Numerische Mathematik (2019) 143:17–83
https://doi.org/10.1007/s00211-019-01047-5

Numerische
Mathematik

On the infinite-dimensional QR algorithm

Matthew J. Colbrook[1] · Anders C. Hansen[1]

Received: 8 June 2018 / Revised: 11 April 2019 / Published online: 18 May 2019

Abstract
Spectral computations of infinite-dimensional operators are notoriously difficult, yet ubiquitous in the sciences. Indeed, despite more than half a century of research, it is still unknown which classes of operators allow for the computation of spectra and eigenvectors with convergence rates and error control. Recent progress in classifying the difficulty of spectral problems into complexity hierarchies has revealed that the most difficult spectral problems are so hard that one needs three limits in the computation, and no convergence rates nor error control is possible. This begs the question: which classes of operators allow for computations with convergence rates and error control? In this paper, we address this basic question, and the algorithm used is an infinite-dimensional version of the QR algorithm. Indeed, we generalise the QR algorithm to infinite-dimensional operators. We prove that not only is the algorithm executable on a finite machine, but one can also recover the extremal parts of the spectrum and corresponding eigenvectors, with convergence rates and error control. This allows for new classification results in the hierarchy of computational problems that existing algorithms have not been able to capture. The algorithm and convergence theorems are demonstrated on a wealth of examples with comparisons to standard approaches (that are notorious for providing false solutions). We also find that in some cases the IQR algorithm performs better than predicted by theory and make conjectures for future study.

Mathematics Subject Classification 47A10 · 65J10 · 46N40 · 03D55

1 Introduction

Spectral computations are ubiquitous in the sciences with applications in solutions to differential and integral equations, spline functions, orthogonal polynomials, quantum

✉ Matthew J. Colbrook
m.colbrook@damtp.cam.ac.uk

[1] DAMTP, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Rd, Cambridge CB3 0WA, UK

mechanics, quantum chemistry, statistical mechanics, Hermitian and non-Hermitian Hamiltonians, optics etc. [10,26,27,34,59,67,68,70]. The computational problem is as follows. Letting T denote a bounded linear operator on the canonical separable Hilbert space $l^2(\mathbb{N})$, one wants to design algorithms to compute the spectrum of T, denoted by $\sigma(T)$. Given the many applications, this problem has been investigated intensely since the 1950s [3,4,6,7,16,17,20–22,25,28,29,35–38,43,46,47,61,63–66,72], and we can only cite a small subset here.

In the paper "On the Solvability Complexity Index, the n-pseudospectrum and approximations of spectra of operators" [38] the Solvability Complexity Index (SCI) was introduced. The SCI provides a classification hierarchy [8,9,38] of spectral problems according to their computational difficulty. The SCI of a class of spectral problems is the least number of limits needed in order to compute the spectrum of operators in this class. From a classical numerical analysis point of view, such a concept may seem foreign. Indeed, the traditional sentiment is that one should have an algorithm, Γ_n, such that for an operator $T \in \mathcal{B}(l^2(\mathbb{N}))$,

$$\Gamma_n(T) \longrightarrow \sigma(T), \quad n \to \infty, \tag{1.1}$$

preferably with some form of error control of the convergence. As this philosophy forms the basics of numerical analysis, it naturally permeates the classical literature on the computational spectral problem. However, as is shown in [8,9,38], an algorithm satisfying (1.1) is impossible even for the class of self-adjoint operators. Indeed, in the general case, the best possible alternative is an algorithm depending on three indices n_1, n_2, n_3 such that

$$\lim_{n_3\to\infty}\lim_{n_2\to\infty}\lim_{n_1\to\infty}\Gamma_{n_3,n_2,n_1}(T) = \sigma(T).$$

In fact, any algorithm with fewer than three limits will fail on the general class of operators. Moreover, no error control nor convergence rate on any of the limits are possible, since any such error control would reduce the number of limits needed. However, for the self-adjoint and normal cases, two limits suffice in order to recover the spectrum. This phenomenon implies that the only way to characterise the computational spectral problem is through a hierarchy classifying the difficulty of computing spectra of different subclasses of operators. This is the motivation behind the SCI hierarchy, which also covers general numerical analysis problems. Indeed, the SCI hierarchy is closely related to Smale's question on the existence of purely iterative generally convergent algorithm for polynomial zero finding [69]. As demonstrated by McMullen [49,50] and Doyle and McMullen [30], this is a case where several limits are needed in the computation, and their results become special cases of classification in the SCI hierarchy [8,9].

Informally, the SCI hierarchy is characterised as follows (see the "Appendix 1" for a more detailed summary describing the SCI hierarchy).

Δ_0: The set of problems that can be computed in finite time, the SCI $= 0$.

Δ_1: The set of problems that can be computed using one limit, the SCI $= 1$, however, one has error control and one knows an error bound that tends to zero as the algorithm progresses.

Δ_2: The set of problems that can be computed using one limit, the SCI $= 1$, but error control may not be possible.

Δ_{m+1}: For $m \in \mathbb{N}$, the set of problems that can be computed by using m limits, the SCI $\leq m$.

The class Δ_1 is of course a highly desired class, however, most spectral problems are much higher in the hierarchy. For example, we have the following known classifications [8,9,38].

(i) The general spectral problem is in $\Delta_4 \backslash \Delta_3$.
(ii) The self-adjoint spectral problem is in $\Delta_3 \backslash \Delta_2$.
(iii) The compact spectral problem is in $\Delta_2 \backslash \Delta_1$.

Here, the notation\indicates the standard "setminus". Note that the SCI hierarchy can be refined. We will not consider the full generalisation in the higher part of the hierarchy in this paper, but recall the class Σ_1 [24]. This class is defined as follows.

Σ_1: We have $\Delta_1 \subset \Sigma_1 \subset \Delta_2$ and Σ_1 is the set of problems that can be computed by passing to one limit. Error control may not be possible, however, there exists an algorithm for these problems that converges and for which its output is included in the spectrum (up to an arbitrarily small accuracy parameter ϵ).

In the context of computing $\sigma(T)$, a Σ_1 classification means the existence of an algorithm Γ_n such that

$$\Gamma_n(T) \subset \sigma(T) + B_{2^{-n}}(0)$$

and Γ_n converges to $\sigma(T)$ in the Hausdorff metric. The Σ_1 class is very important as it allows for algorithms that never make a mistake. In particular, one is always sure that the output is sound but we do not know if we have everything yet. The simplest infinite-dimensional spectral problem is that of computing the spectrum of an infinite diagonal matrix and, as is easy to see, we have the following.

(iv) The problem of computing spectra of infinite diagonal matrices is in $\Sigma_1 \backslash \Delta_1$.

Hence, the computational spectral problem becomes an infinite classification theory in order to characterise the above hierarchy. In order to do so, there will, necessarily, have to be many different types of algorithms. Indeed, characterising the hierarchy will yield a myriad of different approaches, as different structures on the various classes of operators will require specific algorithms. The key contribution of this paper is to investigate the convergence properties of the infinite-dimensional QR (IQR) algorithm, its implementation properties, and how this algorithm provides classification results in the SCI hierarchy.

1.1 Main contribution and novelty of the paper

The main contributions of the paper can be summarised as follows: New convergence results, algorithmic results (the IQR algorithm can be implemented), classification results in the SCI hierarchy and numerical examples.

(1) *Convergence results* We provide new convergence theorems for the IQR algorithm with convergence rates and error control. The results include eigenvalues, eigenvectors and invariant subspaces.
(2) *Algorithmic implementation* We prove that for infinite matrices with finitely many non-zero entries in each column, it is possible to implement the IQR algorithm exactly (on a finite machine) as if one had an infinite computer at one's disposal. This can be extended to implementing the IQR algorithm with error control for general invertible operators.
(3) *SCI hierarchy classifications* As a result of (1) and (2), we provide new classification results for the SCI hierarchy. In particular, the convergence properties of the IQR algorithm capture key structures that allow for sharp Δ_1 classification of the problem of computing extremal points in the spectrum. Moreover, we establish sharp Σ_1 classification of the problem of computing spectra of subclasses of compact operators.
(4) *Numerical examples* Finally, we demonstrate the IQR algorithm and the proven convergence results on a variety of difficult problems in practical computation, illustrating how the IQR algorithm is much more than a theoretical concept. Moreover, the examples demonstrate that the IQR algorithm performs much better than predicted by our theory, working on much larger classes of operators. Hence, we are left with many open problems on the theoretical understanding of the potential and limitations of this algorithm. The computational experiments include examples from

    (i) Toeplitz/Laurent operators and their perturbations,
    (ii) PT-symmetry in quantum mechanics,
    (iii) Hopping sign model in sparse neural networks,
    (iv) NSA Anderson model in superconductors.

1.2 Connection to previous work

Our results connect to many different approaches in the vast literature on spectral computation in infinite dimensions. The infinite-dimensional computational spectral problem is very different from the finite-dimensional computational eigenvalue problem, and even though the IQR algorithm is inspired by the finite-dimensional version, this paper solely focuses on the infinite-dimensional problem. Thus, the paper is aimed at the analysis and numerical analysis audience focusing on infinite-dimensional problems rather than the finite-dimensional numerical linear algebra discipline.

Finite sections The IQR algorithm provides an alternative to the standard finite section method in several cases where it fails. Whereas the finite section method would extract a finite section from the infinite matrix and then apply, for example, the finite-dimensional QR algorithm, the IQR algorithm first performs the infinite QR iterations and then extracts a finite section. In general, these two processes do not commute. The finite section method (or any derivative of it) cannot work in general because of the general classification results in the SCI hierarchy mentioned in Sect. 1. Typically, it may provide false solutions. However, in the cases where it converges, it provides invaluable Δ_2 classifications in the SCI hierarchy. The finite section method has often been viewed in connection with Toeplitz theory and the reader may want to consult the work by Böttcher [14,15], Böttcher and Silberman [18], Böttcher et al. [16], Brunner et al. [22], Hagen et al. [35], Lindner [44], Marletta [46] and Marletta and Scheichl [47]. From the operator algebra point of view, the work of Arveson [5–7] has been influential as well as the work of Brown [21].

Infinite-dimensional Toda flow Deift et al. [28] provided the first results on the IQR algorithm in connection with Toda flows with infinitely many variables. Their results are purely functional analytic and do not take implementation and computability issues into account. However, these results provide the fundamentals of the IQR algorithm. In [36] these results were expanded with a convergence result for eigenvectors corresponding to eigenvalues outside the essential numerical range for normal operators. Yet, this paper did not consider convergence rates, actual numerical calculation nor any classification results.

Infinite-dimensional QL algorithm Olver, Townsend and Webb have provided a practical framework for infinite-dimensional linear algebra and foundational results on computations with infinite data structures [53–56,73]. This includes efficient codes as well as theoretical results. The infinite-dimensional QL (IQL) algorithm is an important part of this program. The IQL algorithm is rather different from the IQR algorithm, although they are similar in spirit. In particular, both the implementation and the convergence results are somewhat contrasting.

Infinite-dimensional spectral computation: The results in this paper follow in the long tradition of infinite-dimensional spectral computations. This field contains a vast literature that spans more than half a century, and the references that we have cited in the first paragraph of Sect. 1 represent a small sample. However, we would like to highlight the recent work by Bögli et al. [13] who were able to computationally confirm, with absolute certainty, a conjecture on a certain oscillatory behaviour of higher auto-ionizing resonances of atoms. Note that problems that are classified as Δ_1 and Σ_1 in the SCI hierarchy may allow for computer assisted proofs.

1.3 Background and notation

Here we briefly recall some definitions used in the paper. We will consider the canonical separable Hilbert space $\mathcal{H} = l^2(\mathbb{N})$ (the set of square summable sequences). Moreover, we write $\mathcal{B}(\mathcal{H})$ for the set of bounded operators on $\mathcal{H}$. For orthogonal projections E, F, we will write $E \leq F$ if the range of E is a subspace of the range of F. We denote the canonical orthonormal basis of $\mathcal{H}$ by $\{e_j\}_{j\in\mathbb{N}}$, and if $\xi \in \mathcal{H}$ we write $\xi(j) = \langle \xi, e_j \rangle$.

Note that $T \in \mathcal{B}(\mathcal{H})$ is uniquely determined by its matrix elements $t_{ij} = \langle Te_j, e_i\rangle$. Hence we will use the words bounded operator and infinite matrix interchangeably. Given a sequence of operators $\{T_n\}$, we will use the notation

$$T_n \xrightarrow{\text{SOT}} T, \qquad T_n \xrightarrow{\text{WOT}} T$$

to mean convergence in the strong and weak operator topology respectively. The spectrum of $T \in \mathcal{B}(\mathcal{H})$ will be denoted by $\sigma(T)$, and $\sigma_d(T)$ denotes the set of isolated eigenvalues with finite multiplicity (the discrete spectrum).

In connection with the spectrum, we need to recall some definitions which will appear in the statement of our theorems. We recall that, for $T \in \mathcal{B}(\mathcal{H})$, the essential spectrum[1] and the essential spectral radius are given by

$$\sigma_{\mathrm{ess}}(T) = \{z \in \mathbb{C} : T - zI \text{ is not Fredholm}\}, \quad r_{\mathrm{ess}}(T) = \sup\{|z| : z \in \sigma_{\mathrm{ess}}(T)\}.$$

Moreover, the numerical range and the essential numerical range of T are defined by

$$W(T) = \{\langle T\xi, \xi\rangle : \|\xi\| = 1\}, \quad W_e(T) = \bigcap_{K \text{ compact}} \overline{W(T+K)}.$$

In addition, we need the Hausdorff metric as defined by the following. Let $\mathcal{S}, \mathcal{T} \subset \mathbb{C}$, be compact. Then their Hausdorff distance is

$$d_H(\mathcal{S}, \mathcal{T}) = \max\{\sup_{\lambda \in \mathcal{S}} d(\lambda, \mathcal{T}), \sup_{\lambda \in \mathcal{T}} d(\lambda, \mathcal{S})\}, \tag{1.2}$$

where $d(\lambda, \mathcal{T}) = \inf_{\rho \in \mathcal{T}} |\rho - \lambda|$. We also recall a generalisation of the spectrum, known as the pseudospectrum. Indeed, for $\epsilon > 0$ define the ϵ-pseudospectrum as

$$\sigma_\epsilon(T) = \left\{z \in \mathbb{C} : \left\|(T - zI)^{-1}\right\| \geq \epsilon^{-1}\right\},$$

where we interpret $\left\|S^{-1}\right\|$ as $+\infty$ if S does not have a bounded inverse. This is easier to compute than the spectrum, converges in the Hausdorff metric to the spectrum as $\epsilon \downarrow 0$ and gives an indication of the instability of the spectrum of T. We shall use it as a comparison for the IQR algorithm and as a means to detect spectral pollution for finite section methods.

Finally, we need a notion of convergence of subspaces. We follow the notation in [41]. Let $M \subset \mathcal{B}$ and $N \subset \mathcal{B}$ be two non-trivial closed subspaces of a Banach space $\mathcal{B}$. The distance between them is defined by

$$\delta(M, N) = \sup_{\substack{x \in M \\ \|x\|=1}} \inf_{y \in N} \|x - y\|, \qquad \hat{\delta}(M, N) = \max[\delta(M, N), \delta(N, M)].$$

[1] Of course in the case of non-normal T there are different definitions of the essential spectrum. However, these differences will not matter regarding the results of this paper.

Given subspaces M and $\{M_k\}$ such that $\hat{\delta}(M_k, M) \to 0$ as $k \to \infty$, we will use the notation $M_k \to M$. If we replace $\mathcal{B}$ with a Hilbert space $\mathcal{H}$, we can express δ and $\hat{\delta}$ conveniently in terms of projections and operator norms. In particular, if E and F are the orthogonal projections onto subspaces $M \subset \mathcal{H}$ and $N \subset \mathcal{H}$ respectively, then

$$\delta(M, N) = \sup_{\substack{x \in M \\ \|x\|=1}} \inf_{y \in N} \|x - y\| = \sup_{\substack{x \in M \\ \|x\|=1}} \|F^{\perp} x\| = \|F^{\perp} E\|.$$

Since the operator $E - F = F^{\perp} E - F E^{\perp}$ is essentially the direct sum of operators $F^{\perp} E \oplus (-F E^{\perp})$, its norm is $\hat{\delta}(M, N)$, i.e.

$$\hat{\delta}(M, N) = \max(\|F^{\perp} E\|, \|E^{\perp} F\|) = \max(\|F^{\perp} E\|, \|F E^{\perp}\|) = \|E - F\|. \quad (1.3)$$

This allows us to extend the definition to allow the trivial subspace $\{0\}$ and gives rise to a metric on the set of all closed subspaces of $\mathcal{H}$ (first introduced by Krein and Krasnoselski in [42]). We also define the (maximal) subspace angle, $\phi(M, N) \in [0, \pi/2]$, between M and N by

$$\sin\big(\phi(M, N)\big) = \hat{\delta}(M, N). \quad (1.4)$$

Finally, we will use two further well-known properties in the Hilbert space setting. First, if M and N are both finite l-dimensional subspaces, then

$$\delta(M, N) \leq l^{\frac{1}{2}} \delta(N, M), \quad (1.5)$$

which shows that to prove convergence of finite-dimensional subspaces, it is enough to prove δ-convergence. Second, suppose we have

$$M = \bigoplus_{j=1}^{n} M_j, \quad N^{(k)} = N_1^{(k)} + \cdots + N_n^{(k)},$$

where the $N_j^{(k)}$ need not be orthogonal. Then a simple application of Hölder's inequality yields

$$\delta(M, N^{(k)}) \leq \left(\sum_{j=1}^{n} \delta(M_j, N_j^{(k)})^2 \right)^{\frac{1}{2}}, \quad (1.6)$$

which shows that if the dimensions of M_j and $N_j^{(k)}$ are finite and equal, then to prove convergence $N^{(k)} \to M$ we only need to prove that $\delta(M_j, N_j^{(k)}) \to 0$ as $k \to \infty$. For further properties (including other notions of distances between subspaces) and a discussion on two projections theory, we refer the reader to the excellent article of Böttcher and Spitkovsky [19].

1.4 Organisation of the paper

The paper is organised as follows. In Sect. 2 we define the IQR algorithm (simple codes are also provided in the appendix). Section 3 contains and proves our main theorems including convergence rates. The outcome is more elaborate than the finite-dimensional case, as the infinite-dimensional setting includes more intricate instances. Our key practical result is that, despite being an algorithm dealing with infinite amount of information, it can be implemented on any standard computer and this is discussed in Sect. 4. The fact that the IQR algorithm can be computed allows for its use in order to provide new classification in the SCI hierarchy as discussed in Sect. 5. In particular, we demonstrate Δ_1 classification for the extremal part of the spectrum and dominant invariant subspaces, as well as Σ_1 results for spectra of certain classes of compact operators. Note that the general spectral problem for compact operators is not in Σ_1. The IQR algorithm and convergence theorems are demonstrated on a large collection of examples from the sciences on difficult computational spectral problems in Sect. 6, with comparisons to the finite section method. The IQR algorithm is also found to perform better than theory predicts and we conjecture conditions on the operator for this to be the case. Finally, we conclude with a discussion of the opportunities and limits of the IQR algorithm in Sect. 7.

2 The infinite-dimensional QR algorithm (IQR)

The IQR algorithm has existed as a pure mathematical concept for more than thirty years and it first appeared in the paper “Toda Flows with Infinitely Many Variables” [28] in 1985. However, the analysis in [28] covers only self-adjoint infinite matrices with real entries, and since the analysis is done from a pure mathematical perspective, the question regarding the actual numerical algorithm is left out. We will extend the analysis to more general operators and answer the crucial question: can one actually implement the IQR algorithm? The answer is affirmative, and we also prove convergence theorems, generalising the well-known finite-dimensional case.

2.1 The QR decomposition

The QR decomposition is the core of the QR algorithm. If $T \in \mathbb{C}^{n\times n}$, one may apply the Gram-Schmidt procedure to the columns of T and store these columns in a matrix Q. This gives us the QR decomposition

$$T = QR, \tag{2.1}$$

where Q is a unitary matrix and R upper triangular. It is no surprise that a QR decomposition should exist in the infinite-dimensional case, however, we need more than just the existence. A key ingredient in the QR algorithm are Householder transformations, used for computational reasons (they are backwards stable). It is crucial that we can adopt these tools in the infinite-dimensional setting. Our goal is to extend the construction of the QR decomposition, via Householder transformations, to infinite matrices and to find a way so that one can implement the procedure on a finite machine. To do this, we need to introduce the concept of Householder reflections in the infinite-dimensional setting.

Definition 2.1 A Householder reflection is an operator $S \in \mathcal{B}(\mathcal{H})$ of the form

$$S = I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}, \qquad \xi \in \mathcal{H}, \tag{2.2}$$

where $\bar{\xi}$ denotes the associated functional in $\mathcal{H}^*$ given by $x \to \langle x, \xi\rangle$. In the case where $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ and I_i is the identity on $\mathcal{H}_i$ then

$$U = I_1 \oplus \left(I_2 - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}\right) \qquad \xi \in \mathcal{H}_2,$$

will be called a Householder transformation.

A straightforward calculation shows that $S^* = S^{-1} = S$ and thus also $U^* = U^{-1} = U$. An important property of the operator S is that if $\{e_j\}$ is an orthonormal basis for $\mathcal{H}$ and $\eta \in \mathcal{H}$, then one can choose $\xi \in \mathcal{H}$ such that

$$\langle S\eta, e_j\rangle = \left\langle \left(I - \frac{2}{\|\xi\|^2}\xi \otimes \bar{\xi}\right)\eta, e_j\right\rangle = 0, \qquad \forall j \neq 1.$$

In other words, one can introduce zeros in the column below the diagonal entry. Indeed, if $\eta_1 = \langle \eta, e_1\rangle \neq 0$ one may choose $\xi = \eta \pm \|\eta\|\zeta$, where $\zeta = \eta_1/|\eta_1|e_1$ and if $\eta_1 = 0$ choose $\xi = \eta \pm \|\eta\|e_1$. The following theorem gives the existence of a QR decomposition, even in the case where the operator is not invertible.

Theorem 2.2 ([36]) *Let T be a bounded operator on a separable Hilbert space $\mathcal{H}$ and let $\{e_j\}_{j\in\mathbb{N}}$ be an orthonormal basis for $\mathcal{H} \cong l^2(\mathbb{N})$. Then there exists an isometry Q such that $T = QR$, where R is upper triangular with respect to $\{e_j\}$. Moreover,*

$$Q = \underset{n\to\infty}{SOT\text{-}lim}\, V_n$$

where $V_n = U_1 \cdots U_n$ are unitary and each U_j is a Householder transformation.

2.2 The IQR algorithm

Let $T \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. By Theorem 2.2 we have $T = QR$, where Q is an isometry and R is upper triangular with respect to $\{e_j\}$. Since T is invertible, Q is in fact unitary. Consider the following construction of unitary operators $\{\hat{Q}_k\}$ and upper triangular (w.r.t. $\{e_j\}$) operators $\{\hat{R}_k\}$. Let $T = Q_1R_1$ be a QR decomposition of T and define $T_1 = R_1Q_1$. Then QR factorize $T_1 = Q_2R_2$ and define $T_2 = R_2Q_2$. The recursive procedure becomes

$$T_{m-1} = Q_mR_m, \quad T_m = R_mQ_m. \tag{2.3}$$

Now define

$$\hat{Q}_m = Q_1 Q_2 \dots Q_m, \quad \hat{R}_m = R_m R_{m-1} \dots R_1. \tag{2.4}$$

This is known as the QR algorithm and is completely analogous to the finite-dimensional case. Note also that we have $T_n = \hat{Q}_n^* T \hat{Q}_n$. In the finite-dimensional case and under favourable conditions, $\hat{Q}_n^* T \hat{Q}_n$ converges to a diagonal operator and the columns of $\hat{Q}_n$ converge to the corresponding eigenvectors as $n \to \infty$ (see Theorem 3.1 below). We will see that the IQR algorithm behaves similarly for the extreme parts of the spectrum.

Definition 2.3 Let $T \in \mathcal{B}(\mathcal{H})$ be invertible and let $\{e_j\}$ be an orthonormal basis for $\mathcal{H}$. The sequences $\{\hat{Q}_j\}$ and $\{\hat{R}_j\}$ constructed as in (2.3) and (2.4) will be called a Q-sequence and an R-sequence of T with respect to $\{e_j\}$.

Remark 2.4 Note that since the Householder transformations used in the proof of Theorem 2.2 are unique up to a ± sign, we will with some abuse of language refer to the QR decomposition constructed as *the* QR decomposition. In general for an invertible operator, the IQR algorithm is uniquely defined up to phase—see Sect. 4.2. This will not be a problem for our theorems or numerical examples.

The following observation will be useful in the later developments. From the construction in (2.3) and (2.4) we get

$$\begin{aligned} &T = Q_1 R_1 = \hat{Q}_1 \hat{R}_1, \\ &T^2 = Q_1 R_1 Q_1 R_1 = Q_1 Q_2 R_2 R_1 = \hat{Q}_2 \hat{R}_2, \\ &T^3 = Q_1 R_1 Q_1 R_1 Q_1 R_1 = Q_1 Q_2 R_2 Q_2 R_2 R_1 = Q_1 Q_2 Q_3 R_3 R_2 R_1 = \hat{Q}_3 \hat{R}_3. \end{aligned}$$

An easy induction gives us that

$$T^m = \hat{Q}_m \hat{R}_m. \tag{2.5}$$

Note that $\hat{R}_m$ must be upper triangular with respect to $\{e_j\}_{j \in \mathbb{N}}$ since $R_j,\ j \le m$ is upper triangular with respect to $\{e_j\}_{j \in \mathbb{N}}$. Also, if T is invertible then $\langle Re_i, e_i \rangle \neq 0$. From this it follows immediately that

$$\operatorname{span}\{T^m e_j\}_{j=1}^J = \operatorname{span}\{\hat{Q}_m e_j\}_{j=1}^J, \quad J \in \mathbb{N}. \tag{2.6}$$

3 Convergence theorems

In finite dimensions we have the following well-known theorem:

Theorem 3.1 (Finite dimensions) *Let $T \in \mathbb{C}^{N \times N}$ be a normal matrix with eigenvalues satisfying $|\lambda_1| > \dots > |\lambda_N|$. Let $\{Q_m\}$ be a Q-sequence of unitary operators. Then (up to re-ordering of the basis)*

$$Q_m^* T Q_m \longrightarrow \bigoplus_{j=1}^{N} \lambda_j e_j \otimes e_j, \qquad \textit{as } m \to \infty.$$

In this section we will address the convergence of the IQR algorithm for normal operators under similar assumptions and prove an analogue of Theorem 3.1 in infinite dimensions (Theorem 3.9). As well as this, and for more general operators T that are not necessarily normal, we address block convergence (Theorem 3.13), relevant when the eigenvalues do not have distinct moduli, and convergence to (dominant) invariant subspaces (Theorem 3.15).

3.1 Preliminary definitions and results

To state and prove our theorems we need some preliminary results. The reader only interested in the results themselves is referred to Sect. 3.2. If T is a normal operator, we will use $\chi_S(T)$ to denote the indicator function of the set S defined via the functional calculus. Without loss of generality, we deal with the Hilbert space $\mathcal{H} = l^2(\mathbb{N})$ and the canonical orthonormal basis $\{e_j\}_{j\in\mathbb{N}}$. Our first set of results concerns the convergence of spanning sets under power iterations and is analogous to the finite-dimensional case. The following proposition can be found in [36] and together with Lemma 3.6 below, these are the only results we will use from [36].

