SIAM J. SCI. COMPUT.
Vol. 48, No. 6, pp. A3293–A3319
© 2026 Society for Industrial and Applied Mathematics

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

A CONTINUOUS APPROACH TO COMPUTING THE PSEUDOSPECTRA OF LINEAR OPERATORS*

KUAN DENG†, XIAOLIN LIU‡, AND KUAN XU§

For Nick Trefethen on the occasion of his 70th birthday

Abstract. We propose a continuous approach to computing the pseudospectra of linear operators with compact or compact-plus-scalar resolvent, following a "solve-then-discretize" strategy. Instead of taking a finite section approach or using a finite-dimensional matrix to approximate the operator of interest, the new method employs an operator analogue of the Lanczos process to work with operators and functions directly. The method is shown to be free of spectral pollution and spectral invisibility, fully adaptive, and nearly optimal in accuracy. The advantages of the method are demonstrated by extensive numerical examples and comparison with the traditional method.

Key words. pseudospectra, linear operators, Lanczos iteration, spectral methods, infinite-dimensional linear algebra

MSC codes. 65F15, 15A18, 47A10, 47E05, 47G10

DOI. 10.1137/24M1720135



See reproducibility of computational results at end of the article.

1. Introduction.

In his 1999 Acta Numerica paper [42], Nick Trefethen wrote,

> *This field (the computation of pseudospectra) will also participate in a broader trend in the scientific computing of the future, the gradual breaking down of the walls between the two steps of discretization (operator → matrix) and solution (matrix → spectrum or pseudospectra).*

Not only serving as a projection for the future, this sentence also succinctly encapsulates the standard routine for computing the pseudospectra of an operator some 25 years ago. At the time of this writing, it is still the default approach to the task. This approach begins by approximating a linear operator $\mathcal{L}$, e.g., a differential operator, using a discretization method, e.g., a spectral method. If we denote by L this finite-dimensional matrix approximation to $\mathcal{L}$, the pseudospectra of $\mathcal{L}$ are then approximated by those of L, following the definition of the ε-pseudospectrum of a matrix

$$\sigma_\varepsilon(L) = \left\{ z \in \mathbb{C} \,|\, \|(zI - L)^{-1}\| > \varepsilon^{-1} \right\}, \tag{1.1}$$

*Submitted to the journal's Numerical Algorithms for Scientific Computing section January 13, 2025; accepted for publication (in revised form) May 26, 2026; published electronically November 3, 2026.

https://doi.org/10.1137/24M1720135

†School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, Anhui, China (dengkuan@mail.ustc.edu.cn).

‡School of Mathematics and Statistics, Hainan University, Haikou 570228, China, and Key Laboratory of Engineering Modeling and Statistical Computation of Hainan Province, Hainan University, Haikou 570228, China (xiaolin9907@hainanu.edu.cn).

§Corresponding author. School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, Anhui, China (kuanxu@ustc.edu.cn).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 1.1. The main part of the core EigTool algorithm.

Input: A matrix $L \in \mathbb{C}^{n\times n}$, $z \in \mathbb{C}$, a normalized vector $u_1 \in \mathbb{C}^n$, and a tolerance δ.

1: Compute the Schur form $\tilde{L}$ so that L and $\tilde{L}$ are unitarily similar.
2: **repeat**
3: Perform Lanczos iteration $T(z)U_k = U_k H_k + \beta_{k+1} u_{k+1} e_k^T$, where $T(z) = (zI - \tilde{L})^{-*}(zI - \tilde{L})^{-1}$, $U_k = [u_1, \ldots, u_k] \in \mathbb{C}^{n\times k}$, and $u_{k+1} \in \mathbb{C}^n$, to obtain the symmetric tridiagonal matrix $H_k \in \mathbb{R}^{k\times k}$.
4: Compute the largest Ritz value $\lambda_1^{(k)} = \lambda_1(H_k)$.
5: $k \leftarrow k + 1$
6: **until** $|\lambda_1^{(k)}/\lambda_1^{(k-1)} - 1| < \delta$

Output: $\|(zI - L)^{-1}\| \approx \sqrt{\lambda_1^{(k)}}$.

where I is the identity. In case of z being an eigenvalue of L, we take the convention that $\|(z\mathcal{I} - L)^{-1}\| = \infty$. If the 2-norm is adopted, (1.1) amounts to the computation of the smallest singular value of $zI - L$:

$$\sigma_\varepsilon(L) = \{z \in \mathbb{C} \mid s_{\min}(zI - L) < \varepsilon\}.$$

The standard method for computing the approximation to the pseudospectra of a matrix is the core EigTool algorithm [44, section 39]. The main part of this algorithm computes the resolvent norm of L and is recapitulated in Algorithm 1.1.

This "discretize-then-solve" paradigm has a few drawbacks: (1) There is no guarantee that the pseudospectra of L serve as a quality approximation to that of the original operator, since L is only finite dimensional. More importantly, the computation may suffer from the so-called *spectral pollution,* i.e., spurious eigenvalues and pseudospectra, and *spectral invisibility,* i.e., missing parts of the spectrum and pseudospectra. See subsection 5.2 for an example. This is known as the "finite section" caveat [15]. (2) It is difficult to effect adaptivity. (3) $T(z)$ might be more ill conditioned than the original problem. See subsection 5.3 for an example. (4) Even though spectral approximations to the operators are used, the convergence of the resolvent norm may fall short of spectral in terms of the discretization size [44, section 43]. See also subsection 5.3 for an example.

More recently, computing the pseudospectra of general bounded infinite matrices is addressed in [15], followed by a significantly improved version [4, 6] that offers a rigorous error control and the ability to deal with a greater range of problems, such as partial differential operators in unbounded domains. It is proposed to represent an operator $\mathcal{L}$ as a matrix of infinite dimension that acts on $\ell^2(\mathbb{N})$. For a sequence of nested operator foldings and rectangular truncations of this infinite matrix, the limit of the smallest singular value is shown to be $\|(z\mathcal{I} - \mathcal{L})^{-1}\|^{-1}$. This iterative method nevertheless also lacks adaptivity in stopping criterion and may have difficulty in dealing with operators restricted by boundary conditions, e.g., differential operators in compact domains.

This paper shows that the pseudospectra of linear operators with compact or compact-plus-scalar resolvent can be computed following a "solve-then-discretize" strategy [3, 5, 7, 10, 18] by taking advantage of the recent development in spectral methods. The new method is an operator analogue of the Lanczos process [20] that computes with operators and functions directly. This leads to a number of advantages: (1) We use effectively infinite-dimensional matrix representation of the operators, which rules out spectral pollution and spectral invisibility. (2) Justified

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

by a priori analysis, both the application of resolvents and the stopping criterion of the Lanczos process are adaptive so that a nearly optimal accuracy can be obtained if a stringent exit condition is used. (3) We use the recently developed fast and well-conditioned spectral methods, making the most of the floating-point precision. (4) There is no involvement of quadrature or weight matrices, and the convergence is spectral. More importantly, the new method offers a unified framework for computing the pseudospectra of linear operators with compact or compact-plus-scalar resolvent. Though we confine ourselves in this article with certain spectral methods only, any method that solves the inverse resolvent equations of the operator of interest (see (3.1)) can be incorporated into the proposed framework straightforwardly in a plug-and-play fashion.

Throughout this paper, we assume that $\mathcal{L} : \mathcal{D}(\mathcal{L}) \subseteq \mathcal{H} \to \mathcal{H}$ is a closed linear operator on a separable Hilbert space $\mathcal{H}$. For $z \in \mathbb{C}$, the resolvent of $\mathcal{L}$ at z is the operator

$$\mathcal{R}(z) = (z\mathcal{I} - \mathcal{L})^{-1},$$

given that $\mathcal{R}(z) \in \mathcal{B}(\mathcal{H})$ exists, where $\mathcal{B}(\mathcal{H})$ denotes the set of bounded operators on $\mathcal{H}$. For $\varepsilon > 0$, the ε-pseudospectrum of $\mathcal{L}$ is

$$\sigma_\varepsilon(\mathcal{L}) = \left\{ z \in \mathbb{C} \mid \|(z\mathcal{I} - \mathcal{L})^{-1}\| > \varepsilon^{-1} \right\} \tag{1.2}$$

with the convention $\|(z\mathcal{I} - \mathcal{L})^{-1}\| = \infty$ if $z \in \sigma(\mathcal{L})$, where $\sigma(\mathcal{L})$ denotes the spectrum of $\mathcal{L}$. This definition can be found in, for example, [44, section 4].

In this paper, $\langle \cdot, \cdot \rangle$ denotes the inner product on the Hilbert space $\mathcal{H}$, unless stated otherwise, and $\|\cdot\|$ denotes a norm whose meaning is clear from the context. For elements of $\mathcal{H}$, $\|\cdot\|$ refers to the norm induced by $\langle \cdot, \cdot \rangle$, while for bounded linear operators on $\mathcal{H}$, $\|\cdot\|$ denotes the operator norm induced by this Hilbert space norm. For vectors and matrices, $\|\cdot\| = \|\cdot\|_2$. $\Lambda(\cdot)$ and $\sigma(\cdot)$ are used to denote the set of eigenvalues and the spectrum of an operator, respectively. The largest eigenvalue, if it exists and is isolated, is denoted by $\lambda_1(\cdot)$. $\Re(\cdot)$ stands for the real part of a complex number, and the upright Roman letter i denotes the imaginary unit. $\mathcal{I}$ and I_∞ are respectively the identity operator and infinite-dimensional identity matrix. An asterisk is used to denote the complex conjugate of a complex variable or the adjoint of an operator. We denote by ϵ_m the machine epsilon, which is, for example, about 2.22×10^{-16} in the double precision floating-point arithmetic. Throughout, we follow [17, p. 68] to take $\gamma_k = ck\epsilon_m/(1 - ck\epsilon_m)$ for an integer k, where c is a small integer constant whose value is unimportant.

The paper is structured as follows. In section 2, we introduce a Lanczos-based algorithm to compute the resolvent norm of an operator. Section 3 discusses the implementation of the Lanczos iteration for operators. In section 4, we provide a careful error analysis of the computed resolvent norm and the resulting pseudospectra over a prescribed region in the complex plane, which in turn leads to an adaptive stopping criterion for the continuous Lanczos process. We validate the proposed method through extensive numerical experiments in section 5 and then demonstrate several important extensions in section 6. Finally, section 7 concludes the paper with a summary.

2. Operator analogue of the core EigTool algorithm. The definition (1.2) offers limited guidance on the practical computation of an operator's pseudospectra. To relate the norm in (1.2) to a computable quantity, we note that

$$\|(z\mathcal{I} - \mathcal{L})^{-1}\|^2 = \|\mathcal{R}^*(z)\mathcal{R}(z)\| = \sup \sigma(\mathcal{R}^*(z)\mathcal{R}(z)).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 2.1. The operator analogue of Algorithm 1.1.

Input: A linear operator $\mathcal{L}:\mathcal{D}(\mathcal{L})\to\mathcal{H}$, $z\in\mathbb{C}$, a normalized function $u_1\in\mathcal{H}$, a tolerance δ

1: **repeat**

2: Perform kth Lanczos iteration (see Algorithm 3.1 for details)

$$\mathcal{T}(z)\mathcal{U}_k=\mathcal{U}_kH_k+\beta_{k+1}u_{k+1}e_k^T, \tag{2.2}$$

where $\mathcal{T}(z)=(z\mathcal{I}-\mathcal{L})^{-*}(z\mathcal{I}-\mathcal{L})^{-1}$, $\mathcal{U}_k=(u_1|\cdots|u_k)$, and $u_1,\dots,u_{k+1}\in\mathcal{H}$, to obtain the symmetric tridiagonal matrix $H_k\in\mathbb{R}^{k\times k}$.

3: Compute the largest Ritz value $\lambda_1^{(k)}=\lambda_1(H_k)$ and the corresponding eigenvector $y_1^{(k)}=[y_{11}^{(k)},\dots,y_{k1}^{(k)}]^T\in\mathbb{R}^k$.

4: $k\leftarrow k+1$

5: **until** $\beta_{k+1}|y_{k1}^{(k)}|<\delta_1(z)$ (see (4.41) for a practical stopping criterion)

Output: $\|(z\mathcal{I}-\mathcal{L})^{-1}\|\approx\sqrt{\lambda_1^{(k)}}$.

If $\|\mathcal{R}^*(z)\mathcal{R}(z)\|$ happens to be the largest eigenvalue and is isolated, i.e.,

$$\sup\sigma(\mathcal{R}^*(z)\mathcal{R}(z))=\lambda_1(\mathcal{R}^*(z)\mathcal{R}(z)), \tag{2.1}$$

then Algorithm 1.1 can be generalized to Algorithm 2.1, an operator analogue of Algorithm 1.1, to deal with $\mathcal{L}$. In other words, the approximation of the resolvent norm of an operator $\mathcal{L}$ boils down to finding the largest eigenvalue of $\mathcal{T}(z)=\mathcal{R}^*(z)\mathcal{R}(z)$. Approximating the pseudospectra of $\mathcal{L}$ over a region of interest further requires applying Algorithm 2.1 repeatedly at multiple points in the complex plane.

The most notable distinction of Algorithm 2.1 from Algorithm 1.1 lies in the substitution of matrix L with the operator $\mathcal{L}$. Additionally, there are two minor changes. First, the preliminary triangularization (line 1 of Algorithm 1.1), which is instrumental in reducing the computational cost [23], has been removed. Second, a new stopping criterion is introduced in line 5 and further elaborated upon in subsection 4.3.

Now we wonder for what kind of linear operator $\mathcal{L}$ (2.1) holds so that the pseudospectra of $\mathcal{L}$ can be computed via Algorithm 2.1. The following lemma shows that this is the case if the resolvent $\mathcal{R}(z)$ is compact or compact-plus-scalar.

LEMMA 2.1. (1) *If $\mathcal{R}(z)$ is compact,* $\sup\sigma(\mathcal{T}(z))=\lambda_1(\mathcal{T}(z))$. (2) *If $\mathcal{R}(z)$ is compact-plus-scalar, i.e., $\mathcal{R}(z)=\mathcal{K}+\tilde{z}\mathcal{I}$, where $\tilde{z}\in\mathbb{C}$ and $\mathcal{K}$ is compact on $\mathcal{H}$, then* $\sup\sigma(\mathcal{T}(z))=\lambda_1(\mathcal{T}(z))$, *except the case of* $\sup\sigma(\mathcal{T}(z))=|\tilde{z}|^2$.

