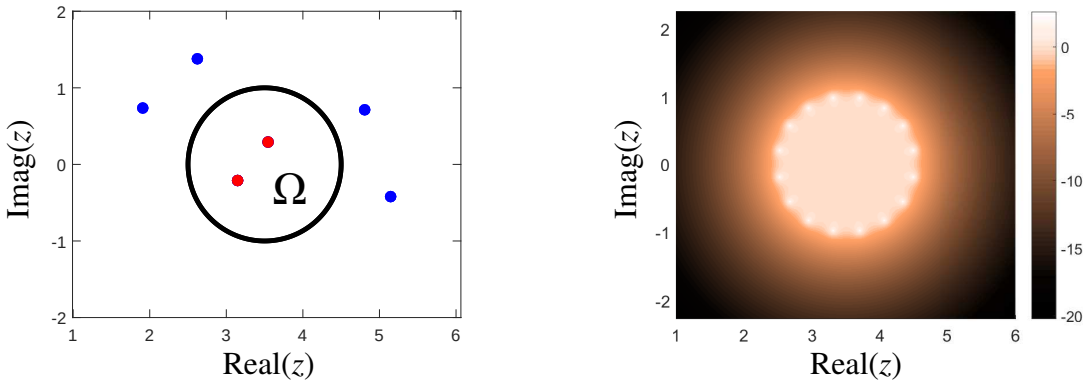

Figure 2.2: FEAST uses an approximation to the spectral projector to compute the eigenvalues that lie inside Ω and project away the eigenvalues outside of Ω (left panel). The approximate spectral projector is often interpreted as a rational map approximating the characteristic function on Ω (right panel).

2.1 The FEAST matrix eigensolver

The FEAST matrix eigensolver uses approximate spectral projection to compute the eigenvalues of a matrix $A \in \mathbb{C}^{n\times n}$ in a region of interest $\Omega \subset \mathbb{C}$ [89] (see Figure 2.2). It is often more computationally efficient than standard eigensolvers when the number of eigenvalues in Ω is much smaller than n and A has data-sparsity that allows fast shifted linear solves. The dominating computational cost of FEAST is solving several independent shifted linear systems, but these can be performed in parallel [89]. There are three essential ingredients to FEAST:

(i) **Spectral projector.** Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of A in Ω and let $\mathcal{V}$ be the associated invariant subspace of A, i.e., $A\mathcal{V} = \mathcal{V}$. The *spectral projector* onto $\mathcal{V}$ is defined as

$$P_{\mathcal{V}} = \frac{1}{2\pi i}\int_{\partial\Omega} (zI - A)^{-1}\, dz. \tag{2.2}$$

The important fact here is that $\text{range}(P_{\mathcal{V}}) = \mathcal{V}$ and so $P_{\mathcal{V}}$ is a projection onto the invariant subspace of A [87].

(ii) **Basis for $\mathcal{V}$.** FEAST uses the spectral projector to construct a basis for $\mathcal{V}$. It

Algorithm 1 The FEAST algorithm for matrix eigenvalue problems [116]. This is often viewed as a single iteration that is repeated to improve the accuracy of the computed eigenvalues and eigenvectors [144].

Input: $A \in \mathbb{C}^{n\times n}$, $\Omega \subset \mathbb{C}$ containing m eigenvalues of A, $Y : \mathbb{C}^{n\times m}$.

1: Compute $V = \mathcal{P}_{\mathcal{V}} Y$.
2: Compute the QR factorization $V = QR$.
3: Compute $A_Q = Q^* A Q$ and solve the eigenvalue problem $A_Q X = \Lambda X$ for $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and $X \in \mathbb{C}^{m\times m}$.

Output: Eigenvalues $\lambda_1, \ldots, \lambda_m$ in Ω and eigenvectors $U = QX$.

begins with a matrix $Y \in \mathbb{C}^{n\times m}$ with linearly independent columns that are not in $\ker(P_{\mathcal{V}})$, then it computes $Z = P_{\mathcal{V}} Y$. The columns of Z span $\mathcal{V}$ and a QR factorization of Z provides an orthonormal basis, Q, for $\mathcal{V}$.

(iii) **Rayleigh–Ritz projection.** Having obtained an orthonormal basis for $\mathcal{V}$, FEAST solves $A_Q x = \lambda x$ using a dense eigensolver [116], where $A_Q = Q^* A Q$. Since $\operatorname{range}(Q) = \mathcal{V}$, the eigenvalues of A_Q are the eigenvalues of A that lie inside Ω. When A_Q is diagonalizable, the eigenvectors of A are given by $u_i = Q x_i$, for $i = 1, \ldots, m$, where $x_1, \ldots, x_m$ are the corresponding eigenvectors of A_Q.

For practical computation, FEAST approximates the contour integral in (2.2) with a quadrature rule. Given a quadrature rule with nodes $z_1, \ldots, z_\ell$ and weights $w_1, \ldots, w_\ell$, one can approximate $P_{\mathcal{V}} Y$ by

$$P_{\mathcal{V}} Y \approx \frac{1}{2\pi i} \sum_{k=1}^{\ell} w_k (z_k I - A)^{-1} Y. \tag{2.3}$$

In this case, the eigenpairs of A_Q provide approximations to the eigenpairs of A, known as Ritz values and vectors [144]. To refine the accuracy of the Ritz values and vectors, a more accurate quadrature rule can be used to compute $P_{\mathcal{V}} Y$ [144]. FEAST also refines the approximate eigenvalues and eigenvectors by applying $P_{\mathcal{V}}$ to the $n \times m$ block of approximate eigenvectors using a quadrature rule and iterating (ii) and (iii) until convergence. This iterative refinement strategy is

best understood as subspace iteration with a rational filter [144]; in Chapter 2, we discover that iteration is crucial to the numerical stability of FEAST.

When the dimension m of the invariant subspace $\mathcal{V}$ is unknown, there are several techniques for estimating m and selecting an appropriate value [89, 102, 144]. These techniques can be incorporated into the operator analogue of FEAST introduced in Chapter 1 without any difficulty. Consequently, we usually assume that a good upper estimate for m is known and focus on the algorithmic and theoretical aspects of FEAST that are salient in the operator setting.

Curiously, the originally proposed FEAST algorithm does not compute an orthonormal basis for $\mathcal{V}$ before performing the Rayleigh–Ritz projection [89, 116].[2] However, when Q has orthonormal columns and $\operatorname{range}(Q)$ is an invariant subspace of A, the eigenvalues of the small matrix Q^*AQ are no more sensitive to perturbations than the original eigenvalues of A. This highly desirable property can be deduced from the structure of the left and right invariant subspaces of Q^*AQ or, alternatively, from the ϵ-pseudospectra of Q^*AQ [157, p. 382].

2.2 An operator analogue of FEAST

The FEAST matrix algorithm provides a natural starting point for an operator analogue because it provides a recipe to construct a small matrix Q^*AQ whose eigenvalues coincide with those of A inside Ω and have related invariant subspaces. Moreover, the eigenstructure of Q^*AQ reflects the eigenstructure of A when the columns of Q are orthonormal. As the sensitivity of the eigenvalues of

[2] The FEAST algorithm for non-Hermitian matrices utilizes dual bases for the left and right eigenspaces to improve stability [89].

A depends intimately on the structure of the associated eigenvectors, Q^*AQ may be used to compute the desired eigenvalues of A efficiently without sacrificing accuracy. Here, we generalize FEAST to construct a matrix whose eigenvalues coincide with those of a closed operator inside Ω.

2.2.1 FEAST for closed operators

In place of a matrix A acting on vectors from $\mathbb{C}^n$, we now consider a closed operator $\mathcal{L}$ acting on functions from a Hilbert space $\mathcal{H}$. As described in section 2.1, the FEAST recipe prescribes a spectral projection to compute a basis for $\mathcal{V}$, which is then used for the Rayleigh–Ritz projection to construct a matrix representation on $\mathcal{V}$.

(i) **Spectral projector.** Although $\mathcal{L}$ may be unbounded, the resolvent $(z\mathcal{I} - \mathcal{L})^{-1}$ is bounded when $z \notin \Lambda(\mathcal{L})$ and the spectral projector onto $\mathcal{V}$ may be defined via contour integral [87, p. 178]. It is given by

$$\mathcal{P}_{\mathcal{V}} = \frac{1}{2\pi i} \int_{\partial\Omega} (z\mathcal{I} - \mathcal{L})^{-1} \, dz. \tag{2.4}$$

(ii) **Basis for $\mathcal{V}$.** With the spectral projector at our disposal, we apply $\mathcal{P}_{\mathcal{V}}$ to functions $f_1, \ldots, f_m$ in $\mathcal{H} \setminus \ker(\mathcal{P}_{\mathcal{V}})$ to obtain a basis of functions $v_1, \ldots, v_m$ for $\mathcal{V}$. Orthonormalizing $v_1, \ldots, v_m$ with respect to the inner product $(\cdot, \cdot)_{\mathcal{H}}$ on $\mathcal{H}$ gives us an $\mathcal{H}$-orthonormal basis $q_1, \ldots, q_m$ for $\mathcal{V}$.

(iii) **Rayleigh–Ritz projection.** To compute a matrix representation L of $\mathcal{L}$ on $\mathcal{V}$, the Rayleigh–Ritz projection is performed using the inner product on $\mathcal{H}$. The elements of L are given by $L_{ij} = (q_i, \mathcal{L}q_j)_{\mathcal{H}}$ for $1 \leq i, j \leq m$. The eigenvalues of L are precisely the eigenvalues $\lambda_1, \ldots, \lambda_m$ of $\mathcal{L}$ that lie inside Ω. The eigenfunc-

Algorithm 2 An operator analogue of FEAST for closed operators.

Input: $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$, $\Omega \subset \mathbb{C}$ containing m eigenvalues of $\mathcal{L}$, $F : \mathbb{C}^m \to \mathcal{H}$.

1: Compute $V = \mathcal{P}_{\mathcal{V}} F$.
2: Compute $V = QR$, where $Q : \mathbb{C}^m \to \mathcal{D}(\mathcal{L}) \subset \mathcal{H}$ has $\mathcal{H}$-orthonormal columns and $R \in \mathbb{C}^{m \times m}$ is upper triangular.
3: Compute $L = Q^* \mathcal{L} Q$ and solve $LX = \Lambda X$ for $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_m]$ and $X \in \mathbb{C}^{m \times m}$.

Output: Eigenvalues $\lambda_1, \ldots, \lambda_m$ in Ω and eigenfunctions $U = QX$.

tions of $\mathcal{L}$ are recovered from the eigenvectors $x_1, \ldots, x_m$ of L by computing $u_i = \sum_{k=1}^{m} x_i^{(k)} q_k$, for $i = 1, \ldots, m$, where $x_i^{(k)}$ is the k^{th} component of x_i.

To avoid a clutter of indices, we employ the notation of quasimatrices.[3] If Q is the quasimatrix with columns $q_1, \ldots, q_m$, then the matrix L whose elements are $L_{ij} = (q_i, \mathcal{L} q_j)_{\mathcal{H}}$ in (iii) is expressed compactly in quasimatrix notation as $L = Q^* \mathcal{L} Q$. Here, Q^* is the conjugate transpose of the quasimatrix Q so its rows are complex conjugates of the functions $q_1, \ldots, q_m$.

The analogue of FEAST for closed operators is summarized in Algorithm 2 using quasimatrix notation so that it resembles its matrix counterpart. Keep in mind that Algorithm 2 is a formal algorithm. In general, we cannot apply the spectral projector exactly, nor represent the basis $\mathcal{V}$ exactly with finite memory. A practical implementation is discussed in section 2.3.

2.2.2 Condition number of the Ritz values

As illustrated in Figure 2.1, the eigenvalues of matrix discretizations of $\mathcal{L}$ can be more sensitive to perturbations than the eigenvalues of $\mathcal{L}$. The advantage of

[3] A *quasimatrix* is a matrix whose columns (or rows) are functions defined on an interval $[a, b]$, in contrast to matrices whose columns (or rows) are vectors [46, Ch.6].

our FEAST approach in section 2.2.1 is that the Ritz values, i.e., the eigenvalues of $Q^*\mathcal{L}Q$, are no more sensitive to perturbations than the original eigenvalues of $\mathcal{L}$ when range(Q) is an invariant subspace of $\mathcal{L}$ (for perturbations of an invariant subspace, see section 2.4).

To see this, let λ be a simple eigenvalue of a differential operator $\mathcal{L}$. Let $u, w \in \mathcal{H}$ satisfy $\mathcal{L}u = \lambda u$ and $\mathcal{L}^*w = \overline{\lambda}w$, where $\overline{\lambda}$ denotes the complex conjugate of λ. The *condition number*[4] of λ is given by [6, Theorem 2.3]

$$\kappa_{\mathcal{H}}(\lambda) = \frac{\|u\|_{\mathcal{H}}\|w\|_{\mathcal{H}}}{(w,u)_{\mathcal{H}}}. \tag{2.5}$$

The condition number $\kappa_{\mathcal{H}}(\lambda)$ quantifies the worst-case first-order sensitivity of λ to perturbations of $\mathcal{L}$. For instance, if we compute λ using a backward stable algorithm in floating point arithmetic, we expect to achieve an accuracy of at least $\kappa_{\mathcal{H}}(\lambda)\epsilon_{\text{mach}}$, where ϵ_{mach} is machine precision [156, Theorem 15.1].

Theorem 2.2.1. *Let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be a closed and densely defined operator on a Hilbert space $\mathcal{H}$, $Q : \mathbb{C}^m \to \mathcal{H}$ be an invariant subspace of $\mathcal{L}$ satisfying $Q^*Q = I$, and $L = Q^*\mathcal{L}Q$. Suppose that $u \in$ range(Q) satisfies $\mathcal{L}u = \lambda u$ and w satisfies $\mathcal{L}^*w = \overline{\lambda}w$, where $\mathcal{L}^*$ denotes the adjoint of $\mathcal{L}$ and λ is a simple eigenvalue of $\mathcal{L}$ with condition number $\kappa_{\mathcal{H}}(\lambda)$. Then,*

1) *$LQ^*u = \lambda Q^*u$ and $L^*Q^*w = \overline{\lambda}Q^*w$,*
2) *$(Q^*w, Q^*u)_{\mathbb{C}^m} = (w,u)_{\mathcal{H}}$, and*
3) *$\kappa_{\mathbb{C}^m}(\lambda) \le \kappa_{\mathcal{H}}(\lambda)$.*

Proof. Denote $x = Q^*u$ and $y = Q^*w$. We prove the statements of the theorem in order. 1) Since $u \in$ range(Q), we can write $u = Qx$. Then, $\mathcal{L}(Qx) = \lambda(Qx)$

[4] Although this formula is usually associated with the condition number for a simple eigenvalue of a matrix, its proof extends to our general setting [149, Theorem 5].

implies that $Q^*\mathcal{L}Qx = \lambda x$ using the fact that $Q^*Q = I$. For the left eigenvector, we write $w = Qy + v$ for some $v \in \text{range}(Q)^\perp$. Rewriting the adjoint equation for w, we find that $\mathcal{L}^*(Qy + v) = \overline{\lambda}(Qy + v)$ and multiplying by Q^* on both sides yields $Q^*\mathcal{L}^*Qy = \overline{\lambda}y$. Here, we have used the fact that $Q^*\mathcal{L}^*v = 0$, which holds because $v^*\mathcal{L}Q = 0$. 2) By calculating $(w, u)_{\mathcal{H}} = (Qy + v, Qx)_{\mathcal{H}}$, we find that $(w, u)_{\mathcal{H}} = (Qy, Qx)_{\mathcal{H}}$ because $v \in \text{range}(Q)^\perp$. Moreover, since $Q^*Q = I$ we conclude that $(w, u)_{\mathcal{H}} = (y, Q^*Qx)_{\mathbb{C}^m} = (y, x)_{\mathbb{C}^m}$. 3) We know that $\|u\|_{\mathcal{H}} = (Qx, Qx)_{\mathcal{H}} = (x, x)_{\mathbb{C}^m} = \|x\|_{\mathbb{C}^m}$ and $\|w\|_{\mathcal{H}} = (Qy + v, Qy + v)_{\mathcal{H}} = \|y\|_{\mathbb{C}^m} + \|v\|_{\mathcal{H}}$. Therefore,

$$\|u\|_{\mathcal{H}}\|w\|_{\mathcal{H}} = \|x\|_{\mathbb{C}^m}(\|y\|_{\mathbb{C}^m} + \|v\|_{\mathcal{H}}) \geq \|x\|_{\mathbb{C}^m}\|y\|_{\mathbb{C}^m}.$$

Referring to 2) for equality of the inner products in the denominator, we have

$$\kappa_{\mathbb{C}^m}(\lambda) = \frac{\|x\|_{\mathbb{C}^m}\|y\|_{\mathbb{C}^m}}{(y, x)_{\mathbb{C}^m}} \leq \frac{\|u\|_{\mathcal{H}}\|w\|_{\mathcal{H}}}{(w, u)_{\mathcal{H}}} = \kappa_{\mathcal{H}}(\lambda),$$

which concludes the proof. □

Theorem 2.2.1 shows that if $\mathcal{L}$ is a normal operator, then $u = w$ and we have $\kappa_{\mathbb{C}^m}(\lambda) = \kappa_{\mathcal{H}}(\lambda) = 1$. For non-normal operators, item 3) of Theorem 2.2.1 may seem to erroneously indicate that ill-conditioning in the eigenvalues of $\mathcal{L}$ can be overcome by a Rayleigh–Ritz projection. However, when $\mathcal{L}$ is non-normal the spectral projector $\mathcal{P}_{\mathcal{V}}$ is an oblique projection and computing the basis Q may be itself an ill-conditioned problem. Theorem 2.2.1 also illustrates why the operator analogue of FEAST leads to a well-conditioned matrix eigenvalue problem when the differential eigenvalue problem is well-conditioned. By computing an $\mathcal{H}$-orthonormal basis for the Rayleigh–Ritz projection, the relevant structure in the eigenspaces of $\mathcal{L}$ and $\mathcal{L}^*$ is preserved. However, the first-order analysis above is limited to simple eigenvalues.

2.2.3 Pseudospectra of $Q^*\mathcal{L}Q$

To go beyond first-order sensitivity analysis, we compare the ϵ-pseudospectra of $\mathcal{L}$ and $Q^*\mathcal{L}Q$. Fix any $\epsilon > 0$ and let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be a closed operator with a domain $\mathcal{D}(\mathcal{L})$ that is dense in $\mathcal{H}$. The ϵ-pseudospectrum of $\mathcal{L}$ is defined as the set [157, p. 31]

$$\Lambda_\epsilon(\mathcal{L}) = \{z \in \mathbb{C} : \|(z\mathcal{I} - \mathcal{L})^{-1}\|_{\mathcal{H}} > 1/\epsilon\}. \tag{2.6}$$

Here, we adopt the usual convention that $\|(z\mathcal{I} - \mathcal{L})^{-1}\|_{\mathcal{H}} = \infty$ when $z \in \Lambda(\mathcal{L})$ so that $\Lambda(\mathcal{L}) \subset \Lambda_\epsilon(\mathcal{L})$. The ϵ-pseudospectrum set of $\mathcal{L}$ bounds the region in which the eigenvalues of the perturbed operator $\mathcal{L} + \mathcal{E}$ with $\|\mathcal{E}\|_{\mathcal{H}} < \epsilon$ can be found [157, p. 31]. This means that $\Lambda(\mathcal{L} + \mathcal{E}) \subset \Lambda_\epsilon(\mathcal{L})$. In fact, there is an equivalence so that [157, p. 31]

$$\bigcup_{\|\mathcal{E}\|_{\mathcal{H}} < \epsilon} \Lambda(\mathcal{L} + \mathcal{E}) = \Lambda_\epsilon(\mathcal{L}). \tag{2.7}$$

This allows us to relate the sensitivity of the eigenvalues of $\mathcal{L}$ and $Q^*\mathcal{L}Q$ by comparing the resolvent norms $\|(z\mathcal{I} - \mathcal{L})^{-1}\|_{\mathcal{H}}$ and $\|(zI - Q^*\mathcal{L}Q)^{-1}\|_{\mathbb{C}^m}$, respectively.

A useful generalization of Theorem 2.2.1 is that the ϵ-pseudospectrum of $Q^*\mathcal{L}Q$ is contained in the ϵ-pseudospectrum of $\mathcal{L}$. Since this holds for any $\epsilon > 0$, it demonstrates that the eigenvalues (even those with multiplicity) of $Q^*\mathcal{L}Q$ are no more sensitive to perturbations than those of $\mathcal{L}$. This inclusion result is well-known in the matrix case where projection methods are a popular method for approximating the ϵ-pseudospectra of large data-sparse matrices [157, p. 381].

Theorem 2.2.2. *Let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be a closed and densely defined operator on a Hilbert space $\mathcal{H}$. For a fixed $\epsilon > 0$, suppose that $Q : \mathbb{C}^m \to \mathcal{H}$ satisfies $Q^*Q = I$ and that* range(Q) *is an invariant subspace of $\mathcal{L}$. Then, $\Lambda_\epsilon(Q^*\mathcal{L}Q) \subset \Lambda_\epsilon(\mathcal{L})$.*

Proof. We follow the proof of Proposition 40.1 in [157, p. 382] for matrices, but with a closed operator. Since $Qx \in \mathcal{H}$ for any $x \in \mathbb{C}^m$, we have that

$$\|(zI - \mathcal{L})^{-1}\|_{\mathcal{H}} = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}}=1} \|(z\mathcal{I} - \mathcal{L})^{-1} f\|_{\mathcal{H}} \geq \max_{x \in \mathbb{C}^m, \|x\|_{\mathbb{C}^m}=1} \|(z\mathcal{I} - \mathcal{L})^{-1} Qx\|_{\mathcal{H}}.$$

Now, when range(Q) is an invariant subspace of $\mathcal{L}$ and $Q^*Q = I$, we have that $\|(z\mathcal{I} - \mathcal{L})^{-1} Qx\|_{\mathcal{H}} = \|Q^*(z\mathcal{I} - \mathcal{L})^{-1} Qx\|_{\mathbb{C}^m}$. Because $QQ^* f = f$ for all $f \in \mathcal{V}$, we can check that $Q^*(z\mathcal{I} - \mathcal{L})^{-1} Q = (Q^*(z\mathcal{I} - \mathcal{L})Q)^{-1}$. Since $Q^*Q = I$, it follows that

$$\|(zI - \mathcal{L})^{-1}\|_{\mathcal{H}} \geq \|(Q^*(z\mathcal{I} - \mathcal{L})Q)^{-1}\|_{\mathbb{C}^m} = \|(zI - Q^*\mathcal{L}Q)^{-1}\|_{\mathbb{C}^m}.$$

Therefore, $z \in \Lambda_\epsilon(\mathcal{L})$ whenever $z \in \Lambda_\epsilon(Q^*\mathcal{L}Q)$. □

The inclusion in Theorem 2.2.2 may be strict, indicating that the eigenvalues of $Q^*\mathcal{L}Q$ are less sensitive than those of $\mathcal{L}$. For example, this may occur when the projection onto range(Q) targets a subset of well-conditioned eigenvalues of $\mathcal{L}$. However, we emphasize that ill-conditioning in the eigenvalues of $\mathcal{L}$ cannot be overcome by a Rayleigh–Ritz projection: in general, the situation is complicated [157, Ch. 40].

Theorem 2.2.2 is useful for studying the stability of Algorithm 2. If an approximate eigenvalue $\hat{\lambda}$ of $Q^*\mathcal{L}Q$ is computed with an error tolerance of $\epsilon > 0$, then

$$\hat{\lambda} \in \Lambda_\epsilon(Q^*\mathcal{L}Q) \subset \Lambda_\epsilon(\mathcal{L}).$$

From this, we know by (2.7) that $\hat{\lambda}$ is an eigenvalue of a perturbed operator $\mathcal{L} + \mathcal{E}$ with $\|\mathcal{E}\|_{\mathcal{H}} < \epsilon$. In other words, the operator analogue of FEAST, Algorithm 2, is backward stable. As we see in section 2.4, Theorem 2.2.2 is also the starting point for a stability analysis when the spectral projection is no longer exact and the Rayleigh–Ritz projection is performed with a matrix $\hat{Q}$ that only approximates a basis for an invariant subspace of $\mathcal{L}$.

2.3 A practical differential eigensolver

The operator analogue of FEAST requires the manipulation of objects such as differential operators, functions, and contour integrals (see Algorithm 2). For a practical implementation, these objects must be discretized; however, we avoid discretizing $\mathcal{L}$ directly. Instead, we construct polynomial approximants to the basis for $\mathcal{V}$ by approximately solving shifted linear ODEs. These polynomial approximants are used in the Rayleigh–Ritz projection to compute the eigenvalues of $\mathcal{L}$ in Ω.

Let $z_1, \ldots, z_\ell$ and $w_1, \ldots, w_\ell$ be a set of quadrature nodes and weights to approximate the integral in (2.4). As FEAST does in the matrix case, we approximate $\mathcal{P}_\mathcal{V}$ in (2.4) with a quadrature rule as follows:

$$\hat{\mathcal{P}}_\mathcal{V} = \frac{1}{2\pi i} \sum_{k=1}^{\ell} w_k (z_k \mathcal{I} - \mathcal{L})^{-1}. \tag{2.8}$$

If F is a quasimatrix with columns $f_1, \ldots, f_m \in \mathcal{H}$, then $\mathcal{P}_\mathcal{V} F$ is replaced by the approximation $\hat{\mathcal{P}}_\mathcal{V} F = \frac{1}{2\pi i} \sum_{k=1}^{\ell} w_k (z_k \mathcal{I} - \mathcal{L})^{-1} F$. Therefore, to compute $\hat{\mathcal{P}}_\mathcal{V} F$ we need to solve ℓ shifted linear ODEs, each with m righthand sides, i.e.,

$$(z_k \mathcal{I} - \mathcal{L}) g_{i,k} = f_i, \quad g_{i,k}(\pm 1) = \cdots = g_{i,k}^{(N/2)}(\pm 1) = 0, \qquad 1 \le i \le m. \tag{2.9}$$

If the quasimatrix with columns $g_{1,k}, \ldots, g_{m,k}$ is denoted by G_k for $k = 1, \ldots, \ell$, then we have $\hat{\mathcal{P}}_\mathcal{V} F = \sum_{k=1}^{\ell} w_k G_k$.

To construct a basis for $\mathcal{V}$, it is important to choose F so that the columns of $\hat{V} = \hat{\mathcal{P}}_\mathcal{V} F$ are linearly independent and, if possible, well-conditioned. By analogy with the implementation of matrix FEAST [89, 116], we obtain the columns of F by selecting m band-limited random functions[5] on $[-1, 1]$ [55]. When $\mathcal{L}$ is a

[5] A periodic band-limited random function on $[-L, L]$ is a periodic function defined by a

Algorithm 3 A practical algorithm for computing the eigenvalues of a differential operator $\mathcal{L}$, which we refer to as `contFEAST`.

Input: $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$, $z_1, \ldots, z_\ell \in \partial\Omega$, $w_1, \ldots, w_\ell \in \mathbb{C}$, $F : \mathbb{C}^m \to \mathcal{H}$, $\epsilon > 0$.

1: **repeat**
2: Solve $(z_k\mathcal{I} - \mathcal{L})G_k = F$, $G_k(\pm 1) = 0, \ldots, G_k^{(N/2)}(\pm 1) = 0$, for $k = 1, \ldots, \ell$.
3: Set $\hat{V} = \sum_{k=1}^{\ell} w_k G_k$.
4: Compute $\hat{V} = \hat{Q}\hat{R}$, where $\hat{Q} : \mathbb{C}^m \to \mathcal{D}(\mathcal{L}) \subset \mathcal{H}$ has $\mathcal{H}$-orthonormal columns and $\hat{R} \in \mathbb{C}^{m\times m}$ is upper triangular.
5: Compute $\hat{L} = \hat{Q}^*\mathcal{L}\hat{Q}$ and solve $\hat{L}\hat{X} = \hat{X}\hat{\Lambda}$ for $\hat{\Lambda} = \mathrm{diag}[\hat{\lambda}_1, \ldots, \hat{\lambda}_m]$ and $\hat{X} \in \mathbb{C}^{m\times m}$. Set $F = \hat{Q}\hat{X}$.
6: **until** $\|\mathcal{L}F - F\hat{\Lambda}\|_{\mathcal{H}} \le \epsilon\|\hat{\Lambda}\|_{\mathbb{C}^m}$.

Output: $\hat{\Lambda}$, $\hat{U} = \hat{Q}\hat{X}$.

normal operator and $\mathcal{V}$ is not orthogonal to the space of band-limited random functions, this generically yields a linearly independent basis $\hat{V}$. We discuss conditions for computing a well-conditioned basis in Chapter 3 and, for more about randomized linear algebra in the infinite-dimensional setting, we recommend Townsend and Boulle's rigorous framework [21].

We now outline the key implementation details of our differential eigensolver:

(i) **Approximate spectral projection.** To compute $\hat{V} = \hat{\mathcal{P}}_{\mathcal{V}}F$, we solve the shifted linear ODEs in (2.9) using the ultraspherical spectral method [103]. The ultraspherical spectral method leads to well-conditioned linear systems and is capable of accurately resolving the functions $g_{i,k}$ even when they are highly oscillatory or have boundary layers. Moreover, an adaptive QR factorization automatically determines the degree of the polynomial interpolants needed to approximate the functions $g_{i,k}$ to near machine precision [104, 105]. After

truncated Fourier series with random (e.g. standard Gaussian distributed) coefficients. In the non-periodic setting, the Fourier series is defined on a larger interval $[-L', L']$ and the domain is then truncated [55].

accurately resolving the functions $g_{i,k}$, we can accurately compute a basis for $\mathcal{V}$ provided that both the spectral projector is well-conditioned (i.e., $\mathcal{L}$ is not highly non-normal) and the quadrature rule is sufficiently accurate.

(ii) **Orthonormal basis.** To compute an orthonormal basis $\hat{Q}$ for the columns of $\hat{V}$, we compute a QR factorization of the quasimatrix $\hat{V}$ by Householder triangularization [154]. The Householder reflectors are constructed with respect to the inner product $(\cdot,\cdot)_{\mathcal{H}}$ so that the columns of $\hat{Q}$ are $\mathcal{H}$-orthonormal.

(iii) **Computing $\hat{\mathbf{Q}}^*\mathcal{L}\hat{\mathbf{Q}}$.** To construct the matrix $\hat{L} = \hat{Q}^*\mathcal{L}\hat{Q}$, we apply $\mathcal{L}$ to the columns of $\hat{Q}$ and then evaluate the action of $\hat{Q}^*$ on $\mathcal{L}\hat{Q}$. Multiplying $\hat{Q}^*$ with $\mathcal{L}\hat{Q}$ involves taking the inner products

$$\hat{L}_{ij} = (\hat{q}_i, \mathcal{L}\hat{q}_j)_{\mathcal{H}}, \qquad 1 \le i, j \le m, \tag{2.10}$$

where $\hat{q}_i$ denotes the ith column of $\hat{Q}$. The eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ and eigenvectors $\hat{x}_1, \ldots, \hat{x}_m$ of the matrix $\hat{L}$ are computed using the QR algorithm [64, p. 385].

Critically, the inner product $(\cdot,\cdot)_{\mathcal{H}}$ used in the QR factorization of $\hat{V}$ and the construction of $\hat{Q}^*\mathcal{L}\hat{Q}$ depends on the choice of the Hilbert space $\mathcal{H}$. As long as we can evaluate $(\cdot,\cdot)_{\mathcal{H}}$, we can exploit the fact that $\mathcal{L}$ is self-adjoint or a normal operator with respect to $(\cdot,\cdot)_{\mathcal{H}}$ so that we can accurately compute the eigenvalues of $\mathcal{L}$ in Ω (see Theorem 2.4.2). For this reason, our algorithm can accurately compute the eigenvalues and eigenfunctions of differential operators that are self-adjoint with respect to non-standard Hilbert spaces (see section 2.3.1).

Evaluating the inner product $(\cdot,\cdot)_{\mathcal{H}}$ usually means computing an integral, which we approximate with a quadrature rule. For example, if $\mathcal{H} = L^2([-1,1])$,

$$(f,g)_{L^2([-1,1])} = \int_{-1}^{1} \overline{f(x)} g(x)\, dx.$$

Given the Gauss–Legendre quadrature nodes $x_1, \ldots, x_p$ and weights $w_1, \ldots, w_p$ on $[-1, 1]$, then one uses the approximation [19]

$$(f, g)_{L^2([-1,1])} \approx \sum_{k=1}^{p} w_k \overline{f(x_k)} g(x_k).$$

A practical implementation of the operator analogue of FEAST is presented in Algorithm 3. As with matrix FEAST, there are two approaches for improving the accuracy of the Ritz values $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ and vectors $\hat{Q}\hat{x}_1, \ldots, \hat{Q}\hat{x}_m$. The first is to improve the accuracy of the quadrature rule in (2.8). The second is to iterate the algorithm by replacing F by the quasimatrix $\hat{U}$ with columns $\hat{u}_i = \hat{Q}\hat{x}_i$ for $1 \leq i \leq m$, repeating the process if necessary.[6] For normal operators, this iteration generates a sequence of quasimatrices $\hat{Q}_k$ with $\mathcal{H}$-orthonormal columns that converge to an $\mathcal{H}$-orthonormal basis for the invariant subspace $\mathcal{V}$ as $k \to \infty$. This can be viewed as a rational subspace iteration and geometric convergence of the Ritz pairs is typical (see section 2.4).

With either refinement strategy, the accuracy of the Ritz pairs may be monitored using the residual norm (see step 6 of Algorithm 3) as a proxy, just as in the matrix case. For normal operators, the error in the eigenvalues and eigenvectors computed by Algorithm 3 is typically $O(\epsilon)$, where ϵ is the threshold for the residual norm in step 6. We defer analyses of convergence and stability for Algorithm 3 to section 2.4 and Chapter 3. Additional resources on residual norm bounds for eigenvalues and eigenvectors of matrices and extensions to closed linear operators are found in [11, 24, 137].

[6] When $\mathcal{L}$ is non-normal the Ritz vectors $\hat{Q}\hat{x}_1, \ldots, \hat{Q}\hat{x}_m$ may become numerically linearly dependent, which can lead to an ill-conditioned basis $\hat{V}$ in susbequent iterations. The robustness of Algorithm 3 may be improved by computing the Schur vectors $v_1, \ldots, v_m$ of $\hat{L}$ and using the orthonormal basis $\hat{Q}\hat{v}_1, \ldots, \hat{Q}\hat{v}_m$ to seed the next iteration [138].

In practice, when $\mathcal{L}$ is non-normal it may be beneficial to use a dual Rayleigh–Ritz projection $\hat{Q}_L^* \mathcal{L} \hat{Q}_R$, where the columns of $\hat{Q}_R$ approximate an orthonormal basis for the target eigenspace of $\mathcal{L}$ and the columns of $\hat{Q}_L$ approximate an orthonormal basis for the associated eigenspace of the adjoint $\mathcal{L}^*$. In the case of matrix FEAST, the use of the dual projection leads to a non-normal matrix eigensolver with improved robustness [89]. Although it is not difficult to adapt Algorithm 3 to an operator analogue of FEAST that uses dual projection, we focus on the implementation and analysis of the one-sided iteration.

Typically, solving the ODEs in (2.9) dominates the computational cost of Algorithm 3. With the ultraspherical spectral method, the computational complexity of solving the linear ODEs with m distinct right hand sides is $O(mMN \log(N))$ floating point operations (flops) [103]. Here, N and M are, respectively, the degrees of the truncated Chebyshev series needed to resolve the columns of G_k and the variable coefficients in $\mathcal{L}$ to within the tolerance ϵ specified in Algorithm 3. In addition to the ODE solve, the QR factorization in (ii) requires $O(m^2 N)$ flops [154], while the dense eigenvalue computation with a small $m \times m$ matrix in (iii) takes $O(m^3)$ flops [64, p. 391]. The complexity of one iteration of Algorithm 3 is therefore $O(mMN \log(N) + m^2 N + m^3)$ flops.

2.3.1 Computing high-frequency eigenmodes

Algorithm 3 adaptively and accurately resolves basis functions for highly oscillatory eigenmodes and preserves the sensitivity of the eigenvalues of the differential operator $\mathcal{L}$, so it is well-suited to computing high-frequency eigenmodes when $\mathcal{L}$ is self-adjoint or normal with respect to $(\cdot, \cdot)_{\mathcal{H}}$. We provide two ex-

amples from Sturm–Liouville theory to illustrate the effectiveness of the solve-then-discretize methodology in the high-frequency regime.

A regular Sturm–Liouville eigenvalue problem

First consider a regular Sturm–Liouville eigenvalue problem (SLEP) given by

$$-\frac{d^2u}{dx^2} + x^2u = \lambda \cosh(x)u, \qquad u(\pm 1) = 0. \tag{2.11}$$

This defines a self-adjoint differential operator with respect to the inner product

$$(v, u)_w = \int_{-1}^{1} \bar{v}u \cosh(x)\, dx. \tag{2.12}$$

Consequently, (2.11) possesses a complete $(\cdot,\cdot)_w$-orthonormal basis of eigenfunctions $u_1, u_2, u_3, \ldots$ for the weighted Hilbert space $\mathcal{H}_w = \{u : \|u\|_w = \sqrt{(u,u)_w} < \infty\}$ and an unbounded set of real eigenvalues $\lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots$.

Asymptotics for the large eigenvalues of (2.11) are given by [2]

$$\sqrt{\lambda_n} \sim \frac{n\pi}{\int_{-1}^{1} \sqrt{\cosh(x)}\, dx}, \qquad n \to \infty. \tag{2.13}$$

To accurately compute the large eigenvalues of (2.11) with Algorithm 3, we prescribe circular search regions with unit radius centered at the values given by the asymptotic formula in (2.13) (see Figure 2.3). Each search region contains one eigenvalue.

An indefinite Sturm–Liouville eigenvalue problem

Next, we consider the following indefinite SLEP:

$$-\frac{d^2u}{dx^2} = \lambda x^3 u, \qquad u(\pm 1) = 0, \tag{2.14}$$

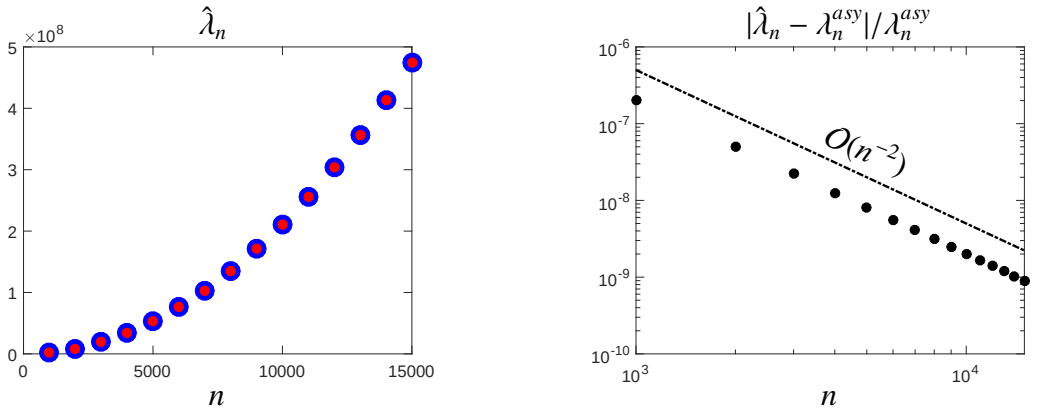

Figure 2.3: Left: The large eigenvalues of (2.11) are computed by `contFEAST` (see Algorithm 3) using search regions given by asymptotic estimates for the eigenvalues (2.13). Right: The relative difference $|\hat{\lambda}_n - \lambda_n^{asy}|/\lambda_n^{asy}$ between the eigenvalues $\hat{\lambda}_n$ computed by `contFEAST` and the asymptotic values λ_n^{asy} from (2.13). The difference is compared to an $O(n^{-2})$ relative error estimate [2].

which is closely related to models of light propagation in a nonhomogeneous material [2, 163]. Since the weight function x^3 changes sign at $x = 0$, (2.14) has a bi-infinite sequence of eigenvalues [7]. We index them in order as $\cdots \leq \lambda_{-2} \leq \lambda_{-1} < 0 < \lambda_1 \leq \lambda_2 \leq \cdots$. The asymptotics for the positive eigenvalues are given by [3]

$$\sqrt{\lambda_n} \sim \frac{(n - 1/4)\pi}{\int_0^1 x^{3/2}\, dx}, \qquad \lambda_n > 0, \qquad n \to \infty. \tag{2.15}$$

A similar expansion holds for the negative eigenvalues [3].