Proposition 3.2 *Suppose that $T \in \mathcal{B}(\mathcal{H})$ is normal, is invertible and that $\sigma(T) = \omega \cup \Psi$ is a disjoint union such that $\omega = \{\lambda_i\}_{i=1}^N$ consists of finitely many isolated eigenvalues of T with $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_N|$. Suppose further that* $\sup\{|z| : z \in \Psi\} < |\lambda_N|$. *Let $l \in \mathbb{N}$ and suppose that $\{\xi_i\}_{i=1}^l$ are linearly independent vectors in $\mathcal{H}$ such that $\{\chi_\omega(T)\xi_i\}_{i=1}^l$ are also linearly independent. Then*

(i) *The vectors $\{T^k\chi_\omega(T)\xi_i\}_{i=1}^l$ are linearly independent and there exists an l-dimensional subspace $B \subset \mathrm{ran}\chi_\omega(T)$ such that*

$$\mathrm{span}\{T^k\xi_i\}_{i=1}^l \to B, \quad \textit{as } k \to \infty.$$

(ii) *If*

$$\mathrm{span}\{T^k\xi_i\}_{i=1}^{l-1} \to D \subset \mathcal{H}, \quad \textit{as } k \to \infty,$$

where D is an $(l-1)$-dimensional subspace, then

$$\mathrm{span}\{T^k\xi_i\}_{i=1}^l \to D \oplus \mathrm{span}\{\xi\}, \quad \textit{as } k \to \infty,$$

where $\xi \in \mathrm{ran}\chi_\omega(T)$ is an eigenvector of T.

In order to extend this proposition to describe rates of convergence and prove our main theorems, we need to describe the space B in more detail. This is done inductively as follows. The first step is to choose $\nu_{1,1} \in \{\lambda_i\}_{i=1}^N$ of maximum modulus such that

$$\mathrm{span}\{\chi_{\nu_{1,1}}(T)\xi_1\} \neq \{0\}.$$

We then let $\xi_{1,1}$ be a linear multiple of ξ_1 such that $\chi_{\nu_{1,1}}(T)\xi_{1,1}$ has norm one. Now suppose that at the m-th stage we have constructed vectors $\{\xi_{m,i}\}_{i=1}^m$ with the same linear span as $\{\xi_i\}_{i=1}^m$ and such that there exist $\{\nu_{m,j}\}_{j=1}^{s_m} \subset \{\lambda_i\}_{i=1}^N$ with the following properties. After re-ordering the vectors $\{\xi_{m,i}\}_{i=1}^m$ if necessary, there exist integers $0 = k_{m,0} < k_{m,1} < k_{m,2} < \cdots < k_{m,s_m} = m$ such that

(1) $\left|\nu_{m,s_m}\right| < \left|\nu_{m,s_m-1}\right| < \cdots < \left|\nu_{m,1}\right|$.
(2) $\chi_\lambda(T)\xi_{m,i} = 0$ if $i > k_{m,j}$ and $\lambda \in \{\lambda_i\}_{i=1}^N$ has $|\lambda| > \left|\nu_{m,j+1}\right|$.
(3) $\{\chi_{\nu_{m,j}}(T)\xi_{m,i}\}_{i=k_{m,j-1}+1}^{k_{m,j}}$ are orthonormal.

We seek to add the space spanned by the vector ξ_{m+1} whilst preserving these properties.

First we deal with (2). Let $\eta_{m+1} \in \{\lambda_i\}_{i=1}^N$ be of maximal modulus such that $\chi_{\{\lambda_1,\ldots,\eta_{m+1}\}}(T)\xi_{m+1} \notin \mathrm{span}\{\chi_{\{\lambda_1,\ldots,\eta_{m+1}\}}(T)\xi_j\}_{j=1}^m$. If $|\eta_{m+1}| < \left|\nu_{m,1}\right|$ then let $t(m+1)$ be maximal such that $|\eta_{m+1}| < \left|\nu_{m,t(m+1)}\right|$. We then choose complex numbers $\{a_{m,j}\}_{j=1}^{k_{m,t(m+1)}}$ such that writing

$$\tilde{\xi}_{m+1,m+1} = \xi_{m+1} + \sum_{j=1}^{k_{m,t(m+1)}} a_{m,j}\xi_{m,j}$$

we have that $\chi_\lambda(T)\tilde{\xi}_{m+1,m+1} = 0$ if $\lambda \in \{\lambda_i\}_{i=1}^N$ has $|\lambda| > |\eta_{m+1}|$. Note that by (2), (3) and the definition of η_{m+1}, the coefficients $a_{m,j}$ are determined uniquely in terms of $\{\xi_{m,i}\}_{i=1}^{k_{m,t(m+1)}}$. If $|\eta_{m+1}| \geq \left|\nu_{m,1}\right|$ then let $t(m+1) = 0$ and we set $\tilde{\xi}_{m+1,m+1} = \xi_{m+1}$. In this case we still have that $\chi_\lambda(T)\tilde{\xi}_{m+1,m+1} = 0$ if $\lambda \in \{\lambda_i\}_{i=1}^N$ has $|\lambda| > |\eta_{m+1}|$.

We then define $\xi_{m+1,j} = \xi_{m,j}$ for $1 \leq j \leq m$ and now deal with (3). If $\eta_{m+1} \notin \{\nu_{m,j}\}_{j=1}^{s_m}$ then let $\xi_{m+1,m+1}$ be a linear multiple of $\tilde{\xi}_{m+1,m+1}$ such that $\chi_{\eta_{m+1}}(T)\xi_{m+1,m+1}$ has norm 1 and we let $\{\nu_{m+1,j}\}_{j=1}^{s_m+1}$ be a re-ordering of $\{\nu_{m,j}\}_{j=1}^{s_m} \cup \{\eta_{m+1}\}$. Otherwise, we have $\eta_{m+1} = \nu_{m,t(m+1)+1}$ and we apply Gram-Schmidt to

$$\{\chi_{\nu_{m,t(m+1)+1}}(T)\xi_{m+1,i}\}_{i=k_{m,t(m+1)}+1}^{k_{m,t(m+1)+1}} \cup \{\chi_{\nu_{m,t(m+1)+1}}(T)\tilde{\xi}_{m+1,m+1}\}$$

(without changing $\{\xi_{m+1,i}\}_{i=k_{m,t(m+1)}+1}^{k_{m,t(m+1)+1}}$). Note that by (2) and the definition of η_{m+1} these vectors are linearly independent. This gives $\xi_{m+1,m+1}$ such that

$$\{\chi_{\nu_{m,t(m+1)+1}}(T)\xi_{m+1,i}\}_{i=k_{m,t(m+1)}+1}^{k_{m,t(m+1)+1}} \cup \{\chi_{\nu_{m,t(m+1)+1}}(T)\xi_{m+1,m+1}\}$$

are orthonormal and $\chi_\lambda(T)\xi_{m+1,m+1} = 0$ if $\lambda \in \{\lambda_i\}_{i=1}^N$ has $|\lambda| > \left|\nu_{m,t(m+1)+1}\right|$. After re-ordering indices if necessary, we see that (1)-(3) now hold for $m+1$.

After l steps the above process terminates giving a new basis $\{\tilde{\xi}_i\}_{i=1}^l = \{\xi_{l,i}\}_{i=1}^l$ for $\operatorname{span}\{\xi_i\}_{i=1}^l$ along with $\{\nu_j\}_{j=1}^n = \{\nu_{l,j}\}_{j=1}^n \subset \{\lambda_i\}_{i=1}^N$ and $0 = k_0 < k_1 < k_2 < \cdots < k_n = l$ such that

(i) $|\nu_n| < |\nu_{n-1}| < \cdots < |\nu_1|$.
(ii) $\chi_\lambda(T)\tilde{\xi}_i = 0$ if $i > k_j$ and $\lambda \in \{\lambda_i\}_{i=1}^N$ has $|\lambda| > \left|\nu_{j+1}\right|$.
(iii) $\{\chi_{\nu_j}(T)\tilde{\xi}_i\}_{i=k_{j-1}+1}^{k_j}$ are orthonormal.

The subspace B can then be described as

$$B = \bigoplus_{j=1}^n \operatorname{span}\{\chi_{\nu_j}(T)\tilde{\xi}_i\}_{i=k_{j-1}+1}^{k_j}.$$

Definition 3.3 With respect to the above construction we define the following:

$$E_j := \operatorname{span}\{\chi_{\nu_j}(T)\tilde{\xi}_i\}_{i=k_{j-1}+1}^{k_j}, \quad Z(T, \{\xi_j\}_{j=1}^l) := \Big(\sum_{i=1}^l (\|\tilde{\xi}_i\|^2 - 1)\Big)^{\frac{1}{2}}. \quad (3.1)$$

Since the Gram-Schmidt process is defined uniquely up to phases we see that $Z(T, \{\xi_j\}_{j=1}^l)$ is well-defined. The above construction also shows that if $\{\chi_\omega(T)\xi_i\}_{i=1}^{l+1}$ are linearly independent then

$$Z(T, \{\xi_j\}_{j=1}^{l+1}) \geq Z(T, \{\xi_j\}_{j=1}^l).$$

We can now prove the following refinement of Proposition 3.2:

Proposition 3.4 *Suppose the assumptions of Proposition 3.2 hold. Let $J \leq N$ be minimal such that $\{\chi_{\{\lambda_1,\ldots,\lambda_J\}}(T)\xi_i\}_{i=1}^l$ are linearly independent. Set*

$$\rho = \sup\{|z| : z \in \Psi \cup \{\lambda_{J+1}, \ldots, \lambda_N\}\},$$
$$r = \max\{|\lambda_2/\lambda_1|, \ldots, |\lambda_J/\lambda_{J-1}|, \rho/|\lambda_J|\}.$$

Then $r < 1$ and $\delta(B, \operatorname{span}\{T^k\xi_i\}_{i=1}^l) \leq Z(T, \{\xi_j\}_{j=1}^l)r^k$. Since the spaces are l-dimensional, it follows from (1.5) that we have the convergence rate

$$\hat{\delta}(B, \operatorname{span}\{T^k\xi_i\}_{i=1}^l) \leq Z(T, \{\xi_j\}_{j=1}^l)l^{\frac{1}{2}}r^k.$$

Proof Consider the subspaces

$$E_j^k = \operatorname{span}\{T^k\tilde{\xi}_i\}_{i=k_{j-1}+1}^{k_j}.$$

Let $\zeta = \sum_{i=k_{j-1}+1}^{k_j} \alpha_i \chi_{\nu_j}(T)\tilde{\xi}_i \in E_j$ be a unit vector (hence $\sum_{i=k_{j-1}+1}^{k_j} |\alpha_i|^2 = 1$) and consider

$$\eta_k = \sum_{i=k_{j-1}+1}^{k_j} \alpha_i T^k \tilde{\xi}_i / \nu_j^k \in E_j^k.$$

By construction, we have for any such $\tilde{\xi}_i$ in the above sum that

$$\tilde{\xi}_i = (\chi_{\nu_j}(T) + \chi_{\theta_j}(T))\tilde{\xi}_i, \qquad \theta_j = \{\lambda \in \sigma(T) : |\lambda| < |\nu_j|\}.$$

This gives $T^k\tilde{\xi}_i = \nu_j^k \chi_{\nu_j}(T)\tilde{\xi}_{j,i} + T^k \chi_{\theta_j}(T)\tilde{\xi}_i$. Now, by the assumption on $\sigma(T)$, we have

$$\rho_j = \sup\{|z| : z \in \theta_j\} < |\nu_j|.$$

Thus, since

$$\|T^k \chi_{\theta_j}(T)\tilde{\xi}_i\| / |\nu_j^k| < |\rho_j/\nu_j|^k \|\chi_{\theta_j}(T)\tilde{\xi}_i\|,$$

we have

$$\|\zeta - \eta_k\| \leq |\rho_j/\nu_j|^k \sum_{i=k_{j-1}+1}^{k_j} |\alpha_i| \, \|\chi_{\theta_j}(T)\tilde{\xi}_i\| \leq \Big(\sum_{i=k_{j-1}+1}^{k_j} (\|\tilde{\xi}_i\|^2 - 1) \Big)^{\frac{1}{2}} r^k.$$

Here we have used Hölder's inequality together with the fact that $\|\chi_{\theta_j}(T)\tilde{\xi}_i\|^2 = \|\tilde{\xi}_i\|^2 - 1$ by orthonormality of $\{\chi_{\nu_j}(T)\tilde{\xi}_i\}_{i=k_{j-1}+1}^{k_j}$. The right-hand side gives an upper bound for $\delta(E_j, E_j^k)$. Analogous rates of convergence hold for the other subspaces and from (1.6) we have

$$\delta(B, \operatorname{span}\{T^k\tilde{\xi}_i\}_{i=1}^l) \leq Z(T, \{\xi_j\}_{j=1}^l) r^k, \tag{3.2}$$

since the spaces E_j are orthogonal. □

For the rest of this section we shall assume the following:

(A1) $T \in \mathcal{B}(\mathcal{H})$ is an invertible normal operator and $\{e_j\}_{j\in\mathbb{N}}$ an orthonormal basis for $\mathcal{H}$. $\{Q_k\}$ and $\{R_k\}$ are Q- and R-sequences of T with respect to the basis $\{e_j\}_{j\in\mathbb{N}}$.

(A2) $\sigma(T) = \omega \cup \Psi$ such that $\omega \cap \Psi = \emptyset$ and $\omega = \{\lambda_i\}_{i=1}^N$, where the λ_is are isolated eigenvalues with (possibly infinite) multiplicity m_i. Let $M = m_1 + \cdots + m_N = \dim(\operatorname{ran}\chi_\omega(T))$ and suppose that $|\lambda_1| > \ldots > |\lambda_N|$. Suppose further that $\sup\{|\theta| : \theta \in \Psi\} < |\lambda_N|$.

To apply Propositions 3.2 and 3.4 to prove the main result Theorem 3.9, we need to take care of the case that some of the e_j may have $\chi_\omega(T)e_j = 0$.

Definition 3.5 Suppose that (A1) and (A2) hold and let $K \in \mathbb{N} \cup \{\infty\}$ be minimal with the property that $\dim(\mathrm{span}\{\chi_\omega(T)e_j\}_{j=1}^K) = M$. Define

$$\begin{aligned}
\Lambda_\omega &= \{e_j : \chi_\omega(T)e_j \neq 0, j \leq K\},\\
\Lambda_\Psi &= \{e_j : \chi_\omega(T)e_j = 0, j \leq K\},\\
\tilde{\Lambda}_\omega &= \{e_j \in \Lambda_\omega : \chi_\omega(T)e_j \in \mathrm{span}\{\chi_\omega(T)e_i\}_{i=1}^{j-1}\}.
\end{aligned}$$

Define also the corresponding subset $\{\hat{e}_j\}_{j=1}^M \subset \{e_j\}_{j=1}^K$ such that $\{\hat{e}_j\}_{j=1}^M = \Lambda_\omega \backslash \tilde{\Lambda}_\omega$ and such that upon writing $\hat{e}_j = e_{p_j}$, the p_j are increasing.

Note that we have the following decomposition of T into

$$T = \left(\sum_{j=1}^M \lambda_{c_j}\, \xi_j \otimes \bar{\xi}_j\right) \oplus \chi_\Psi(T)T, \quad \lambda_{c_j} \in \omega,$$

where $\{\xi_j\}_{j=1}^M$ is an orthonormal set of eigenvectors of T. The following simple lemma extends Lemma 39 in [36] to infinite M but the proof is verbatim so omitted.

Lemma 3.6 *If $e_m \in \Lambda_\Psi \cup \tilde{\Lambda}_\omega$, then*

$$\mathrm{span}\{\chi_\omega(T)q_{k,j}\}_{j=1}^m = \mathrm{span}\{\chi_\omega(T)\hat{q}_{k,j}\}_{j=1}^{s(m)}, \quad q_{k,j} = Q_k e_j, \quad \hat{q}_{k,j} = Q_k\hat{e}_j,$$

where $s(m)$ is the largest integer such that $\{\hat{e}_j\}_{j=1}^{s(m)} \subset \{e_j\}_{j=1}^m$.

The following theorem is the key step of the proof of Theorem 3.9 and concerns convergence to the eigenvectors of T.

Theorem 3.7 *Assume (A1) and (A2) and define*

$$\rho = \sup\{|z| : z \in \Psi\}, \quad r = \max\{|\lambda_2/\lambda_1|, \ldots, |\lambda_N/\lambda_{N-1}|, \rho/|\lambda_N|\}.$$

Then there exists a collection of orthonormal eigenvectors $\{\hat{q}_j\}_{j=1}^M \subset \mathrm{ran}\chi_\omega(T)$ of T and collections of constants $A(m)$, $B(j)$ and $C(\mu)$ such that

(a) *If $e_m \in \Lambda_\Psi \cup \tilde{\Lambda}_\omega$ and μ is maximal with $p_\mu < m$ (recall that $\hat{e}_j = e_{p_j}$), then we have*

$$\left\|\chi_\omega(T)q_{k,m}\right\| \leq A(m)Z(T, \{\hat{e}_j\}_{j=1}^\mu)r^k. \tag{3.3}$$

In the case that $m < p_1$, we interpret this as $\left\|\chi_\omega(T)q_{k,m}\right\| = 0$ which holds from Lemma 3.6.

(b) *For any $j < M + 1$,*

$$\hat{\delta}(\mathrm{span}\{\hat{q}_j\}, \mathrm{span}\{\hat{q}_{k,j}\}) \leq B(j)Z(T, \{\hat{e}_i\}_{i=1}^j)r^k. \tag{3.4}$$

(c) *For any* $\mu < M+1$,

$$\delta(\mathrm{span}\{\hat{q}_{j,k}\}_{j=1}^{\mu}, \mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu}) \leq C(\mu)Z(T, \{\hat{e}_j\}_{j=1}^{\mu})r^k \tag{3.5}$$

and hence

$$\hat{\delta}(\mathrm{span}\{\hat{q}_{j,k}\}_{j=1}^{\mu}, \mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu}) \leq \mu^{\frac{1}{2}}C(\mu)Z(T, \{\hat{e}_j\}_{j=1}^{\mu})r^k. \tag{3.6}$$

Here, as in Lemma 3.6, $q_{k,j} = Q_k e_j$ *and* $\hat{q}_{k,j} = Q_k\hat{e}_j$*. Finally, if* M *is finite then we must have* $\mathrm{span}\{\hat{q}_j\}_{j=1}^{M} = \mathrm{ran}\,\chi_\omega(T)$.

We will provide an inductive proof of Theorem 3.7 which requires the following for the inductive step of part (a).

Lemma 3.8 *Assume the conditions in the statement of Theorem 3.7. Suppose also that (b) in Theorem 3.7 holds for* $j = 1, \ldots, \mu$ *and that (c) holds for a given* $\mu < M$*. Let* $e_{p_{\mu+1}} = \hat{e}_{\mu+1}$*, then if* $e_m \in \Lambda_\Psi \cup \tilde{\Lambda}_\omega$*, where* $m < p_{\mu+1}$*, (3.3) also holds with*

$$A(m) = \left\{\sum_{j=1}^{\mu}[C(\mu)+B(j)]^2\right\}^{\frac{1}{2}} + C(\mu).$$

Proof First note that from (2.6), invertibility of T and the fact that $\{\chi_\omega(T)\hat{e}_j\}_{j=1}^{\mu}$ are linearly independent, it must hold that $\{\chi_\omega(T)\hat{q}_{k,j}\}_{j=1}^{\mu}$ are linearly independent also. Then by using the assumptions stated and the fact that $\chi_\omega(T)\hat{q}_j = \hat{q}_j$ we have

$$\begin{aligned}\delta(\mathrm{span}\{\chi_\omega(T)\hat{q}_{k,j}\}_{j=1}^{\mu}, \mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu}) &\leq \delta(\mathrm{span}\{\hat{q}_{k,j}\}_{j=1}^{\mu}, \mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu})\\ &\leq C(\mu)Z(T, \{\hat{e}_j\}_{j=1}^{\mu})r^k.\end{aligned}$$

Also, we have that $s(m) \leq \mu$ and Lemma 3.6 implies

$$\mathrm{span}\{\chi_\omega(T)q_{k,j}\}_{j=1}^{m} = \mathrm{span}\{\chi_\omega(T)\hat{q}_{k,j}\}_{j=1}^{s(m)} \subset \mathrm{span}\{\chi_\omega(T)\hat{q}_{k,j}\}_{j=1}^{\mu}.$$

Using the fact that $\left\|\chi_\omega(T)q_{k,m}\right\| \leq 1$ and the definition of δ (along with the fact that $\mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu}$ is finite-dimensional), it follows that there exists some $v_k = \sum_{j=1}^{\mu}\beta_{j,k}\hat{q}_j \in \mathrm{span}\{\hat{q}_j\}_{j=1}^{\mu}$ with $\|v_k\| \leq 1$ and

$$\left\|\chi_\omega(T)q_{k,m} - v_k\right\| \leq C(\mu)Z(T, \{\hat{e}_j\}_{j=1}^{\mu})r^k. \tag{3.7}$$

We also have from assumption (b) that

$$\begin{aligned}\left|\langle\chi_\omega(T)q_{k,m}, \hat{q}_j\rangle\right| = \left|\langle q_{k,m}, \hat{q}_j\rangle\right| &\leq B(j)Z(T, \{\hat{e}_i\}_{i=1}^{j})r^k\\ &+\left|\langle q_{k,m}, \hat{q}_{k,j}\rangle\right| = B(j)Z(T, \{\hat{e}_i\}_{i=1}^{j})r^k,\end{aligned} \tag{3.8}$$

since $q_{k,m}$ is orthogonal to $\hat{q}_{k,j}$. This together with (3.7) gives that $\left|\beta_{j,k}\right| \leq \left[C(\mu) + B(j)\right] Z(T, \{\hat{e}_j\}_{j=1}^{\mu}) r^k$. Hence we must have

$$\|v_k\| \leq \left\{ \sum_{j=1}^{\mu} \left[C(\mu) + B(j)\right]^2 \right\}^{\frac{1}{2}} Z(T, \{\hat{e}_j\}_{j=1}^{\mu}) r^k.$$

Using (3.7) again then gives the result. Note that we have used orthonormality of $\{\hat{q}_j\}_{j=1}^{\mu}$ which will be proven as part of the induction. □

Proof of Theorem 3.7 We begin with the initial step of the induction for (b) and (c). Note that (a) trivially holds by construction with $A(m) = 0$ for any $m < p_1$ where $e_{p_1} = \hat{e}_1$ and this provides the initial step for (a).

By Propositions 3.2 and 3.4, there exists a unit eigenvector $\hat{q}_1 \in \operatorname{ran}\chi_{\omega}(T)$ such that

$$\delta(\operatorname{span}\{\hat{q}_1\}, \operatorname{span}\{T^k \hat{e}_1\}) \leq Z(T, \{\hat{e}_1\}) r^k.$$

Since $\operatorname{span}\{T^k \hat{e}_1\} \subset \operatorname{span}\{T^k e_i\}_{i=1}^{p_1}$, this implies that

$$\delta(\operatorname{span}\{\hat{q}_1\}, \operatorname{span}\{T^k e_i\}_{i=1}^{p_1}) \leq Z(T, \{\hat{e}_1\}) r^k.$$

Thus, it follows that

$$\delta(\operatorname{span}\{\hat{q}_1\}, \operatorname{span}\{q_{k,i}\}_{i=1}^{p_1}) = \delta(\operatorname{span}\{\hat{q}_1\}, \operatorname{span}\{T^k e_i\}_{i=1}^{p_1}) \leq Z(T, \{\hat{e}_1\}) r^k, \tag{3.9}$$

from (2.6). Note that $\{q_{k,i}\}_{i=1}^{p_1}$ are orthonormal (recall that Q_k is unitary) and hence by (3.9) there exists some coefficients $\alpha_{k,i}$ with $\sum_{i=1}^{p_1} |\alpha_{k,i}|^2 \leq 1$ such that defining $\tilde{\eta}_k = \sum_{i=1}^{p_1} \alpha_{k,i} q_{k,i}$ we have

$$\left\|\hat{q}_1 - \tilde{\eta}_k\right\| \leq Z(T, \{\hat{e}_1\}) r^k. \tag{3.10}$$

If $e_m \in \Lambda_{\Psi} \cup \tilde{\Lambda}_{\omega}$, where $m < p_1$ then by Lemma 3.6 $\langle q_{k,m}, \hat{q}_1 \rangle = 0$. It follows that we must have

$$\delta(\operatorname{span}\{\hat{q}_1\}, \operatorname{span}\{\hat{q}_{k,1}\}) \leq \left\|\hat{q}_1 - \alpha_{k,p_1} \hat{q}_{k,1}\right\| \leq Z(T, \{\hat{e}_1\}) r^k.$$

Hence we can take $B(1) = 1$ and $C(1) = 1$ in (b) and (c) respectively which completes the initial step.