Proof. Statement (1) follows from the fact that $\mathcal{T}(z)=\mathcal{R}(z)^*\mathcal{R}(z)$ is a compact self-adjoint operator and the spectral theorem [11, section 4.5]. To show (2), let

$$\mathcal{T}(z)=(\mathcal{K}+\tilde{z}\mathcal{I})^*(\mathcal{K}+\tilde{z}\mathcal{I})=\mathcal{J}+|\tilde{z}|^2\mathcal{I},$$

where $\mathcal{J}=\tilde{z}\mathcal{K}^*+\tilde{z}^*\mathcal{K}+\mathcal{K}^*\mathcal{K}$. Noting that $\mathcal{J}$ is a compact self-adjoint operator, we use the spectral theorem again to have

$$\sup\sigma(\mathcal{T}(z))=\sup\Lambda(\mathcal{J})+|\tilde{z}|^2,$$

where $\Lambda(\mathcal{J})$ forms a real countably infinite sequence with a unique accumulation point at 0. When $\sup\Lambda(\mathcal{J})\neq 0$, $\sup\sigma(\mathcal{T}(z))=\lambda_1(\mathcal{J})+|\tilde{z}|^2=\lambda_1(\mathcal{T}(z))$. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The results we collect in the following subsections reveal that virtually all of the common linear operators defined on a one-dimensional bounded domain have compact or compact-plus-scalar resolvent. These include differential operators restricted by proper boundary conditions and integral operators of Volterra and Fredholm types, which we now discuss individually and show how the corresponding $\mathcal{R}^*(z)$ is formulated. Without loss of generality, we consider only the operators acting on $L^2([-1,1])$. We defer to section 6 the discussion of operators acting on Hilbert spaces defined on unbounded or higher-dimensional domains.

2.1. Differential operators. For the differential expression

$$\tau = \frac{\mathrm{d}^N}{\mathrm{d}x^N} + a_{N-1}(x)\frac{\mathrm{d}^{N-1}}{\mathrm{d}x^{N-1}} + \cdots + a_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + a_0(x), \tag{2.3}$$

where $a_k(x)$ for $k = 0, 1, \ldots, N-1$ are locally integrable on $[-1,1]$, the *maximal operator* $\mathcal{Q}_{\max} : L^2([-1,1]) \to L^2([-1,1])$ is defined by $\mathcal{Q}_{\max} g = \tau g$ with $\mathcal{D}(\mathcal{Q}_{\max}) = \{g \in \mathcal{AC}_N([-1,1]) \cap L^2([-1,1]) \mid \tau g \in L^2([-1,1])\}$.[1] Suppose that $\mathcal{Q}$ is a restriction of $\mathcal{Q}_{\max}$ by proper boundary conditions [12, section XIV.3]. It can be shown that $\mathcal{R}(z) = (z\mathcal{I} - \mathcal{Q})^{-1}$ is compact for $z \in \mathbb{C}\backslash\sigma(\mathcal{Q})$ [12, section XIV.3]. For $\mathcal{R}^*(z)$, we use the fact that $(z\mathcal{I} - \mathcal{Q})^{-*} = (z^*\mathcal{I} - \mathcal{Q}^*)^{-1}$, where $\mathcal{Q}^*$ can be found by definition. See, for example, [12, section XIV.4].

2.2. Fredholm integral operator. The Fredholm integral operator $\mathcal{F} : L^2([-1,1]) \to L^2([-1,1])$ is defined as

$$(\mathcal{F}u)(s) = \int_{-1}^{1} K(s,t)u(t)\mathrm{d}t, \quad -1 \le s \le 1, \tag{2.4}$$

where the kernel $K(s,t) \in L^2([-1,1]^2)$. For any nonzero $z \in \mathbb{C}$ that is not an eigenvalue of $\mathcal{F}$, $\mathcal{R}(z) = (z\mathcal{I} - \mathcal{F})^{-1}$ is compact-plus-scalar [35, Theorem 2.1.2]. To evaluate $\mathcal{R}^*(z)$, we use $(z\mathcal{I} - \mathcal{F})^{-*} = (z^*\mathcal{I} - \mathcal{F}^*)^{-1}$, where $(\mathcal{F}^*u)(s) = \int_{-1}^{1} K^*(t,s)u(t)\mathrm{d}t$.

One of the most important subcases of the Fredholm integral operator is the Fredholm convolution integral operator

$$(\mathcal{F}_c u)(s) = \int_{-1}^{1} K(s-t)u(t)\mathrm{d}t, \quad -1 \le s \le 1. \tag{2.5}$$

2.3. Volterra integral operator. The Volterra integral operator $\mathcal{V} : L^2([-1,1]) \to L^2([-1,1])$ given by

$$(\mathcal{V}u)(s) = \int_{-1}^{s} K(s,t)u(t)\mathrm{d}t, \quad -1 \le s \le 1, \tag{2.6}$$

can be deemed as a subcase of the Fredholm integral operator [35, section 2.7] with the kernel $K(s,t) \in L^2([-1,1]^2)$ zero-valued for $-1 \le s < t \le 1$. For any nonzero $z \in \mathbb{C}$, $\mathcal{R}(z) = (z\mathcal{I} - \mathcal{V})^{-1}$ is compact-plus-scalar [35, Theorem 2.7.1]. For $\mathcal{R}^*(z)$, we use the fact that $(z\mathcal{I} - \mathcal{V})^{-*} = (z^*\mathcal{I} - \mathcal{V}^*)^{-1}$, where $(\mathcal{V}^*u)(s) = \int_{s}^{1} K^*(t,s)u(t)\mathrm{d}t$. Similarly, an important subcase of the Volterra integral operator is the Volterra convolution integral operator

$$(\mathcal{V}_c u)(s) = \int_{-1}^{s} K(s-t)u(t)\mathrm{d}t, \quad -1 \le t \le s \le 1. \tag{2.7}$$

[1] $\mathcal{AC}_N([-1,1])$ denotes the set of functions on $[-1,1]$ whose $(N-1)$th derivative exists and is absolutely continuous.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 3.1. Lanczos process.

Input: $\mathcal{T}(z)$, H_{k-1}, β_k, u_k, u_{k-1}
1: $w_k = \mathcal{T}(z)u_k$
2: **if** $k > 1$ **then**
3: $\quad p_k = w_k - \beta_k u_{k-1}$
4: **else**
5: $\quad p_k = w_k$
6: **end if**
7: $\alpha_k = \Re(\langle u_k, p_k\rangle)$
8: $q_k = p_k - \alpha_k u_k$
9: $\beta_{k+1} = \|q_k\|$
10: $u_{k+1} = q_k/\beta_{k+1}$
Output: H_k, β_{k+1}, u_{k+1}

3. Implementation of the Lanczos iteration. To facilitate the discussion, we elaborate upon the Lanczos iteration in Algorithm 3.1, which can be embedded into Algorithm 2.1 at line 2 there.

To implement Algorithm 3.1 for a general Hilbert space $\mathcal{H}$, we represent the iterates $u_k, w_k, p_k, q_k \in \mathcal{H}$ as series in a suitable basis $\{\varphi_i\}_{i=1}^{\infty}$ of $\mathcal{H}$. A natural choice is an orthonormal basis, with which all the operations in Algorithm 3.1, except the application of $\mathcal{T}(z)$ (line 1), can be effected directly using the coefficients of these series. In particular, the inner product $\langle u_k, p_k\rangle$ in line 7 is evaluated as the ℓ^2 inner product of the corresponding coefficient vectors, and the norm $\|q_k\|$ in line 9 is obtained as the ℓ^2-norm of the coefficient vector of q_k.

For line 1 of Algorithm 3.1, applying $\mathcal{T}(z) = (z\mathcal{I} - \mathcal{L})^{-*}(z\mathcal{I} - \mathcal{L})^{-1}$ to $u \in \mathcal{H}$ amounts to solving

$$(z\mathcal{I} - \mathcal{L})v = u, \tag{3.1a}$$

$$(z^*\mathcal{I} - \mathcal{L}^*)w = v \tag{3.1b}$$

in sequence. We assume that, with suitably chosen bases $\{\phi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ for the Hilbert space $\mathcal{H}$, the operators $z\mathcal{I} - \mathcal{L}$ and $z^*\mathcal{I} - \mathcal{L}^*$ admit infinite-dimensional banded matrix representations. Under this representation, the operator equations in (3.1) can be formulated as infinite-dimensional banded linear systems. Specifically, let $\mathbf{v}$ and $\mathbf{u}$ be the coefficient vectors of v and u with respect to the bases $\{\phi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$, respectively, and (3.1a) is written as the infinite-dimensional linear system

$$A\mathbf{v} = \mathbf{u}. \tag{3.2}$$

Suppose that the bandwidth of A is b and that the first n_u elements of $\mathbf{u}$ are nonzero. Following [25], we use the adaptive QR to triangularize A with a sequence of Householder reflectors and apply the same reflectors to the right-hand side $\mathbf{u}$ simultaneously. Assuming $n \geq n_u$, we denote by $Q_n \in \mathbb{C}^{(n+b)\times(n+b)}$ the product of the first n Householder reflectors. After the first n steps, the left-hand side of the partially triangularized system is

$$\begin{pmatrix} Q_n^* & \\ & I_\infty \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ & A_{22} \end{pmatrix} = \begin{pmatrix} R_{11} & Q_n^* A_{12} \\ & A_{22} \end{pmatrix}, \tag{3.3}$$

where $A_{11} \in \mathbb{C}^{(n+b)\times n}$ and $R_{11} \in \mathbb{C}^{n\times n}$ is upper triangular. The resulting infinite vector $\mathbf{r}$ on the right-hand side is

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\mathbf{r} = \begin{pmatrix} Q_n^* & \\ & I_\infty \end{pmatrix} \begin{pmatrix} \mathbf{u}_{1:n_u} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{1:n} \\ \mathbf{r}_{n+1:n+b} \\ 0 \end{pmatrix}, \tag{3.4}$$

whose first $n+b$ elements are nonzero. We can then obtain a truncated solution

$$\bar{\mathbf{v}} = \begin{pmatrix} R_{11}^{-1}\mathbf{r}_{1:n} \\ 0 \end{pmatrix} \tag{3.5}$$

by back substitution. In exact arithmetic, $\mathbf{r}_{n+1:n+b}$ is exactly the error of $\bar{\mathbf{v}}$ in residual [25]. Thus, we terminate the QR process if $\|\mathbf{r}_{n+1:n+b}\| < \delta_{\mathbf{r}}\|\mathbf{r}\|$ for a preset tolerance $\delta_{\mathbf{r}}$. Otherwise, we keep applying Householder reflectors until $\|\mathbf{r}_{n+1:n+b}\| < \delta_{\mathbf{r}}\|\mathbf{r}\|$ is satisfied.

Operator equation (3.1b) can be solved in an analogous manner.

3.1. Spectral methods for infinite-dimensional banded system. As we concentrate below on the operators discussed in section 2, we henceforth assume that $\mathcal{H} = L^2([-1,1])$. Accordingly, we choose $\varphi_i = \tilde{P}_{i-1}$ for $i = 1, 2, \ldots$, where

$$\tilde{P}_i(x) = \sqrt{\frac{2i+1}{2}}\, P_i(x), \tag{3.6}$$

and P_i denotes the standard Legendre polynomial of degree i. With this normalization, $\{\tilde{P}_i\}_{i=0}^{\infty}$ forms an orthonormal basis of $L^2([-1,1])$.

If the variable coefficients or kernel functions are general functions rather than polynomials, we assume that they possess sufficient regularity to be approximated to any prescribed accuracy by polynomials, thereby leading to infinite-dimensional banded systems. In principle, to minimize the error in the ∞- and 2-norms, one should use best approximation [28, 43] and truncated Legendre series [45], respectively. In practice, however, computing the best approximation or the truncated Legendre series of a given function is expensive and, more importantly, yields no significant improvement in accuracy over the Chebyshev interpolant of the same degree in the ∞- and 2-norms, respectively [43, section 16].

Thus, we approximate variable coefficients and univariate kernel functions by Chebyshev series computed via the fast cosine transform, with the degree selected automatically, for example, by the chopping algorithm [1]. Since coefficient-based spectral methods for integral equations require kernel functions to be represented in terms of Legendre polynomials, for Volterra integral operators we further convert the Chebyshev series approximating the univariate kernel function into a Legendre expansion of the same degree using the fast polynomial transform [19]. For Fredholm integral operators, the bivariate kernel function is approximated by a low-rank adaptive cross approximation (ACA) [40]. Such an ACA is based on one-dimensional Chebyshev expansions, which are also subsequently converted into Legendre expansions. Although both the chopping algorithm and the ACA construction are heuristic, they perform reliably in practice.

Now we detail the construction of the infinite-dimensional banded systems for (3.1a) with the operators discussed in section 2; the same procedure applies, *mutatis mutandis,* to (3.1b).

3.1.1. Differential operator. When $\mathcal{L}$ is the differential operator $\mathcal{Q}$ defined in subsection 2.1, we represent or approximate the variable coefficients $a_j(x)$ by Chebyshev series $\tilde{a}_j(x)$ of degree m^{a_j} for $j = 0, 1, \ldots, N-1$ and denote the difference between $a_j(x)$ and $\tilde{a}_j(x)$ by $\Delta a_j(x)$. That is,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$\tilde{a}_j(x) = a_j(x) + \Delta a_j(x). \tag{3.7}$$

This yields a differential operator $\tilde{\mathcal{Q}}$ defined by

$$\tilde{\mathcal{Q}} = \frac{\mathrm{d}^N}{\mathrm{d}x^N} + \tilde{a}_{N-1}(x)\frac{\mathrm{d}^{N-1}}{\mathrm{d}x^{N-1}} + \cdots + \tilde{a}_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + \tilde{a}_0(x), \tag{3.8}$$

which is subject to the same boundary conditions as $\mathcal{Q}$.

To solve (3.1a), we represent v in a recombined basis [29], formed by suitable linear combinations of neighboring normalized Legendre polynomials that satisfy the boundary conditions,[2] while representing the range of $z\mathcal{I} - \tilde{\mathcal{Q}}$ in the ultraspherical basis $\{C_{i-1}^{(N+1/2)}(x)\}_{i=1}^{\infty}$ of half-integer parameter. Consequently, the ultraspherical spectral method[3] [25, 2] yields an infinite-dimensional matrix representation of $z\mathcal{I} - \tilde{\mathcal{Q}}$ that is strictly banded. Notably, the multiplication operators corresponding to premultiplication of a normalized Legendre series by a Chebyshev series can be constructed using the operator-valued Clenshaw algorithm [38, section 6.3.1].

On the right-hand side of the infinite-dimensional system, the coefficients of u in the $\{\tilde{P}_{i-1}(x)\}_{i=1}^{\infty}$ basis are also promoted to those in the $\{C_{i-1}^{(N+1/2)}(x)\}_{i=1}^{\infty}$ basis to match the left-hand side [25]. After solving the resulting system, the solution—expressed in coefficients of recombined normalized Legendre polynomials—is transformed back to coefficients in the $\{\tilde{P}_{i-1}(x)\}_{i=1}^{\infty}$ basis [29].

