

Convergence of the isometric Arnoldi process

S. Helsen
A.B.J. Kuijlaars
M. Van Barel

Report TW 373, November 2003

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Convergence of the isometric Arnoldi process

S. Helsen
A.B.J. Kuijlaars
M. Van Barel

Report TW 373, November 2003

Department of Computer Science, K.U.Leuven

Abstract

It is well known that the performance of eigenvalue algorithms such as the Lanczos and the Arnoldi method depends on the distribution of eigenvalues. Under fairly general assumptions we characterize the region of good convergence for the Isometric Arnoldi Process. We also determine bounds for the rate of convergence and we prove sharpness of these bounds. The distribution of isometric Ritz values is obtained as the minimizer of an extremal problem. We use techniques from logarithmic potential theory in proving these results.

Keywords : Isometric Arnoldi Process, Ritz values, equilibrium distribution, potential theory.

AMS(MOS) Classification : Primary : 15A18, Secondary : 31A05, 31A15, 65F15.

CONVERGENCE OF THE ISOMETRIC ARNOLDI PROCESS*

S. HELSEN[†], A.B.J. KUIJLAARS[‡], AND M. VAN BAREL[§]

Abstract. It is well known that the performance of eigenvalue algorithms such as the Lanczos and the Arnoldi method depends on the distribution of eigenvalues. Under fairly general assumptions we characterize the region of good convergence for the Isometric Arnoldi Process. We also determine bounds for the rate of convergence and we prove sharpness of these bounds. The distribution of isometric Ritz values is obtained as the minimizer of an extremal problem. We use techniques from logarithmic potential theory in proving these results.

Key words. Isometric Arnoldi Process, Ritz values, equilibrium distribution, potential theory

AMS subject classifications. 15A18, 31A05, 31A15, 65F15

1. Introduction. Unitary eigenvalue problems arise in a number of different fields, for example signal processing and trigonometric approximation problems (for references, see [10]). There exist numerical methods specifically designed to solve such eigenvalue problems. In this article we examine the convergence of one such method: the Isometric Arnoldi Process (IAP), which was introduced by Gragg [15]. Recently, Stewart proved numerical stability of a variant in [25]. Other useful references include [9, 16].

The Arnoldi iteration method is a very popular method to compute some eigenvalues of a matrix. For a unitary matrix $U \in \mathbb{C}^{N \times N}$, the method can be adapted to exploit the structure. Here we give an outline of the method. An orthonormal basis q_1, q_2, \dots, q_N is created for \mathbb{C}^N based on a Gram-Schmidt orthogonalisation of the vectors $b, Ub, U^2b, \dots, U^{N-1}b$ for some starting vector $b \in \mathbb{C}^N$. If Q is the unitary matrix with the q_j as its columns, we get $UQ = QH$ for some unitary Hessenberg matrix H , which necessarily has the same eigenvalues as U . The Arnoldi idea is to look at the $n \times n$ leading principal submatrix H_n of H (for some $n \leq N$) and to compute the eigenvalues of H_n . It is hoped that some of these eigenvalues are good approximants to some of the eigenvalues of U . If the required eigenvalues are indeed approximated and if $n \ll N$, then operating on H_n instead of H can save a considerable amount of computing time.

The matrix H_n is not unitary anymore. In fact, all eigenvalues are strictly *inside* the unit circle. The numbers we want to calculate are *on* the unit circle, so it is very natural to take the approximants also on the unit circle. To this end we modify the matrix H_n . To make it a unitary matrix, it suffices to rescale the last column, and

*The research was partially supported by the Research Council K.U.Leuven, project OT/00/16 (SLAP: Structured Linear Algebra Package), by the Fund for Scientific Research–Flanders (Belgium), projects G.0078.01 (SMA: Structured Matrices and their Applications), G.0176.02 (ANCILA: Asymptotic aNalysis of the Convergence behavior of Iterative methods in numerical Linear Algebra), and G.0455.04 (RHPH: Riemann-Hilbert problems, random matrices and Padé-Hermite approximation), and by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, project IUAP V-22 (Dynamical Systems and Control: Computation, Identification & Modelling). The scientific responsibility rests with the authors.

[†]Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium. (steff@wis.kuleuven.ac.be)

[‡]Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium. (arno@wis.kuleuven.ac.be)

[§]Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium. (Marc.VanBarel@cs.kuleuven.ac.be)

take the eigenvalues of the modified submatrices as approximants. This is the basic idea of the Isometric Arnoldi Process (IAP). In actual implementations of the IAP the computations are done implicitly and involve only the Schur parameters $(\gamma_n)_n$ that are associated with a unitary Hessenberg matrix.

For the convergence of the IAP, it is important to know in what sense the approximation of eigenvalues takes place, and which eigenvalues are well approximated. We will consider this question from the point of view of logarithmic potential theory. Polynomial minimization problems provide the connection between Krylov subspace methods in numerical linear algebra and potential theory, which is clearly explained by Driscoll, Toh, and Trefethen [12]. See also [26, p.279] where one finds the rule of thumb that the Lanczos iteration tends to converge to eigenvalues in regions of “too little charge” for an equilibrium distribution. This rule of thumb for the Lanczos method was made more precise in [5, 18]. It is the aim of this paper to apply similar ideas to the IAP.

Note that potential theory was also used in the recent papers [4, 7, 6, 8, 24] for the convergence analysis of other iterative methods in numerical linear algebra.

The rest of the paper is organized as follows. In the next section we state our main results. Then we collect the properties of unitary Hessenberg matrices and para-orthogonal polynomials that we need for our purposes. In particular we mention a polynomial minimization problem, which is crucial for the link to potential theory. We have not seen this minimization problem in the literature before, but it may be known to specialists in the field. Section 4 contains the proofs of the main results. In the last section we will discuss some numerical experiments that illustrate our theoretical results.

2. Statement of results. The results we obtain will be of an asymptotic nature. We do not investigate the eigenvalues of a single unitary matrix U , but instead we look at a sequence of unitary matrices $(U_N)_N$, with $U_N \in \mathbb{C}^{N \times N}$. This setting reflects for example the discretization of a continuous problem with decreasing mesh size. The eigenvalues and orthonormal eigenvectors of U_N are denoted by $\{\lambda_{k,N}\}_{k=1}^N$ and $\{v_{k,N}\}_{k=1}^N$ respectively. We also take a unit starting vector $b_N \in \mathbb{C}^N$ for every N . For our results, we have to impose a number of mild conditions on the sequence of matrices.

In the conditions, and also in the rest of the paper, the logarithmic potential U^μ of a measure μ appears. This is the function

$$U^\mu(z) = \int \log \frac{1}{|z - z'|} d\mu(z'),$$

which is a harmonic function outside the support of μ . The logarithmic potential U^μ may take the value $-\infty$. Further, δ_λ denotes the Dirac point mass in λ and $\|\cdot\|$ denotes the Euclidian two-norm of a vector. The unit circle in the complex plane is denoted by \mathbb{T} .

CONDITIONS 2.1.