For the induction step we will argue simultaneously for (a), (b) and (c) using induction on μ. Suppose that (a) holds for $m < p_{\mu}$ with $e_{p_{\mu}} = \hat{e}_{\mu}$ together with (b) and (c) for $j \leq \mu$ and some $\mu < M$. Let $e_{p_{\mu+1}} = \hat{e}_{\mu+1}$ then we can use Lemma 3.8 to extend (a) to all $m < p_{\mu+1}$ and this provides the step for (a). For (b), we note that Propositions 3.2 and 3.4 imply that

$$\begin{aligned}&\delta\left(\operatorname{span}\{\hat{q}_i\}_{i=1}^{\mu}\oplus\operatorname{span}\{\xi\},\operatorname{span}\{T^k\hat{e}_i\}_{i=1}^{\mu+1}\right),\\&\quad\le Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k,\quad \xi\in\operatorname{ran}\chi_\omega(T),\end{aligned}\tag{3.11}$$

where ξ is a unit eigenvector of T. We may also assume without loss of generality that ξ is orthogonal to $\hat{q}_j$ for $j=1,\ldots,\mu$. As before, since $\operatorname{span}\{T^k\hat{e}_i\}_{i=1}^{\mu+1}\subset\operatorname{span}\{T^k e_i\}_{i=1}^{p_{\mu+1}}$ we have

$$\delta(\operatorname{span}\{\hat{q}_i\}_{i=1}^{\mu}\oplus\operatorname{span}\{\xi\},\operatorname{span}\{T^k e_i\}_{i=1}^{p_{\mu+1}})\le Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k,$$

and hence by invertibility of T

$$\begin{aligned}&\delta(\operatorname{span}\{\hat{q}_i\}_{i=1}^{\mu}\oplus\operatorname{span}\{\xi\},\operatorname{span}\{q_{k,i}\}_{i=1}^{p_{\mu+1}})\\&\quad=\delta(\operatorname{span}\{\hat{q}_i\}_{i=1}^{\mu}\oplus\operatorname{span}\{\xi\},\operatorname{span}\{T^k e_i\}_{i=1}^{p_{\mu+1}})\\&\quad\le Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k.\end{aligned}\tag{3.12}$$

Again, using that$\{q_{k,i}\}_{i=1}^{p_{\mu+1}}$ are orthonormal, there exists some coefficients $\alpha_{k,i}$ with $\sum_{i=1}^{p_{\mu+1}}|\alpha_{k,i}|^2\le 1$ such that defining $\tilde{\eta}_k=\sum_{i=1}^{p_{\mu+1}}\alpha_{k,i}q_{k,i}$ we have

$$\|\xi-\tilde{\eta}_k\|\le Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k.\tag{3.13}$$

If $e_m\in\Lambda_\Psi\cup\tilde{\Lambda}_\omega$, where $m<p_{\mu+1}$ then as shown above we have

$$\left|\langle q_{k,m},\xi\rangle\right|=\left|\langle\chi_\omega(T)q_{k,m},\xi\rangle\right|\le A(m)Z(T,\{\hat{e}_j\}_{j=1}^{\mu})r^k\le A(m)Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k.$$

Taking the inner product of $\xi-\tilde{\eta}_k$ with $q_{k,m}$ and using (3.13) together with the orthonormality of the $q_{k,j}$s, it follows that $\left|\alpha_{k,m}\right|\le\big(A(m)+1\big)Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k$. Similarly, if $j\le\mu$ then for any $c\in\mathbb{C}$

$$\left|\langle\hat{q}_{k,j},\xi\rangle\right|\le\left|\langle c\hat{q}_j,\xi\rangle\right|+\left|c\hat{q}_j-\hat{q}_{k,j}\right|=\left|c\hat{q}_j-\hat{q}_{k,j}\right|,$$

since ξ is orthogonal to $\hat{q}_j$. Minimising over c, we can bound this by $B(j)Z(T,\{\hat{e}_j\}_{j=1}^{\mu})r^k$. In the same way, it then follows that $|\alpha_{k,p_j}|\le\big(B(j)+1\big)Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k$ where $\hat{e}_j=e_{p_j}$. Together, these imply that

$$\begin{aligned}\left\|\xi-\alpha_{k,p_{\mu+1}}\hat{q}_{k,\mu+1}\right\|\le\Bigg[1+\Bigg\{&\sum_{m=1,e_m\in\Lambda_\Psi\cup\tilde{\Lambda}_\omega}^{p_{\mu+1}}\big[A(m)+1\big]^2\\&+\sum_{j=1}^{\mu}[B(j)+1]^2\Bigg\}^{\frac{1}{2}}\Bigg]Z(T,\{\hat{e}_j\}_{j=1}^{\mu+1})r^k.\end{aligned}$$

To finish the inductive step, we define $\hat{q}_{\mu+1} = \xi$. Recall that ξ is orthogonal to any $\hat{q}_l$ with $l \leq \mu$. Hence it follows that $\{\hat{q}_i\}_{i=1}^{\mu+1}$ are orthonormal and we can take

$$B(\mu+1) = 1 + \left\{ \sum_{m=1, e_m \in \Lambda_\Psi \cup \tilde{\Lambda}_\omega}^{p_{\mu+1}} \left[A(m)+1\right]^2 + \sum_{j=1}^{\mu} \left[B(j)+1\right]^2 \right\}^{\frac{1}{2}}$$

in (b). For the induction step for (c), the fact that $\{\hat{q}_{k,i}\}_{i=1}^{\mu+1}$ are orthonormal and (1.6) imply we can take

$$C(\mu+1) = \left(\sum_{j=1}^{\mu+1} B(j)^2 \right)^{\frac{1}{2}}.$$

Finally, if M is finite we demonstrate that $\operatorname{span}\{\hat{q}_j\}_{j=1}^M = \operatorname{span}\{\xi_j\}_{j=1}^M$. Since the $\{\hat{q}_i\}_{i=1}^M$ are orthogonal and are eigenvectors of $\sum_{j=1}^M \lambda_{c_j} \xi_j \otimes \bar{\xi}_j$, it follows that $\operatorname{span}\{\hat{q}_j\}_{j=1}^M = \operatorname{span}\{\xi_j\}_{j=1}^M = \operatorname{ran}\chi_\omega(T)$. □

3.2 Main results

Our first result generalises Theorem 3.1 to infinite dimensions and relies on Theorem 3.7 (which concerns convergence to eigenvectors).

Theorem 3.9 (Convergence theorem for normal operators in infinite dimensions) *Let $T \in \mathcal{B}(l^2(\mathbb{N}))$ be an invertible normal operator with $\sigma(T) = \omega \cup \Psi$ and $\omega = \{\lambda_i\}_{i=1}^N$, where the λ_i's are isolated eigenvalues with (possibly infinite) multiplicity m_i satisfying $|\lambda_1| > \cdots > |\lambda_N|$. Suppose further that* $\sup\{|\theta| : \theta \in \Psi\} < |\lambda_N|$, *and let $\{e_j\}_{j\in\mathbb{N}}$ be the canonical orthonormal basis. Let $\{Q_n\}_{n\in\mathbb{N}}$ and $\{R_n\}_{n\in\mathbb{N}}$ be Q- and R-sequences of T with respect to $\{e_j\}_{j\in\mathbb{N}}$. Let $\{\hat{e}_j\}_{j=1}^M \subset \{e_j\}_{j\in\mathbb{N}}$, where $M = m_1 + \cdots + m_N$, be the subset described in Definition 3.5 and Theorem 3.7, i.e.* $\operatorname{span}\{Q_k\hat{e}_j\} \to \operatorname{span}\{\hat{q}_j\}$ *where $\{\hat{q}_j\}_{j=1}^M \subset \operatorname{ran}\chi_\omega(T)$ is a collection of orthonormal eigenvectors of T and if $e_j \notin \{\hat{e}_j\}_{j=1}^M$, then $\chi_\omega(T)Q_k e_j \to 0$. Then:*

(i) *Every subsequence of $\{Q_n^* T Q_n\}_{n\in\mathbb{N}}$ has a convergent subsequence $\{Q_{n_k}^* T Q_{n_k}\}_{k\in\mathbb{N}}$ such that*

$$Q_{n_k}^* T Q_{n_k} \xrightarrow{WOT} \left(\bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right) \bigoplus \sum_{j\in\Theta} \xi_j \otimes e_j,$$

as $k \to \infty$, where

$$\Theta = \{j : e_j \notin \{\hat{e}_l\}_{l=1}^M\}, \quad \xi_j \in \overline{\operatorname{span}\{e_i\}_{i\in\Theta}}$$

and only $\sum_{j\in\Theta}\xi_j\otimes e_j$ depends on the choice of subsequence. Furthermore, if T has only finitely many non-zero entries in each column then we can replace WOT convergence by SOT convergence.

(ii) *We have the following convergence of sections:*

$$\widehat{P}_M Q_n^* T Q_n \widehat{P}_M \xrightarrow{SOT} \bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j, \qquad \text{as } n\to\infty,$$

where $\widehat{P}_M$ denotes the orthogonal projection onto $\overline{\mathrm{span}\{\hat{e}_j\}_{j=1}^M}$. Furthermore, if we define

$$\rho = \sup\{|z| : z\in\Psi\}, \quad r = \max\{|\lambda_2/\lambda_1|, \ldots, |\lambda_N/\lambda_{N-1}|, \rho/|\lambda_N|\}$$

then $r < 1$ and for any fixed $x \in \mathrm{span}\{\hat{e}_j\}_{j=1}^M$ we have the following rate of convergence

$$\left\| \widehat{P}_M Q_n^* T Q_n \widehat{P}_M x - \left(\bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right) x \right\| = O(r^n), \quad \text{as } n\to\infty. \tag{3.14}$$

If M is finite then we can write (after possibly re-ordering)

$$\bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j = \bigoplus_{k=1}^{N} \left(\lambda_k \bigoplus_{j=1+\sum_{l<k} m_l}^{\sum_{l\leq k} m_l} \hat{e}_j \otimes \hat{e}_j \right), \tag{3.15}$$

and in part (ii) we have the rate of convergence

$$\left\| \widehat{P}_M Q_n^* T Q_n \widehat{P}_M - \bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right\| = O(r^n), \quad \text{as } n\to\infty. \tag{3.16}$$

If $\{\chi_\omega(T)e_l\}_{l=1}^M$ are linearly independent, then we can take $\hat{e}_j = e_j$.

Remark 3.10 What Theorem 3.9 essentially says is that if we take the n-th iteration of the IQR algorithm and truncate to an $m\times m$ matrix (i.e. $P_m Q_n^* T Q_n P_m$) then, as n grows, the eigenvalues of this matrix will converge to the *extremal parts* of the spectrum of T. In particular, the theorem suggests that the IQR algorithm can locate the *extremal parts* of the spectrum.

Proof of Theorem 3.9 To prove (i), since a closed ball in $\mathcal{B}(l^2(\mathbb{N}))$ is weakly sequentially compact, it follows that any subsequence of $\{Q_n^* T Q_n\}_{n\in\mathbb{N}}$ must have a weakly convergent subsequence $\{Q_{n_k}^* T Q_{n_k}\}_{k\in\mathbb{N}}$. In particular, there exists a $W \in \mathcal{B}(l^2(\mathbb{N}))$ such that

$$Q_{n_k}^* T Q_{n_k} \xrightarrow{\text{WOT}} W, \qquad k\to\infty.$$

Let $\widehat{P}_M$ denote the projection onto $\overline{\mathrm{span}\{\hat{e}_j\}_{j=1}^M}$. Note that part (i) of the theorem will follow if we can show that

$$\widehat{P}_M W \widehat{P}_M = \bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j, \tag{3.17}$$

and

$$\widehat{P}_M^{\perp} W \widehat{P}_M = 0, \quad \widehat{P}_M W \widehat{P}_M^{\perp} = 0.$$

We will indeed show this, and we start by observing that, due to the weak convergence and the standard functional calculus, we have that

$$\langle W\hat{e}_j, e_i\rangle = \lim_{k\to\infty} \langle TQ_{n_k}\hat{e}_j, \chi_\omega(T)Q_{n_k}e_i\rangle + \lim_{k\to\infty} \langle TQ_{n_k}\hat{e}_j, \chi_\Psi(T)Q_{n_k}e_i\rangle, \tag{3.18}$$
$$\langle We_i, \hat{e}_j\rangle = \lim_{k\to\infty} \langle \chi_\omega(T)Q_{n_k}e_i, T^*Q_{n_k}\hat{e}_j\rangle + \lim_{k\to\infty} \langle TQ_{n_k}e_i, \chi_\Psi(T)Q_{n_k}\hat{e}_j\rangle. \tag{3.19}$$

We then have the following

$$\chi_\omega(T)Q_n e_i \to 0, \quad n\to\infty, \quad i\in\Theta$$
$$\Longrightarrow \begin{cases} \lim_{k\to\infty}\langle TQ_{n_k}\hat{e}_j, \chi_\omega(T)Q_{n_k}e_i\rangle = 0, & i\in\Theta, \\ \lim_{k\to\infty}\langle \chi_\omega(T)Q_{n_k}e_i, T^*Q_{n_k}\hat{e}_j\rangle = 0, & i\in\Theta, \end{cases} \tag{3.20}$$
$$\mathrm{span}\{Q_n\hat{e}_j\} \to \mathrm{span}\{\hat{q}_j\}, \quad n\to\infty, \quad T\hat{q}_j = \lambda\hat{q}_j,\ \lambda\in\omega,$$
$$\Longrightarrow \begin{cases} \lim_{k\to\infty}\langle TQ_{n_k}\hat{e}_j, \chi_\Psi(T)Q_{n_k}e_i\rangle = 0, & i\in\mathbb{N}, \\ \lim_{k\to\infty}\langle TQ_{n_k}e_i, \chi_\Psi(T)Q_{n_k}\hat{e}_j\rangle = 0, & i\in\mathbb{N}, \\ \lim_{k\to\infty}\langle TQ_{n_k}\hat{e}_j, \chi_\omega(T)Q_{n_k}\hat{e}_l\rangle = \delta_{j,l}\lambda. \end{cases} \tag{3.21}$$

Thus, by (3.18), (3.20), (3.21) and Theorem 3.7 we get (3.17) and also that $\widehat{P}_M^{\perp} W \widehat{P}_M = 0$. Also, by (3.19), (3.20), (3.21) and Theorem 3.7 we get that $\widehat{P}_M W \widehat{P}_M^{\perp} = 0$. Note that in all of these cases, Theorem 3.7 implies that the rate of convergence is such that the difference between $\langle W\hat{e}_j, e_i\rangle$, $\langle We_i, \hat{e}_j\rangle$ and their limiting values is $O(r^{n_k})$ (however, not necessarily uniformly over the indices). Now suppose that T has finitely many non-zero entries in each column. This can be described by a function $f:\mathbb{N}\to\mathbb{N}$ non-decreasing with $f(n)\geq n$ such that $\langle Te_j, e_i\rangle = 0$ when $i > f(j)$ as in Definition 4.1. Proposition 4.2 shows that this is preserved under the iteration in the IQR algorithm, i.e. $Q_{n_k}^* T Q_{n_k}$ also has this property. So let $x\in l^2(\mathbb{N})$ and $\epsilon > 0$. Choose y of finite support such that $\|x-y\|\leq\epsilon$. It is then clear that $\|Q_{n_k}^* TQ_{n_k}y - Wy\|\to 0$ as $n_k\to\infty$ (since we only require convergence in finitely many entries). Hence

$$\limsup_{n_k\to\infty} \|Q_{n_k}^* TQ_{n_k}x - Wx\| \leq (\|T\| + \|W\|)\epsilon.$$

Since $\epsilon > 0$ and x were arbitrary, we have $Q_{n_k}^* TQ_{n_k} \xrightarrow{\mathrm{SOT}} W$.

To prove (ii), suppose that $x \in \text{span}\{\hat{e}_j\}_{j=1}^M$, then x can be written as

$$x = \sum_{j=1}^{M} x_j \hat{e}_j,$$

with at most finitely many x_j non-zero. We have that $\hat{\delta}(\text{span}\{Q_n\hat{e}_j\}, \text{span}\{\hat{q}_j\}) = O(r^n)$ and hence there exists some $a_{n,j}$ of unit modulus such that $\|Q_n\hat{e}_j - a_{n,j}\hat{q}_j\| = O(r^n)$. Since Q_n is unitary, we then have

$$\begin{aligned}
&\left\| \widehat{P}_M Q_n^* T Q_n \widehat{P}_M x - \left(\bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right) x \right\| \\
&\leq \left\| Q_n^* T Q_n \widehat{P}_M x - \left(\bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right) Q_n^* Q_n x \right\| \\
&= \left\| \sum_{j=1}^{M} x_j (T - \langle T\hat{q}_j, \hat{q}_j\rangle I) Q_n \hat{e}_j \right\| = O(r^n),
\end{aligned}$$

where we have used the fact that T is bounded in the last line. We therefore have convergence on $\text{span}\{\hat{e}_j\}_{j=1}^M$, and, since the operators are uniformly bounded, we must have convergence on $\overline{\text{span}\{\hat{e}_j\}_{j=1}^M}$ which implies that

$$\widehat{P}_M Q_n^* T Q_n \widehat{P}_M \xrightarrow{\text{SOT}} \bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j, \qquad \text{as } n \to \infty.$$

For the last parts, suppose that M is finite. Theorem 3.7 then implies (3.15) after a possible re-ordering. The rate of convergence in (3.14) also implies that

$$\left\| \widehat{P}_M Q_n^* T Q_n \widehat{P}_M - \bigoplus_{j=1}^{M} \langle T\hat{q}_j, \hat{q}_j\rangle \hat{e}_j \otimes \hat{e}_j \right\| = O(r^n).$$

More generally, let $K \in \mathbb{N} \cup \{\infty\}$ be minimal such that $\dim(\text{span}\{\chi_\omega(T)e_j\}_{j=1}^K) = M$. Recall that we defined

$$\begin{aligned}
&\Lambda_\omega = \{e_j : \chi_\omega(T)e_j \neq 0, j \leq K\}, \quad \Lambda_\Psi = \{e_j : \chi_\omega(T)e_j = 0, j \leq K\} \\
&\text{and } \tilde{\Lambda}_\omega = \{e_j \in \Lambda_\omega : \chi_\omega(T)e_j \in \text{span}\{\chi_\omega(T)e_i\}_{i=1}^{j-1}\}.
\end{aligned}$$

Recall also from the proof of Theorem 3.7 that $\{\hat{e}_j\}_{j=1}^M = \Lambda_\omega \backslash \tilde{\Lambda}_\omega$. If $\{\chi_\omega(T)e_j\}_{j=1}^M$ are linearly independent then $\tilde{\Lambda}_\omega = \emptyset$, and therefore $\{\hat{e}_j\}_{j=1}^M = \{e_j\}_{j=1}^M$, which yields that the projection $\widehat{P}_M$ in (3.17) is the projection onto $\overline{\text{span}\{e_j\}_{j=1}^M}$. □

Theorems 3.9 and 3.7 also give us convergence to the eigenvectors. With the use of (possibly countably many) shifts and rotations, the above theorem allows us to find all eigenvalues, their multiplicities and eigenspaces outside the convex hull of the essential spectrum, i.e. outside the essential numerical range.

Example 3.11 It is possible in the case of infinite M that the $\hat{q}_j$ do not form an orthonormal basis of $\operatorname{ran}\chi_\omega(T)$ and we can even lose part of ω in the convergence of $\widehat{P}_M Q_n^* T Q_n \widehat{P}_M$ to a diagonal operator. This is to be contrasted to the finite-dimensional case. For example, suppose that with respect to an initial orthonormal basis $\{v_j\}_{j\in\mathbb{N}}$, T is given by the diagonal matrix $\mathrm{Diag}(1/2, 1, 1, \ldots)$. Now define $f_j = v_1 + (1/j)v_{j+1}$ and apply Gram-Schmidt to the sequence $\{f_j\}_{j\in\mathbb{N}}$ to generate orthonormal vectors $\{e_j\}_{j\in\mathbb{N}}$. It is easy to see that any v_j can be approximated to arbitrary accuracy using finite linear combinations of e_j and hence $\{e_j\}_{j\in\mathbb{N}}$ is an orthonormal basis of our Hilbert space. We also have that the $\chi_1(T)(f_j) = (1/j)v_{j+1}$ are linearly independent and hence so are $\chi_1(T)(e_j)$. It follows that the IQR iterates converge in the strong operator topology to the identity operator. However, we could equally take $\omega = \{1, 1/2\}$ in Theorem 3.9. Hence we have the curious case that $\overline{\mathrm{span}\{\hat{q}_j\}_{j\in\mathbb{N}}} \subset \overline{\mathrm{span}\{\hat{v}_j\}_{j>1}}$ and we lose the eigenvalue $1/2$.

The following corollary is entirely analogous to the finite-dimensional case.

Corollary 3.12 *Suppose that the conditions of Theorem 3.9 hold with M finite. Suppose also that for $j = 1, \ldots, N$ the vectors $\{\chi_{\{\lambda_1,\ldots,\lambda_j\}}(T)e_i\}_{i=1}^{\sum_{l\leq j} m_l}$ are linearly independent. In the notation of Theorem 3.9, let $\rho = \sup\{|z| : z \in \Psi\}$. For $j < N$ define $r_j = \max\{|\lambda_{k+1}/\lambda_k| : k \leq j\}$ and for $j = N$ define $r_N = \max\{|\lambda_{k+1}/\lambda_k|, |\lambda_N/\rho| : k \leq j\}$. We then have the following rates of convergence to the diagonal operator for $i, j \leq M$:*

1. $\left|\langle Q_n^* T Q_n e_j, e_i\rangle\right| = O(r_k^n)$ *as $n \to \infty$ if $i > j$ and k is minimal such that $i \leq \sum_{l\leq k} m_l$,*
2. $\left|\langle Q_n^* T Q_n e_i, e_i\rangle - \lambda_k\right| = O(r_k^n)$ *as $n \to \infty$ if k is minimal such that $i \leq \sum_{l\leq k} m_l$.*

Proof The result follows from Theorem 3.9 applied successively to $\omega_1, \omega_2, \ldots, \omega_N$ where $\omega_j = \{\lambda_k : k \leq j\}$. In general, analogous results follows from Theorem 3.9 when M is infinite and with other linear independence conditions on $\chi_{\omega'}(T)e_i$ with $\omega' \subset \omega$ but the statements become less succinct. □

In the finite-dimensional case and the case of distinct eigenvalues of the *same magnitude*, the QR algorithm applied to a normal matrix will ‘converge’ to a block diagonal matrix (without necessarily converging in each block). This can be extended to infinite dimensions by inductively using the following theorem which also extends to *non-normal* operators.

Theorem 3.13 (Block convergence theorem in infinite dimensions) *Let $T \in \mathcal{B}(l^2(\mathbb{N}))$ be an invertible operator (not necessarily normal) and suppose that there exists an orthogonal projection P of rank M (possibly infinite) such that both the ranges of P and of $I - P$ are invariant under T. Suppose also that there exists $\alpha > \beta > 0$ such that*

- $\|Tx\| \geq \alpha\|x\| \quad \forall x \in \text{ran}(P)$,
- $\|Tx\| \leq \beta\|x\| \quad \forall x \in \text{ran}(I-P)$.

Let $\{Q_n\}_{n\in\mathbb{N}}$ and $\{R_n\}_{n\in\mathbb{N}}$ be Q- and R-sequences of T with respect to $\{e_i\}$. Then there exists a subset $\{\hat{e}_j\}_{j=1}^M \subset \{e_i\}_{i\in\mathbb{N}}$ such that

(i) *For any finite $\mu \leq M$ we have $\delta(\text{span}\{Q_n\hat{e}_j\}_{j=1}^{\mu}, \text{ran}(P)) = O(\beta^n/\alpha^n)$ as $n \to \infty$. If M is finite this implies full convergence $\hat{\delta}(\text{span}\{Q_n\hat{e}_j\}_{j=1}^{M}, \text{ran}(P)) = O(\beta^n/\alpha^n)$ as $n \to \infty$.*

(ii) *Every subsequence of $\{Q_n^*TQ_n\}_{n\in\mathbb{N}}$ has a convergent subsequence $\{Q_{n_k}^*TQ_{n_k}\}_{k\in\mathbb{N}}$ such that*

$$Q_{n_k}^*TQ_{n_k} \xrightarrow{WOT} \sum_{j=1}^{M} \xi_j \otimes \hat{e}_j \bigoplus \sum_{i\in\Theta} \zeta_i \otimes e_i,$$

as $k \to \infty$, where

$$\Theta = \{j : e_j \notin \{\hat{e}_l\}_{l=1}^M\}, \quad \xi_j \in \overline{\text{span}\{\hat{e}_l\}_{l=1}^M}, \quad \zeta_i \in \overline{\text{span}\{e_l\}_{l\in\Theta}}.$$

If $\{Pe_l\}_{l=1}^M$ are linearly independent then we can take $\hat{e}_j = e_j$. Furthermore, if T has only finitely many non-zero entries in each column then we can replace WOT convergence by SOT convergence.

Remark 3.14 Theorem 3.13 essentially says that the IQR algorithm can compute the invariant subspace $\text{ran}(P)$ of such an operator if there is enough separation between T restricted to $\text{ran}(P)$ and $\text{ran}(I-P)$. In other words, provided the existence of a *dominant* invariant subspace.

Proof of Theorem 3.13 The main ideas of the proof of Theorem 3.13 have already been presented so we sketch the proof. We first define the vectors $\{\hat{e}_j\}_{j=1}^M$ in a similar way to Definition 3.5 inductively by $\hat{e}_j = e_{p_j}$ where

$$p_j = \min\{i : Pe_i \notin \text{span}\{P\hat{e}_k\}_{k=1}^{j-1}\}.$$

Let $r = \beta/\alpha < 1$. We will prove inductively that

(a) $\hat{\delta}(\text{span}\{Q_n\hat{e}_j\}_{j=1}^{\mu}, \text{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu}) \leq C_1(\mu)r^n$ for any finite $\mu \leq M$,

(b) $\|PQ_ne_j\| \leq C_2(j)r^n$ for any $j \in \Theta$,

for some constants $C_1(\mu)$ and $C_2(j)$. Suppose that this has been done. Part (i) of Theorem 3.13 now follows since $\text{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu} \subset \text{ran}(P)$. We then argue as in the proof of Theorem 3.9 to gain

$$Q_{n_k}^*TQ_{n_k} \xrightarrow{\text{WOT}} W, \qquad k \to \infty.$$

Then by studying the inner products $\langle TQ_{n_k}e_j, Q_{n_k}e_i\rangle$ using the invariance of $\text{ran}(P)$, $\text{ran}(I-P)$ under T and from (b), part (ii) of Theorem 3.13 easily follows (note that (a) implies that $\|(I-P)Q_n\hat{e}_j\| \leq C_1(j)r^n$). The final part of the theorem then follows from the same arguments in the proof of Theorem 3.9. Hence we only need to prove (a) and (b).

We first claim that

$$\delta(\mathrm{span}\{PT^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}) \leq C_3(\mu)r^n. \tag{3.22}$$

P commutes with T which is invertible and hence both of these spaces have dimension μ by the construction of the $\hat{e}_j$. It follows that (3.22) implies

$$\hat{\delta}(\mathrm{span}\{PT^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}) \leq \mu^{\frac{1}{2}}C_3(\mu)r^n = C_4(\mu)r^n. \tag{3.23}$$

To show (3.22), let $x_1^n, \ldots, x_\mu^n$ be an orthonormal basis for $\mathrm{span}\{PT^n\hat{e}_j\}_{j=1}^{\mu}$ and let $\xi = \sum_{j=1}^{\mu}\alpha_j x_j^n$ have norm at most 1. Now, we may choose coefficients $\beta_{j,n}$ such that $T^n\sum_{j=1}^{\mu}\beta_{j,n}x_j^n = \xi$ since $T|_{\mathrm{ran}(P)}$ is invertible when viewed as an operator acting on $\mathrm{ran}(P)$. By the assumptions on T we must have that

$$\left(\sum_{j=1}^{m}|\beta_{j,n}|^2\right)^{1/2} \leq \frac{1}{\alpha^n}.$$

We may change basis from $\{\hat{e}_j\}_{j=1}^{\mu}$ to $\{\tilde{e}_j\}_{j=1}^{\mu}$ such that $P\tilde{e}_j = x_j^n$. Form the vector

$$\eta_n = T^n\left(\sum_{j=1}^{\mu}\beta_{j,n}\tilde{e}_j\right) \in \mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}.$$

Then clearly by Hölder's inequality

$$\|\xi - \eta_n\| \leq \frac{\left(\sum_{j=1}^{\mu}\|T^n(I-P)\tilde{e}_j\|^2\right)^{1/2}}{\alpha^n} \leq C_3(\mu)\frac{\beta^n}{\alpha^n},$$

proving (3.22) and hence (3.23).