3.1.2. Fredholm operator. When $\mathcal{L}$ is a general Fredholm operator $\mathcal{F}$ defined by (2.4), we first represent (or approximate) the kernel $K(s,t)$ by a low-rank cross representation [40] of rank r, that is,

$$\tilde{K}(s,t) = \sum_{j=1}^{r} f_j(s)g_j(t) = K(s,t) + \Delta K(s,t), \tag{3.9}$$

where f_j and g_j $(j = 1, 2, \ldots, r)$ are normalized Legendre series of degrees m^f and m^g, respectively. Let $\tilde{\mathcal{F}}$ denote the Fredholm operator defined by $\tilde{K}(s,t)$. Then

$$(z\mathcal{I} - \tilde{\mathcal{F}})v = zv(s) - \sum_{j=1}^{r} \langle g_j^*(t), v(t)\rangle f_j(s);$$

that is, the matrix representation of $z\mathcal{I} - \tilde{\mathcal{F}}$ with respect to the normalized Legendre basis is the sum of an infinite-dimensional diagonal matrix and a rank-r semiseparable matrix,

$$M_{\tilde{\mathcal{F}}}(z) = zI_\infty - CR^*,$$

where $C = [C_1, \ldots, C_r] \in \mathbb{C}^{\infty\times r}$ and $R = [R_1, \ldots, R_r] \in \mathbb{C}^{\infty\times r}$. The first $m^f + 1$ entries of C_j are the normalized Legendre coefficients of $f_j(s)$, with the remaining entries equal to zero. Similarly, R_j is zero, except for its first $m^g + 1$ entries, which store the normalized Legendre coefficients of $g_j(t)$. Consequently, only the top-left $(m^f + 1) \times (m^g + 1)$ block and the diagonal of $M_{\tilde{\mathcal{F}}}(z)$ are nonzero.

[2] Such a basis is commonly referred to as a recombined basis, and is also known by various other names in different contexts, including integrated Legendre polynomials (of Babuška), bubble functions, modal bases, and boundary-adapted bases. The use of such recombined bases in spectral methods was pioneered by Shen [33, 34].

[3] The method used here is a variant of the original ultraspherical spectral method, in that the latter works with Chebyshev polynomials and ultraspherical polynomials of integer orders, i.e., $C^{(N)}(x)$ for $N = 1, 2, \ldots$, whereas we employ ultraspherical polynomials of half-integer orders, i.e., $C^{(N+1/2)}(x)$ for $N = 0, 1, \ldots$, since the solution is represented in normalized Legendre polynomials.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

3.1.3. Fredholm convolution operator. When $\mathcal{L}$ is a Fredholm convolution integral operator $\mathcal{F}_c$ given by (2.5), we represent (or approximate) the univariate kernel $K(x)$ by a finite-degree normalized Legendre series $\tilde{K}(x)$ of degree m^K, which introduces a perturbation $\Delta K(x)$, that is,

$$\tilde{K}(x) = K(x) + \Delta K(x). \tag{3.10}$$

The kernel function $\tilde{K}(x)$ defines the convolution operator

$$(\tilde{\mathcal{F}}_c v)(s) = \int_{-1}^{1} \tilde{K}(s-t)\, v(t)\, \mathrm{d}t. \tag{3.11}$$

The matrix representation of $\tilde{\mathcal{F}}_c$ with respect to the normalized Legendre basis is infinite dimensional, with only the upper skew-triangular part of the top-left $(m^K + 1) \times (m^K + 1)$ block being nonzero. These entries can be constructed from the $m^K + 1$ coefficients of $\tilde{K}(x)$ using a four-term recurrence [22]. Adding the infinite-dimensional scalar matrix representing $z\mathcal{I}$ yields an infinite-dimensional arrow-shaped representation of $z\mathcal{I} - \tilde{\mathcal{F}}_c$.

3.1.4. Volterra operator. When $\mathcal{L}$ is the Volterra operator $\mathcal{V}$ defined by (2.6), we follow the same strategy as for the Fredholm operator by representing (or approximating) the kernel $K(s,t)$ using the low-rank representation (3.9), so that the resulting operator $\tilde{\mathcal{V}}$ acts as

$$(\tilde{\mathcal{V}} v)(s) = \int_{-1}^{s} \tilde{K}(s,t)\, v(t)\, \mathrm{d}t = \sum_{j=1}^{r} f_j(s) \int_{-1}^{s} g_j(t)\, v(t)\, \mathrm{d}t.$$

In the normalized Legendre basis, the infinite-dimensional matrix representation of $z\mathcal{I} - \tilde{\mathcal{V}}$ is banded and can be written as

$$M_{\tilde{\mathcal{V}}}(z) = zI_\infty - \sum_{j=1}^{r} M_{\frac{1}{2}}[f_j]\, J\, M_{\frac{1}{2}}[g_j],$$

where J denotes the indefinite integration matrix for normalized Legendre series [37, section 4.7] and $M_{\frac{1}{2}}[f_j]$ and $M_{\frac{1}{2}}[g_j]$ are the normalized Legendre multiplication matrices [25, section 3.1].

3.1.5. Volterra convolution operator. When $\mathcal{L}$ is the Volterra convolution integral operator $\mathcal{V}_c$ defined by (2.7), we follow the same strategy as in subsection 3.1.3 by representing (or approximating) the univariate kernel $K(x)$ with a finite-degree normalized Legendre series $\tilde{K}(x)$. This yields

$$(\tilde{\mathcal{V}}_c v)(s) = \int_{-1}^{s} \tilde{K}(s-t)\, v(t)\, \mathrm{d}t.$$

In the normalized Legendre basis, a banded infinite-dimensional matrix representation of $\tilde{\mathcal{V}}_c$ can be constructed using a three-term recurrence relation together with a scaled symmetry relation [14, 46]. Adding zI_∞ does not alter the sparsity structure.

4. Error analysis. In this section, we shall still use the notation $\lambda_1(\cdot)$, but now to denote the largest eigenvalue of $\mathcal{T}(z)$. Thus, we reserve its argument for the location $z \in \mathbb{C}$, writing $\lambda_1(z)$. In addition, we denote by $\lambda_1^{(k)}(z)$ the largest Ritz value obtained after k iterations of Algorithm 2.1, which serves as an approximation to

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$\lambda_1(z)$. Likewise, let $\sigma_1(z)$ and $\sigma_1^{(k)}(z)$ denote the corresponding resolvent norm and its k-step estimate, respectively, i.e.,

$$\sigma_1(z) = \sqrt{\lambda_1(z)}, \quad \sigma_1^{(k)}(z) = \sqrt{\lambda_1^{(k)}(z)}.$$

We begin the error analysis of Algorithm 2.1 with a result on the convergence and an error bound for the Lanczos iteration, under the assumption that $\mathcal{T}(z)$ is applied exactly, i.e., that (3.1a) and (3.1b) are solved without error. The following proposition is a straightforward generalization of the well-known result due to Saad [32].

PROPOSITION 4.1. *Let $\mathcal{L}$ be an operator on a Hilbert space $\mathcal{H}$, and let $z \in \mathbb{C}$ be such that the resolvent $\mathcal{R}(z)$ exists and* (2.1) *holds. Let $\lambda_1(z) = \|\mathcal{T}(z)\|$ be the largest eigenvalue corresponding to the eigenmode $\varphi_1(z)$, and let $\lambda_{\inf}(z) := \inf \sigma(\mathcal{T}(z))$ denote the infimum of the spectrum of $\mathcal{T}(z)$. Denote by $P_1(z)$ the orthogonal projection onto the eigenspace corresponding to $\lambda_1(z)$.*

(1) *For an initial function u_1 satisfying $P_1(z)u_1 \neq 0$, the first k Lanczos iterations produce an approximation $\lambda_1^{(k)}(z)$ whose deviation from $\lambda_1(z)$ is bounded by*

$$0 \le \lambda_1(z) - \lambda_1^{(k)}(z) \le (\lambda_1(z) - \lambda_{\inf}(z)) \frac{\tan^2 \theta(\varphi_1(z), u_1)}{T_{k-1}^2(\omega)}, \tag{4.1}$$

where $\theta(\varphi_1(z), u_1)$ denotes the angle between $\varphi_1(z)$ and u_1, T_m is the Chebyshev polynomial of the first kind[4] *of degree m, and*

$$\omega = 1 + 2\frac{\lambda_1(z) - \lambda_2(z)}{\lambda_2(z) - \lambda_{\inf}(z)}, \quad \lambda_2(z) = \sup\big(\sigma(\mathcal{T}(z)) \setminus \{\lambda_1(z)\}\big).$$

(2) *For a given tolerance $\delta_1(z) > 0$, if Algorithm* 2.1 *terminates with $\beta_{k+1}|y_{k1}^{(k)}| < \delta_1(z)$ and $\lambda_1^{(k)}(z)$ is well converged, i.e.,*

$$\lambda_1^{(k)}(z) > \lambda_2(z) + \delta_1(z), \tag{4.2}$$

then it follows that

$$0 \le \lambda_1(z) - \lambda_1^{(k)}(z) \le \delta_1(z). \tag{4.3}$$

The results given in Proposition 4.1 are based on the assumption that $\mathcal{T}(z)$ is applied exactly. The following proposition provides an error bound for Algorithm 2.1 when $\mathcal{T}(z)$ is applied inexactly in a controlled manner. Since the proof closely follows the analyses in [26, 27], we omit it for brevity.

PROPOSITION 4.2. *Suppose that Algorithm* 2.1 *is terminated with $\beta_{k+1}|y_{k1}^{(k)}| < \delta_1(z)$ for some k and that $\mathcal{T}(z)$ is applied inexactly with the error bounded by $\delta_2(z)$, i.e.,*

$$\|w_i - \mathcal{T}(z)u_i\| \le \delta_2(z), \quad i = 1, 2, \ldots, k. \tag{4.4}$$

If the largest Ritz value $\lambda_1^{(k)}(z)$ is well separated and converged, i.e.,

$$\lambda_1^{(k)}(z) - \lambda_2^{(k)}(z) \ge \sqrt{2}\, k^{5/2} \delta_2(z), \quad \lambda_1^{(k)}(z) > \lambda_2(z) + \frac{5}{2}\big(\delta_1(z) + \sqrt{k}\, \delta_2(z)\big), \tag{4.5}$$

then

$$|\lambda_1(z) - \lambda_1^{(k)}(z)| \le \frac{5}{2}\big(\delta_1(z) + \sqrt{k}\, \delta_2(z)\big). \tag{4.6}$$

[4]Since $\omega > 1$, the Chebyshev polynomial T_{k-1} in (4.1) is evaluated outside the canonical domain $[-1, 1]$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

4.1. Error in the application of $\mathcal{T}(z)$. In this subsection, we bound the error in the solutions of (3.1), when (3.1) are solved as described in section 3. Our analysis first assumes that the computation is carried out in exact arithmetic. Instead of (3.1), we actually solve

$$(z\mathcal{I} - \mathcal{L} - \Delta\mathcal{L})\tilde{v} = u + \Delta u, \tag{4.7a}$$

$$(z^*\mathcal{I} - \mathcal{L}^* - \Delta\mathcal{L}^*)\tilde{w} = \tilde{v} + \Delta\tilde{v}, \tag{4.7b}$$

where $\Delta\mathcal{L}$ and $\Delta\mathcal{L}^*$ are perturbations, if applicable, caused by using the approximate infinite-dimensional banded matrix representation of $z\mathcal{I} - \mathcal{L}$ and $z^*\mathcal{I} - \mathcal{L}^*$. The terms Δu and $\Delta\tilde{v}$ are the residual errors on the right-hand side caused by controlled termination of adaptive QR.

The following lemma bounds $\Delta\mathcal{Q}$, $\Delta\mathcal{F}$, $\Delta\mathcal{V}$, $\Delta\mathcal{F}_c$, and $\Delta\mathcal{V}_c$, which are the concrete instances of the abstract perturbation $\Delta\mathcal{L}$, for the operators considered in section 2 in terms of the errors in the finite series of orthogonal polynomials used to represent or approximate the variable coefficients and kernel functions.

LEMMA 4.3. (1) *Suppose that, for the differential operators defined in subsection* 2.1*, the coefficients $a_j(x) \in L^\infty([-1,1])$ for $j = 0, 1, \ldots, N-1$. Let $C^j_{\mathrm{reg}}(z) > 0$ be constants such that*

$$\left\| \frac{\mathrm{d}^j w}{\mathrm{d}x^j} \right\| \le C^j_{\mathrm{reg}}(z)\big(\|w\| + \|(z^*\mathcal{I} - \mathcal{Q}^*)w\|\big) \tag{4.8}$$

for all $w \in \mathcal{D}(\mathcal{Q}^*)$. Then the norm of $\mathcal{R}(z)\Delta\mathcal{Q} : \mathcal{D}(\mathcal{Q}) \to L^2([-1,1])$ is bounded by

$$\|\mathcal{R}(z)\Delta\mathcal{Q}\| \le \delta_{\mathcal{Q}}(1 + \sigma_1(z)), \qquad \delta_{\mathcal{Q}} = \sum_{j=0}^{N-1} C^j_{\mathrm{reg}}(z)\|\Delta a_j(x)\|_{L^\infty([-1,1])}. \tag{4.9}$$

(2) *For the general Fredholm and Volterra integral operators defined in subsections* 2.2 and 2.3*, the norms of $\Delta\mathcal{F}$ and $\Delta\mathcal{V}$ satisfy*

$$\|\Delta\mathcal{F}\| \le \delta_{\mathcal{F}}, \qquad \|\Delta\mathcal{V}\| \le \delta_{\mathcal{V}}, \qquad \delta_{\mathcal{F}} = \delta_{\mathcal{V}} = \|\Delta K(s,t)\|_{L^2([-1,1]^2)}.$$

(3) *For the Fredholm and Volterra convolution operators defined in subsections* 2.2 and 2.3*, the norms of $\Delta\mathcal{F}_c$ and $\Delta\mathcal{V}_c$ satisfy*

$$\|\Delta\mathcal{F}_c\| \le \delta_{\mathcal{F}_c}, \qquad \delta_{\mathcal{F}_c} = \sqrt{2}\,\|\Delta K(x)\|_{L^2([-2,2])}, \tag{4.10a}$$

$$\|\Delta\mathcal{V}_c\| \le \delta_{\mathcal{V}_c}, \qquad \delta_{\mathcal{V}_c} = \sqrt{2}\,\|\Delta K(x)\|_{L^2([0,2])}. \tag{4.10b}$$

Proof. (1) For differential operators, we first bound $(\Delta\mathcal{Q})^*\mathcal{R}^*(z)$. The condition (4.8) can be found in Lemma 1.4 of [12, section XVII.1], where the graph norm associated with the operator $z^*\mathcal{I} - \mathcal{Q}^*$ is defined by

$$\|w\|_{z^*\mathcal{I}-\mathcal{Q}^*} = \|w\| + \|(z^*\mathcal{I} - \mathcal{Q}^*)w\|, \qquad w \in \mathcal{D}(\mathcal{Q}^*).$$

For any $v \in L^2([-1,1])$, the graph norm of $w = \mathcal{R}^*(z)v$ satisfies

$$\|w\|_{z^*\mathcal{I}-\mathcal{Q}^*} \le (1 + \sigma_1(z))\|v\|. \tag{4.11}$$

It follows from (4.8) that

$$\begin{aligned} \|(\Delta\mathcal{Q})^*w\| &\le \sum_{j=0}^{N-1} \|\Delta a_j(x)\|_{L^\infty([-1,1])} \left\| \frac{\mathrm{d}^j w}{\mathrm{d}x^j} \right\| \\ &\le \sum_{j=0}^{N-1} C^j_{\mathrm{reg}}(z)\|\Delta a_j(x)\|_{L^\infty([-1,1])}\|w\|_{z^*\mathcal{I}-\mathcal{Q}^*}. \end{aligned} \tag{4.12}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Combining (4.11) and (4.12) yields a bound for $\|(\Delta\mathcal{Q})^*\mathcal{R}^*(z)\|$. Since $\|\mathcal{R}(z)\Delta\mathcal{Q}\| \le \|(\Delta\mathcal{Q})^*\mathcal{R}^*(z)\|$, the estimate (4.9) follows.