1) There exists a probability measure σ on \mathbb{T} whose logarithmic potential U^σ is real valued and continuous, such that

$$(2.1) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N \delta_{\lambda_{j,N}} = \sigma.$$

2) For every $\varepsilon > 0$ there exists a $\delta \in (0, 1)$ so that for all sufficiently large N and

for all $k \leq N$

$$(2.2) \quad \prod_{\substack{j=1 \\ 0 < |\lambda_{j,N} - \lambda_{k,N}| < \delta}}^N |\lambda_{j,N} - \lambda_{k,N}| > e^{-N\varepsilon}.$$

3) For every N , we have that $\|b_N\| = 1$ and

$$(2.3) \quad \lim_{N \rightarrow \infty} \left(\min_{1 \leq k \leq N} |\langle b_N, v_{k,N} \rangle| \right)^{1/N} = 1.$$

The limit in (2.1) is in the sense of weak*-convergence of measures. In this paper convergence of measures will always be in the weak*-sense, i.e., if ν and ν_n are Borel probability measures on \mathbb{T} then $\nu_n \rightarrow \nu$ if and only if

$$\int f d\nu_n \rightarrow \int f d\nu$$

for every continuous function f on \mathbb{T} . Thus the first condition states that the eigenvalues have a limiting distribution σ . The condition that U^σ is continuous and real valued (and so does not take the value $-\infty$) is a regularity condition on σ . It is satisfied, for example, if σ has a bounded density with respect to the Lebesgue measure on \mathbb{T} . The second condition is a technical one that prevents the eigenvalues from being too close to each other. Beckermann [5, Lemma 2.4(a)] proved that under Condition 1) the second condition is equivalent with

2b) For all sequences $(k_N)_N$ with $k_N \in \{1, \dots, N\}$ such that $\lim_{N \rightarrow \infty} \lambda_{k_N, N} = \lambda$ for some λ , we have

$$(2.2b) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\substack{j=1 \\ j \neq k_N}}^N \log |\lambda_{k_N, N} - \lambda_{j, N}| = \int \log |\lambda - z| d\sigma(z).$$

A discussion about this condition can be found in [19]. The third condition imposes that the starting vectors are sufficiently random, i.e., their eigenvector components are not exponentially small. Since the numbers $|\langle b_N, v_{j,N} \rangle|$ will be used frequently, we introduce a shorter notation:

$$(2.4) \quad w_{j,N} := |\langle b_N, v_{j,N} \rangle|.$$

For every N , we consider the IAP on U_N with starting vector b_N . Iteratively an orthonormal basis is created for the Krylov subspaces

$$\mathcal{K}_{n,N} = \text{span}\{b_N, U_N b_N, U_N^2 b_N, \dots, U_N^{n-1} b_N\}.$$

If we compute a basis for whole \mathbb{C}^N in this way, U_N is represented by a Hessenberg matrix in this basis. The $n \times n$ principal left upper block $H_{n,N}$ of this matrix is the representation of the orthogonal projection of U_N onto $\mathcal{K}_{n,N}$. By modifying the last column of $H_{n,N}$ we can obtain a unitary Hessenberg matrix $\tilde{H}_{n,N}$. The modification depends on a unimodular constant $\rho_{n,N}$, see also §3. The precise value of $\rho_{n,N}$ is not

important for our results, and we will not indicate the dependence on $\rho_{n,N}$ in our notation. Let

$$(2.5) \quad \psi_{n,N}(z) = \det(zI_n - \tilde{H}_{n,N}),$$

where I_n denotes the $n \times n$ identity matrix and let $\theta_{1,n,N}, \theta_{2,n,N}, \dots, \theta_{n,n,N}$ be the zeros of $\psi_{n,N}$. We call these numbers the *Ritz values for the IAP* or the *isometric Ritz values*. Since they are the eigenvalues of $\tilde{H}_{n,N}$, which is a unitary matrix, the isometric Ritz values are on the unit circle. We take the eigenvalues of the matrices U_N and the isometric Ritz values to be numbered counterclockwise, but we do not specify a starting point. We also take $\lambda_{0,N} := \lambda_{N,N}$ and $\theta_{0,n,N} := \theta_{n,n,N}$.

In §3 (see Proposition 3.4 below) we will prove that the isometric Ritz values are separated by the eigenvalues, by which we mean that on the open arc between two consecutive isometric Ritz values there is at least one eigenvalue, or put differently, on the closed arc between any two consecutive eigenvalues there is at most one isometric Ritz value.

We consider the convergence of isometric Ritz values along ray sequences, i.e., we let N approach infinity, and with it also n , in such a fashion that $n/N \rightarrow t$ for some $t \in (0, 1)$. If we consider the points (N, n) in a triangular array then the convergence is taken along a sequence of (N, n) values that are asymptotic to a line with slope t in the N - n plane. We denote a limit in this sense by $\lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}}$.

THEOREM 2.2. *Let (U_N) and (b_N) be such that Conditions 2.1 hold. Then for every $t \in (0, 1)$, there exists a Borel probability measure μ_t , depending only on t and σ , such that*

$$(2.6) \quad \lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \frac{1}{n} \sum_{j=1}^n \delta_{\theta_{j,n,N}} = \mu_t$$

and a real constant F_t such that

$$(2.7) \quad \lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|\psi_{n,N}(U_N)b_N\|^{1/n} = \exp(-F_t).$$

The measure μ_t satisfies

$$(2.8) \quad 0 \leq t\mu_t \leq \sigma, \quad \int d\mu_t = 1$$

and minimizes the logarithmic energy

$$(2.9) \quad I(\mu) = \iint \log \frac{1}{|z - z'|} d\mu(z) d\mu(z')$$

among all measures μ satisfying $0 \leq t\mu \leq \sigma$ and $\int d\mu = 1$. The logarithmic potential U^{μ_t} of μ_t is a continuous function on \mathbb{C} , and the constant F_t is such that

$$(2.10) \quad \begin{cases} U^{\mu_t}(z) = F_t & \text{for } z \in \text{supp}(\sigma - t\mu_t), \\ U^{\mu_t}(z) \leq F_t & \text{for } z \in \mathbb{C}. \end{cases}$$

Furthermore, the relations (2.8) and (2.10) characterize the pair (μ_t, F_t) .

This theorem tells us that the isometric Ritz values have a limiting distribution μ_t if we let $n, N \rightarrow \infty$ in such a way that $n/N \rightarrow t$. The measure μ_t is the minimizer of the logarithmic energy (2.9) under the constraints (2.8). Conditions (2.10) are the Euler-Lagrange variational conditions for this minimization problem and together with (2.8) they also characterize μ_t .

The next theorem shows that in a certain region the isometric Ritz values converge exponentially fast to eigenvalues.

THEOREM 2.3. *Let (U_N) and (b_N) be such that Conditions 2.1 hold and let F_t be as in Theorem 2.2. Then we have for every $t \in (0, 1)$,*

$$(2.11) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \leq \exp(U^{\mu_t}(\lambda) - F_t)$$

for every sequence of indices (k_N) with $1 \leq k_N \leq N$, such that $(\lambda_{k_N, N})_N$ converges to $\lambda \in \mathbb{T}$.