Note that the proof of Lemma 3.6 carries over (replacing the projection $\chi_\omega(T)$ by P) to prove that

$$\mathrm{span}\{PQ_ne_j\}_{j=1}^{m} = \mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{s(m)} \tag{3.24}$$

where $s(m)$ is maximal with $\{\hat{e}_j\}_{j=1}^{s(m)} \subset \{e_j\}_{j=1}^{m}$. It follows that

$$\begin{aligned}\delta(\mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu}) &= \delta(\mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{PQ_ne_j\}_{j=1}^{p_\mu})\\ &= \delta(\mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{PT^ne_j\}_{j=1}^{p_\mu})\end{aligned}$$

$$\begin{aligned} &\leq \delta(\mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{PT^n\hat{e}_j\}_{j=1}^{\mu}) \\ &\leq C_4(\mu)r^n, \end{aligned}$$

where we have used (2.6) to reach the second line and the fact that $\mathrm{span}\{PT^n\hat{e}_j\}_{j=1}^{\mu} \subset \mathrm{span}\{PT^n e_j\}_{j=1}^{p_\mu}$ to reach the third line. Again, both spaces have dimension μ so we have

$$\begin{aligned} \delta(\mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{Q_n e_j\}_{j=1}^{p_\mu}) &= \delta(\mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{T^n e_j\}_{j=1}^{p_\mu}) \\ &\leq \delta(\mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu}, \mathrm{span}\{T^n\hat{e}_j\}_{j=1}^{\mu}) \\ &\leq C_5(\mu)r^n. \end{aligned} \tag{3.25}$$

With these arguments out of the way (these are the analogue of Proposition 3.4) we can now form our inductive argument, similar to the proof of Theorem 3.7. Suppose first that (a) holds for μ (allowing $\mu = 0$ for the initial step) and let $j \in \Theta$ have $j < p_{\mu+1}$ (where $p_{\mu+1} = \infty$ if $\mu = M$). From (a) for μ and (3.24) we have that

$$PQ_n e_j = v_n + \sum_{i=1}^{\mu} a_{n,i} Q_n\hat{e}_i$$

for some v_n with $\|v_n\| \leq C_1(\mu)r^n$. Then we must have

$$a_{n,i} + \langle v_n, Q_n\hat{e}_i\rangle = \langle PQ_n e_j, Q_n\hat{e}_i\rangle = \langle Q_n e_j, PQ_n\hat{e}_i\rangle.$$

Using (a) again, along with the fact that $Q_n e_j$ is orthogonal to $\{Q_n\hat{e}_i\}_{i=1}^{\mu}$, we must have $\left|a_{n,i}\right| \leq 2C_1(\mu)r^n$. It follows that we can take $C_2(j) = (2\sqrt{\mu}+1)C_1(\mu)$ for $j \in [p_\mu + 1, \dots, p_{\mu+1})$ in (b). Now we use (3.25). Let $\xi \in \mathrm{span}\{PQ_n\hat{e}_j\}_{j=1}^{\mu+1}$ have unit norm and assume that $p_{\mu+1} < \infty$ (else there is nothing to prove since then $\mu = M$). Then there exists $b_{n,j}$ and w_n such that

$$\xi = \sum_{j=1}^{p_{\mu+1}} b_{n,j} Q_n e_j + w_n$$

and $\|w_n\| \leq C_5(\mu+1)r^n$. Now let $j \in \Theta$ with $j < p_{\mu+1}$ then we must have

$$\langle \xi, PQ_n e_j\rangle = \langle \xi, Q_n e_j\rangle = b_{n,j} + \langle w_n, Q_n e_j\rangle.$$

We have proven (b) for such j and hence we have $\left|b_{n,j}\right| \leq \big(C_2(j) + C_5(\mu+1)\big)r^n$. It follows that we can take

$$C_1(\mu+1) = \mu^{\frac{1}{2}}\left[C_5(\mu+1) + \left\{\sum_{j=1, j\in\Theta}^{p_{\mu+1}} [C_2(j) + C_5(\mu+1)]^2\right\}^{\frac{1}{2}}\right],$$

where the square root factor appears since the relevant spaces are μ-dimensional. This completes the inductive step (the initial step is identical) and hence the proof of the theorem. □

Theorem 3.13 can be made sharper (under a slightly stricter assumption on the linear independence of $\{e_j\}_{j=1}^M$) with the following theorem which includes the case that $\mathrm{ran}(I-P)$ is not necessarily invariant.

Theorem 3.15 (Convergence to invariant subspace in infinite dimensions) *Let $T \in \mathcal{B}(l^2(\mathbb{N}))$ be an invertible operator (not necessarily normal) and suppose that there exists an orthogonal projection P of finite rank M such that the range of P is invariant under T. Suppose also that there exists $\alpha > \beta > 0$ such that*

- $\|Tx\| \geq \alpha\|x\| \quad \forall x \in \mathrm{ran}(P)$,
- $\|(I-P)T(I-P)\| \leq \beta$.

Under these conditions, there exists a canonical M-dimensional T^*-invariant subspace S and we let $\tilde{P}$ denote the orthogonal projection onto S (in the special case that $\mathrm{ran}(I-P)$ is also T-invariant such as in Theorems 3.9 and 3.13, then $S = \mathrm{ran}(P)$). Suppose also that $\{\tilde{P}e_j\}_{j=1}^M$ are linearly independent. Let $\{Q_n\}_{n\in\mathbb{N}}$ and $\{R_n\}_{n\in\mathbb{N}}$ be Q- and R-sequences of T with respect to $\{e_i\}$. Then

(i) *The subspace angle $\phi(\mathrm{span}\{e_j\}_{j=1}^M, S) < \pi/2$ and we have*

$$\begin{aligned}&\hat{\delta}(\mathrm{span}\{Q_n e_j\}_{j=1}^M, \mathrm{ran}(P))\\ &\quad\leq \frac{\sin\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, \mathrm{ran}(P))\right)}{\cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right)}\frac{\beta^n}{\alpha^n}\left(1+\frac{\|PT(I-P)\|}{\alpha-\beta}\right),\end{aligned} \tag{3.26}$$

(ii) *Every subsequence of $\{Q_n^* T Q_n\}_{n\in\mathbb{N}}$ has a convergent subsequence $\{Q_{n_k}^* T Q_{n_k}\}_{k\in\mathbb{N}}$ such that*

$$Q_{n_k}^* T Q_{n_k} \overset{WOT}{\longrightarrow} \sum_{j=1}^M \xi_j \otimes e_j \bigoplus \sum_{i=M+1}^\infty \zeta_i \otimes e_i,$$

as $k \to \infty$, where

$$\xi_j \in \overline{\mathrm{span}\{e_l\}_{l=1}^M}, \quad \zeta_i \in \mathcal{H}.$$

Furthermore, if T has only finitely many non-zero entries in each column then we can replace WOT convergence by SOT convergence.

Remark 3.16 Theorem 3.15 says that the IQR algorithm can be used to approximate dominant invariant subspaces. In particular, we shall use the bound (3.26) to build a Δ_1 algorithm in Sect. 5. Note in the normal case that Theorem 3.9 is more precise, both in giving convergence of individual vectors to eigenvectors and in the less restrictive assumptions on spanning sets and M. In the normal case (and that of Theorem 3.13) we also have that the limit operator has a block diagonal form.

3.3 Proof of Theorem 3.15

In this section we will prove Theorem 3.15. The proof technique is different from those used above, and hence we have given it a separate section. Throughout, we will denote the ratio β/α by r. Note that since M is finite, the bound α implies that $T|_{\mathrm{ran}(P)}: \mathrm{ran}(P) \rightarrow \mathrm{ran}(P)$ is invertible with $\|T|_{\mathrm{ran}(P)}^{-1}\| \leq 1/\alpha$. First, let Q denote a unitary change of basis matrix from $\{e_j\}$ to $\{\tilde{e}_j\}$ where $\{\tilde{e}_j\}_{j=1}^{M}$ is a basis for $\mathrm{ran}(P)$. Then as matrices with respect to the original basis we can write

$$Q = [P_1, P_2], \quad Q^* T Q = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix},$$

where $T_{11} \in \mathbb{C}^{M\times M}$ and T_{12} has M rows. Our assumptions imply that $\|T_{11}^{-1}\| \leq 1/\alpha$ and $\|T_{22}\| \leq \beta$. The next lemma shows that we can change the basis further to eliminate the sub-block T_{12}. This is needed to apply a power iteration type argument.

Lemma 3.17 *Define the linear function $F : \mathcal{B}(l^2(\mathbb{N}), \mathbb{C}^M) \rightarrow \mathcal{B}(l^2(\mathbb{N}), \mathbb{C}^M)$ by*

$$F(A) = T_{11}^{-1} A T_{22},$$

where we identify elements of $\mathcal{B}(l^2(\mathbb{N}), \mathbb{C}^M)$ as matrices. Then we can define $A \in \mathcal{B}(l^2(\mathbb{N}), \mathbb{C}^M)$ by $A - F(A) = -T_{11}^{-1}T_{12}$. Furthermore, if we define

$$B(A) = \begin{pmatrix} I & A \\ 0 & I \end{pmatrix},$$

then $B(A)$ has inverse $B(-A)$ and

$$B(-A)\begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} B(A) = \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix}. \tag{3.27}$$

Proof Our assumptions on T ensure that F is a contraction with $\|F\| \leq r < 1$. Hence we can define A via the series

$$A = \sum_{k=0}^{\infty} F^k(-T_{11}^{-1}T_{12}).$$

It is then straightforward to check $A - F(A) = -T_{11}^{-1}T_{12}$, $B(A)B(-A) = B(-A)B(A) = I$ and the identity (3.27). □

Let

$$Y = Q\begin{pmatrix} I & 0 \\ -A^* & I \end{pmatrix}$$

then we have the matrix identity

$$Y^{-1}T^*Y = \begin{pmatrix} T_{11}^* & 0 \\ 0 & T_{22}^* \end{pmatrix}.$$

The canonical T^*-invariant subspace alluded to in Theorem 3.15 is then simply $S = \operatorname{span}\{Ye_j\}_{j=1}^M$. The space is canonical since it is easily seen that it is unchanged if we use a different basis for $\operatorname{ran}(P_1)$ and $\operatorname{ran}(P_2)$ in the definition of Q.

Now let $P_0 = \begin{pmatrix} e_1 & e_2 & \dots & e_M \end{pmatrix} \in \mathcal{B}(\mathbb{C}^M, l^2(\mathbb{N}))$ denote the matrix whose columns are the first M basis elements $\{e_j\}_{j=1}^M$. Since the $\{R_i\}$ are upper triangular, it is easy to see that

$$T^n P_0 = Q_n R_n P_0 = Q_n P_0 P_0^* R_n P_0.$$

We will denote the (invertible) matrix $P_0^* R_n P_0 \in \mathbb{C}^{M\times M}$ by Z_n. Now define

$$V_n^1 = P_1^* Q_n P_0 \in \mathcal{B}(\mathbb{C}^M), \quad V_n^2 = P_2^* Q_n P_0 \in \mathcal{B}(\mathbb{C}^M, l^2(\mathbb{N})),$$

then we have the relation

$$\begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}^n \begin{pmatrix} V_0^1 \\ V_0^2 \end{pmatrix} = \begin{pmatrix} V_n^1 \\ V_n^2 \end{pmatrix} Z_n.$$

But by Lemma 3.17 we have

$$\begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}^n = B(A) \begin{pmatrix} T_{11}^n & 0 \\ 0 & T_{22}^n \end{pmatrix} B(-A).$$

Unwinding the definitions, this implies the matrix identities

$$T_{11}^n (V_0^1 - AV_0^2) = (V_n^1 - AV_n^2) Z_n, \tag{3.28}$$

$$T_{22}^n V_0^2 = V_n^2 Z_n. \tag{3.29}$$

Lemma 3.18 *The following identity holds*

$$\hat{\delta}(\operatorname{span}\{Q_n e_j\}_{j=1}^M, \operatorname{ran}(P)) = \|V_n^2\|. \tag{3.30}$$

Proof Note that $\operatorname{span}\{Q_n e_j\}_{j=1}^M = \operatorname{ran}(Q_n P_0)$ and $\operatorname{ran}(P) = \operatorname{ran}(P_1)$. Since $P_1 P_1^*$ and $Q_n P_0 P_0^* Q_n^*$ are orthogonal projections, it follows that

$$\begin{aligned} \hat{\delta}(\operatorname{span}\{Q_n e_j\}_{j=1}^M, \operatorname{ran}(P)) &= \|Q_n P_0 P_0^* Q_n^* - P_1 P_1^*\| \\ &= \|Q_n^*(Q_n P_0 P_0^* Q_n^* - P_1 P_1^*) Q\| \\ &= \left\| \begin{pmatrix} 0 & P_0^* Q_n^* P_2 \\ -(I - P_0)^* Q_n^* P_1 & 0 \end{pmatrix} \right\|. \end{aligned}$$

But we have that $\|P_0^* Q_n^* P_2\| = \|V_n^2\|$ and hence we are done if we can show $\|P_0^* Q_n^* P_2\| = \|(I - P_0)^* Q_n^* P_1\|$. Consider the unitary matrix

$$U := Q_n^* Q = \begin{pmatrix} P_0^* Q_n^* P_1 & P_0^* Q_n^* P_2 \\ (I - P_0)^* Q_n^* P_1 & (I - P_0)^* Q_n^* P_2 \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}.$$

Now let $x \in \mathbb{C}^M$ be of unit norm, then $\|U_{11}x\|^2 + \|U_{21}x\|^2 = 1$. It follows that $\|U_{21}\|^2 = 1 - \sigma_0(U_{11})^2$, where σ_0 denotes the smallest singular value. Applying the same argument to U^* we see that $\|U_{12}\|^2 = 1 - \sigma_0(U_{11})^2 = \|U_{21}\|^2$, completing the proof. □

Lemma 3.19 *The matrix* $(V_0^1 - AV_0^2)$ *is invertible with*

$$\|(V_0^1 - AV_0^2)^{-1}\| \leq \frac{1}{\cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right)}. \tag{3.31}$$

Proof First note that since $\{\tilde{P}e_j\}_{j=1}^M$ are linearly independent, we must have $\phi(\mathrm{span}\{e_j\}_{j=1}^M, S) < \pi/2$ and hence the bound in (3.31) is finite. Let $W = (P_1 - P_2 A^*)(I + AA^*)^{-1/2} \in \mathcal{B}(\mathbb{C}^M, l^2(\mathbb{N}))$. By considering $W^* W = I \in \mathbb{C}^{M \times M}$, we see that the columns of W are orthonormal. In fact, expanding Y we have

$$Y = [P_1 - P_2 A^* \quad P_2]$$

and hence the columns of W are a basis for the subspace S. Arguing as in the proof of Lemma 3.18, we have that

$$\hat{\delta}(\mathrm{span}\{e_j\}_{j=1}^M, S) = \sqrt{1 - \sigma_0(W^* P_0)^2} < 1.$$

This implies that $W^* P_0$ is invertible with

$$\sigma_0(W^* P_0) = \cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right) > 0.$$

We also have the identity

$$V_0^1 - AV_0^2 = (I + AA^*)^{1/2}(W^* P_0).$$

Since $(I + AA^*)^{-1/2}$ has norm at most 1, we see that $(V_0^1 - AV_0^2)$ is invertible and (3.31) holds. □

Proof of Theorem 3.15 Using Lemma 3.19 and the matrix identities (3.28) and (3.29), we can write

$$V_n^2 = T_{22}^n V_0^2 (V_0^1 - AV_0^2)^{-1} T_{11}^{-n} (V_n^1 - AV_n^2).$$

Using (3.30) and (3.31), this implies

$$\begin{aligned}\hat{\delta}(\mathrm{span}\{Q_n e_j\}_{j=1}^M, \mathrm{ran}(P)) &\leq \frac{\|V_0^2\|\|V_n^1 - AV_n^2\|r^n}{\cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right)} \\ &= \frac{\sin\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, \mathrm{ran}(P))\right)\|V_n^1 - AV_n^2\|r^n}{\cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right)}.\end{aligned} \tag{3.32}$$

It is clear by summing a geometric series that

$$\|A\| \leq \frac{\|T_{12}\|}{\alpha(1-r)} = \frac{\|PT(I-P)\|}{\alpha - \beta}.$$

It follows that $\|V_n^1 - AV_n^2\| \leq 1 + \|PT(I-P)\|/(\alpha - \beta)$. Substituting this into (3.32) proves part (i) of the theorem.

Next we argue that if $i > M$ then $\|PQ_n e_i\| \to 0$ as $n \to \infty$. We have that

$$PQ_n e_i = \sum_{j=1}^{M} \alpha_{j,n} Q_n e_j + v_n$$

with $\|v_n\| \to 0$ by part (i). Note that we then have

$$\alpha_{j,n} = \langle PQ_n e_i, Q_n e_j\rangle + \epsilon_{j,n} = \langle Q_n e_i, PQ_n e_j\rangle + \epsilon_{j,n}$$

with $\{\epsilon_{j,n}\}$ null. But again by (i) we have that $PQ_n e_j$ approaches $\mathrm{span}\{Q_n e_k\}_{k=1}^M$ which is orthogonal to $Q_n e_i$ and hence $\{\alpha_{j,n}\}$ is null. The proof of part (ii) now follows the same argument as in the proof of part (i) of Theorem 3.9 and of the final part of Theorem 3.13. The key property being that if $j \leq M$ and $i > M$ then $\langle Q_n^* T Q_n e_j, e_i\rangle \to 0$ due to the invariance of $\mathrm{ran}(P)$ under T. Note that it does not necessarily follow (as is easily seen by considering upper triangular T) that $\langle Q_n^* T Q_n e_i, e_j\rangle \to 0$ for such i, j. □

4 The IQR algorithm can be computed

The previous section gives a theoretical justification for why the IQR algorithm may work, but we are faced with the possibly unpleasant problem of how to compute with infinite data structures on a computer. Fortunately, there is a way to overcome such a problem. The key is to impose some structural requirements on the infinite matrix.

4.1 Quasi-banded subdiagonals

Definition 4.1 Let T be an infinite matrix acting as a bounded operator on $l^2(\mathbb{N})$ with basis $\{e_j\}_{j\in\mathbb{N}}$. For $f : \mathbb{N} \to \mathbb{N}$ non-decreasing with $f(n) \geq n$ we say that T has quasi-banded subdiagonals with respect to f if $\langle Te_j, e_i\rangle = 0$ when $i > f(j)$.

This is the class of infinite matrices with a finite number of non-zero elements in each column (and not necessarily in each row) which is captured by the function f. It is for this class that the computation of the IQR algorithm is feasible on a finite machine. For this class of operators one can actually compute (without any approximation or any extra discretisation) the matrix elements of the n-th iteration of the IQR algorithm as if it was done on an infinite computer (meaning the computation collapses to a finite one). The following result of independent interest is needed in the proof and generalises the well-known fact in finite dimensions that the QR algorithm preserves bandwidth (see [57] for a good discussion of the tridiagonal case).

Proposition 4.2 *Let $T \in \mathcal{B}(l^2(\mathbb{N}))$ and let T_n be the n-th element in the IQR iteration, such that $T_n = Q_n^* \cdots Q_1^* T Q_1 \cdots Q_n$, where*

$$Q_j = \underset{l\to\infty}{SOT\text{-}lim}\, U_1^j \cdots U_l^j$$

and U_l^j is a Householder transformation. If T has quasi-banded subdiagonals with respect to f then so does T_n.

Proof By induction, it is enough to prove the result for $n = 1$. From the construction of the Householder reflections $U_m^1 = P_{m-1} \oplus S_m$, the chosen η_m (see Theorem 2.2) have

$$\langle \eta_m, e_j \rangle = 0, \quad j > f(m). \tag{4.1}$$

Using the fact that f is increasing, it follows that each U_m^1 has quasi-banded subdiagonals with respect to f, as does the product $U_1^1 \cdots U_m^1$. It follows that Q_1 must have quasi-banded subdiagonals with respect to f and hence so does $T_1 = R_1 Q_1$ since R_1 is upper triangular. □

Theorem 4.3 *Let $T \in \mathcal{B}(l^2(\mathbb{N}))$ have quasi-banded subdiagonals with respect to f and let T_n be the n-th element in the IQR iteration, i.e. $T_n = Q_n^* \cdots Q_1^* T Q_1 \cdots Q_n$, where*

$$Q_j = \underset{l\to\infty}{SOT\text{-}lim}\, U_1^j \cdots U_l^j$$

and U_l^j is a Householder transformation (the superscript is not a power, but an index). Let P_m be the usual projection onto $\mathrm{span}\{e_j\}_{j=1}^m$ and denote the a-fold iteration of f by $\underbrace{f \circ f \circ \ldots \circ f}_{a\ times} = f_a$. Then

$$\begin{aligned} P_m T_n P_m = {} & P_m U_m^n \cdots U_1^n U_{f_1(m)}^{n-1} \cdots U_1^{n-1} \cdots U_{f_{(n-2)}(m)}^2 \cdots U_1^2 U_{f_{(n-1)}(m)}^1 \cdots U_1^1 \\ & \cdot P_{f_n(m)} T P_{f_n(m)} \\ & \cdot U_1^1 \cdots U_{f_{(n-1)}(m)}^1 U_1^2 \cdots U_{f_{(n-2)}(m)}^2 \cdots U_1^{n-1} \cdots U_{f_1(m)}^{n-1} U_1^n \cdots U_m^n P_m. \end{aligned} \tag{4.2}$$

Remark 4.4 What Theorem 4.3 says is that to compute the finite section of size m of the n-th iteration of the IQR algorithm (i.e. $P_m T_n P_m$), one only needs information from the finite section of size $f_n(m)$ (i.e. $P_{f_n(m)} T P_{f_n(m)}$) since the relevant Householder reflections can be computed from this information. In other words, the IQR algorithm can be computed.

Proof of Theorem 4.3 By induction it is enough to prove that

$$P_m T_n P_m = P_m U_m^n \dots U_1^n P_{f(m)} T_{n-1} P_{f(m)} U_1^n \dots, U_1^m P_m \tag{4.3}$$

To see why this is true, note that by the assumption that T has quasi-banded subdiagonals with respect to f, Proposition 4.2 shows that T_n has quasi-banded subdiagonals with respect to f for all $n \in \mathbb{N}$. Thus, it follows from the construction in the proof of Theorem 2.2 that each U_l^j is of the form

$$U_l^j = I_{l,j,1} \oplus \left(I_{l,j,2} - \frac{2}{\|\xi_{l,j}\|^2} \xi_{l,j} \otimes \bar{\xi}_{l,j} \right) \oplus I_{l,j,3},$$

where $I_{l,j,1}$ denotes the identity on $P_{l-1}\mathcal{H}$, $I_{l,j,2}$ denotes the identity on $\operatorname{span}\{e_k : l \leq k \leq f(l)\}$, $I_{l,j,3}$ denotes the identity on $P_{f(l)}^{\perp}\mathcal{H}$ and $\xi_{l,j} \in \operatorname{span}\{e_k : l \leq k \leq f(l)\}$. Since P_m is compact, it then follows that

$$\begin{aligned} P_m T_n P_m &= (\operatorname*{SOT-lim}_{l\to\infty} P_m U_l^n, \dots, U_1^n) P_{f(m)} T_{n-1} P_{f(m)} (\operatorname*{SOT-lim}_{l\to\infty} U_1^n, \dots, U_l^n P_m) \\ &= P_m U_m^n, \dots, U_1^n P_{f(m)} T_{n-1} P_{f(m)} U_1^n, \dots, U_1^m P_m. \end{aligned} \tag{4.4}$$

□

Remark 4.5 This result allows us to implement the IQR algorithm because each U_l^j only affects finitely many columns or rows of A if multiplied either on the left or the right. In computer science, it is often referred to as "Lazy evaluation" when one computes with infinite data structures, but defers the use of the information until needed. A simple implementation is shown in the appendix for the case that the matrix has k subdiagonals (i.e. we have $f(n) = n + k$).

The next question is how restrictive is the assumption in Definition 4.1? In particular, suppose that $T \in \mathcal{B}(\mathcal{H})$ and that $\xi \in \mathcal{H}$ is a cyclic vector for T (i.e. $\operatorname{span}\{\xi, T\xi, T^2\xi, \dots\}$ is dense in $\mathcal{H}$). Then by applying the Gram-Schmidt procedure to $\{\xi, T\xi, T^2\xi, \dots\}$ we obtain an orthonormal basis $\{\eta_1, \eta_2, \eta_3, \dots\}$ for $\mathcal{H}$ such that the matrix representation of T with respect to $\{\eta_1, \eta_2, \eta_3, \dots\}$ is upper Hessenberg, and thus the matrix representation has only one subdiagonal. The question is therefore about the existence of a cyclic vector. Note that if T does not have invariant subspaces then every vector $\xi \in \mathcal{B}(\mathcal{H})$ is a cyclic vector. Now what happens if ξ is not cyclic for T? We may still form $\{\eta_1, \eta_2, \eta_3, \dots\}$ as above, however, $\mathcal{H}_1 = \overline{\operatorname{span}\{\eta_1, \eta_2, \eta_3, \dots\}}$ is now an invariant subspace for T and $\mathcal{H}_1 \neq \mathcal{H}$. We may still form a matrix representation of T with respect to $\{\eta_1, \eta_2, \eta_3, \dots\}$, but this will now be a matrix representation of $T|_{\mathcal{H}_1}$. Obviously, we can have that $\sigma(T|_{\mathcal{H}_1}) \subsetneq \sigma(T)$.

However, the following example shows that the class of matrices for which we can compute the IQR algorithm covers a wide number of applications. In particular, it includes all finite interaction Hamiltonians on graphs. Such operators play a prominent role in solid state physics [48,51] describing propagation of waves and spin waves as well as encompassing Jacobi operators studied in many physical models and integrable lattices [71].

Example 4.6 Consider a connected, undirected graph G, such that each vertex degree is finite and the set of vertices $V(G)$ is countably infinite. Consider the set of all bounded operators A on $l^2(V(G)) \cong l^2(\mathbb{N})$ such that the set $S(v) := \{w \in V : \langle w, Av\rangle \neq 0\}$ is finite for any $v \in V$. Suppose our enumeration of the vertices obeys the following pattern. e_1's neighbours (including itself) are $S_1 = \{e_1, e_2, \dots, e_{q_1}\}$ for some finite q_1. The set of neighbours of these vertices is $S_2 = \{e_1, \dots, e_{q_2}\}$ for some finite q_2 where we continue the enumeration of S_1 and this process continues inductively enumerating S_m. If we know $S(v)$ for all $v \in V$ then we can find an $f : \mathbb{N} \rightarrow \mathbb{N}$ such that $A_{j,m} = 0$ if $|j| > f(m)$. We simply choose $f(n) = q_{r_n}$ where r_n is minimal such that $\cup_{j \leq n} S(e_j) \subset S_{r_n}$.

4.2 Invertible operators

More generally, given an invertible operator T with information on how its columns decay at infinity, we can compute finite sections of the IQR iterates with *error control*. For computing spectral properties, we can assume, by shifting $T \rightarrow T + \lambda I$ then translating by $-\lambda$ back, that the operator we are interested in is invertible, hence the invertibility criterion is not that restrictive. Throughout, we will use the following lemma which says that for invertible operators, the QR decomposition is essentially unique.

Lemma 4.7 *Let T be an invertible operator (viewed as a matrix acting on $l^2(\mathbb{N})$), then there exists a unique decomposition $T = QR$ with Q unitary and R invertible, upper triangular such that $R_{ii} \in \mathbb{R}_{>0}$. Furthermore, any other "QR" decomposition $T = Q'R'$ has a diagonal matrix $D = \mathrm{Diag}(t_1, t_2, \dots)$ such that $|t_i| = 1$ and $Q = Q'D$. In other words, the QR decomposition is unique up to phase choices.*

Proof Consider the QR decomposition already discussed in this paper, $T = Q''R''$. T is invertible, and hence Q'' is a surjective isometry so is unitary. Hence $R'' = Q''^{*}T$ is invertible. Being upper triangular, it follows that $R''_{ii} \neq 0$ for all i. Choose $t_i \in \mathbb{T}$ such that $t_i R''_{ii} \in \mathbb{R}_{>0}$ and set $D = \mathrm{Diag}(t_1, t_2, \dots)$. Letting $Q = Q''D^*$ and $R = DR''$ we clearly have the decomposition as claimed.

Now suppose that $T = Q'R'$ then we can write $Q = Q'R'R^{-1}$. It follows that $R'R^{-1}$ is a unitary upper triangular matrix and hence must be of the form $D = \mathrm{Diag}(t_1, t_2, \dots)$ with $|t_i| = 1$. □

Another way to see this result is to note that the columns of Q are obtained by applying the Gram–Schmidt procedure to the columns of T. The restriction that $R_{ii} \in \mathbb{R}_{>0}$ can also be incorporated into Theorem 4.3. Theorem 4.3 (in this subcase of invertibility) is then a consequence of the fact that if T has quasi-banded subdiagonals with respect to f then

$$P_m T^n P_m = P_m (P_{f_n(m)} T P_{f_n(m)})^n P_m$$

and the relations (2.5)—we can apply Gram-Schmidt (or a more stable modified version) to the columns of $P_{f_n(m)} T P_{f_n(m)}$ and truncate the resulting matrix.