(2) For Fredholm and Volterra integral operators, the operators $\Delta\mathcal{F}$ and $\Delta\mathcal{V}$ are bounded by $\|\Delta K(s,t)\|_{L^2([-1,1]^2)}$, since

$$(\Delta\mathcal{F}u)(s) = \int_{-1}^{1} \Delta K(s,t)\, u(t)\,\mathrm{d}t, \quad (\Delta\mathcal{V}u)(s) = \int_{-1}^{s} \Delta K(s,t)\, u(t)\,\mathrm{d}t, \quad -1 \le s \le 1.$$

(3) For Fredholm convolution operators, the change of variables $x = s - t \in [-2,2]$ yields

$$\|\Delta K(s,t)\|^2_{L^2([-1,1]^2)} = \int_{-2}^{2} |\Delta K(x)|^2 (2 - |x|)\,\mathrm{d}x \le 2\int_{-2}^{2} |\Delta K(x)|^2\,\mathrm{d}x,$$

which implies (4.10a). For Volterra convolution operators, (4.10b) follows from an analogous derivation with the change of variables $x = s - t \in [0,2]$. ∎

The purpose of Lemma 4.3 goes far beyond the specific result it establishes. Rather, it serves to motivate a more general assumption—for linear operators $\mathcal{L}$ whose resolvent is compact or compact-plus-scalar, the perturbation $\Delta\mathcal{L}$ can be expected to be bounded in a similar manner. This insight underpins a general framework for the error analysis of Algorithm 2.1. Accordingly, we make the general assumption

$$\|\mathcal{R}(z)\Delta\mathcal{L}\| \le \delta_{\mathcal{L}}(z)\sigma_1(z), \quad \|\mathcal{R}^*(z)\Delta\mathcal{L}^*\| \le \delta_{\mathcal{L}}(z)\sigma_1(z). \tag{4.13}$$

We now turn to bounds on Δu and $\Delta\tilde{v}$. For $x \in \Phi_n := \operatorname{span}\{\phi_i\}_{i=1}^n$ and $y \in \Psi_n := \operatorname{span}\{\psi_i\}_{i=1}^n$, we denote by $\mathbf{x}$ and $\mathbf{y}$ the corresponding coefficient vectors of x and y with respect to their respective bases. The norms of $\mathbf{x}$ and $\mathbf{y}$ are bounded above and below by those of x and y as

$$\alpha_n^{\phi}\|x\| \le \|\mathbf{x}\| \le \beta_n^{\phi}\|x\| \quad \forall x \in \Phi_n, \tag{4.14a}$$

$$\alpha_n^{\psi}\|y\| \le \|\mathbf{y}\| \le \beta_n^{\psi}\|y\| \quad \forall y \in \Psi_n. \tag{4.14b}$$

Here, the parameters α_n^{ϕ}, β_n^{ϕ}, α_n^{ψ}, and β_n^{ψ} depend on the chosen bases. We further define the condition numbers of these bases by

$$\kappa_n^{\phi} = \frac{\beta_n^{\phi}}{\alpha_n^{\phi}}, \quad \kappa_n^{\psi} = \frac{\beta_n^{\psi}}{\alpha_n^{\psi}}.$$

For differential operators, since ϕ_i is a combination of normalized Legendre polynomials that satisfies boundary conditions, the condition number κ_n^{ϕ} also depends on the boundary conditions. For example, $\kappa_n^{\phi} = O(n)$ for a Dirichlet boundary condition and $\kappa_n^{\phi} = O(n^2)$ for clamped boundary conditions [24, section 4.1.5]. If the differential operator of interest is an Nth-order one, $\{\psi_i\}_{i=1}^{\infty}$ is the ultraspherical basis $\{C_{i-1}^{(N+1/2)}(x)\}_{i=1}^{\infty}$ and the corresponding condition number κ_n^{ψ} is then $O(n^N)$. Of course, these estimates are rather pessimistic in that the effective condition numbers in practice are much smaller than these asymptotic results. For integral operators, $\kappa_n^{\phi} = \kappa_n^{\psi} = 1$ because $\{\phi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ are simply the normalized Legendre basis.

Suppose that the adaptive QR process terminates with tolerance $\delta_{\mathbf{r}}$. Then the norms of Δu and $\Delta\tilde{v}$ in (4.7) can be bounded using (4.14b), i.e.,

$$\|\Delta u\| \le \delta_{\mathbf{r}} \frac{\beta^{\psi}_{n_u}}{\alpha^{\psi}_{\tilde{n}_v+b}}\|u\|, \quad \|\Delta\tilde{v}\| \le \delta_{\mathbf{r}} \frac{\beta^{\psi}_{\tilde{n}_v}}{\alpha^{\psi}_{\tilde{n}_w+b}}\|\tilde{v}\|,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where n_u, $\tilde{n}_v$, and $\tilde{n}_w$ denote the truncation degree of u, $\tilde{v}$, and $\tilde{w}$, respectively. Since $n_u \le \tilde{n}_v \le \tilde{n}_w$, we define $\tilde{n} := \tilde{n}_w + b$ to be the maximum truncation degree of the series encountered in the course of applying adaptive QR to solve (4.7). We further obtain

$$\frac{\|\Delta u\|}{\|u\|} \le \delta_{\mathbf{r}} \kappa^{\psi}_{\tilde{n}}, \qquad \frac{\|\Delta \tilde{v}\|}{\|\tilde{v}\|} \le \delta_{\mathbf{r}} \kappa^{\psi}_{\tilde{n}}. \tag{4.15}$$

With (4.13) and (4.15), we are ready to bound $\tilde{w} - w$.

LEMMA 4.4. *Assume that* (4.13) *and* (4.15) *hold and that* $\delta_{\mathcal{L}}(z)\sigma_1(z) < 1/2$. *Then*

$$\|\tilde{w} - w\| \le 6\left(\delta_{\mathcal{L}}(z)\sigma_1(z) + \delta_{\mathbf{r}}\kappa^{\psi}_{\tilde{n}}\right)\lambda_1(z)\|u\|, \tag{4.16}$$

where higher-order terms in $\delta_{\mathcal{L}}(z)$ *and* $\delta_{\mathbf{r}}$ *are neglected.*

Proof. Subtracting (3.1a) from (4.7a) yields

$$(z\mathcal{I} - \mathcal{L})(\mathcal{I} - \mathcal{R}(z)\Delta\mathcal{L})(\tilde{v} - v) = \Delta\mathcal{L}v + \Delta u. \tag{4.17}$$

Since $\|\mathcal{R}(z)\Delta\mathcal{L}\| < 1/2$, the operator $(\mathcal{I} - \mathcal{R}(z)\Delta\mathcal{L})$ is invertible and satisfies $\|(\mathcal{I} - \mathcal{R}(z)\Delta\mathcal{L})^{-1}\| \le 2$. Thus, (4.17) can be rewritten as

$$\tilde{v} - v = (\mathcal{I} - \mathcal{R}(z)\Delta\mathcal{L})^{-1}\mathcal{R}(z)(\Delta\mathcal{L}v + \Delta u).$$

The bounds (4.13) and (4.15), together with $\|v\| \le \sigma_1(z)\|u\|$, imply that

$$\|\tilde{v} - v\| \le 2\big(\delta_{\mathcal{L}}(z)\sigma_1(z) + \delta_{\mathbf{r}}\kappa^{\psi}_{\tilde{n}}\big)\sigma_1(z)\|u\|. \tag{4.18}$$

Subtracting (3.1b) from (4.7b) gives

$$(z^*\mathcal{I} - \mathcal{L}^*)(\mathcal{I} - \mathcal{R}^*(z)\Delta\mathcal{L}^*)(\tilde{w} - w) = \Delta\tilde{v} + \Delta\mathcal{L}^* w + \tilde{v} - v. \tag{4.19}$$

Analogously, we have $\|(\mathcal{I} - \mathcal{R}^*(z)\Delta\mathcal{L}^*)^{-1}\| \le 2$, and hence (4.19) can be recast as

$$\tilde{w} - w = (\mathcal{I} - \mathcal{R}^*(z)\Delta\mathcal{L}^*)^{-1}\big(\mathcal{R}^*(z)\Delta\mathcal{L}^* w + \mathcal{R}^*(z)\Delta\tilde{v} + \mathcal{R}^*(z)(\tilde{v} - v)\big). \tag{4.20}$$

Finally, (4.16) follows from (4.13), (4.15), and (4.18), the bound $\|w\| \le \lambda_1(z)\|u\|$, and neglecting higher-order terms. ∎

In practice, it is impossible to solve (4.7) exactly in floating-point arithmetic. Let $\hat{v}$ and $\hat{w}$ denote the numerically computed solutions of (4.7a) and (4.7b), respectively. The forward errors in these computed solutions can be significant, if not enormous, since moderate backward errors are magnified by the ill-conditioning of (4.7). Accordingly, we write the equations that are actually solved by $\hat{v}$ and $\hat{w}$ as

$$(z\mathcal{I} - \mathcal{L} - \Delta^{\mathrm{fp}}\mathcal{L})\hat{v} = u + \Delta^{\mathrm{fp}}u, \tag{4.21a}$$

$$(z^*\mathcal{I} - \mathcal{L}^* - \Delta^{\mathrm{fp}}\mathcal{L}^*)\hat{w} = \hat{v} + \Delta^{\mathrm{fp}}\hat{v}, \tag{4.21b}$$

where $\Delta^{\mathrm{fp}}\mathcal{L} = \Delta\mathcal{L} + \mathcal{E}$ and $\Delta^{\mathrm{fp}}\mathcal{L}^* = \Delta\mathcal{L}^* + \mathcal{E}^*$. Here, $\mathcal{E}$ and $\mathcal{E}^*$ denote the backward errors caused by floating-point rounding, while $\Delta^{\mathrm{fp}}u$ and $\Delta^{\mathrm{fp}}\hat{v}$ incorporate perturbations arising from both the preset tolerance $\delta_{\mathbf{r}}$ and the backward errors on the right-hand sides. Notably, $(\mathcal{E})^* \ne \mathcal{E}$.

The following lemma collects bounds analogous to (4.13) and (4.15), with the effects of rounding errors taken into account.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 4.5. *Suppose that* (4.7a) *and* (4.7b) *are solved by adaptive QR in floating-point arithmetic with tolerance* $\delta_{\mathbf{r}}$. *Let* $\hat{n}$ *be the maximum truncation degree of the series involved in the course of applying adaptive QR to solve* (4.21). *If* (4.13) *holds and* $\delta_{\mathcal{L}}(z)\sigma_1(z) < 1/2$, *then*

$$\|\mathcal{R}(z)\Delta^{\mathrm{fp}}\mathcal{L}\| \le \delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z)\sigma_1(z), \qquad \|\mathcal{R}^*(z)\Delta^{\mathrm{fp}}\mathcal{L}^*\| \le \delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z)\sigma_1(z), \tag{4.22}$$

where $\delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z) = \delta_{\mathcal{L}}(z) + \delta^{\mathrm{fp}}_{\mathcal{E},\hat{n}}(z)$ *and*

$$\delta^{\mathrm{fp}}_{\mathcal{E},\hat{n}}(z) = \kappa^{\phi}_{\hat{n}}\kappa^{\psi}_{\hat{n}}\gamma_{b\hat{n}^{3/2}}\|(z\mathcal{I}-\mathcal{L})|_{\Phi_{\hat{n}}}\|. \tag{4.23}$$

Moreover,

$$\frac{\|\Delta^{\mathrm{fp}}u\|}{\|u\|} \le \delta^{\mathrm{fp}}_{\mathbf{r},\hat{n}}\kappa^{\psi}_{\hat{n}}, \qquad \frac{\|\Delta^{\mathrm{fp}}\hat{v}\|}{\|\hat{v}\|} \le \delta^{\mathrm{fp}}_{\mathbf{r},\hat{n}}\kappa^{\psi}_{\hat{n}}, \tag{4.24}$$

where $\delta^{\mathrm{fp}}_{\mathbf{r},\hat{n}} = \delta_{\mathbf{r}} + \gamma_{b\hat{n}}$ *and we ignore the higher-order terms in* $\delta_{\mathbf{r}}$ *and* $\gamma_{b\hat{n}}$.