We define the set

$$\Lambda(t, \sigma) := \{\lambda \in \mathbb{T} \mid U^{\mu_t}(\lambda) < F_t\}.$$

This is the region of good convergence of the IAP in the regime we are considering. Inside this set, the right-hand side of (2.11) is strictly less than 1, which indicates that for large N , an eigenvalue $\lambda_{k_N, N}$ of U_N in $\Lambda(t, \sigma)$ is approximated by an isometric Ritz value at a geometric rate. Outside $\Lambda(t, \sigma)$, the right-hand side is just one, and then no convergence can be guaranteed.

In the next theorem, we will show that the convergence rate is actually twice as big, except for perhaps one eigenvalue. It is also proven that this convergence bound is sharp. In the theorem there will appear ‘exceptional indices’: the sharper convergence rate will hold for all indices except for these ‘exceptional indices’.

THEOREM 2.4. *Let (U_N) and (b_N) be such that Conditions 2.1 hold, let F_t be as in Theorem 2.2 and let $\lambda \in \Lambda(t, \sigma)$. Then for every N , there exists at most one index $k_N^*(\lambda) \in \{1, 2, \dots, N\}$ such that*

$$(2.12) \quad \lim_{n/N \rightarrow t} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} = \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right)$$

for every sequence of indices (k_N) with $1 \leq k_N \leq N$, such that $(\lambda_{k_N, N})_N$ converges to λ and satisfying $k_N \neq k_N^*(\lambda)$ for N large enough.

Remark 2.5. The fact that the convergence rate can be doubled was first realized by Beckermann [5] in the context of the convergence of the Lanczos method. He also introduced the exceptional indices. The proof of Theorem 2.4 is based on the proof of [5, Theorem 2.1], but we have streamlined some of the arguments, see Section 4.4 below.

Remark 2.6. It is possible to prove the estimate

$$\limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \leq \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right)$$

under weaker conditions. See [5].

There are two types of possible exceptional behavior for the index $k_N^*(\lambda)$ in Theorem 2.4, namely

$$(2.13a) \quad \min_j |\lambda_{k_N^*(\lambda), N} - \theta_{j, n, N}|^{1/n} \gg \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right),$$

$$(2.13b) \quad \min_j |\lambda_{k_N^*(\lambda), N} - \theta_{j, n, N}|^{1/n} \ll \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right).$$

According to Theorem 2.4 at most one of them can occur for a fixed N . So we have three possible situations: no exception, exception (2.13a) or exception (2.13b). It will depend on the choice of parameter $\rho_{n, N}$ which situation occurs. To show what happens, we will make a classification of the relative positioning of the isometric Ritz values and the eigenvalues in a closed arc $I \subset \Lambda(t, \sigma)$.

It will be shown in Proposition 3.4 that the isometric Ritz values are separated by the eigenvalues, and (2.11) tells us that each eigenvalue in $\Lambda(t, \sigma)$ is approximated at an exponential rate. Since the gaps between eigenvalues are *not* exponentially small (see Lemma 4.2), each Ritz value can be close to a single eigenvalue only, if N is large enough. From this information we can make a complete classification of the relative positions of eigenvalues and isometric Ritz values on the arc $I \subset \Lambda(t, \sigma)$:

- case 1: Each eigenvalue in I is close to exactly one isometric Ritz value and the isometric Ritz value follows closely after the eigenvalue (when looking in the counterclockwise direction).
- case 2: One eigenvalue in I is close to two isometric Ritz value, one on each side of it.
- case 3: One isometric Ritz value in I is not close to an eigenvalue.
- case 4: Each eigenvalue in I is close to exactly one isometric Ritz value and the isometric Ritz value precedes the eigenvalue (when looking in the counterclockwise direction).
- case 5: One isometric Ritz value in I coincides with an eigenvalue.
- case 6: One arc between two consecutive eigenvalues in I contains no isometric Ritz values.

The different cases are illustrated in Figure 2.1. The six cases are mutually exclusive and cover all possibilities. In [2] and [3, Theorem 2.12] one can find a similar description of the zeros of discrete orthogonal polynomials on the real line.

Recall that the IAP depends on the choice of a unimodular constant $\rho_{n, N}$. If we move $\rho_{n, N}$ around the unit circle in the counterclockwise direction, the isometric Ritz values also move in the counterclockwise direction, as shown in Figure 2.1. If we start in case 1, no isometric Ritz value can leave ‘its’ eigenvalue, until an extra isometric Ritz value enters the arc I from the right. Then we are in case 2. Then one isometric Ritz value is free to move away from its eigenvalue, and we pass via case 3 to case 2 again. This process is shown in parts (a)–(d) of Figure 2.1. Continuing this way, we see that the eigenvalue that is well approximated by two isometric Ritz values ‘moves’ through I , until it drops off and we reach case 4 (part (h) of Figure 2.1). We stay in case 4 until the left-most isometric Ritz value reaches ‘its’ eigenvalue. Then one isometric Ritz value exactly coincides with an eigenvalue and we are in case 5. The left-most isometric Ritz value then passes the eigenvalue and we are in case 6, where there are two consecutive eigenvalues without an isometric Ritz value on the arc between them. We refer to this arc as a gap. The gap moves to the right as shown in parts (j)–(n) of Figure 2.1, until we reach case 1 again, see part (p).

Now we turn to the exceptional cases. In cases 1, 3, and 4, there are no exceptions. The exception (2.13a) may occur in case 2. In case 2 there are two isometric Ritz

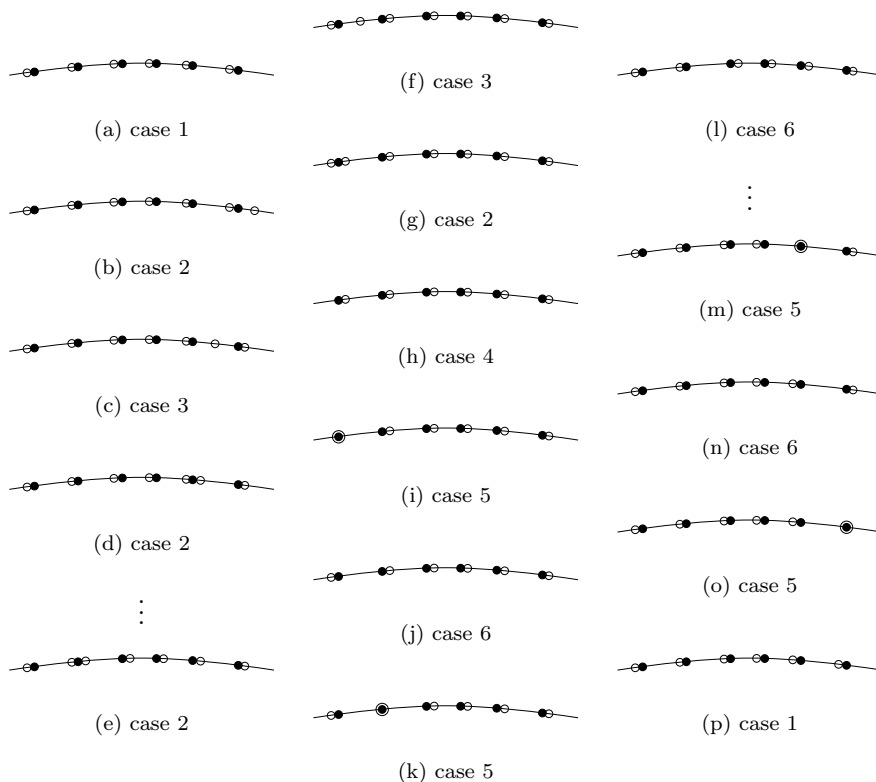


FIGURE 2.1. The evolution of the isometric Ritz values in a closed arc $I \subset \Lambda(t, \sigma)$, when $\rho_{n,N}$ moves counterclockwise around \mathbb{T} . The full dots are the eigenvalues, the open circles are the isometric Ritz values. The possibilities of their location are:

- case 1: An isometric Ritz value follows after each eigenvalue at close distance.
- case 2: Two isometric Ritz values are close to the same eigenvalue.
- case 3: One isometric Ritz value is not close to any eigenvalue.
- case 4: An isometric Ritz value precedes each eigenvalue at close distance.
- case 5: One isometric Ritz value coincides with an eigenvalue.
- case 6: One arc between two eigenvalues contains no isometric Ritz values.

values close to the same eigenvalue. In this case the doubling of the exponent in (2.12) need not take place.

In cases 5 and 6 the exception (2.13b) appears. This is clear if an eigenvalue and an isometric Ritz value coincide, which corresponds to case 5. In case 6 there is a gap and this case arises out of case 5 after a small perturbation of the parameter. For a sufficiently small perturbation, the isometric Ritz value is still closer to the eigenvalue than predicted by (2.12). So in case 6 there may be one eigenvalue around the gap with an isometric Ritz value that is too close to it. This eigenvalue corresponds to the exceptional index. It may be somewhat surprising that only one of the eigenvalues around the gap may be an exception, the other one is not.

3. Unitary Hessenberg matrices and para-orthogonal polynomials. In this section we collect a number of results that can be found in various sources and we put it in a form that is convenient for our purposes. The size N is fixed throughout

this section and will not be indicated in the notation.

We consider a unitary matrix U of size $N \times N$ with simple eigenvalues $\lambda_1, \dots, \lambda_N$ and corresponding normalized eigenvectors v_1, \dots, v_N . We also consider a unit starting vector $b \in \mathbb{C}^N$ with a non-zero component in the direction of every eigenvector. We define a measure

$$\nu = \sum_{j=1}^N w_j^2 \delta_{\lambda_j} = \sum_{j=1}^N |\langle b, v_j \rangle|^2 \delta_{\lambda_j}.$$

Since b is a unit vector and the v_j form an orthonormal basis of \mathbb{C}^N , we have that

$$\int d\nu = \sum_{j=1}^N |\langle b, v_j \rangle|^2 = \|b\|^2 = 1,$$

so that ν is a discrete probability measure supported on the eigenvalues λ_j .

LEMMA 3.1. *For every function $f : \mathbb{T} \rightarrow \mathbb{C}$, we have*

$$\|f(U)b\|^2 = \int |f|^2 d\nu.$$

Proof. Let V be the unitary matrix with the v_j as columns, and Λ the diagonal matrix with the λ_j on the diagonal, so $U = V\Lambda V^*$ is the eigenvalue decomposition of U . Then $f(U) = Vf(\Lambda)V^*$ and, since V is unitary,

$$\|f(U)b\| = \|Vf(\Lambda)V^*b\| = \|f(\Lambda)V^*b\|.$$

Now $f(\Lambda)$ is a diagonal matrix with $f(\lambda_j)$ on the diagonal and V^*b is a vector whose j th component is $v_j^*b = \langle b, v_j \rangle$. Hence

$$\|f(U)b\|^2 = \sum_{j=1}^N |f(\lambda_j)\langle b, v_j \rangle|^2 = \int |f|^2 d\nu,$$

which proves the lemma. \square

If carried out to the end, the IAP transforms the unitary matrix U to the $N \times N$ unitary upper Hessenberg matrix H

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1N} \\ h_{21} & h_{22} & & \\ & h_{32} & \ddots & \vdots \\ & & \ddots & \\ & & & h_{N,N-1} & h_{NN} \end{bmatrix}$$

with real and positive subdiagonal elements $h_{j+1,j} > 0$. The eigenvalues of H and U are the same. The principal leading submatrix of size $n \times n$ will be denoted by H_n . The matrices H_n , $n < N$, are not unitary, since the norm of the last column of H_n is strictly less than one. We define the characteristic polynomials

$$\phi_n(z) = \det(zI_n - H_n).$$

LEMMA 3.2. *The polynomial ϕ_n is the monic polynomial of degree n that is orthogonal with respect to ν .*

Proof. We define polynomials φ_n , $n = 0, \dots, N$, recursively by $\varphi_0(z) \equiv 1$ and

$$(3.1) \quad z\varphi_k(z) = \sum_{j=0}^{k+1} h_{j+1,k+1}\varphi_j(z) \quad \text{for } k = 0, \dots, N-1,$$

where we have put (somewhat arbitrarily) $h_{N+1,N} = 1$. Then we have for $n \leq N$,

$$(3.2) \quad \begin{aligned} & [\varphi_0(z) \quad \varphi_1(z) \quad \cdots \quad \varphi_{n-1}(z)] H_n \\ & = z [\varphi_0(z) \quad \cdots \quad \varphi_{n-1}(z)] - [0 \quad \cdots \quad 0 \quad h_{n+1,n}\varphi_n(z)] \end{aligned}$$

It follows that z is an eigenvalue of H_n if and only if $\varphi_n(z) = 0$ and we have

$$(3.3) \quad \varphi_n(z) = \left(\prod_{j=1}^n h_{j+1,j}^{-1} \right) \det(zI_n - H_n) = \left(\prod_{j=1}^n h_{j+1,j}^{-1} \right) \phi_n(z),$$

see also [13].

From (3.2) with $n = N$, it follows that $[\varphi_0(\lambda_j) \quad \varphi_1(\lambda_j) \quad \cdots \quad \varphi_{N-1}(\lambda_j)]$ is a left eigenvector of H for the eigenvalue λ_j . Let

$$\tilde{w}_j = \|[\varphi_0(\lambda_j) \quad \varphi_1(\lambda_j) \quad \cdots \quad \varphi_{N-1}(\lambda_j)]\|^{-1}$$

so that $[\tilde{w}_j\varphi_0(\lambda_j) \quad \tilde{w}_j\varphi_1(\lambda_j) \quad \cdots \quad \tilde{w}_j\varphi_{N-1}(\lambda_j)]$ is a normalized eigenvector of H . Since the matrix H is unitary (hence normal) with simple spectrum, its normalized eigenvectors form an orthonormal basis of \mathbb{C}^n . Thus

$$S = \begin{bmatrix} \tilde{w}_1\varphi_0(\lambda_1) & \cdots & \tilde{w}_1\varphi_{N-1}(\lambda_1) \\ \vdots & \ddots & \vdots \\ \tilde{w}_N\varphi_0(\lambda_N) & \cdots & \tilde{w}_N\varphi_{N-1}(\lambda_N) \end{bmatrix}$$

is unitary. Then $S^*S = I$ and if we look at the individual matrix entries of this last expression, we find

$$\sum_{j=1}^N \tilde{w}_j^2 \varphi_k(\lambda_j) \varphi_l(\lambda_j) = \delta_{k,l}, \quad \text{for } k, l = 0, 1, \dots, N-1.$$

So the polynomials φ_n are orthonormal polynomials with respect to the measure $\sum_{j=1}^N \tilde{w}_j^2 \delta_{\lambda_j}$, and because of (3.3) we have that the polynomials ϕ_n are the monic orthogonal polynomials with respect to this measure.