Assume that given $T \in \mathcal{B}(l^2(\mathbb{N}))$ invertible (not necessarily with quasi-banded subdiagonals), we can evaluate an increasing family of increasing functions $g^j : \mathbb{N} \to \mathbb{N}$ such that defining the matrix $T_{(j)}$ with columns $\{P_{g^j(n)} T e_n\}$ we have that $T_{(j)}$ is invertible and

$$\left\| (P_{g^j(n)} - I) T e_n \right\| \leq \frac{1}{j}. \tag{4.5}$$

It is easy to see that such a sequence of functions must exist since any S with $\|S - T\| < \left\| T^{-1} \right\|^{-1}$ is invertible. Given this information, without loss of generality by increasing the g^js pointwise if necessary, applying Hölder's inequality and taking subsequences, we may assume that $\left\| T_{(j)} - T \right\| \leq 1/j$. In other words, given a sequence of functions satisfying (4.5) we can evaluate a sequence of functions with this stronger condition. The following says that given such a sequence of functions, we can compute the truncations $P_m T_n P_m$ to a given precision.

Theorem 4.8 *Suppose $T \in \mathcal{B}(l^2(\mathbb{N}))$ is invertible and the family of functions $\{g^j\}$ are as above. Suppose also that we are given a bound C such that $\|T\| \leq C$. Let $\epsilon > 0$ and $m, n \in \mathbb{N}$, then we can choose j such that applying Theorem 4.3 (with the diagonal operators to ensure $R_{ii} > 0$) to $T_{(j)}$ using the function g^j instead of f, we have the guaranteed bound*

$$\left\| P_m T_n P_m - P_m T_{(j),n} P_m \right\| \leq \epsilon,$$

where $T_{(j),n}$ denotes the n-th IQR iterate of $T_{(j)}$.

Proof of Theorem 4.8 First consider the error when applying Theorem 4.3 to $T_{(j)}$ with g^j for any fixed j. We will show that we can compute an error bound which converges to zero as $j \to \infty$ and from this the theorem easily follows by successively computing the bound and halting when this bound is less than ϵ.

Write the QR decompositions

$$T^n = \hat{Q}_n \hat{R}_n, \quad (T_{(j)})^n = \hat{Q}_{(j),n} \hat{R}_{(j),n}.$$

We have $\left\| T - T_{(j)} \right\| \leq 1/j$ and hence, by writing $T_{(j)} = T + (T_{(j)} - T)$, that

$$\left\| T^n - (T_{(j)})^n \right\| \leq \sum_{k=1}^{n} \binom{n}{k} \frac{1}{j^k} C^{n-k} \leq \frac{(C+1)^n}{j} = \frac{\tilde{C}}{j},$$

where $\tilde{C} = (C+1)^n$. The columns of $\hat{Q}_n$ and $\hat{Q}_{(j),n}$ are simply the columns of the matrices T^n and $(T_{(j)})^n$ after the application of Gram-Schmidt. Let the first m columns of T^n and $(T_{(j)})^n$ be denoted by $\{t_k\}_{k=1}^m$ and $\{\tilde{t}_k^j\}_{k=1}^m$ respectively and let $\{q_k\}_{k=1}^m$ and $\{\tilde{q}_k^j\}_{k=1}^m$ be the vectors obtained after applying Gram-Schmidt to these sequences of vectors. We then have

$$\begin{aligned}\|q_1 - \tilde{q}_1^j\| &= \left\| \frac{t_1}{\|t_1\|} - \frac{\tilde{t}_1^j}{\|\tilde{t}_1^j\|} \right\| \\ &= \left\| \frac{t_1(\|\tilde{t}_1^j\| - \|t_1\|)}{\|t_1\|\|\tilde{t}_1^j\|} - \frac{(\tilde{t}_1^j - t_1)\|t_1\|}{\|t_1\|\|\tilde{t}_1^j\|} \right\| \leq \frac{2\|t_1 - \tilde{t}_1^j\|}{\|\tilde{t}_1^j\|} \leq \frac{2\tilde{C}}{j\|\tilde{t}_1^j\|}.\end{aligned} \tag{4.6}$$

For a vector v of unit norm, let $P_{\perp v}$ denote the orthogonal projection onto the space of vectors perpendicular to v. Note that for two such vectors v, w, we have $\|P_{\perp v} - P_{\perp w}\| \leq \|v - w\|$. Let

$$v_k = P_{\perp q_{k-1}} \cdots P_{\perp q_1} t_k, \quad \tilde{v}_k^j = P_{\perp \tilde{q}_{k-1}^j} \cdots P_{\perp \tilde{q}_1^j} \tilde{t}_k^j, \tag{4.7}$$

then q_k are just the normalised version of v_k and likewise $\tilde{q}_k^j$ are just the normalised version of $\tilde{v}_k^j$. Suppose that for $\mu < k$ we have $\|q_\mu - \tilde{q}_\mu^j\| \leq \delta$ for some $\delta > 0$. Then applying the above products of projections we have

$$\begin{aligned}\|v_k - \tilde{v}_k^j\| &\leq \|P_{\perp q_{k-1}} \cdots P_{\perp q_1}(t_k - \tilde{t}_k^j)\| + \|P_{\perp q_{k-1}} \cdots P_{\perp q_1}\tilde{t}_k^j - \tilde{v}_k^j\| \\ &\leq \|t_k - \tilde{t}_k^j\| + \|P_{\perp q_{k-1}} \cdots P_{\perp q_1} - P_{\perp \tilde{q}_{k-1}^j} \cdots P_{\perp \tilde{q}_1^j}\|\|\tilde{t}_k^j\| \\ &\leq \|t_k - \tilde{t}_k^j\| + (k-1)\delta\|\tilde{t}_k^j\|.\end{aligned}$$

In the last line we have used the fact that if the operators $\{A_l\}_{l=1}^m$ and $\{B_l\}_{l=1}^m$ have norm bounded by 1, then

$$\left\| \prod_{l=1}^m A_l - \prod_{l=1}^m B_l \right\| \leq \sum_{l=1}^m \|A_l - B_l\|.$$

Applying the same argument as in the inequalities (4.6) we see that

$$\|q_k - \tilde{q}_k^j\| \leq \frac{2(\|t_k - \tilde{t}_k^j\| + (k-1)\delta\|\tilde{t}_k^j\|)}{\|\tilde{v}_k^j\|} \leq \frac{2(\tilde{C}/j + 2(k-1)\delta\tilde{C})}{\|\tilde{v}_k^j\|}, \tag{4.8}$$

since $\|\tilde{t}_k^j\| \leq C + \tilde{C}/j \leq 2\tilde{C}$. Now note that we can compute the $\|\tilde{v}_k^j\|$ from the proof of Theorem 4.3. Set $\delta_1(j) = \frac{2\tilde{C}}{j\|\tilde{t}_1^j\|}$ and for $1 < k \leq m$ define iteratively

$$\delta_k(j) = \max\left\{\delta_{k-1}(j), \frac{2(\tilde{C}/j + 2(k-1)\delta_{k-1}(j)\tilde{C})}{\|\tilde{v}_k^j\|}\right\}.$$

We must have $\|q_k - \tilde{q}_k^j\| \leq \delta_m(j)$ for $1 \leq k \leq m$ where we have now shown the j dependence as an argument.

It follows that $\|(\hat{Q}_n - \hat{Q}_{(j),n})P_m\| \leq \sqrt{m}\delta_m(j)$ and hence that

$$\begin{aligned}\|P_m T_n P_m - P_m T_{(j),n} P_m\| &\leq \|P_m(\hat{Q}_n - \hat{Q}_{(j),n})^* T \hat{Q}_n P_m\| \\ &\quad + \|P_m \hat{Q}_{(j),n}^* (T\hat{Q}_n - T_{(j)}\hat{Q}_{(j),n})P_m\| \\ &\leq \sqrt{m}\delta_m(j)C + \|(T - T_{(j)})\hat{Q}_{(j),n}P_m\| \\ &\quad + \|T(\hat{Q}_n - \hat{Q}_{(j),n})P_m\| \\ &\leq 2\sqrt{m}\delta_m(j)C + \frac{1}{j}.\end{aligned}$$

So we need only show that $\delta_m(j) \to 0$ as $j \to \infty$. Note that as $j \to \infty$, the columns of $(T_{(j)})^n$ converge to that of T^n. It follows that $\tilde{t}_k^j$ converge to t_k and $\tilde{q}_1^j$ converges to q_1. An easy inductive argument using (4.7) and (4.8) shows that the vectors $\tilde{q}_k^j$ converge to q_k and $\|\tilde{v}_k^j\|$ are bounded below. The convergence $\delta_m(j) \to 0$ now follows. □

5 SCI classification theorems

In this section we will apply the above results to prove three new classification theorems in the SCI hierarchy. First, assume that $T \in \mathcal{B}(l^2(\mathbb{N}))$ is an invertible normal operator with $\sigma(T) = \omega \cup \Psi$, where $\omega \cap \Psi = \emptyset$, $\omega = \{\lambda_i\}_{i=1}^N$, and the λ_i's are isolated eigenvalues with multiplicity m_i satisfying $|\lambda_1| > \cdots > |\lambda_N|$. As usual, we also assume that $\sup\{|\theta| : \theta \in \Psi\} < |\lambda_N|$ and set

$$M := m_1 + \cdots + m_N \in \mathbb{N} \cup \{\infty\}. \tag{5.1}$$

In this section we will assume for simplicity that all the m_i except possibly m_N are finite. To be able to obtain the classification results we need two key assumptions.

(I) *Column decay* We assume a much weaker condition than bandedness of the infinite matrix. Indeed, we suppose a known decay of the elements in the columns of T that is described through a family of increasing functions $\{g^j\}_{j\in\mathbb{N}}$. In particular, $g^j : \mathbb{N} \to \mathbb{N}$ is such that defining the infinite matrix $T_{(j)}$ with columns $\{P_{g^j(n)}Te_n\}_{n\in\mathbb{N}}$ we have that $T_{(j)}$ is invertible and

$$\left\|(P_{g^j(n)} - I)Te_n\right\| \leq \frac{1}{j}, \qquad n \in \mathbb{N}. \tag{5.2}$$

(II) *Distance to span of eigenvectors* In order to obtain error control (Δ_1 classification) one needs to control the hidden constant in the $O(r^n)$ estimate in (3.16). This is done as follows, where $\{Q_n\}_{n\in\mathbb{N}}$ is a Q-sequence of T with respect to $\{e_j\}_{j\in\mathbb{N}}$. Given finite $k < M+1$ with $m_1+\cdots+m_{N-1} < k$, we will assume that if $l < N$ then $\{\chi_{\{\lambda_1,\ldots,\lambda_l\}}(T)e_j\}_{j=1}^{m_1+\cdots+m_l}$ are linearly independent. We also assume that $\{\chi_{\{\lambda_1,\ldots,\lambda_N\}}(T)e_j\}_{j=1}^{k}$ are linearly independent. This simply ensures that the IQR algorithm converges with the expected ordering (largest eigenvalue in the first diagonal entry then in descending order). It follows from Theorems 3.9 and 3.7, that there exist eigenspaces $E_1,\ldots,E_N$ (with the last space depending on k and the vectors $\{e_j\}$) corresponding to the eigenvalues $\lambda_1,\ldots,\lambda_N$ such that

- $E_i = \ker(T-\lambda_i I)$ is the full eigenspace if $i < N$
- $\hat{\delta}\left(\bigoplus_{i=1}^{l} E_i, \mathrm{span}\{Q_n e_j\}_{j=1}^{\min\{m_1+\cdots+m_l,k\}}\right) \to 0$ as $n\to\infty$ for $l=1,\ldots,N$.

We then define the initial supremum subspace angle by

$$\Phi(T,\{e_j\}_{j=1}^{k}) := \sup_{l=1,\ldots,N} \phi\left(\bigoplus_{i=1}^{l} E_i, \mathrm{span}\{e_j\}_{j=1}^{\min\{m_1+\cdots+m_l,k\}}\right), \tag{5.3}$$

where ϕ, defined by (1.4), denotes the subspace angle. Our assumptions and the proofs in Sect. 3 show that $\Phi(T,\{e_j\}_{j=1}^{k}) < \pi/2$ and hence the key quantity $\tan\left(\Phi(T,\{e_j\}_{j=1}^{k})\right)$ is finite.

Remark 5.1 The quantity $\tan\left(\Phi(T,\{e_j\}_{j=1}^{k})\right)$ can be viewed as a measure of how far $\{e_j\}_{j=1}^{k}$ is from $\{q_j\}_{j=1}^{k}$, the k eigenvectors of T corresponding to the first k eigenvalues (including multiplicity and preserving order). Hence it gives an estimate of how good the initial approximation $\{e_j\}_{j=1}^{k}$ to $\{q_j\}_{j=1}^{k}$ is. Indeed, we know from (3.16) that the convergence rate is $O(r^n)$, and the hidden constant C depends exactly on this behaviour. In particular, if $e_j = q_j$ for $j\leq k$ then $C=0$.

Define also

$$r(T) = \max\{|\lambda_2/\lambda_1|,\ldots,|\lambda_N/\lambda_{N-1}|, \rho(T)/|\lambda_N|\}, \qquad \rho(T) = \sup\{|z| : z\in\Psi\}.$$

We can now define the class of operators $\Omega_{t,L}^{k}$ for the classification theorem.

Definition 5.2 Given $k\in\mathbb{N}$, $t\in(0,1)$ and $L>0$, let $\Omega_{t,L}^{k}$ denote the class of invertible normal operators T acting on $l^2(\mathbb{N})$ with $\|T\|\leq L$ such that:

1. There exists the decomposition $\sigma(T)=\omega\cup\Psi$ as above with $m_1+\cdots+m_{N-1} < k \leq M$, where M is defined in (5.1).
2. If $m_1+\cdots+m_l < k$ then $\{\chi_{\{\lambda_1,\ldots,\lambda_l\}}(T)e_j\}_{j=1}^{m_1+\cdots+m_l}$ are linearly independent. Also, the vectors $\{\chi_{\{\lambda_1,\ldots,\lambda_N\}}(T)e_j\}_{j=1}^{k}$ are linearly independent.
3. We have access to functions $g^j:\mathbb{N}\to\mathbb{N}$ with (5.2).
4. It holds that $r(T)\leq t$ and $\tan\left(\Phi(T,\{e_j\}_{j=1}^{k})\right)\leq L$.

We can now define the computational problem that we want to classify in the SCI hierarchy. Consider for any $T\in\Omega_{t,L}^{k}$, the problem of computing the k-th largest eigenvalues (including multiplicity) and the corresponding eigenspaces. In other words, we consider the set-valued mapping

$$\Xi_1(T) = \mathcal{S} \subset \mathcal{M} = \mathbb{C}^k \times \left(l^2(\mathbb{N})\right)^k$$

where we define

$$\mathcal{S} := \Bigg\{ (\underbrace{\lambda_1, \ldots, \lambda_1}_{m_1 \text{ times}}, \ldots, \underbrace{\lambda_N, \ldots, \lambda_N}_{k-(m_1+\cdots+m_{N-1}) \text{ times}}) \times (\hat{q}_1, \ldots, \hat{q}_k) :$$
$$\text{s.t. } \{\hat{q}_j\}_{j=m_1+\cdots+m_{l-1}+1}^{m_1+\cdots+m_l} \text{ is an orthonormal basis of } \mathrm{ran}(\chi_{\lambda_l}(T)) \text{ for } l < N$$
$$\text{and } \{\hat{q}_j\}_{j=m_1+\cdots+m_{N-1}+1}^{k} \text{ is an orthonormal basis for a subspace of } \mathrm{ran}(\chi_{\lambda_N}(T)) \Bigg\}.$$

As discussed in Remark A.6 in an Appendix, where we review the SCI hierarchy, when we speak of convergence of $\mathcal{M} \ni \Gamma_n(T)$ to $\Xi_1(T)$, we define, with a slight abuse of notation,

$$\mathrm{dist}(\Gamma_n(T), \Xi_1(T)) := \inf_{y \in \Xi_1(T)} d_{\mathcal{M}}(\Gamma_n(T), y) \to 0.$$

Having established the basic definition we can now present the classification theorem.

Theorem 5.3 (Δ_1 classification for the extremal part of the spectrum) *Given the above set-up we have $\{\Xi_1, \Omega^k_{t,L}\} \in \Delta_1$. In other words, for all $n \in \mathbb{N}$, there exists a general tower using radicals, $\Gamma_n(T)$, such that for all $T \in \Omega^k_{t,L}$,*

$$\mathrm{dist}(\Gamma_n(T), \Xi_1(T)) \leq 2^{-n}.$$

Remark 5.4 Note that this means we converge to the k largest magnitude eigenvalues *in order* with error control, and not just arbitrary points of the spectrum. This is in contrast to most Σ_1 classifications in the SCI hierarchy where the best we can hope for is to bound $\mathrm{dist}(z, \sigma(T))$ for $z \in \mathbb{C}$.

Proof of Theorem 5.3 Let $T \in \Omega^k_{t,L}$ then by the definition of $\Omega^k_{t,L}$, we may take $\hat{e}_j = e_j$ for $j = 1, \ldots, k$ in the arguments in Sect. 3.1. The first step is to bound $Z(T, \{e_j\}_{j=1}^k)$ in terms of $\Phi(T, \{e_j\}_{j=1}^k)$. Let $\{\tilde{e}_j\}_{j=1}^k$ denote the basis described in Sect. 3.1. In our case:

- For any $1 \leq i \leq k$, $\mathrm{span}\{\tilde{e}_j\}_{j=1}^i = \mathrm{span}\{e_j\}_{j=1}^i$.
- If $j > m_1 + \cdots + m_l$ then $\chi_{\lambda_l}(T)\tilde{e}_j = 0$.
- The vectors $\{\chi_{\lambda_l}(T)\tilde{e}_j\}_{j=m_1+\cdots+m_{l-1}+1}^{\min\{m_1+\cdots+m_l,k\}}$ are orthonormal.

Let $\delta_j = \|\tilde{e}_j\|$ then we must have that if $m_1 + \cdots, m_{l-1} < j \leq m_1 + \cdots + m_l$ then

$$\begin{aligned}\frac{\delta_j^2 - 1}{\delta_j^2} &\leq \delta\Big(\mathrm{span}\{\tilde{e}_j\}, \bigoplus_{i=1}^{l} \mathrm{span}\{\chi_{\{\lambda_i\}}(T)\tilde{e}_j\}_{j=m_1+\cdots,m_{i-1}+1}^{\min\{m_1+\cdots+m_i,k\}}\Big)^2 \\ &\leq \delta\Big(\mathrm{span}\{\tilde{e}_j\}_{j=1}^{\min\{m_1+\cdots+m_l,k\}}, \bigoplus_{i=1}^{l} \mathrm{span}\{\chi_{\{\lambda_i\}}(T)\tilde{e}_j\}_{j=m_1+\cdots,m_{i-1}+1}^{\min\{m_1+\cdots+m_i,k\}}\Big)^2 \\ &= \delta\Big(\mathrm{span}\{e_j\}_{j=1}^{\min\{m_1+\cdots+m_l,k\}}, \bigoplus_{i=1}^{l} E_i\Big)^2 \\ &\leq \sin^2\left(\Phi(T, \{e_j\}_{j=1}^k)\right)\end{aligned}$$

Where the first line holds since the nearest point to $\tilde{e}_j$ in $\bigoplus_{i=1}^{l} \mathrm{span}\{\chi_{\{\lambda_i\}}(T)\tilde{e}_j\}_{j=m_1+\cdots,m_{i-1}+1}^{\min\{m_1+\cdots+m_i,k\}}$ is simply $\chi_{\lambda_l}(T)\tilde{e}_j$ and the E_i are defined as above and in (3.1). Rearranging, this implies that

$$\delta_j^2 \leq \frac{1}{1 - \sin^2\left(\Phi(T, \{e_j\}_{j=1}^k)\right)} = \frac{1}{\cos^2\left(\Phi(T, \{e_j\}_{j=1}^k\right)}.$$

Hence it follows that

$$Z(T, \{e_j\}_{j=1}^k) = \Big(\sum_{j=1}^{k} \delta_j^2 - 1\Big)^{\frac{1}{2}} \leq \Big(\sum_{j=1}^{k} \tan^2\left(\Phi(T, \{e_j\}_{j=1}^k)\right)\Big)^{\frac{1}{2}} \leq \sqrt{k}L.$$

In particular, Theorem 3.7 and its proof now implies that

$$\hat{\delta}(\mathrm{span}\{\hat{q}_j\}, \mathrm{span}\{Q_m e_j\}) \leq B(j)\sqrt{k}Lt^m,$$

where $\{\hat{q}_j\}_{j=1}^k$ are orthonormal eigenvectors of T and Q_m is a $Q-$sequence of T. In particular, $\{B(j)\}_{j=1}^k$ can be computed in finitely many arithmetic operations from the induction proof of Theorem 3.7. It follows that there exists $z_{j,m} \in \mathbb{C}$ of unit modulus such that defining $\beta = \max\{B(1), \ldots, B(k)\}\sqrt{k}L$, we have

$$\left\| Q_m e_j - z_{j,m}\hat{q}_j \right\| \leq \beta t^m.$$

Note that we do not need to assume knowledge of N for this bound (trivially $N \leq k$). Using that Q_m is an isometry, this implies that

$$\left|\langle Q_m^* T Q_m e_j, e_j\rangle - \lambda_{a_j}\right| \leq 2\left\|T\right\| \beta t^m \leq 2L\beta t^m,$$

where $T\hat{q}_j = \lambda_{a_j}$. Note that we must have $\{\lambda_{a_j}\}_{j=m_1+\cdots+m_{l-1}+1}^{m_1+\cdots+m_l} = \lambda_l$ and $\{\lambda_{a_j}\}_{j=m_1+\cdots+m_{N-1}+1}^{k} = \lambda_N$ by 3. in the definition of $\Omega_{t,L}^k$.

Given any $\epsilon > 0$, choose m large enough so that $2L\beta t^m \leq \epsilon$ and $\beta t^m \leq \epsilon$. The fact that $\|T\| \leq L$ and (5.2) hold implies that we can compute $\langle Q_m^* T Q_m e_j, e_j \rangle$ to accuracy ϵ using finitely many arithmetical and square root operations using Theorem 4.8. Call these approximations $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_k$. Furthermore, the proof of Theorem 4.8 also makes clear that we can compute $Q_m e_j \in l^2(\mathbb{N})$ to accuracy ϵ using finitely many arithmetical and square root operations (the approximations have finite support). Call these approximations $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_k$. Then set

$$\Gamma^{\epsilon}(T) = (\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_k) \times (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_k).$$

The above estimates show that $\mathrm{dist}(\Gamma^{\epsilon}(T), \Xi_1(T)) \leq 4k\epsilon$. The proof is completed by setting $\Gamma_n(T) = \Gamma^{2^{-(n+2)}/k}(T)$. □

Next, suppose that we have a continuous increasing function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ diverging at ∞ such that $g(0) = 0$ and $g(x) \leq x$. Let $\Omega_{\mathrm{IQR}}^{g}$ be the set of all operators T acting on $l^2(\mathbb{N})$ (i.e. we fix the representation w.r.t. the canonical basis) for which the IQR algorithm converges in the weak operator topology to a diagonal matrix with the same spectrum as T and such that

$$\left\| (T - zI)^{-1} \right\|^{-1} \geq g\big(\mathrm{dist}(z, \sigma(T))\big).$$

Note that by Theorem 3.9 this includes all normal compact operators, T, such that $\{z \in \sigma(T) : |z| = s\}$ has size at most 1 for all $s > 0$ (where we can take $g(x) = x$).[2] We will allow evaluations of g in our algorithms and also assume that we are given functions that satisfy (5.2) and have an upper bound for $\|T\|$. We consider computing $\Xi_2(T) = \sigma(T)$ in the space of compact non-empty subsets of $\mathbb{C}$ with the Hausdorff metric.

Theorem 5.5 (Σ_1 classification for spectrum) *Given the above set-up we have* $\{\Xi_2, \Omega_{\mathrm{IQR}}^{g}\} \in \Sigma_1$. *In other words, there is a convergent sequence of general towers using radicals,* $\Gamma_n(T)$, *such that* $\Gamma_n(T) \to \Xi_2(T) = \sigma(T)$ *for any* $T \in \Omega_{\mathrm{IQR}}^{g}$ *and for all n we have*

$$\Gamma_n(T) \subset \sigma(T) + B_{2^{-n}}(0).$$

Proof of Theorem 5.5 Let $T \in \Omega_{\mathrm{IQR}}^{g}$ and Q_m be a Q−sequence of T. Fix $n \in \mathbb{N}$. Then Theorem 4.8 shows that we can compute any finite number of the diagonal entries of $Q_m^* T Q_m$ to any given accuracy using finitely many arithmetical and square root operations. Similarly, the proof shows that we can compute $T Q_m e_j$ and $Q_m e_j$ to any given accuracy in $l^2(\mathbb{N})$ (the approximations have finite support). Now let $\alpha_{j,m}$ be the computed approximations of $\langle Q_m^* T Q_m e_j, e_j \rangle$ to accuracy $1/m$, then since $T \in \Omega_{\mathrm{IQR}}^{g}$ we have that $\lim_{m\to\infty} \alpha_{j,m} = \alpha_j \in \sigma(T)$. Furthermore, $\{\alpha_j : j \in \mathbb{N}\}$ is dense in

[2] A simple compactness argument shows that for any bounded operator T there is a corresponding function g that works.

$\sigma(T)$. We have that

$$\left\|(T-\alpha_{j,m}I)^{-1}\right\|^{-1} \leq \left\|TQ_m e_j - \alpha_{j,m}Q_m e_j\right\|$$

and hence that

$$\mathrm{dist}(\alpha_{j,m},\sigma(T)) \leq g^{-1}(\left\|TQ_m e_j - \alpha_{j,m}Q_m e_j\right\|). \tag{5.4}$$

Given m, j, we can compute an upper bound $h_{j,m}$ for the right-hand side of (5.4) by approximating the norm $\left\|TQ_m e_j - \alpha_{j,m}Q_m e_j\right\|$ from above to accuracy $1/m$ and finitely many evaluations of g. Namely, let $x_{j,m}$ be the approximation of $\left\|TQ_m e_j - \alpha_{j,m}Q_m e_j\right\|$ and set

$$h_{j,m} = \frac{\min\{l \in \mathbb{N} : g(l/m) \geq x_{j,m}\}}{m}.$$

It is then clear that $\lim_{m\to\infty} h_{j,m} = 0$ and $h_{j,m} \geq g^{-1}(\left\|TQ_m e_j - \alpha_{j,m}Q_m e_j\right\|)$.

We set $\Gamma_n(T) = \{\alpha_{j,m(n,T)} : j = 1,\ldots,n\}$ where $m(n,T)$ is minimal such that $h_{j,m} \leq 2^{-n}$ for $j = 1,\ldots,n$. By (5.4), we must have that

$$\Gamma_n(T) \subset \sigma(T) + B_{2^{-n}}(0).$$

It is also clear that $\Gamma_n(T) \to \sigma(T)$ in the Hausdorff metric. □

The final result considers dominant invariant subspaces discussed in Theorem 3.15. Let $M \in \mathbb{N}$, $t \in (0,1)$ and $L > 0$. We let $\tilde{\Omega}^M_{t,L}$ denote the class of operators such that the assumptions of Theorem 3.15 hold (same M) and such that:

1. $\beta/\alpha < t$
2. $\max\left\{\|T\|, \frac{\sin\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, \mathrm{ran}(P))\right)}{\cos\left(\phi(\mathrm{span}\{e_j\}_{j=1}^M, S)\right)}\left(1+\frac{\|PT(I-P)\|}{\alpha-\beta}\right)\right\} \leq L$

We also assume that we are given functions that satisfy (5.2) and consider computing the dominant invariant subspace $\Xi_3(T) = \mathrm{ran}(P)$ in the space of M-dimensional subspaces of $l^2(\mathbb{N})$ equipped with the metric $\hat{\delta}$.