Proof. With the same notation as in (3.3)–(3.5), except that n is replaced by $\hat{n}_v$, it follows from [17, Theorem 19.5] that

$$(A_{11}+E_{11})\hat{\mathbf{v}}_{1:\hat{n}_v} = Q_{\hat{n}_v}\begin{pmatrix}\hat{\mathbf{r}}_{1:\hat{n}_v}\\ 0\end{pmatrix} = Q_{\hat{n}_v}\begin{pmatrix}\mathbf{r}_{1:\hat{n}_v}+\Delta^{\mathrm{fp}}\mathbf{r}_{1:\hat{n}_v}\\ 0\end{pmatrix}, \tag{4.25}$$

where $\|E_{11}\| \le \gamma_{b\hat{n}_v^{3/2}}\|A_{11}\|$ and $\|\mathbf{r}-\hat{\mathbf{r}}\| \le \gamma_{b\hat{n}_v}\|\mathbf{r}\|$. Embedding $E_{11} \in \mathbb{C}^{(\hat{n}_v+b)\times\hat{n}_v}$ into an infinite-dimensional matrix yields the representation of $\mathcal{E}$ under the bases $\{\phi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$. Using (4.14), we obtain

$$\|\mathcal{E}\| \le \frac{\beta^{\phi}_{\hat{n}_v}}{\alpha^{\psi}_{\hat{n}_v+b}}\gamma_{b\hat{n}_v^{3/2}}\|A_{11}\|, \qquad \|(z\mathcal{I}-\tilde{\mathcal{L}})|_{\Phi_{\hat{n}_v}}\| \ge \frac{\alpha^{\phi}_{\hat{n}_v}}{\beta^{\psi}_{\hat{n}_v+b}}\|A_{11}\|.$$

Combining these inequalities and using $\hat{n}_v + b \le \hat{n}$ gives

$$\|\mathcal{E}\| \le \kappa^{\phi}_{\hat{n}}\kappa^{\psi}_{\hat{n}}\gamma_{b\hat{n}^{3/2}}\|(z\mathcal{I}-\tilde{\mathcal{L}})|_{\Phi_{\hat{n}}}\|. \tag{4.26}$$

Since $\|\mathcal{R}(z)\Delta\mathcal{L}\| < 1/2$, we have

$$\|(z\mathcal{I}-\tilde{\mathcal{L}})|_{\Phi_{\hat{n}}}\| \le \|(z\mathcal{I}-\mathcal{L})|_{\Phi_{\hat{n}}}\|\|\mathcal{I}+\mathcal{R}(z)\Delta\mathcal{L}\| \le \frac{3}{2}\|(z\mathcal{I}-\mathcal{L})|_{\Phi_{\hat{n}}}\|. \tag{4.27}$$

Using (4.26) and (4.27) gives the first inequality in (4.22), with the factor $3/2$ absorbed into $\gamma_{b\hat{n}^{3/2}}$. Prolonging the right-hand side of (4.25) to an infinite vector with zeros and comparing it against (3.4) (with n replaced by $\hat{n}_v$) yields

$$\Delta^{\mathrm{fp}}\mathbf{u} = \begin{pmatrix}Q^*_{\hat{n}_v} & \\ & I_{\infty}\end{pmatrix}\begin{pmatrix}\Delta^{\mathrm{fp}}\mathbf{r}_{1:\hat{n}_v}\\ -\mathbf{r}_{\hat{n}_v+1:\hat{n}_v+b}\\ 0\end{pmatrix}. \tag{4.28}$$

Since $Q_{\hat{n}_v}$ is unitary, the stopping criterion $\|\hat{\mathbf{r}}_{\hat{n}_v+1:\hat{n}_v+b}\| \le \delta_{\mathbf{r}}\|\hat{\mathbf{r}}\|$ implies

$$\|\Delta^{\mathrm{fp}}\mathbf{u}\| \le \|\Delta^{\mathrm{fp}}\mathbf{r}_{1:\hat{n}_v}\| + \|\Delta^{\mathrm{fp}}\mathbf{r}_{\hat{n}_v+1:\hat{n}_v+b}\| + \|\hat{\mathbf{r}}_{\hat{n}_v+1:\hat{n}_v+b}\| \le (\gamma_{b\hat{n}_v}+\delta_{\mathbf{r}})\|\mathbf{r}\|,$$

where higher-order terms are neglected. Thus, (4.14b) yields the first inequality in (4.24). The bounds for $\|\mathcal{E}^*\|$ and $\|\Delta^{\mathrm{fp}}\hat{v}\|$ follow analogously, giving the second inequalities in (4.22) and (4.24). ∎

We are now in a position to bound the error in $\hat{w}$. The following corollary follows readily, and we omit its proof.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

COROLLARY 4.6. *If* (4.22) *and* (4.24) *hold and* $\delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z)\sigma_1(z) < 1/2$, *then*

$$\|\hat{w} - w\| \le 6\left(\delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z)\sigma_1(z) + \delta^{\mathrm{fp}}_{\mathbf{r},\hat{n}}\kappa^{\psi}_{\hat{n}}\right)\lambda_1(z)\|u\|.$$

Since the rounding errors introduced and amplified by the application of $\mathcal{T}(z)$ are significantly larger than those arising in other parts of Algorithm 2.1, we do not further quantify the latter and proceed to bound the errors in the computed resolvent norm and pseudospectra.

4.2. Error bound for the computed pseudospectra. With the error bounds for $\hat{w}$ established, we now assess the accuracy of the computed resolvent norm for a given z. In the following lemma, we recycle the notations $\tilde{n}$ and $\hat{n}$ to denote, respectively, the maximum truncation degree of the series used throughout the k Lanczos iterations in exact and floating-point arithmetic.

THEOREM 4.7. *For* $\mathcal{L}:\mathcal{D}(\mathcal{L})\to\mathcal{H}$, *let* $z\in\mathbb{C}$ *be such that the resolvent* $\mathcal{R}(z)$ *exists and* (2.1) *holds. Let* $\tilde{\mathcal{L}} = \mathcal{L} + \Delta\mathcal{L}$ *be a perturbed operator whose matrix representations of* $z\mathcal{I} - \tilde{\mathcal{L}}$ *and* $z^*\mathcal{I} - \tilde{\mathcal{L}}^*$ *under the bases* $\{\phi_i\}_{i=1}^{\infty}$ *and* $\{\psi_i\}_{i=1}^{\infty}$ *are infinite-dimensional banded matrices with bandwidth* b. *Assume that the perturbation* $\Delta\mathcal{L}$ *satisfies* (4.13) *and that the bases* $\{\phi_i\}_{i=1}^{\infty}$ *and* $\{\psi_i\}_{i=1}^{\infty}$ *satisfy* (4.14).
(1) *Suppose that Algorithm* 2.1 *is implemented in exact arithmetic with* $\mathcal{T}(z)$ *applied exactly. If the Lanczos iteration terminates at the* k*th step with tolerance* $\delta_1(z)$ *and returns* $\sigma_1^{(k)}(z)$ *satisfying* (4.2), *then*

$$\frac{|\sigma_1^{(k)}(z) - \sigma_1(z)|}{\sigma_1(z)} \le \frac{\delta_1(z)}{\lambda_1(z)}. \tag{4.29}$$

(2) *Suppose that Algorithm* 2.1 *is implemented in exact arithmetic, with* $\mathcal{T}(z)$ *applied by solving the infinite-dimensional banded systems corresponding to* (4.7) *via adaptive QR with relative tolerance* $\delta_{\mathbf{r}}$. *Assume that the Lanczos iteration terminates at the* k*th step with tolerance* $\delta_1(z)$ *and returns* $\tilde{\sigma}_1^{(k)}(z)$. *Let*

$$\delta_{\mathcal{T},\tilde{n}}(z) = 6\left(\delta_{\mathcal{L}}(z)\sigma_1(z) + \delta_{\mathbf{r}}\kappa^{\psi}_{\tilde{n}}\right). \tag{4.30}$$

If $\delta_{\mathcal{L}}(z)\sigma_1(z) \le 1/2$ *and* (4.5) *holds with* $\delta_1(z)$ *and* $\delta_2(z) := \delta_{\mathcal{T},\tilde{n}}(z)\lambda_1(z)$, *then*

$$\frac{|\tilde{\sigma}_1^{(k)}(z) - \sigma_1(z)|}{\sigma_1(z)} \le \frac{5}{2}\left(\frac{\delta_1(z)}{\lambda_1(z)} + \sqrt{k}\,\delta_{\mathcal{T},\tilde{n}}(z)\right). \tag{4.31}$$

(3) *Let* $\hat{\sigma}_1^{(k)}(z)$ *denote the resolvent norm computed by Algorithm* 2.1 *under the same assumptions as in* (2) *but in floating-point arithmetic. Let*

$$\delta^{fp}_{\mathcal{T},\hat{n}}(z) = 6\left(\delta^{fp}_{\mathcal{L},\hat{n}}(z)\sigma_1(z) + \delta^{fp}_{\mathbf{r},\hat{n}}\kappa^{\psi}_{\hat{n}}\right). \tag{4.32}$$

If $\delta^{fp}_{\mathcal{L},\hat{n}}(z)\sigma_1(z) \le 1/2$ *and* (4.5) *holds with* $\delta_1(z)$ *and* $\delta_2(z) := \delta^{fp}_{\mathcal{T},\hat{n}}(z)\lambda_1(z)$, *then*

$$\frac{|\hat{\sigma}_1^{(k)}(z) - \sigma_1(z)|}{\sigma_1(z)} \le \frac{5}{2}\left(\frac{\delta_1(z)}{\lambda_1(z)} + \sqrt{k}\,\delta^{fp}_{\mathcal{T},\hat{n}}(z)\right). \tag{4.33}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. It follows from Proposition 4.1 that $0 \le (\sigma_1^{(k)}(z))^2 - (\sigma_1(z))^2 \le \delta_1(z)$. Part (1) then follows from the inequality

$$\frac{|\sigma_1^{(k)}(z) - \sigma_1(z)|}{\sigma_1(z)} \le \frac{|(\sigma_1^{(k)}(z))^2 - (\sigma_1(z))^2|}{\sigma_1(z)^2} \le \frac{\delta_1(z)}{\lambda_1(z)}. \tag{4.34}$$

Combining Proposition 4.2 with Lemma 4.4, and using (4.30), we obtain

$$\frac{|\tilde{\sigma}_1^{(k)}(z) - \sigma_1(z)|}{\sigma_1(z)} \le \frac{|(\tilde{\sigma}_1^{(k)}(z))^2 - (\sigma_1(z))^2|}{\sigma_1(z)^2} \le \frac{5}{2}\left(\frac{\delta_1(z)}{\lambda_1(z)} + \sqrt{k}\,\delta_{\mathcal{T},\tilde{n}}(z)\right), \tag{4.35}$$

which establishes part (2). Part (3) follows analogously from Proposition 4.2 and Corollary 4.6. ∎

Part (3) of the preceding theorem is sufficient to derive a bound for the backward error of the computed ε-pseudospectrum in terms of the Hausdorff distance. For a compact set $G \subset \mathbb{C}$ and $\varepsilon > 0$, we define

$$\sigma_\varepsilon(\mathcal{L}; G) := \left\{ z \in G \,\middle|\, \|(z\mathcal{I} - \mathcal{L})^{-1}\| > \varepsilon^{-1} \right\} \tag{4.36}$$

and

$$\hat{\sigma}_\varepsilon(\mathcal{L}; G) := \left\{ z \in G \,\middle|\, \hat{\sigma}_1^{(k)}(z) > \varepsilon^{-1} \right\} \tag{4.37}$$

to be the ε-pseudospectrum of $\mathcal{L}$ and its numerical approximation returned by Algorithm 2.1 after k Lanczos iterations, respectively, both restricted to G. We now reuse the notation $\hat{n}$ to denote the maximum truncation degree of the series employed throughout the set G in floating-point arithmetic and define

$$\delta_G := 5\sqrt{k} \sup_{z \in G} \delta^{\mathrm{fp}}_{\mathcal{T},\hat{n}}(z).$$

The corollary below bounds the Hausdorff distance between the true ε-pseudospectrum $\sigma_\varepsilon(\mathcal{L}; G)$ and its numerical approximation $\hat{\sigma}_\varepsilon(\mathcal{L}; G)$.

COROLLARY 4.8. *Suppose that for every $z \in G$ the assumptions of part* (3) *of Theorem* 4.7 *hold and the tolerance $\delta_1(z)$ is sufficiently small enough such that*

$$\delta_1(z) \le \sqrt{k}\,\delta^{fp}_{\mathcal{T},\hat{n}}(z)\lambda_1(z). \tag{4.38}$$

If $\delta_G < 1$, it then follows a backward-error enclosure, i.e.,

$$\sigma_{(1-\delta_G)\varepsilon}(\mathcal{L}; G) \subset \hat{\sigma}_\varepsilon(\mathcal{L}; G) \subset \sigma_{(1+\delta_G)\varepsilon}(\mathcal{L}; G).$$

Consequently, with $d_H(\cdot,\cdot)$ denoting the Hausdorff distance in $\mathbb{C}$, we have

$$d_H(\hat{\sigma}_\varepsilon(\mathcal{L}; G), \sigma_\varepsilon(\mathcal{L}; G)) \le d_H\left(\sigma_{(1+\delta_G)\varepsilon}(\mathcal{L}; G), \sigma_{(1-\delta_G)\varepsilon}(\mathcal{L}; G)\right). \tag{4.39}$$

4.3. Stopping criterion. Part (3) in Theorem 4.7 also suggests a practical stopping criterion for Algorithm 2.1.

Since $\delta_{\mathcal{L}}(z)$ measures the perturbation $\Delta\mathcal{L}$, it is usually a moderate multiple of ϵ_m entrywise. For the highest possible accuracy, the tolerance $\delta_{\mathbf{r}}$ of the adaptive QR process is also set to be ϵ_m. It follows from Lemma 4.5 that $\delta^{\mathrm{fp}}_{\mathcal{L},\hat{n}}(z)$ and $\delta^{\mathrm{fp}}_{\mathbf{r},\hat{n}}$ are also $\mathcal{O}(\epsilon_m)$. Thus, we may write $\sqrt{k}\delta^{\mathrm{fp}}_{\mathcal{T},\hat{n}}(z) \approx C_L\sigma_1(z)\epsilon_m$, where C_L is a constant independent of z. For a user-prescribed tolerance δ, part (3) of Theorem 4.7 implies

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

that, to ensure the relative error of the computed resolvent norm is $\mathcal{O}(\delta)$, we need to set $\delta_1(z)$ as

$$\delta_1(z) = \delta\lambda_1(z), \tag{4.40}$$

in case of $\delta \geq C_L\sigma_1(z)\epsilon_m$. However, if $\delta < C_L\sigma_1(z)\epsilon_m$, then (4.40) does not yield any further improvement in accuracy. We therefore adopt the stopping criterion

$$\beta_{k+1}|y_{k1}^{(k)}| < \delta_1(z) = \max\Big(C_L\epsilon_m\left(\lambda_1(z)\right)^{3/2},\ \delta\lambda_1(z)\Big).$$

Of course, $\lambda_1(z)$ is unknown in practice. We therefore replace $\lambda_1(z)$ by its approximation $\lambda_1^{(k)}(z)$, leading to the stopping criterion

$$\beta_{k+1}|y_{k1}^{(k)}| < \delta_1(z) = \max\left(C_L\epsilon_m\left(\lambda_1^{(k)}(z)\right)^{3/2},\ \delta\lambda_1^{(k)}(z)\right). \tag{4.41}$$

This criterion should be used at line 5 of Algorithm 2.1.