In contrast to the previous example, the indefinite weight function x^3 means that (2.14) is not immediately associated with a self-adjoint operator on a Hilbert space. Instead, (2.14) is usually studied through the lens of a Krein space and the eigenfunctions form a Riesz basis for the Hilbert space with the inner product [36]

$$(v, u)_{|w|} = \int_{-1}^{1} \bar{v} u |x|^3\, dx. \tag{2.16}$$

We use the leading order asymptotics in (2.15) to identify search regions that are likely to contain an eigenvalue of (2.14). Because the ultraspherical spectral

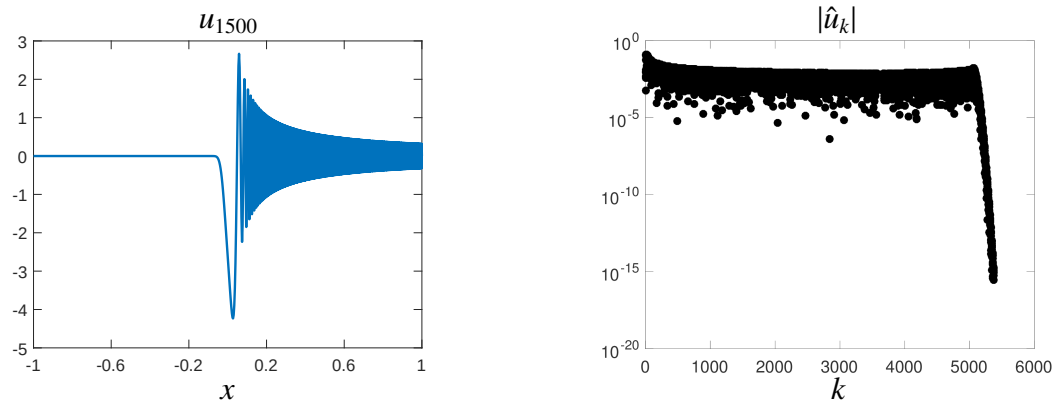

Figure 2.4: Left: The high-frequency eigenfunction associated to λ_{1500} of the indefinite Sturm–Liouville eigenvalue problem (2.14) computed by `contFEAST` (see Algorithm 3). Right: The Chebyshev coefficients $\{\hat{u}_k\}$ in a series expansion used to represent the eigenfunction. About 5371 Chebyshev coefficients are needed to accurately resolve the eigenfunction. The rapid decay in the coefficients to essentially machine precision is a good indication that the solution is fully resolved.

method used to solve the ODEs in step 2 of Algorithm 3 is efficient when applied to ODEs with smooth variable coefficients, it is convenient to treat (2.14) as a generalized eigenvalue problem, i.e., as $\mathcal{L}_1 u = \lambda \mathcal{L}_2 u$, where $\mathcal{L}_1 u = -\frac{d^2u}{dx^2}$ and $\mathcal{L}_2 u = x^3 u$. The eigenvalues of the pencil $z\mathcal{L}_2 - \mathcal{L}_1$ are then computed with a straightforward generalization of Algorithm 3 that is based on the spectral projector for the generalized eigenvalue problem, i.e.,

$$\mathcal{P}_{\mathcal{V}} = \frac{1}{2\pi i} \int_{\partial\Omega} (z\mathcal{L}_2 - \mathcal{L}_1)^{-1} \mathcal{L}_2 \, dz. \tag{2.17}$$

The eigenvalues and eigenfunctions are automatically resolved to essentially machine precision because of the use of the adaptive QR solver (see Figure 2.4).

2.4 Convergence and stability

The primary consequence of the approximations introduced in Algorithm 3 is that the spectral projector is no longer applied exactly. Therefore, the basis $\hat{Q}$

computed for the Rayleigh–Ritz projection is not an exact basis for the invariant subspace $\mathcal{V}$ of $\mathcal{L}$ and may require further refinement. Here, we view the iterative refinement procedure used in Algorithm 3 as a rational subspace iteration applied to a normal differential operator $\mathcal{L}$ in order to provide a preliminary analysis of the stability of the iteration and the sensitivity of the Ritz values. The main results may be summarized as follows.

(i) Algorithm 3 yields a sequence of quasimatrices $\hat{Q}_1, \ldots, \hat{Q}_k$ that (generically) converge geometrically to an orthonormal basis for the eigenspace $\mathcal{V}$ (see Theorem 2.4.1).

(ii) If $\hat{Q}_k$ is a sufficiently good approximation to an orthonormal basis for $\mathcal{V}$, then the ϵ-psuedospectrum of $\hat{Q}_k^* \mathcal{L} \hat{Q}_k$ is contained in the 2ϵ-psuedospectrum of $\mathcal{L}$ itself (see Theorem 2.4.2).

(iii) Under mild conditions on the initial quasimatrix F in Algorithm 3, the sequence $\|\mathcal{L}(\hat{Q}_k - Q)\|_{\mathbb{C}^m \to \mathcal{H}}$ is uniformly bounded as $k \to \infty$ (see Lemma 2.4.3).

Taken together, these results demonstrate that each iteration of Algorithm 3 yields uniformly consistent Ritz pairs that converge linearly to the desired eigenpair and that the unboundedness of $\mathcal{L}$ does not lead to instability. Note that this analysis does not take into account the impact of finite-precision arithmetic or the fact that the shifted differential equations in (2.9) are not solved exactly at each iteration (see the discussion at the end of section 2.4.1). However, (ii) ensures that the eigenvalues of the small matrix $\hat{Q}_k^* \mathcal{L} \hat{Q}_k$ are not much more sensitive than the eigenvalues of $\mathcal{L}$. Therefore, provided that the eigenvalue problem for $\mathcal{L}$ is well-conditioned and we compute a sufficiently accurate approximation to a basis for $\mathcal{V}$, then we expect that the eigenvalues computed with Algorithm 3 provide an accurate approximation to the desired eigenvalues

of $\mathcal{L}$.

2.4.1 Rational subspace iteration for differential operators

In analogy to the matrix case [144], Algorithm 3 may be interpreted as a filtered subspace iteration. Filtered subspace iteration is a variant of standard subspace iteration for computing a target subset of eigenvalues of a matrix A [124, Ch. 5]. The main idea is to choose a filter function $s(\cdot)$ that is large on the targeted eigenvalues of A and small on the unwanted eigenvalues of A. Applying the spectral transformation $s(A)$,[7] one uses standard subspace iteration to compute a basis for the eigenspace of $s(A)$ corresponding to its largest eigenvalues, i.e., the targeted eigenvalues of A. With an approximate basis for the eigenspace available, the eigenvalues and eigenvectors can be extracted with a Rayleigh–Ritz step.

From this perspective, Algorithm 3 computes the eigenvalues of $\mathcal{L}$ in Ω with the aid of a rational filter function induced by the quadrature rule in (2.8), i.e., $s(\cdot)$ is given by

$$s(z) = \sum_{k=1}^{\ell} \frac{w_k}{z_k - z}, \qquad z \in \mathbb{C} \setminus \{z_1, \ldots, z_l\}. \tag{2.18}$$

The functional calculus for unbounded normal operators ensures that if λ_i is an eigenvalue of $\mathcal{L}$ with eigenfunction u_i, then $s(\lambda_i)$ is an eigenvalue of $s(\mathcal{L})$ with eigenfunction u_i [119, VIII.5].[8] As the degree of the quadrature rule is in-

[7] A spectral transformation $s(\cdot)$ may be applied to A via the eigendecomposition of A, or more generally the Jordan decomposition. For example, if A has eigendecomposition $A = X\Lambda X^{-1}$ with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, then $s(A) = X\, s(\Lambda) X^{-1}$, where $s(\Lambda) = \operatorname{diag}(s(\lambda_1), \ldots, s(\lambda_n))$.

[8] The result [119, VIII.5] is stated for closed self-adjoint operators on $\mathcal{H}$, however, it extends immediately to closed normal operators on $\mathcal{H}$ if the spectral decomposition of a closed normal operator [122, Theorem 13.33] is used in place of the spectral decomposition of a self-adjoint

creased, the rational function becomes an increasingly good approximation to the Cauchy integral

$$\chi(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{dw}{w - z}, \qquad z \in \mathbb{C} \setminus \partial\Omega. \tag{2.19}$$

Therefore, the eigenvalues of $\mathcal{L}$ in Ω are usually $O(1)$ in size under the spectral transformation $s(\cdot)$ while the eigenvalues outside of Ω are much smaller.

We now turn to the convergence of the iteration described in Algorithm 3, which we interpret as a subspace iteration applied to the bounded linear operator $\hat{\mathcal{P}}_{\mathcal{V}} = s(\mathcal{L})$. It is helpful to introduce the notions of the *spectral radius* and a *dominant eigenspace* of a bounded linear operator $\mathcal{B}$. The spectral radius of a bounded linear operator $\mathcal{B}$ on a Hilbert space $\mathcal{H}$ is defined as [41, p. 99]

$$\rho(\mathcal{B}) = \max\{|z| : z \in \Lambda(\mathcal{B})\}. \tag{2.20}$$

The spectral radius is useful because it characterizes the asymptotic behavior of $\|\mathcal{B}^k\|_{\mathcal{H}}$, in the sense that $\rho(\mathcal{B}) = \lim_{k\to\infty} \|\mathcal{B}^k\|_{\mathcal{H}}^{1/k}$ [41, Theorem 4.1.3]. Let $\mathcal{V}$ be an invariant subspace of $\mathcal{B}$ associated with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ and a spectral projector $\mathcal{P}_{\mathcal{V}}$. We say that $\mathcal{B}$ has dominant eigenspace $\mathcal{V}$ if

$$\rho((\mathcal{I} - \mathcal{P}_{\mathcal{V}})\mathcal{B}) < |\lambda_m|. \tag{2.21}$$

The following theorem is an extension of a convergence analysis [124, p. 119] for matrix subspace iteration to the setting of bounded linear operators with a dominant eigenspace. We omit the details of the proof, as they are identical to those found in the proof of Lemma 3.1 and Lemma 3.2 of [66].

Theorem 2.4.1. *Let $\mathcal{B}$ be a bounded linear operator on a Hilbert space $\mathcal{H}$ with dominant eigenspace $\mathcal{V}$, defined in* (2.21)*, having* $\dim(\mathcal{V}) = m$*. Select a quasimatrix $F : \mathbb{C}^m \to \mathcal{H}$*

operator. For information on the spectral decomposition of unbounded normal operators and the associated functional calculus, see [122, Ch. 13] and [47].

such that the columns of $\mathcal{P}_{\mathcal{V}}F$ are linearly independent and suppose the columns of the quasimatrix $\hat{Q}_k : \mathbb{C}^m \to \mathcal{H}$ form an orthonormal basis for $\text{range}(\mathcal{B}^k F)$*, for $k = 1, 2, 3, \ldots$. If $u \in \mathcal{V}$ is an eigenvector of $\mathcal{B}$ with eigenvalue λ, then there is a function $\hat{u}_k \in \text{range}(\hat{Q}_k)$ such that*

$$\|\hat{u}_k - u\|_{\mathcal{H}} \leq (|\rho/\lambda| + \epsilon_k)^k \, \|(I - \mathcal{P}_{\mathcal{V}})Fx\|_{\mathcal{H}}, \qquad k = 1, 2, 3, \ldots,$$

where $\rho = \rho((\mathcal{I} - \mathcal{P}_{\mathcal{V}})\mathcal{B})$, $\epsilon_k \to 0$ as $k \to \infty$, and $u = \mathcal{P}_{\mathcal{V}}Fx$.

Although we have neglected the effects of approximately solving the ODEs in (2.9) and the impact of round-off errors in our brief analysis of rational subspace iteration for normal differential operators, we mention two recent results for rational subspace iteration with matrices [126] and self-adjoint differential operators [66, 67].

- For matrices, small errors made during application of the spectral projector generally do not alter the convergence behavior of subspace iteration [126]. In this case, the sequence $\hat{Q}_k$ no longer converges to an exact basis for $\mathcal{V}$. However, the matrices $\hat{Q}_k$ approximate a basis for $\mathcal{V}$ and the approximation error converges geometrically to a constant determined by the sizes of the errors introduced at each iteration [126].
- For self-adjoint differential operators (closed and densely defined on $\mathcal{H}$), rational subspace iteration converges to a subspace even when the resolvent operator is discretized to solve the ODEs in (2.9) [66, 67]. The distance between the computed subspace and the target eigenspace (in a distance metric between subspaces) is proportional to the approximation error in the discretized resolvent [66, 67].

We expect that similar statements hold for normal operators on $\mathcal{H}$, but a rigor-

ous and detailed convergence analysis is more subtle and beyond the scope of this chapter.

2.4.2 A pseudospectral inclusion theorem

As range($\hat{Q}$) is not an invariant subspace of $\mathcal{L}$, the ϵ-pseudospectrum of $\hat{Q}^*\mathcal{L}\hat{Q}$ is not, in general, contained in the ϵ-pseudospectrum of $\mathcal{L}$. However, if $\|\hat{Q} - Q\|_{\mathbb{C}^m \to \mathcal{H}}$ is sufficiently small, then the ϵ-pseudospectrum of $\hat{Q}^*\mathcal{L}\hat{Q}$ is contained in the 2ϵ-pseudospectrum of $\mathcal{L}$.

Theorem 2.4.2. *Consider a closed operator $\mathcal{L}$ with domain $\mathcal{D}(\mathcal{L})$ that is densely defined on a Hilbert space $\mathcal{H}$ and fix $\epsilon > 0$. Let $Q : \mathbb{C}^m \to \mathcal{D}(\mathcal{L}) \bigcap \mathcal{D}(\mathcal{L}^*)$ satisfy $Q^*Q = I$, and let* range(Q) *be an m-dimensional invariant subspace of $\mathcal{L}$. If $\hat{Q} : \mathbb{C}^m \to \mathcal{D}(\mathcal{L})$ satisfies*

$$\|\hat{Q} - Q\|_{\mathbb{C}^m \to \mathcal{H}} \left(\|\mathcal{L}^*Q\|_{\mathbb{C}^m \to \mathcal{H}} + \|\mathcal{L}Q\|_{\mathbb{C}^m \to \mathcal{H}} + \|\mathcal{L}(\hat{Q} - Q)\|_{\mathbb{C}^m \to \mathcal{H}} \right) < \frac{\epsilon}{2},$$

then $\Lambda_\epsilon(\hat{Q}^*\mathcal{L}\hat{Q}) \subset \Lambda_{2\epsilon}(\mathcal{L})$.

Proof. Consider $z \in \Lambda_\epsilon(\hat{Q}^*\mathcal{L}\hat{Q})$. If $z \in \Lambda_\epsilon(\hat{Q}^*\mathcal{L}\hat{Q}) \cap \Lambda_\epsilon(\mathcal{L})$, there is nothing to prove, so assume without loss of generality that $z \notin \Lambda_\epsilon(\mathcal{L})$. If we denote $R_Q(z) = (zI - Q^*\mathcal{L}Q)^{-1}$, $R_{\hat{Q}}(z) = (zI - \hat{Q}^*\mathcal{L}\hat{Q})^{-1}$, and $E = \hat{Q} - Q$, then we have that $R_{\hat{Q}}(z) = [R_Q(z)^{-1} - B]^{-1}$, where $B = Q^*\mathcal{L}E + E^*\mathcal{L}Q + E^*\mathcal{L}E$. Employing a formula for the inverse of the sum of two matrices, we obtain $R_{\hat{Q}}(z) = R_Q(z) + R_Q(z)[I - BR_Q(z)]^{-1}BR_Q(z)$ [77].

Now, $\|B\|_{\mathbb{C}^m} \leq \|Q^*\mathcal{L}E\|_{\mathbb{C}^m} + \|E^*\mathcal{L}Q\|_{\mathbb{C}^m} + \|E^*\mathcal{L}E\|_{\mathbb{C}^m}$. Since $\|E^*\|_{\mathcal{H} \to \mathbb{C}^m} = \|E\|_{\mathbb{C}^m \to \mathcal{H}}$ and $\|Q^*\mathcal{L}\|_{\mathcal{H} \to \mathbb{C}^m} = \|\mathcal{L}^*Q\|_{\mathbb{C}^m \to \mathcal{H}}$ [87, p. 256], our hypothesis indicates that the sum of the three terms comprising B is bounded in norm by

$$\|B\|_{\mathbb{C}^m} \leq \|E\|_{\mathbb{C}^m \to \mathcal{H}} \left(\|\mathcal{L}^*Q\|_{\mathbb{C}^m \to \mathcal{H}} + \|\mathcal{L}Q\|_{\mathbb{C}^m \to \mathcal{H}} + \|\mathcal{L}E\|_{\mathbb{C}^m \to \mathcal{H}} \right) < \frac{\epsilon}{2}.$$

Moreover, since $z \notin \Lambda_\epsilon(\mathcal{L})$, we have that $\|R_Q(z)\|_{\mathbb{C}^m} \leq 1/\epsilon$ by Theorem 2.2.2. Therefore, $\|BR_Q(z)\|_{\mathbb{C}^m} \leq 1/2$.

Because $\|BR_Q(z)\|_{\mathbb{C}^m} \leq 1/2$, we may use the Neumann series to compute $(I - BR_Q(z))^{-1} = \sum_{k=0}^{\infty}(BR_Q(z))^k$. We see that $R_{\hat{Q}}(z) = R_Q(z)\left(I + \sum_{k=1}^{\infty}(BR_Q(z))^k\right)$ and therefore,

$$\|R_{\hat{Q}}(z)\|_{\mathbb{C}^m} \leq \left(1 + \sum_{k=1}^{\infty}\frac{1}{2^k}\right)\|R_Q(z)\|_{\mathbb{C}^m} = 2\|R_Q(z)\|_{\mathbb{C}^m}.$$

Now, if $z \in \Lambda_\epsilon(\hat{Q}^*\mathcal{L}\hat{Q})$, then $\|R_Q(z)\|_{\mathbb{C}^m} \geq \|R_{\hat{Q}}(z)\|_{\mathbb{C}^m}/2 > 1/(2\epsilon)$. By Theorem 2.2.2, we have that $\|(zI - \mathcal{L})^{-1}\|_{\mathcal{H}} \geq \|R_Q(z)\|_{\mathbb{C}^m}$. Collecting inequalities yields the result $\|(zI - \mathcal{L})^{-1}\|_{\mathcal{H}} > 1/(2\epsilon)$, i.e., $z \in \Lambda_{2\epsilon}(\mathcal{L})$. □

A consequence of Theorem 2.4.2 is that Algorithm 3 possesses a type of stability provided that $\mathcal{L}$ is uniformly bounded on the sequence $E_1, E_2, E_3, \ldots$, where $E_k = \hat{Q}_k - Q$ for $k \geq 1$. If $\mathcal{L}$ is uniformly bounded on $\{E_k\}_{k=1}^{\infty}$, then there is a $\Lambda \geq 0$ such that $\sup_{k\geq 1}\|\mathcal{L}E_k\|_{\mathbb{C}^m\to\mathcal{H}} \leq \Lambda$. Applying Theorem 2.4.2, we see that Algorithm 3 computes elements in the 2ϵ-pseudospectrum of $\mathcal{L}$ provided that a basis for $\mathcal{V}$ is resolved to within $\epsilon/(2(\|\mathcal{L}^*Q\|_{\mathbb{C}^m\to\mathcal{H}} + \|\mathcal{L}Q\|_{\mathbb{C}^m\to\mathcal{H}} + \Lambda))$.

We now verify, with two mild constraints placed on the choice of the initial quasimatrix F, that $\mathcal{L}$ is uniformly bounded on the sequence $\{\hat{Q}_k\}_{k=1}^{\infty}$ generated by Algorithm 3. Note that this implies that $\mathcal{L}$ is uniformly bounded on $\{E_k\}_{k=1}^{\infty}$ because $E_k = \hat{Q}_k - Q$ and $\text{range}(Q) \subset \mathcal{D}(\mathcal{L})$. The constraints on F are generically satisfied when F is selected as in section 2.3. In the statement of the bound on $\|\mathcal{L}\hat{Q}_k\|_{\mathbb{C}^m\to\mathcal{H}}$, we use the notation $\sigma_{\min}(\mathcal{P}_{\mathcal{V}}F)$ and $\sigma_{\min}((\mathcal{I} - \mathcal{P}_{\mathcal{V}})F)$ to denote the smallest singular values of the quasimatrices $\mathcal{P}_{\mathcal{V}}F$ and $(\mathcal{I} - \mathcal{P}_{\mathcal{V}})F$, respectively.[9]

[9]The singular value decomposition of a quasimatrix $A : \mathbb{C}^m \to \mathcal{H}$ is the decomposition $A = U\Sigma V^*$, where $U : \mathbb{C}^m \to \mathcal{H}$ is a quasimatrix with $\mathcal{H}$-orthonormal columns, $\Sigma \in \mathbb{C}^{m\times m}$ is a diagonal matrix with non-negative entries $\sigma_1 \geq \cdots \geq \sigma_m$, and $V \in \mathbb{C}^{m\times m}$ is a unitary matrix [153].

Lemma 2.4.3. *Consider a closed, normal operator $\mathcal{L}$ with domain $\mathcal{D}(\mathcal{L})$ that is densely defined on a Hilbert space $\mathcal{H}$. Let $\hat{\mathcal{P}}_{\mathcal{V}}$ be the bounded operator on $\mathcal{H}$ defined in* (2.8) *and suppose that $\hat{\mathcal{P}}_{\mathcal{V}}$ has a dominant eigenspace of $\mathcal{V}$ (see* (2.21)*) with* $\dim(\mathcal{V}) = m$*. Let F, $\mathcal{P}_{\mathcal{V}}$, and $\{\hat{Q}_k\}_{k=1}^{\infty}$ be as in Theorem 2.4.1 with $\mathcal{B} = \hat{\mathcal{P}}_{\mathcal{V}}$. Suppose that $\hat{\mathcal{P}}_{\mathcal{V}}^k F$ (for each $k \geq 1$) and $(\mathcal{I} - \mathcal{P}_{\mathcal{V}})F$ each have linearly independent columns and that* $\text{range}(F) \subset \mathcal{D}(\mathcal{L})$*. Then, we have that*

$$\|\mathcal{L}\hat{Q}_k\|_{\mathbb{C}^m \to \mathcal{H}} \leq 2M\|\mathcal{L}F\|_{\mathbb{C}^m \to \mathcal{H}}, \qquad k = 1, 2, 3 \ldots,$$

where $M = \max\{1/\sigma_{\min}(\mathcal{P}_{\mathcal{V}}F), 1/\sigma_{\min}((\mathcal{I} - \mathcal{P}_{\mathcal{V}})F)\}$.

Proof. Since $\hat{Q}_k$ is an orthonormal basis for $\hat{\mathcal{P}}_{\mathcal{V}}^k F$, there is a matrix $R_k \in \mathbb{C}^{m \times m}$ such that $\hat{\mathcal{P}}_{\mathcal{V}}^k F = \hat{Q}_k R_k$. By the assumption that $\hat{\mathcal{P}}_{\mathcal{V}}^k F$ has linearly independent columns, we know that R_k is invertible. We obtain that

$$\hat{Q}_k = \hat{\mathcal{P}}_{\mathcal{V}}^k F R_k^{-1}. \tag{2.22}$$

We use the spectral projector $\mathcal{P}_{\mathcal{V}}$ to rewrite (2.22) as

$$\hat{Q}_k = \hat{\mathcal{P}}_{\mathcal{V}}^k \left(\mathcal{P}_{\mathcal{V}}F + (\mathcal{I} - \mathcal{P}_{\mathcal{V}})F\right) R_k^{-1}. \tag{2.23}$$

Now, $\text{range}(\mathcal{P}_{\mathcal{V}}F)$ and $\text{range}((\mathcal{I} - \mathcal{P}_{\mathcal{V}})F)$ are invariant under $\hat{\mathcal{P}}_{\mathcal{V}}$ [87, p. 178]. Consequently, there are matrices $D_1, D_2 \in \mathbb{C}^{m \times m}$ such that

$$\hat{\mathcal{P}}_{\mathcal{V}}^k \mathcal{P}_{\mathcal{V}}F = \mathcal{P}_{\mathcal{V}}FD_1^k, \qquad \hat{\mathcal{P}}_{\mathcal{V}}^k(\mathcal{I} - \mathcal{P}_{\mathcal{V}})F = (\mathcal{I} - \mathcal{P}_{\mathcal{V}})FD_2^k. \tag{2.24}$$

Substituting (2.24) into (2.23) yields the following useful equation for $\hat{Q}_k$:

$$\hat{Q}_k = \left(\mathcal{P}_{\mathcal{V}}FD_1^k + (\mathcal{I} - \mathcal{P}_{\mathcal{V}})FD_2^k\right) R_k^{-1}. \tag{2.25}$$

Applying $\mathcal{L}$ to both sides of (2.25) and commuting with the spectral projectors $\mathcal{P}_{\mathcal{V}}$ and $\mathcal{I} - \mathcal{P}_{\mathcal{V}}$ [87, p. 179], we obtain

$$\mathcal{L}\hat{Q}_k = \left(\mathcal{P}_{\mathcal{V}}\mathcal{L}FD_1^k + (\mathcal{I} - \mathcal{P}_{\mathcal{V}})\mathcal{L}FD_2^k\right) R_k^{-1}. \tag{2.26}$$

Since range$(F) \subset \mathcal{D}(\mathcal{L})$, we have that $\|\mathcal{L}F\|_{\mathbb{C}^m \to \mathcal{H}} < \infty$. Additionally, since $\mathcal{L}$ is normal, the spectral projectors have norms equal to 1 [87, p. 277]. Therefore, it remains to find a uniform bound for $\|D_1^k R_k^{-1}\|_{\mathbb{C}^m}$ and $\|D_2^k R_k^{-1}\|_{\mathbb{C}^m}$ as $k \to \infty$.

For brevity, we prove uniform boundedness of $\|D_1^k R_k^{-1}\|_{\mathbb{C}^m}$ and note that the proof for $\|D_2^k R_k^{-1}\|_{\mathbb{C}^m}$ is essentially identical. We begin by commuting $\hat{\mathcal{P}}_{\mathcal{V}}$ with the spectral projectors in (2.24) and substituting the QR factorization of $\hat{\mathcal{P}}_{\mathcal{V}}^k F$ to see that

$$\mathcal{P}_{\mathcal{V}} \hat{Q}_k = (\mathcal{P}_{\mathcal{V}} F) D_1^k R_k^{-1}. \tag{2.27}$$

Using the psuedoinverse $(\mathcal{P}_{\mathcal{V}} F)^+$ of the quasimatrix[10] $\mathcal{P}_{\mathcal{V}} F$ and noting that $\mathcal{P}_{\mathcal{V}} F$ has linearly independent columns, (2.27) implies that

$$D_1^k R_k^{-1} = (\mathcal{P}_{\mathcal{V}} F)^+ \mathcal{P}_{\mathcal{V}} \hat{Q}_k. \tag{2.28}$$

Now, we know that $\|\mathcal{P}_{\mathcal{V}} \hat{Q}_k\|_{\mathcal{H}} \leq 1$, because $\|\mathcal{P}_{\mathcal{V}}\|_{\mathcal{H}} = 1$ and $\hat{Q}_k$ has orthonormal columns. We conclude that

$$\|D_1^k R_k^{-1}\|_{\mathbb{C}^m} \leq \frac{1}{\sigma_{\min}(\mathcal{P}_{\mathcal{V}} F)}. \tag{2.29}$$

A similar argument shows that

$$\|D_2^k R_k^{-1}\|_{\mathbb{C}^m} \leq \frac{1}{\sigma_{\min}((\mathcal{I} - \mathcal{P}_{\mathcal{V}}) F)}. \tag{2.30}$$

Taking norms in (2.26) and substituting the bounds from (2.29) and (2.30), we find

$$\|\mathcal{L}\hat{Q}_k\|_{\mathbb{C}^m \to \mathcal{H}} \leq \|\mathcal{L}F\|_{\mathbb{C}^m \to \mathcal{H}} \left(\frac{1}{\sigma_{\min}(\mathcal{P}_{\mathcal{V}} F)} + \frac{1}{\sigma_{\min}((\mathcal{I} - \mathcal{P}_{\mathcal{V}}) F)} \right). \tag{2.31}$$

The lemma follows immediately from (2.31). □

[10] The pseudoinverse of a quasimatrix $A : \mathbb{C}^m \to \mathcal{H}$ may be defined via the SVD as $A^+ = V \Sigma^+ U^*$, where Σ^+ is the diagonal matrix with entries $\Sigma_{ii}^+ = 1/\sigma_i$ if $\sigma_i \neq 0$ and 0 otherwise. It is easy to verify familiar properties from the matrix case [64, p. 290], i.e., if A has linearly independent columns, then $A^+ A = I$ and $\|A^+\|_{\mathcal{H} \to \mathbb{C}^m} = 1/\sigma_{\min}(A)$.

Theorem 2.4.1, Theorem 2.4.2, and Lemma 2.4.3 provide a preliminary analysis to explain why Algorithm 3 accurately computes the eigenvalues of normal operators with a dominant eigenspace $\mathcal{V}$. Theorem 2.4.1 allows us to accurately resolve an orthonormal basis Q for $\mathcal{V}$ by refining the quasimatrix $\hat{Q}_k$ with subspace iteration. Lemma 2.4.3 confirms that $\mathcal{L}\hat{Q}_k$ does not grow without bound as $\hat{Q}_k$ is refined. Finally, Theorem 2.4.2 demonstrates that the eigenvalues are computed to the expected accuracy, provided that the basis for $\mathcal{V}$ has been resolved.

2.5 An operator analogue of the Rayleigh Quotient Iteration

It is useful to have operator analogues for other eigensolvers too; particularly, when the eigenvalues of interest are difficult to target with a pre-selected search region $\Omega \subset \mathbb{C}$. The Rayleigh Quotient Iteration (RQI) is a generalization of the inverse iteration that incorporates dynamic shifting to obtain cubic (for Hermitian problems) or quadratic (non-Hermitian problems) convergence [109]. Given a matrix $A \in \mathbb{C}^{n\times n}$ and an initial vector $\tilde{y}_0 \in \mathbb{C}^n$, RQI computes the iterates

$$\tilde{y}_{k+1} = (A - \beta_k I)^{-1} y_k, \qquad \beta_k = y_k^* A y_k, \qquad y_k = \frac{\tilde{y}_k}{\|\tilde{y}_k\|_2}, \qquad k = 0, 1, 2, \ldots. \tag{2.32}$$

The vectors y_k typically converge to a nearby eigenvector of A, while the sequence β_k converges to the associated eigenvalue of A [107]. In the matrix setting, (2.32) is often used to compute interior eigenvalues or refine an estimate of an invariant subspace [108, 109].

Replacing a matrix A by a differential operator $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$, as in (1.1), and the vectors $\tilde{y}_k$ by functions $f_k \in \mathcal{D}(\mathcal{L})$, we obtain an operator analogue of RQI. One needs to select an initial function $f_0 \in \mathcal{D}(\mathcal{L})$ and solve a sequence of

ODEs, i.e.,

$$(\mathcal{L} - \beta_k I) f_{k+1} = f_k, \qquad f_{k+1}(\pm 1) = \cdots = f_{k+1}^{(N/2)}(\pm 1) = 0. \tag{2.33}$$

At each iteration, the shift β_k is computed from the Rayleigh Quotient $(f_k, \mathcal{L} f_k)_{\mathcal{H}}$ (in strong form) and the solution f_{k+1} is normalized after each iteration. Analogous to the matrix setting, we observe that the operator analogue of the Rayleigh Quotient Iteration converges cubically for self-adjoint operators and quadratically otherwise [71].

We note that block generalizations of RQI (RSQR and GRQI [1]) are also easily extended to the differential operator setting. In this case, a sequence of quasimatrices $\hat{Q}_k$ with $\mathcal{H}$-orthonormal columns are generated to approximate an invariant subspace of $\mathcal{L}$ and a Rayleigh–Ritz projection is performed to compute approximate eigenvalues and eigenvectors. As with the operator analogue of FEAST, Theorem 2.4.2 implies that the iteration (2.33) accurately computes eigenvalues of normal differential operators when the basis for the target eigenspace is sufficiently resolved.

2.5.1 Free vibrations of an airplane wing

The improved convergence rate of RQI can offer much faster computation time than subspace iteration, often requiring only 3 or 4 ODE solves to reach an accuracy of essentially machine precision [71]. We now employ (2.33) for the rapid computation of vibrational modes of an airplane wing.

An airplane wing may be crudely modeled as a thin, cantilevered beam of

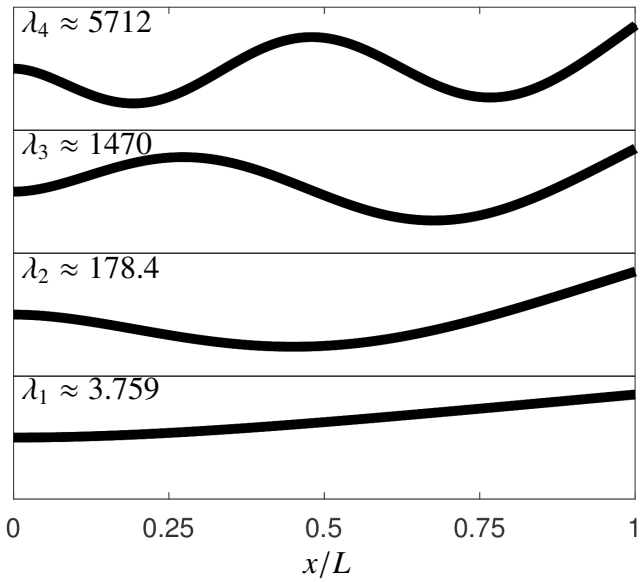

Figure 2.5: Selected free-vibration modes of an airplane wing modeled by (2.34).

length L with a linear taper. The governing equation for free vibrations is [74]

$$\frac{d^2}{dx^2}\left((1+x)\frac{d^2u}{dx^2}\right) = \lambda u, \quad u(0) = u'(0) = 0, \quad u''(L) = u'''(L) = 0. \tag{2.34}$$

The variable coefficient $1 + x$ accounts for the linear taper of the wing, while the boundary conditions on u'' and u''' at $x = 1$ express the natural requirement that the bending moment and shear force vanish at the endpoint.

To compute a few of the smoothest modes of (2.34) we use the eigenfunctions w_n of the cantilevered beam equation with constant coefficients, given in closed form by [74]

$$w_n(x) = \cosh\beta_n x - \cos\beta_n x + \frac{\cos\beta_n L + \cosh\beta_n L}{\sin\beta_n L + \sinh\beta_n L}(\sin\beta_n x + \sinh\beta_n x). \tag{2.35}$$

Here β_n is the nth root of $g(\beta) = \cosh(\beta L)\cos(\beta L) + 1$ [74]. We target a mode of (2.34) by setting $f_0(x) = w_n(x)$. Figure 2.5 shows the modes that are computed using initial guesses $w_1, \ldots, w_4$, corresponding to the smallest four positive roots of $g(\beta)$.

2.6 Computing eigenvalues in unbounded regions

Stability analyses of solutions to time-dependent partial differential equations (PDEs) provide an abundant source of differential eigenvalue problems (recall the example in section 1.2). When $\mathcal{L}$ is a self-adjoint or normal operator, linear stability analysis often reduces to determining whether or not the eigenvalues of $\mathcal{L}$ are contained in one half-plane [4,94,128,157]. We now show how to modify the spectral projector in (2.4) to derive a practical rational filter to compute (finitely many) eigenvalues of $\mathcal{L}$ in the right half-plane.

2.6.1 A rational filter for the half-plane

Suppose that $\mathcal{L}$ is a normal operator with a spectrum in the left half-plane $\mathrm{Re}(z) < 0$ except for finitely many eigenvalues $\lambda_1, \ldots, \lambda_m$ (including multiplicities) such that $\mathrm{Re}(\lambda_i) > 0$ for $1 \leq i \leq m$. Denote the eigenspace associated with $\lambda_1, \ldots, \lambda_m$ by $\mathcal{V}$ and consider search regions that are semi-circles of radius R, i.e.,

$$\Omega_R = \{z \in \mathbb{C} : |z| < R, \mathrm{Re}(z) > 0\}, \qquad R > \max_{1 \leq i \leq m} |\lambda_i|. \tag{2.36}$$

To construct a computable spectral projector onto the right half-plane we consider taking $R \to \infty$. We adopt the following strategy:

(i) Introduce a $1/R$ decay into the integrand of the spectral projector (2.4) as $R \to \infty$, while preserving the projection onto $\mathcal{V}$.

(ii) Split the projector into an integral over the vertical part of $\partial\Omega_R$ and an integral over the circular arc of $\partial\Omega_R$. By taking $R \to \infty$, we observe that the contribution from the circular arc goes to 0 due to the additional $1/R$ decay in the integrand.

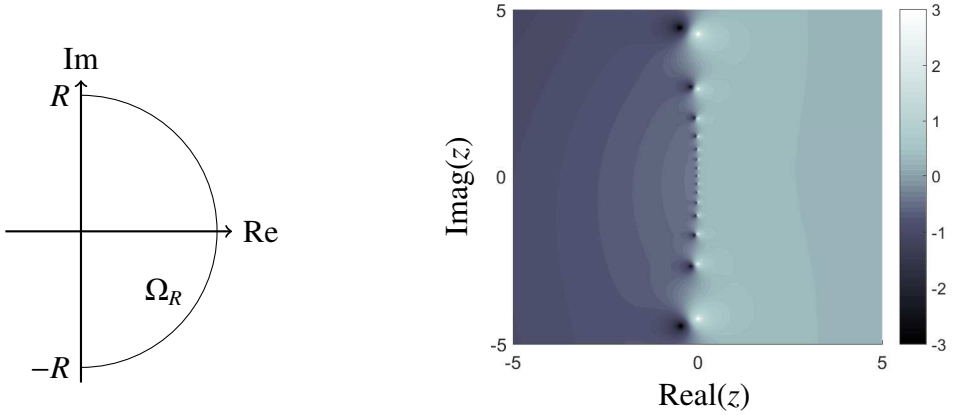

Figure 2.6: Left: The region Ω_R from (2.36) used in the derivation of the rational filter over the right half-plane. Right: The constructed rational filter (2.40) for the right half-plane with $\ell = 20$.

(iii) Map the imaginary axis to the interval $[-1, 1]$ and approximate the spectral projector by a quadrature rule.

Select $a \in \mathbb{R}^+$ and consider the family of functions that are analytic in the right half-plane defined by

$$\mathcal{P}_R(\lambda) = \frac{1}{2\pi i} \int_{\partial\Omega_R} (z + a)^{-1}(z - \lambda)^{-1}\, dz. \tag{2.37}$$

By Cauchy's Integral Formula, we know that $\mathcal{P}_R(\lambda) = (\lambda + a)^{-1}$ if $\lambda \in \Omega_R$ and is zero otherwise [134]. Taking the limit $R \to \infty$, we obtain

$$\mathcal{P}(\lambda) = \lim_{R\to\infty} \mathcal{P}_R(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (iy + a)^{-1}(iy - \lambda)^{-1}\, dy. \tag{2.38}$$

Using functional calculus for unbounded normal operators we can extend $\mathcal{P}(\lambda)$ to an operator-valued function $\mathcal{P}(\mathcal{L})$ [122, Theorem 13.24].[11] Moreover, we have that $\mathcal{P}(\mathcal{L})u = \mathcal{P}(\lambda)u$ when $\mathcal{L}u = \lambda u$ [119, VIII.5]. Consequently, range$(\mathcal{P}(\mathcal{L})) = \mathcal{V}$.