Theorem 5.6 (Δ_1 classification for dominant invariant subspace) *Given the above set-up we have $\{\Xi_3, \tilde{\Omega}^M_{t,L}\} \in \Delta_1$. In other words, for all $n \in \mathbb{N}$, there exists a general tower using radicals, $\Gamma_n(T)$, each an M-dimensional subspace of $l^2(\mathbb{N})$, such that for all $T \in \tilde{\Omega}^M_{t,L}$,*

$$\hat{\delta}(\Gamma_n(T), \Xi_3(T)) \leq 2^{-n}.$$

Proof of Theorem 5.6 Let $n \in \mathbb{N}$ and $T \in \tilde{\Omega}^M_{t,L}$. Then from Theorem 3.15, we can choose m large so that $t^m L < 2^{-(n+1)}$, and hence

$$\hat{\delta}(\mathrm{span}\{Q_m e_j\}_{j=1}^M, \mathrm{ran}(P)) < 2^{-(n+1)}.$$

Using Theorem 4.8 and its proof, given ϵ we can compute in finitely many arithmetical and square root operations, approximations $v_{m,j}(\epsilon)$ (of finite support) such that

$$\|v_{m,j}(\epsilon) - Q_m e_j\| \leq \epsilon.$$

The vectors $\{Q_m e_j\}_{j=1}^{M}$ are orthonormal, as are the approximations $\{v_{m,j}(\epsilon)\}_{j=1}^{M}$. A simple application of Hölder's inequality then yields

$$\hat{\delta}(\mathrm{span}\{v_{m,j}(\epsilon)\}_{j=1}^{M}, \mathrm{span}\{Q_m e_j\}_{j=1}^{M}) \leq \sqrt{M}\epsilon.$$

By the triangle inequality, the proof of the theorem is complete by choosing ϵ such that $\sqrt{M}\epsilon \leq 2^{-(n+1)}$ and then setting $\Gamma_n(T) = \mathrm{span}\{v_{m,j}(\epsilon)\}_{j=1}^{M}$. □

6 Examples and numerical simulations

The aim of the is section is threefold:

1. To demonstrate the convergence and implementation results of Sects. 3–5 on practical examples.
2. To demonstrate that, as well as the proven results, the IQR algorithm performs better than theoretically expected in many cases. In particular, we conjecture that for normal operators whose essential spectrum has exactly one extremal point, the IQR algorithm will also converge to this point. We also demonstrate cases where this seems to hold even if there are multiple extreme points of the essential spectrum and even in non-normal cases.
3. To compare the IQR algorithm to the finite section method and show that in some cases it considerably outperforms it. In general, one can view $\sigma(P_m Q_n^* T Q_n|_{P_m\mathcal{H}})$ as a generalised version of the finite section method, now with two parameters (m and n) that can be varied with n controlling the number of IQR iterates. In some cases we find this avoids spectral pollution whilst still converging to the entire spectrum.

Before embarking with some numerical examples, two remarks are in order. First, extra care has been taken in the case of non-self-adjoint operators whose finite truncations can be non-normal, and hence the computation of their spectra can be numerically unstable. Unless stated otherwise, all calculations were performed in double precision (in MATLAB) and have been checked against extended precision [32] to ensure that none of the results are due to numerical artefacts. Second, when dealing with operators acting on $l^2(\mathbb{Z})$ we use $\mathbb{N}$ as an index set by listing the canonical basis as $e_0, e_1, e_{-1}, e_2, e_{-2}, \ldots$, allowing us to apply the IQR algorithm on $l^2(\mathbb{N})$. Of course different indexing is possible and in general, this would lead to different implementations of the IQR algorithm,[3] but we stick with this ordering throughout.

[3] A discussion of this is beyond the scope of this paper. In effect, for invertible operators, this corresponds to choosing the order of columns on which to perform a Gram-Schmidt type procedure.

6.1 The finite section method

We first briefly say a few words on the finite section method, the standard means to discretise infinite matrices, since comparisons will be made later. If $\{P_m\}_{m\in\mathbb{N}}$ is a sequence of finite-rank projections such that $P_{m+1} \geq P_m$ and $P_m \to I$ strongly, where I is the identity, then the idea is to replace T by the finite square matrix $P_mT|_{P_m\mathcal{H}}$ (typically, one takes P_m to be the orthogonal projection onto $\mathrm{span}\{e_1, \ldots, e_m\}$). Thus, to find $\sigma(T)$, we instead compute $\sigma(P_mT|_{P_m\mathcal{H}})$. However, there can be significant issues when using the finite section method. In general, there is no guarantee that the computed spectra $\sigma(P_mT|_{P_m\mathcal{H}})$ need converge to $\sigma(T)$.

For example, consider the shift operator $Se_j = e_{j+1}$ on $l^2(\mathbb{N})$. If P_m projects onto $\mathrm{span}\{e_1, \ldots, e_m\}$, we would get that $\sigma((P_mS|_{P_m\mathcal{H}}) = \{0\}$ for all m, whereas $\sigma(S)$ is the closed unit disc. We can also have that $\sigma(P_mT|_{P_m\mathcal{H}}) \nsubseteq \sigma(T)$. For example, let

$$A = \begin{pmatrix} a_1 & i & & & \\ 1 & a_2 & i & & \\ & 1 & a_3 & i & \\ & & 1 & a_4 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \tag{6.1}$$

where $a_j = 5\cos(j)/4 + 2i\sin(j)$. To gain an accurate picture of the spectrum, note that A is banded and hence we can compute approximates to the pseudospectrum [38]. In order to approximate the spectrum in the best possible way we must take ϵ as small as possible. Unfortunately, there is a restriction to how small ϵ can be depending on ϵ_{mach} (machine precision) of the software used. To illustrate this, observe that the approximates are given by (a discrete version of)

$$\begin{aligned}\Gamma_m(A) = \{z \in \mathbb{C} : \min\{\sqrt{\lambda} : \lambda \in \sigma(P_m(A-z)^*(A-z)|_{P_m\mathcal{H}})\} \leq \epsilon\} \\ \cup \{z \in \mathbb{C} : \min\{\sqrt{\lambda} : \lambda \in \sigma(P_m(A-z)(A-z)^*|_{P_m\mathcal{H}})\} \leq \epsilon\}.\end{aligned} \tag{6.2}$$

Thus, ignoring the additional error in computing the smallest eigenvalues of the squared operator and assuming A to have matrix entries of order 1, computing $\Gamma_m(A)$ will have the same challenges as if one squares a real number and then takes its square root. In particular, due to the floating point arithmetic used in the software and (6.2), we must ensure that

$$\epsilon \gtrsim \sqrt{\epsilon_{\mathrm{mach}}},$$

and this puts a serious restriction on our computation, particularly for the non-normal case where the distance $d_H(\sigma(T), \sigma_\epsilon(T))$ may be large (though we always have $\sigma(T) \subset \sigma_\epsilon(T)$). However, it is possible to detect spectral pollution outside of $\sigma_\epsilon(T)$ if we can approximate it well.

The phenomenon of "spectral pollution" occurs for A: namely, the computed spectrum $\sigma(P_mA|_{P_m\mathcal{H}})$ contains elements that have nothing to do with $\sigma(A)$. This is

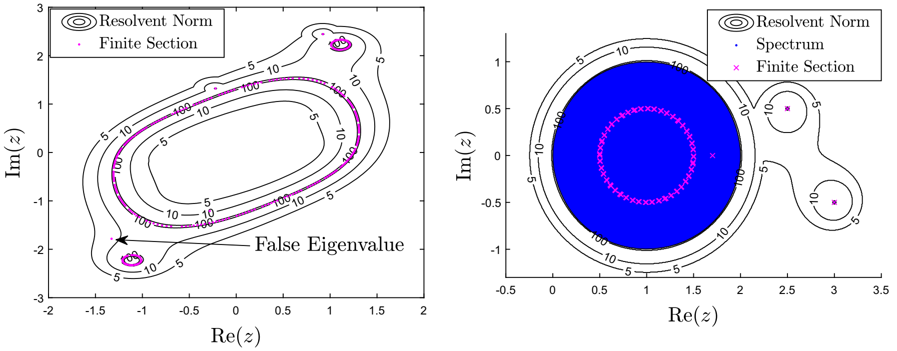

Fig. 1 Left: $\sigma_\epsilon(A)$ plotted as contours of the resolvent norm, as well as $\sigma(P_m A|_{P_m\mathcal{H}})$ for $m = 300$ with the false eigenvalue (recall that $\sigma(A) \subseteq \sigma_\epsilon(A)$). Right: $\sigma_\epsilon(T)$, $\sigma(T)$ and $\sigma(P_m T|_{P_m\mathcal{H}})$ for $m = 100$

visualised in Fig. 1, an example with spectral pollution $z \notin \sigma_{1/10}(A)$, where the same phenomenon occurs for larger m. The spectral pollution phenomenon is well-known. As the following theorem suggests, such pollution can be arbitrarily bad.

Theorem 6.1 (Pokrzywa [58]) *Let $A \in \mathcal{B}(\mathcal{H})$ and $\{P_m\}$ be a sequence of finite-dimensional projections converging strongly to the identity. Suppose that $S \subset W_e(A)$. Then there exists a sequence $\{\tilde{P}_m\}$ of finite-dimensional projections such that $P_m < \tilde{P}_m$ (so $\tilde{P}_m \to I$ strongly) and*

$$d_H(\sigma(A_m) \cup S, \sigma(\tilde{A}_m)) \to 0, \quad \textit{as } m \to \infty,$$

where

$$A_m = P_m A|_{P_m\mathcal{H}}, \qquad \tilde{A}_m = \tilde{P}_m A|_{\tilde{P}_m\mathcal{H}}$$

and d_H denotes the Hausdorff metric.

Despite this result, the finite section can perform quite well. This is the case for self-adjoint operators [6,21,36] and it is also well suited for the computation of pseudospectra of Toeplitz operators [14,18]. Moreover, in general, we have the following (recall that $W_e(T)$ is the convex hull of the essential spectrum for T normal):

Theorem 6.2 (Pokrzywa [58]) *Let $T \in \mathcal{B}(\mathcal{H})$ and $\{P_m\}$ be a sequence of finite-dimensional projections converging strongly to the identity. If $\lambda \notin W_e(T)$ then $\lambda \in \sigma(T)$ if and only if*

$$\operatorname{dist}(\lambda, \sigma(P_m T|_{P_m\mathcal{H}})) \longrightarrow 0, \qquad \textit{as } m \to \infty.$$

However, if we want to use the finite section method and rely on Theorem 6.2, we must know $W_e(T)$, and that may be unpleasant to compute. Alternatively, we could hope that $\sigma_{\mathrm{ess}}(T)$ is close to $W_e(T)$. For example, if T is hypo-normal ($T^*T - TT^* \geq 0$) then

$$\operatorname{conv}(\sigma_{\mathrm{ess}}(T)) = W_e(T),$$

where $\operatorname{conv}(\sigma_{\mathrm{ess}}(T))$ denotes the convex hull of $\sigma_{\mathrm{ess}}(T)$. But what if we have a "very non-normal" operator?

Another problem we may encounter when using the finite section method is that even though $\sigma_d(T)$ may be recovered, one may get a very misleading picture of the rest of the spectrum. Such problems are illustrated in the following simple example. Let

$$T = \left(\begin{array}{cccc|cccc} 2.5+0.5i & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 3-0.5i & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 1.7 & 0.05 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0.05 & t_4 & 0 & 0 & 0 & \cdots \\ \hline 0 & 0 & 0 & 0 & t_5 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & t_6 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & t_7 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{array}\right), \tag{6.3}$$

where $t_j = 1 + 0.5(\sin(j) + i\cos(j))$ for $j \geq 4$. This operator decomposes into an upper 4×4 block and an operator acting on the perpendicular subspace. It is also possible to compute the spectrum analytically (it consists of a disc of radius 1 centred at 1 together with two isolated eigenvalues). Again, we can compute the pseudospectrum of T (Fig. 1) to reveal that whilst the eigenvalues produced by the finite section method are correct, they do not capture the entire spectrum. It is straightforward to adapt this example (e.g. by changing basis) to have the same phenomena without an obvious decomposition of the operator into a finite part and triangular part. Without the support from the picture of the pseudospectrum, the finite section method does not provide information regarding the boundary of the essential numerical range of T—there is a misleading circle of eigenvalues of $P_m T|_{P_m\mathcal{H}}$ which do not occur along the boundary of the essential spectrum but are simply given by the diagonal entries $\{t_5, t_6, \ldots, t_m\}$.

Remark 6.3 The previous examples demonstrated that, in general, the finite section method is not always suitable for computing spectra. Rather then working with square sections of the infinite matrix T, one should work with *uneven sections* $P_n T P_m$, where the parameters n and m are allowed to vary independently. Indeed, the algorithms presented in [24,38] use this method. In effect, we need to know how large n should be to retain enough information of the operator TP_m. This type of idea is also used implicitly in the IQR algorithm (see Sect. 4).

6.2 Numerical examples I: normal operators

Example 6.4 (*Convergence of the IQR algorithm*) We begin with two simple examples that demonstrate the linear (or exponential) convergence proven in Theorem 3.9 and Corollary 3.12 (and its generalisations). Consider first the one-dimensional discrete Schrödinger operator given by

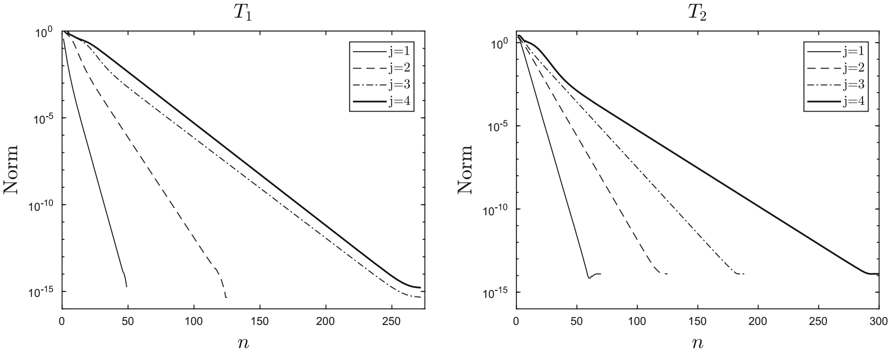

Fig. 2 Exponential convergence to the diagonal blocks for T_1 and T_2

$$T_1 = \begin{pmatrix} v_1 & 1 & & & \\ 1 & v_2 & 1 & & \\ & 1 & v_3 & 1 & \\ & & 1 & v_4 & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

where $v_j = 5\sin(j)^2/\sqrt{j}$ if $j \le 10$ and $v_j = 0$ otherwise. As a compact (in fact finite rank) perturbation of the free Laplacian, $\sigma(T_1)$ consists of the interval $[-2, 2]$ together with isolated eigenvalues of finite multiplicity which can be computed [73]. The second operator, T_2, consists of taking the operator

$$T_0 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3i}{2} & 0 & 0 \\ 0 & 0 & -\frac{5}{4} & 0 \\ 0 & 0 & 0 & -\frac{9i}{8} \end{pmatrix} \bigoplus U_1,$$

where U_k denotes the bilateral shift $e_j \to e_{j+k}$, writing this as an operator on $l^2(\mathbb{N})$ and then mixing the spaces via a random unitary transformation on the span of the first 9 basis vectors. This ensures T_2 is not written in block form but has known eigenvalues. We have plotted the difference in norm between the first $j \times j$ block of each $Q_n^* T_l Q_n$ and the diagonal operator formed via the largest j eigenvalues for $j = 1, 2, 3$ and 4 in Fig. 2. The plot clearly shows the exponential convergence.

Example 6.5 (*Convergence to extremal parts of the spectrum*) To see why we may need some condition on $\sigma(T)$ for convergence of the IQR algorithm to the extreme parts of the spectrum, we consider Laurent and Toeplitz operators with symbol given by a trigonometric polynomial

$$a(t) = \sum_{j=-k}^{j=k} a_j t^j.$$

Given such a symbol, we define Laurent and Toeplitz operators

$$L(a) = \left(\begin{array}{ccc|cccc} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & a_0 & a_{-1} & a_{-2} & a_{-3} & a_{-4} & \cdots \\ \cdots & a_1 & a_0 & a_{-1} & a_{-2} & a_{-3} & \cdots \\ \hline \cdots & a_2 & a_1 & a_0 & a_{-1} & a_{-2} & \cdots \\ \cdots & a_3 & a_2 & a_1 & a_0 & a_{-1} & \cdots \\ \cdots & a_4 & a_3 & a_2 & a_1 & a_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right), \quad T(a) = \begin{pmatrix} a_0 & a_{-1} & a_{-2} & \cdots \\ a_1 & a_0 & a_{-1} & \cdots \\ a_2 & a_1 & a_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

acting on $l^2(\mathbb{Z})$ and $l^2(\mathbb{N})$ respectively. Note that $L(a)$ is always normal whereas $T(a)$ need not be (see for example [18]). A simple example already mentioned is $a(t) = t$ which gives rise to the bilateral and unilateral shifts $L(a) = U_1$ and $T(a) = S$. In this case, both of these operators are invariant under iterations of the IQR algorithm and hence their finite sections $P_m Q_n^* T Q_n|_{P_m\mathcal{H}}$ always have spectrum $\{0\}$. In the case of $L(a)$ this is an example of spectral pollution, whereas in the case of $T(a)$ this does not capture the extremal parts of the spectrum. Regarding pure finite section, the following beautiful result is known:

Theorem 6.6 (Schmidt and Spitzer [62]) *If a is a trigonometric polynomial then we have the following convergence in the Hausdorff metric:*

$$\lim_{m\to\infty} \sigma(P_m L(a)|_{P_m\mathcal{H}}) = \lim_{m\to\infty} \sigma(P_m T(a)|_{P_m\mathcal{H}}) = \bigcap_{r\in(0,\infty)} \sigma(T(a_r)) =: \Upsilon(a),$$

where $a_r(t) = a(rt)$. Furthermore, this limit set is a connected finite union of analytic arcs, each pair of which has at most endpoints in common.

It is straightforward to construct examples where it appears that both $\lim_{n\to\infty} P_m Q_n^* T(a) Q_n|_{P_m\mathcal{H}}$ and $\lim_{n\to\infty} P_m Q_n^* L(a) Q_n|_{P_m\mathcal{H}}$ exist and are either the extreme parts of $\sigma(L(a))$ or of $\Upsilon(a)$. For example, consider the symbols

$$a(t) = \frac{t^3 + t^{-1}}{2}, \quad \tilde{a}(t) = t + it^{-2}.$$

Figure 3 shows the outputs of the IQR algorithm and plain finite section for the corresponding Laurent and Toeplitz operators for $m = 50$ and $n = 1$ and $n = 300$. In the case of a, it appears that both limit sets are the extremal parts of $\sigma(L(a))$ (together with 0 if m is not a multiple of 4). Whereas in the case of $\tilde{a}$ it appears that $\lim_{n\to\infty} P_m Q_n^* T(\tilde{a}) Q_n|_{P_m\mathcal{H}}$ is the extremal parts of $\Upsilon(\tilde{a})$ and $\lim_{n\to\infty} P_m Q_n^* L(\tilde{a}) Q_n|_{P_m\mathcal{H}}$ is the extremal parts of $\sigma(L(\tilde{a}))$ (again together with a finite collection of points depending on the value of m modulo 3). Curiously, in both cases we observed convergence in the strong operator topology to block diagonal operators (up to unitary equivalence in each sublock), whose blocks have spectra corresponding to the limiting sets (hence the dependence on the remainder of m modulo 2 or 3). However, in contrast to convergence to points in the discrete spectrum,

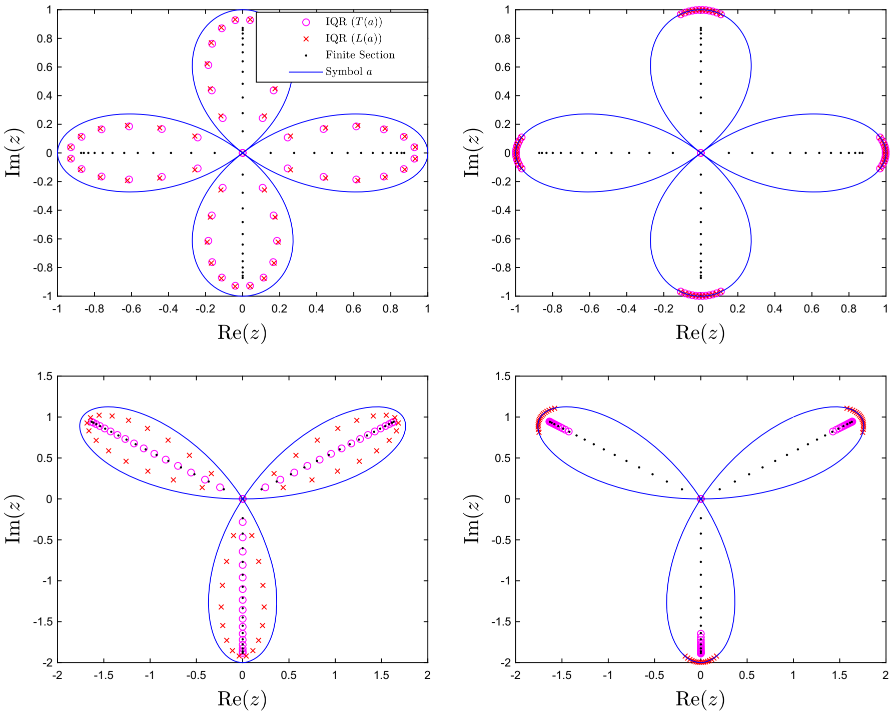

Fig. 3 Top: output of IQR and finite section on $T(a)$ and $L(a)$ for $m = 50$ and $n = 1$ (left), $n = 300$ (right). Bottom: same but for the symbol $\tilde{a}$. In both cases for a given symbol b, $\sigma(L(b))$ is given by $\{b(z) : z \in \mathbb{T}\}$ (shown) and $\sigma(T(b))$ is given by $\sigma(L(b)) \cup \{z \in \mathbb{C} \backslash b(\mathbb{T}) : \mathrm{wind}(b, z) \neq 0\}$

convergence to these operators was only algebraic. This is shown in Fig. 4 where we have plotted the Hausdorff distance between the limiting set and the eigenvalues of the first diagonal block. We also shifted the operators ($+\,1.1I$ for a and $-\,1.5iI$ for $\tilde{a}$) so that the extremal points correspond to exactly one point. In this scenario and for all operators (Laurent or Toeplitz) the IQR algorithm converges strongly to a diagonal operator whose diagonal entries are the corresponding extremal point of $\sigma(L(a))$. This convergence is also shown in Fig. 4 and we observed a slower rate of convergence than before. This is possibly due to points from the other tips of the petals of $\sigma(L(a))$ converging as we increase n. It would be interesting to see if some form of Theorem 6.6 holds for the IQR algorithm (now taking $n \to \infty$). Given the examples presented here, such a statement would likely be quite complicated. However, we conjecture that if a normal operator has exactly one extreme point of its essential spectrum (and finitely many eigenvalues of magnitude greater than r_{ess}) then this extreme point will be recovered in the limit $n \to \infty$ for large enough m.

Example 6.7 (*IQR and avoiding spectral pollution*) In this example we consider whether the IQR algorithm may be used as a tool to avoid spectral pollution. Sometimes when considering $\sigma(P_m T|_{P_m\mathcal{H}})$, spectral pollution can be detected by changing m (edge states which correspond to spectral pollution are often unstable, but this is not

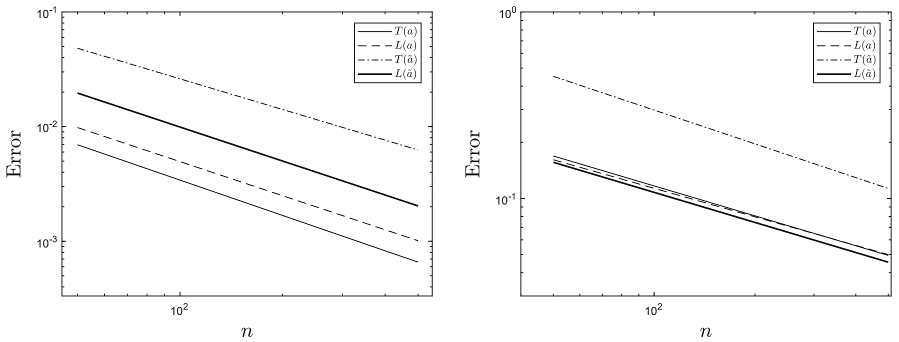

Fig. 4 Left: algebraic convergence to block diagonal operators. Right: algebraic convergence to diagonal operators. In both cases, we have plotted the difference in eigenvalues of the first block as we increase n

always the case). In general, $\sigma(P_m Q_n^* T Q_n|_{P_m\mathcal{H}})$ can be considered as a generalised version of finite section with a finite number (n) of IQR iterates being performed on the infinite-dimensional operator before truncation. If Q_n is unitary, then this simply changes the basis before truncation and such a change may reduce (or change) spectral pollution allowing it to be detected. Here we consider

$$T_3 = \begin{pmatrix} 0 & 3 & & & \\ 3 & 0 & 1 & & \\ & 1 & 0 & 3 & \\ & & 3 & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}.$$

The spectrum of T_3 is $[-4,-2]\cup[2,4]$. However, if m is odd then $0 \in \sigma(P_m T_3|_{P_m\mathcal{H}})$. We shifted the operator by considering $T_3 + 0.2I$ (and then shifted back for the spectrum). Figure 5 shows the Hausdorff distance between $\sigma(P_m Q_n^*(T_3+0.2I)Q_n|_{P_m\mathcal{H}}) - 0.2I$ and $\sigma(T_3)$ as n varies for different m. The spikes in the distance correspond to eigenvalues leaving the interval $[-4,-2]$ and crossing to $[2,4]$ (also shown in Fig. 5). The increase in distance as m decreases (for large n) is due less of the interval $[-4,-2]$ being approximated. It appears that the IQR algorithm can be an effective tool at detecting spectral pollution - certainly a mixture of varying m and n will be more effective than just varying m.