Now we show $\tilde{w}_j = |\langle b, v_j \rangle|$ for $j = 1, \dots, N$, to complete the proof of the lemma. We know that $UQ = QH$ where Q is a unitary matrix whose first column is b . From the eigenvalue decomposition $U = V\Lambda V^*$ we get that $V^*QH = \Lambda V^*Q$, which means that v_j^*Q is a normalized left eigenvector of H of the eigenvalue λ_j . Also $[\tilde{w}_j\varphi_0(\lambda_j) \quad \cdots \quad \tilde{w}_j\varphi_{N-1}(\lambda_j)]$ is a normalized left eigenvector with λ_j . Then the first components have the same absolute values. The first column of Q is equal to b so that the first component of v_j^*Q is equal to $v_j^*b = \langle b, v_j \rangle$. Thus we have $\tilde{w}_j = |\tilde{w}_j\varphi_0(\lambda_j)| = |\langle b, v_j \rangle|$. \square

The previous lemma connects the Arnoldi process to the theory of orthogonal polynomials and in particular to the Arnoldi minimization problem, see for example [26]:

Arnoldi minimization problem:

Minimize $\|p_n(U)b\|$ among all monic polynomials p_n of degree n .

It is a general fact that the monic polynomial ϕ_n of degree n which is orthogonal with respect to μ minimizes the $L^2(\mu)$ norm $(\int |p_n|^2 d\mu)^{1/2}$ among all monic polynomials p_n of degree n . Because of Lemma 3.1 it is then clear that ϕ_n is the minimizer in the Arnoldi minimization problem.

We want to establish a similar minimization problem for the isometric Arnoldi process. To that end we first recall that H can be decomposed as a product of Givens reflectors [15]:

$$H = G_1(\gamma_1)G_2(\gamma_2)\cdots G_{N-1}(\gamma_{N-1})\tilde{G}_N(\gamma_N),$$

for some complex parameters γ_j satisfying $|\gamma_j| < 1$ for $j = 1, \dots, N-1$ and $|\gamma_N| = 1$. The matrices $G_j(\alpha)$ are given by

$$G_j(\alpha) = \begin{bmatrix} I_{j-1} & & & & & \\ & -\alpha & \sqrt{1-|\alpha|^2} & & & \\ & \sqrt{1-|\alpha|^2} & \bar{\alpha} & & & \\ & & & & & \\ & & & & & \\ & & & & & I_{N-j-1} \end{bmatrix},$$

and $\tilde{G}_N(\gamma_N)$ is given by

$$\tilde{G}_N(\alpha) = \begin{bmatrix} I_{N-1} & \\ & -\alpha \end{bmatrix}.$$

The numbers γ_j are called the Schur parameters for the unitary Hessenberg matrix H . We use the notation $H = H(\gamma_1, \dots, \gamma_N)$. If we define $\sigma_j := \sqrt{1-|\gamma_j|^2}$ and write out the above product, we get an explicit expression for H in terms of the Schur parameters:

$$H = \begin{bmatrix} -\gamma_1 & -\sigma_1\gamma_2 & -\sigma_1\sigma_2\gamma_3 & \cdots & -\sigma_1\cdots\sigma_{N-2}\gamma_{N-1} & -\sigma_1\cdots\sigma_{N-1}\gamma_N \\ \sigma_1 & -\tilde{\gamma}_1\gamma_2 & -\tilde{\gamma}_1\sigma_2\gamma_3 & & & \\ & \sigma_2 & -\tilde{\gamma}_2\gamma_3 & & & \\ & & \sigma_3 & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma_{N-1} & -\tilde{\gamma}_{N-2}\sigma_{N-1}\gamma_N \\ & & & & & -\tilde{\gamma}_{N-1}\gamma_N \end{bmatrix}.$$

From this expression for the matrix, it is easy to see that $H_n = H(\gamma_1, \dots, \gamma_n)$. Since the matrices $G_j(\alpha)$ have determinant -1 , and $\tilde{G}_n(\alpha)$ has determinant $-\alpha$, it easily follows that [14]

$$(3.4) \quad \phi_n(0) = \det(-H_n) = \gamma_n, \quad \text{for } n = 1, \dots, N.$$

As mentioned before, the IAP modifies the matrix H_n in order to make it unitary. The only thing that needs to change is the length of the last column. To rescale that last column, we construct

$$\tilde{H}_n := H(\gamma_1, \dots, \gamma_{n-1}, \rho_n)$$

with ρ_n a unimodular number. This transformation amounts to multiplying the last column of H_n by the number $\frac{\rho_n}{\gamma_n}$ (provided $\gamma_n \neq 0$). Note that the parameter ρ_n can be anywhere on the unit circle. The matrices \tilde{H}_n do depend on the precise choice of ρ_n , but its location will not be of any importance to us, as can be seen from the theorems. As a consequence, we do not include the dependence on ρ_n in the notation.

We will need the concept of para-orthogonal polynomials. To that end, we recall their definition, see for example [17]. For a polynomial p of degree n , let

$$p^*(z) = z^n \overline{p(1/\bar{z})}$$

be the reciprocal polynomial. The (monic) para-orthogonal polynomials ψ_n are then defined by

$$(3.5) \quad \psi_n(z) := \frac{\phi_n(z) + \omega_n \phi_n^*(z)}{1 + \omega_n \bar{\gamma}_n}$$

where ϕ_n is the monic orthogonal polynomial with respect to the measure ν and $\omega_n \in \mathbb{T}$. Note that in the literature the para-orthogonal polynomials are usually defined as $\phi_n + \omega_n \phi_n^*$ so that they are not monic.

We have to be careful here, since we have already defined a set of polynomials $\psi_{n,N}$ in (2.5). In fact, the two definitions are the same. More precisely, for every $\rho_n \in \mathbb{T}$ there exists an $\omega_n \in \mathbb{T}$, and vice versa for every $\omega_n \in \mathbb{T}$ there exists a $\rho_n \in \mathbb{T}$, such that

$$\psi_n(z) = \det(zI_n - \tilde{H}_n),$$

where ψ_n is defined as in (3.5). The 1-1 correspondence between ρ_n and ω_n is given by

$$(3.6) \quad \rho_n = \omega_n \left(\frac{1 + \bar{\omega}_n \gamma_n}{1 + \omega_n \bar{\gamma}_n} \right), \quad \omega_n = \rho_n \left(\frac{1 - \bar{\rho}_n \gamma_n}{1 - \rho_n \bar{\gamma}_n} \right).$$

This is a consequence of a remark in [1] and is easily verified using the recurrence relations for the orthogonal polynomials and their reciprocals which are stated in [15].