Now, take the change-of-variables $x = \frac{2}{\pi} \tan^{-1} y$ in (2.38) to obtain

$$\mathcal{P}(\mathcal{L}) = \frac{1}{4} \int_{-1}^{1} \left(i \tan\left(\frac{\pi x}{2}\right) + a \right)^{-1} \left(i \tan\left(\frac{\pi x}{2}\right) \mathcal{I} - \mathcal{L} \right)^{-1} \sec^2\left(\frac{\pi x}{2}\right) dx. \tag{2.39}$$

[11] In [122, Theorem 13.24], $E_{x,y}$ is the spectral measure of $\mathcal{L}$ [122, Theorem 13.33].

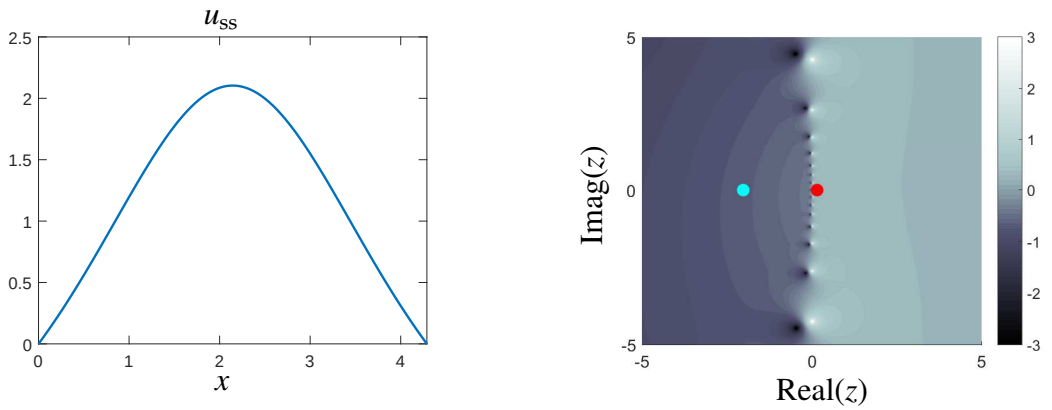

Figure 2.7: Left: A droplet, u_{ss}, which is a steady-state solution to (2.41), computed from the IVP (2.43). Right: Two rightmost eigenvalues (blue and red dots) of (2.42) together with a log-scale colormap of the rational filter in (2.40) with $\ell = 20$, which is used in place of (2.8). The eigenvalue with a positive real part (red dot) indicates that this steady-state droplet is unstable.

Using Gauss–Legendre quadrature nodes $x_1, \ldots, x_\ell$ and weights $w_1, \ldots, w_\ell$ on $[-1, 1]$, we can approximate $\mathcal{P}(\mathcal{L})$ by

$$\hat{\mathcal{P}}(\mathcal{L}) = \frac{1}{4} \sum_{k=1}^{\ell} w_k \frac{1 - z_k^2}{z_k + a} (z_k \mathcal{I} - \mathcal{L})^{-1}, \qquad z_k = i \tan\left(\frac{\pi x_k}{2}\right). \tag{2.40}$$

Figure 2.6 (right) shows the derived rational filter $\hat{\mathcal{P}}(\lambda)$ in the complex plane.

2.6.2 Stability of thin fluid films

To demonstrate the utility of the filter in (2.40), we assess the stability of the steady-state solutions to a PDE governing the motion of a thin film of fluid supported below by a flat substrate (see section 1.2). The PDE is

$$u_t = \partial_x^4 u + \partial_x (u \partial_x u), \tag{2.41}$$

where $u(x, t)$ is a positive, periodic function representing the thickness of the fluid [95]. The nonlinear term models gravitational effects and substrate-fluid interactions [95].

A droplet steady-state $u_{\text{ss}}(x)$ of (2.41), rescaled so that it is supported on $[0, l]$ with contact angle $\pi/4$, is stable if all the eigenvalues of a fourth-order differential operator are in the left half-plane. The associated differential eigenproblem is [94]

$$\frac{d^4u}{dx^4} + \frac{d}{dx}\left(u_{\text{ss}}\frac{du}{dx}\right) = \lambda u, \qquad u(0) = u(l) = 0, \quad u''(0) = u''(l) = 0. \tag{2.42}$$

We compute the steady-state $u_{\text{ss}}(x)$ by solving the second order nonlinear ODE [95]

$$\frac{du_{\text{ss}}}{dx} + \frac{1}{2}u_{\text{ss}}^2 - \delta = 0, \qquad u_{\text{ss}}(0) = 0, \quad u'_{\text{ss}}(0) = 1. \tag{2.43}$$

Here, δ is a dimensionless quantity relating the rescaled problem to the original contact angle [95]. The length l of the droplet's base and δ may be calculated analytically [95].

In Figure 2.7, we show an approximation to the rescaled steady-state u_{ss} along with the right-most eigenvalues of (2.42). Using the rational filter in (2.40) with $\ell = 20$ (the degree of the quadrature rule defining the filter) to perform the approximate spectral projection in Algorithm 3, we are able to identify an eigenvalue of (2.42) in the right half-plane, which indicates that the droplet (see Figure 2.7 (left)) is unstable.

Techniques for selecting the dimension m of the subspace $\mathcal{V}$ [89, 144] are important in stability analysis as one is trying to determine the number of eigenvalues in the right half-plane. To select m, we monitor the singular values of the matrix $\hat{V}^*\hat{V}$ after each iteration and adjust the number of basis functions by removing columns of $\hat{V}$ associated with singular values that are close to machine precision (relative to the largest singular value) [144]. This procedure usually allows us to capture the dominant eigenspace of the filtered operator $\hat{\mathcal{P}}(\mathcal{L})$ that includes the target eigenspace as well as any eigenvalues clustered

near the imaginary axis. We then determine whether there are any eigenvalues in the right half-plane by sorting through the computed eigenvalues. However, this strategy may break down, for instance, if there is an eigenvalue close to a quadrature node. Additionally, the sharp decay of the filter (2.40) across the imaginary axis is softened as $|\mathrm{Im}(z)| \to \infty$, which can lead to difficulties when there are clusters of eigenvalues near the imaginary axis with large imaginary part. In this case, one may need to take a large number of basis functions to accurately resolve the dominant eigenvalues of $\hat{\mathcal{P}}(\mathcal{L})$.

CHAPTER 3

STABILITY OF CONTOUR INTEGRAL EIGENSOLVERS

This chapter[1] examines the numerical stability of the FEAST matrix eigensolver and, more generally, eigensolvers that use rational filters of the form

$$r(A) = \sum_{j=1}^{\ell} \omega_j (z_j I - A)^{-1}, \tag{3.1}$$

to target interior eigenvalues of a large matrix $A \in \mathbb{C}^{n\times n}$ [9,70,83,88,116,127,144]. When a pole z_{j_*} in (2.18) is close to an eigenvalue of A, computing the product $r(A)x$ for some $x \in \mathbb{C}^n$ requires the numerical solution of an ill-conditioned linear system. Consequently, round-off errors may pollute the basis computed for the target space, and the associated eigenpairs may be computed inaccurately. We now show that iteratively refining the eigenpairs, as in section 2.4.1, plays an important role when such *dangerous eigenvalues* are present.

Throughout the chapter, $\|\cdot\|$ denotes the spectral norm of a matrix (Euclidean norm for vectors) and A denotes an $n\times n$ diagonalizable matrix with eigenvalues and eigenvectors satisfying $Av_i = \lambda_i v_i$, for $1 \le i \le n$. Except in section 3.6, we assume that A has a complete orthonormal set of eigenvectors (i.e., A is normal), in which case it is convenient to write the eigendecomposition of A in the form

$$A = V_1 \Lambda_1 V_1^* + V_2 \Lambda_2 V_2^*. \tag{3.2}$$

Here, $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ contains a set of target eigenvalues that we wish to compute and $\Lambda_2 = \mathrm{diag}(\lambda_{m+1}, \ldots, \lambda_n)$ contains the remaining unwanted eigenvalues (usually, $m \ll n$). We denote the target eigenspace by $\mathcal{V} = \mathrm{span}(V_1)$,

[1] This chapter is based on a paper with Yuji Nakatsukasa [82], who suggested this project after we uncovered the 'twice-is-enough' phenomenon in a series of numerical experiments in July 2019. I was the lead author, and I developed the main theoretical results and numerical experiments with Yuji's guidance during weekly meetings.

the full eigenvalue matrix by $\Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2)$, and the eigenvector matrix by $V = [V_1\ V_2]$. Our analysis is focused on a single dangerous eigenvalue λ located at a distance $d = |z_{j^*} - \lambda| \ll 1$ from a pole of the filter in (3.1) (we discuss numerical experiments with clusters of dangerous eigenvalues in section 3.8).

For simplicity, we always assume that $r(\Lambda)$ is invertible, that there is a nonzero spectral gap between $r(\Lambda_1)$ and $r(\Lambda_2)$, and index the eigenvalues in order of decreasing modulus under the filter so that

$$|r(\lambda_1)| \geq \cdots \geq |r(\lambda_m)| > |r(\lambda_{m+1})| \geq \cdots \geq |r(\lambda_n)|. \tag{3.3}$$

Here, $r(\lambda) = \sum_{j=1}^{\ell} \omega_j (z_j - \lambda)^{-1}$ is the scalar form of the filter in (3.1).[2] Under the ordering in (3.3), the dangerous eigenvalue is λ_1. Without loss of generality, we assume that the weight w_j associated with a pole near the dangerous eigenvalue λ_1 is equal to one (by scaling $r(\cdot)$ if necessary). This simplifies the analysis and usually implies that the other weights w_i are also modest in size.

3.1 The power of iteration

When combined with shift-and-invert enhancement, subspace iteration and Arnoldi are two classic iterative schemes for computing a few interior eigenvalues of an $n \times n$ matrix A. Each method constructs an orthonormal basis for a search subspace by iteratively applying the spectral filter

$$s(A) = (zI - A)^{-1} \tag{3.4}$$

to a set of vectors. Approximate eigenpairs can then be extracted from the search subspace with a projection step, e.g., Rayleigh–Ritz. The shift z is selected to

[2] We refer to the scalar function $r(z)$ and its matrix companion $r(A)$ with the same symbol. We always include the argument when it is necessary to clarify which we mean.

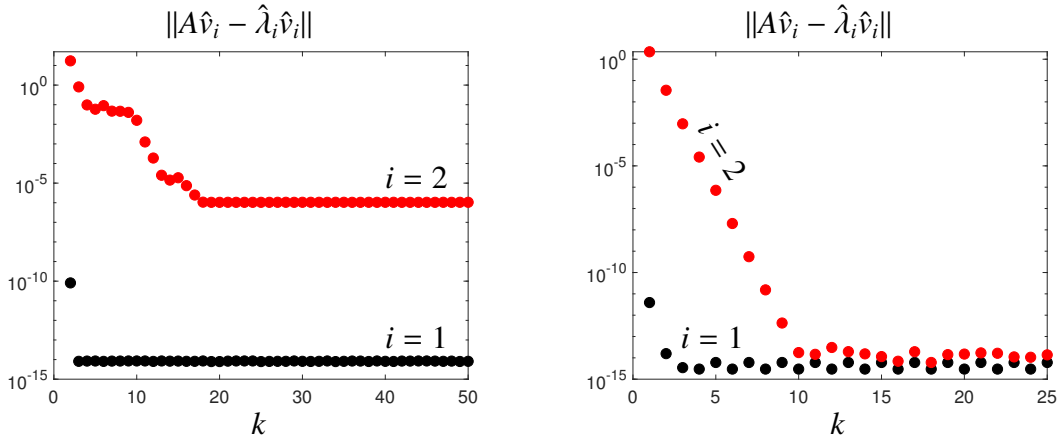

Figure 3.1: The residuals for two approximate eigenpairs of a real-symmetric 100×100 matrix at iterations $k = 2, \ldots, 50$ of Arnoldi (left) and iterations $k = 1, \ldots, 25$ of subspace iteration (right), both with shift-and-invert enhancement. The approximate eigenpairs correspond to a dangerous eigenvalue (black) with $|z - \lambda_1| = 10^{-12}$ and a second target eigenvalue (red) with $|z - \lambda_2| \approx 0.1$.

target a region of interest, and both methods typically approximate eigenvalues of A closest to z.

In his 2001 volume on matrix algorithms for eigenvalue problems, Stewart noted that shift-and-invert Arnoldi encounters difficulties in floating-point arithmetic when the shift lies too close to an eigenvalue of A [139, p. 309]. Although the eigenvalue adjacent to the shift is rapidly approximated to the order of the unit round-off u, the residuals of other computed eigenpairs stagnate near the order of u/d, where d is the distance between the "dangerous" eigenvalue and the shift. This phenomenon has also recently been observed in the context of Krylov methods, where the subspace is constructed with contour integrals or rational approximations [9].

Curiously, dangerous eigenvalues do not inflict the same stagnation in the residuals of the other target eigenpairs during subspace iteration. Figure 3.1 compares the residuals of two target eigenpairs computed with Arnoldi (left) and subspace iteration (right), using the shift-and-invert filter in (3.4) with

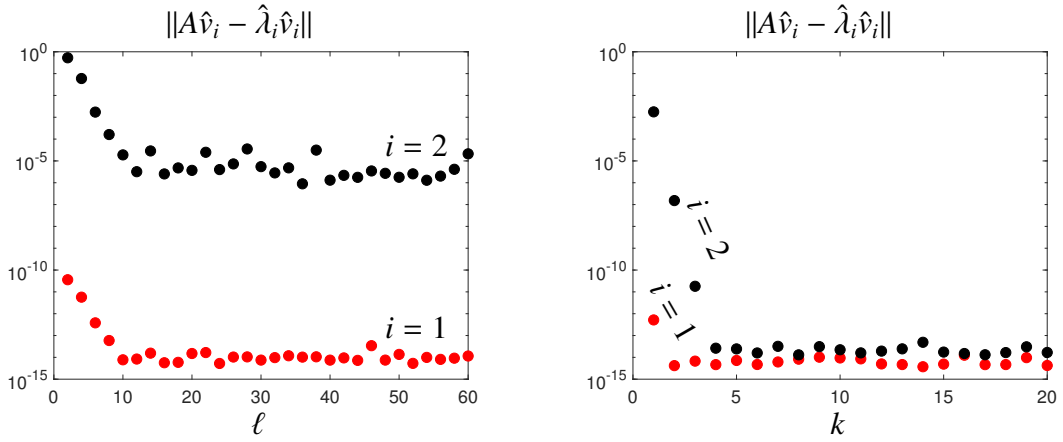

Figure 3.2: The left panel displays the residuals for two approximate eigenpairs of a 100×100 real-symmetric matrix computed with the contour integral eigensolver described in [144], as the quadrature rule approximating the contour integral is refined. One of the target eigenvalues ($i = 1$, red) is a distance of 10^{-10} from the contour. The right panel displays the residuals for the two approximate eigenpairs when refined via iteration rather than quadrature rule. The quadrature rule used corresponds to a rational filter in (3.1) with $\ell = 8$.

$z = 10$. The approximation to the dangerous eigenvalue $\lambda_1 = 10+10^{-12}$ converges rapidly to unit round-off accuracy in both cases. However, only subspace iteration computes an approximation to the second target eigenvalue $\lambda_2 \approx 10.1$ to unit round-off accuracy.

A similar story unfolds in Figure 3.2, where we compute two target eigenpairs with the contour integral eigensolver described in [144], one of them located at a distance of 10^{-10} from the contour. As we refine the quadrature along the contour, the poles of a rational filter with form (3.1) cluster near the dangerous eigenvalue, and we observe the residual of the dangerous eigenpair converge rapidly to unit round-off, while the residuals of the remaining target pairs stagnate near 10^{-5}. On the other hand, if we fix the number of quadrature points (i.e., poles) and refine via filtered subspace iteration, the residuals of all target eigenpairs converge geometrically to order u.

This chapter is about explaining Figures 3.1 and 3.2. We first examine how

rational filtered subspace iteration disarms dangerous eigenvalues after the first iteration. When A has a complete set of orthonormal eigenvectors, orthogonal bases for the search subspace play a special role and "twice-is-enough" to recover full precision in the computed iterates (see sections 3.3 to 3.5).[3] In the non-normal case, iterating on approximate eigenvectors (obtained from a Rayleigh–Ritz step, for instance) is the key to overcoming round-off errors incurred by the dangerous eigenvalue, while iterations based on orthogonal bases (such as approximate Schur vectors) suffer stagnation in the remaining target eigenpairs (see section 3.6). Our analysis generalizes that of Peters and Wilkinson [115], who demonstrated that single-vector inverse iterations converge rapidly when the shift is very close to an eigenvalue, to the substantially more complex case of subspace iteration and rational filters with multiple poles.

To obtain full precision in the remaining target eigenpairs for Arnoldi and related Krylov schemes, the prevailing consensus is to alter the rational filter by moving or removing the offending poles [9, 139]. Unfortunately, this usually means settling for a less efficient filter or starting over with a new filter. Informed by our analysis of subspace iteration and its immunity to dangerous eigenvalues, we offer simple restart strategies that fix stagnation in shift-and-invert Arnoldi (see section 3.7).

[3] Aspects of our analysis are similar to Parlett and Kahan's "twice-is-enough" algorithm and analysis for Gram-Schmidt reorthogonalization [110, pp. 107–109].

3.2 Principle angles between subspaces

Given an $n \times m$ matrix Q_0 with orthonormal columns, the simplest practical form of subspace iteration with a rational filter, as in (3.1), computes the iterates

$$X_k = r(A)Q_{k-1}, \qquad Q_k = \mathrm{qf}(X_k). \tag{3.5}$$

Here, $\mathrm{qf}(X_k)$ denotes the orthogonal factor from a QR decomposition of X_k and we write $\mathcal{S}_k = \mathrm{span}(Q_k)$. As in section 2.1, the eigenpairs of $Q_k^* A Q_k$ provide approximations to the target eigenpairs of A.

In practice, there are many modifications one can make to (3.5) to improve convergence, enhance stability, or increase computational efficiency. Nevertheless, when A is normal, (3.5) is enough to capture both the dangers and the self-correcting effects of eigenvalues that are close to the poles in (3.1). When A is non-normal, iterations that incorporate the Ritz vectors when forming Q_{k-1} play a special role, while other variants (including (3.5) itself) typically fail to converge to full precision (see Figure 3.7). We discuss these modifications further in section 3.6.

The principal angles between the subspaces $\mathcal{S}_k$ and $\mathcal{V}$ provide a natural framework with which to characterize the refinement of the iterates in (3.5). Generalizing the notion of an angle between two vectors, the principal angles tell us how close $\mathcal{S}_k$ and $\mathcal{V}$ are in a geometric sense [18].

Definition 3.2.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two m-dimensional subspaces with orthonormal bases X and Y, respectively, and let $\sigma_i(Y^*X)$ denote the ith singular value of Y^*X. The principal angles between $\mathcal{X}$ and $\mathcal{Y}$ are the acute angles $\theta_1(\mathcal{X}, \mathcal{Y}) \geq \cdots \geq \theta_m(\mathcal{X}, \mathcal{Y})$ satisfying*

$$\cos \theta_i(\mathcal{X}, \mathcal{Y}) = \sigma_{m+1-i}(Y^*X), \qquad i = 1, \ldots, m. \tag{3.6}$$

The sine of the largest principal angle, given by $\sin\theta_1(\mathcal{X},\mathcal{Y}) = \|(I - P_{\mathcal{Y}})X\|$, defines a metric on the set of m-dimensional subspaces. However, the tangents of the principal angles, which are the singular values of the matrix [169]

$$T(X,Y) = (I - P_{\mathcal{Y}})X(Y^*X)^+, \tag{3.7}$$

are better equipped to describe the behavior of the iterates in (3.5). In (3.7), $(Y^*X)^+$ denotes the Moore–Penrose pseudoinverse of Y^*X and, crucially, X need not be orthonormal.

A subspace analogue of Theorem 2.4.1, based on the largest principal angle between $\mathcal{S}_k$ and $\mathcal{V}$, is easy to derive with (3.7) (c.f. [110, Thm 14.4.1]).

Theorem 3.2.2. *Let normal $A \in \mathbb{C}^{n\times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.2) *and* (3.3)*, respectively, and let $\mathcal{S}_k = \operatorname{span}(Q_k)$ in* (3.5)*. If $\cos\theta_1(\mathcal{S}_0,\mathcal{V}) > 0$, then*

$$\tan\theta_1(\mathcal{S}_k,\mathcal{V}) \le \left|\frac{r(\lambda_{m+1})}{r(\lambda_m)}\right| \tan\theta_1(\mathcal{S}_{k-1},\mathcal{V}) \le \left|\frac{r(\lambda_{m+1})}{r(\lambda_m)}\right|^k \tan\theta_1(\mathcal{S}_0,\mathcal{V}). \tag{3.8}$$

Proof. We prove the first inequality with a direct calculation using (3.7); the second follows immediately by induction and the fact that $\cos\theta > 0$ when $\tan\theta < \infty$. We compute that $(I - P_{\mathcal{V}})X_k = V_2 r(\Lambda_2) V_2^* Q_{k-1}$ and that $V_1^* X_k = r(\Lambda_1) V_1^* Q_{k-1}$. Using the induction hypothesis that $\cos\theta_1(\mathcal{S}_{k-1},\mathcal{V}) > 0$, which implies $V_1^* Q_{k-1}$ is invertible, we obtain

$$(I - P_{\mathcal{V}})X_k(V_1^*X_k)^+ = V_2 r(\Lambda_2) V_2^* Q_{k-1}(V_1^*Q_{k-1})^{-1} r(\Lambda_1)^{-1}. \tag{3.9}$$

The theorem follows by taking norms and noting that $\|V_2^*Q_{k-1}(V_1^*Q_{k-1})^{-1}\| = \tan\theta_1(\mathcal{S}_{k-1},\mathcal{V})$, $\|r(\Lambda_2)\| = |r(\lambda_{m+1})|$, and $\|r(\Lambda_1)^{-1}\| = |r(\lambda_m)|^{-1}$. □

The tangents (and sines) of the principal angles play an important role in the perturbation theory of eigenpairs and, consequently, the bounds in Theorem 3.2.2 are useful when determining the accuracy in the computed Ritz

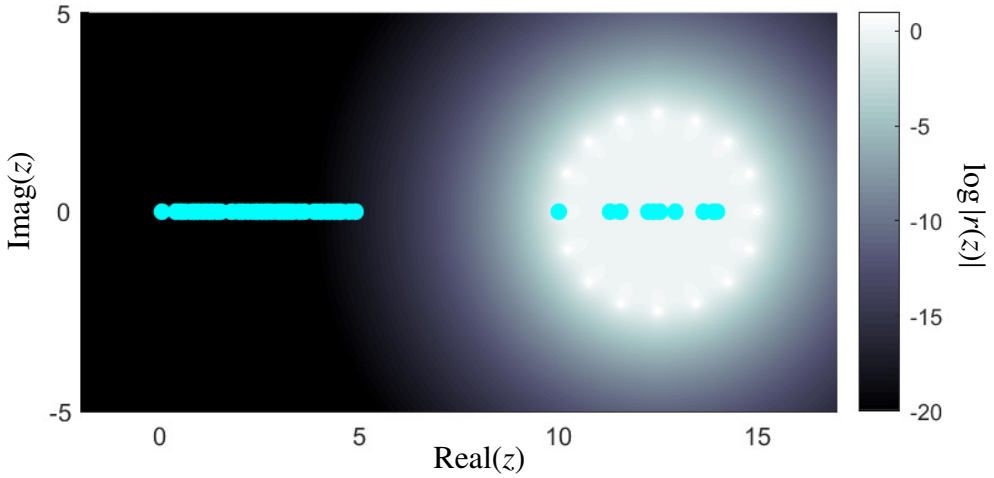

Figure 3.3: The eigenvalues of a 100×100 real-symmetric matrix overlaid on a complex color plot of the magnitude of a rational approximation to the characteristic function on $[10, 15]$. A dangerous eigenvalue is located at distance $d = 10^{-10}$ from the pole at $z = 10$.

pairs [125, 139, 140]. For our purposes, Theorem 3.2.2 and its proof are useful tools when analyzing subspace iterations subject to perturbations (see section 3.5), because $\tan\theta_1(\mathcal{S}_k, \mathcal{V})$ is computed directly from the iterate X_k.

3.3 Dangerous eigenvalues

When an eigenvalue of A is very close to a pole of the rational filter in (3.1), $r(A)$ disproportionately amplifies components in the direction of the associated eigenvector. Given any vector $x \in \mathbb{C}^n$, we estimate

$$r(A)x = \sum_{i=1}^{n} r(\lambda_i) v_i v_i^* x = \frac{v_1^* x}{d e^{i\theta}} v_1 + O(1), \quad \text{as} \quad d \to 0. \tag{3.10}$$

(It is convenient to write the complex-valued difference between λ_1 and the nearest pole z_{j_*} in the polar notation $z_{j_*} - \lambda_1 = d e^{i\theta}$, with argument $0 \le \theta < 2\pi$.) This amplification is precisely the reason that shift-and-invert power iterations are so effective when the shift is close to the target eigenvalue. If we apply $r(A)$

to a random vector with unit norm and normalize, the result approximates v_1 with relative accuracy $O(d)$, under the generic assumption that the random vector is not nearly orthogonal to v_1. Similarly, when $r(A)$ is applied to a random orthonormal matrix Q_0, $\operatorname{span}(r(A)Q_0)$ contains good approximations to v_1 when $\|v_1^* Q_0\|$ is not too small.

However, the amplifying effect of a dangerous eigenvalue may cause issues when computing the iterates in (3.5) in floating-point arithmetic. Figure 3.3 shows the eigenvalues of a 100×100 real symmetric matrix plotted in the complex plane over the magnitude (indicated by color) of a rational filter targeting the interval $[10, 15]$. The matrix has a large cluster of eigenvalues in the interval $[0, 5]$, where the filter has decayed to less than unit round-off, and a small set of eigenvalues in the target region, where the filter has magnitude close to 1. One eigenvalue of the matrix is very close to the pole at $z = 10$, separated by a distance of 10^{-10}. By Theorem 3.2.2, we expect that (in exact arithmetic) all of the eigenvalues in the target region are resolved to accuracy on the order of u after one iteration. However, Figure 3.4 (left) shows that only the dangerous eigenpair has been computed accurately. The residuals of the remaining target eigenpairs are on the order of 10^{-5}, that is, roughly u/d.

The large residuals are best explained with a look at the computed orthonormal basis $\hat{Q}_1$ (we denote computed quantities with a hat throughout, so $\hat{Q}_1$ is the computed approximant to Q_1) in the eigenvector coordinates in Figure 3.4. The first column of $\hat{Q}_1$ (circular markers) looks as expected: the dangerous eigenvector dominates and the unwanted components are near the unit round-off in magnitude. However, the magnitude of the unwanted components is much larger, on the order of u/d, in the remaining columns $\hat{q}_2^{(1)}, \ldots, \hat{q}_m^{(1)}$. The 10th col-

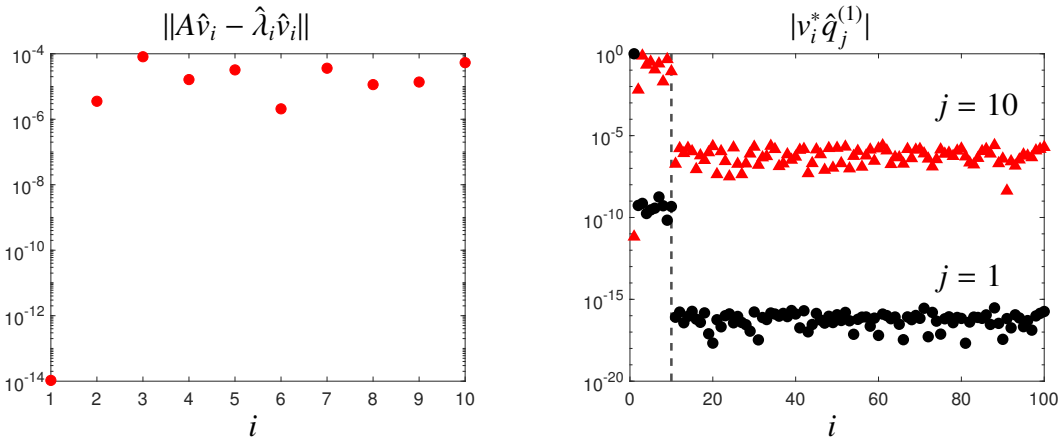

Figure 3.4: The residuals of 10 target eigenpairs of a 100×100 real-symmetric matrix after one iteration of subspace iteration with the rational filter in Figure 3.3 are plotted on the left. On the right are the eigenvector coordinates of the 1st (black circles) and 10th (red triangles) columns of $\hat{Q}_1$. The dangerous component and the remaining target components dominate in columns 1 and 10, respectively. The unwanted eigenvector components (right of dashed line) are filtered out almost entirely to order u in the 1st column, but are orders of magnitude larger in the 10th column, with magnitude near u/d.

umn (triangular markers in Figure 3.4) is representative of this observation. Although the quality of the filter means that the unwanted components should be on the order of u in the columns of $\hat{Q}_1$, they are polluted with noise on the order of u/d in all but the first column. Consequently, the accuracy in the remaining Ritz pairs computed from $\hat{Q}_1$ is also degraded to u/d.

There are two potential sources of error degrading the accuracy in $\hat{Q}_1$. The first is the most obvious: round-off errors are amplified when solving the ill-conditioned linear system associated with the dangerous eigenvalue. The second source is more subtle: the overwhelming dominance of the dangerous eigenvector in each column of X_1 leads to an ill-conditioned basis for $\mathcal{S}_1$. Remarkably, the heart of the story in Figures 3.1 and 3.2 is contained in the latter, subtler effect and we can learn a great deal without mentioning errors incurred while applying $r(A)$. Of course, a thorough understanding requires a careful treatment of the ill-conditioned linear systems and the accumulation of errors

at each iteration. We address both points in section 3.5, where we study the convergence and stability of the iteration in (3.5) when computed in floating-point arithmetic. For now, we focus on the influence of ill-conditioning in the iterates $X_1, X_2, \ldots$, noting that round-off errors in the computed iterates have little effect on their condition number (see section 3.5 for a full explanation).

3.3.1 Accuracy of the computed orthonormal basis

When a basis $X \in \mathbb{C}^{n\times m}$ is ill-conditioned, small perturbations to the columns can have a large effect on their span. This is reflected in the sensitivity of the orthogonal factor in the QR factorization, $Q = \mathrm{qf}(X)$. If, for some small $\epsilon > 0$, X is perturbed by ΔX with $\|\Delta X\| \leq \epsilon\|X\|$, then there is a ΔQ such that $Q + \Delta Q = \mathrm{qf}(X + \Delta X)$ and [78, p. 382]

$$\|\Delta Q\| \leq c_m \kappa(X)\|\Delta X\|/\|X\|. \tag{3.11}$$

Here, c_m is a modest constant depending only on the dimension m and $\kappa(\cdot)$ denotes the 2-norm condition number of a rectangular matrix. (3.11) tells us that when X is highly ill-conditioned, the QR factorization may be extremely sensitive to perturbations. When we compute an orthonormal basis $\hat{Q}$ in floating-point arithmetic, we are not guaranteed accuracy much better than $\|\hat{Q} - Q\| \leq c_m\kappa(X)u$ (at least, as long as the columns of X do not vary significantly in magnitude).

Because the rational filter amplifies the v_1 component in each column of Q_0 by $1/d$ in (3.5), X_1 is usually extremely ill-conditioned. Intuitively, $\kappa(X_1)$ cannot be much worse than $|r(\lambda_1)|/|r(\lambda_m)|$ and not much better than $|r(\lambda_1)|/|r(\lambda_2)|$ because v_1 is present in each column with magnitude near $|r(\lambda_1)|$ while the rest of the

target eigenpairs are present with magnitude at least $|r(\lambda_m)|$ and no greater than $|r(\lambda_2)|$. Proposition 3.3.1 makes this intuition precise in the form of an upper bound and asymptotic lower bound. The implication is that the error in the computed orthonormal basis $\hat{Q}_1$ is on the order of u/d as long as the columns of Q_0 are not orthogonal to the dangerous eigenvector, as we observed in Figure 3.4.

We use the shorthand notation $f(x) \lesssim g(x)$ to denote the asymptotic relation

$$f(x) \leq g(x)(1 + o(1)), \qquad \text{as} \qquad x \to 0. \tag{3.12}$$

Note that this is slightly sharper than $f(x) = O(g(x))$, but weaker than $f(x) \sim g(x)$ [81, Def. 1.1-1.2].[4]

Proposition 3.3.1. *Let normal $A \in \mathbb{C}^{n\times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.2) *and* (3.3)*, respectively, and given orthonormal $Q_0 \in \mathbb{C}^{n\times m}$, let $X_1 = r(A)Q_0$. If $V_1^* Q_0$ has full rank, then the condition number of X_1 satisfies*

$$\frac{\|v_1^* Q_0\|}{d|r(\lambda_2)|} \lesssim \kappa(X_1) \leq \left|\frac{r(\lambda_1)}{r(\lambda_m)}\right| \|(V_1^* Q_0)^{-1}\|, \qquad \text{as} \qquad d \to 0. \tag{3.13}$$

Proof. The condition number of X_1 may be written as $\kappa(X_1) = \sigma_1(X_1)/\sigma_m(X_1)$, where $\sigma_1(X_1) \geq \cdots \geq \sigma_m(X_1)$ are the singular values of X_1. To bound $\sigma_1(X_1)$ above, we substitute the spectral decomposition $r(A) = V r(\Lambda) V^*$ into the definition of X_1 and estimate $\sigma_1(X_1) \leq |r(\Lambda_1)| \|V^* Q_0\| \leq |r(\lambda_1)|$ (recall (3.3)). To bound $\sigma_m(X_1)$ below, we use the spectral decomposition in (3.2) to write

$$X_1 = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \tag{3.14}$$

[4] This definition of $f \lesssim g$ is sharper than its common usage in the analysis of partial differential equations, where it means $f \leq Cg$ for some constant $C > 0$ [146, p. xiv].

where $M_1 = r(\Lambda_1)V_1^*Q_0$ and $M_2 = r(\Lambda_2)V_2^*Q_0$. Because V is unitary, the singular values of X_1 are precisely those of $\left[\begin{smallmatrix} M_1 \\ M_2 \end{smallmatrix}\right]$. Furthermore, $\sigma_m(X_1) \geq \sigma_m(M_1)$ since adding rows can only increase the singular values of a matrix. Finally, since $\sigma_m(M_1) = \|M_1^{-1}\|^{-1}$, we have that $\kappa(X_1) \leq \sigma_1(X_1)\|M_1^{-1}\|$. We estimate that

$$\|M_1^{-1}\| = \|(V_1^*Q_0)^{-1}r(\Lambda_1)^{-1}\| \leq \|(V_1^*Q_0)^{-1}\|\|r(\lambda_m)|^{-1}.$$

Collecting the bounds on $\sigma_1(X_1)$ and $\|M_1^{-1}\|$ establishes the upper bound in (3.13).

To establish the asymptotic lower bound, we apply (3.10) to $r(A)Q_0$, obtaining

$$X_1 = \sum_{i=1}^{n} r(\lambda_i)v_i v_i^* Q_0 = v_1 \frac{v_1^* Q_0}{de^{i\theta}} + O(1), \quad \text{as} \quad d \to 0. \tag{3.15}$$

Taking norms provides the asymptotic lower bound on $\sigma_1(X_1)$. To obtain a lower bound on $\sigma_m(X_1)^{-1}$, we can bound $\sigma_m(X_1)$ from above with an interlacing property for singular values of matrices subject to rank one perturbations. We rewrite (3.15) as

$$X_1 = r(\lambda_1)v_1 v_1^* Q_0 + V\text{diag}(0, r(\lambda_2), \ldots, r(\lambda_n))V^*Q_0 = N_1 + N_2.$$

Now, $\sigma_2(N_1) = 0$ and $\sigma_1(N_2) \leq |r(\lambda_2)|$ so, by interlacing [148], we obtain the estimate

$$\sigma_2(X_1) \leq \sigma_1(N_2) + \sigma_2(N_1) \leq |r(\lambda_2)|.$$

As $\sigma_m(X_1) \leq \sigma_2(X_1)$ implies that $1/\sigma_m(X_1) \geq |r(\lambda_2)|^{-1}$, collecting lower bounds concludes the proof of (3.13). □

The factor $\|(V_1^*Q_0)^{-1}\|$ in Proposition 3.3.1 appears naturally in connection with subspace iteration, and we will encounter it again in section 3.5. It is precisely the reciprocal of $\cos\theta_1(\mathcal{S}_0, \mathcal{V})$ (see Definition 3.2.1), approaching unity when $\mathcal{V}$ and $\mathcal{S}_0$ are nearby and blowing up quadratically when they are made

orthogonal. In Proposition 3.3.1 it indicates that X_1 may suffer additional ill-conditioning if the initial subspace $\mathcal{S}_0$ is accidentally chosen to be too near orthogonal to $\mathcal{V}$.[5]

3.4 Twice is enough

In Proposition 3.3.1, the asymptotic lower bound in (3.13) plummets if the columns of Q_0 are taken nearly orthogonal to v_1, the dangerous eigenvector. This is because the rational filter has nothing to amplify when v_1 is absent in the columns of Q_0. If v_1 is present with magnitude no greater than $O(d)$ in Q_0, then the columns of X_1 are not strongly aligned along any single eigenvector and the conditioning of X_1 is likely to improve. Crucially, this intuition holds even if v_1 dominates in one column but not the others. The main point is that the columns of X_1 are no longer necessarily close to a linearly dependent set.

Let us return to the example of Figure 3.3. If we print out the residual norms of the target eigenpairs after the second iteration of (3.5), we see remarkable improvement:

```
6.7997e-15   2.5942e-14   2.2680e-13   4.3433e-14   9.1978e-14
1.3716e-14   9.7045e-14   3.4121e-14   1.4594e-13   4.0235e-14
```

Now all the target pairs have been resolved to within 13 or 14 digits of accuracy,

[5]When Q_0 is selected so that its entries are independent, identically distributed Gaussian random variables, $\|(V_1^* Q_0)^{-1}\|$ is roughly $\sqrt{m}$ in expectation, but can be an order of magnitude or so larger with nontrivial probability. A powerful workaround is to work with a slightly larger subspace and take m larger than the number of target eigenvalues; this dramatically reduces the probability of large $\|(V_1^* Q_0)^{-1}\|$ [40].

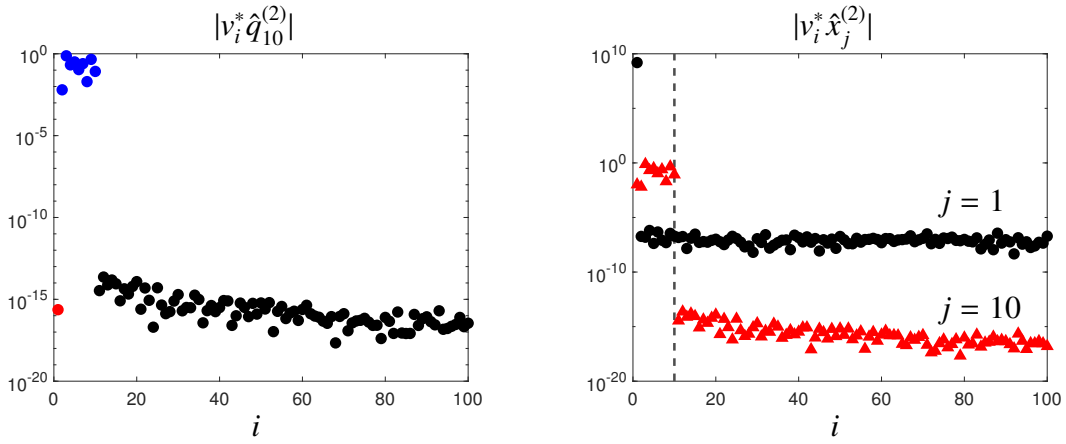

Figure 3.5: The structure of the iterates $\hat{X}_2$ and $\hat{Q}_2$ after the second iteration of subspace iteration with a rational filter. On the left, the eigenvector coordinates of the 10th column of the computed orthonormal basis color-coded for dangerous component (red), remaining target components (blue), and unwanted components (black). On the right, the eigenvector coordinates of the 1st (black circles) and 10th (red triangles) columns of the computed basis $\hat{X}_2$ with dashed line dividing target and unwanted eigenvector coordinates.

in contrast to Figure 3.4 (left). If we examine the computed orthonormal basis used to extract the Ritz pairs, we observe that the noise in the direction of the unwanted eigenvectors has also been reduced to the order of u, compared with u/d in the first iteration. Figure 3.5 (left) illustrates the composition of the 10th column of $\hat{Q}_2$, which is representative of the last $m - 1$ columns.