Another example of this is given by the operator $L(a)$ considered previously. For fixed m we found that

$$\lim_{n\to\infty} \sup_{z\in\sigma(P_m Q_n^* L(a) Q_n|_{P_m\mathcal{H}})} \mathrm{dist}(z,\sigma(L(a))) = 0.$$

However, for finite section, spectral pollution occurs for all large m

$$\lim_{m\to\infty} \sup_{z\in\sigma(P_m L(a)|_{P_m\mathcal{H}})} \mathrm{dist}(z,\sigma(L(a))) > 0$$

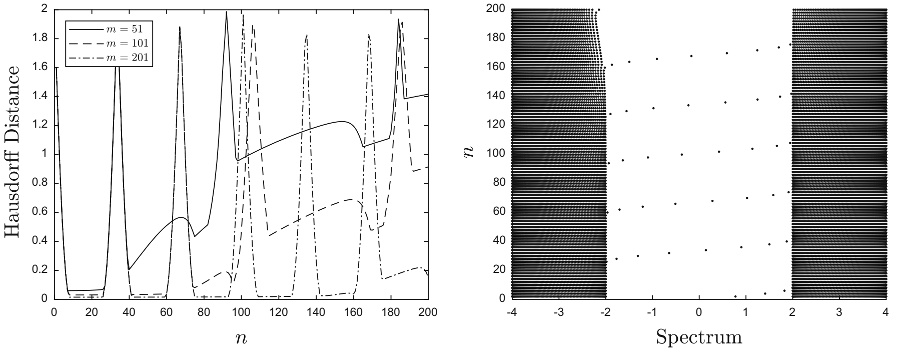

Fig. 5 Left: $d_H(\sigma(P_m Q_n^*(T_3 + 0.2I)Q_n|_{P_m\mathcal{H}}) - 0.2I, \sigma(T_3))$ as a function of n for different m. Right: $\sigma(P_m Q_n^*(T_3 + 0.2I)Q_n|_{P_m\mathcal{H}}) - 0.2I$ as a function of n for $m = 201$. Note the crossing of eigenvalues across the spectral gap

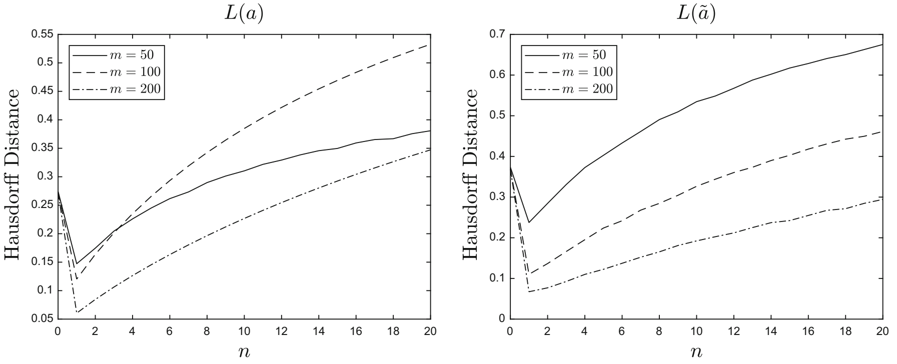

Fig. 6 Left: $d_H(\sigma(P_m Q_n^* L(a) Q_n|_{P_m\mathcal{H}}), \sigma(L(a)))$ as a function of n for different m. $d_H(\sigma(P_m Q_n^* L(\tilde{a}) Q_n|_{P_m\mathcal{H}}), \sigma(L(\tilde{a})))$ as a function of n for different m

and the IQR algorithm can only recover the extreme parts of the spectrum

$$\lim_{n\to\infty} d_H(\sigma(P_m Q_n^* L(a) Q_n|_{P_m\mathcal{H}}), \sigma(L(a))) > 0.$$

Despite this, we found that for small fixed $n > 0$ it appears that

$$\lim_{m\to\infty} d_H(\sigma(P_m Q_n^* L(a) Q_n|_{P_m\mathcal{H}}), \sigma(L(a))) = 0.$$

This is shown in Fig. 6 with similar results for $L(\tilde{a})$.

6.3 Numerical examples II: non-normal operators

Although Theorem 3.9 considers normal operators, Theorems 3.13 and 3.15 suggest the IQR algorithm may also be useful for non-normal operators. Indeed, the results presented here demonstrate that in practice the IQR algorithm can work very well for

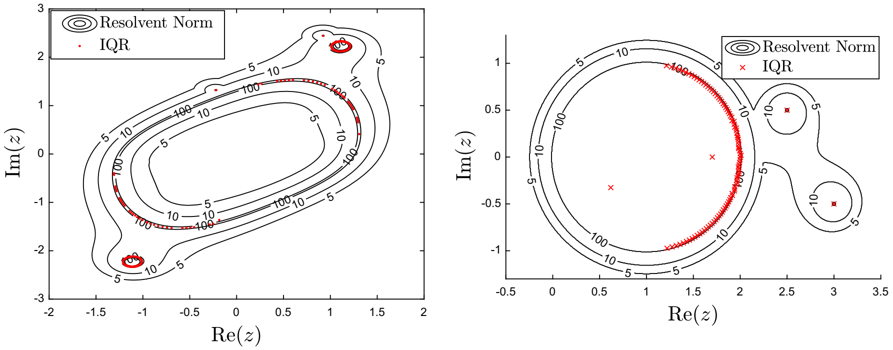

Fig. 7 Left: output of the IQR algorithm $\sigma(P_m Q_n^* A Q_n|_{P_m})$ for $m = 300$ and $n = 1000$. Right: output of the IQR algorithm $\sigma(P_m Q_n^* T Q_n|_{P_m})$ for $m = 100$ and $n = 300$

non-normal problems. If an infinite matrix T has m isolated eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$ (repeated according to multiplicity) outside $r_{\text{ess}}(T)$ (the essential spectral radius), then Theorems 3.13 and 3.15 suggest that the eigenvalues will appear on the diagonal of $P_m Q_n^* T Q_n|_{P_m \mathcal{H}}$ as $n \to \infty$, i.e.

$$\sigma(P_m Q_n^* T Q_n|_{P_m \mathcal{H}}) \longrightarrow \{\lambda_1, \ldots, \lambda_m\}, \qquad \text{as } n \to \infty.$$

We will verify this numerically in the next examples. However, we will see that not only do we get convergence to the eigenvalues, but often we also pick up parts of the boundary of the essential spectrum (this was the case when considering $T(a)$ but appeared not to be the case for $T(\tilde{a})$). This phenomenon is not accounted for in the previous exposition where normality was crucial for proving Theorem 3.9.

Example 6.8 (*Recovering the extremal part of the spectrum*) Let us return to the infinite matrices A in (6.1) and T in (6.3) from Sect. 6.1. We have run the IQR algorithm with $n = 1000$ and $n = 300$ for A and T respectively, shown in Fig. 7. We see that if one takes a finite section after running the IQR algorithm, then part of the boundary of the essential spectrum also appears, along with the discrete spectrum $\sigma_d(A)$. Note that the part of the boundary that is captured is the extreme part (points with largest modulus). It seems that after running the IQR algorithm, the spectral information from the largest isolated eigenvalues and the largest approximate point spectrum is "squeezed up" to the upper and leftmost portions of the matrix. This is not completely counter-intuitive given (2.5) and is what normally happens in finite dimensions. For both examples, we found that the IQR iterates converge to an upper triangular matrix (analogous to the finite-dimensional case) in agreement with Theorems 3.13 and 3.15. The convergence of the upper 1×1 block for A (corresponding to the dominant eigenvalue) and 4×4 non-diagonal block for T are shown in Fig. 8 where we have plotted the difference in norm.

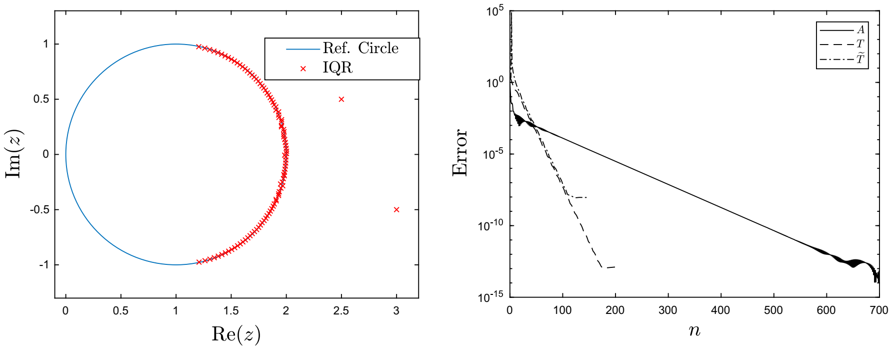

Fig. 8 Left: output of the IQR algorithm $\sigma(P_m Q_n^* \widetilde{T} Q_n|_{P_m})$ for $m = 100$ and $n = 300$. The reference circle is the boundary of the essential spectrum. Right: convergence of upper diagonal blocks for operators A, T and $\widetilde{T}$

We also discuss another drawback of the algorithm, Γ_m, to compute the pseudospectrum by perturbing the operator T. Let $\widetilde{T}$ be the operator obtained from T if we set $\widetilde{T}_{5,4} = 5 \times 10^7$ (note that this gets rid of the block form). The computation of Γ_m involves squaring the operator and hence leads to matrices of norm of order 10^{15}, making it impossible to compute $\sigma_\epsilon(\widetilde{T})$ using double precision for $\epsilon \lesssim 10^{-1}$. However, the IQR algorithm shares the pleasant feature of finite section in allowing a wider range of magnitudes of the matrix entries of the operator. The output for $n = 300$, $m = 100$ is shown in Fig. 8 as well as convergence of the upper 2×2 block (corresponding to the dominant eigenvalues). Note in this case we can only compute this upper block to an accuracy of about 10^{-8} in double precision due to the large perturbed entry. However, this is still much better than pseudospectral techniques. All the errors in Fig. 8 were obtained via comparison with converged matrices computed using quadruple precision.

Example 6.9 (*PT-symmetry in quantum mechanics*) Finally, we consider a so-called PT-symmetric operator (non-normal), demonstrating the same phenomena. A Hamiltonian $H = p^2/2 + V(x)$ is said to be PT-symmetric if it commutes with the action of the operator PT where P is the parity operator $\hat{x} \rightarrow -\hat{x}$, $\hat{p} \rightarrow -\hat{p}$ and T the time operator $\hat{p} \rightarrow -\hat{p}$, $i \rightarrow -i$. Further distinction can be made between exact (unbroken) PT symmetry, when H shares common "eigenfunctions" with PT, and broken PT symmetry, when they possess different eigenfunctions. Many PT Hamiltonians possess the remarkable property that their spectra are real for small enough $\mathrm{Im}(V)$, but that the spectrum becomes complex above a certain threshold [11]. This phase transition from exact to broken PT phase is known as symmetry breaking. There has been a lot of interest in recent years, both theoretically and experimentally, in non-Hermitian PT-symmetric Hamiltonians [45,60].

We consider an operator on $l^2(\mathbb{Z})$ of the form

$$(H_1 x)_n = x_{n-1} + x_{n+1} + V_n x_n. \tag{6.4}$$

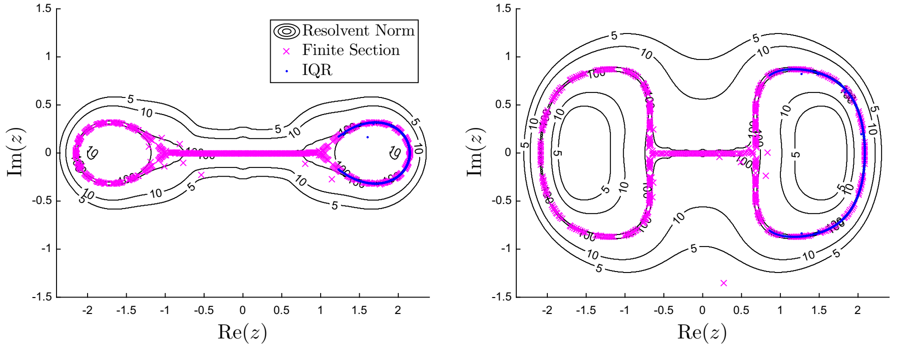

Fig. 9 The figures show finite sections $\sigma(P_m H_1|_{P_m\mathcal{H}})$ (magenta) and (shifted) $\sigma(P_m Q_n^* H_1 Q_n|_{P_m\mathcal{H}})$ IQR iterates (blue) along with converged resolvent norm contours for $\gamma = 1$ (left) and $\gamma = 2$ (right). Both figures are for $m = 500$, $n = 3000$ and show the convergence to the extremal parts of the spectrum

This commutes with (the discrete version of) PT precisely when the potential has even real part and odd imaginary part. We tested the IQR algorithm on the potential

$$V_n = \begin{cases} \cos(n) + i\gamma \sin(n), & \bmod\ (n, 2) = 0 \\ 0, & \bmod\ (n, 2) = 1 \end{cases}, \tag{6.5}$$

and found similar results for other potentials. Figure 9 shows the same qualitative behaviour as the last example for $\gamma = 1, 2$ at $m = 500, n = 3000$. We shifted by 2.2 and 2.15 for $\gamma = 1, 2$ respectively. For comparison, we have shown converged resolvent norms. We found that spectral pollution with no IQR iterates was consistent as we varied m. However, for a fixed m, increasing the number of iterates ($n \to \infty$) caused $\sigma(P_m Q_n^* H_1 Q_n|_{P_m\mathcal{H}})$ to approach the extremal part of the spectrum.

6.4 Numerical examples III: random non-Hermitian operators and boundary conditions

In this final section, we explore examples where the $P_m Q_n^* T Q_n P_m$ naturally give rise to periodic boundary conditions (this was already seen for some examples of Laurent operators in Sect. 6.2). Both examples discussed here are physically motivated random tridiagonal operators on the lattice $\mathbb{Z}$. One of the key applications of studying such random operators can be found in condensed matter physics. The discrete models below have been used to study conductivity of disordered media, flux lines in superconductors and asymmetric hopping particles. Many such operators are also the discretisation of certain stochastic differential equations. As we will demonstrate, the IQR method can be a powerful way of avoiding spectral pollution caused by unnatural "open" boundary conditions in forming the finite section $P_m T P_m$. In both of these examples, periodic boundary conditions are natural and we find that taking finite sections after iterating the IQR algorithm captures periodic boundary conditions.

Example 6.10 (*Hopping sign model in sparse neural networks*) The first example is a non-normal operator with random sub and superdiagonals, first studied by Feinberg and Zee [23,31,40]. The usual "Hopping Sign Model" is defined via

$$(H_2x)_n = x_{n-1} + b_n x_{n+1},$$

with $b_n \in \{\pm 1\}$ (say independent Bernoulli with parameter $p = 1/2$). This describes a particle "'hopping" on $\mathbb{Z}$ and can be mapped into a (complex-valued) random walk. We will consider a slightly different operator described by

$$(H_3x)_n = s^-_{n-1} \exp(-g)x_{n-1} + s^+_n \exp(g)x_{n+1}, \tag{6.6}$$

and appearing in [1] in the context of sparse neural networks. We shall assume that g is real and non-negative and that $s^\pm_j$ are i.d.d. random variables with Bernoulli distribution p. In other words

$$\mathbb{P}(s^\pm_j = 1) = 1 - \mathbb{P}(s^\pm_j = -1) = p.$$

We will only consider $g = 1/10$ and $p = 1/2$, but will vary p in an effort to compute the spectrum of H_3 which only depends on the support of the distribution of the $s^\pm_j$'s. It is easy to prove that the spectrum (and pseudospectrum) of H_3 is almost surely constant and that there is no inessential spectrum. Furthermore, one can show that $\sigma(H_3)$ is contained in the annulus $\{z \in \mathbb{C} : 2\sinh(g) \leq |z| \leq 2\cosh(g)\}$.

Finite section calculations associated with this operator have some interesting properties and are extensively studied in [1]. If one projects using the standard basis of $l^2(\mathbb{Z})$ then one obtains matrices of the form

$$M^1_n = \begin{pmatrix} 0 & s^-_{-n+1}\exp(-g) & & \\ s^+_{-n+1}\exp(g) & 0 & \ddots & \\ & \ddots & \ddots & s^-_{n-1}\exp(-g) \\ & & s^+_{n-1}\exp(g) & 0 \end{pmatrix}.$$

If we use open boundary conditions (i.e. we simply project onto the space spanned by $\{e_{-n}, \ldots, e_n\}$) then one can "gauge" away g by a similarity transformation, leading to

$$M'_n = \begin{pmatrix} 0 & s^-_{-n+1} & & \\ s^+_{-n+1} & 0 & \ddots & \\ & \ddots & \ddots & s^-_{n-1} \\ & & s^+_{n-1} & 0 \end{pmatrix}.$$

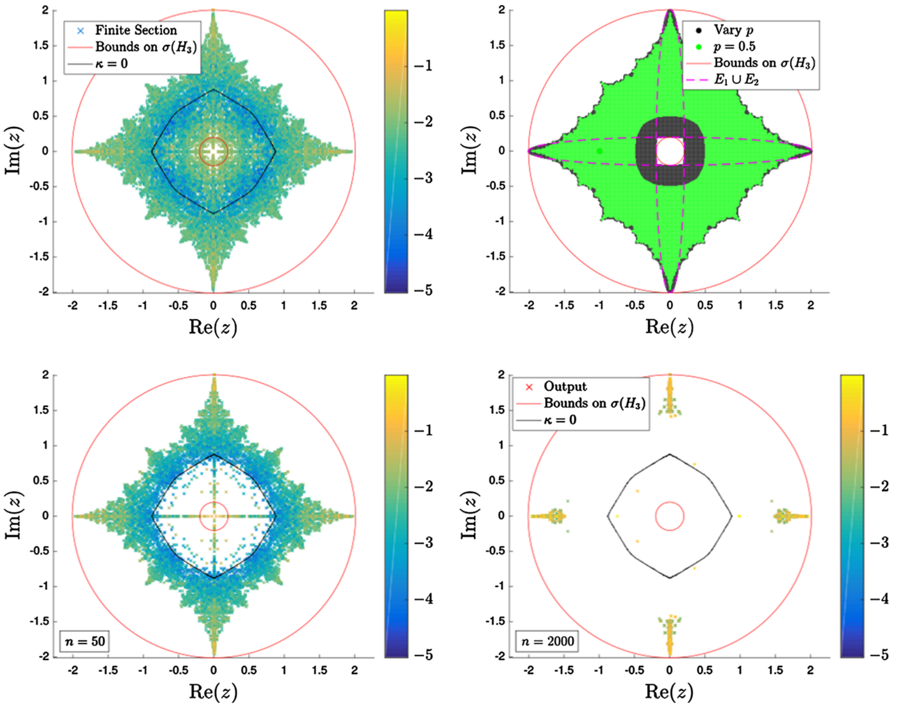

Fig. 10 Top: output of finite section over a random sample of 200 matrices of size 200 (left) and the estimates using pseudospectral techniques (right). Bottom: the output of IQR over 200 samples computing $\sigma(P_m Q_n^* H_3 Q_n|_{P_m\mathcal{H}})$ for $m = 200$ and $n = 50$ (left), $n = 2000$ (right). Note that after a few iterates, the output seems to agree with periodic boundary conditions and then increasing the number of iterates leads to convergence to the extremal parts of the essential spectrum

On the other hand, the use of periodic boundary conditions leads to the matrix

$$M_n^2 = \begin{pmatrix} 0 & s^-_{-n+1}\exp(-g) & & s^+_n\exp(g) \\ s^+_{-n+1}\exp(g) & 0 & \ddots & \\ & \ddots & \ddots & s^-_{n-1}\exp(-g) \\ s^-_n\exp(-g) & & s^+_{n-1}\exp(g) & 0 \end{pmatrix},$$

which does not suffer from this setback.

In [1] this phenomenon was studied via localisation of the eigenvalues of M_n^2, in particular using the Lyapunov exponent $\kappa(z)$ which is equal to the inverse of the localisation length. An eigenfunction ψ with eigenvalue z localised around x_0 behaves approximately as

$$|\psi(x)| \sim \exp(-\kappa(z)\,|x - x_0|).$$

If one defines recursively

$$y_{n+1}(z) = \exp(g)\frac{\psi_{n+2}}{\psi_{n+1}} = -(s_{n-1}^{-}/s_n^{+})/y_n(z) + z/s_n^{+}$$

then (in the limit of large system sizes)

$$\kappa(z; g) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{j=-N}^{N} \left(\log\left|y_j(z)\right| - g\right).$$

This is known as the transfer matrix approach. For fixed z, as we increase g, $\kappa(z; g)$ becomes negative. The heuristic is that a hole opens up in the spectrum corresponding to a negative Lyapunov exponent. Eigenvalues of M_n^2 inside the hole are swept up and become delocalised moving to the rim of the hole, whereas those outside remain largely undisturbed. Eigenvalues of M_n^1 inside the negative κ zone correspond to edge states due to the finite system size approximation.

Figure 10 shows the output of a sample of 200 finite sections with open boundary conditions and matrix size 200. We have also shown the annular region that bounds the spectrum, as well as the contour $\kappa = 0$. In order to calculate κ, we calculated the above sum on a grid with large N to ensure convergence. The colour bar corresponds to the inverse participation ratio (log scale) of normalised eigenfunctions defined by

$$1/P \equiv \frac{\sum_j |\psi_i|^4}{\sum_j |\psi_i|^2}.$$

Note that this has a maximum value of 1 (localised) and a minimum value of $1/N$ (delocalised), N being the size of the matrix. Open boundary conditions produce spectral pollution in the hole with localised eigenfunctions and the contour $\kappa = 0$ corresponds to the delocalised region. In order to compare to the spectrum of the infinite operator on $l^2(\mathbb{Z})$ we have plotted $\sigma_\epsilon(H_3)$, for $\epsilon = 10^{-2}$, calculated using matrix sizes of order 10^5. We note that the spectrum is independent of $p \in (0, 1)$ so we have also shown the union of these estimates over $p = \{k/100\}_{k=1}^{99}$. Although the algorithm used to compute the pseudospectrum is guaranteed to converge to $\sigma_\epsilon(H_3)$, there are regions in the complex plane where this convergence is very slow. Taking unions over p is simply a way to speed up this convergence. Upon taking ϵ smaller, we found that the spectrum appeared to have a fractal-like nature. It also appears that the hole in the spectrum corresponds to the boundary of two ellipses. It is easy to prove that the ellipse

$$E_1 = \{\exp(g + i\theta) + \exp(-g - i\theta) : \theta \in [0, 2\pi)\}$$

is contained in $\sigma(H_3)$ and that the spectrum (and pseudospectrum) of H_3 has fourfold rotational symmetry. Denoting the rotation of E_1 by $\pi/4$ as E_2 we have shown $E_1 \cup E_2$ in the figure.

Figure 10 also shows the effect of IQR iterations over random samples of size 200 for $m = 200$ and $n = 50$ and 2000. Remarkably, as we increase n, a few iterations is enough to capture periodic boundary conditions and sweep away the localised edge states. We have also shown the inverse participation ratio which, although now is defined with respect to a new basis, still gives an indication of how "diagonal" the matrix $P_m Q_n^* H_3 Q_n|_{P_m\mathcal{H}}$ is. If we increase n further, the output approaches the edge of the spectrum with eigenvectors becoming more localised (in the new basis). We found exactly the same phenomena to occur if we shifted the operator H_3, with convergence to the corresponding extremal part of the essential spectrum.

Example 6.11 (*NSA Anderson model in superconductors*) Finally, we consider a non-normal operator with no inessential spectrum where the IQR algorithm does not seem to converge to the boundary of the essential spectrum, but rather to a curve associated with periodic boundary conditions in the large system size limit.

Over the past twenty years there has been considerable interest in non-self-adjoint random operators, sparked by Hatano and Nelson studying a non-self-adjoint Anderson model in the context of vortex pinning in type-II superconductors [39]. Their model showed that an imaginary gauge field in a disordered one-dimensional lattice can induce a delocalisation transition. The operator in $\mathcal{B}(l^2(\mathbb{Z}))$ can be written as

$$(H_4 x)_n = \exp(-g)x_{n-1} + \exp(g)x_{n+1} + V_n x_n \tag{6.7}$$

where $g > 0$ and V is a random potential. This operator also has applications in population biology [52] and the self-adjoint version of this model is widely studied for the phenomenon of Anderson localisation (absence of diffusion of waves) [2,12]. In the non-self-adjoint case, complex values of the spectrum indicate delocalisation. Note that we now have randomness on the diagonal with fixed coupling coefficients $\exp(\pm g)$.

Standard finite section produces real eigenvalues since the matrix $P_m H_4|_{P_m\mathcal{H}}$ is similar to a real symmetric matrix. However, truncating the operator and adopting periodic boundary conditions gives rise to the famous "bubble and wings". If $V = 0$ then the spectrum is an ellipse $E = \{\exp(g + i\theta) + \exp(-g - i\theta) : \theta \in [0, 2\pi)\}$, but as we increase the randomness wings appear on the real axis. For a study of this phenomenon and the described phase transition we refer the reader to [31]. Goldsheid and Khoruzhenko have studied the convergence of the spectral measure in the periodic case as $N \to \infty$ in [33], N being the number of sites. In general, the support of these measures as $N \to \infty$ can be very different from the spectrum of the operator on $l^2(\mathbb{Z})$ given by (6.7), highlighting the difficulty in computing the spectrum.

We consider the case $g = 1/2$ with V_n i.i.d. Bernoulli random variables taking values in $\{\pm 1\}$ with equal probability $p = 1/2$. Again, there is no inessential spectrum and the spectrum/pseudospectrum is constant almost surely, depending only on the support of the distribution of the V_n. The following inclusion is also known, which bounds the spectrum:

$$\sigma(H_4) \subset (\overline{\mathrm{conv}}(E) + [-1, 1]) \cap (E + B_1),$$

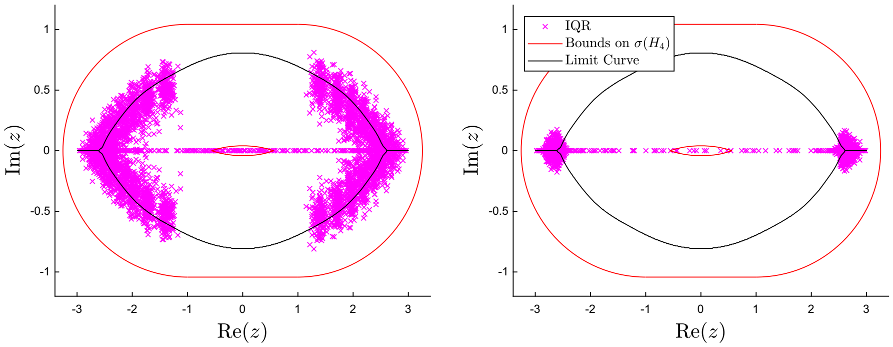

Fig. 11 The output of IQR over 200 samples computing $\sigma(P_m Q_n^* H_4 Q_n|_{P_m\mathcal{H}})$ for $m = 30$ and $n = 15$ (left), $n = 300$ (right). Note that we appear to recover the periodic limit curve and increasing the number of iterates converges to the extremal parts. Applying shifts allowed us to recover the extremal parts of the limit curves

where $\overline{\text{conv}}(E)$ its closed convex hull of E and B_1 denotes the closed unit disk. The choice of g ensures the spectrum has a hole in it. One may calculate the Lyapunov exponent, either by the transfer matrix approach or by calculating a potential related to the density of states. The limiting distribution of the eigenvalues of finite section with periodic boundary conditions is given by the complex curve

$$\{z \in \mathbb{C}\backslash\mathbb{R} : \kappa(z) = 0\} \cup \{x \in \text{supp}(dN) : \kappa(x + i0) > 0\}.$$

The output of the IQR algorithm for $m = 30$ and $n = 15$ and $n = 300$ over 200 random samples are shown in Fig. 11. Note that if we took $n = 0$, the spectrum would be real in stark contrast to Fig. 11. Taking a small number of IQR iterates approximates the bubble and wings with a few remaining real eigenvalues. However, upon increasing n, the output does not seem to converge to the extremal parts of the spectrum, but seems to remain stuck on the limit curve with the operator $P_m Q_n^* H_4 Q_n|_{P_m\mathcal{H}}$. Shifting by $+4iI$ caused the output to recover the top part of the limit curve.

Remark 6.12 For any operator T that has Q_n unitary, the essential spectrum and spectrum of $Q_n^* T Q_n$ is equal to that of T. As the above two examples suggest, taking a small value of n could be used as a method of testing eigenvalues of finite section methods that correspond to finite system size effects, such as open boundary conditions. This could be used in quasi-periodic systems or systems with very few symmetries, where there is no obvious choice of appropriate boundary conditions. However, detecting isolated eigenvalues of finite multiplicity within the convex hull of the essential spectrum still remains a challenge.