These polynomials are called para-orthogonal since they are orthogonal with respect to ν to all polynomials of degree less than n without constant term [17], that is,

$$(3.7) \quad \int_{\mathbb{T}} \psi_n(z) \bar{z}^k d\nu(z) = 0, \quad k = 1, 2, \dots, n-1.$$

It is known that the zeros of ψ_n are simple and lie on the unit circle [15, 17]. We denote them by $\theta_1, \dots, \theta_n$. We recall that these zeros are the basis of the Gauss quadrature formula on the unit circle [17]

$$(3.8) \quad \sum_{j=1}^n \beta_j p(\theta_j) = \int p d\nu, \quad \beta_j > 0,$$

which is valid for Laurent polynomials p of degree $\leq n-1$ (i.e., for linear combinations of z^k with $k = -n+1, -n+2, \dots, n-2, n-1$).

Our next result, which is the main result of this section, states that the para-orthogonal polynomials solve a minimization problem, similar to the Arnoldi minimization problem. We call it the isometric Arnoldi minimization problem. While this result may be known already, we have not seen it in the literature.

Isometric Arnoldi minimization problem:

Minimize $\|p_n(U)b\|$ among all monic polynomials p_n of degree n satisfying $p_n(0) = \rho_n$, where $\rho_n \in \mathbb{T}$ is given.

THEOREM 3.3. *The minimizer of the Isometric Arnoldi minimization problem is unique and it is given by the monic para-orthogonal polynomial ψ_n where ω_n is related to ρ_n as in (3.6).*

Proof. Let ψ_n be the monic para-orthogonal polynomial of degree n with parameter $\omega_n = \rho_n \left(\frac{1 - \bar{\rho}_n \gamma_n}{1 - \rho_n \bar{\gamma}_n} \right)$. Making the same reasoning as the one leading to (3.4), we find $\psi_n(0) = \rho_n$.

If p_n is an arbitrary monic polynomial of degree n with $p_n(0) = \rho_n$, then $p_n - \psi_n$ is a linear combination of z, z^2, \dots, z^{n-1} , so that by the para-orthogonality property (3.7) we have

$$\int \psi_n(z) \overline{(p_n(z) - \psi_n(z))} d\nu(z) = 0.$$

Thus

$$\int \psi_n \bar{p}_n d\nu = \int |\psi_n|^2 d\nu.$$

This leads to

$$\int |p_n - \psi_n|^2 d\nu = \int |p_n|^2 d\nu - \int |\psi_n|^2 d\nu,$$

from which we deduce that

$$\int |p_n|^2 d\nu \geq \int |\psi_n|^2 d\nu,$$

with equality if and only if $\int |p_n - \psi_n|^2 d\nu = 0$. Since $p_n - \psi_n$ is a polynomial of degree $n - 1$ and the measure ν is carried on N points, equality can only hold if $p_n = \psi_n$. Thus by Lemma 3.1

$$\|p_n(U)b\| \geq \|\psi_n(U)b\|$$

with equality if and only if $p_n = \psi_n$. This proves the theorem. \square

Using Theorem 3.3 we will now prove that the zeros of the para-orthogonal polynomial ψ_n (which are on the unit circle) are separated by the eigenvalues of U .

PROPOSITION 3.4. *Let $n < N$. Then the zeros of ψ_n are separated by the eigenvalues of U .*

Proof. Let θ_1 and θ_2 be two consecutive zeros of ψ_n and assume that there are no eigenvalues on the open arc between θ_1 and θ_2 . Without loss of generality we may restrict ourselves to the case that $\theta_1 = e^{-is_0}$, $\theta_2 = e^{is_0}$ where $s_0 \in (0, \pi)$, and all eigenvalues are of the form $\lambda_j = e^{is_j}$ with $s_0 \leq |s_j| \leq \pi$. Then

$$(3.9) \quad \psi_n(z) = (z - e^{-is_0})(z - e^{is_0})q_{n-2}(z),$$

where q_{n-2} is a polynomial of degree $n - 2$. We know from Theorem 3.3 that ψ_n minimizes

$$\|p_n(U)b\|^2 = \sum_{j=1}^N w_j^2 |p_n(\lambda_j)|^2,$$

among all monic polynomials of degree n with $p_n(0) = \rho_n$ (see also Lemma 3.1). For each s , we have that $(z - e^{-is})(z - e^{is})q_{n-2}(z)$ is a monic polynomial with value ρ_n at $z = 0$. Thus

$$I(s) := \sum_{j=1}^N w_j^2 (|\lambda_j - e^{-is}| |\lambda_j - e^{is}|)^2 |q_{n-2}(\lambda_j)|^2$$

is minimal for $s = s_0$. Observe that

$$\begin{aligned} |\lambda_j - e^{-is}| |\lambda_j - e^{is}| &= |e^{is_j} - e^{-is}| |e^{is_j} - e^{is}| \\ &= 4 \left| \sin \frac{s_j - s}{2} \sin \frac{s_j + s}{2} \right| = 2 |\cos s - \cos s_j|, \end{aligned}$$

so that

$$I(s) = 4 \sum_{j=1}^N w_j^2 (\cos s - \cos s_j)^2 |q_{n-2}(\lambda_j)|^2$$

and therefore

$$(3.10) \quad I'(s_0) = 8 \sum_{j=1}^N -w_j^2 (\cos s_0 - \cos s_j) \sin s_0 |q_{n-2}(\lambda_j)|^2.$$

Since $0 < s_0 \leq |s_j| \leq \pi$, we have $\sin s_0 > 0$ and $\cos s_0 - \cos s_j \geq 0$ for every $j = 1, 2, \dots, N$. There are at least $N - 2$ values of j with $0 < s_0 < |s_j| \leq \pi$ so that $\cos s_0 - \cos s_j > 0$ for at least one j . (We suppose $N > 2$ since otherwise $n \leq N - 1 \leq 1$ and there is nothing to prove.) It follows that all terms in the right-hand side of (3.10) are non-positive and at least one is negative. Hence $I'(s_0) < 0$, which contradicts the fact that $I(s)$ has a minimum for $s = s_0$. The proposition is proved. \square

Remark 3.5. The fact that $I'(s_0) = 0$ means that

$$\sum_{\lambda_j \in [\theta_1, \theta_2]} w_j^2 |\lambda_j - \theta_1| |\lambda_j - \theta_2| |q_{n-2}(\lambda_j)|^2 = \sum_{\lambda_j \in [\theta_2, \theta_1]} w_j^2 |\lambda_j - \theta_1| |\lambda_j - \theta_2| |q_{n-2}(\lambda_j)|^2,$$

where we use $[\theta_1, \theta_2]$ to denote the circular arc from θ_1 to θ_2 and $[\theta_2, \theta_1]$ to denote the complementary arc from θ_2 to θ_1 . We rewrite this in terms of the para-orthogonal polynomial ψ_n as

$$(3.11) \quad \sum_{\lambda_j \in [\theta_1, \theta_2]} w_j^2 \frac{|\psi_n(\lambda_j)|^2}{|\lambda_j - \theta_1| |\lambda_j - \theta_2|} = \sum_{\lambda_j \in [\theta_2, \theta_1]} w_j^2 \frac{|\psi_n(\lambda_j)|^2}{|\lambda_j - \theta_1| |\lambda_j - \theta_2|}$$

There is an exact balance between the contributions from both arcs.