We set $C_L = 100$ for all experiments in section 5 and find that this value works uniformly well. It is not difficult to realize that (4.41) is analogous to the standard stopping criterion used in ARPACK [21, section 4.6] and the `eigs` function of MATLAB. In fact, (4.41) is virtually optimal in the sense that the best possible accuracy is achieved once δ is chosen smaller than $\sqrt{k}\,\delta^{\mathrm{fp}}_{\mathcal{T},\hat{n}}(z)$. See subsection 5.1 for an example demonstrating this.

When we are bailed out of the Lanczos process, the smallest possible relative error in $\lambda_1^{(k)}(z)$ is $\mathcal{O}(\sigma_1(z)\epsilon_m)$. Thus, we can always have an idea on roughly how many faithful digits we have at most in $\sqrt{\lambda_1^{(k)}(z)}$ by its magnitude when the resolvent norm $\sigma_1(z)$ is no greater than $\mathcal{O}(1/\epsilon_m)$, i.e., $\sigma_1(z)^{-1} \gtrsim \mathcal{O}(\epsilon_m)$. If $\sigma_1(z)^{-1}$ turns out to be about 10^{-10}, we then have roughly 3 or 4 digits to trust. If $\sigma_1(z)^{-1} \approx 10^{-20}$ is what we need, double precision is then off the table, and we have to look at quadruple or higher precisions. In such a case, we should repeat the entire computation with the extended precision, including the approximations to the variable coefficients and the kernel functions by normalized Legendre series. Usually, the quantities $m^{a_{N-1}}, \ldots, m^{a_0}$, m^K, m^f, m^g, etc., are expected to become larger, since the tail of the series deemed to be negligible emerges (much) later in the extended precision arithmetic.

5. Numerical experiments. To demonstrate the advantages of the proposed method, we test it on a wide range of problems, covering all the operators we discussed in section 2. The "exact" resolvent norm that we compare against in the remainder of this section is obtained using octuple precision[5] in Julia. In the rest of this paper, we refer to the degree of u_k at the termination of the Lanczos iteration (see Algorithm 2.1) as the degrees of freedom (DoF) at a given z.

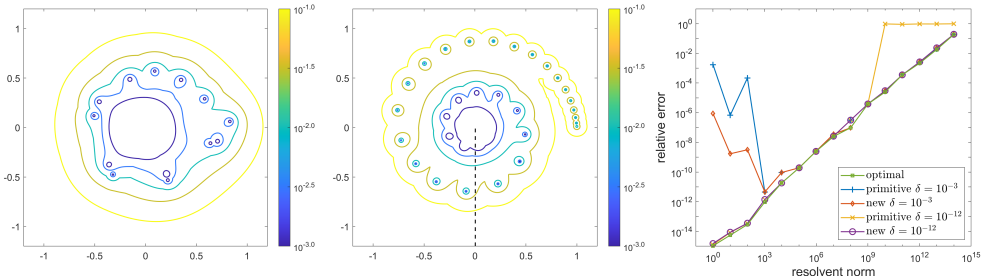
5.1. Lasers. Our first problems are the Huygens–Fresnel operators

$$(\mathcal{F}^L u)(s) = \sqrt{\frac{\mathrm{i}F}{\pi}}\int_{-1}^{1} e^{-\mathrm{i}FM(s/M-t)^2}u(t)\mathrm{d}t, \quad (\mathcal{F}_c^L u)(s) = \sqrt{\frac{\mathrm{i}F}{\pi}}\int_{-1}^{1} e^{-\mathrm{i}F(s-t)^2}u(t)\mathrm{d}t$$

from modeling the laser problem [44, section 60]. These integral operators acting on $u \in L^2([-1,1])$ correspond to unstable and stable resonators, respectively,

[5]The 256-bit octuple precision is the default format of Julia's `BigFloat` type of floating point number. `BigFloat`, based on the GNU MPFR library, is the arbitrary precision floating point number type in Julia.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) ε-pseudospectra of $\mathcal{F}^L$ (b) ε-pseudospectra of $\mathcal{F}_c^L$ (c) relative error

FIG. 1. *Huygens–Fresnel operators. (Color for all figures is available online only.)*

with the former being a general Fredholm operator and the latter a Fredholm operator of convolution type. In Figures 1(a) and 1(b), we reproduce by Algorithm 2.1 the second panels of Figures 60.6 and 60.2 from [44, section 60] for Fresnel number $F = 16\pi$ and magnification $M = 2$. The contours at $\varepsilon = 10^{-1}, 10^{-1.5}, \ldots, 10^{-3}$ are a spot-on-match with the ones that Trefethen and Embree obtained using the traditional method.

Figure 1(c) shows the resolvent norm versus its relative error obtained with Algorithm 2.1 at the 15 intersection points of the vertical dashed line $x = 0$ in Figure 1(b) and the boundaries of ε-pseudospectra for $\varepsilon = 10^0, 10^{-1}, 10^{-2}, \ldots, 10^{-14}$ (not shown). The purple and red curves correspond to $\delta = 10^{-12}$ and $\delta = 10^{-3}$, respectively. We also search for the smallest relative error that can be achieved at each of these intersection points and have it plotted (green) for comparison. The green squares lie roughly on a straight line of slope 1, confirming the analysis at the end of section 4 that the best relative error of $\lambda_1^{(k)}(z)$ is $\mathcal{O}(\sigma_1(z)\epsilon_m)$. It also can be seen that the deviation of the purple curve from the green one is tiny, suggesting that our adaptive stopping criterion is nearly optimal when δ is set to a sufficiently small value. The first 5 diamonds on the red curve show early exits from the Lanczos process due to the low target accuracies, as expected. For comparison, we also include two curves obtained by replacing the adaptive stopping criterion (4.41) by the primitive one in line 6 of Algorithm 1.1 for $\delta = 10^{-3}$ (blue) and $\delta = 10^{-12}$ (yellow), respectively. The blue curve behaves similarly like the red one—an early bail-out to save computation for a loose δ. However, a δ that is excessively stringent like 10^{-12} results in nonconvergence for z with a large resolvent norm. This clearly shows that why the adaptive stopping criterion is indispensable for the success of Algorithm 2.1.

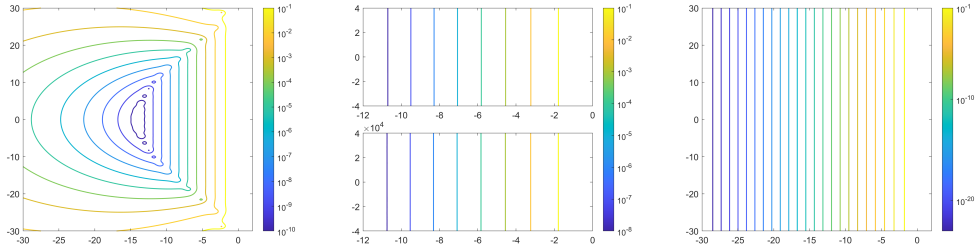
5.2. First derivative operator. Our second example is the first derivative operator

$$\mathcal{Q}^F u = \frac{\mathrm{d}u}{\mathrm{d}x},$$

acting on $L^2([0,2])$ and subject to $u(2) = 0$ [30, 41, 44]. This simple yet interesting operator features the nonexistence of eigenvalues and eigenmodes, and its resolvent norm depends only on $\Re(z)$, implying that the pseudospectral contours are straight vertical lines. The pseudospectral plot obtained by the traditional method is shown in Figure 2(a). The discretize-then-solve paradigm is at least partially failed for this example, if not a fiasco—only the contours in a triangular region to the left of the origin resemble the true pseudospectra as vertical lines, and the rest of the plot is an exhibition of spectral pollution.

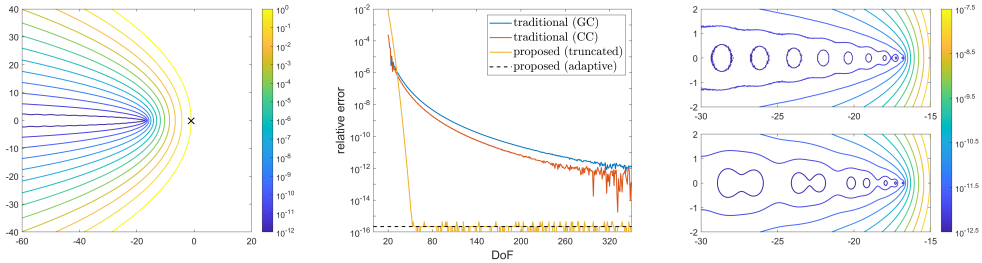
Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy



(a) traditional method (b) proposed method (c) extending to the left

FIG. 2. *ε-pseudospectra of the first derivative operator.*



(a) ε-pseudospectra (b) spectral convergence (c) close-up

FIG. 3. *Advection-diffusion operator.*

Whereas Figure 5.2 in [44, section 5] and the upper panel of Figure 3 in [41] are obtained from cropping a plot similar to Figure 2(a), the upper panel of Figure 2(b) reproduces the same plot directly using the proposed method. In fact, we can go much beyond the y-scale in Figure 2(b)—the lower panel of Figure 2(b) also displays the same level sets, obtained by Algorithm 2.1, but with the y-axis limit 10,000 times greater.

In the upper panel of Figure 2(b), the left limit is set to $\Re(z) = -12$. We actually can push it a little further to $\Re(z) = -16.2$ (not shown), where the resolvent norm there is about $10^{12.5}$, the largest that we can compute with double precision arithmetic so that the two- or three-digit accuracy still allows the lines generated by the contour plotter to be straight to eyes. Since $\|(z\mathcal{I} - \mathcal{Q}^F)^{-1}\| = e^{2|\Re(z)|}/(2|\Re(z)|) + \mathcal{O}(1/|\Re(z)|)$ for $\Re(z) < 0$, the resolvent norm grows virtually exponentially toward $-\infty$. As per the discussion of subsection 4.3, if we intend to calculate the resolvent norm to the left of the vertical line corresponding to $\varepsilon = 10^{-12.5}$, we have to resort to higher precision arithmetic. Figure 2(c) shows the pseudospectra up to $x = -30$ and is meant to be compared with Figure 2(c); the calculation is done in quadruple precision with $\epsilon_m \approx 1.93 \times 10^{-34}$.

5.3. Advection-diffusion operator. The advection-diffusion operator

$$\mathcal{Q}^{AD} u = \eta \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + \frac{\mathrm{d}u}{\mathrm{d}x}$$

acting on $L^2([0,1])$ and subject to the homogeneous Dirichlet boundary conditions serves in [41] and [44, section 12] as an example for transient effects caused by non-normality. The pseudospectra of $\mathcal{Q}^{AD}$ for $\eta = 0.015$ are shown as Figure 12.4 in [44, section 12], whereas Figure 3(a) is a reproduction by Algorithm 2.1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The advection-diffusion operator is also employed to demonstrate the quadrature issue related to the evaluation of the function norms [44, section 43]. Trefethen and Embree suggest using a weight matrix corresponding to either Gauss–Chebyshev (GC) or Clenshaw–Curtis (CC) quadrature when the Chebyshev pseudospectral method is used in a nonperiodic setting. However, they caution the loss of spectral convergence and a degeneration to algebraic convergence caused by the quadrature difficulty. This is confirmed by the blue and red curves in Figure 3(b) which are the error-DoF plot of the resolvent norm at $-1.05 - 0.10\mathrm{i}$ (cross in Figure 3(a)). These curves are obtained using the traditional method with the GC and CC quadratures and very much resemble those given in Figure 43.5 of [44, section 43] for the computed numerical abscissa. To the contrary, spectral convergence is gained by the proposed method using normalized Legendre polynomials. This can be seen from the yellow curve which is obtained by solving (3.1) with a fixed n throughout the Lanczos iterations and plotting the corresponding error for various n—the error decays exponentially to machine precision at about $n = 50$ and then levels off, signaling adequate resolution. Of course, we do not have to choose n manually in practice. The adaptivity in DoF (section 3) helps us choose an optimal n in each solution of (3.1) for complete resolutions. The resolvent norm at $-1.05 - 0.10\mathrm{i}$ returned by our fully automated algorithm has an error no greater than 2.19×10^{-16}, which is indicated by a dashed line in Figure 3(b) for comparison.

Figure 3(c) contains close-ups of Figure 3(a) obtained using the proposed (upper) and the traditional (lower) methods, respectively. Whereas the innermost contours in the upper panel barely hold their oval shape and start to wiggle, some of them coalesce in the lower panel, which is totally erroneous. This is because the collocation-based matrix approximation to $\mathcal{Q}^{AD}$ is ill-conditioned. It worsens the conditioning of the original problem and puts the calculation at even higher risk. Thanks to the well-conditioned ultraspherical spectral methods, the precision is made the most of with only minimal loss of digits in solving (3.1).

5.4. GO and Wiener–Hopf operators. Figure 4(a) shows the ε-pseudospectra of the GO operator [13]

$$(\mathcal{V}^{GO}u)(s) = \int_0^s e^{-10\left(s-\frac{1}{3}\right)^2 - 10\left(t-\frac{1}{3}\right)^2} u(t)\mathrm{d}t, \qquad s \in [0,1],$$

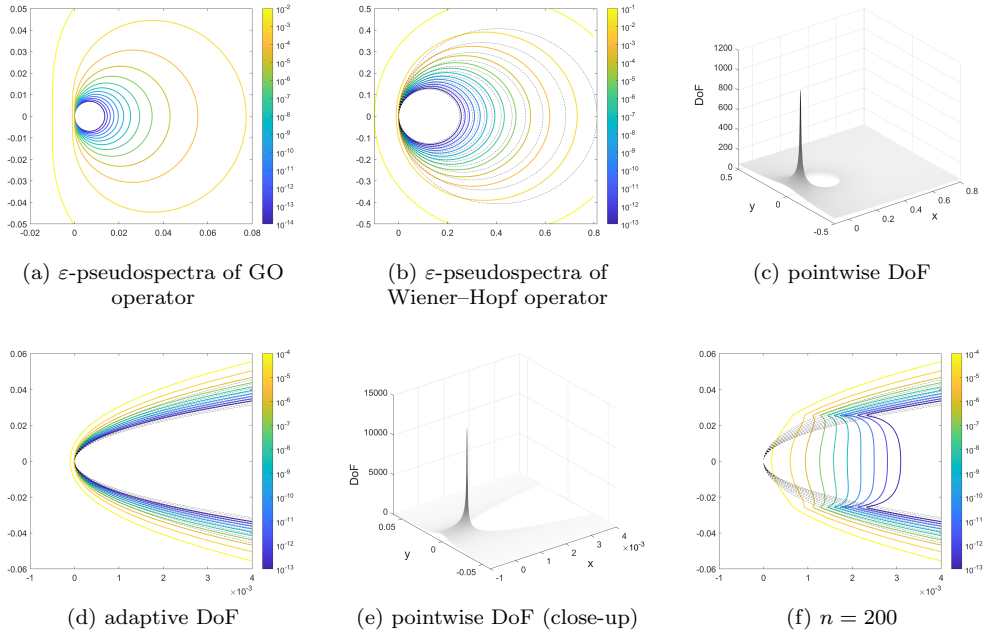
which is computed via the approach elaborated in subsection 3.1.4. Interestingly, we have not found any pseudospectral plot of a general Volterra operator in the literature. This pseudospectral plot resembles very much that of the Wiener–Hopf operator, which is well-studied. The Wiener–Hopf operator

$$(\mathcal{V}_c^{WH}u)(s) = \int_s^d e^{s-t}u(t)\mathrm{d}t, \qquad s \in [0,d],$$

is a Volterra operator of convolution type. Previous study on its nonnormality and pseudospectra includes, e.g., [30, 41].