7 Concluding remarks and open problems

This paper discussed the generalisation of the famous QR algorithm to infinite dimensions. It was shown that for a large class of operators, encompassing many in scientific applications, the iterates of the IQR algorithm can be computed efficiently on a computer. For matrices with finitely many entries in each column, the computation collapses to a finite one. In general, for an invertible operator, we can compute the iterates to any given accuracy in finite time. Furthermore, it was proven that for normal operators, the algorithm converges to the discrete spectrum outside the convex hull of the essential spectrum, with the rate of convergence generalising the well-known result in finite dimensions. These were extended to more general invariant subspaces and non-normal operators in Theorems 3.13 and 3.15. Unfortunately, the IQR algorithm cannot, in general, be sped up with the use of shift strategies, which considerably speed up the finite-dimensional algorithm [57]. This is due to two reasons. The first is simply that there is no final column of an infinite matrix, hence the usual link with inverse iteration cannot be made. Second, it is also possible for part of the spectrum to be lost in the limit (see Example 3.11), and below the essential spectral radius there is no guarantee of convergence.

Despite these inherent drawbacks of the infinite-dimensional setting, we showed how the IQR algorithm can be used to gain new classification results and convergent algorithms in the SCI hierarchy. In particular, we showed how to compute eigenvalues and eigenspaces outside the essential spectrum with error control for normal operators. This was extended to dominant invariant subspaces for general (possibly non-normal) operators as well as the spectrum of a large class of operators that includes compact normal operators with eigenvalues of distinct magnitude. These results present the first such algorithms that tackle these problems with error control.

Finally, we demonstrated that the IQR algorithm can be implemented both in theory and in practice. We demonstrated the convergence theorems in Sect. 3 as well as some examples (normal and non-normal) where the extremal parts of the essential spectrum also appear to be recovered. Based on this, we conjecture that there may be a large class of operators for which the IQR algorithm converges to the extreme parts of the essential spectrum.[4] In particular, we conjecture that this holds for normal operators if the set of extremal points of the essential spectrum has size one. However, an example was given where convergence to the essential spectrum was only algebraic $O(n^{-\alpha})$ as $n \rightarrow \infty$ as opposed to the linear convergence rate $O(r^n)$ to the discrete spectrum/eigenvalues. It was also demonstrated that the truncations of the IQR algorithm have spectra agreeing with periodic boundary conditions for a range of operators in the class of "pseudoergodic" NSA random operators. In some cases, the algorithm performed much better than standard finite section methods. We should stress that, as in the case of the finite section method, for fixed n and $m \rightarrow \infty$, the output $\sigma(P_m Q_n^* T Q_n|_{P_m})$ will in general still suffer from the spectral pollution phenomenon and in some cases not recover the full spectrum. This was apparent in the numerical examples and is likely to hold for many operators even when taking a mixture of double limits $m, n \rightarrow \infty$. However, examples of Laurent operators were given where it appears $\sigma(P_m Q_n^* T Q_n|_{P_m})$ converges to the spectrum as $m \rightarrow \infty$ for fixed $n > 0$ but not $n = 0$. We hope that the algorithm's potential use in sifting out spectral pollution/complying with appropriate boundary conditions via a canonical unitary transformation can also be exploited.

[4] This is false in general as is easily seen by considering the shift operator.

Based on our findings, we end with a list of open problems for further study on the theoretical properties of the IQR algorithm:

- Which conditions are needed on a possibly non-normal operator in order for the IQR algorithm to pick up the extreme points of the essential spectrum?
- Is the convergence rate to non-isolated points of the spectrum algebraic?
- For operators which do not have a trivial QR decomposition, is there a way of choosing $n = n(m)$ such that $\sigma(P_m Q_{n(m)}^* T Q_{n(m)}|_{P_m})$ converges to the spectrum as $m \to \infty$? If not, then for which classes of operators does such a choice exist?
- Is there a link between the IQR algorithm and the finite section method with periodic boundary conditions for the class of pseudoergodic operators?
- Are there other cases where the IQR algorithm alleviates the need to provide natural boundary conditions when applying the finite section method?
- Extending the IQR algorithm to unbounded operators. Can the IQR algorithm also be extended to a continuous version for differential operators?

Acknowledgements MJC acknowledges support from the UK Engineering and Physical Sciences Research Council (EPSRC) Grant EP/L016516/1. ACH acknowledges support from a Royal Society University Research Fellowship as well as EPSRC Grant EP/L003457/1.We would also like to thank the referees whose comments and suggestions led to the improvement of the manuscript.

A Appendix

A.1 Example codes

Here we show example code for the IQR algorithm in the case that the matrix has k subdiagonals. The code can easily be adapted for the more general case considered in Sect. 4.1.

Algorithm A.1

```
% The Infinite_QR(A,n,k,m) takes a section P_{nk+m}AP_{nk+m}
% of an infinite matrix A with k subdiagonals, performs n iterations
% of the infinite-dimensional QR algorithm and returns
% J = P_mQ_n*AQ_nP_m.

function J = Infinite_QR(A,n,k,m)
d = size(A,2);
 for j=1:n
   A = Inf_QR(A,d-j*k,k);   % The output in each loop is actually
 end                        % U_(d-j*k)...U_1A_(j-1)U_1...U_(d-j*k)
J = A(1:m,1:m);             % if A_j is the j-th term in the QR iteration.
```

Algorithm A.2

```
% Inf_QR(A,n,k) takes a matrix A with k subdiagonals and performs
% multiplication by n Householder transformation from the left and
% right, i.e. B = U_n...U_1AU_1...U_n.

function B = Inf_QR(A,n,k)
B = A; d = size(A,1);
 for j = 1:n
    u = House(A(j:j+k,j));
    A(j:j+k,j:d) = A(j:j+k,j:d) - 2*u*(u'*A(j:j+k,j:d));
    B(j:j+k,1:d) = B(j:j+k,1:d) - 2*u*(u'*B(j:j+k,1:d));
    B(1:d,j:j+k) = B(1:d,j:j+k) - 2*(B(1:d,j:j+k)*u)*u';
 end
```

Algorithm A.3

```
% House(x) takes a vector x and creates a unit vector u
% such that (I - 2u*u')x = ce_1 where c is some complex
% number (depending on x) and e_1 = [1,0,\ldots].

function u = House(x)
v = x;
if v(1) == 0
   v(1) = v(1) + norm(v);             %This is the classical way
else                                  %of creating Householder reflections
   v(1) = x(1) + sign(x(1))*norm(x);  %as in finite dimensions.
end
u = v/norm(v);
```

A.2 Recalling the basics of the SCI hierarchy

The cornerstone in the SCI hierarchy is the definition of a computational problem, a general algorithm and towers of algorithms. The basic objects in a computational problem are as follows:

(i) Ω is some set, called the *domain*.
(ii) Λ is a set of complex-valued functions on Ω called the *evaluation* set.
(iii) $\mathcal{M}$ is a metric space with metric $d_{\mathcal{M}}$.
(iv) $\Xi : \Omega \to \mathcal{M}$ is called the *problem function*.

The set Ω is the set of objects that give rise to our computational problems. The problem function $\Xi : \Omega \to \mathcal{M}$ is what we are interested in computing. Moreover, the set Λ is the collection of functions that provide us with the information we are allowed to read. This leads to the following definition.

Definition A.4 (*Computational Problem*) Given a primary set Ω, an evaluation set Λ, a metric space $\mathcal{M}$ and a problem function $\Xi : \Omega \to \mathcal{M}$, we call the collection $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ a *computational problem*.

For instance, when computing the spectrum of bounded operators on $l^2(\mathbb{N})$, we let Ω be a subset of $\mathcal{B}(l^2(\mathbb{N}))$ (for example the set of self-adjoint operators or compact operators), $(\mathcal{M}, d)$ be the set of all non-empty compact subsets of $\mathbb{C}$ provided with the Hausdorff metric $d = d_H$ in (1.2). The evaluation functions in Λ consist of the family of all functions $f_{i,j} : A \mapsto \langle Ae_j, e_i\rangle$, $i, j \in \mathbb{N}$, which provide the entries of the matrix representation of A with respect to the canonical basis $\{e_i\}_{i\in\mathbb{N}}$. Finally, $\Xi : A \mapsto \sigma(A)$.

The goal is to find algorithms which approximate the function Ξ. More generally, the main pillar of our framework is the concept of a tower of algorithms, which is needed to describe problems that need several limits in the computation. However, first one needs the definition of a general algorithm.

Definition A.5 (*General Algorithm*) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, a *general algorithm* is a mapping $\Gamma : \Omega \to \mathcal{M}$ such that for each $A \in \Omega$

(i) there exists a finite subset of evaluations $\Lambda_\Gamma(A) \subset \Lambda$,
(ii) the action of Γ on A only depends on $\{A_f\}_{f \in \Lambda_\Gamma(A)}$ where $A_f := f(A)$,
(iii) for every $B \in \Omega$ such that $B_f = A_f$ for every $f \in \Lambda_\Gamma(A)$, it holds that $\Lambda_\Gamma(B) = \Lambda_\Gamma(A)$.

Note that the definition of a general algorithm is more general than the definition of a Turing machine or a Blum–Shub–Smale (BSS) machine. A general algorithm has no restrictions on the operations allowed. The only restriction is that it can only take a finite amount of information, though it is allowed to *adaptively* choose the finite amount of information it reads depending on the input. Condition (iii) assures that the algorithm reads the information in a consistent way. Note that the purpose of such a general definition is to get strong lower bounds. In particular, the more general the definition is, the stronger a proven lower bound will be.

With a definition of a general algorithm we can define the concept of towers of algorithms. However, before we define that, we will discuss the cases for which we may have a set-valued function.

Remark A.6 (*Set-valued functions*) Occasionally we will consider a function Ξ such that for $T \in \Omega$ we have that $\Xi(T) \subset \mathcal{M}$. In this case, we will still require that a general algorithm produces a single valued out put i.e $\Gamma(T) \in \mathcal{M}$ for $T \in \Omega$. However, we replace the metric in order to define convergence. In particular, $\Gamma_n(T) \to \Xi(T)$, as $n \to \infty$ means

$$\inf_{y \in \Xi(T)} d_{\mathcal{M}}(\Gamma_n(T), y) \to 0.$$

Definition A.7 (*Tower of Algorithms*) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, a *tower of algorithms of height k for* $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ is a family of sequences of functions

$$\Gamma_{n_k} : \Omega \to \mathcal{M},\ \Gamma_{n_k, n_{k-1}} : \Omega \to \mathcal{M}, \dots,\ \Gamma_{n_k, \dots, n_1} : \Omega \to \mathcal{M},$$

where $n_k, \dots, n_1 \in \mathbb{N}$ and the functions $\Gamma_{n_k, \dots, n_1}$ at the "lowest level" of the tower are general algorithms in the sense of Definition A.5. Moreover, for every $A \in \Omega$,

$$\Xi(A) = \lim_{n_k \to \infty} \Gamma_{n_k}(A), \quad \Gamma_{n_k, \dots, n_{j+1}}(A) = \lim_{n_j \to \infty} \Gamma_{n_k, \dots, n_j}(A) \quad j = k-1, \dots, 1.$$

In addition to a general tower of algorithms (defined above), we will focus on radical towers. The definition of a general algorithm allows for strong lower bounds, however, to produce upper bounds we must add structure to the algorithm and towers of algorithms. A radical tower allows for arithmetic operations, comparisons and radicals.

Definition A.8 (*Radical Towers*) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, a *Radical Tower of Algorithms* of height k for $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ is a tower of algorithms where the lowest level functions

$$\Gamma = \Gamma_{n_k,\ldots,n_1} : \Omega \to \mathcal{M}$$

satisfy the following: For each $A \in \Omega$ the action of Γ on A consists of only finitely many arithmetic operations, comparisons and radicals ($\sqrt{\cdot}$) of positive numbers on $\{A_f\}_{f\in\Lambda_\Gamma(A)}$, where $A_f = f(A)$.

In other words one may say that for the finitely many steps of the computation of the lowest functions $\Gamma = \Gamma_{n_k,\ldots,n_1} : \Omega \to \mathcal{M}$ only the four arithmetic operations $+, -, \cdot, /$ within the smallest (algebraic) field which is generated by the input $\{A_f\}_{f\in\Lambda_\Gamma(A)}$ are allowed. In addition, we allow the extraction of radicals of positive real numbers. We implicitly assume that any complex number can be decomposed into a real and an imaginary part, and moreover we can determine whether $a = b$ or $a > b$ for all real numbers a, b which can occur during the computations. Given the definitions above we can now define the key concept, namely, the Solvability Complexity Index:

Definition A.9 (*Solvability Complexity Index*) A computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ is said to have *Solvability Complexity Index* $\mathrm{SCI}(\Xi, \Omega, \mathcal{M}, \Lambda)_\alpha = k$, with respect to a tower of algorithms of type α, if k is the smallest integer for which there exists a tower of algorithms of type α of height k. If no such tower exists then $\mathrm{SCI}(\Xi, \Omega, \mathcal{M}, \Lambda)_\alpha = \infty$. If there exists a tower $\{\Gamma_n\}_{n\in\mathbb{N}}$ of type α and height one such that $\Xi = \Gamma_{n_1}$ for some $n_1 < \infty$, then we define $\mathrm{SCI}(\Xi, \Omega, \mathcal{M}, \Lambda)_\alpha = 0$. We may sometimes write $\mathrm{SCI}(\Xi, \Omega)_\alpha$ to simplify notation when $\mathcal{M}$ and Λ are obvious.

The definition of the SCI immediately induces the SCI hierarchy:

Definition A.10 (*The Solvability Complexity Index Hierarchy*) Consider a collection $\mathcal{C}$ of computational problems and let $\mathcal{T}$ be the collection of all towers of algorithms of type α for the computational problems in $\mathcal{C}$. Define

$$\begin{aligned}\Delta_0^\alpha &:= \{\{\Xi, \Omega\} \in \mathcal{C} \mid \mathrm{SCI}(\Xi, \Omega)_\alpha = 0\}\\ \Delta_{m+1}^\alpha &:= \{\{\Xi, \Omega\} \in \mathcal{C} \mid \mathrm{SCI}(\Xi, \Omega)_\alpha \leq m\}, \qquad m \in \mathbb{N},\end{aligned}$$

as well as

$$\Delta_1^\alpha := \{\{\Xi, \Omega\} \in \mathcal{C} \mid \exists\, \{\Gamma_n\}_{n\in\mathbb{N}} \in \mathcal{T} \text{ s.t. } \forall A \in \Omega,\ d(\Gamma_n(A), \Xi(A)) \leq 2^{-n}\}.$$

Remark A.11 (*The Δ_k notation*) Note that in this paper we only consider radical towers and hence the superscript α will be omitted throughout. Thus we will always write Δ_k.

Finally, we recall the definition of Σ_1^α.

$$\begin{aligned}\Sigma_1^\alpha = \{\{\Xi, \Omega\} \in \Delta_2^\alpha \mid \exists\, \{\Gamma_n\}_{n\in\mathbb{N}} \in \mathcal{T}\\ \text{s.t. } \Gamma_n(A) \subset \mathcal{N}_{2^{-n}}(\Xi(A)) \text{ and } \Gamma_n(A) \to \Xi(A)\ \forall A \in \Omega\}\end{aligned}$$

where $\mathcal{N}_\delta(\omega)$ denotes the δ-neighbourhood of $\omega \subset \mathcal{M}$.

References

1. Amir, A., Hatano, N., Nelson, D.R.: Non-Hermitian localization in biological networks. Phys. Rev. E **93**(4), 042310 (2016)
2. Anderson, P.W.: Absence of diffusion in certain random lattices. Phys. Rev. **109**(5), 1492 (1958)
3. Aronszajn, N.: Approximation methods for Eigenvalues of completely continuous symmetric operators. In: Proceedings of the Symposium on Spectral Theory and Differential Problems, pp. 179–202 (1951)
4. Arveson, W.: Improper filtrations for C^*-algebras: spectra of unilateral tridiagonal operators. Acta Sci. Math. (Szeged) **57**(1–4), 11–24 (1993)
5. Arveson, W.: Noncommutative spheres and numerical quantum mechanics. In: Operator Algebras, Mathematical Physics, and Low-dimensional Topology (Istanbul, 1991), Research Notes in Mathematics, vol. 5, A K Peters, Wellesley, pp. 1–10 (1993)
6. Arveson, W.: C^*-algebras and numerical linear algebra. J. Funct. Anal. **122**(2), 333–360 (1994)
7. Arveson, W.: The role of C^*-algebras in infinite-dimensional numerical linear algebra. In: C^*-algebras: 1943–1993 (San Antonio, TX, 1993), Contemporary Mathematics, vol. 167, Amer. Math. Soc., Providence, RI, pp. 114–129 (1994)
8. Ben-Artzi, J., Colbrook, M.J., Hansen, A.C., Nevanlinna, O., Seidel, M.: On the Solvability Complexity Index Hierarchy and Towers of Algorithms (Preprint) (2018)
9. Ben-Artzi, J., Hansen, A.C., Nevanlinna, O., Seidel, M.: New barriers in complexity theory: on the solvability complexity index and the towers of algorithms. Comput. Rend. Math. **353**(10), 931–936 (2015)
10. Bender, C.M.: Making sense of non-Hermitian Hamiltonians. Rep. Prog. Phys. **70**(6), 947 (2007)
11. Bender, C.M., Boettcher, S.: Real spectra in non-Hermitian Hamiltonians having PT symmetry. Phys. Rev. Lett. **80**(24), 5243 (1998)
12. Billy, J., Josse, V., Zuo, Z., Bernard, A., Hambrecht, B., Lugan, P., Clément, D., Sanchez-Palencia, L., Bouyer, P., Aspect, A.: Direct observation of Anderson localization of matter waves in a controlled disorder. Nature **453**(7197), 891–894 (2008)
13. Bögli, S., Brown, B.M., Marletta, M., Tretter, C., Wagenhofer, M.: Guaranteed resonance enclosures and exclosures for atoms and molecules. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 470, no. 2171 (2014)
14. Böttcher, A.: Pseudospectra and singular values of large convolution operators. J. Integral Equ. Appl. **6**(3), 267–301 (1994)
15. Böttcher, A.: Infinite matrices and projection methods. In: Lectures on Operator Theory and Its Applications (Waterloo, ON, 1994), Fields Institute Monographs, vol. 3, Amer. Math. Soc., Providence, pp. 1–72 (1996)
16. Böttcher, A., Brunner, H., Iserles, A., Nørsett, S.P.: On the singular values and eigenvalues of the Fox-Li and related operators. N. Y. J. Math. **16**, 539–561 (2010)
17. Böttcher, A., Chithra, A.V., Namboodiri, M.N.N.: Approximation of approximation numbers by truncation. Integral Equ. Oper. Theory **39**(4), 387–395 (2001)
18. Böttcher, A., Silbermann, B.: Introduction to Large Truncated Toeplitz Matrices. Springer, New York (1999)
19. Böttcher, A., Spitkovsky, I.M.: A gentle guide to the basics of two projections theory. Linear Algebra Appl. **432**(6), 1412–1459 (2010)
20. Boulton, L.: Projection methods for discrete Schrödinger operators. Proc. Lond. Math. Soc. **88**(2), 526–544 (2004)
21. Brown, N.: Quasi-diagonality and the finite section method. Math. Comput. **76**(257), 339–360 (2007)
22. Brunner, H., Iserles, A., Nørsett, S.P.: The computation of the spectra of highly oscillatory Fredholm integral operators. J. Integral Equ. Appl. **23**(4), 467–519 (2011)
23. Chandler-Wilde, S., Chonchaiya, R., Lindner, M.: Eigenvalue problem meets Sierpinski triangle: computing the spectrum of a non-self-adjoint random operator. Oper. Matrices **5**(4), 633–648 (2011)
24. Colbrook, M.J., Roman, B., Hansen, A.: How to Compute Spectra with Error Control (Preprint) (2019)
25. Davies, E.B.: Spectral enclosures and complex resonances for general self-adjoint operators. LMS J. Comput. Math. **1**, 42–74 (1998)

26. Davies, E.B.: Linear Operators and Their Spectra, vol. 106. Cambridge University Press, Cambridge (2007)
27. Dean, C., Wang, L., Maher, P., Forsythe, C., Ghahari, F., Gao, Y., Katoch, J., Ishigami, M., Moon, P., Koshino, M., et al.: Hofstadter's butterfly and the fractal quantum Hall effect in moire superlattices. Nature **497**(7451), 598–602 (2013)
28. Deift, P., Li, L., Tomei, C.: Toda flows with infinitely many variables. J. Funct. Anal. **64**(3), 358–402 (1985)
29. Digernes, T., Varadarajan, V.S., Varadhan, S.: Finite approximations to quantum systems. Rev. Math. Phys. **6**(04), 621–648 (1994)
30. Doyle, P., McMullen, C.: Solving the quintic by iteration. Acta Math. **163**(3–4), 151–180 (1989)
31. Feinberg, J., Zee, A.: Non-Hermitian localization and delocalization. Phys. Rev. E **59**(6), 6433 (1999)
32. M. C. T. for MATLAB 4.5.3.12856. Advanpix LLC., Yokohama, Japan
33. Goldsheid, I.Y., Khoruzhenko, B.A.: Distribution of eigenvalues in non-Hermitian Anderson models. Phys. Rev. Lett. **80**(13), 2897 (1998)
34. Gray, R.M., et al.: Toeplitz and circulant matrices: a review. Found. Trends Commun. Inf. Theory **2**(3), 155–239 (2006)
35. Hagen, R., Roch, S., Silbermann, B.: C^*-algebras and numerical analysis. In: Monographs and Textbooks in Pure and Applied Mathematics, vol. 236, Marcel Dekker Inc., New York (2001)
36. Hansen, A.C.: On the approximation of spectra of linear operators on Hilbert spaces. J. Funct. Anal. **254**(8), 2092–2126 (2008)
37. Hansen, A.C.: Infinite-dimensional numerical linear algebra: theory and applications. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **466**(2124), 3539–3559 (2010)
38. Hansen, A.C.: On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators. J. Am. Math. Soc. **24**(1), 81–124 (2011)
39. Hatano, N., Nelson, D.R.: Localization transitions in non-Hermitian quantum mechanics. Phys. Rev. Lett. **77**(3), 570 (1996)
40. Holz, D.E., Orland, H., Zee, A.: On the remarkable spectrum of a non-Hermitian random matrix model. J. Phys. A Math. Gen. **36**(12), 3385 (2003)
41. Kato, T.: Perturbation Theory for Linear Operators, vol. 132. Springer, Berlin (2013)
42. Krein, M., Krasnoselski, M.: Fundamental theorems concerning the extension of Hermitian operators and some of their applications to the theory of orthogonal polynomials and the moment problem. Uspekhi Mat. Nauk. **2**, 60–106 (1947)
43. Levitin, M., Shargorodsky, E.: Spectral pollution and second-order relative spectra for self-adjoint operators. IMA J. Numer. Anal. **24**(3), 393–416 (2004)
44. Lindner, M.: Infinite matrices and their finite sections. In: Frontiers in Mathematics: An Introduction to the Limit Operator Method, Birkhäuser Verlag, Basel (2006)
45. Makris, K.G., El-Ganainy, R., Christodoulides, D.N., Musslimani, Z.H.: Beam dynamics in PT symmetric optical lattices. Phys. Rev. Lett. **100**(10), 103904 (2008)
46. Marletta, M.: Neumann–Dirichlet maps and analysis of spectral pollution for non-self-adjoint elliptic PDEs with real essential spectrum. IMA J. Numer. Anal. **30**(4), 917–939 (2010)
47. Marletta, M., Scheichl, R.: Eigenvalues in spectral gaps of differential operators. J. Spectr. Theory **2**(3), 293–320 (2012)
48. Mattis, D.C.: The few-body problem on a lattice. Rev. Mod. Phys. **58**(2), 361 (1986)
49. McMullen, C.: Families of rational maps and iterative root-finding algorithms. Ann. Math. **125**(3), 467–493 (1987)
50. McMullen, C.: Braiding of the attractor and the failure of iterative algorithms. Invent. Math. **91**(2), 259–272 (1988)
51. Mogilner, A.: Hamiltonians in solid state physics as multiparticle discrete Schrödinger operators. Adv. Soc. Math. **5**, 139–194 (1991)
52. Nelson, D.R., Shnerb, N.M.: Non-Hermitian localization and population biology. Phys. Rev. E **58**(2), 1383 (1998)
53. Olver, S.: ApproxFun.jl v0.8. github (online). https://github.com/JuliaApproximation/ApproxFun.jl (2018)
54. Olver, S., Townsend, A.: A fast and well-conditioned spectral method. SIAM Rev. **55**(3), 462–489 (2013)

55. Olver, S., Townsend, A.: A practical framework for infinite-dimensional linear algebra. In: Proceedings of the 1st First Workshop for High Performance Technical Computing in Dynamic Languages, HPTCDL '14, Piscataway, NJ, USA, IEEE Press, pp. 57–62 (2014)
56. Olver, S., Webb, M.: SpectralMeasures.jl. github (online). https://github.com/JuliaApproximation/SpectralMeasures.jl (2018)
57. Parlett, B.N.: The Symmetric Eigenvalue Problem, vol. 20. siam, Bangkok (1998)
58. Pokrzywa, A.: Method of orthogonal projections and approximation of the spectrum of a bounded operator. Stud. Math. **65**(1), 21–29 (1979)
59. Ponomarenko, L., Gorbachev, R., Yu, G., Elias, D., Jalil, R., Patel, A., Mishchenko, A., Mayorov, A., Woods, C., Wallbank, J., et al.: Cloning of Dirac fermions in graphene superlattices. Nature **497**(7451), 594–597 (2013)
60. Regensburger, A., Bersch, C., Miri, M.-A., Onishchukov, G., Christodoulides, D.N., Peschel, U.: Parity-time synthetic photonic lattices. Nature **488**(7410), 167–171 (2012)
61. Riddell, R.: Spectral concentration for self-adjoint operators. Pac. J. Math. **23**(2), 377–401 (1967)
62. Schmidt, P., Spitzer, F.: The Toeplitz matrices of an arbitrary Laurent polynomial. Math. Scand. **8**(1), 15–38 (1960)
63. Seidel, M.: On (N, ϵ)-pseudospectra of operators on Banach spaces. J. Funct. Anal. **262**(11), 4916–4927 (2012)
64. Seidel, M., Silbermann, B.: Finite sections of band-dominated operators—norms, condition numbers and pseudospectra. In: Operator Theory, Pseudo-differential Equations, and Mathematical Physics, Operator Theory: Advances and Applications, vol. 228, Birkhauser/Springer Basel AG, Basel, pp. 375–390 (2013)
65. Shargorodsky, E.: Geometry of higher order relative spectra and projection methods. J. Oper. Theory **44**(1), 43–62 (2000)
66. Shargorodsky, E.: On the limit behaviour of second order relative spectra of self-adjoint operators. J. Spectr. Theory **3**, 535–552 (2013)
67. Shivakumar, P., Sivakumar, K., Zhang, Y.: Infinite Matrices and Their Recent Applications. Springer, Berlin (2016)
68. Simon, B.: The classical moment problem as a self-adjoint finite difference operator. Adv. Math. **137**(1), 82–203 (1998)
69. Smale, S.: The fundamental theorem of algebra and complexity theory. Bull. Am. Math. Soc. (N.S.) **4**(1), 1–36 (1981)
70. Szabo, A., Ostlund, N.S.: Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory. Courier Corporation, Chelmsford (2012)
71. Teschl, G.: Jacobi Operators and Completely Integrable Nonlinear Lattices. American Mathematical Soc, Providence (2000)
72. Trefethen, L.N., Embree, M.: Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators. Princeton University Press, Princeton (2005)
73. Webb, M., Olver, S.: Spectra of Jacobi Operators Via Connection Coefficient Matrices. arXiv preprint. arXiv:1702.03095 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.