Remark 3.6. By now it is clear that the structure of unitary Hessenberg matrices with positive subdiagonal elements (connected to the IAP) is very similar to the structure of Jacobi matrices (connected to the Lanczos process). We have para-orthogonal

polynomials instead of orthogonal polynomials, but both kind of polynomials are characterized by a minimization problem and for both there is a separation property for their zeros. Since these properties of orthogonal polynomials were among the main tools in the study of the convergence of the Lanczos process in [18], we can use similar ideas for the convergence of the IAP, as will be clear from the proofs of the theorems that we give in the next section.

4. Proofs of Theorems 2.2, 2.3, and 2.4. Here we give the proofs of our main Theorems 2.2, 2.3 and 2.4. We will also make essential use of properties of logarithmic potentials U^μ . We refer the reader to [22, 23] for background information on logarithmic potential theory.

In what follows we use χ_p to denote the normalized zero counting measure of a polynomial p . So if p has degree n then

$$\chi_p = \frac{1}{n} \sum_{p(\lambda)=0} \delta_\lambda$$

where the sum is over all zeros of p and the zeros are counted according to their multiplicity.

Note that in §3 we dropped the index N . Here it will re-appear and we will use the properties and results of §3 with no further comment.

4.1. Proof of Theorem 2.2. Theorem 2.2 was established for orthogonal polynomials whose zeros are on the real line by Rakhmanov [21]. Dragnev and Saff [11] used similar ideas to prove a more general theorem (including external fields), and weakened one of the conditions of Rakhmanov. Although these papers do not mention matrix iterations, we can nicely fit our setting in their results. The proof follows along arguments given in [11, 21]. We will indicate how we can modify them to the case of para-orthogonal polynomials, who have their zeros on the unit circle.

Proof of Theorem 2.2. Rakhmanov [21] showed that there exists a unique Borel probability measure μ_t that minimizes the logarithmic energy (2.9) among all Borel probability measures μ satisfying $0 \leq t\mu \leq \sigma$. He also showed that there exists a constant F_t such that (2.10) is satisfied, and that (2.8) and (2.10) characterize the pair (μ_t, F_t) . It remains to show that (2.6) and (2.7) hold.

The first step is to show that

$$(4.1) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|\psi_{n,N}(U_N)b_N\|^{1/n} \leq e^{-F_t}.$$

The proof of (4.1) follows the proof of Lemma 5.3 in [11]. For a given $\varepsilon > 0$, a monic polynomial q_N of degree n is constructed for every large enough N , so that

$$(4.2) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|q_N(U_N)b_N\|^{1/n} \leq e^{-F_t + \varepsilon}.$$

There is a set A so that every eigenvalue of U_N outside A is a zero of q_N , and the rest of the zeros of q_N are taken in such a way that $\chi_{q_N} \rightarrow \mu_t$. We need to modify this construction a little bit in order to guarantee that

$$(4.3) \quad q_N(0) = \psi_{n,N}(0) = \rho_{n,N}.$$

We can achieve (4.3) by moving one of the zeros in A to a different position on the unit circle. This will not affect the estimate (4.2). Having (4.2) and (4.3) we use Theorem 3.3 to conclude that

$$\limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|\psi_{n, N}(U_N)b_N\|^{1/n} \leq e^{-F_t + \varepsilon}.$$

Since ε can be chosen arbitrarily small (4.1) follows.

In the second step we show the following. Suppose we are given a sequence $(q_N)_N$ of monic polynomials such that q_N has degree n , the zeros of q_N are separated by the eigenvalues of U_N and the normalized zero counting measures χ_{q_N} have a weak*-limit μ . Then

$$(4.4) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|q_N(U_N)b_N\|^{1/n} \geq e^{-F^\mu},$$

where

$$(4.5) \quad F^\mu = \min_{z \in \text{supp}(\sigma - t\mu)} U^\mu(z).$$

Dragnev and Saff [11, Lemma 5.5] showed this for the case of the real line. The exact same proof works here.

In the third step we show that

$$(4.6) \quad F^\mu \leq F_t$$

for every Borel probability measure μ with $0 \leq t\mu \leq \sigma$ and equality in (4.6) holds if and only if $\mu = \mu_t$. Thus let μ be a Borel probability measure such that $0 \leq t\mu \leq \sigma$. Let $z \in \text{supp}(\sigma - t\mu)$. We know from (2.10) that $U^{\mu_t}(z) \leq F_t$ and from (4.5) that $F^\mu \leq U^\mu(z)$. Hence

$$(4.7) \quad U^{\sigma - t\mu}(z) - U^{\sigma - t\mu_t}(z) = t(U^{\mu_t}(z) - U^\mu(z)) \leq t(F_t - F^\mu).$$

On $\mathbb{C} \setminus \text{supp}(\sigma - t\mu)$ we have that $U^{\sigma - t\mu}$ is harmonic and $U^{\sigma - t\mu_t}$ superharmonic, so that $U^{\sigma - t\mu} - U^{\sigma - t\mu_t}$ is a subharmonic function there. Since $U^{\sigma - t\mu} - U^{\sigma - t\mu_t}$ is bounded at infinity (it has limit 0 at infinity), we can apply the maximum principle for subharmonic functions [22, Theorem 2.3.1], [23, Theorem 0.5.2] and it follows that (4.7) holds for every $z \in \mathbb{C}$. At infinity the left-hand side is 0, so that $F^\mu \leq F_t$.

If $F^\mu = F_t$, then we get $U^{\mu_t} - U^\mu \leq 0$ everywhere. Since at infinity these two functions are equal, and their difference is a harmonic function on $\mathbb{C} \setminus \mathbb{T}$ we can conclude that it is zero outside the unit disc. By continuity, it is also zero on the unit circle. Inside the unit disc it is harmonic, and applying the maximum principle again, we find that it is zero inside the unit disc. So $U^{\mu_t} = U^\mu$ everywhere, which means that $\mu_t = \mu$ [22, Corollary 3.7.5], [23, Corollary II.2.2].