We show in Figure 4(b) the pseudospectra of the Wiener–Hopf operator $\mathcal{V}_c$ for $d = 10$ by the boundaries of ε-pseudospectra for $\varepsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-13}$. These boundaries are in good agreement with their respective lower bounds (dashed curves) [30]. Specifically, the resolvent norm is calculated on a grid of 600×600 points before the results are sent to the contour plotter. For each grid point, the largest n used throughout the Lanczos process is shown in Figure 4(c). It can be seen that n grows as z gets closer to the origin. At $-8.35 \times 10^{-4} \pm 8.35 \times 10^{-4}\mathrm{i}$, the closest grid points

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) ε-pseudospectra of GO operator

(b) ε-pseudospectra of Wiener–Hopf operator

(c) pointwise DoF

(d) adaptive DoF

(e) pointwise DoF (close-up)

(f) $n = 200$

FIG. 4. *GO and Wiener–Hopf operators.*

to the origin, the computed solution to (3.1) is expressed by a normalized Legendre series of degree 1031, the highest among all the grid points.

Figure 4(d) is a close-up near the origin with the resolvent norm computed on a 800×800 grid. As shown in Figure 4(e), $n = 13280$ for a complete resolution at $1.25 \times 10^{-6} \pm 7.51 \times 10^{-5}$i, the grid points nearest to the origin. Thanks to the adaptivity, the determination of n is fully automated and sufficient resolution is guaranteed. In principle, the proposed method can compute the resolvent norm at a point arbitrarily close to the spectra, provided that sufficiently powerful hardware is available.

For comparison, Figure 4(f) shows the boundaries of ε-pseudospectra obtained with $n = 200$ throughout the entire grid. This "finite section" type implementation produces contours that exhibit conspicuous violation to the lower bounds for all the ε considered, highlighting the importance of the adaptivity in DoF n.

6. Extensions. In this section, we discuss several extensions of the proposed method which, together with those presented in the previous sections, cover a large portion of the pseudospectral problems encountered in practice. Although the error analysis developed in section 4 is not immediately applicable to the operators arising in these extensions, it can be readily adapted to obtain analogous results, which we omit for brevity.

6.1. Generalized eigenvalue problem. For the generalized eigenvalue problem (GEP)

$$\mathcal{A}x = \lambda \mathcal{B}x, \tag{6.1}$$

we consider the case where both $\mathcal{A}$ and $\mathcal{B}$ are differential operators defined in subsection 2.1 subject to appropriate boundary conditions so that $\mathcal{B}$ is invertible and $\mathcal{D}(\mathcal{A}) \subset \mathcal{D}(\mathcal{B})$. Specifically, we assume that $\mathcal{A}$ and $\mathcal{B}$ have, respectively, the differential expressions

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$\begin{aligned}\tau_{\mathcal{A}} &= \frac{\mathrm{d}^N}{\mathrm{d}x^N} + a_{N-1}(x)\frac{\mathrm{d}^{N-1}}{\mathrm{d}x^{N-1}} + \cdots + a_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + a_0(x),\\ \tau_{\mathcal{B}} &= b_M(x)\frac{\mathrm{d}^M}{\mathrm{d}x^M} + b_{M-1}(x)\frac{\mathrm{d}^{M-1}}{\mathrm{d}x^{M-1}} + \cdots + b_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + b_0(x).\end{aligned} \tag{6.2}$$

Without loss of generality, we also assume $M < N$. If we denote by $\Lambda(\mathcal{A},\mathcal{B}) = \{\lambda \in \mathbb{C} \mid \mathcal{A}x = \lambda\mathcal{B}x,\ x \neq 0\}$ the set of eigenvalues of (6.1), then (6.1) can be solved as the standard eigenvalue problem with $\Lambda(\mathcal{A},\mathcal{B}) = \Lambda(\mathcal{G})$ by the invertibility of $\mathcal{B}$, where $\mathcal{G} = \mathcal{B}^{-1}\mathcal{A}$. Moreover, this is followed by the ε-pseudospectrum of the GEP

$$\sigma_\varepsilon(\mathcal{A},\mathcal{B}) = \sigma_\varepsilon(\mathcal{G}),$$

which is the operator extension of the definition for the pseudospectra of matrix GEPs [44, section 45].

The lemma below formulates the resolvent and its adjoint of a GEP operator.

LEMMA 6.1. *For the GEP* (6.1)*, we assume that* $a_j(x), b_k(x) \in L^\infty([-1,1])$ *for* $j = 0,1,\ldots,N-1$ *and* $k = 0,1,\ldots,M$*, respectively, and* $\mathcal{D}(\mathcal{A}^*) \subset \mathcal{D}(\mathcal{B}^*)$*. Then* $\mathcal{R}(z) = (z\mathcal{I} - \mathcal{G})^{-1}$ *and*

$$\mathcal{R}^*(z) = \mathcal{B}^*(z^*\mathcal{B}^* - \mathcal{A}^*)^{-1} \tag{6.3a}$$

are both compact for $z \in \mathbb{C}\backslash\sigma(\mathcal{G})$*. In addition, the restriction of* $\mathcal{R}(z)$ *to* $\mathcal{D}(\mathcal{B}_{\max})$ *is*

$$\mathcal{R}(z)\big|_{\mathcal{D}(\mathcal{B}_{\max})} = (z\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}_{\max}. \tag{6.3b}$$

Proof. We note that $(z\mathcal{I} - \mathcal{G})^* = (z^*\mathcal{B}^* - \mathcal{A}^*)\mathcal{B}^{-*}$, and its domain $\mathcal{D}((z\mathcal{I} - \mathcal{G})^*) = \{u \in L^2([-1,1]) | \mathcal{B}^{-*}u \in \mathcal{D}(\mathcal{A}^*)\}$. When $z^*\mathcal{B}^* - \mathcal{A}^*$ is invertible, so is $(z\mathcal{I} - \mathcal{G})^*$, and $\mathcal{R}^*(z) = \big((z\mathcal{I} - \mathcal{G})^{-1}\big)^* = ((z\mathcal{I} - \mathcal{G})^*)^{-1} = \mathcal{B}^*(z^*\mathcal{B}^* - \mathcal{A}^*)^{-1}$ by Proposition 2.6 in [12, section XIV.2]. Since $\{a_k\}_{k=0}^{N-1}, \{b_k\}_{k=0}^{M} \in L^\infty([-1,1])$ and $M < N$, $\mathcal{B}^*$ is $(z^*\mathcal{B}^* - \mathcal{A}^*)$-compact by Theorem 1.3 in [12, section XVII.1]. It then follows from the invertibility of $z^*\mathcal{B}^* - \mathcal{A}^*$ and the argument used in the proof of Theorem 5.1 in [12, section XVII.5] that $\mathcal{B}^*(z^*\mathcal{B}^* - \mathcal{A}^*)^{-1}$ is compact, i.e., $\mathcal{R}^*(z)$ is compact, which, in turn, leads to the compactness of $\mathcal{R}(z)$.

To show (6.3b), we first consider $\mathcal{R}(z)\big|_{\mathcal{D}(\mathcal{B})}$. For $u \in \mathcal{D}(\mathcal{B})$ and $v \in L^2([-1,1])$,

$$\langle (z\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}u, v\rangle = \langle \mathcal{B}u, (z^*\mathcal{B}^* - \mathcal{A}^*)^{-1}v\rangle = \langle u, \mathcal{B}^*(z^*\mathcal{B}^* - \mathcal{A}^*)^{-1}v\rangle, \tag{6.4}$$

from which it follows that $\mathcal{R}(z)|_{\mathcal{D}(\mathcal{B})} = (z\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}$. By the extension theorem for bounded linear operators, we have $\mathcal{R}(z)|_{\mathcal{D}(\mathcal{B}_{\max})} = (z\mathcal{B} - \mathcal{A})^{-1}\mathcal{B}_{\max}$. ∎

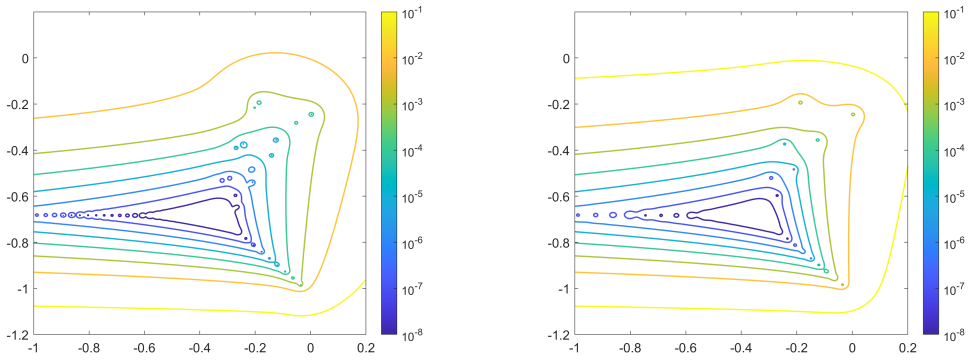
It follows from Lemma 6.1 that the analogue of (3.1a) for a GEP operator is

$$(z\mathcal{B} - \mathcal{A})v = \mathcal{B}_{\max}u, \tag{6.5}$$

where $u \in \mathcal{D}(\mathcal{B}_{\max})$. To solve (6.5), we follow the procedure described in subsection 3.1.1 to represent (or approximate) the variable coefficients in (6.2) by normalized Legendre series. This yields the (approximate) operators $z\tilde{\mathcal{B}} - \tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}_{\max}$ together with their banded infinite-dimensional matrix representations. We then apply $\tilde{\mathcal{B}}_{\max}$ to u and convert the resulting coefficients to the $C^{(M+1/2)}$ basis so as to match the range of $z\tilde{\mathcal{B}} - \tilde{\mathcal{A}}$. Finally, we solve the resulting infinite-dimensional banded linear system by adaptive QR to obtain an approximate solution v.

To apply $\mathcal{R}^*(z)$, we solve $(z^*\mathcal{B}^* - \mathcal{A}^*)h = v$ and subsequently compute $w = \mathcal{B}^*h$. As before, we form the (approximate) operators $z^*\tilde{\mathcal{B}}^* - \tilde{\mathcal{A}}^*$ and $\tilde{\mathcal{B}}^*$ together

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) L^2 norm (b) energy norm

FIG. 5. *ε-pseudospectra of the Orr–Sommerfeld operator.*

with their infinite-dimensional banded matrix representations. The banded system associated with $z^*\tilde{\mathcal{B}}^* - \tilde{\mathcal{A}}^*$ is solved by adaptive QR, after which we apply the matrix representation of $\tilde{\mathcal{B}}^*$ to the coefficient vector of h to obtain the approximate coefficients of w.

A classical GEP example is the Orr–Sommerfeld operator, which plays a pivotal role in analyzing the temporal stability of fluid flows; see [31] and [44, section 22]. In the standard form of (6.1), $\mathcal{A}^{OS}$ and $\mathcal{B}^{OS}$ are given by

$$\mathcal{A}^{OS}u = \frac{1}{R}\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2} - \alpha^2 u\right)^2 - \mathrm{i}\alpha(1-x^2)\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + \alpha^2 u\right) - 2\mathrm{i}\alpha u, \quad \mathcal{B}^{OS}u = \frac{\mathrm{d}^2u}{\mathrm{d}x^2} - \alpha^2 u,$$

with $\mathcal{A}^{OS}$ attached with $u(\pm 1) = 0$ and $u'(\pm 1) = 0$ and $\mathcal{B}^{OS}$ with $u(\pm 1) = 0$. Figure 5(a) displays the 2-norm pseudospectra for $R = 10000$ and $\alpha = 1.02$, computed by Algorithm 2.1, where the application of $\mathcal{T}(z)$ is described above. If it is the pseudospectra in the energy norm [31] defined by

$$\langle u, v\rangle_E = \langle \mathcal{B}^{OS}u, v\rangle, \quad u, v \in \mathcal{D}(\mathcal{B}^{OS}),$$

that we desire, we should note that $\mathcal{R}^*(z)$ is different from (6.3a), as (6.3a) is obtained via the Euclidean inner product. Reworking out as we do in (6.4) and noting $\mathcal{B}^{OS} = (\mathcal{B}^{OS})^*$ for this particular example, we have

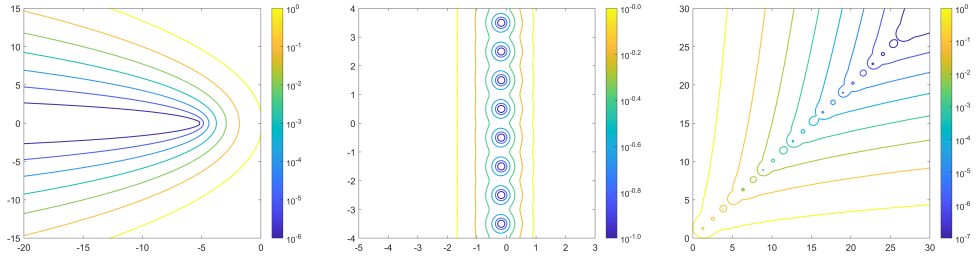
$$\langle \mathcal{R}(z)u, v\rangle_E = \langle \mathcal{B}^{OS}u, (z^*\mathcal{B}^{OS} - \mathcal{A}^{OS*})^{-1}\mathcal{B}^{OS}v\rangle = \langle u, (z^*\mathcal{B}^{OS} - \mathcal{A}^{OS*})^{-1}\mathcal{B}^{OS}u\rangle_E.$$

Thus, $\mathcal{R}^*(z) = (z^*\mathcal{B}^{OS} - (\mathcal{A}^{OS})^*)^{-1}\mathcal{B}^{OS}$. Since $\mathcal{R}(z)$ is not changed with the norm, (6.3b) still holds. The application of $\mathcal{R}^*(z)$, i.e., $w = (z^*\mathcal{B}^{OS} - (\mathcal{A}^{OS})^*)^{-1}\mathcal{B}^{OS}v$, is done by solving $(z^*\mathcal{B}^{OS} - (\mathcal{A}^{OS})^*)w = \mathcal{B}^{OS}v$. Figure 5(b) is a reproduction of Figure 2 in [31], showing the ε-pseudospectra in energy norm for $\varepsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-8}$.