The reason for the restored accuracy in the computed orthonormal basis is that, unlike X_1, the basis X_2 has an even blend of the target eigenvector directions in all but the first of its columns. Figure 3.5 (right) displays the magnitude of the eigenvector coordinates for the first (circular markers) and last (triangular markers) columns of the computed basis, $\hat{X}_2$. In the first column, the dangerous direction is effectively the only direction present, since all other components appear with relative magnitude near u. In contrast, the last column of $\hat{X}_2$ contains order one components in each target direction with the unwanted directions completely filtered out. The remaining columns of $\hat{X}_2$ are similar in composition to the last. Without v_1 dominating in every column, we can accurately extract

an orthonormal basis.

The clue to the stark difference in the composition of $\hat{X}_1$ and $\hat{X}_2$ is contained in Figure 3.4. We see that the first column of $\hat{Q}_1$ is dominated by the dangerous eigenvector, up to the 9th or 10th digit. Consequently, the remaining columns of $\hat{Q}_1$ are nearly orthogonal to v_1. We observe this in Figure 3.4 (right), where $v_1^* q_{10}^{(1)} \approx 10^{-11}$. When the rational filter is applied to $\hat{Q}_1$ in the second iteration, the amplification of v_1 restores an even blend of the target eigenvectors in the last $m - 1$ columns of $\hat{X}_2$, rather than boosting v_1 above the others.

3.4.1 A well-conditioned basis

Motivated by the preceding discussion, we now examine the condition number of the second iterate, X_2, when $q_1^{(1)}$ is a good approximation to v_1 and, consequently, all but one of the columns of Q_1 are deficient in the dangerous direction. We then briefly explain the structure of the eigenvector coordinates for the computed orthonormal basis $\hat{Q}_1$ observed in Figure 3.4.

To investigate how a weak presence of v_1 in columns of Q_1 improves the conditioning of X_2, we break the target eigenvector coordinates of Q_1 into blocks, as

$$V_1^* Q_1 = \begin{bmatrix} v_1^* q_1^{(1)} & v_1^* \tilde{Q}_1 \\ \tilde{V}_1^* q_1^{(1)} & \tilde{V}_1^* \tilde{Q}_1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & D \end{bmatrix}. \tag{3.16}$$

Here, we use $\tilde{V}_1$ and $\tilde{Q}_1$ to denote the $n \times (m-1)$ matrices formed by removing the first columns of V_1 and Q_1, respectively. When $q_1^{(1)}$ closely approximates v_1, $\|c\|$ is small due to the orthogonality of the eigenvectors. Moreover, $\|b\|$ is also small because the columns of $\tilde{Q}_1$ are nearly orthogonal to v_1. Let us suppose that

$q_1^{(1)} = v_1 + O(d)$, so that b and c are $O(d)$ (we explain why this holds momentarily, following Theorem 3.4.1).

After applying the rational filter to Q_1, the eigenvector coordinates $V_1^* X_2$ inherit a natural block structure from (3.16). Letting $\tilde{\Lambda}_1 = \mathrm{diag}(\lambda_2, \ldots, \lambda_m)$, we have that

$$V_1^* X_2 = r(\Lambda_1) V_1^* Q_1 = \begin{bmatrix} r(\lambda_1)a & r(\lambda_1)b \\ r(\tilde{\Lambda}_1)c & r(\tilde{\Lambda}_1)D \end{bmatrix}. \tag{3.17}$$

While the bottom left block remains small, with norm no greater than $|r(\lambda_2)|\|c\| = O(d)$, the entire first row is amplified by $|r(\lambda_1)|$ so that $\|r(\lambda_1)b\| \approx \|b\|/d$.

To estimate the condition number of X_2, we scale the columns of X_2 with the $m \times m$ diagonal matrix

$$T = \mathrm{diag}(r(\lambda_1)^{-1}, 1, \ldots, 1). \tag{3.18}$$

This diagonal scaling does not alter $\mathrm{span}(X_2)$ or the sensitivity of the orthonormal basis, $Q_2 = \mathrm{qf}(X_2)$. (Note that scaling the columns of X_1 has no effect in Proposition 3.3.1, as the columns of X_1 all have magnitude $\approx 1/d$.) However, it conveniently puts the diagonal blocks in (3.17) on equal footing and ensures that $\sigma_1(X_2 T) = O(1)$, so that any ill-conditioning due to the dangerous eigenvalue is captured in the smallest singular value of $X_2 T$. This allows us to focus on computing a lower bound for $\sigma_m(X_2 T)$ or, equivalently, an upper bound for $1/\sigma_m(X_2 T)$. Just as in the proof of Proposition 3.3.1, it suffices to bound $\|(V_1^* X_2 T)^{-1}\|$ from above (assuming as usual that $V_1^* Q_1$, and therefore $V_1^* X_2$, has full rank).

With all of the ingredients in place, the estimate is fairly straightforward. After the column scaling, $V_1^* X_2 T$ is approximately block upper triangular. We

have that

$$V_1^* X_2 T = \begin{bmatrix} r(\lambda_1)a & r(\lambda_1)b \\ r(\tilde{\Lambda}_1)c & r(\tilde{\Lambda}_1)D \end{bmatrix} T = \begin{bmatrix} a & b/(de^{i\theta}) \\ & r(\tilde{\Lambda}_1)D \end{bmatrix} + O(d). \tag{3.19}$$

We can apply the formula for 2×2 block upper triangular matrix inversion and the fact that matrix inversion is locally Lipschitz continuous to compute (for d sufficiently small)

$$(V_1^* X_2 T)^{-1} = \begin{bmatrix} a^{-1} & -a^{-1} D^{-1} r(\tilde{\Lambda}_1)^{-1} b/(de^{i\theta}) \\ & D^{-1} r(\tilde{\Lambda}_1)^{-1} \end{bmatrix} (I + O(d)) \,. \tag{3.20}$$

The norm of the block upper triangular matrix in (3.20) is bounded by the sum of the norms of the blocks, so we conclude that $1/\sigma_m(X_2 T) \leq \|(V_1^* X_2 T)^{-1}\| = O(1)$ when $\|b\| = O(d)$. Estimating the norms of these blocks individually and combining with an estimate for $\sigma_1(X_2 T)$ leads to the following upper bound on $\kappa(X_2 T)$.

Theorem 3.4.1 (Twice-is-enough)**.** *Let normal $A \in \mathbb{C}^{n \times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.2) *and* (3.3)*, respectively, and given orthonormal $Q_1 \in \mathbb{C}^{n \times m}$, let $X_2 = r(A)Q_1$. Let b and D denote the blocks of $V_1^* Q_1$ in* (3.16)*. If D is invertible and the first column of Q_1 satisfies $\|q_1^{(1)} - v_1\| = O(d)$, then $\|b\| = O(d)$ and*

$$\kappa(X_2 T) \leq M \left(\left(\frac{\|b\|}{d} + 1 \right) \frac{\|D^{-1}\|}{|r(\lambda_m)|} + 1 \right) + O(d) \qquad \text{as} \qquad d \to 0. \tag{3.21}$$

Here, $T = \operatorname{diag}(r(\lambda_1)^{-1}, 1, \ldots, 1) \in \mathbb{C}^{m \times m}$ and $M = \|b\|/d + \max\{1, |r(\lambda_2)|\}$.

Proof. First, the hypothesis $\|q_1^{(1)} - v_1\| = O(d)$ immediately implies that $|a| = 1 + O(d)$ and $\|b\| = O(d)$. Then, following the discussion above, it suffices to bound $\|X_2 T\|$ and the norms of the blocks in (3.20). The condition number of $X_2 T$ is bounded above by the product of these two estimates. Since $\|r(\tilde{\Lambda}_1)^{-1}\| = |r(\lambda_m)|^{-1}$, we obtain that

$$1/\sigma_m(X_2 T) \leq \|(V_1^* X_2 T)^{-1}\| \leq 1 + \frac{\|D^{-1}\|}{|r(\lambda_m)|} \left(1 + \frac{\|b\|}{d} \right) + O(d). \tag{3.22}$$

On the other hand, we can write V^*X_2T in block form analogous to (3.19), as

$$V^*X_2T = \begin{bmatrix} a & b/(de^{i\theta}) \\ & r(\tilde{\Lambda})\tilde{D} \end{bmatrix} + O(d),$$

where $\tilde{\Lambda} = \text{diag}(\lambda_2, \ldots, \lambda_n)$ and $\tilde{D} = \tilde{V}^*\tilde{Q}_1$ (here, $\tilde{V}$ is V with first column removed). Calculating the norm of the block diagonal component and the off-diagonal component separately and applying the triangle inequality yields $\|X_2T\| \leq M$. Collecting with the bound in (3.22) concludes the proof. □

Theorem 3.4.1 tells us that X_2 is only a simple column scaling away from a well-conditioned basis when the first column of Q_1 approximates v_1 with accuracy $O(d)$. Since the sensitivity (and numerical computation) of the QR factorization is not affected by column scaling, the $O(1)$ bound on $\kappa(X_2T)$ in (3.21) explains why the computed orthonormal basis for $\mathcal{S}_2$ is accurate to unit round-off. This line of analysis follows naturally from our observation about the eigenvector coordinates of $\hat{Q}_1$ in Figure 3.4 (right), but one question remains. Why is the first column of the computed orthonormal basis such a good approximation to v_1?

The answer is that $\hat{q}_1^{(1)}$ is essentially the first column of X_1 after normalization, up to the unit round-off u. In particular, $\hat{q}_1^{(1)}$ is unaffected by the u/d errors in $\hat{Q}_1$ caused by ill-conditioning in X_1 (see Proposition 3.3.1). These errors are concentrated in the later columns of $\hat{Q}_1$ because of the nested structure of Householder reflections (or Givens rotations) used to make X_1 upper triangular. We have that $x_1/\|x_1\| = v_1 + O(d)$ by (3.10), so we expect that $\hat{q}_1^{(1)} = v_1 + O(d)$ also, as observed in Figure 3.4.

Finally, if the orthogonal factor is computed with modified Gram-Schmidt instead of Householder reflections or Givens rotations, the columns of $\hat{Q}_1$ lose

orthogonality in proportion to the condition number of the ill-conditioned basis X_1. The consequence of this is that the block $v_1^*(\hat{Q}_1)_{(2:m)}$ from (3.16) may be as large as u/d instead of $O(d)$, even though $\hat{q}_1^{(1)} = v_1 + O(d)$. This may alter the order of magnitude of $\kappa(X_2 T)$ when $d \ll \sqrt{u}$ (since then, $u/d \gg d$) as the balance in Theorem 3.4.1 is disrupted. In particular, twice may no longer be enough to correct ill-conditioning in $\hat{X}_2$. A similar effect is observed for non-normal matrices in section 3.6 even when Householder reflections or Givens rotations are employed in the QR factorizations.

3.5 Convergence and stability

So far, our analysis of dangerous eigenvalues has focused on the conditioning of the iterates $X_1, X_2, \ldots$ in (3.5) and the corresponding accuracy in the computed orthonormal bases. Indeed, this perspective explains the u/d errors observed in the first iteration (see Figure 3.4) and provides the essential insight into the restored accuracy observed in the second iteration (see Figure 3.5). But we have not yet explained how the round-off errors incurred while applying the ill-conditioned rational filter enter the picture. Nor have we discussed how these round-off errors, together with the error in the computed orthonormal basis, accumulate during the iterations in (3.5).

To apply the rational filter $r(A)$ to an $n \times m$ matrix Q in practice, one solves linear systems with a shift at each pole and takes a weighted average of the solutions:

$$r(A)Q = \sum_{j=1}^{\ell} \omega_j X^{(j)}, \quad \text{where} \quad (z_j I - A)X^{(j)} = Q, \quad j = 1, \ldots, \ell. \tag{3.23}$$

If the linear systems are solved with a backward stable algorithm, then the computed solutions $\hat{X}^{(j)}$ satisfy, for each $j = 1, \ldots, \ell$,

$$(z_j I - A - \mathcal{E}_j)\hat{X}^{(j)} = Q, \qquad \|\mathcal{E}_j\| \leq \gamma \|A\| u. \tag{3.24}$$

Here, $\mathcal{E}_j$ is the backward error and γ is a constant, with modest dependence on $z_1, \ldots, z_\ell$ and the dimension of Q, such that $\gamma u \ll 1$ for typical situations.[6]

Now, if we neglect errors made while forming the linear combination on the left hand side of (3.23), then the forward error in $r(A)Q$ can be written as[7]

$$\hat{X} - r(A)Q = \sum_{j=1}^{\ell} \omega_j (z_j I - A)^{-1} \mathcal{E}_j \hat{X}^{(j)}, \qquad \text{where} \qquad \hat{X} = \sum_{j=1}^{\ell} \omega_j \hat{X}^{(j)}. \tag{3.25}$$

Due to the appearance of $\mathcal{E}_j$, the terms in the left hand sum are all on the order of u except for the term corresponding to the pole near the dangerous eigenvalue, whose index we call $j = j_*$. In the dangerous term, $(z_{j_*} I - A)^{-1}$ amplifies the v_1 components in the columns of $\mathcal{E}_{j_*}$ by a factor of $1/d$. Similarly, the components of v_1 in the columns of Q are amplified to order $1/d$ in the corresponding columns of $\hat{X}^{(j_*)}$ (this is made precise by expanding (3.24) in a Neumann series). Therefore, the relative errors in the columns of $\hat{X}$ are on the order of u/d.

Thus, every time $r(A)$ is applied in (3.5), relative errors of order u/d are accrued in the columns of $\hat{X}_k$. On the one hand, our understanding of accuracy in the computed orthonormal basis $\hat{Q}_k$ (developed in sections 3.3 and 3.4) remains

[6] This characterization can be modified to accommodate inexact solution techniques, such as iterative methods, but γ may be much larger, depending on the stability properties of the particular numerical method [139, p. 339].

[7] For expositional clarity, we neglect round-off errors accrued when forming the linear combination in the right hand side of (3.23) to focus on the effect of the ill-conditioned linear systems. For typical choices of the weights and nodes in $r(A)$, this amounts to discarding a term on the order of u relative to the largest column norm of the $\hat{X}^{(j)}$.

intact, because perturbations of relative order u/d to the columns of X_k have little effect on the leading-order estimates for $\kappa(X_k)$. On the other hand, we may wonder: what effect do such perturbations have on $\operatorname{span}(X_k)$ and the geometric convergence implied in Theorem 3.2.2?

Recent analyses of subspace iteration accelerated with a rational filter suggest that $\operatorname{span}(\hat{X}_k)$ tends to $\mathcal{V}$ geometrically at roughly the expected rate until a threshold of accuracy is reached, at which point convergence plateaus [144]. This threshold is usually the same order of magnitude as the error accrued in the subspace at each iteration, i.e., in the columns of $\hat{X}_k$. Similar results have been derived for perturbations in the entries of the matrix $r(A)$ (this work does not consider filters explicitly) [126]. However, the evidence of the experiments in Figures 3.1 and 3.2 and in section 3.4 indicates that errors in $\operatorname{span}(\hat{X}_k)$ caused by dangerous eigenvalues do not prevent the Rayleigh–Ritz procedure from finding vectors in $\operatorname{span}(\hat{X}_k)$ that approximate the target eigenvectors to unit round-off accuracy. We now show that errors in $\hat{X}_k$ caused by the dangerous eigenvalue do not lead to early stagnation or instability in the computed iterates. In the worst case, they may slow the geometric convergence rate by a factor of roughly $(1 - u/d)^{-1}$. Moreover, the iteration is stable as long as the columns of the initial guess Q_0 are not too near to $\mathcal{V}^\perp$ (see Figure 3.6).

3.5.1 One-step refinement bounds

The amplifying power of the dangerous eigenvalue leads to large relative errors in the columns of $\hat{X}_k$. However, the errors possess an important quality: the amplification is entirely in the direction of v_1 so that the relative errors in

the unwanted direction are still small. To understand how these structured perturbations influence $\hat{\mathcal{S}}_k = \operatorname{span}(\hat{X}_k)$, we gather the errors accrued during the kth iteration into a perturbation to the orthonormal basis for $\hat{\mathcal{S}}_{k-1}$ and construct a one-step refinement bound as in Theorem 3.2.2. Formulated precisely, we replace (3.5) with the perturbed form

$$\hat{X}_k = r(A)(Q'_{k-1} + R_k), \qquad Q'_k = \operatorname{qf}(\hat{X}_k). \tag{3.26}$$

Note that we include any errors in the computed orthonormal factor in R_k, placing the emphasis on $\hat{\mathcal{S}}_k = \operatorname{span}(\hat{X}_k) = \operatorname{span}(Q'_k)$ rather than on $\operatorname{span}(\hat{Q}_k)$. This causes no difficulty since, as we know from section 3.4, the error $\hat{Q}_k - Q'_k$ is on the order of u for $k \geq 2$. Since Q'_k is an orthonormal basis, $\hat{\mathcal{S}}_k$ and $\operatorname{span}(Q'_k)$ only differ by a term not much larger than u.

To begin, we establish the form (3.26) by way of the residuals of the linear systems in (3.25) and study the structure of R_k. To measure the columns of R_k relative to the columns of $\hat{X}_k$, it is convenient to apply the diagonal scaling

$$C_k = \operatorname{diag}(\|(\hat{X}_k)_i\|^{-1}, \ldots, \|(\hat{X}_k)_m\|^{-1})/\sqrt{m}, \tag{3.27}$$

so that $\|\hat{X}_k C_k\| \leq 1$. We also need the majorization of the rational filter, denoted

$$\tilde{r}(\lambda) = \sum_{j=1}^{\ell} |\omega_j| |(z_j - \lambda)^{-1}|. \tag{3.28}$$

As usual, $\tilde{r}(\Lambda_1)$ and $\tilde{r}(\Lambda_2)$ are the matrices when the function in (3.28) is applied to the diagonal matrices Λ_1 and Λ_2. Observe that $\|\tilde{r}(\Lambda_1) r(\Lambda_1)^{-1}\| = O(1)$ as $d \to 0$, because the poles near the dangerous eigenvalue cancel.

Lemma 3.5.1. *Let normal $A \in \mathbb{C}^{n\times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.2) *and* (3.3)*, respectively. Given $Q \in \mathbb{C}^{n\times m}$, let $\hat{X} = \sum_{j=1}^{\ell} \omega_j \hat{X}^{(j)}$ with each $\hat{X}^{(j)}$ satisfying* (3.24)*. Then, there is an $R \in \mathbb{C}^{n\times m}$ such that $\hat{X} = r(A)(Q + R)$ and*

$$\|P_{\mathcal{V}} R\| \leq \gamma_1 \|A\| u/d \qquad \text{and} \qquad \|(I - P_{\mathcal{V}}) r(A) R C\| \leq \gamma_2 \|A\| u. \tag{3.29}$$

Here, $\gamma_1 = \gamma \|r(\Lambda_1)^{-1}\tilde{r}(\Lambda_1)\|$, $\gamma_2 = \gamma \|\tilde{r}(\Lambda_2)\|$, *and* C *is the diagonal scaling in* (3.27) *(with index* k *suppressed).*

Proof. Because each $\hat{X}^{(j)}$ satisfies (3.24) and $\hat{X} = \sum_{j=1}^{\ell} \omega_j \hat{X}^{(j)}$, we collect like terms in (3.25) and compute

$$\hat{X} = r(A)Q + \sum_{j=1}^{\ell} \omega_j (z_j I - A)^{-1} R^{(j)}, \tag{3.30}$$

where $R^{(j)} = \mathcal{E}^{(j)}\hat{X}$. Note that $\|R^{(j)}C\| \leq \gamma \|A\| u$, for $j = 1, \ldots, \ell$, by (3.24).

We compute R directly by comparing (3.30) with $\hat{X} = r(A)(Q + R)$ and noting that we need $r(A)R = \sum_{j=1}^{\ell} \omega_j (z_j I - A)^{-1} R^{(j)}$. Inserting the eigenvalue decomposition $A = V\Lambda V^*$ into both sides and inverting $r(A) = Vr(\Lambda)V^*$, we obtain

$$R = Vr(\Lambda)^{-1} \left(\sum_{j=1}^{\ell} \omega_j (z_j I - \Lambda)^{-1} V^* R^{(j)} \right). \tag{3.31}$$

Calculating $P_{\mathcal{V}}R$ and $(I - P_{\mathcal{V}})r(A)RC$ directly from (3.31) and applying the backward error bounds in (3.24) to bound the residuals $\|R^{(j)}\|$ uniformly, we obtain the bounds in (3.29). □

Lemma 3.5.1 demonstrates that the perturbations R_k in (3.26) capture the essential structure of the errors in $\hat{X}_k$. First, R_k perturbs Q'_{k-1} with relative magnitude u/d and direction in the subspace $\mathcal{V}$. Second, $r(A)R_k$ perturbs the columns of X_k with relative magnitude u and direction in the subspace $\mathcal{V}^{\perp}$. We note that $\|V_2^* R_k C_k\|$ itself is not small when the filter is very good, i.e., close to unit-round off on the unwanted eigenvalues, as $\|r(\Lambda_2)^{-1}\tilde{r}(\Lambda_2)\|$ may be extremely large. However, the forward application of the filter cancels any large factors in $r(\Lambda_2)^{-1}$ exactly.

With Lemma 3.5.1 in hand, we can calculate a one-step refinement bound generalizing Theorem 3.2.2 to the perturbed iteration in (3.26). While the u/d

relative errors in $\hat{X}_k$ are felt in the refinement factor in (3.32), they do not appear in the additive perturbation to $\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$. This point is crucial because, as we show in section 3.5.2, the size of the additive term determines the threshold for stagnation in the worst-case accumulation of errors.

Theorem 3.5.2. *Let normal $A \in \mathbb{C}^{n\times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.2) *and* (3.3)*, respectively, and let $\hat{\mathcal{S}}_k = \operatorname{span}(\hat{X}_k)$, with $\hat{X}_k$ defined in* (3.26) *and R_k satisfying* (3.29)*. If $\cos\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V}) > \gamma_1\|A\|u/d$ and $\cos\theta_1(\hat{\mathcal{S}}_k, \mathcal{V}) > 0$, then*

$$\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V}) \leq \left|\frac{r(\lambda_{m+1})}{r(\lambda_m)}\right| \frac{\tan\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V})}{1-\alpha_k} + \beta_k, \tag{3.32}$$

where $\alpha_k \leq \gamma_1\|A\|u/(d\cos\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V}))$ and $\beta_k \leq \gamma_2\|A\|\kappa(\hat{X}_k C_k)u/\cos\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$.

Proof. Calculating directly as in the proof of Theorem 3.2.2, we have that

$$T(\hat{X}_k C_k, V_1) = (I - P_{\mathcal{V}})r(A)(Q'_{k-1} + R_k)C_k(V_1^*\hat{X}_k C_k)^{-1}. \tag{3.33}$$

We proceed by bounding the two terms in (3.33) corresponding to Q'_{k-1} and R_k. By Lemma 3.5.1, $\|(I-P_{\mathcal{V}})r(A)R_k C_k\| \leq \gamma_2\|A\|u$. If $\hat{X}_k C_k = Q'_k S_k$ is an economy-sized QR factorization, then the singular values of S_k and $\hat{X}_k C_k$ coincide, and

$$\|(V_1^*\hat{X}_k C_k)^{-1}\| = \|S_k^{-1}(V_1^* Q'_k)^{-1}\| \leq \left(\sigma_m(\hat{X}_k C_k)\cos\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})\right)^{-1}. \tag{3.34}$$

Since $\|\hat{X}_k C_k\| \leq 1$, we conclude that $\|(I - P_{\mathcal{V}})r(A)R_k C_k(V_1^*\hat{X}_k C_k)^{-1}\| \leq \beta_k$.

Now, rewrite $(V_1^*\hat{X}_k)^{-1} = (V_1^*(Q'_{k-1} + R_k))^{-1} r(\Lambda_1)^{-1}$ and expand

$$(V_1^*(Q'_{k-1} + R_k))^{-1} = \left(I + \sum_{j=1}^{\infty}(V_1^* Q'_{k-1})^{-j}(V_1^* R_k)^j\right)(V_1^* Q'_{k-1})^{-1}.$$

The Neumann series converges absolutely because $\|V_1^* R_k\| \leq \gamma_1\|A\|u/d$ by Lemma 3.5.1 and $\|(V_1^* Q'_{k-1})^{-1}\| = (\cos\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V}))^{-1} < d/(\gamma_1\|A\|u)$ by hypothesis; consequently, $\|(V_1^* Q'_{k-1})^{-1} V_1^* R_k\| < 1$. Since $T(Q'_{k-1}, V_1) = (I -$

$P_{\mathcal{V}})Q'_{k-1}(V_1^*Q'_{k-1})^{-1}$, we have

$$\begin{aligned}(I - P_{\mathcal{V}})r(A)Q'_{k-1}(V_1^*\hat{X}_k)^{-1} &= r(A)(I - P_{\mathcal{V}})Q'_{k-1}(V_1^*(Q'_{k-1} + R_k))^{-1}r(\Lambda_1)^{-1} \\ &= r(A)T(Q'_{k-1}, V_1)\left(I + \sum_{j=1}^{\infty}(V_1^*Q'_{k-1})^{-j+1}(V_1^*R_k)^j(V_1^*Q'_{k-1})^{-1}\right)r(\Lambda_1)^{-1}.\end{aligned}$$

Because the range of $T(Q'_{k-1}, V_1)$ is $\mathcal{V}^\perp$, the first factor on the right hand side is bounded by $\|r(A)T(Q'_{k-1}, V_1)\| \leq \|r(\Lambda_2)\| \tan\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V})$. The factor in parentheses is bounded above by $\sum_{j=0}^{\infty}\alpha_k^j = (1-\alpha_k)^{-1}$, where $\alpha_k = \|(V_1^*Q'_{k-1})^{-1}\|\|V_1^*R_k\|$. Therefore, we have the upper bound

$$\|(I - P_{\mathcal{V}})r(A)Q'_{k-1}(V_1^*\hat{X}_k)^{-1}\| \leq \|r(\Lambda_2)\|\|r(\Lambda_1)^{-1}\|\frac{\tan\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V})}{1-\alpha_k}.$$

Noting that $\|r(\Lambda_2)\| = |r(\lambda_{m+1})|$, $\|r(\Lambda_1)^{-1}\| = |r(\lambda_m)|^{-1}$, and collecting the bounds for the two terms in (3.33) establishes (3.32). □

The significance of Theorem 3.5.2 is that the errors in $\hat{X}_k$ that lie in the target subspace $\mathcal{V}$ impact only the refinement rate, and do not contribute to the additive term β_k in (3.32). This worst-case scenario occurs over one iteration only when the perturbations are aligned to maximally cancel the components of $\mathcal{V}$ present in the basis Q'_{k-1}. In fact, such errors are just as likely to align perfectly with the $\mathcal{V}$ components of Q'_{k-1} and improve the refinement rate by $(1+\alpha_k)^{-1}$, so the impact on the geometric convergence rate implied by (3.32) is probably not observed in practice.

On the other hand, the errors in $\hat{X}_k$ that lie in $\mathcal{V}^\perp$ degrade the expected refinement through the additive term β_k and, due to orthogonality, have a tangible effect on the convergence of subspace iteration in floating-point arithmetic. Note that the magnitude of β_k is proportional to the condition number of the basis $\hat{X}_k$ after column scaling. From sections 3.3 and 3.4, we know that $\beta_1 \approx u/d$ and

$\beta_k \approx u$ for $k \geq 2$, provided that $\hat{\mathcal{S}}_k$ and $\mathcal{V}$ do not become too close to orthogonal during the iteration.

3.5.2 Stability and stagnation

According to Theorem 3.5.2, the search subspace is refined by a factor comparable to Theorem 3.2.2, up to the size of the errors β_k introduced in $\mathcal{V}^\perp$, at each iteration. As we accumulate iterations, the errors in $\mathcal{V}^\perp$ are filtered out by $r(A)$ and, in the apt words of the authors of [144], "the dominant error term is the one most recently introduced." As long as $\cos\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ is bounded sufficiently far from zero for $k \geq 0$, the sequences α_k and β_k remain stable at the order of u/d and u (respectively) after the first iteration. In this case, we expect that $\hat{\mathcal{S}}_k$ converges geometrically toward $\mathcal{V}$ until a threshold of about u is reached, after which convergence stagnates. This is what we observe in Figures 3.1 and 3.2.

If $\cos\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ does become very small at some step in the iteration, then the one-step refinement bound may not imply any refinement in the search subspace at all: the iteration in (3.26) is potentially unstable. With a slight change of perspective, we now characterize the behavior of the iterates in (3.26) as $k \to \infty$, addressing both the stability and the threshold for stagnation in subspace refinement.

Let us introduce the constants $\rho = |r(\lambda_{m+1})|/|r(\lambda_m)|$, $\epsilon_1 = \gamma_1\|A\|u/d$, and $\epsilon_2 = \gamma_2\|A\|\hat{M}u$, where $\hat{M}$ is an $O(1)$ uniform bound on $\kappa(X_k C_k)$ for $k \geq 2$ (i.e., from Theorem 3.4.1). Consider the function

$$\Phi(\eta) = \frac{1}{1-\epsilon_2}\left(\frac{\rho\,\eta}{1-\epsilon_1(1+\eta)} + \epsilon_2\right). \tag{3.35}$$

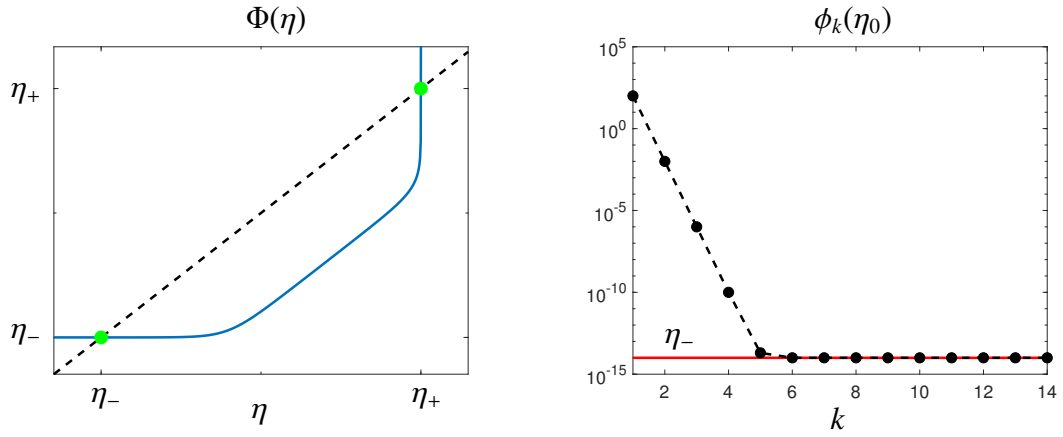

Figure 3.6: The dynamics of perturbed subspace iteration from (3.36). In the left panel, the solid line is the graph of $\Phi(\eta)$ and its fixed points (green circles) are marked at the intersections $\Phi(\eta_\pm) = \eta_\pm$. If $\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V})$ falls between the two fixed points (green circles), then the $\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ must converge geometrically to a threshold near the lower fixed point (see Theorem 3.5.4). In the right panel, the iterated map $\phi_k(\eta_0)$ (circles) is compared with the upper bound in Theorem 3.5.4 (dashed line) for $k = 0, \ldots, 14$. For this experiment, $\eta_0 = 100$, $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 10^{-14}$, and $\rho = 10^{-4}$.

Because $1/\cos\theta \leq 1 + \tan\theta$ when $0 \leq \theta \leq \pi/2$, we can rewrite Theorem 3.5.2 in the form $\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V}) \leq \Phi(\tan\theta_1(\hat{\mathcal{S}}_{k-1}, \mathcal{V}))$ when $k \geq 2$. We can understand the "worst-case" behavior of subspace iteration by studying the trajectory of $\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V})$, for some initial subspace $\hat{\mathcal{S}}_0$, obtained by iterating the map Φ.

Given $\eta_0 > 0$, let $\phi_k(\eta_0)$ denote the k-fold iteration of the map Φ on the point η_0, so that (letting $f \circ g$ denote the composition of two functions) we have

$$\phi_k(\eta_0) = \underbrace{\Phi \circ \cdots \circ \Phi}_{k}(\eta_0). \tag{3.36}$$

We call η_* a fixed point of Φ if $\Phi(\eta_*) = \eta_*$ and say that η_* is monotone attracting for $\Omega \subset [0, \infty)$ if $\phi_k(\eta) \to \eta_*$ monotonically as $k \to \infty$ for all $\eta \in \Omega$. After applying Φ to $\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V})$ k times, we see that (since $\Phi(\eta)$ is non-decreasing on $0 \leq \eta < -1 + 1/\epsilon_1$)

$$\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V}) \leq \phi_k\left(\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V})\right), \tag{3.37}$$

as long as $\phi_j(\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V})) < -1 + 1/\epsilon_1$ for each $j \geq 1$. Consequently, the fixed points of Φ and their attracting sets provide insight into the behavior of

$\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ in the limit $k \to \infty$, that is, about the convergence and stability of the iteration in (3.26).

Lemma 3.5.3. *Define the map* $\Phi : [0, -1 + 1/\epsilon_1) \to [0, \infty)$ *as in* (3.35)*, with constants* $0 < \rho < 1$ *and* $0 < \epsilon_1, \epsilon_2 < 1$*. Let*

$$\delta = \frac{1}{2\epsilon_1}\left[1 - \frac{\rho}{1-\epsilon_2} - \epsilon_1\left(1 - \frac{\epsilon_2}{1-\epsilon_2}\right)\right], \qquad \textit{and} \qquad \sigma = \frac{\epsilon_2(1-\epsilon_1)}{\epsilon_1(1-\epsilon_2)}.$$

If $\delta^2 > \sigma$*, then* Φ *has precisely two fixed points, given by* $\eta_\pm = \delta \pm \sqrt{\delta^2 - \sigma}$*. Moreover, the fixed point* η_- *is monotone attracting on* $[0, \eta_+)$*.*

Proof. Starting from the fixed point equation $\Phi(\eta_*) = \eta_*$, we multiply through by $(1 - \epsilon_1(1 + \eta_*))$ and collect powers of η_* to obtain the quadratic equation

$$\epsilon_1\eta_*^2 - \left[1 - \frac{\rho}{1-\epsilon_2} - \epsilon_1\left(1 - \frac{\epsilon_2}{1-\epsilon_2}\right)\right]\eta_* + \frac{\epsilon_2(1-\epsilon_1)}{1-\epsilon_2} = 0. \tag{3.38}$$

Applying the quadratic formula for the roots and rewriting in terms of δ and σ concludes the fixed-point calculation. Now, the quadratic on the left hand side of (3.38) is negative between the roots, which implies that $\Phi(\eta) > \eta$ for $0 < \eta < \eta_-$ and $\Phi(\eta) < \eta$ for $\eta_- < \eta < \eta_+$. This change at each fixed point implies that η_- attracts nearby points and that η_+ repels nearby points. Because Φ is non-decreasing and has no other fixed points, we conclude that η_- is monotone attracting on $[0, \eta_+)$. □

Lemma 3.5.3 shows that if $\tan\theta_1(\hat{\mathcal{S}}_0, \mathcal{V}) < \eta_+$, then $\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ must eventually be on the order of η_- or better for all sufficiently large k. Recalling that the constants ϵ_1 and ϵ_2 are on the order of u/d and u, respectively, and that ρ is the filtered spectral ratio, we estimate the size of the fixed points to be

$$\eta_- \approx \frac{\epsilon_2}{1-\rho}, \qquad \text{and} \qquad \eta_+ \approx -1 + \frac{1-\rho}{\epsilon_1}. \tag{3.39}$$

Crucially, the lower fixed point η_- is on the order of u, not u/d. Having established stability properties of the perturbed iteration in (3.26), we can now estimate the rate of convergence to the fixed point η_-.

Theorem 3.5.4. *Define the map* $\Phi : [0, -1+1/\epsilon_1) \to [0, \infty)$ *as in* (3.35)*, with constants* $0 < \rho < 1$ *and* $0 < \epsilon_1, \epsilon_2 < 1$*. Let* ϕ_k *denote the k-fold iteration of* Φ *as in* (3.36)*. If* Φ *satisfies the hypotheses of Lemma 3.5.3, then given* $0 \le \eta_0 < \eta_+$*, it holds that*

$$\phi_k(\eta_0) \le \tilde{\rho}^k \eta_0 + \tilde{\epsilon}_2 (1 - \tilde{\rho})^{-1}, \qquad k \ge 1. \tag{3.40}$$

Here, $\tilde{\rho} = \rho(1 - \epsilon_2)^{-1}(1 - \epsilon_1(1 + \eta_0))^{-1}$ *and* $\tilde{\epsilon}_2 = \epsilon_2(1 - \epsilon_2)^{-1}$.

Proof. Denote $\eta_k = \phi_k(\eta_0)$, for each $k \ge 1$. From the definitions of Φ and ϕ_k in (3.35) and (3.36), respectively, we compute that

$$\eta_k = \Phi(\eta_{k-1}) = \tilde{\rho}\eta_{k-1} + \tilde{\epsilon}_2 \qquad k \ge 1. \tag{3.41}$$

By hypothesis, Lemma 3.5.3 applies, so $\eta_k \to \eta_-$ monotonically as $k \to \infty$ and, consequently, $\tilde{\rho} < 1$. Therefore, we iterate (3.41) $k - 1$ times to obtain

$$\eta_k = \tilde{\rho}^k \eta_0 + \tilde{\epsilon}_2 \sum_{j=0}^{k-1} \tilde{\rho}^j \le \tilde{\rho}^k \eta_0 + \tilde{\epsilon}_2/(1 - \tilde{\rho}).$$

Plugging the original parameters back in to $\tilde{\rho}$ and $\tilde{\epsilon}_2$ establishes (3.40). □

Thus, Theorem 3.5.4 and (3.37) demonstrate that the reduction of $\tan\theta_1(\hat{\mathcal{S}}_k, \mathcal{V})$ down to the order of η_- is approximately geometric with rate close to ρ. So (accounting for the fact that the additive perturbation term is actually on the order of u/d in the first iteration) it takes approximately $1 + \log(\eta_-)/\log(\rho)$ steps for $\hat{\mathcal{S}}_k$ to converge to within order u of $\mathcal{V}$, as measured by the tangent of the principal angle between the two subspaces.

3.6 Non-normal matrices

We now consider the case of an $n \times n$ diagonalizable matrix A whose eigenvectors are not orthogonal. Although a straightforward extension of Proposition 3.3.1 shows that the condition number of X_1 still scales, generically, like $1/d$ (see Proposition 3.6.1 below), the effect of a dangerous eigenvalue on subsequent iterates, $X_2, X_3, \ldots$, computed via (3.5) is distinct in the non-normal case due to interactions among non-orthogonal modes. In fact, the condition numbers of the computed iterates do not improve during subsequent iterations unless approximate eigenvectors (i.e., from Ritz vectors) are incorporated into the subspace iteration (see Algorithm 4). Even with this modification, the condition numbers may remain large after one iteration when d is very small (loosely, when $d \ll \sqrt{u}$), unlike the normal case. Here, we demonstrate that $\kappa(X_k)$ is typically reduced in step with the error in the Ritz vectors and that $\kappa(X_k) \approx (u/d)^k$ in the best case (i.e., when $|r(\lambda_{m+1})|/|r(\lambda_m)| \approx u$ and the Ritz vectors are well-conditioned at each iteration).