6.2. Multivariate operators. The existing strategies for handling an operator $\mathcal{L}$ acting in multiple space dimensions include decoupling the pseudospectra problem into problems of lower dimensions [44, section 43] and representing $\mathcal{L}$ as an operator acting on $\ell^2(\mathbb{N})$ [4, 6]. On the other hand, the framework presented in this paper applies perfectly to an operator in multiple dimensions, provided that its resolvent is compact or compact-plus-scalar. We now take as an example the Dirichlet operator [12, section XIV.6], noting that most of the elliptic operators fall into this category.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy



(a) 2D advection-diffusion operator

(b) block-structured wave operator

(c) Davies' operator in an unbounded domain

FIG. 6. *ε-pseudospectra of multivariate operator, block operator, and operator on an unbounded domain.*

Given a bounded open set Ω in $\mathbb{R}^d$, consider the partial differential expression

$$\tau_e = \sum_{i,j=1}^{d} a_{ij}(\mathbf{x}) \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} + \sum_{i=1}^{d} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} + c(\mathbf{x}),$$

where a_{ij}, b_i, and c belong to $C^\infty(\bar{\Omega})$ with $\bar{\Omega}$ denoting the closure of Ω. The Dirichlet operator $\mathcal{P} : L^2(\Omega) \to L^2(\Omega)$ is defined as $\mathcal{P}u = \tau_e u$ with $\mathcal{D}(\mathcal{P}) = H_0^1(\Omega) \cap H^2(\Omega)$. If $\mathcal{P}$ is uniformly elliptic, $\mathcal{R}(z) = (z\mathcal{I} - \mathcal{P})^{-1}$ is compact for $z \in \mathbb{C} \backslash \sigma(\mathcal{P})$ [12, section XIV.6].

In implementing Algorithms 2.1 and 3.1, we solve (3.1) by the multivariate ultraspherical spectral method [36, 39]. The basis functions for such a problem are the tensor product of recombined normalized Legendre polynomials that satisfy boundary conditions in each spatial dimension. After reordering these tensor-product bases by the total degree, we obtain an infinite-dimensional matrix representation of $\mathcal{P}$ with a growing bandwidth, which we solve by adaptive QR. Figure 6(a) shows the pseudospectra of the 2D advection-diffusion operator [16]

$$\mathcal{P}^{AD} u = \eta \Delta u + \mathbf{V} \cdot \nabla u$$

in $\Omega = [-1, 1]^2$ with $\eta = 0.05$ and $\mathbf{V} = [0, -1]^T$.

6.3. Block operators. It is not uncommon to consider the pseudospectra of block operators that act on Cartesian product of Hilbert spaces. Such an operator may come from reducing a high-order problem to a system of low-order ones. For example, Trefethen and Driscoll in [9, 41] investigated the wave operator

$$\mathcal{Q}^B = \begin{pmatrix} 0 & \mathrm{d}/\mathrm{d}x \\ \mathrm{d}/\mathrm{d}x & 0 \end{pmatrix}$$

with $\mathcal{D}(\mathcal{Q}^B) = \{(u, v) \in L^2([0, \pi]) \times L^2([0, \pi]) \mid v(0) = 0,\ u(\pi) + \eta v(\pi) = 0\}$ and $\eta = 1/2$. The compactness of $(z\mathcal{I} - \mathcal{Q}^B)^{-1}$ can be established following the proof in [12, section XI.1]. Figure 6(b) displays the pseudospectra of $\mathcal{Q}^B$ obtained using the proposed method. This replicates Figure 3 in [9], which was originally produced from a finite-dimensional discretization of $\mathcal{Q}^B$ using the finite difference method. In solving (3.1) by the ultraspherical spectral method, each occurrence of $\mathrm{d}/\mathrm{d}x$ gives rise to an infinite-dimensional banded matrix. Consequently, $\mathcal{Q}^B$ is represented by a 2×2 block matrix of infinite dimensions, whose $(1, 2)$ and $(2, 1)$ blocks are banded. A

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

suitable reordering then transforms the resulting system into an infinite-dimensional banded system. Moreover, the boundary conditions can be enforced through basis recombination without destroying the banded structure.

6.4. Operators on unbounded domains. We sometimes need to consider operators defined on unbounded domains, such as the Schrödinger operator. As an example, consider the 2D Davies operator of complex harmonic oscillator [8]

$$\mathcal{P}^D u = -h^2 \Delta u + \mathrm{i}(x^2 + y^2)u,$$

which acts on $L^2(\mathbb{R}^2)$. For such a problem, we have to ditch the Legendre polynomials and take the Hermite functions as the basis [6]. After reordering the basis functions in the tensor product of the Hermite functions, the infinite-dimensional matrix representation of $\mathcal{P}^D$ is banded with a growing bandwidth. This system is again solved by adaptive QR. Figure 6(c) is an illustration of the pseudospectra of Davies' operator for $h = 0.8$, which much resembles the pseudospectra of the one-dimensional version given in [8] and [44, section 43].

7. Conclusion. We have proposed a continuous framework for computing the pseudospectra of linear operators, provided that the linear operator has a compact or compact-plus-scalar resolvent. The new method is free of spectral pollution and invisibility and features adaptivity in both the stopping criterion for the operator Lanczos iteration and the DoF for fully resolved solutions to the inverse resolvent equations. The adoption of the well-conditioned spectral methods prevents any deterioration in the conditioning of the problem. In addition, a thorough error analysis is supplied for the computed resolvent norm at a single point and the pseudospectra in a prescribed region of the complex plane.

We are pleased to see what was once anticipated by Nick Trefethen a quarter century ago has now come to pass.

Reproducibility of computational results. This paper has been awarded the "SIAM Reproducibility Badge: Code and Data Available" as a recognition that the authors have followed reproducibility principles valued by SISC and the scientific computing community. Code and data that allow readers to reproduce the results in this paper are available at https://github.com/kuandeng/PseSp and in the supplementary materials (PseSp-master.zip [local/web 2.29MB]), linked from the main article webpage.

Acknowledgments. We are grateful to Matthew Colbrook (Cambridge) and Anthony P. Austin (NPS) for sharing with us their insights and valuable feedback on a draft of this paper which led us to significantly improve our work. We have also benefited from discussions with our team members Lu Cheng, Ouyuan Qin, and Kaining Deng. Finally, we would like to thank the two constructive anonymous referees.

REFERENCES

[1] J. L. Aurentz and L. N. Trefethen, *Chopping a Chebyshev series*, ACM Trans. Math. Software, 43 (2017), 33, https://doi.org/10.1145/2998442.

[2] K. J. Burns, G. M. Vasil, J. S. Oishi, D. Lecoanet, and B. P. Brown, *Dedalus: A flexible framework for numerical simulations with spectral methods*, Phys. Rev. Res., 2 (2020), 023068, https://doi.org/10.1103/PhysRevResearch.2.023068.

[3] M. J. Colbrook, *Computing semigroups with error control*, SIAM J. Numer. Anal., 60 (2022), pp. 396–422, https://doi.org/10.1137/21M1398616.

[4] M. J. Colbrook and A. C. Hansen, *The foundations of spectral computations via the solvability complexity index hierarchy*, J. Eur. Math. Soc., 25 (2022), pp. 4639–4718, https://doi.org/10.4171/jems/1289.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[5] M. Colbrook, A. Horning, and A. Townsend, *Computing spectral measures of self-adjoint operators*, SIAM Rev., 63 (2021), pp. 489–524, https://doi.org/10.1137/20M1330944.

[6] M. J. Colbrook, B. Roman, and A. C. Hansen, *How to compute spectra with error control*, Phys. Rev. Lett., 122 (2019), 250201, https://doi.org/10.1103/PhysRevLett.122.250201.

[7] M. J. Colbrook and A. Townsend, *Avoiding discretization issues for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 46 (2025), pp. 648–675, https://doi.org/10.1137/23M1569927.

[8] E. B. Davies, *Pseudospectra, the harmonic oscillator and complex resonances*, Proc. R. Soc. A, 455 (1999), pp. 585–599, https://doi.org/10.1098/rspa.1999.0325.

[9] T. A. Driscoll and L. N. Trefethen, *Pseudospectra for the wave equation with an absorbing boundary*, J. Comput. Appl. Math., 69 (1996), pp. 125–142, https://doi.org/10.1016/0377-0427(95)00021-6.

[10] M. A. Gilles and A. Townsend, *Continuous analogues of Krylov subspace methods for differential operators*, SIAM J. Numer. Anal., 57 (2019), pp. 899–924, https://doi.org/10.1137/18M1177810.

[11] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Basic Classes of Linear Operators*, Birkhäuser, 2012.

[12] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Birkhäuser, 2013.

[13] T. S. Gutleb and S. Olver, *A sparse spectral method for Volterra integral equations using orthogonal polynomials on the triangle*, SIAM J. Numer. Anal., 58 (2020), pp. 1993–2018, https://doi.org/10.1137/19M1267441.

[14] N. Hale and A. Townsend, *An algorithm for the convolution of Legendre series*, SIAM J. Sci. Comput., 36 (2014), pp. A1207–A1220, https://doi.org/10.1137/140955835.

[15] A. C. Hansen, *Infinite-dimensional numerical linear algebra: Theory and applications*, Proc. R. Soc. A, 466 (2010), pp. 3539–3559.

[16] L. Hemmingsson, *A semi-circulant preconditioner for the convection-diffusion equation*, Numer. Math., 81 (1998), pp. 211–248, https://doi.org/10.1007/s002110050390.

[17] N. J. Higham, *Accuracy and Stability of Numerical Alorithms*, SIAM, 1996.

[18] A. Horning and A. Townsend, *FEAST for differential eigenvalue problems*, SIAM J. Numer. Anal., 58 (2020), pp. 1239–1262, https://doi.org/10.1137/19M1238708.

[19] J. Keiner, *Fast Polynomial Transforms*, Logos Verlag Berlin GmbH, 2011.

[20] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Natl. Inst. Stand. Technol., 45 (1950), pp. 255–282, https://doi.org/10.6028/jres.045.026.

[21] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, 1998, https://doi.org/10.1137/1.9780898719628.

[22] X. Liu, K. Deng, and K. Xu, *Spectral approximation of convolution operators of Fredholm type*, SIAM J. Sci. Comput., 47 (2025), pp. A1964–A1982, https://doi.org/10.1137/24M1652532.

[23] S.H. Lui, *Computation of pseudospectra by continuation*, SIAM J. Sci. Comput., 18 (1997), pp. 565–573, https://doi.org/10.1137/S1064827594276035.

[24] M. Mortensen, *A generic and strictly banded spectral Petrov–Galerkin method for differential equations with polynomial coefficients*, SIAM J. Sci. Comput., 45 (2023), pp. A123–A146, https://doi.org/10.1137/22M1492842.

[25] S. Olver and A. Townsend, *A fast and well-conditioned spectral method*, SIAM Rev., 55 (2013), pp. 462–489, https://doi.org/10.1137/120865458.

[26] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, IMA J. Appl. Math., 18 (1976), pp. 341–349, https://doi.org/10.1093/imamat/18.3.341.

[27] C. C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258, https://doi.org/10.1016/0024-3795(80)90167-6.

[28] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, 1981.

[29] O. Qin, L. Cheng, and K. Xu, *A New Banded Petrov–Galerkin Spectral Method*, preprint, arXiv:2502.11652, 2025.

[30] S. C. Reddy, *Pseudospectra of Wiener-Hopf integral operators and constant-coefficient differential operators*, J. Integral Equ. Appl., 5 (1993), pp. 369–403, https://doi.org/10.1216/jiea/1181075761.

[31] S. C. Reddy, P. J. Schmid, and D. S. Henningson, *Pseudospectra of the Orr–Sommerfeld operator*, SIAM J. Appl. Math., 53 (1993), pp. 15–47, https://doi.org/10.1137/0153002.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[32] Y. Saad, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706, https://doi.org/10.1137/0717059.
[33] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505, https://doi.org/10.1137/0915089.
[34] J. Shen, *Efficient spectral-Galerkin method* II. *Direct solvers of second- and fourth-order equations using Chebyshev polynomials*, SIAM J. Sci. Comput., 16 (1994), pp. 74–87, https://doi.org/10.1137/0916006.
[35] F. Smithies, *Integral Equations*, Vol. 49, CUP Archive, 1958.
[36] C. Strössner and D. Kressner, *Fast global spectral methods for three-dimensional partial differential equations*, IMA J. Numer. Anal., 43 (2023), pp. 1519–1542, https://doi.org/10.1093/imanum/drac030.
[37] G. Szegő, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, 1939.
[38] A. Townsend, *Computing with Functions in Two Dimensions*, Ph.D. thesis, Oxford University, UK, 2014.
[39] A. Townsend and S. Olver, *The automatic solution of partial differential equations using a global spectral method*, J. Comput. Phys., 299 (2015), pp. 106–123, https://doi.org/10.1016/j.jcp.2015.06.031.
[40] A. Townsend and L. N. Trefethen, *An extension of Chebfun to two dimensions*, SIAM J. Sci. Comput., 35 (2013), pp. C495–C518, https://doi.org/10.1137/130908002.
[41] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406, https://doi.org/10.1137/S0036144595295284.
[42] L. N. Trefethen, *Computation of pseudospectra*, Acta Numer., 8 (1999), pp. 247–295, https://doi.org/10.1017/S0962492900002932.
[43] L. N. Trefethen, *Approximation Theory and Approximation Practice*, extended ed., SIAM, 2019, https://doi.org/10.1137/1.9781611975949.
[44] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behaviour of Non-normal Matrices and Operators*, Princeton University Press, 2005.
[45] H. Wang and S. Xiang, *On the convergence rates of Legendre approximation*, Math. Comp., 81 (2012), pp. 861–877, https://doi.org/10.1090/S0025-5718-2011-02549-4.
[46] K. Xu and A. F. Loureiro, *Spectral approximation of convolution operators*, SIAM J. Sci. Comput., 40 (2018), pp. A2336–A2355, https://doi.org/10.1137/17M1149249.

Downloaded 11/03/26 to 195.245.219.41 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.