When A does not have an orthogonal basis of eigenvectors (but is still diagonalizable), the orthogonal spectral projectors $v_i v_i^*$ that diagonalize the filter in (3.10) are replaced by oblique spectral projectors, so that

$$r(A)x = \sum_{i=1}^{n} r(\lambda_i) \frac{w_i^* x}{w_i^* v_i} v_i = \frac{w_1^* x}{(de^{i\theta})(w_1^* v_1)} v_1 + O(1), \qquad \text{as} \qquad d \to 0. \tag{3.42}$$

Here, $w_1, \ldots, w_n$ are the left eigenvectors of A, satisfying $w_i^* A = \lambda_i w_i^*$ with $\|w_i\| = 1$ for $i = 1, \ldots, n$. Likewise, the spectral decomposition in (3.2) is replaced by

$$A = V_1 \Lambda_1 W_1^* + V_1 \Lambda_2 W_2^*, \tag{3.43}$$

where the ith column of $W = [W_1 \; W_2]$ is $(w_i^* v_i)^{-1} w_i$. With this normalization, V and W form a biorthogonal system, meaning that $W^* V = I$, I being the $n \times n$

identity matrix. In the biorthogonal system, the dangerous eigenvalue amplifies the w_1 component in the input x along the v_1 direction in the output $r(A)x$. Due to biorthogonality, v_1 and w_1 are parallel only when v_1 is orthogonal to $v_2, \ldots, v_n$.

3.6.1 First iteration

To develop a sense of how non-normality impacts the conditioning of the iterates, it is worthwhile to revisit the analysis of $\kappa(X_1)$ in Proposition 3.3.1 when A is only diagonalizable. While the condition number of X_1 is still $O(1/d)$ as $d \to 0$, the constants in the bound now depend on the structure of the left and right eigenvectors. This is because the stretching and shrinking actions of A no longer belong solely to its eigenvalues, but can be enhanced or attenuated by interactions among non-orthogonal eigenvectors. We denote the smallest singular values of V_1 and W_1 by $\sigma_m(V_1)$ and $\sigma_m(W_1)$, respectively.

Proposition 3.6.1. *Let diagonalizable $A \in \mathbb{C}^{n\times n}$ and $r : \Lambda \to \mathbb{C}$ satisfy* (3.43) *and* (3.3)*, respectively, and given orthonormal $Q_0 \in \mathbb{C}^{n\times m}$, let $X_1 = r(A)Q_0$. If $U_1 = \mathrm{qf}(W_1)$ and $U_1^* Q_0$ has full rank, then the condition number of X satisfies*

$$\frac{\|w_1^* Q_0\|/|w_1^* v_1|}{d\kappa(V)|r(\lambda_2)|} \lesssim \kappa(X_1) \leq \left|\frac{r(\lambda_1)}{r(\lambda_m)}\right| \frac{\kappa(V)\|(U_1^* Q_0)^{-1}\|}{\sigma_m(V_1)\sigma_m(W_1)}, \quad \text{as} \quad d \to 0. \tag{3.44}$$

Proof. The steps of the proof are essentially identical to those in Proposition 3.3.1 if (3.42) and (3.43) are used in place of (3.2) and (3.10), so we emphasize the adaptations made for non-orthogonal eigenvectors. For the largest singular value of X_1, we bound $\sigma_1(X_1) = \|r(A)Q_0\| \leq \kappa(V)|r(\lambda_1)|$, since $|r(\lambda_1)| \leq \|r(A)\| \leq \kappa(V)\|r(\Lambda)\|$ in the non-normal case. If we use (3.43) to decompose X_1 as in (3.14), the singular values of $r(A)W_1^* Q_0$ do not tell us directly about the singular values of X_1 because V is not unitary. However, if $\Omega_1 R_1 = V_1$ and $\Omega_2 R_2 = V_2$ are

economy-sized QR factorizations, we can decompose

$$r(A)Q_0 = \begin{bmatrix} \Omega_1 & \Omega_2 \end{bmatrix} \begin{bmatrix} R_1 r(\Lambda_1) W_1^* Q_0 \\ R_2 r(\Lambda_2) W_2^* Q_0 \end{bmatrix}.$$

Since Ω_1 and Ω_2 have orthonormal columns, we apply the argument in the proof of Proposition 3.3.1 to obtain the bound $1/\sigma_m(X_1) \leq \|(R_1 r(\Lambda_1) W_1^* Q_0)^{-1}\|$. Now, R_1 has the same singular values as V_1 and $\|R_1^{-1}\| = 1/\sigma_m(R_1)$, so we have that

$$\kappa(X_1) \leq \frac{|r(\lambda_1)|}{|r(\lambda_m)|} \frac{\kappa(V)\|(W_1^* Q_0)^{-1}\|}{\sigma_m(V_1)}. \tag{3.45}$$

The upper bound in (3.44) follows by substituting the QR decomposition $U_1 S_1 = W_1$ into (3.45) and noting that $\|S_1^{-1}\| = 1/\sigma_m(W_1)$.

A lower bound on $\sigma_1(X_1)$ follows directly from (3.42), analogous to (3.15). For the lower bound on $1/\sigma_m(X_1)$, we can use (3.42) to write X_1 as a rank one perturbation of the matrix

$$\tilde{N}_2 = V \mathrm{diag}(0, \lambda_2, \ldots, \lambda_n) W^* Q_0.$$

We have that $\sigma_1(N_2) \leq \|V\|\|W^*\||r(\lambda_2)| = \kappa(V)|r(\lambda_2)|$, where the equality is due to biorthogonality, which implies that $W^* = V^{-1}$. By interlacing, we find that $1/\sigma_m(X_1) \geq 1/(\kappa(V)|r(\lambda_2)|)$, establishing the asymptotic lower bound in (3.44). □

When A is normal, Proposition 3.6.1 reduces to Proposition 3.3.1. In the non-normal case, ill-conditioning in the eigenvectors, reflected in $\kappa(V)$, widens the interval between the upper and lower bounds. Similarly, ill-conditioning in the target eigenvectors, captured by the smallest singular values of V_1 and W_1 (since the columns of both matrices have unit norm), may further widen the gap. On the other hand, the dangerous eigenvalue itself is ill-conditioned when

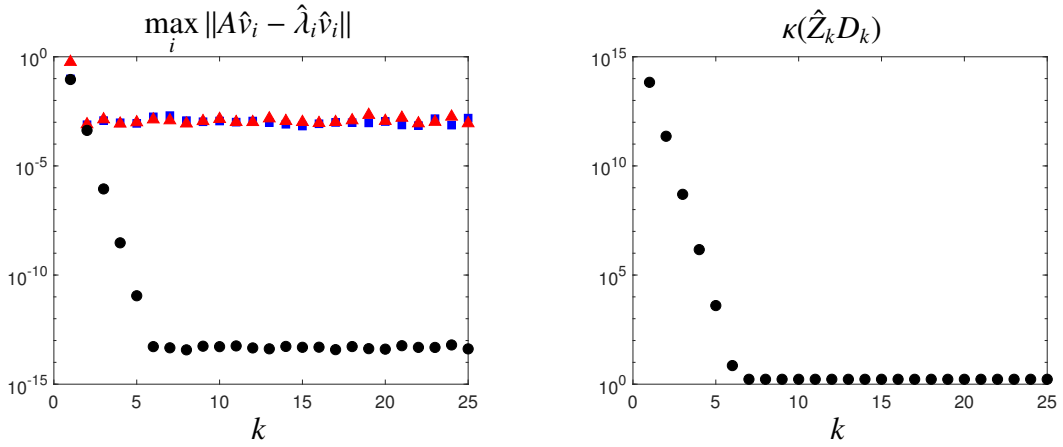

Figure 3.7: *Dangerous eigenvalues of a non-normal matrix.* The eigenvalues and rational filter are identical to the setup displayed in Figure 3.3, however, this matrix has non-orthogonal eigenvectors and the dangerous eigenvalue has been moved to distance $d = 10^{-13}$ from the pole at $z = 10$. On the left, the maximum residual of 10 target eigenpairs after each iteration of (3.5) (blue squares), a variant of subspace iteration based on Schur vectors [125, ch. 5.2] (red triangles), and a variant based on approximate eigenvectors, described in Algorithm 4 (black circles). On the right, the condition number of the iterates $\hat{Z}_k D_k$ (D_k scales the columns of $\hat{Z}_k$ to have unit norm) decreases in step with residuals from Algorithm 4, at a rate of about u/d per iteration.

$|w_1^* v_1|$ is small.[8] The left hand side of (3.44) illustrates how this may enhance the amplifying effects of the dangerous eigenvalue, increasing the asymptotic lower bound to $d|w_1^* v_1|^{-1}$. Broadly speaking, the widening gap between upper and lower bounds indicates that our picture is blurred in the non-normal case because the structure of the eigenvectors plays a key role. The extent of the damage may depend on where the ill-conditioning in V is concentrated.

3.6.2 Iterating with orthonormal bases

Now that we understand the interaction between non-normality and dangerous eigenvalues in the initial iteration, we are ready to examine subsequent iterations. As in section 3.4, we focus on the coordinates of Q_1 in the eigenvector

[8] With $\|v_i\| = \|w_i\| = 1$, the quantity $|w_i^* v_i|^{-1}$ is Wilkinson's condition number for λ_i, measuring the first-order sensitivity of the eigenvalue to infinitesimal perturbations in A [168, pp. 88–89].

basis, partitioned into blocks as

$$W_1^* Q_1 = \begin{bmatrix} w_1^* q_1^{(1)} & w_1^* \tilde{Q}_1 \\ \tilde{W}_1^* q_1^{(1)} & \tilde{W}_1^* \tilde{Q}_1 \end{bmatrix}. \tag{3.46}$$

The critical observation about (3.46) is that, in contrast to the normal case, the upper right block is not small (the lower-left block remains small). Although the columns of $\tilde{Q}_1$ are still nearly orthogonal to v_1, the eigenvectors v_1 and w_1 are only parallel in the special case that v_1 is orthogonal to $v_2, \ldots, v_n$. Consequently, $w_1^* \tilde{Q}_1$ is typically $O(1)$ and, when we compute $X_2 = r(A)Q_1$, the components in each column of Q_1 in the w_1 direction will be amplified according to (3.42). Each column of X_2 will be dominated by v_1 at magnitude $O(1/d)$ and X_2 is just as ill-conditioned as X_1 in the first iteration. This line of thinking seems to indicate that, when $r(A)$ is repeatedly applied to an orthonormal basis, subspace iteration for non-normal matrices must stagnate at an accuracy of $\approx u/d$ due to ill-conditioning in the iterates $X_1, X_2, \ldots$.

To illustrate, we return to the experimental setup illustrated in (3.3). We select the same rational filter and a matrix with the same eigenvalues, but now the eigenvector matrix is not orthogonal. The condition number of the eigenvector matrix is $\approx 10^2$, but the target eigenvectors themselves are not far from orthogonal. Figure 3.7 shows the maximum residual of the computed target eigenpairs after each of the first 10 iterations of (3.5). We also compare with a modified subspace iteration based on Schur vectors that is commonly used to compute eigenvalues of non-normal matrices [125, ch. 5.2]. Both iterations apply the rational filter directly to an orthonormal basis for the search space and the residuals stagnate near u/d in both cases.

Algorithm 4 Filtered subspace iteration with Rayleigh–Ritz projection.

Input: Given $A \in \mathbb{C}^{n\times n}$, $r : \Lambda \to \mathbb{C}$, and $Y_0 \in \mathbb{C}^{n\times m}$.

1: **for** $k = 1, 2, \ldots$ **do**
2: Apply the filter $Z_k = r(A)Y_{k-1}$.
3: Compute orthonormal basis $Q_k = \mathrm{qf}(Z_k)$.
4: Form $A_k = Q_k^* A Q_k$ and diagonalize $A_k = U_k \Theta_k U_k^{-1}$.
5: Set $Y_k = Q_k U_k$.
6: **end for**

Output: Approximate eigenvalue matrix Θ_k and eigenvector matrix Y_k.

3.6.3 Iterating with approximate eigenvectors

What can we do to improve the conditioning of the iterates and the accuracy in the target eigenpairs? Consider another common variant of subspace iteration shown in Algorithm 4, which forms the iterates $Z_1, Z_2, \ldots$ by applying $r(A)$ to approximate eigenvectors constructed from the Ritz vectors at each iteration. Let us partition $W_1^* Y_1$ in the usual way,

$$W_1^* Y_1 = \begin{bmatrix} w_1^* y_1^{(1)} & w_1^* \tilde{Y}_1 \\ \tilde{W}_1^* y_1^{(1)} & \tilde{W}_1^* \tilde{Y}_1 \end{bmatrix} = \begin{bmatrix} e & f \\ g & H \end{bmatrix}, \tag{3.47}$$

where $\tilde{W}_1$ and $\tilde{Y}_1$ denote the last $m-1$ columns of W_1 and Y_1, respectively. Now, because the left and right eigenvectors are biorthogonal, w_1^* annihilates the remaining target eigenvectors $v_2, \ldots, v_m$, so the upper right block f in (3.47) is small when the columns of Y_1 are a good approximation to the target eigenvectors. In turn, small $\|f\|$ mitigates the amplification of v_1 in the last $m-1$ columns of Z_2.

Unfortunately, the behavior of approximate eigenvectors computed with (4) may vary widely for general non-normal matrices. In exact arithmetic, their accuracy will depend on the rational filter through the eigenvalues of $r(A)$ and on interactions among non-orthogonal eigenvectors. This can delay convergence

and may lead to instability on a computer. In floating-point arithmetic, it is further limited by the accuracy in the computed orthonormal basis and Ritz vectors. Despite these difficulties, we can glean some practical insight into a distinct feature of the non-normal setting by examining a "best-case" situation.

Let us suppose that the non-normal effects are relatively mild, that $r(\cdot)$ filters out the unwanted eigenvalues to unit round-off or better (as in Figure 3.3), and that the Ritz vectors are computed accurately at each iteration. In this regime, the accuracy of the approximate eigenvectors Y_1 is limited mainly by the accuracy in the computed orthonormal basis, $\hat{Q}_1$, and we can focus on the influence of the dangerous eigenvalue in the second iteration (and beyond). From our analysis of the first iteration in section 3.6.1, we expect that $\|\hat{Q}_1 - Q_1\| \approx u/d$ and, therefore, (by our assumptions on the filter and the Ritz vectors) that $\|\hat{Y}_1 - V_1\| \approx u/d$.

Interestingly, the order of magnitude of block f in (3.47) is distinctly different from the analogous block b in the normal case. Instead of the perfect balancing between b and $r(\lambda_1)$ when the filter is applied (leading to perfectly well-conditioned columns of X_2), we have the order-of-magnitude estimate $\|f\||r(\lambda_1)| \approx u/d^2$. In other words, v_1 may still dominate each column of Z_2 when $d \ll \sqrt{u}$, but the gap in magnitude between the v_1 component and the remaining target components in the last $m - 1$ columns is reduced by a factor of u/d at the second iteration. Figure 3.7 illustrates this phenomenon in action with the same matrix and rational filter used for the experiments in section 3.6.2. The residuals in the target eigenpairs decrease geometrically with rate u/d (left panel), mirroring the reduction in the condition number of the iterates Z_k (after scaling columns to have unit norm, right panel).

Thus, for a mildly non-normal matrix with a dangerous eigenvalue at distance $d \ll \sqrt{u}$ from a pole of $r(\cdot)$, two iterations are not usually enough to remove the adverse influence of the dangerous eigenvalue. Instead, the target residuals and the errors in the computed orthonormal basis are often refined in step down to the unit round-off (depending on the sensitivity of the target eigenpairs). As in the normal case, round-off errors caused by the dangerous eigenvalue may even go unnoticed when the rational filter is mediocre so that the noise in the unwanted directions is dominated by poor filtering.

3.7 Restarting Arnoldi

Now that we understand the right hand side of Figure 3.1, let us examine the stagnation of Arnoldi with shift-and-invert enhancement illustrated in the left hand panel of the same figure. Unlike subspace iteration, which applies $r(A)$ iteratively to a subspace of fixed dimension, Arnoldi refines the subspace by expanding it. Given an initial unit vector $q_1 \in \mathbb{C}^n$, shift-and-invert Arnoldi computes the iterates

$$y_k = s(A)q_{k-1}, \qquad q_k = \mathrm{mgsr}(y_k\,;\, q_1, \ldots, q_{k-1}), \tag{3.48}$$

with the expression $\mathrm{mgrs}(\cdot)$ indicating that y_k is orthogonalized against $q_1, \ldots, q_{k-1}$ using modified Gram–Schmidt with full reorthogonalization [139, pp. 307–308].

After k steps of (3.48), we have an $n \times k$ orthonormal basis $Q_k = [q_1 \cdots q_k]$ and we can approximate eigenpairs of A in one of two ways:

- Directly from the eigenpairs of the upper Hessenberg matrix H_k generated

from the weights calculated during modified Gram–Schmidt [156, p.253].

- A Rayleigh–Ritz step by computing eigenpairs of $A_k = Q_k^* A Q_k$.

Usually, the upper Hessenberg matrix is the method of choice because it does not require any additional matrix-vector products. However, when a dangerous eigenvalue is present, the upper Hessenberg matrix in the Arnoldi decomposition of $s(A)$ typically has norm $\|H_k\| = O(d^{-1})$: this makes the accurate calculation of the remaining target eigenvalues challenging for standard dense solvers. To focus on the accuracy in the computed basis Q_k, we work with A_k, but we revisit H_k at the end of this section.

In keeping with the analysis in sections 3.3 and 3.4, we can understand the accuracy in the computed orthonormal basis $\hat{Q}_k$ through the conditioning of the matrix

$$Y_k = \begin{bmatrix} q_1 & \cdots & q_{k-1} & y_k \end{bmatrix}, \qquad k = 2, 3, 4, \ldots. \tag{3.49}$$

The matrix Q_k from the Arnoldi iterations is precisely the QR factorization of Y_k obtained by orthogonalizing y_k against the previous $(k-1)$ columns, which are already an orthonormal set. If y_k is not too closely aligned with $\operatorname{span}(q_1, \ldots, q_{k-1})$, then the matrix Y_k is well-conditioned, at least after a simple column scaling. Consequently, $Q_k = \operatorname{qf}(Y_k)$ is not too sensitive to perturbations caused by round-off in Y_k, as discussed in section 3.3.1. However, if y_k is closely aligned with any of the previous columns, the smallest singular value of Y_k will be close to zero and Q_k will be very sensitive to round-off in Y_k.

This perspective provides an explanation for the stagnation observed in Figure 3.1. When q_1 is chosen randomly, $v_1^* q_1$ is generically $O(1)$ (as $d \to 0$). After

applying the shift-and-invert filter, we calculate (as usual) that

$$y_2 = s(A)q_1 = \frac{v_1^* q_1}{de^{i\theta}} v_1 + O(1).$$

After we orthogonalize y_2 against q_1 to compute q_2, then for some constant h_2, we have

$$q_2 = h_2 \frac{v_1^* q_1}{de^{i\theta}} (v_1 - (v_1^* q_1) q_1) + O(1). \tag{3.50}$$

In other words, q_2 may not be dominated by v_1, but $\text{span}(q_1, q_2)$ contains approximations to v_1 that are accurate to $O(d)$.

Now, note that q_2 is not near orthogonal to v_1 unless q_1 happens to be very closely aligned with v_1. This means that the subsequent iterate y_3 is also aligned with v_1, and therefore with a vector in $\text{span}(q_1, q_2)$, to about order d. Consequently, the matrix Y_3 is ill conditioned and we expect that Q_3, and in particular q_3, can only be accurate to about order u/d when computed in floating-point precision. Moreover, q_3 is not dominated by v_1 and this process repeats, so that each iterate y_k is closely aligned with v_1 in $\text{span}(q_1, q_2)$, leading to errors in q_k on the order of u/d.

In our discussion above, note that y_3 was only aligned with v_1, and thus close to $\text{span}(q_1, q_2)$, because q_2 was not nearly orthogonal to v_1. Unlike in subspace iteration, the dangerous direction is never rendered harmless by orthogonalizing directly against it! The geometric picture of the iterates $y_2, y_3, y_4, \ldots$ being attracted to v_1 as a result of $q_2, q_3, q_4, \ldots$ not being sufficiently orthogonal to v_1 suggests an interesting fix. If we restart the Arnoldi iteration with the Ritz approximation associated to v_1 after the second iteration, the picture changes drastically. Again, y_2 is aligned with v_1, but now it is orthogonalized against $q_1 = v_1 + O(d)$. The corresponding q_2 may not be particularly accurate, but this doesn't matter much: the point is that all subsequent iterates are orthogonalized

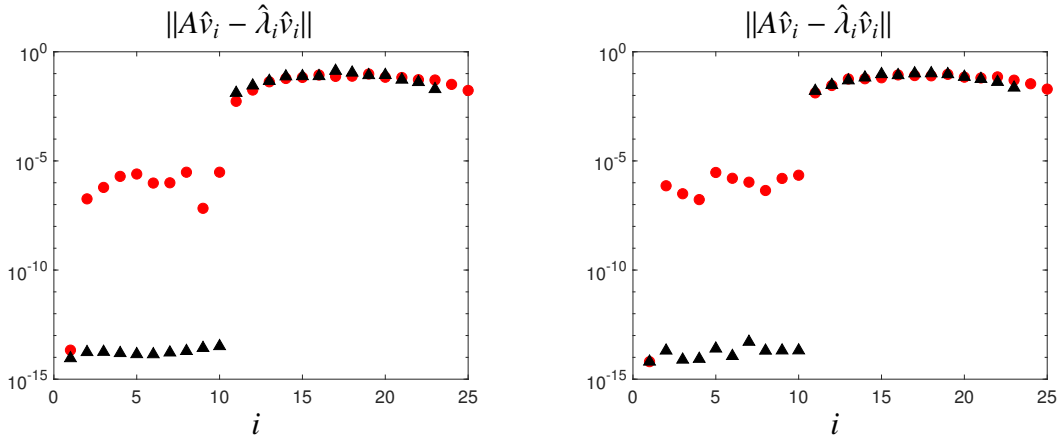

Figure 3.8: After restarting Arnoldi with Ritz vectors that are nearly orthogonal to the dangerous direction, Arnoldi produces approximations to the target eigenpairs with accuracy near the unit round-off. Both plots compare eigenpair residuals after 25 steps of shift-and-invert Arnoldi with no restart (red circles) to eigenpair residuals obtained after 25 total steps of shift-and-invert Arnoldi with the Ritz restart. The eigenpairs were extracted from $Q_{25}^* A Q_{25}$ in the left panel and from the Hessenberg matrix H_{25} in the right panel.

against the dangerous direction (via q_1) up to order $O(d)$. Analogous to the situation encountered in subspace iteration, the iterates $y_3, y_4, y_5, \ldots$, are no longer dominated by v_1 and, consequently, $q_3, q_4, q_5, \ldots$ can be computed accurately. In a sense, we are tricking Arnoldi into running in the orthogonal complement of the dangerous direction.

Figure 3.8 demonstrates this restart strategy in action. As we saw earlier, 25 iterations of shift-and-invert Arnoldi leads to stagnation in 9 of the 10 target eigenpairs. However, we can resolve all 10 target eigenpairs to unit round-off accuracy in 25 iterations if we restart with the Ritz vector corresponding to the dangerous direction after the second iteration. The right Ritz vector is easy to identify: it is most closely aligned with the second iteration y_2. It is worth noting that the Ritz restart strategy seems to be equally successful when eigenpairs are extracted from the Hessenberg matrix H_k instead of $Q_k^* A Q_k$ (see the right panel in Figure 3.8).

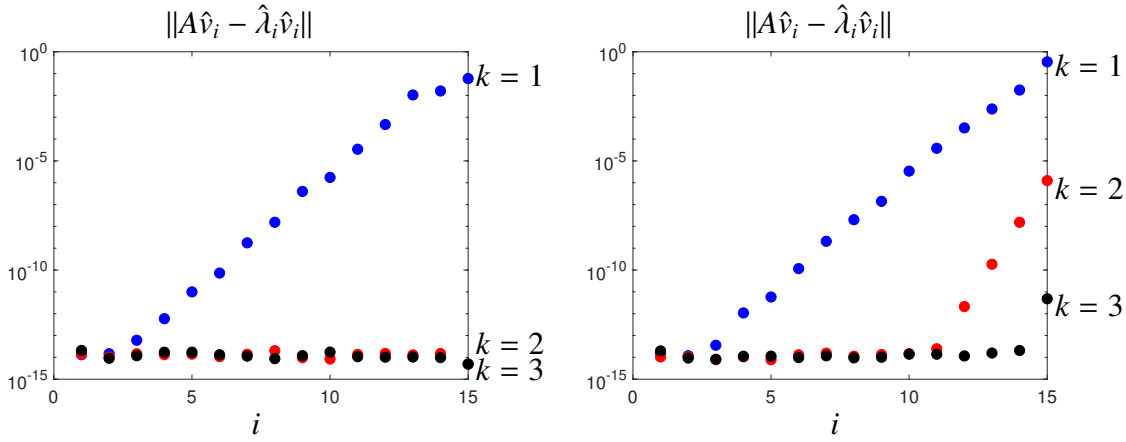

Figure 3.9: Convergence with multiple dangerous eigenvalues. On the left, two iterations of rational subspace iteration with a high-quality filter ($\ell = 32$ poles) reduce the residuals of 15 target eigenpairs to the order of u, despite exponential clustering of the target eigenvalues at a pole. On the right, three iterations of rational subspace iteration with a medium-quality filter ($\ell = 8$) reduce the residuals of 15 target eigenpairs geometrically (see Theorem 3.5.4), with no observable interference from the exponentially clustered eigenvalues.

3.8 Multiple dangerous eigenvalues

For simplicity, we have analyzed the case where there is only one dangerous eigenvalue, however, other situations may arise naturally in practice. When eigenvalues are heavily clustered, many dangerous eigenvalues may surround a single pole at various distances. The analysis in sections 3.3 to 3.5 may be adapted to these cases and our numerical experiments conform accordingly. To illustrate, we generate a 200×200 symmetric matrix with 15 target eigenvalues in $[10, 15]$, and employ a filter with equally-spaced poles on a circular contour centered at 12.5. There are two dangerous eigenvalues at $10 + 10^{-13}$, and the other 13 target eigenvalues are clustered exponentially at the pole, taking the values $10 + 10^{-i}$, $i = 0, 1, 2, \ldots, 12$. Figure 3.9 shows the results with two rational filters: one excellent and one of medium quality. Just as in sections 3.3 and 3.4, we see that twice is enough if the filter quality is high; with a poorer filter, all iterates beyond the first behave as if there were no dangerous eigenvalues.

CHAPTER 4

COMPUTING SPECTRAL MEASURES

As we saw in section 1.4, operators with continuous spectra are not diagonalized by their eigenfunctions alone; instead, spectral measures (c.f. section 1.4.1) provide a useful analogue of diagonalization. Computing spectral measures is subtle, and previous efforts have mainly focused on operators where analytical formulas or heuristics are available (see section 4.1). In this chapter,[1] we develop a general framework for computing approximations to spectral measures of operators, building on [30, 83], that only requires two capabilities:

1. A numerical solver for shifted linear equations, i.e., $(\mathcal{L}-z)u = f$ with $z \in \mathbb{C}$.
2. Numerical approximations to inner products of the form $(u, f)_{\mathcal{H}}$.

Our algorithms are designed to evaluate smoothed approximations of a spectral measure, μ_f, of $\mathcal{L}$ with respect to $f \in \mathcal{H}$ (see section 1.4.1). This means that we compute samples from a smooth function g_ϵ, with smoothing parameter $\epsilon > 0$, that converges weakly to μ_f [16, Ch. 1]. That is,

$$\int_{\mathbb{R}} \phi(y) g_\epsilon(y)\, dy \to \int_{\mathbb{R}} \phi(y)\, d\mu_f(y), \qquad \text{as} \qquad \epsilon \downarrow 0,$$

for any bounded, continuous function ϕ. Approximation properties and explicit convergence bounds are studied in sections 4.2 and 4.3.

[1] This chapter is based on sections 1-6 of a paper by Alex Townsend, Matt Colbrook, and me [34]. The collaboration began during Matt's visit to Cornell in 2018, when I proposed a heuristic spectral density algorithm based on Stone's formula. Matt led the development of the convergence theory for high-order kernels, while I led the development of stable and efficient numerical implementations for differential and integral operators. Matt has also written on theoretical foundations in [30].

The chapter is organized as follows. We survey applications and existing algorithms in section 4.1, introduce our computational framework in section 4.2 and develop versions with improved convergence properties in section 4.3, and discuss algorithmic issues and practical implementations in section 4.4.

4.1 Applications of spectral measures

Spectral measures appear in many traditional topics of applied analysis, such as ordinary (ODEs) and partial differential equations (PDEs), stochastic processes, orthogonal polynomials, and random matrix theory. Here, we give a brief survey of existing algorithms for computing μ_f and closely related quantities.

4.1.1 Particle and condensed matter physics

Spectral measures are prominent in quantum mechanics [72, 119], where a self-adjoint operator $\mathcal{L}$ represents an observable quantity, and μ_f describes the likelihood of different outcomes when the observable is measured (see section 4.3.2). In this setting, $f \in \mathcal{H}$ with $\|f\| = 1$ represents a quantum state. For example, in quantum models of interacting particles, spectral measures of many-body Hamiltonians are used to study the response of a quantum system to perturbations [50]. In condensed matter physics, spatially-resolved statistical properties of materials are analyzed using the local density-of-states[2] (LDOS) of an $n \times n$ matrix A_n [90, Ch. 6.4], which is the spectral measure of A_n taken with respect to

[2]This is distinct from the global density-of-states (DOS), which is formally obtained from the LDOS via an averaging procedure [90, Ch. 6.4].

a vector b [98]. Here, A_n is typically a discretized or truncated Hamiltonian and one is interested in the thermodynamic limit $n \to \infty$, so that A_n is too large to compute a full eigenvalue decomposition.

There are two main classes of numerical methods for computing these measures. One class constructs smooth global approximations of the measure with explicit moment-matching procedures [93,130,166], while another class exploits a connection between the spectral measure and the resolvent operator to evaluate samples from a smoothed approximation to the measure [15, 51, 75]. For example, the so-called recursion method [15,75] evaluates the resolvent of tridiagonal Hamiltonians using associated continued-fraction expansions. Resolvent techniques to compute the DOS of finite matrices also appear in the study of random matrices and Schrödinger operators, where the connection is made through the Stieltjes transform [12,23].

The resolvent of an operator $\mathcal{L}$ with spectrum $\Lambda(\mathcal{L})$ is given by [87, p. 173]

$$\mathcal{R}_{\mathcal{L}}(z) = (\mathcal{L} - z)^{-1}, \qquad z \in \mathbb{C} \setminus \Lambda(\mathcal{L}). \tag{4.1}$$

In section 4.2, we evaluate a smoothed approximation of μ_f by evaluating the resolvent function $(\mathcal{R}_{\mathcal{L}}(z)f, f)$ in the upper half-plane, i.e., $\mathrm{Im}(z) > 0$. Our approach is closely related to the second class of methods developed for operators in quantum mechanics. A key theme in the above moment-matching and resolvent-based approaches is smoothing, which is introduced by convolution with a smoothing kernel to avoid difficulties associated with the singular part of the measure [98]. The smoothed approximations of the spectral measures that we compute in sections 4.2 and 4.3 also have the form of $K_\epsilon * \mu_f$, where K_ϵ is a smoothing kernel with smoothing parameter $\epsilon > 0$.

Our framework is "discretization-oblivious," in the sense that it directly re-

solves the spectral measure of an infinite dimensional $\mathcal{L}$, and not an underlying discretization. This means that our algorithms do not suffer from spectral pollution.[3] Moreover, our framework can be used with any accurate numerical method for solving linear operator equations and computing inner products, making it applicable to differential, integral, and lattice operators. Achieving a discretization-oblivious framework requires balancing refinement in the computation of $(\mathcal{R}_{\mathcal{L}}(z)f, f)$ and refinement in the smoothing parameter, which we do in a principled way (see section 4.2.3).

4.1.2 Time evolution and spectral density estimation

Spectral measures provide a useful lens when studying processes that evolve over time. Suppose that $u : [0, T] \to \mathcal{H}$ evolves over time according to the abstract Cauchy problem

$$\frac{du}{dt} = -i\mathcal{L}u, \qquad u(0) = f \in \mathcal{H}, \tag{4.2}$$

where $\mathcal{L}$ is a self-adjoint operator. For example, (4.2) could describe the evolution of a quantum system according to the Schrödinger equation [99]. Semigroup theory [114] shows that the solution to (4.2) is given by the operator exponential $e^{-i\mathcal{L}t}f$. The autocorrelation function of u is of interest, i.e.,

$$(u(t), f) = (e^{-i\mathcal{L}t}f, f) = \int_{\mathbb{R}} e^{-iyt}\, d\mu_f(y), \qquad t \in [0, T],$$

which can reveal features that persist over time [145]. This interpretation of a time evolution process is quite flexible and can be adapted to describe many

[3] Spectral pollution is the phenomenon of eigenvalues of finite discretizations/truncations clustering at points not in the spectrum of $\mathcal{L}$ as the truncation size increases.

signals, u, generated by PDEs [43, 84, 132] and stochastic processes [62, 86] [121, Ch. 7].

In certain evolution processes, μ_f is referred to as the spectral density of u [35]. The task of spectral density estimation is to recover μ_f from samples of $u(t)$ [141, Ch. 1.5]. A popular technique used in spectral density estimation, related to statistical kernel density estimation [161, 162], reconstructs a smoothed approximation to μ_f by convolving the empirical measure (a discrete measure supported on the observed samples) with a smoothing kernel [111, 117]. The particular choice of smoothing kernel affects the convergence properties of the smoothed spectral density [112].

In analogy to the variance-bias tradeoff encountered when selecting the smoothing parameter in statistical kernel density estimation [113, 120], our smoothed approximations, $K_\epsilon * \mu_f$, exhibit a tradeoff between numerical cost and smoothing (see section 4.2.3). In section 4.3, we adapt arguments from kernel density estimation to determine what properties a smoothing kernel needs to achieve a high-order of convergence in the smoothing parameter.

4.2 Resolvent-based approach to evaluate the spectral measure

The key to our framework for computing spectral measures is the resolvent of $\mathcal{L}$ (see (4.1)). A classical result in operator theory is Stone's formula, which says that the spectral measure of $\mathcal{L}$ can be recovered from the jump in the resolvent $\mathcal{R}_{\mathcal{L}}(z)$ across the real axis [142] [119, Thm. VII.13]. More precisely, if we select

$\epsilon > 0$ and regard $\mathcal{R}_{\mathcal{L}}(x + i\epsilon)$ as a function of the real variable x, then we have that

$$\tfrac{1}{2\pi i}((\mathcal{R}_{\mathcal{L}}(\cdot + i\epsilon) - \mathcal{R}_{\mathcal{L}}(\cdot - i\epsilon))f, f) = \tfrac{1}{\pi}\mathrm{Im}((\mathcal{R}_{\mathcal{L}}(\cdot + i\epsilon)f, f)) \to \mu_f \text{ as } \epsilon \downarrow 0. \tag{4.3}$$

Here, the equality is due to the conjugate symmetry of $\mathcal{R}_{\mathcal{L}}(z)$ across the real axis and the limit should be understood in the sense of weak convergence of measures.

Stone's formula is a consequence of the functional calculus identity

$$(\mathcal{R}_{\mathcal{L}}(x + i\epsilon)f, f) = \int_{\mathbb{R}} \frac{d\mu_f(y)}{y - (x + i\epsilon)}. \tag{4.4}$$

By using (4.4) to rewrite (4.3), we arrive at an expression for the jump over the real axis as a convolution of the spectral measure with the Poisson kernel, i.e.,

$$\frac{1}{\pi}\mathrm{Im}((\mathcal{R}_{\mathcal{L}}(x + i\epsilon)f, f)) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{\epsilon}{\epsilon^2 + (x - y)^2} d\mu_f(y). \tag{4.5}$$

The Poisson kernel is one of the most common kernels used to smooth approximations of measures in particle and condensed matter physics (see the discussion in section 4.1.1). When $\mathcal{L}$ has no singular continuous spectrum, substituting the spectral measure given in (1.8) into the expression (4.5) shows that $\mathcal{R}_{\mathcal{L}}(x + i\epsilon)$ provides an approximation to both the discrete and continuous components of the measure μ_f for $\epsilon > 0$. That is,

$$\frac{1}{\pi}\mathrm{Im}((\mathcal{R}_{\mathcal{L}}(x + i\epsilon)f, f)) = \frac{1}{\pi}\sum_{\lambda \in \Lambda^{\mathrm{p}}(\mathcal{L})} \frac{\epsilon\,(\mathcal{P}_\lambda f, f)}{\epsilon^2 + (x - \lambda)^2} + \frac{1}{\pi}\int_{\mathbb{R}} \frac{\epsilon\,\rho_f(y)}{\epsilon^2 + (x - y)^2}\, dy. \tag{4.6}$$

The contribution from the sum in (4.6) is a series of Poisson kernels centered at the eigenvalues and scaled by the corresponding coefficients $(\mathcal{P}_\lambda f, f)$ for $\lambda \in \Lambda^{\mathrm{p}}(\mathcal{L})$. As $\epsilon \downarrow 0$, the sum converges to a series of Dirac delta distributions representing the discrete part of the measure in (1.8). Meanwhile, the integral in (4.6) contributes a smoothed approximation to the Radon–Nikodym derivative ρ_f.

Motivated by (4.6), we select $\epsilon > 0$ and approximate samples of μ_f by evaluating

$$\mu_f^\epsilon(x) := \tfrac{1}{\pi}\mathrm{Im}((\mathcal{R}_{\mathcal{L}}(x + i\epsilon)f, f)). \tag{4.7}$$

From (4.3), we know that as $\epsilon \downarrow 0$ we have $\mu_f^\epsilon \to \mu_f$ in the sense of weak convergence of measures. Moreover, if μ_f has some additional local regularity about a point $x_0 \in \mathbb{R}$, then $\mu_f^\epsilon(x_0) \to \rho_f(x_0)$ as $\epsilon \downarrow 0$ (see Theorem 4.2.1). There is a two-step procedure for evaluating $\mu_f^\epsilon(x_0)$ at some $x_0 \in \mathbb{R}$, which is immediate from (4.7):

1. Solve the shifted linear equation for u^ϵ:

$$(\mathcal{L} - x_0 - i\epsilon)u^\epsilon = f, \qquad u^\epsilon \in \mathcal{D}(\mathcal{L}). \tag{4.8}$$

2. Compute the inner product $\mu_f^\epsilon(x_0) = \frac{1}{\pi}\mathrm{Im}((u^\epsilon, f))$.

In practice, the smaller $\epsilon > 0$, the more computationally expensive it is to evaluate (4.7) because if $x_0 \in \Lambda(\mathcal{L})$ then the resolvent operator $\mathcal{R}_{\mathcal{L}}(x_0 + i\epsilon)$ is unbounded in the limit $\epsilon \downarrow 0$. One often computes $\mu_f^\epsilon(x_0)$ for successively smaller ϵ to obtain a sequence that converges to $\mu_f(x_0)$. For example, Richardson's extrapolation can improve the convergence rate in ϵ [30], which can be proven using the machinery of section 4.3.

Typically, one wants to sample μ_f^ϵ at several points $x_1, \ldots, x_m \in \mathbb{R}$, and then construct a local or global representation of μ_f^ϵ for visualization or further computations. If one wants to visualize μ_f^ϵ in an interval, then we recommend evaluating at equispaced points in that interval. However, when one wants to calculate an integral with respect to μ_f^ϵ, it is better to evaluate μ_f^ϵ at quadrature nodes (see section 4.3.2). Note that if $x_j \notin \Lambda(\mathcal{L})$, then $\mu_f^\epsilon(x_j) \to 0$ as $\epsilon \downarrow 0$ (for example, see Figure 4.1).

Although singular continuous spectrum may appear to be an exotic phenomenon, it occurs in applications of practical interest. For example, discrete Schrödinger operators with aperiodic potentials on $\ell^2(\mathbb{Z})$ (such as the Fibonacci Hamiltonian) can have spectra that are Cantor sets with purely singular continuous spectral measures (see [10, 37, 38, 68, 118, 143]). When $\Lambda(\mathcal{L})$ has a non-zero singular continuous component, $\mu_f^\epsilon \to \mu_f$ weakly as $\epsilon \downarrow 0$ and our algorithms can compute $\mu_f(U)$ (for open sets U) and the functional calculus of $\mathcal{L}$.[4]

4.2.1 Evaluating the spectral measure of an integral operator

To illustrate our evaluation strategy, consider the integral operator defined by

$$[\mathcal{L}u](x) = xu(x) + \int_{-1}^{1} e^{-(x^2+y^2)} u(y)\, dy, \qquad x \in [-1,1], \qquad u \in L^2([-1,1]). \tag{4.9}$$

The integral operator $\mathcal{L}$ in (4.9) has continuous spectrum in $[-1,1]$, due to the $xu(x)$ term, and discrete spectrum in $\mathbb{R} \setminus [-1,1]$ from the integral term (a compact perturbation [87]). Figure 4.1 (left) shows three smoothed approximations of μ_f with $f(x) = \sqrt{3/2}\, x$, for smoothing parameter $\epsilon = 0.1, 0.01$, and 0.001. We see the presence of an eigenvalue near $x \approx 1.37$ from a spike in the smoothed measure that approximates a Dirac delta .

To perform the two-step procedure described above on a computer, one must discretize the operator $\mathcal{L}$, and we do this by discretizing $\mathcal{L}$ with an $N \times N$ matrix corresponding to an adaptive Chebyshev collocation scheme.[5] While the precise discretization details are delayed until section 4.4.2, Figure 4.1 illustrates the

[4] In general, it is also impossible to design a black-box method that separates the singular continuous component of μ_f from the other components. This is made precise in [30], which uses the framework of the Solvability Complexity Index (SCI) hierarchy [28, 29, 31, 32].

[5] While $N \times N$ discretizations converge for Fredholm operators [85], square truncations of

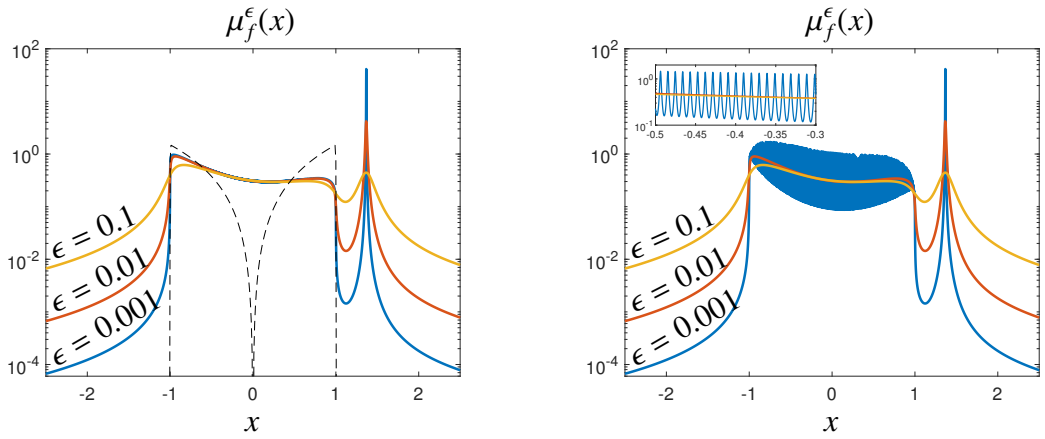

Figure 4.1: Left: The smoothed approximation μ_f^ϵ for the integral operator in (4.9) and different ϵ. The discretization sizes for solving the shifted linear systems are adaptively selected. The dashed line corresponds to the spectral measure of the operator given by $u(x) \to xu(x)$. Adding the compact perturbation (the integral term) alters the shape of the measure over $[-1, 1]$ and there is an additional eigenvalue near $x \approx 1.37$. Right: The same computation except with a fixed discretization size of $N = 300$ to solve (4.8). The magnified region shows spurious high-frequency oscillations for $\epsilon = 0.001$, an artifact caused by the discrete spectrum of the underlying discretization.

critical role that N plays when evaluating μ_f^ϵ. In particular, there are two limits to take in theory: $N \to \infty$ and $\epsilon \downarrow 0$. It is known that these two limits must be taken with considerable care [30]. If N is kept fixed as one takes $\epsilon \downarrow 0$, then the computed samples of μ_f^ϵ do not converge (see Figure 4.1 (right)) because the computed samples get polluted by the discrete spectrum of the discretization. Instead, as one takes $\epsilon \downarrow 0$, one must appropriately increase N too. In practice, we increase N by selecting it adaptively to ensure that we adequately resolve solutions to (4.8) (see Figure 4.1 (left)). The precise details on how we adequately resolve solutions are given in section 4.4.2.

spectral discretizations of operators may not always converge. Instead, one may need to take rectangular truncations to ensure that discretizations of $\mathcal{R}_{\mathcal{L}}(z)f$ converge [30].

4.2.2 Pointwise convergence of smoothed measures

It is known that if μ_f is locally absolutely continuous with continuous Radon–Nikodym derivative ρ_f (see (1.8)), then μ_f^ϵ converges pointwise to ρ_f [80, p. 22]. However, under additional smoothness assumptions on μ_f, it is useful to understand how rapidly μ_f^ϵ converges to μ_f. The connection between μ_f^ϵ and the Poisson kernel in (4.5) allows us to do this on intervals for which μ_f possesses some local regularity so that ρ_f is Hölder continuous. We let $C^{k,\alpha}(I)$ denote the Hölder space of functions that are k times continuously differentiable on an interval I with an α-Hölder continuous kth derivative [54]. For $h_1 \in C^{0,\alpha}(I)$ and $h_2 \in C^{k,\alpha}(I)$ we define the seminorm and norm, respectively, as

$$|h_1|_{C^{0,\alpha}(I)} = \sup_{x \neq y \in I} \frac{|h_1(x) - h_1(y)|}{|x-y|^\alpha}, \quad \|h_2\|_{C^{k,\alpha}(I)} = |h_2^{(k)}|_{C^{0,\alpha}(I)} + \max_{0 \leq j \leq k} \|h_2^{(j)}\|_{\infty, I}.$$

Theorem 4.2.1. *Suppose that the measure μ_f in (1.8) is absolutely continuous on the interval $I = (x_0 - \eta, x_0 + \eta)$ for some $x_0 \in \mathbb{R}$ and $\eta > 0$, let μ_f^ϵ be defined as in* (4.7)*, and let $0 \leq \alpha < 1$. If $\rho_f \in C^{0,\alpha}(I)$, then*

$$|\rho_f(x_0) - \mu_f^\epsilon(x_0)| = O(\epsilon^\alpha), \qquad \text{as} \quad \epsilon \downarrow 0.$$

Proof. First, decompose ρ_f into two non-negative parts so that $\rho_f = \rho_1 + \rho_2$, where the support of ρ_1 is in I and ρ_2 vanishes on $(x_0 - \eta/2, x_0 + \eta/2)$. Since $\rho_f(x_0) = \rho_1(x_0)$ and the Poisson kernel integrates to 1, we can use the convolution representation for μ_f^ϵ (see (4.5) and (4.7)) and the commutativity of convolution to bound the approximation error as

$$\begin{aligned} \pi|\rho_f(x_0) - \mu_f^\epsilon(x_0)| &= \left| \int_{\mathbb{R}} \frac{\epsilon}{\epsilon^2 + y^2} \rho_1(x_0) dy - \int_{\mathbb{R}} \frac{\epsilon d\mu_f(y)}{\epsilon^2 + (x_0 - y)^2} \right| \\ &\leq \left| \int_{\mathbb{R}} \frac{\epsilon}{\epsilon^2 + y^2} \left(\rho_1(x_0) - \rho_1(x_0 - y)\right) dy \right| + \int_{\mathbb{R}} \frac{\epsilon d\mu_f^{(\mathrm{r})}(y)}{\epsilon^2 + (x_0 - y)^2}. \end{aligned} \tag{4.10}$$

Here, $d\mu_f^{(\mathrm{r})}(y) := d\mu_f(y) - \rho_1(y)dy$ is a non-negative measure with support in $\mathbb{R} \setminus (x_0 - \eta/2, x_0 + \eta/2)$. Since μ_f is a probability measure, we have that $\int_{\mathbb{R}} d\mu_f^{(\mathrm{r})}(y) \leq 1$, and the second term in (4.10) is bounded via

$$\int_{\mathbb{R}} \frac{\epsilon d\mu_f^{(\mathrm{r})}(y)}{\epsilon^2 + (x_0 - y)^2} = \int_{|x_0 - y| \geq \eta/2} \frac{\epsilon d\mu_f^{(\mathrm{r})}(y)}{\epsilon^2 + (x_0 - y)^2} \leq \frac{\epsilon}{\epsilon^2 + \frac{\eta^2}{4}}. \tag{4.11}$$

Since $\rho_f \in C^{0,\alpha}(I)$, standard arguments using cutoff functions [54] show that we can choose ρ_1 so that $|\rho_1|_{C^{0,\alpha}(I)} \leq |\rho_f|_{C^{0,\alpha}(I)} + C\eta^{-\alpha}\|\rho_f\|_{\infty,I}$ for some universal constant C. Consequently, we have that

$$|\rho_1(x_0) - \rho_1(x_0 - y)| \leq |\rho_1|_{C^{0,\alpha}(I)}|y|^\alpha \leq (|\rho_f|_{C^{0,\alpha}(I)} + C\eta^{-\alpha}\|\rho_f\|_{\infty,I})|y|^\alpha.$$

Substituting this bound into the first term on the right hand side of (4.10) and combining with (4.11), yields

$$|\rho_f(x_0) - \mu_f^\epsilon(x_0)| \leq \frac{|\rho_f|_{C^{0,\alpha}(I)} + C\eta^{-\alpha}\|\rho_f\|_{\infty,I}}{\pi} \int_{\mathbb{R}} \frac{\epsilon}{\epsilon^2 + y^2}|y|^\alpha \, dy + \frac{\epsilon}{\pi\left(\epsilon^2 + \frac{\eta^2}{4}\right)}.$$

Calculating the integral explicitly leads to

$$|\rho_f(x_0) - \mu_f^\epsilon(x_0)| \leq \left(|\rho_f|_{C^{0,\alpha}(I)} + C\eta^{-\alpha}\|\rho_f\|_{\infty,I}\right)\sec\left(\frac{\alpha\pi}{2}\right)\epsilon^\alpha + \frac{\epsilon}{\pi\left(\epsilon^2 + \frac{\eta^2}{4}\right)}. \tag{4.12}$$

The right hand side of (4.12) is $O(\epsilon^\alpha)$ as $\epsilon \downarrow 0$, which concludes the proof. □

In Theorem 4.2.1, we see that the convergence rate of $|\rho_f(x_0) - \mu_f^\epsilon(x_0)|$ as $\epsilon \downarrow 0$ depends on the local regularity of μ_f. One can also show (see Theorem 4.3.2) that $|\rho_f(x_0) - \mu_f^\epsilon(x_0)| = O(\epsilon \log(1/\epsilon))$ if $\rho_f \in C^1(I)$ as well as the fact that any additional smoothness assumptions on ρ_f no longer improve the convergence rate.[6] Since our procedure is local, the convergence rate is not affected by far away discrete and singular continuous components of μ_f. However, the convergence degrades

[6]The logarithmic term occurs due to the non-integrability of $x/(\pi(x^2+1))$. One can also show that the error rate of $O(\epsilon \log(1/\epsilon))$ is achieved if $\rho_f \in C^{0,1}(I)$ is Lipschitz continuous.

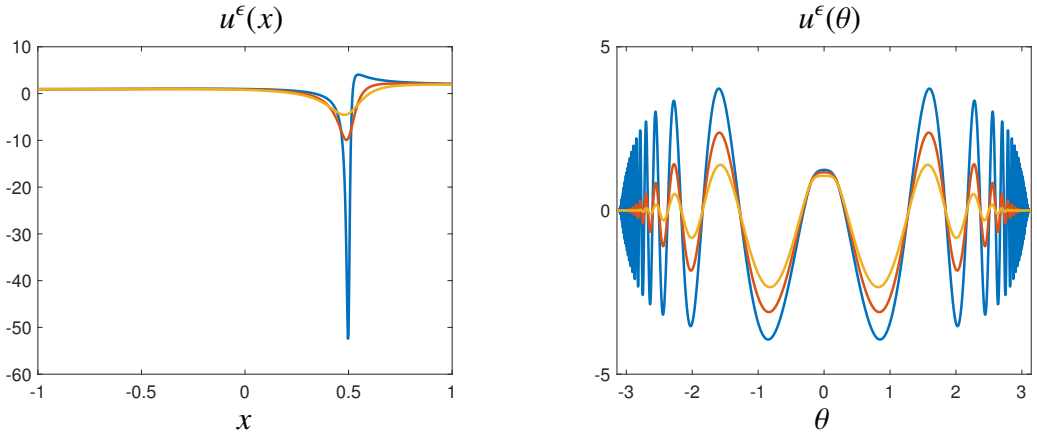

Figure 4.2: Real part of the numerical solutions to the shifted linear equations in (4.8) for the integral operator in (4.9) (left) and the Schrödinger operator in (4.13) (right), with $\epsilon = 0.1$ (yellow), $\epsilon = 0.05$ (orange), and $\epsilon = 0.01$ (blue). The solutions to (4.13) are mapped to $[-\pi, \pi]$ via $x = 10i(1 - e^{i\theta})/(1 + e^{i\theta})$. We discretize using sparse, well-conditioned spectral methods and the discretization sizes are selected adaptively to accurately resolve $u^\epsilon(x)$ and $u^\epsilon(\theta)$.

near singular points in the spectral measure because the constants in (4.12) blow up as $\eta \to 0$. While $|\rho_f(x_0) - \mu_f^\epsilon(x_0)| = \mathcal{O}(\epsilon^\alpha)$ in Theorem 4.2.1 is stated as an asymptotic statement, we can also obtain explicit bounds for adaptive selection of ϵ (see Theorem 4.3.2).

4.2.3 A numerical balancing act

To explore the practical importance of the convergence rates in Theorem 4.2.1, we examine the numerical cost associated with solving the shifted linear systems in (4.8). When the real component of the shift is in the continuous spectrum of $\mathcal{L}$ and ϵ is small, we typically require large discretizations to avoid the situation observed in Figure 4.1 (right). There are many potential reasons why we require large discretization sizes as $\epsilon \downarrow 0$. Here are two illustrative examples:

1) Interior layers. Revisiting the integral operator example in (4.9), we select $x_0 = 1/2$ in the continuous spectrum of $\mathcal{L}$, and $f(x) = \sqrt{3/2}\, x$. In Figure 4.2 (left), we observe that the solution $u^\epsilon(x)$ develops an interior layer and blows up at $x_0 = 1/2$ as $\epsilon \downarrow 0$. The blow-up occurs because the multiplicative term in $\mathcal{L} - (x_0 + i\epsilon)$ has a root at $x_0 = 1/2$ when $\epsilon = 0$, giving rise to a pole in $u^\epsilon(x)$. For $\epsilon > 0$, the pole of $u^\epsilon(x)$ is located at a distance of $O(\epsilon)$ away from the real axis. A large discretization size is needed to resolve $u^\epsilon(x)$ for small ϵ due to the thin interior layer in $u^\epsilon(x)$.

2) Oscillatory behavior. Consider the second-order differential operator given by

$$[\mathcal{L}u](x) = -\frac{d^2u}{dx^2}(x) + \frac{x^2}{1+x^6}u(x), \qquad x \in \mathbb{R}. \tag{4.13}$$

We select $x_0 = 0.3$ in the continuous spectrum of $\mathcal{L}$, and $f(x) = \sqrt{9/\pi} \cdot x^2/(1 + x^6)$. In Figure 4.2 (right), we plot solutions mapped onto the domain $[-\pi, \pi]$ by the change-of-variables $x = 10i(1 - e^{i\theta})/(1 + e^{i\theta})$. The solutions $u^\epsilon(x)$ are highly oscillatory with slow decay as $\theta \to \pm\pi$. As $\epsilon \downarrow 0$ the decay degrades and the persistent oscillations correspond to a transition in the nature of the singular points of (4.8) at $\pm\infty$. This means a large discretization is needed to resolve $u^\epsilon(x)$ for small ϵ.

The dominating computational expense in evaluating μ_f^ϵ is solving the shifted linear systems in (4.8), and the cost of computing $u^\epsilon(x)$ generally increases as $\epsilon \downarrow 0$. There is a balancing act. On the one hand, we wish to stay as far away from the spectrum as possible, so that the evaluation of μ_f^ϵ is computationally efficient. On the other hand, we desire samples of μ_f^ϵ to be good approximations to ρ_f, which requires a small $\epsilon > 0$. Even though we use sparse, well-conditioned spectral methods to discretize (4.8) (see section 4.4), the trade-

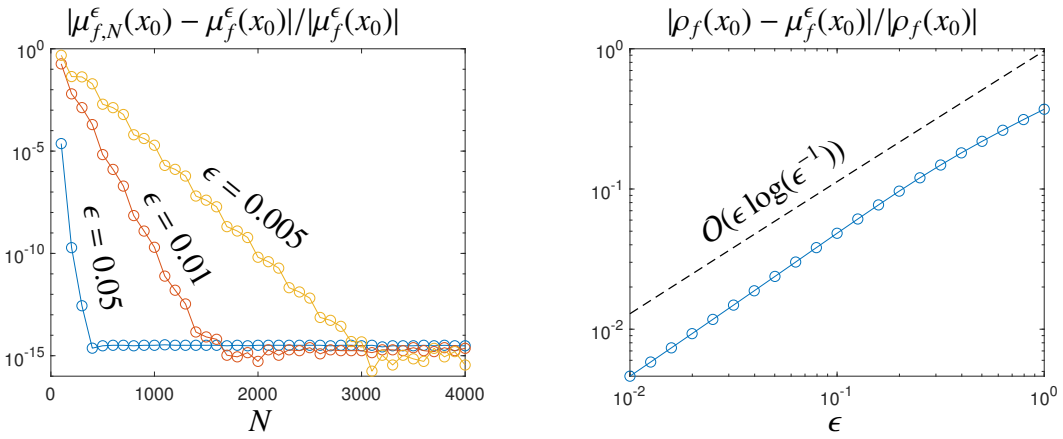

Figure 4.3: Left: The relative error in the numerical approximation $\mu^\epsilon_{f,N}$, corresponding to discretization size N, of the smoothed measure in (4.7) for the integral operator in (4.9) with $\epsilon = 0.05$, $\epsilon = 0.01$, and $\epsilon = 0.005$. Right: The pointwise relative difference between the smoothed measure $\mu^\epsilon_f(x)$ and the density $\rho_f(x)$, evaluated at $x_0 = 1/2$, compared with the $O(\epsilon \log(\epsilon^{-1}))$ error bound in Theorem 4.3.2 for the integral operator in (4.9). The relative error is computed by comparing with a numerical solution that has been adaptively resolved to machine precision.

off between computational cost and accuracy means that the slow convergence rate determined in Theorem 4.2.1 is a severe limitation. In Figure 4.3, we explore the discretization sizes that are needed to evaluate spectral measures with the Poisson kernel accurately. For the integral operator in (4.9) and $\epsilon = 0.05$, 0.01, and 0.005, we observe that we need $N = 400$, 1700, and 3100, respectively (see Figure 4.3 (left)). Unfortunately, to obtain samples of the spectral measure with two digits of relative accuracy, we require that $\epsilon \approx 0.01$ (see Figure 4.3). For this example, we observe that we require $N \approx 20/\epsilon$ for small $\epsilon > 0$, so it is computationally infeasible to obtain more than five or six digits of accuracy with the Poisson kernel.

In addition to the computational cost of increasing N, the discretizations used to solve the linear systems in (4.8) become increasingly ill-conditioned when $x_0 \in \Lambda(\mathcal{L})$ and $\epsilon \downarrow 0$ (a reflection of $\|\mathcal{R}_\mathcal{L}(x_0 + i\epsilon)\| = \epsilon^{-1}$). This can limit the attainable accuracy. Moreover, the performance of iterative methods, if used to accelerate the solution of the large shifted linear systems, may also suffer. In our

experience, the cost of increasing N is usually the limiting factor and we rarely take $\epsilon < 10^{-2}$.

4.3 High-order kernels

Theorem 4.2.1 demonstrates that $\mu_f^\epsilon \to \rho_f$ pointwise in intervals for which μ_f is absolutely continuous with Hölder continuous density ρ_f, where the rate of convergence depends on the Hölder exponent of ρ_f. However, even when ρ_f possesses additional regularity, the best rate of convergence for smoothed measures using the Poisson kernel is $O(\epsilon \log(1/\epsilon))$. A natural question is:

"Can we use other kernels to exploit additional regularity in μ_f?"

In this section, we construct kernels that can be used to compute smoothed measures that approximate ρ_f to high-order in ϵ when ρ_f is smooth. This allows us to obtain accurate samples of μ_f while avoiding extremely small ϵ and the associated computational cost of solving the shifted linear equations in (4.8) when the shifts are close to the real line. We use $K(x)$ to denote a kernel for which $K_\epsilon(x) = \epsilon^{-1}K(x/\epsilon)$ is an approximation to the identity, i.e., $K_\epsilon \to \delta$ as $\epsilon \downarrow 0$ in the sense of distributions [136, Ch. 3], where δ is the Dirac delta distribution.

To gain intuition about the conditions that $K(x)$ must satisfy so that $K_\epsilon * \mu_f$ approximates μ_f to high-order, consider an absolutely continuous probability measure μ with density ρ supported on an interval $I = (x_0-\eta, x_0+\eta)$, for some $x_0 \in \mathbb{R}$ and $\eta > 0$. The following argument is common in statistical non-parametric regression [161, 162]. Since we want K_ϵ to be an approximation to the identity, our first property is that $\int_{\mathbb{R}} K(x)dx = 1$. For further properties, we examine the

approximation error

$$[K_\epsilon * \mu](x_0) - \rho(x_0) = \int_{\mathbb{R}} K_\epsilon(y)(\rho(x_0 - y) - \rho(x_0))\, dy.$$

Assuming that $\rho \in C^{n,\alpha}(I)$ for some $0 < \alpha < 1$, we can use an nth order Taylor expansion of $\rho(x_0 - y) - \rho(x_0)$ to rewrite the approximation error as

$$[K_\epsilon * \mu](x_0) - \rho(x_0) = \sum_{k=1}^{n-1} \frac{(-1)^k \rho^{(k)}(x_0)}{k!} \int_{\mathbb{R}} K_\epsilon(y) y^k\, dy + \int_{\mathbb{R}} K_\epsilon(y) R_n(x_0, y)\, dy,$$

where $R_n(x_0, y)$ denotes the $O(|y|^n)$ remainder term in the Taylor series and $\rho^{(k)}$ is the kth derivative of ρ. The change-of-variables $y \to \epsilon y$ reveals that the kth term in the series is of size $O(\epsilon^k)$, provided that $K(y)y^k$ is integrable. Meanwhile, the Hölder continunity of $\rho^{(n)}$ shows that the term involving $R_n(x_0, y)$ is of size $O(\epsilon^{n+\alpha})$ provided that $K(y)y^{n+\alpha}$ is integrable and $\int_{\mathbb{R}} K(y)y^n\, dy = 0$. Therefore, a kernel that achieves an $O(\epsilon^{n+\alpha})$ approximation error has vanishing moments, i.e., $\int_{\mathbb{R}} K(y)y^k\, dy = 0$ for $1 \le k \le n$.

In practice, μ may not be absolutely continuous and its absolutely continuous part may have a density ρ with singular points or unbounded support. As in Theorem 4.2.1, we can deal with the general case by decomposing $\rho = \rho_1 + \rho_2$ into two non-negative parts, where ρ_1 is sufficiently smooth and compactly supported on I, and where ρ_2 vanishes in a neighborhood of x_0. The cost of this decomposition is a second term in the approximation error (analogous to the second term on the right hand side of (4.10))

$$[K_\epsilon * \mu](x_0) - \rho(x_0) = \int_{\mathbb{R}} K_\epsilon(y)(\rho_1(x_0 - y) - \rho_1(x_0))\, dy + \int_{\mathbb{R}} K_\epsilon(x_0 - y)\, d\mu^{(\mathrm{r})}(y),$$

where $d\mu^{(\mathrm{r})}(y) = d\mu(y) - \rho_1(y)dy$. To ensure that this additional term does not dominate as $\epsilon \downarrow 0$, it is necessary that the kernel $K(y)$ decays at an appropriate rate as $|y| \to \infty$. This ensures that $K_\epsilon(x_0 - y)$ is sufficiently small on the support

of $d\mu^{(r)}(y)$ (see (4.11) for the decay in the Poisson kernel). Motivated by this discussion, we make the following definition (similar to [161, Def. 1.3]).

Definition 4.3.1 (mth order kernel)**.** *Let m be a positive integer and $K \in L^1(\mathbb{R})$. We say K is an mth order kernel if it satisfies the following properties:*

(i) Normalized: $\int_{\mathbb{R}} K(x)dx = 1$.

(ii) Zero moments: $K(x)x^j$ *is integrable and* $\int_{\mathbb{R}} K(x)x^j dx = 0$ *for* $0 < j < m$.

(iii) Decay at $\pm\infty$*: There is a constant* C_K*, independent of* x*, such that*

$$|K(x)| \leq \frac{C_K}{(1+|x|)^{m+1}}, \qquad x \in \mathbb{R}. \tag{4.14}$$

It is straightforward to verify that the Poisson kernel is a first-order kernel and the Gaussian kernel, i.e., $h(x) = (2\pi)^{-1/2}e^{-x^2/2}$, is a second-order kernel. While the Gaussian kernel plays an important role in DOS calculations [98] and kernel density estimation [131], it is not as useful in our framework since the evaluation of $h_\epsilon * \mu_f$ is not immediately related to pointwise evaluations of the resolvent (see section 4.3.1).

Since an mth order kernel, K, is an approximation to the identity, one can show that $K_\epsilon * \mu_f$ converges weakly to μ_f. Moreover, in intervals where μ_f is absolutely continuous and sufficiently regular, $K_\epsilon * \mu_f$ converges pointwise to ρ_f and the rate of convergence increases with the smoothness of ρ_f, up to a maximum of $O(\epsilon^m \log(1/\epsilon))$.

Theorem 4.3.2. *Let K be an mth order kernel and suppose that the measure μ_f is absolutely continuous on $I = (x_0 - \eta, x_0 + \eta)$ for $\eta > 0$ and a fixed $x_0 \in \mathbb{R}$. Let ρ_f be the Radon–Nikodym derivative of the absolutely continuous component of μ_f, and suppose that $\rho_f \in C^{n,\alpha}(I)$ with $\alpha \in [0, 1)$. Denote the pointwise error by $E_\epsilon(x) = \left|\rho_f(x) - [K_\epsilon * \mu_f](x)\right|$. Then it holds that*

(i) If $n + \alpha < m$, then, for a constant $C(n, \alpha)$ depending only on n and α,

$$E_\epsilon(x_0) \leq \frac{C_K \epsilon^m}{(\epsilon + \frac{\eta}{2})^{m+1}} + C(n, \alpha) \|\rho_f\|_{C^{n,\alpha}(I)} \int_{\mathbb{R}} |K(y)||y|^{n+\alpha} dy (1 + \eta^{-n-\alpha}) \epsilon^{n+\alpha}. \tag{4.15}$$

(ii) If $n + \alpha \geq m$, then, for a constant $C(m)$ depending only on m,

$$E_\epsilon(x_0) \leq \frac{C_K \epsilon^m}{(\epsilon + \frac{\eta}{2})^{m+1}} + C(m) \|\rho_f\|_{C^m(I)} \left(C_K + \int_{-\frac{\eta}{\epsilon}}^{\frac{\eta}{\epsilon}} |K(y)||y|^m dy \right) (1 + \eta^{-m}) \epsilon^m. \tag{4.16}$$

Here, C_K is from (4.14)*.*

Proof. See Appendix A.1 of [34]. □

Using (4.14) to bound $|K(y)|$ in (4.15) and (4.16), Theorem 4.3.2 shows that, under local regularity conditions near $x_0 \in \mathbb{R}$ and for fixed $\eta > 0$, an mth order kernel has

$$\left| \rho_f(x_0) - [K_\epsilon * \mu_f](x_0) \right| = O(\epsilon^{n+\alpha}) + O(\epsilon^m \log(1/\epsilon)), \quad \text{as} \quad \epsilon \downarrow 0.$$

The logarithmic term appears in the case that $K(x)x^m$ is not integrable. The upper bounds on $E_\epsilon(x_0)$ in Theorem 4.3.2 deteriorate as the interval of regularity shrinks ($\eta \to 0$), which is to be expected.[7]

4.3.1 Rational kernels

Now that we know the necessary properties of a kernel K so that $K_\epsilon * \mu_f$ achieves high-order convergence (see Definition 4.3.1), we can develop a resolvent-based

[7] Similar results to Theorem 4.3.2, without the first term on the right hand side of (4.15) and (4.16), for absolutely continuous probability measures with globally Hölder continuous density functions are used in kernel density estimation in statistics (see, for example, [161, Prop. 1.2]).

approach to approximately evaluate a spectral measure more efficiently. The key to our computational framework (see section 4.2) is the connection between the smoothed measure and the resolvent in (4.5). This relation allows us to compute the convolution of the measure μ_f with the Poisson kernel by evaluating the resolvent operator at the poles of the (rescaled) Poisson kernel. In other words, we can sample the smoothed measure by solving the shifted linear equations in (4.8).

Using the identity in (4.4), we can build generalizations of (4.5) for convolutions with rational functions. Suppose that the kernel K is of the form

$$K(x) = \frac{1}{2\pi i} \sum_{j=1}^{n_1} \frac{\alpha_j}{x - a_j} - \frac{1}{2\pi i} \sum_{j=1}^{n_2} \frac{\beta_j}{x - b_j}, \tag{4.17}$$

where $a_1, \ldots, a_{n_1}$ are distinct points in the upper half-plane and $b_1, \ldots, b_{n_2}$ are distinct points in the lower half-plane. We restrict K to have only simple poles to avoid having to compute powers of the resolvent. Using (4.4), the convolution $K_\epsilon * \mu_f$ is given by

$$[K_\epsilon * \mu_f](x) = \frac{-1}{2\pi i} \left[\sum_{j=1}^{n_1} \alpha_j (\mathcal{R}_{\mathcal{L}}(x - \epsilon a_j) f, f) - \sum_{j=1}^{n_2} \beta_j (\mathcal{R}_{\mathcal{L}}(x - \epsilon b_j) f, f) \right]. \tag{4.18}$$

Our goal is to choose the poles and residues in (4.17) so that K is an mth order kernel. Given an integer $m \geq 1$, we are interested in finding the smallest possible n_1 and n_2 in (4.17) so that (4.18) is as efficient to evaluate as possible.

We want $K(x) = O(|x|^{-(m+1)})$ as $|x| \to \infty$, which forces linear constraints to hold between the $\alpha_1, \ldots, \alpha_{n_2}$ and $\beta_1, \ldots, \beta_{n_2}$ parameters, as follows. Generically, K in (4.17) is a type $(n_1 + n_2 - 1, n_1 + n_2)$ rational function, which means it can be written as the quotient of a degree $n_1 + n_2 - 1$ polynomial and a degree $n_1 + n_2$ polynomial. In this form, the coefficient of highest power of x in the numerator

is a multiple of

$$\sum_{j=1}^{n_1} \alpha_j - \sum_{j=1}^{n_2} \beta_j,$$

which must vanish for K to have sufficient decay. Under this condition, we find that

$$K(x)x = \frac{1}{2\pi i} \sum_{j=1}^{n_1} \frac{\alpha_j a_j}{x - a_j} - \frac{1}{2\pi i} \sum_{j=1}^{n_2} \frac{\beta_j b_j}{x - b_j}.$$

We can apply the same argument as before to see that when $m \geq 2$, we require that

$$\sum_{j=1}^{n_1} \alpha_j a_j - \sum_{j=1}^{n_2} \beta_j b_j = 0.$$

We repeat this process $m - 1$ times (each time multiplying each term in the sum by the appropriate a_j or b_j) to find that $K(x) = O(|x|^{-(m+1)})$ as $|x| \to \infty$ if and only if

$$\sum_{j=1}^{n_1} \alpha_j a_j^k = \sum_{j=1}^{n_2} \beta_j b_j^k, \qquad k = 0, \ldots, m-1. \tag{4.19}$$

Assuming (4.19) is satisfied, the normalization and zero moment conditions (see Definition 4.3.1 (i) and (ii)) provide us with m linear conditions on the moments of K, which can be computed explicitly via contour integration. Employing a semi-circle contour in the upper half-plane, applying Cauchy's residue theorem, and taking the radius of the semi-circle to infinity, we find that the moments are given in terms of the poles and residues of K, i.e.,

$$\int_{\mathbb{R}} K(y) y^k \, dy = \sum_{j=1}^{n_1} \alpha_j a_j^k = \sum_{j=1}^{n_2} \beta_j b_j^k, \qquad k = 0, \ldots, m-1,$$

where the second equality follows from (4.19) or closing the contour in the lower half-plane. Therefore, the rational kernel in (4.17) is an mth order kernel pro-

vided that the following (transposed) Vandermonde systems are satisfied:

$$\begin{pmatrix} 1 & \dots & 1 \\ a_1 & \dots & a_{n_1} \\ \vdots & \ddots & \vdots \\ a_1^{m-1} & \dots & a_{n_1}^{m-1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n_1} \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ b_1 & \dots & b_{n_2} \\ \vdots & \ddots & \vdots \\ b_1^{m-1} & \dots & b_{n_2}^{m-1} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n_2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.20}$$

The systems in (4.20) are guaranteed to have solutions when $n_1, n_2 \geq m$. For computational efficiency, we select $n_1 = n_2 = m$ poles in the upper and lower half-planes. The Poisson kernel fits into this setting with $m = 1$, $a_1 = \overline{b_1} = i$ and $\alpha_1 = \beta_1 = 1$.

It may appear from (4.18) that we need $2m$ resolvent evaluations to evaluate $K_\epsilon * \mu_f$ at a single point x. However, if the poles are selected so that $b_j = \overline{a_j}$ and $\beta_j = \overline{\alpha_j}$, then the conjugate symmetry of the resolvent, i.e., $(\mathcal{R}_{\mathcal{L}}(\bar{z})f, f) = \overline{(\mathcal{R}_{\mathcal{L}}(z)f, f)}$, reduces the number of resolvent evaluations to m. With this choice, we find that

$$[K_\epsilon * \mu_f](x) = \frac{-1}{\pi} \sum_{j=1}^{m} \operatorname{Im}\left(\alpha_j \, (\mathcal{R}_{\mathcal{L}}(x - \epsilon a_j)f, f)\right),$$

which is analogous to (4.7). While the properties of an mth order kernel determine the number of poles and the residues of K (see (4.20)), the locations of the poles in the upper half-plane are left to our discretion.

Equispaced poles

As a natural extension of the Poisson kernel, whose two poles are at $\pm i$, we consider the family of mth order kernels with equispaced poles in the upper and lower half-planes given by

$$a_j = \frac{2j}{m+1} - 1 + i, \qquad b_j = \overline{a_j}, \qquad 1 \leq j \leq m. \tag{4.21}$$

We then determine the residues by solving the Vandermonde system in (4.20). The first six kernels are plotted in Figure 4.4 (left).

Empirically, we found that the choice in (4.21) performed slightly better than other natural choices such as Chebyshev points with an offset $+i$, rotated roots of unity or dyadic poles $a_j = i2^{-j}$. Dyadic poles have the advantage that if ϵ is halved, the resolvent only needs to be computed at one additional point. The ill-conditioning of the Vandermonde system did not play a role for the values of m here. Moreover, equispaced poles are particularly useful when one wishes to sample the smoothed measure $K_\epsilon * \mu_f$ over an interval since samples of the resolvent can be reused for different points in the interval. Finally, if ϵ is found to be insufficiently small, instead of re-evaluating the resolvent at m points, one can add poles closer to the real axis (with a smaller ϵ) and reuse the old resolvent evaluations. This effectively increases m, and hence the coefficients α_j need to be recomputed. This may be computationally beneficial since the cost of solving the Vandermonde system is typically negligible compared to the cost of evaluating the resolvent close to the real axis.

To demonstrate the practical advantage of high-order kernels, we revisit the examples from section 4.2 and compute the smoothed measure $K_\epsilon * \mu_f$ using mth order kernels with equispaced poles. In Figure 4.4 (right) and Figure 4.5 (right), we observe the convergence rates predicted in Theorem 4.3.2 for the integral operator in (4.9) and the differential operator in (4.13), respectively. While the Poisson kernel requires us to solve linear equations with shifts extremely close to the continuous spectrum to achieve a few digits of accuracy in our approximation to ρ_f, a sixth-order kernel enables us to achieve about 11 and 9 digits of accuracy, respectively, without decreasing ϵ below 0.01. Figure 4.5 (left) shows

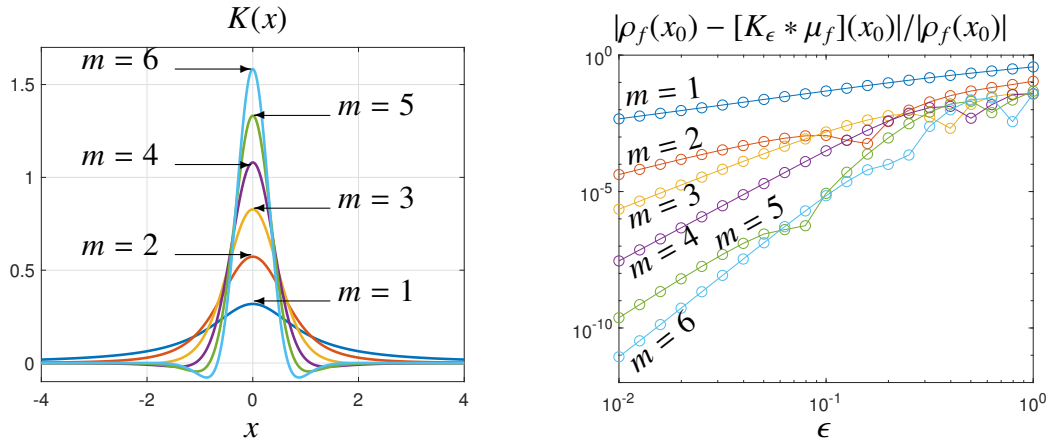

Figure 4.4: Left: The mth order kernels constructed from (4.20) with poles in (4.21) for $1 \leq m \leq 6$. Right: The pointwise relative error in smoothed measures of the integral operator in (4.9) computed using the high-order kernels with poles in (4.21) for $1 \leq m \leq 6$. The relative error is computed by comparing with a numerical solution that is resolved to machine precision.

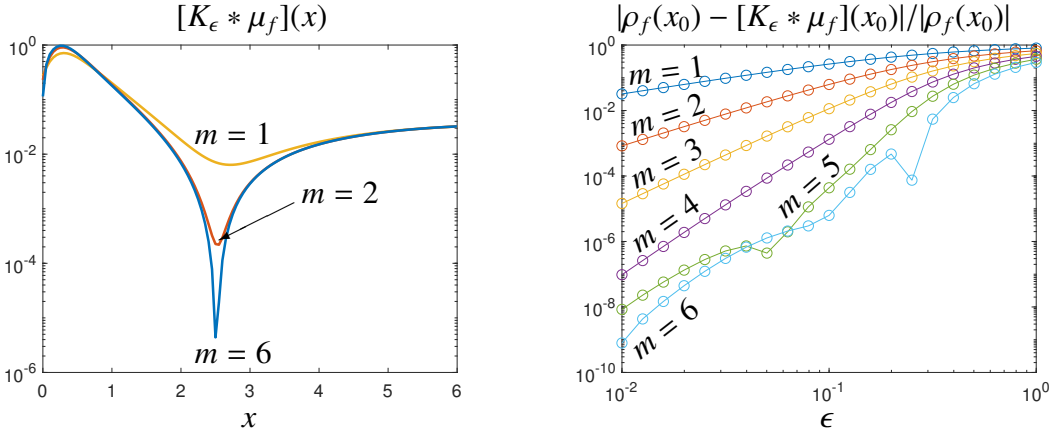

Figure 4.5: Results for the Schrödinger operator in (4.13) using mth order kernels with equispaced poles (see (4.21)). Left: Smoothed approximations to the spectral measure. Right: Pointwise relative error, computed by comparing with a numerical solution resolved to machine precision.

the increased resolution obtained when using high-order kernels for the differential operator in (4.13) with smoothing parameter $\epsilon = 0.1$. Although using a sixth-order kernel requires six times as many resolvent evaluations as that of the Poisson kernel, this is typically favorable because the cost of evaluating the resolvent near the continuous spectrum of $\mathcal{L}$ increases as $\epsilon \downarrow 0$ (see section 4.2.3).

4.3.2 Other types of convergence

Consider the radial Schrödinger operator with a Hellmann potential and angular momentum quantum number ℓ, given by [76]

$$[\mathcal{L}u](r) = -\frac{d^2u}{dr^2}(r) + \left(\frac{\ell(\ell+1)}{r^2} + \frac{1}{r}(e^{-r}-1)\right)u(r), \qquad r > 0. \tag{4.22}$$

The spectral properties of $\mathcal{L}$ are of interest in quantum chemistry, where the Hellman potential models atomic and molecular ionization processes [73]. Ionization rates and related transition probabilities are usually studied by computing bound and resonant states of $\mathcal{L}$; however, we compute this information directly from the spectral measure.

For example, if $f(r) = Ce^{-(r-r_0)^2}$ (where C is chosen so that $\|f\|_{L^2(\mathbb{R}_+)} = 1$) is the radial component of the wave function of an electron interacting with an atomic core via the Hellmann potential in (4.22), then we can calculate the probability that the electron escapes from the atomic core with energy $E \in [a,b]$ (with $0 < a < b$) via

$$\mathbb{P}(a \le E \le b) = \mu_f([a,b]) \approx \int_a^b [K_\epsilon * \mu_f](y)\,dy, \qquad \epsilon \ll 1. \tag{4.23}$$

The error for the approximation in (4.23) is bounded above by

$$\left|\mu_f([a,b]) - \int_a^b [K_\epsilon * \mu_f](y)\,dy\right| \le \int_a^b |\rho_f(y) - [K_\epsilon * \mu_f](y)|\,dy = \|\rho_f - K_\epsilon * \mu_f\|_{L^1([a,b])}.$$

This leads us naturally to the notion of L^p convergence on an interval. The smoothed measure always converges to ρ_f in $L^1([a,b])$ when μ_f is absolutely continuous on $[a,b]$. However, in analogy with the pointwise results in section 4.2.2 and section 4.3, we need to impose some additional regularity on ρ_f to obtain rates of convergence. We let $\mathcal{W}^{k,p}(I)$ denote the Sobolev space of functions in $L^p(I)$ such that f and its weak derivatives up to order k have a finite L^p norm [54].

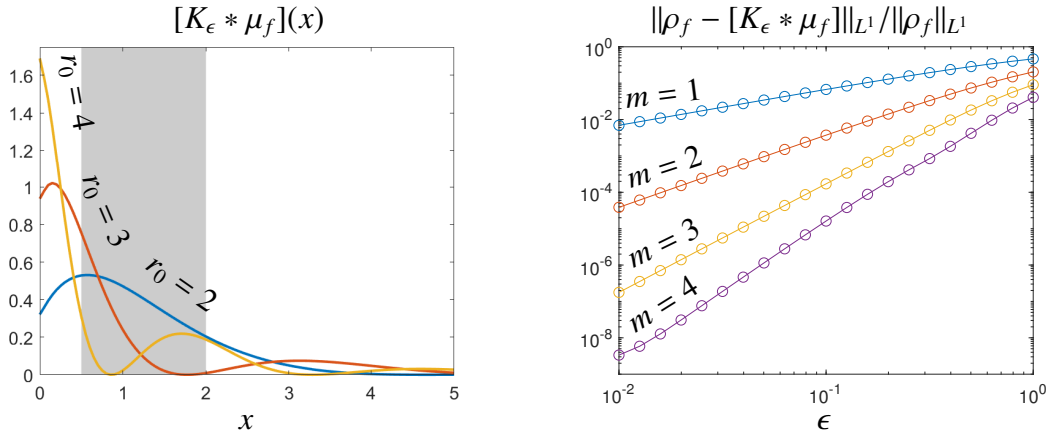

Figure 4.6: Left: The smoothed approximation to the density on the absolutely continuous spectrum of $\mathcal{L}$ in (4.22), with $f_{r_0}(r) = C_{r_0} e^{-(r-r_0)^2}$ and $\ell = 1$, for $r_0 = 2$, $r_0 = 3$, and $r_0 = 4$ (C_{r_0} is a normalization constant so that $\|f_{r_0}\|_{L^2(\mathbb{R}_+)} = 1$). The shaded area under each curve corresponds to $\mathbb{P}(1/2 \le E \le 2)$ in (4.23) for the particle with wave function $f_{r_0}(r)$. Right: The $L^1((1/2, 2))$ relative error in smoothed measures for the radial Schrödinger operator in (4.22). The relative error is computed by comparing with a numerical solution that is resolved to machine precision.

Theorem 4.3.3. *Let K be an mth order kernel and $1 \le p < \infty$. Suppose that the measure μ_f is absolutely continuous on the interval $I = (a - \eta, b + \eta)$ for $\eta > 0$ and some $a < b$. Let ρ_f denote the Radon–Nikodym derivative of the absolutely continuous component of μ_f, and suppose that $\rho_I := \rho_f|_I \in \mathcal{W}^{m,p}(I)$. Then,*

$$\|\rho_I - [K_\epsilon * \mu_f]\|_{L^p((a,b))} \le \frac{C_K (b-a)^{1/p}}{\left(\epsilon + \frac{\eta}{2}\right)^{m+1}} \epsilon^m + C(m) C_K \|\rho_I\|_{\mathcal{W}^{k,p}(I)} (1 + \eta^{-m}) \log\left(1 + \frac{b-a+2\eta}{\epsilon}\right) \epsilon^m,$$

where $C(m)$ is a constant depending only on m, and C_K is from (4.14).

Proof. See Appendix A.2 of [34]. □

Theorem 4.3.3 implies the asymptotic error rate[8]

$$\left\|\rho_I - [K_\epsilon * \mu_f]\right\|_{L^p(I)} = O(\epsilon^m \log(1/\epsilon)), \qquad \text{as} \quad \epsilon \downarrow 0.$$

[8]Theorem 4.3.3 for $p = 2$ without the first term on the right hand side and for absolutely continuous probability measures with $\mathcal{W}^{m,2}(\mathbb{R})$ density function is used in kernel density estimation in statistics [161, Prop. 1.5]. In this context, the L^2 error is used to bound the bias term

The L^1 convergence for the approximation to the probabilities in (4.23) is shown in Figure 4.6 (right), which agrees with the asymptotic rates implied by Theorem 4.3.3.

If one wishes to compute the dynamics of the electron interacting with the atomic core via the Hellman potential, then we need a slightly weaker form of convergence. For instance, the time autocorrelation of the electron's wave function can be computed by integrating the function $F_t(E) = e^{-iEt}$ against the measure μ_f, so that

$$\mu_f(F_t) = (e^{-i\mathcal{L}t}f, f) = \int_{-\infty}^{\infty} e^{-iyt}\, d\mu_f(y) \approx \int_{-\infty}^{\infty} e^{-iyt}\, [K_\epsilon * \mu_f](y)\, dy, \quad \epsilon \ll 1.$$

Unlike the previous cases of pointwise and L^1 convergence, we do not need any additional requirements on the measure μ_f, which may be singular and have discrete components, to obtain convergence rates. Instead, we require that the function F be sufficiently smooth. For example, if $F \in C^{n,\alpha}(\mathbb{R})$ and K is an mth order kernel, then approximating F via convolutions and applying Fubini's theorem shows that

$$|\mu_f(F) - [K_\epsilon * \mu_f](F)| = O(\epsilon^{n+\alpha}) + O(\epsilon^m \log(1/\epsilon)), \qquad \text{as} \quad \epsilon \downarrow 0.$$

Finally, note that a kernel cannot be non-negative everywhere and have an order greater than two. This is not a problem in practice since we can replace $[K_\epsilon * \mu_f](x)$ by $\max\{0, [K_\epsilon * \mu_f](x)\}$ with the same error bounds in Theorems 4.3.2 and 4.3.3.

in the mean integrated squared error. The case of L^1 convergence requires a different proof technique.

4.4 The resolvent framework in practice

Given an mth order rational kernel, defined by distinct poles $a_1, \ldots, a_m$ in the upper half-plane, the resolvent-based framework for evaluating an approximate spectral measure is summarized in Algorithm 5. This algorithm, which can be performed in parallel for several x_0, forms the foundation of `SpecSolve`. `SpecSolve` uses equispaced poles (see section 4.3.1) by default, but users may select other options with the name-value pair `'PoleType'`.

In practice, the resolvent in Algorithm 5 is discretized before being applied. We compute an accurate value of μ_f^ϵ provided that the resolvent is applied with sufficient accuracy (see Figure 4.1), which can be done *adaptively* with *a posteriori* error bounds [30]. For an efficient adaptive implementation, `SpecSolve` constructs a fixed discretization, solves linear systems at each required complex shift, and checks the approximation error at each shift. If further accuracy is needed at a subset of the shifts, then the discretization is refined geometrically, applied at these shifts, and the error is recomputed. This process is repeated until the resolvent is computed accurately at all shifts. The user may (optionally) specify initial and maximum discretization sizes with the name-value pairs `'DiscMin'` and `'DiscMax'`.

`SpecSolve` supports three types of operators: (1) ordinary differential operators, (2) integral operators, and (3) infinite matrices with finitely many nonzeros per column. For more general operators and inner products, the user must supply a command that solves the shifted linear equations in Algorithm 5 and a command that evaluates the inner products, allowing a user to evaluate spectral measures for exotic problems and employ their favorite discretization.

Algorithm 5 A practical framework for evaluating an approximate spectral measure of an operator $\mathcal{L}$ at $x_0 \in \mathbb{R}$ with respect to a vector $f \in \mathcal{H}$.

Input: $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$, $f \in \mathcal{H}$, $x_0 \in \mathbb{R}$, $a_1, \ldots, a_m \in \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$, and $\epsilon > 0$.

1: Solve the Vandermonde system (4.20) for the residues $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$.
2: Solve $(\mathcal{L} - (x_0 - \epsilon a_j))u_j^\epsilon = f$ for $1 \le j \le m$.
3: Compute $\mu_f^\epsilon(x_0) = \frac{-1}{\pi}\mathrm{Im}\left(\sum_{j=1}^m \alpha_j (u_j^\epsilon, f)\right)$.

Output: $\mu_f^\epsilon(x_0)$.

4.4.1 Ordinary differential operators

In `SpecSolve`, the function `diffMeas` computes samples from a smoothed approximation to the spectral measure of a self-adjoint, regular ordinary differential operator on the real-line or on the half-line, i.e.,

$$[\mathcal{L}u](x) = c_p(x)\frac{d^p u}{dx^p}(x) + \cdots + c_1(x)\frac{du}{dx}(x) + c_0(x)u(x), \qquad p \ge 0, \tag{4.24}$$

with the standard inner products. Here, the variable coefficients $c_0, \ldots, c_p$ are smooth functions and $c_p \neq 0$ on the relevant domain (real-line or half-line). Note that $\mathcal{L}$ in (4.24) is not necessarily self-adjoint: the user provides the variable coefficients $c_0, \ldots, c_p$ to `diffMeas` and must verify that $\mathcal{L}$ is self-adjoint.

To demonstrate, recall the Schrödinger operator defined on the real line in (4.13). We can compute a smoothed approximation to its spectral measure as follows.

```
xi = linspace(0,6,121);                          % Evaluation pts
f = @(x) x.^2./(1+x.^6) * sqrt(9/pi);            % Measure wrt f(x)
c = {@(x) x.^2./(1+x.^6), @(x) 0, @(x) -1};      % Schrodinger op
mu = diffMeas(c, f, xi, 0.1, 'order', 1);        % epsilon=0.1, m=1
```

The differential operator is specified by its coefficients $c_0, \ldots, c_2$, which are input as a cell array of function handles. Given evaluation points `xi` and function

handle `f`, `diffMeas` computes the smoothed measure, with respect to `f`, using the specified smoothing parameter and kernel order (the default kernel is $m = 2$). To work on the half-line, the user simply adds `'dom','half'` to the argument list for `diffMeas`.

To apply the resolvent of a differential operator acting on functions on the real line, the associated differential equation (see Algorithm 5) is automatically transplanted to the periodic interval $[-\pi, \pi]$ with an analytic map and solved with an adaptive Fourier spectral method [22]. Typically, the differential equation has singular points at $\pm\pi$ after mapping, and the Fourier spectral method usually converges to a bounded analytic solution [22, Ch. 17.8]. Similarly, on the half-line, the differential equation is mapped to the unit interval $[-1, 1]$ with an analytic map and solved with an adaptive nonperiodic analogue of the Fourier spectral method, which is known as the ultraspherical spectral method [103]. After solving the differential equation on the mapped domain, the inner products in (4.7) are computed using a trapezoidal rule (for the unit circle) [159] or a Clenshaw–Curtis rule (for the unit interval) [155, Ch. 19].

In many applications, differential operators on the half-line may have a singular point at the origin. This makes an efficient and automatic representation of variable coefficients somewhat subtle. For example, the radial Schrödinger operator in (4.22) has a singular point at the origin for $\ell \geq 1$, and the shifted linear equations in Algorithm 5 should be multiplied through by r^2 so that subsequent discretizations yield sparse, banded matrices [103]. In addition to `diffMeas`, `SpecSolve` contains a small gallery of functions that sample smoothed spectral measures for common operators with singular points, such as `rseMeas`, which samples the smoothed measure of the radial Schrödinger operator with

a user-specified potential.

To illustrate, we use `rseMeas()` to compute $\mathbb{P}(1/2 \le E \le 2)$ from (4.23).

```
normf = sqrt(pi/8)*(2-igamma(1/2,8)/gamma(1/2)); % Normalization
f = @(r) exp(-(r-2).^2)/sqrt(normf);          % Measure wrt f(r)
V={@(r) 0, @(r) exp(-r)-1, 1};                % Potential, l=1
[xi, wi] = chebpts(20,  [1/2 2]);             % Quadrature rule
mu = rseMeas(V, f, xi, 0.1, 'Order', 4)       % epsilon=0.1, m=4
ion_prob = wi * mu;                           % Ionization prob
```

The user specifies the potential of the radial Schrödinger operator through a cell array of function handles: `V{1}` is the nonsingular part of the potential, `V{2}` is the variable coefficient for the r^{-1} Coulomb term, and `V{3}` is the quantum angular momentum number that defines the coefficient for the r^{-2} centrifugal term.

4.4.2 Integral operators

In `SpecSolve`, the function `intMeas` computes samples from a smoothed approximation of the spectral measure of an integral operator, acting on functions defined on $[-1, 1]$, of the form

$$[\mathcal{L}u](x) = a(x)u(x) + \int_{-1}^{1} g(x, y)\, u(y)\, dy, \qquad x \in [-1, 1], \qquad u \in L^2([-1, 1]).$$

We assume that the multiplicative coefficient $a(x)$ and the kernel $g(x, y)$ are smooth functions (well-approximated by polynomials), and that $g(x, y) = \overline{g(y, x)}$ so that $\mathcal{L}$ is self-adjoint with respect to the standard inner product. Revisiting

the integral operator from (4.9), we can compute the smoothed measure with a few simple commands.

```
xi = linspace(-2.5,2.5,501);                    % Evaluation pts
f = @(x) sqrt( 3/2 ) * x;                       % Measure wrt f(x)
a = { @(x) x, @(x,y) exp(-(x.^2+y.^2)) };       % Integral operator
mu = intMeas(a, f, xi, 0.1, 'Order', 1);        % epsilon=0.1, m=1
```

The integral operator is specified by a cell array containing function handles for the kernel and multiplicative coefficient. Given a smoothing parameter and kernel order, the smoothed measure is approximated at the evaluation points `xi`.

To apply the resolvent, we use an adaptive Chebyshev collocation scheme to solve the shifted linear systems in Algorithm 5. For efficient storage and computation, we exploit low numerical rank structure in the discretization of the smooth kernels when possible [152]. We apply a Clenshaw–Curtis quadrature rule to compute the inner products required to sample μ_f^ϵ [155].

CHAPTER 5

COMPUTING GENERALIZED EIGENFUNCTIONS

In the study of time-harmonic wave propagation, self-adjoint operators with absolutely continuous spectra play a central role. In this setting, the absolutely continuous spectrum is typically associated with a continuum of scattering states. These states encode and quantify important physical behavior, such as the transmission and reflection coefficients associated with an interface or the effective cross-section of a scattering process, e.g., see [151, Ch. 10]. In the language of spectral theory, scattering states may be understood as generalized eigenfunctions of the underlying self-adjoint operator [42] (see section 1.4.2).

In this chapter, we propose a practical framework for computing generalized eigenfunctions of a self-adjoint operator $\mathcal{L}$ on a rigged Hilbert space $\Phi \subset \mathcal{H} \subset \Phi^*$. Building on the methodology of the previous chapters, we construct approximations to the generalized eigenfunctions of $\mathcal{L}$ by solving linear operator equations with complex shifts. These approximations are naturally interpreted as wave-packets with a narrow band of spectral content centered on the target mode; we demonstrate their convergence in two relevant topologies as the spectral bandwidth vanishes. The basic facets of the method are illustrated through two worked examples. We conclude with a study of scattering modes and related phenomena in a 2-dimensional quantum waveguide.

5.1 Wave-packet approximations

By the nuclear spectral theorem, $\mathcal{L}$ has a complete system of generalized eigenfunctions in Φ^* that satisfy (1.12) and (1.13) in Chapter 1. Denote these by $\psi_{\lambda,k} \in \Phi^*$ for $\lambda \in \Lambda(\mathcal{L})$ and $k \leq M_\lambda$, where M_λ is the multiplicity of the generalized eigenvalue λ. To approximate these generalized eigenfunctions, we return to Stone's relation between the spectral measure and the resolvent (see section 4.2). For a convenient notation in the following sections, we always assume that ϕ and $\mathcal{L}\phi$ take real values, so that $(\mathcal{L} - z)^{-1}\phi - (\mathcal{L} - \bar{z})^{-1}\phi = 2\mathrm{Im}(\mathcal{L} - z)^{-1}\phi$.

By expanding the resolvent with generalized eigenfunctions rather than a spectral measure, i.e., employing the nuclear spectral theorem in place of the classical spectral theorem, we have (for any $z \notin \Lambda(\mathcal{L})$) that

$$(\mathcal{L} - z)^{-1}\phi = \sum_{k=1}^{M} \int_{\Lambda(\mathcal{L})} \frac{(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}}{\lambda - z} \psi_{\lambda,k}\, d\mu_k(\lambda), \qquad \text{for all} \qquad \phi \in \Phi.$$

Now, pick $\lambda_* \in \lambda(\mathcal{L})$, $\epsilon > 0$, and set $z = \lambda_* + i\epsilon$. We calculate that

$$\frac{1}{\pi}\mathrm{Im}(\mathcal{L} - \lambda_* - i\epsilon)^{-1}\phi = \frac{1}{\pi} \sum_{k=1}^{M} \int_{\lambda(\mathcal{L})} \frac{\epsilon(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}}{(\lambda - \lambda_*)^2 + \epsilon^2} \psi_{\lambda,k}\, d\mu_k(\lambda). \tag{5.1}$$

The integrals on the right hand side are convolutions of $(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}\psi_{\lambda,k}$ with the Poisson kernel, taken with respect to the Borel measures μ_k. By analogy with Stone's formula for measures, the intuition behind our approach is that the contribution of $\psi_{\lambda,k}$ to the integral vanishes in the limit $\epsilon \to 0$ unless $\lambda = \lambda_*$. Therefore, we expect that (5.1) approximates an element of the generalized eigenspace associated with λ_* in an appropriate sense when ϵ is small.

While the generalized eigenfunctions are distributions in Φ^*, the approximation constructed from the shifted linear system in (5.1) is an element of the test space Φ, formed from a "wave-packet" of generalized modes that is

strongly concentrated around λ_*. Before considering convergence in two natural topologies in section 5.1.3, we examine the wave-packet approximation scheme through two simple, illustrative examples (see Figure 5.1).

5.1.1 Example 1: multiplication operator

First, let's revisit the "multiplication by x" example from section 1.4.2, i.e., $[\mathcal{L}u](x) = x\,u(x)$. Given $\phi \in C^\infty[-1,1]$, inverting $\mathcal{L} - z$ directly shows that

$$\frac{1}{\pi}\mathrm{Im}(\mathcal{L} - \lambda_* - i\epsilon)^{-1}\phi = \frac{1}{\pi}\frac{\epsilon\phi(x)}{(x-\lambda_*)^2 + \epsilon^2}. \tag{5.2}$$

Note that this is precisely (5.1) with $d\mu_k(\lambda) = d\lambda$ (Lebesgue measure) because $(\phi, \psi_\lambda)_{\Phi,\Phi^*} = \int_{-1}^{1} \phi(y)\delta(y-\lambda)\,dy = \phi(\lambda)$, so the right hand side of (5.1) is

$$\frac{1}{\pi}\int_{\lambda(\mathcal{L})} \frac{\epsilon\phi(\lambda)}{(\lambda-\lambda_*)^2+\epsilon^2}\delta(x-\lambda)\,d\lambda = \frac{1}{\pi}\frac{\epsilon\phi(x)}{(x-\lambda_*)^2+\epsilon^2}.$$

Since the Poisson kernel is an approximation to the identity, (5.2) converges to $\phi(\lambda_*)\delta(x-\lambda_*)$ in the sense of compactly supported distributions on $[-1,1]$. This is the projection of ϕ onto the generalized eigenfunction associated with λ_*.

5.1.2 Example 2: differential operator

Next, consider the differential operator $\mathcal{L}u = -u''$ defined on $H^2(\mathbb{R})$, the space of square-integrable functions on the real line possessing two weak square-integrable derivatives. Then, $\mathcal{L} : H^2(\mathbb{R}) \to L^2(\mathbb{R})$ and we can rig $L^2(\mathbb{R})$ with the dense subspace of functions in the Schwartz space $\mathcal{S}(\mathbb{R})$ and its topological dual $\mathcal{S}^*(\mathbb{R})$, the space of tempered distributions. The spectrum of $\mathcal{L}$ is continuous, filling the real axis, and a straightforward calculation in the Fourier domain reveals that $\mathcal{L}$ has no eigenfunctions in $L^2(\mathbb{R})$.

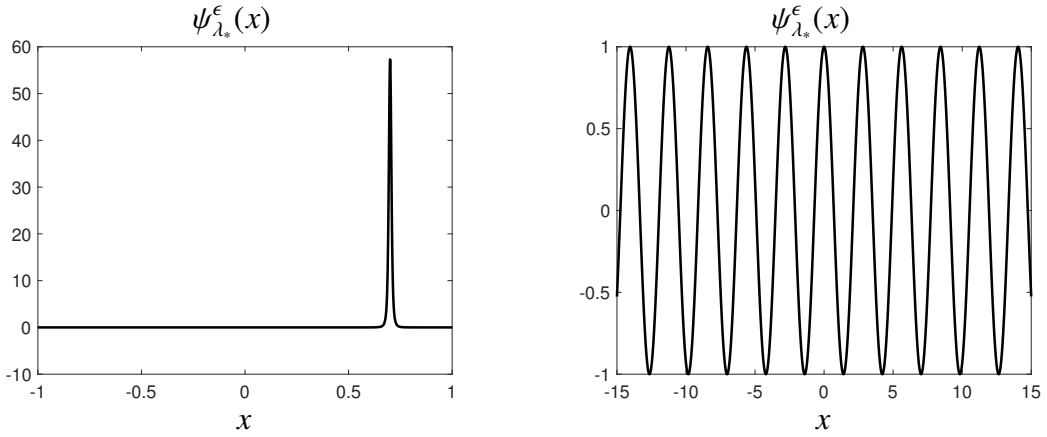

Figure 5.1: Wave-packet approximation of a generalized eigenfunction of (left panel) the multiplication operator in section 5.1.1, corresponding to $\lambda_* = 0.7$ and $\epsilon = 10^{-2}$ and (right panel) the differential operator in section 5.1.2, corresponding to $\lambda_* = 5$ and $\epsilon = 10^{-2}$. The wave-packet approximations converge in the sense of compactly supported distributions for the multiplication operator, but converge pointwise for the differential operator due to additional regularity in the generalized eigenfunctions.

On the other hand, we can calculate the generalized eigenfunctions of $\mathcal{L}$ with two sequential integration-by-parts. We find that

$$\int_{-\infty}^{\infty} e^{\pm 2\pi i k x} \mathcal{L}\phi(x)\, dx = 4\pi^2 k^2 \int_{-\infty}^{\infty} e^{\pm 2\pi i k x} \phi(x)\, dx,$$

whenever $\phi \in \mathcal{S}(\mathbb{R})$. This means that the generalized eigenfunctions of $\mathcal{L}$ are Fourier modes, which form a complete system as both (1.12) and (1.13) hold. The generalized eigenvector coordinates of $\phi \in \Phi = \mathcal{S}(\mathbb{R})$ are closely related to its Fourier transform $\hat{\phi}$, as

$$(\phi, \psi_{4\pi^2 k^2})_{\Phi,\Phi^*} = \int_{-\infty}^{\infty} e^{-2\pi i k x} \phi(x)\, dx + \int_{-\infty}^{\infty} e^{2\pi i k x} \phi(x)\, dx = \hat{\phi}(k) + \hat{\phi}(-k).$$

In other words, $(\phi, \psi_\lambda)_{\Phi,\Phi^*} = \hat{\phi}(\sqrt{\lambda}/2\pi) + \hat{\phi}(-\sqrt{\lambda}/2\pi)$.

Note that there are two generalized eigenfunctions $\exp(\pm i\sqrt{\lambda}x)$ associated with each point $\lambda \in \lambda(\mathcal{L})$. Each generalized eigenvalue has multiplicity 2, and $M = 2$ in (1.12) and (1.13). Inverting $\mathcal{L} - z$ explicitly in the Fourier domain and applying the Fourier inversion theorem, we compute that

$$\psi^{\epsilon}_{\lambda_*} = \frac{1}{\pi}\mathrm{Im}(\mathcal{L} - \lambda_* - i\epsilon)^{-1} f = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\epsilon \hat{f}(k)}{(4\pi^2 k^2 - \lambda_*)^2 + \epsilon^2} e^{2\pi i k x}\, dk. \tag{5.3}$$

By taking an $L^2(\mathbb{R})$ inner product with $\phi \in \mathcal{S}(\mathbb{R})$ and applying Fubini's theorem to interchange the order of integration, we find that

$$(\phi, \psi^\epsilon_{\lambda_*})_{\Phi,\Phi^*} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\epsilon \hat{f}(k)\hat{\phi}(k)}{(4\pi^2 k^2 - \lambda_*)^2 + \epsilon^2}\, dk.$$

Because the Poisson kernel is an approximation to the identity and $\hat{f}(k)\hat{\phi}(k) \in \mathcal{S}(\mathbb{R})$ when $f, \phi \in \mathcal{S}(\mathbb{R})$, the right hand side converges pointwise as $\epsilon \to 0$:

$$(\phi, \psi^\epsilon_{\lambda_*})_{\Phi,\Phi^*} \to \hat{f}\left(\frac{\sqrt{\lambda_*}}{2\pi}\right)\hat{\phi}\left(\frac{\sqrt{\lambda_*}}{2\pi}\right) + \hat{f}\left(\frac{-\sqrt{\lambda_*}}{2\pi}\right)\hat{\phi}\left(\frac{-\sqrt{\lambda_*}}{2\pi}\right).$$

Therefore, we again have convergence in the sense of distributions, i.e.,

$$\psi^\epsilon_{\lambda_*} \to \hat{f}\left(\frac{\sqrt{\lambda_*}}{2\pi}\right) e^{i\sqrt{\lambda_*}x} + \hat{f}\left(\frac{-\sqrt{\lambda_*}}{2\pi}\right) e^{-i\sqrt{\lambda_*}x} \quad \text{as} \quad \epsilon \to 0. \tag{5.4}$$

As before, the wave-packet converges to the projection of $f \in \Phi$ onto the generalized eigenspace associated with λ_*, which has multiplicity two in this case.

5.1.3 Weak* and pointwise convergence

Given a point λ_* in the interior of the absolutely continuous spectrum of $\mathcal{L}$, we now demonstrate that the wave-packet approximation defined in (5.1),

$$\psi^\epsilon_{\lambda_*} = \frac{1}{\pi}\mathrm{Im}(\mathcal{L} - \lambda_* - i\epsilon)^{-1} f,$$

converges to a generalized eigenfunction ψ_{λ_*} of $\mathcal{L}$ in the weak* topology on Φ^*, provided that the generalized eigenvector coordinates are continuous at λ_*. Recall that $\psi_\epsilon \to \psi$ in the weak* topology on Φ^* if and only if it holds that

$$(\phi, \psi_\epsilon)_{\Phi,\Phi^*} \to (\phi, \psi)_{\Phi,\Phi^*}, \qquad \text{for all} \qquad \phi \in \Phi.$$

We denote the interior of the absolutely continuous spectrum of $\mathcal{L}$ by $\Lambda^{\text{a.c.}}_{\text{int}}(\mathcal{L})$.

Theorem 5.1.1. *Let $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{H}$ be a self-adjoint operator on a rigged Hilbert space $\Phi \subset \mathcal{H} \subset \Phi^*$, where Φ is nuclear and $\mathcal{L}\Phi \subset \Phi \subset \mathcal{D}(\mathcal{L})$. If $\psi^{\epsilon}_{\lambda_*}$ is defined as in* (5.1)*, and $(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}$ is continuous at $\lambda_* \in \Lambda^{\text{a.c.}}_{\text{int}}(\mathcal{L})$ for each $k = 1, \ldots, M$ and $\phi \in \Phi$, then*

$$(\phi, \psi^{\epsilon}_{\lambda_*})_{\Phi,\Phi^*} \to \sum_{k=1}^{M_{\lambda_*}} (f, \psi_{\lambda_*,k})_{\Phi,\Phi^*} (\phi, \psi_{\lambda_*,k})_{\Phi,\Phi^*}, \qquad \text{as} \qquad \epsilon \to 0, \tag{5.5}$$

holds for every $\phi \in \Phi$. Here, each $\psi_{\lambda_*,k} \in \Phi^*$ satisfies $(\phi, \mathcal{L}\psi_{\lambda_*,k})_{\Phi,\Phi^*} = \lambda_* (\phi, \psi_{\lambda_*,k})_{\Phi,\Phi^*}$.

In the proof below, it is worth noting that approximations to the individual generalized eigenfunction coordinates converge for almost every $\lambda_* \in \Lambda^{\text{a.c.}}_{\text{int}}(\mathcal{L})$ because Poisson integrals recover integrable functions at their Lebesgue points [14]. In essence, weak* convergence requires that approximations to the generalized eigenfunction coordinates of every $\phi \in \Phi$ converge at λ_*.

Proof. To begin, we expand the duality pairing in generalized eigenfunctions, as

$$(\phi, \psi^{\epsilon}_{\lambda_*})_{\Phi,\Phi^*} = \frac{1}{\pi} \sum_{k=1}^{M} \int_{\Lambda(\mathcal{L})} \frac{\epsilon (f, \psi_{\lambda,k})_{\Phi,\Phi^*} (\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}}{(\lambda - \lambda_*)^2 + \epsilon^2} d\mu_k(\lambda). \tag{5.6}$$

Consider each integral in (5.6) individually. It is useful to note that the Borel measures in Theorem 1.4.2 are uniquely determined up to equivalence of measures, i.e., measures with the same measure-zero sets.[1] Without loss of generality, take them to be positive and set $d\mu_k(\lambda) = d\lambda$ (Lebesgue measure) on $\Lambda^{\text{a.c.}}_{\text{int}}(\mathcal{L}) \bigcup \text{supp}(\mu_k)$. Recall that μ_k vanishes on the spectrum with multiplicity $> k$.

Now, let χ_1 and χ_2 denote the characteristic functions on $\Lambda^{\text{a.c.}}_{\text{int}}(\mathcal{L})$ and its complement in $\Lambda(\mathcal{L})$, respectively. When $k \leq M_{\lambda_*}$, first observe that as $\epsilon \to 0$,

$$\frac{1}{\pi} \int_{\mathbb{R}} \frac{\epsilon (f, \psi_{\lambda,k})_{\Phi,\Phi^*} (\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}}{(\lambda - \lambda_*)^2 + \epsilon^2} \chi_1(\lambda)\, d\mu_k(\lambda) \to (f, \psi_{\lambda_*,k})_{\Phi,\Phi^*} (\phi, \psi_{\lambda_*,k})_{\Phi,\Phi^*}.$$

[1] This reflects the possibility of a change of measure in the integrals in (1.12) and (1.13).

This follows because convolution with the Poisson kernel recovers integrable functions at points of continuity [14]. By hypothesis, $\chi_1(\lambda)(f, \psi_{\lambda,k})_{\Phi,\Phi^*}(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}$ is continuous at λ_* and it is integrable as a consequence of Hölder's inequality and the identity $\|\phi\|_{\mathcal{H}}^2 = \sum_{k=1}^{M} \int_{\lambda(\mathcal{L})} |(\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}|^2 \, d\mu_k(\lambda) < \infty$ [57].

On the other hand, denote the distance between λ_* and $\Lambda(\mathcal{L}) \setminus \Lambda_{\text{int}}^{\text{a.c.}}(\mathcal{L})$ by d. Applying the identity for $\|\phi\|_{\mathcal{H}}^2$ and Hölder's inequality in reverse order now, we bound the integral over the complement of $\Lambda_{\text{int}}^{\text{a.c.}}(\mathcal{L})$ by

$$\frac{1}{\pi} \left| \int_{\mathbb{R}} \frac{\epsilon (f, \psi_{\lambda,k})_{\Phi,\Phi^*} (\phi, \psi_{\lambda,k})_{\Phi,\Phi^*}}{(\lambda - \lambda_*)^2 + \epsilon^2} \chi_2(\lambda) \, d\mu_k(\lambda) \right| \leq \frac{\epsilon}{\pi(d^2 + \epsilon^2)} \|f\|_{\mathcal{H}} \|\phi\|_{\mathcal{H}}.$$

The right hand side vanishes as $\epsilon \to 0$ and a similar bound holds for the integral over $\lambda_{\text{int}}^{\text{a.c.}}(\mathcal{L})$ when $k > M_{\lambda_*}$. Therefore, the right hand side of (5.6) converges to the right hand side of (5.5), which establishes the theorem. □

Although weak* convergence is a natural characterization of convergence in Φ^*, generalized eigenfunctions may possess additional regularity that implies convergence in finer topologies. An important instance of this appears in section 5.1.2, where both generalized eigenfunctions and coordinates are continuous with respect to the spectral parameter $\lambda > 0$. In this case, the wave-packet approximations converge pointwise to elements in the generalized eigenspace.

The use of the Poisson kernel to construct a wave-packet approximation is not particularly important, except in its connection to the resolvent of $\mathcal{L}$. We recover similar approximations if we substitute other approximations to the identity; in particular, we can use the rational kernels developed in [34]. It is straightforward to show that higher-order rational kernels also accelerate convergence rates in this setting when the generalized eigenvector coordinates have additional regularity in the parameter λ_*, although we do not pursue this here.

5.2 Scattering modes in a 2-dimensional quantum device

Over the last several decades, nanoscale devices have become small enough that quantum mechanical models play an essential role in analyzing their functionality [26, 96]. In a typical setup, electrons propagating at the interface of two semiconductors are confined to a strip using large potential barriers, while current is supplied by leads whose length is orders of magnitude greater than the device itself [96]. Material or structural modifications to the heterogeneous interface modify the electronic properties of the device by inducing variations in the local electric field, or equivalently, potential.

The performance of these nanoscale devices can be understood by studying the scattering modes of the (non-dimensionalized) Schrodinger operator [96]

$$[\mathcal{L}u](x,y) = -\Delta u(x,y) + V(x,y)u(x,y), \qquad (x,y) \in \Omega. \tag{5.7}$$

Here, $\Delta = \partial_x^2 + \partial_y^2$ is the Laplacian, $V(x,y)$ is the potential energy within the device, and the 2-dimensional domain Ω models the device and lead geometry. We consider an infinite strip $\Omega = \mathbb{R} \times (-1,1)$, with $V(x,y)$ supported in $[-L,L] \times (-1,1)$ and two semi-infinite leads held at constant potential (without loss of generality, shifted to zero). The confining potential outside the strip is enforced by homogeneous Dirichlet boundary conditions along its edges, i.e., $u(x,\pm 1) = 0$.

The scattering modes that characterize the electronic behavior of the device are classical eigenfunctions of the Schrodinger operator in that they satisfy

$$[\mathcal{L}\psi_{\lambda,k}](x,y) = \lambda\psi_{\lambda,k}(x,y), \qquad (x,y) \in \Omega. \tag{5.8}$$

However, because of propagation through the device leads, scattering modes do not decay and are not square integrable. Instead, they satisfy λ-dependent

asymptotic radiation conditions that are derived from conservation laws. Consequently, numerical methods for scattering modes usually incorporate exact or asymptotic knowledge about the solution in the leads.

In the framework of this chapter, scattering modes can also be formulated as generalized eigenfunctions of $\mathcal{L}$ acting on the rigged Hilbert space $\mathcal{D}(\Omega) \subset L^2(\Omega) \subset \mathcal{D}^*(\Omega)$, where $\mathcal{D}(\Omega)$ is the space of infinitely differentiable functions with compact support in Ω and $\mathcal{D}^*(\Omega)$ is the corresponding dual space of distributions. Going forward, we assume that $V \in \mathcal{D}(\Omega)$ so that $\mathcal{L}\phi \in \mathcal{D}(\Omega)$ whenever $\phi \in \mathcal{D}(\Omega)$. Note that our framework does not require *a priori* knowledge about $\psi_{\lambda,k}(x, y)$ in the leads. (However, incorporating asymptotic information when solving the shifted linear systems sometimes increases numerical efficiency.)

5.2.1 Free scattering modes

Before considering the influence of the potential $V(x, y)$, it is instructive to examine the scattering modes when $V(x, y) = 0$ in Ω. In this case, we can calculate them directly by separation of variables. They can be written as

$$\psi^{\pm}_{n,k}(x) = \begin{cases} \exp(\pm 2\pi i k x)\cos(\pi n y/2), & n \text{ odd}, \\ \exp(\pm 2\pi i k x)\sin(\pi n y/2), & n \text{ even}. \end{cases} \tag{5.9}$$

Here, $k > 0$ is any real number and $n \geq 1$ is an integer. In other words, they are tensor products of the scattering modes of the free particle and the bound states of the particle in a box. The spectrum of $-\Delta$ with the specified boundary conditions on Ω is absolutely continuous, filling the ray $(\pi^2/4, +\infty)$.

Notice that the multiplicity of the generalized eigenspace changes as the continuous spectrum is traversed. Each point λ_* in the spectrum is associated with

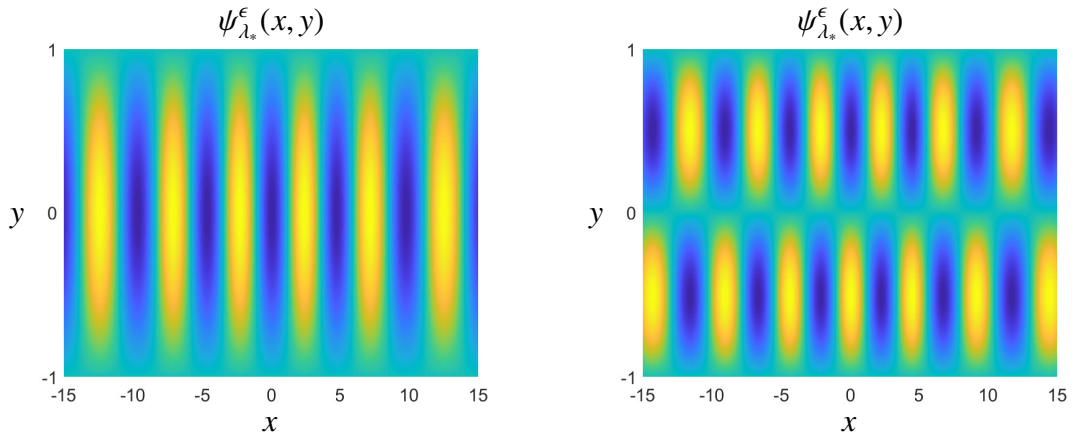

Figure 5.2: The $\epsilon = 0.05$ wave-packet approximations to scattering modes in (5.9) corresponding to $\lambda_* = 7$ ($n = 1$, $k \approx 4.53$) and $\lambda_* = 15$ ($n = 2$, $k \approx 5.13$) in the left and right panels, respectively.

all modes in (5.9) for which $\lambda_* = (2\pi k)^2 + (\pi n/2)^2$. For example, the multiplicity is two between the two lowest transverse modes ($n = 1, 2$), but jumps to four between the second and third lowest transverse modes ($n = 2, 3$). As usual, $\pm k$ in the complex exponential corresponds to, respectively, right and left propagating plane waves in the time domain. Two wave packet approximations to the scattering modes in (5.9), with $\epsilon = 0.05$, are shown in Figure 5.2.

5.2.2 Resonance phenomena

Having examined the scattering states when $V(x, y) = 0$, we now turn on the local potential $V(x, y)$ shown in Figure 5.3. The addition of a finite-range potential does not change the location of the continuous spectrum of $\mathcal{L}$, but it does distort the scattering modes. This particular choice of $V(x, y)$ also supports several bound states that are localized around the potential well. These bound states are associated with eigenvalues of $\mathcal{L}$ below the continuous spectrum.

Now, suppose that we adjust the bias of $V(x, y)$ relative to the leads, so that

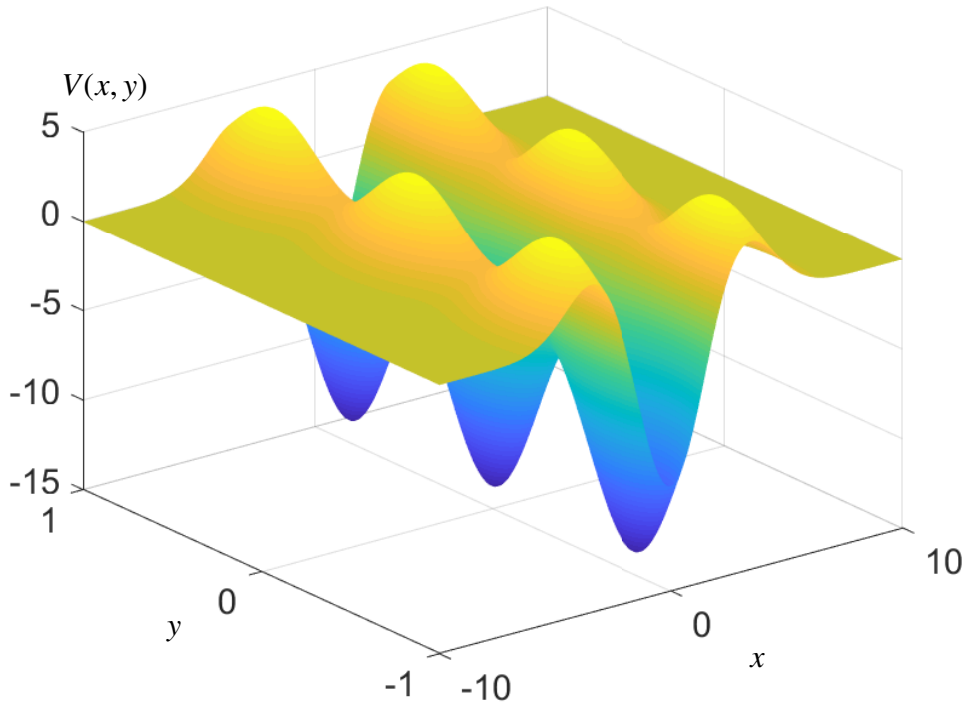

Figure 5.3: The potential energy $V(x, y)$ applied within the 2-dimensional quantum device.

both peaks and valleys are shifted upwards in Figure 5.3. This effectively raises the energy of the bound states, and the corresponding eigenvalues are shifted toward the continuous spectrum. If the energy of a state exceeds the potential in the leads, the bound state becomes a metastable resonant state with a finite lifetime due to the quantum phenomenon of tunneling [100]. This transition occurs when the right-most eigenvalue collides with the continuous spectrum.

We can explore these bound states and resonance states numerically by examining the spectral measure of $\mathcal{L}$ with respect to a Gaussian probe. (Note that in the rigged Hilbert space setting, the density of the spectral measure with respect to f coincides with the generalized eigenvector coordinates of f on the absolutely continuous spectrum.) The spectral measure in the left panel Fig-

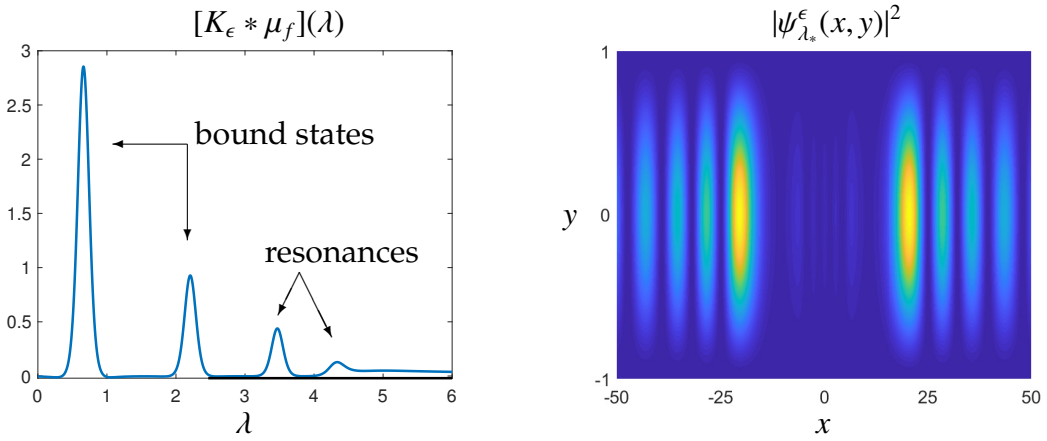

Figure 5.4: A smoothed approximation (computed with smoothing parameter $\epsilon = 0.1$ and an order $m = 3$ rational kernel) to the spectral measure of $\mathcal{L}$, taken with respect to a centered Gaussian probe, reveals bound states and resonances induced by the potential $V(x, y)$ (left panel). The modulus of the wave-packet approximation, $|\psi^\epsilon_{\lambda_*}(x, y)|^2$, shows concentrations of probability mass on both sides of the potential barrier that reflect prolonged interactions with the barrier, i.e., resonance.

ure 5.4 shows four distinct peaks with decreasing height and increasing width from left to right. The two peaks with the lowest energy lie to the left of the continuous spectrum and correspond to bound states of $\mathcal{L}$, but the two peaks on the right are located in the lower region of the continuous spectrum and correspond to resonant states. The right panel of Figure 5.4 shows the squared modulus of a wave-packet approximation to the scattering state associated with the resonance peak at $\lambda_* \approx 3.4$.

In quantum mechanics, the physical properties of particles (or systems) are encoded in wave functions $\psi \in \mathcal{H}$. In the position-space representation, $|\psi(x)|^2$ defines a probability density associated with the particle's location [69, Ch. 1]. Scattering states cannot be directly interpreted in this manner because they are not normalizable. However, the wave-packet approximations $\psi^\epsilon_\lambda(x, y)$ can be because they are elements of $\mathcal{H}$. While the unobstructed scattering states in section 5.2.1 were completely delocalized, notice that the scattering mode associated with the resonance peak is strongly concentrated near the barrier. This

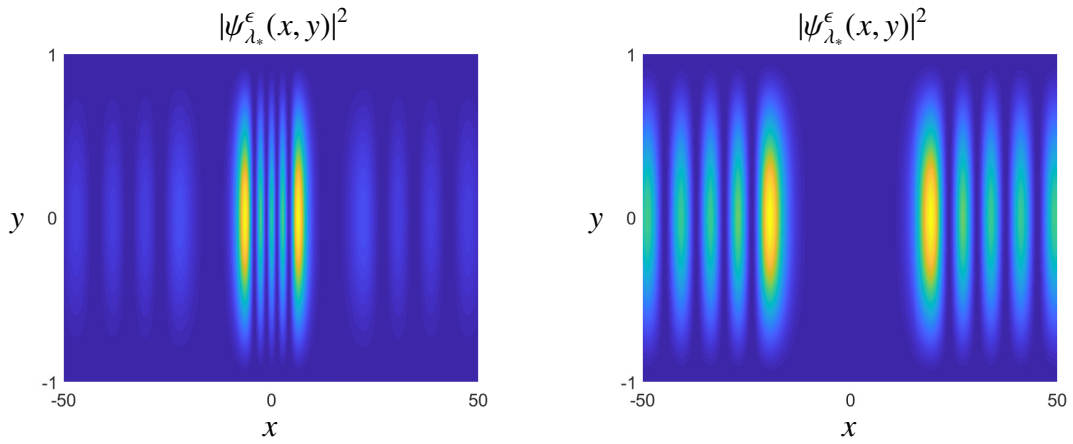

Figure 5.5: The modulus $|\psi^{\epsilon}_{\lambda_*}(x,y)|^2$ of two more wave-packet approximations to the left ($\lambda_* = 3.2$) and right ($\lambda_* = 3.6$) of the resonance peak at $\lambda_* = 3.4$ in Figure 5.4.

concentration reflects prolonged interactions with the potential $V(x,y)$, which drives the resonance phenomenon. A characteristic "lifetime" of these interactions, which measures the average time a particle spends near the potential compared with a free particle, can be estimated from the width of the resonance peaks in the left panel of Figure 5.4 (e.g., see [100] or [49, Ch. 1]).[2]

Finally, note that there is a small concentration of $|\psi^{\epsilon}_{\lambda_*}(x,y)|^2$ within the potential well. By examining other scattering modes in the vicinity of the third peak in Figure 5.4 (left), we can probe the fine structure of the resonance phenomena. To the left side of the peak at lower energies, the scattering modes concentrate inside the potential well (see Figure 5.5, left). To the right side of the peak at higher energies, the scattering modes delocalize and concentrate outside the well (see Figure 5.5, right).

[2]Note that to accurately assess the width d of a peak, the smoothed measure computed with an order m kernel must use a smoothing parameter with $\epsilon \ll d^{1/m}$.

CHAPTER 6

CONCLUSIONS

For the last 70 years, computing the spectral properties of differential and integral operators has usually meant discretizing the operator and computing spectral properties of a matrix. This paradigm is flexible and integrates well with powerful software packages for numerical linear algebra. However, relying on the spectral properties of discretizations can degrade accuracy, pollute the spectrum, or altogether fail to capture infinite-dimensional phenomena related to the continuous spectrum. In this thesis, we have explored solutions to these shortcomings through an alternative "discretization-oblivious" paradigm.

The key ingredient in our "discretization-oblivious" algorithms was the resolvent operator. By sampling its range at strategic complex nodes, we developed robust approximations to eigenvalues, generalized eigenfunctions, and spectral measures. Computationally, we required two tasks: (1) solve shifted linear systems and (2) take inner products. This approach pairs well with existing software systems, such as Chebfun [46] and ApproxFun [105], which encapsulate state-of-the-art spectral methods in a user-friendly interface. Our algorithms have been implemented in a GitHub repository called `SpecSolve` [33].

In Chapter 2, we derived an operator analogue of the FEAST matrix eigensolver to solve differential eigenvalue problems without discretizing the operator. This approach exploits spectrally accurate techniques for computing with functions while preserving key properties, like well-conditioned eigenvalues, of $\mathcal{L}$. The result is an efficient, automated, and accurate eigensolver for differential operators that is adept in the high-frequency regime. We believe the operator analogue of FEAST may also be useful for computing eigenvalues that cluster in

gaps in the continuous spectrum, a challenging problem that arises frequently in condensed matter physics and quantum chemistry, where Green's functions and contour integral eigensolvers have recently been applied successfully [147].

When contour integral methods like FEAST do encounter tightly clustered eigenvalues, dangerous eigenvalues near the quadrature nodes may severely degrade the accuracy of any floating-point computations. Chapter 3 studied this phenomenon in detail and explained why methods that incorporate subspace iteration have a clear advantage here. In particular, large-round off errors incurred during the first iteration are corrected in the second iteration when the matrix is normal; a similar self-correction occurs for non-normal matrices, but the situation is more complicated. This understanding of dangerous eigenvalues in rational subspace iteration opens the door to related questions about the numerical stability of iterative methods for Sylvester's equation, which often require the solution of linear systems with ill-conditioned shifts (e.g., see [133]).

Chapters 4 and 5 introduced a systematic method to compute spectral measures and generalized eigenfunctions, two tools for studying the continuous spectra of self-adjoint operators. Working directly with Stone's formula in the Hilbert space, we showed how to carefully balance adaptive discretization and a smoothing parameter to resolve the spectral measure of the full infinite-dimensional operator. By studying convergence in the smoothing parameter, we overcame limitations in efficiency and accuracy associated with complex shifts near the continuous spectrum. In the rigged Hilbert space setting, we modified this procedure to construct wave-packet approximations to generalized eigenfunctions. This approximation scheme produces elements in a well-behaved test space, requires no *a priori* knowledge about the generalized eigen-

functions, and is accompanied by a simple but rigorous convergence analysis.

Infinite-dimensional spectral problems abound in applied mathematics and there are a few areas where the ideas behind `SpecSolve` are particularly well-poised to contribute. In condensed-matter physics, the electronic properties of topological insulators are frequently studied via spectral properties of the tight-binding Hamiltonian. Conventional algorithms struggle in many physically interesting settings, for example, when the material is disordered (c.f. Anderson localization). In this case, the Hamiltonian has an exotic blend of continuous and discrete spectrum which cannot be reduced using periodicity. Systems that scatter or radiate energy in the presence of long-range interactions are another source of challenging computational spectral problems, due to difficulties determining or enforcing the correct asymptotic boundary behavior. In this case, our (rigged) Hilbert space setting may prove advantageous as the correct behavior is implicit in the limit $\epsilon \to 0$ (e.g., the limiting absorption principle), but no explicit enforcement is needed during computation (see Chapter 5).

The efficiency of `SpecSolve`'s resolvent-based algorithms is intimately linked with the approximation power of rational functions. We have limited ourselves to approximations with fixed poles, but the introduction of the AAA algorithm has ushered in an exciting new era of practical rational approximation with adaptive pole selection [101]. Incorporating these nonlinear approximation schemes into the computational paradigm of this thesis may offer major improvements in efficiency for operators with heavily clustered eigenvalues or nearly singular spectral measures. However, significant mathematical advances are needed for a rigorous understanding of their convergence properties.

Although we have focused on spectral computations, tensions between

infinite-dimensional operators and their discretizations are not uncommon when computing with differential and integral operators. The design of accurate integrators for time-evolution PDEs and efficient iterative methods for PDE optimization are two such areas. Once one has begun it is hard not to wonder, where else might "discretization-oblivious" algorithms be useful?

BIBLIOGRAPHY

[1] P.-A. Absil, R. Sepulchre, P. Van Dooren, and R. Mahony. Cubically convergent iterations for invariant subspace computation. *SIAM J. Matrix Anal. Appl.*, 26(1):70–96, 2004.

[2] A. J. Akbarfam and A. Mingarelli. Higher order asymptotics of the eigenvalues of Sturm–Liouville problems with a turning point of arbitrary order. *Canadian Applied Mathematics Quarterly*, 12:275–301, 2004.

[3] A. J. Akbarfam and A. B. Mingarelli. Higher order asymptotic distribution of the eigenvalues of nondefinite Sturm–Liouville problems with one turning point. *J. Comput. Appl. Math.*, 149(2):423–437, 2002.

[4] N. Alikakos, P. Bates, and X. Chen. Periodic traveling waves and locating oscillating patterns in multidimensional domains. *Trans. Amer. Math. Soc.*, 351(7):2777–2805, 1999.

[5] W. O. Amrein and V. Georgescu. Characterization of bound states and scattering states in quantum mechanics. Technical report, Univ., Geneva, 1973.

[6] L. M. Anguas, M. I. Bueno, and F. M. Dopico. A comparison of eigenvalue condition numbers for matrix polynomials. *Linear Algebra Appl.*, 564:170–200, 2019.

[7] F. V. Atkinson and A. B. Mingarelli. Asymptotics of the number of zeros and of the eigenvalues of general weighted Sturm–Liouville problems. *J. Reine Ang. Math.*, 375:380–393, 1987.

[8] A. P. Austin. Eigenvalues of differential operators by contour integral projection. `http://www.chebfun.org/examples/ode-eig/ContourProjEig.html`, May 2013.

[9] A. P. Austin and L. N. Trefethen. Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic. *SIAM J. Sci. Comp.*, 37(3):A1365–A1387, 2015.

[10] A. Avila and S. Jitomirskaya. The ten martini problem. *Annals of Mathematics*, 170:303–342, 2009.

[11] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM, 2000.

[12] Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010.

[13] A. Barnett and A. Hassell. Fast computation of high-frequency Dirichlet eigenmodes via spectral flow of the interior Neumann-to-Dirichlet map. *Commun. Pure Appl. Math.*, 67(3):351–407, 2014.

[14] H. Bear. Approximate identities and pointwise convergence. *Pac. J. Math.*, 81(1):17–27, 1979.

[15] N. Beer and D. G. Pettifor. The recursion method and the estimation of local densities of states. In *The Electronic Structure of Complex Systems*, pages 769–777. Springer, 1984.

[16] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, second edition, 1999.

[17] D. Bindel and M. Zworski. Theory and computation of resonances in 1D scattering. `http://www.cs.cornell.edu/~bindel/cims/resonant1d/theo2.html`, October 2006.

[18] A. Bjorck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Math. Comput.*, 27(123):579–594, 1973.

[19] I. Bogaert. Iteration-free computation of Gauss–Legendre quadrature nodes and weights. *SIAM J. Sci. Comput.*, 36(3):A1008–A1026, 2014.

[20] A. Böttcher and B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices*. Springer-Verlag, New York, 1999.

[21] N. Boullé and A. Townsend. Learning elliptic partial differential equations with randomized linear algebra. *arXiv preprint arXiv:2102.00491*, 2021.

[22] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Courier Corporation, 2001.

[23] R. Carmona and J. Lacroix. *Spectral Theory of Random Schrödinger Operators*. Prob. Appl. Birkhäuser Boston, 1990.

[24] F. Chatelin. *Spectral Approximation of Linear Operators*. SIAM, 2011.

[25] L. Chen and H.-P. Ma. Approximate solution of the Sturm–Liouville problems with Legendre–Galerkin–Chebyshev collocation method. *Applied Mathematics and Computation*, 206(2):748–754, 2008.

[26] C. Cheng, J.-H. Lee, K. H. Lim, H. Z. Massoud, and Q. H. Liu. 3D quantum transport solver based on the perfectly matched layer and spectral element methods for the simulation of semiconductor nanodevices. *J. Comp. Phys.*, 227(1):455–471, 2007.

[27] P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*, volume 130. SIAM, 2013.

[28] M. J. Colbrook. The foundations of spectral computations via the solvability complexity index hierarchy: Part II. *arXiv:1908.09598*, 2019.

[29] M. J. Colbrook. *The Foundations of Infinite-Dimensional Spectral Computations*. PhD thesis, University of Cambridge, 2020.

[30] M. J. Colbrook. Computing spectral measures and spectral types. *Commun. Math. Phys.*, 384(1):433–501, 2021.

[31] M. J. Colbrook and A. C. Hansen. The foundations of spectral computations via the solvability complexity index hierarchy: Part I. *arXiv:1908.09592*, 2019.

[32] M. J. Colbrook and A. C. Hansen. On the infinite-dimensional QR algorithm. *Numer. Math.*, 143(1):17–83, 2019.

[33] M. J. Colbrook, A. Horning, and A. Townsend. SpecSolve. *github (online) https://github.com/SpecSolve*, 2020.

[34] M. J. Colbrook, A. Horning, and A. Townsend. Computing spectral measures of self-adjoint operators. *SIAM Rev.*, 63(3):489–524, 2021.

[35] H. Cramér. On some classes of nonstationary stochastic processes. In *Proceedings of the Fourth Berkeley symposium on mathematical statistics and probability*, volume 2, pages 57–78. University of Los Angeles, Press Berkeley and Los Angeles, 1961.

[36] B. Ćurgus, A. Fleige, and A. Kostenko. The Riesz basis property of an indefinite Sturm–Liouville problem with non-separated boundary conditions. *Integr. Equ. Oper. Theory*, 77(4):533–557, 2013.

[37] D. Damanik. Singular continuous spectrum for a class of substitution Hamiltonians. *Letters in Mathematical Physics*, 46(4):303–311, 1998.

[38] D. Damanik, M. Embree, and A. Gorodetski. Spectral properties of Schrödinger operators arising in the study of quasicrystals. In *Mathematics of aperiodic order*, pages 307–370. Springer, 2015.

[39] D. Damanik and B. Simon. Jost functions and Jost solutions for Jacobi matrices, I. A necessary and sufficient condition for Szegő asymptotics. *Invent. Math.*, 165(1):1–50, 2006.

[40] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1(131):317–366, 2001.

[41] E. B. Davies. *Linear Operators and Their Spectra*, volume 106. Cambridge University Press, 2007.

[42] R. de la Madrid Modino. *Quantum mechanics in rigged Hilbert space language*. PhD thesis, Ph.D. Thesis, Universidad de Valladolid, 2001.

[43] F. Dell'Oro and V. Pata. Second order linear evolution equations with general dissipation. *Appl. Math. Opt.*, 2019.

[44] J. Dombrowski and P. Nevai. Orthogonal polynomials, measures and recurrence relations. *SIAM J. Math. Anal.*, 17(3):752–759, 1986.

[45] J. J. Dongarra, B. Straughan, and D. W. Walker. Chebyshev tau-QZ algorithm methods for calculating spectra of hydrodynamic stability problems. *Appl. Numer. Math.*, 22(4):399–434, 1996.

[46] T. A. Driscoll, N. Hale, and L. N. Trefethen. Chebfun Guide, 2014.

[47] N. Dunford. A survey of the theory of spectral operators. *Bull. Amer. Math. Soc.*, 64(5):217–274, 1958.

[48] N. Dunford and J. T. Schwartz. *Linear Operators: Part II: Spectral Theory: Self Adjoint Operators in Hilbert Space*. Interscience Publishers, 1963.

[49] S. Dyatlov and M. Zworski. *Mathematical Theory of Scattering Resonances*, volume 200. American Mathematical Soc., 2019.

[50] V. D. Efros, W. Leidemann, and G. Orlandini. Response functions from integral transforms with a Lorentz kernel. *Phys. Lett. B*, 338(2-3):130–133, 1994.

[51] V. D. Efros, W. Leidemann, G. Orlandini, and N. Barnea. The Lorentz integral transform (LIT) method and its applications to perturbation-induced reactions. *J. Phys. G*, 34(12):R459, 2007.

[52] V. D. Efros, W. Leidemann, and V. Y. Shalamova. On calculating response functions via their Lorentz integral transforms. *Few-Body Sys.*, 60(2):35, 2019.

[53] V. Enss. Asymptotic completeness for quantum mechanical potential scattering. *Comm. Math. Phys.*, 61(3):285–291, 1978.

[54] L. C. Evans. *Partial Differential Equations*, volume 19. Amer. Math. Soc., second edition, 2010.

[55] S. Filip, A. Javeed, and L. N. Trefethen. Smooth random functions, random ODEs, and Gaussian processes. *SIAM Rev.*, 61(1):185–205, 2019.

[56] K. O. Friedrichs. On the perturbation of continuous spectra. *Commun. Pure Appl. Math.*, 1(4):361–406, 1948.

[57] M. Gadella and F. Gómez. A measure-theoretical approach to the nuclear and inductive spectral theorems. *Bull. des Sci. Math.*, 129(7):567–590, 2005.

[58] J. Gary. Computing eigenvalues of ordinary differential equations by finite differences. *Math. Comput.*, 19(91):365–379, 1965.

[59] I. M. Gel'fand and Y. N. Vilenkin. *Generalized Functions: Applications of Harmonic Analysis*, volume 4. Academic Press, 2014.

[60] C.-I. Gheorghiu. *Spectral methods for non-standard eigenvalue problems: fluid and structural mechanics and beyond*. Springer Science & Business, 2014.

[61] M. A. Gilles and A. Townsend. Continuous analogues of Krylov subspace methods for differential operators. *SIAM J. Numer. Anal.*, 57(2):899–924, 2019.

[62] V. Girardin and R. Senoussi. Semigroup stationary processes and spectral representation. *Bernoulli*, 9(5):857–876, 2003.

[63] I. M. Glazman. *Direct Methods of Qualitative Spectral Analysis of Singular Differential Operators*. Israel Program for Scientific Translations, 1965.

[64] G. H. Golub and C. F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

[65] T. R. Goodman. The numerical solution of eigenvalue problems. *Math. Comput.*, 19(91):462–466, 1965.

[66] J. Gopalakrishnan, L. Grubišić, and J. Ovall. Filtered subspace iteration for self-adjoint operators. *arXiv preprint arXiv:1709.06694v1*, 2017.

[67] J. Gopalakrishnan, L. Grubišić, and J. Ovall. Spectral discretization errors in filtered subspace iteration. *Math. Comput.*, 89(321):203–228, 2020.

[68] A. Y. Gordon, S. Jitomirskaya, Y. Last, and B. Simon. Duality and singular continuous spectrum in the almost Mathieu equation. *Acta Mathematica*, 178(2):169–183, 1997.

[69] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, third edition, 2018.

[70] S. Güttel, E. Polizzi, P. T. P. Tang, and G. Viaud. Zolotarev quadrature rules and load balancing for the FEAST eigensolver. *SIAM J. Sci. Comp.*, 37(4):A2100–A2122, 2015.

[71] N. Hale and Y. Nakatsukasa. Rayleigh quotient iteration for an operator. `http://www.chebfun.org/examples/ode-eig/RayleighQuotient.html`, March 2017.

[72] B. C. Hall. *Quantum Theory for Mathematicians*, volume 267 of *Graduate Texts in Mathematics*. Springer, 2013.

[73] M. Hamzavi, K.-E. Thylwe, and A. Rajabi. Approximate bound states solution of the Hellmann potential. *Commun. Theor. Phys.*, 60(1):1, 2013.

[74] S. M. Han, H. Benaroya, and T. Wei. Dynamics of transversely vibrating beams using four engineering theories. *JSV*, 225(5):935–988, 1999.

[75] R. Haydock, V. Heine, and M. J. Kelly. Electronic structure based on the local atomic environment for tight-binding bands. *J. Phys. C: Solid State Phys.*, 5(20):2845, 1972.

[76] H. Hellmann. A new approximation method in the problem of many electrons. *J. Chem. Phys*, 3(1):61–61, 1935.

[77] H. V. Henderson and S. Searle. On deriving the inverse of a sum of matrices. *SIAM Rev.*, 23(1):53–60, 1981.

[78] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*, volume 80. SIAM, 2002.

[79] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

[80] K. Hoffman. *Banach Spaces of Analytic Functions*. Prentice–Hall, 1962.

[81] M. H. Holmes. *Introduction to Perturbation Methods*, volume 20. Springer Science & Business Media, 2013.

[82] A. Horning and Y. Nakatsukasa. Twice is enough for dangerous eigenvalues. *To appear in SIAM J. Matrix Anal. Appl., arXiv preprint arXiv:2010.09710*, 2020.

[83] A. Horning and A. Townsend. FEAST for differential eigenvalue problems. *SIAM J. Numer. Anal.*, 58(2):1239–1262, 2020.

[84] D. Hundertmark, M. Meyries, L. Machinek, and R. Schnaubelt. Operator semigroups and dispersive equations. In *16th Internet Seminar on Evolution Equations*, 2013.

[85] S. Joe. Discrete collocation methods for second kind Fredholm integral equations. *SIAM J. Numer. Anal.*, 22(6):1167–1177, 1985.

[86] G. Kallianpur and V. Mandrekar. Spectral theory of stationary H-valued processes. *J. Multivar. Anal.*, 1(1):1–16, 1971.

[87] T. Kato. *Perturbation Theory for Linear Operators*, volume 132. Springer Science & Business Media, 1976.

[88] J. Kestyn, V. Kalantzis, E. Polizzi, and Y. Saad. PFEAST: a high performance sparse eigenvalue solver using distributed-memory linear solvers. In *SC'16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 178–189. IEEE, 2016.

[89] J. Kestyn, E. Polizzi, and P. T. P. Tang. FEAST eigensolver for non-Hermitian problems. *SIAM J. Sci. Comput.*, 38(5):S772–S799, 2016.

[90] A. Kiejna and K. F. Wojciechowski. *Metal Surface Electron Physics*. Elsevier, 1996.

[91] R. Killip and B. Simon. Sum rules for Jacobi matrices and their applications to spectral theory. *Ann. Math.*, 158:253–321, 2003.

[92] W. Koppelman. On the spectral theory of singular integral operators. *Trans. Am. Math. Soc.*, 97(1):35–63, 1960.

[93] P. W. Langhoff. Stieltjes–Tchebycheff moment-theory approach to photoeffect studies in Hilbert space. In *Theory and Applications of Moment Methods in Many-Fermion Systems*, pages 191–212. Springer, 1980.

[94] R. S. Laugesen and M. C. Pugh. Linear stability of steady states for thin film and Cahn–Hilliard type equations. *Archive for rational mechanics and analysis*, 154(1):3–51, 2000.

[95] R. S. Laugesen and M. C. Pugh. Properties of steady states for thin film equations. *European J. Appl. Math.*, 11(3):293–351, 2000.

[96] C. S. Lent and D. J. Kirkner. The quantum transmitting boundary method. *J. Appl. Phys.*, 67(10):6353–6359, 1990.

[97] B. M. Levitan and I. S. Sargsian. *Introduction to Spectral Theory: Selfadjoint Ordinary Differential Operators*, volume 39 of *Translations of Mathematical Monographs*. Amer. Math. Soc., 1975.

[98] L. Lin, Y. Saad, and C. Yang. Approximating spectral densities of large matrices. *SIAM Rev.*, 58(1):34–65, 2016.

[99] C. Lubich. *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2008.

[100] N. Moiseyev. Quantum theory of resonances: calculating energies, widths and cross-sections by complex scaling. *Phys. Rep.*, 302(5-6):212–293, 1998.

[101] Y. Nakatsukasa, O. Sète, and L. N. Trefethen. The AAA algorithm for rational approximation. *SIAM J. Sci. Comput.*, 40(3):A1494–A1522, 2018.

[102] E. D. Napoli, E. Polizzi, and Y. Saad. Efficient estimation of eigenvalue counts in an interval. *Numer. Linear Algebra Appl.*, 23(4):674–692, 2016.

[103] S. Olver and A. Townsend. A fast and well-conditioned spectral method. *SIAM Rev.*, 55(3):462–489, 2013.

[104] S. Olver and A. Townsend. A practical framework for infinite-dimensional linear algebra. In *Proc. 1st Workshop High Perf. Tech. Comput. Dyn. Lang.*, pages 57–62. IEEE Press, 2014.

[105] S. Olver et al. Approxfun v0.10.8 julia package. `https://github.com/JuliaApproximation/ApproxFun.jl`, 2018.

[106] S. A. Orszag. Accurate solution of the Orr–Sommerfeld stability equation. *J. Fluid Mech.*, 50(4):689–703, 1971.

[107] A. M. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. I. *Archive for Rational Mechanics and Analysis*, 1(1):233–241, 1957.

[108] R. D. Pantazis and D. B. Szyld. Regions of convergence of the Rayleigh quotient iteration method. *Numer. Linear Algebra Appl.*, 2(3):251–269, 1995.

[109] B. N. Parlett. The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Math. Comput.*, 28(127):679–693, 1974.

[110] B. N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, 1998.

[111] E. Parzen. On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Stat.*, 28:329–348, 1957.

[112] E. Parzen. Mathematical considerations in the estimation of spectra. *Technometrics*, 3(2):167–190, 1961.

[113] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stats.*, 33(3):1065–1076, 1962.

[114] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Sciences*. Springer Science & Business Media, 2012.

[115] G. Peters and J. H. Wilkinson. Inverse iteration, ill-conditioned equations and Newton's method. *SIAM Review*, 21(3):339–360, 1979.

[116] E. Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B*, 79(11):115112, 2009.

[117] M. B. Priestley. Basic considerations in the estimation of spectra. *Technometrics*, 4(4):551–564, 1962.

[118] C. Puelz, M. Embree, and J. Fillman. Spectral approximation for quasiperiodic jacobi operators. *Integral Equations and Operator Theory*, 82(4):533–554, 2015.

[119] M. Reed and B. Simon. *Methods of Modern Mathematical Physics. Vol. 1. Functional Analysis*. Academic Press, second edition, 1980.

[120] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stats.*, 27(3):832–837, 1956.

[121] M. Rosenblatt. *Stochastic curve estimation*, volume 3 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. IMS, 1991.

[122] W. Rudin. *Functional Analysis. International series in pure and applied mathematics*. McGraw-Hill, Inc., New York, 1991.

[123] D. Ruelle. A remark on bound states in potential-scattering theory. *Il Nuovo Cimento A (1965-1970)*, 61(4):655–662, 1969.

[124] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, 1992.

[125] Y. Saad. *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. SIAM, 2011.

[126] Y. Saad. Analysis of subspace iteration for eigenvalue problems with evolving matrices. *SIAM J. Matrix Anal. Appl.*, 37(1):103–122, 2016.

[127] T. Sakurai and H. Sugiura. A projection method for generalized eigenvalue problems using numerical integration. *J. Comput. Appl. Math.*, 159(1):119–128, 2003.

[128] N. Sanford, K. Kodama, J. D. Carter, and H. Kalisch. Stability of traveling wave solutions to the Whitham equation. *Phys. Lett. A*, 378(30-31):2100–2107, 2014.

[129] J. Shen and L.-L. Wang. Fourierization of the Legendre–Galerkin method and a new space–time spectral method. *Appl. Numer. Math.*, 57(5-7):710–720, 2007.

[130] R. N. Silver and H. Röder. Densities of states of mega-dimensional Hamiltonian matrices. *Inter. J. Modern Phys. C*, 5(04):735–753, 1994.

[131] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[132] B. Simon. Schrödinger semigroups. *Bull. Am. Math. Soc.*, 7(3):447–526, 1982.

[133] V. Simoncini. Computational methods for linear matrix equations. *SIAM Rev.*, 58(3):377–441, 2016.

[134] E. M. Stein and R. Shakarchi. *Complex Analysis*, volume 2 of *Princeton Lectures in Analysis*. Princeton University Press, 2003.

[135] E. M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, volume 3 of *Princeton Lectures in Analysis*. Princeton University Press, 2005.

[136] E. M. Stein and R. Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*, volume 4 of *Princeton Lectures in Analysis*. Princeton University Press, 2011.

[137] G. W. Stewart. Error bounds for approximate invariant subspaces of closed linear operators. *SIAM J. on Num. Anal.*, 8(4):796–808, 1971.

[138] G. W. Stewart. Simultaneous iteration for computing invariant subspaces of non-hermitian matrices. *Numer. Math.*, 25(2):123–136, 1976.

[139] G. W. Stewart. *Matrix Algorithms: Volume II, Eigensystems*. SIAM, 2001.

[140] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990.

[141] P. Stoica and R. L. Moses. *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[142] M. H. Stone. *Linear Transformations in Hilbert Space*, volume 15 of *Amer. Math. Soc. Colloq. Pub.* Amer. Math. Soc., Providence, RI, 1990.

[143] A. Sütő. Singular continuous spectrum on a Cantor set of zero Lebesgue measure for the Fibonacci Hamiltonian. *Journal of statistical physics*, 56(3-4):525–531, 1989.

[144] P. T. P. Tang and E. Polizzi. FEAST as a subspace iteration eigensolver accelerated by approximate spectral projection. *SIAM J. Matrix Anal. Appl.*, 35(2):354–390, 2014.

[145] D. J. Tannor. *Introduction to Quantum Mechanics: a Time-Dependent Perspective*. University Science Books, 2007.

[146] T. Tao. *Nonlinear Dispersive Equations: Local and Global Analysis*. Number 106. AMS, 2006.

[147] K. Thicke, A. B. Watson, and J. Lu. Computing edge states without hard truncation. *SIAM J. Sci. Comput.*, 43(2):B323–B353, 2021.

[148] R. C. Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra Appl.*, 13(1-2):69–78, 1976.

[149] F. Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309(1-3):339–361, 2000.

[150] E. C. Titchmarsh. *Eigenfunction Expansions Associated With Second Order Differential Equations, Part I*. Oxford University Press, second edition, 1962.

[151] D. Tong. Applications of quantum mechanics. *University of Cambridge Part II Mathematical Tripos*, 2017.

[152] A. Townsend and L. N. Trefethen. An extension of Chebfun to two dimensions. *SIAM J. Sci. Comput.*, 35(6):C495–C518, 2013.

[153] A. Townsend and L. N. Trefethen. Continuous analogues of matrix factorizations. *Proc. R. Soc. A*, 471(2173):20140585, 2015.

[154] L. N. Trefethen. Householder triangularization of a quasimatrix. *IMA J. Num. Anal.*, 30(4):887–897, 2009.

[155] L. N. Trefethen. *Approximation Theory and Approximation Practice*, volume 164 of *Other Titles in Applied Mathematics*. SIAM, second edition, 2019.

[156] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. SIAM, 1997.

[157] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: the Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

[158] L. N. Trefethen and M. R. Trummer. An instability phenomenon in spectral methods. *SIAM J. Num. Anal.*, 24(5):1008–1023, 1987.

[159] L. N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. *SIAM Rev.*, 56(3):385–458, 2014.

[160] T. Trogdon, S. Olver, and B. Deconinck. Numerical inverse scattering for the Korteweg–de Vries and modified Korteweg–de Vries equations. *Phys. D: Nonlinear Pheno.*, 241(11):1003–1025, 2012.

[161] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

[162] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall/CRC, 1994.

[163] W. Wasow. Linear turning point theory. *Bull. Amer. Math. Soc.*, 15:252–254, 1986.

[164] M. Webb and S. Olver. Spectra of Jacobi operators via connection coefficient matrices. *Commun. Math. Phys.*, 382(2):657–707, 2021.

[165] M. D. Webb. *Isospectral algorithms, Toeplitz matrices and orthogonal polynomials*. PhD thesis, University of Cambridge, 2017.

[166] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske. The kernel polynomial method. *Rev. Modern Phys.*, 78(1):275, 2006.

[167] J. Wilkening and A. Cerfon. A spectral transform method for singular Sturm–Liouville problems with applications to energy diffusion in plasma physics. *SIAM J. Appl. Math.*, 75(2):350–392, 2015.

[168] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*, volume 87. Oxford University Press, 1965.

[169] P. Zhu and A. V. Knyazev. Angles between subspaces and their tangents. *J. Numer. Math.*, 21(4):325–340, 2013.