Now collecting all the pieces finishes the proof. By Proposition 3.4, we know that the zeros of $\psi_{n, N}$ are separated by the eigenvalues of U_N . Let μ be a weak*-limit of a subsequence of the sequence of normalized zero counting measures $(\chi_{\psi_{n, N}})$. Then we find by (4.1) and (4.4) that

$$(4.8) \quad e^{-F^\mu} \leq \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|\psi_{n, N}(U_N)b_N\|^{1/n} \leq \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \|\psi_{n, N}(U_N)b_N\|^{1/n} \leq e^{-F_t},$$

hence $F^\mu \geq F_t$. From the separation property of the zeros of $\psi_{n,N}$ it also follows that $0 \leq t\mu \leq \sigma$. By (4.6) we must have $F^\mu = F_t$ so that $\mu = \mu_t$. Hence the inequalities in (4.8) are all equalities, which proves (2.7). We also see that μ_t is the only possible limit of a weak*-convergent subsequence of $(\chi_{\psi_{n,N}})$. Since the unit circle is compact, the set of Borel probability measures on \mathbb{T} is compact in the weak*-topology. Hence the full sequence $(\chi_{\psi_{n,N}})$ converges to μ_t , which gives (2.6).

This concludes the proof of Theorem 2.2. \square

4.2. Three lemmas. For the proof of Theorems 2.3 and 2.4 we need a number of lemmas. We will use the approach of Beckermann [5] who established these theorems for the Lanczos process. We will assume that the Conditions 2.1 hold.

The first lemma is borrowed from [6].

LEMMA 4.1 ([6]). *Let σ be a Borel probability measure on the unit circle and suppose $(\Lambda_N)_N$ is a sequence of sets, all contained in \mathbb{T} , such that*

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\lambda \in \Lambda_N} f(\lambda) = \int f(\lambda) d\sigma(\lambda)$$

for every continuous function f on \mathbb{T} .

Let $t \in (0, 1)$ and let μ be a Borel probability measure such that $t\mu \leq \sigma$. Let $n = n_N \leq \#\Lambda_N$ such that $n/N \rightarrow t$. Then there exists a sequence of sets $(Z_N)_N$ such that

- (a) $\#Z_N = n$,
- (b) $Z_N \subset \Lambda_N$, and
- (c) for all continuous functions f ,

$$\lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \frac{1}{n} \sum_{\lambda \in Z_N} f(\lambda) = \int f(\lambda) d\mu(\lambda).$$

Furthermore, if K is a closed set such that $\sigma(\partial K) = 0$ and $\sigma(K) = t\mu(K)$, then the sets Z_N can be chosen such that in addition to (a), (b), and (c), we also have for N large enough,

- (d) $\Lambda_N \cap K \subset Z_N$.

Proof. In [6, Lemma A.1] this lemma is proven for the case where the sets Λ_N are contained in the real line. The same proof works here. \square

The following lemma tells us that the eigenvalues inside $\Lambda(t, \sigma)$ are not exponentially close.

LEMMA 4.2 ([5]). *We have*

$$\lim_{N \rightarrow \infty} \min\{|\lambda_{k\pm 1, N} - \lambda_{k, N}|^{1/N} : k = 1, 2, \dots, N\} = 1$$

Proof. In [5, Lemma 2.4(b)] this lemma is proven for the case of points on the real line. The same proof works here. \square

The next lemma gives an estimate for $|\lambda_{k, N} - \theta_{\kappa-1, n, N}| |\lambda_{k, N} - \theta_{\kappa, n, N}|$, where $\lambda_{k, N}$ is on the closed arc between $\theta_{\kappa-1, n, N}$ and $\theta_{\kappa, n, N}$. Recall that the isometric Ritz values are numbered counterclockwise, and that $\theta_{0, n, N} := \theta_{n, n, N}$. We introduce the function

$$(4.9) \quad r_{\kappa, n, N}(z) = (z^{-1} - \bar{\theta}_{\kappa-1, n, N})(z - \theta_{\kappa, n, N}) \sqrt{\theta_{\kappa-1, n, N} / \theta_{\kappa, n, N}}, \quad \kappa = 1, \dots, n,$$

where we choose the branch of the square root belonging to the lower half plane. Thus, if $\theta_{\kappa-1,n,N} = e^{i\tau_1}$ and $\theta_{\kappa,n,N} = e^{i\tau_2}$ with $0 < \tau_2 - \tau_1 < 2\pi$, then

$$(4.10) \quad \sqrt{\theta_{\kappa-1,n,N}/\theta_{\kappa,n,N}} = e^{-i\frac{\tau_2-\tau_1}{2}}.$$

Observe that $|\lambda_{k,N} - \theta_{\kappa-1,n,N}||\lambda_{k,N} - \theta_{\kappa,n,N}| = |r_{\kappa,n,N}(\lambda_{k,N})|$.

LEMMA 4.3. *Let $r_{\kappa,n,N}(z)$ be defined as in (4.9)–(4.10). Then the following hold.*

(a) *The function $r_{\kappa,n,N}(z)$ is real and negative for z on the open arc from $\theta_{\kappa-1,n,N}$ to $\theta_{\kappa,n,N}$ and real and positive on the complementary open arc.*

(b) *Let $\lambda_{k,N}$ be on the closed arc from $\theta_{\kappa-1,n,N}$ to $\theta_{\kappa,n,N}$. Then for every polynomial q of degree at most $n-2$,*

$$(4.11) \quad w_{k,N}^2 |q(\lambda_{k,N})|^2 |r_{\kappa,n,N}(\lambda_{k,N})| \leq \sum_{j \neq k} w_{j,N}^2 |q(\lambda_{j,N})|^2 r_{\kappa,n,N}(\lambda_{j,N}).$$

(c) *Equality holds in (4.11) for the polynomial*

$$(4.12) \quad q(z) = \frac{\psi_{n,N}(z)}{(z - \theta_{\kappa-1,n,N})(z - \theta_{\kappa,n,N})},$$

where $\psi_{n,N}$ is the monic para-orthogonal polynomial.

Proof. Let $z \in \mathbb{T}$ and choose $\tau_1 = \arg \theta_{\kappa-1,n,N}$, $\tau_2 = \arg \theta_{\kappa,n,N}$ such that $0 < \tau_2 - \tau_1 < 2\pi$. Then we have for $z \in \mathbb{T}$,

$$\begin{aligned} r_{\kappa,n,N}(z) &= (z - e^{i\tau_2})(\bar{z} - e^{-i\tau_1})e^{-i\frac{\tau_2-\tau_1}{2}} \\ &= (e^{-i\frac{\tau_1+\tau_2}{2}}z - e^{i\frac{\tau_2-\tau_1}{2}})(e^{i\frac{\tau_1+\tau_2}{2}}\bar{z} - e^{i\frac{\tau_2-\tau_1}{2}})e^{-i\frac{\tau_2-\tau_1}{2}} \\ &= e^{-i\frac{\tau_2-\tau_1}{2}} - e^{-i\frac{\tau_1+\tau_2}{2}}z - e^{i\frac{\tau_1+\tau_2}{2}}\bar{z} + e^{i\frac{\tau_2-\tau_1}{2}} \\ &= -2\operatorname{Re}(e^{-i\frac{\tau_1+\tau_2}{2}}z) + 2\cos\frac{\tau_2-\tau_1}{2}. \end{aligned}$$

This shows that $r_{\kappa,n,N}(z)$ is real for $z \in \mathbb{T}$. Moreover, $r_{\kappa,n,N}$ is the composition of the mappings $z \mapsto e^{-i\frac{\tau_1+\tau_2}{2}}z$, $z \mapsto \operatorname{Re} z$, and $z \mapsto -2z + 2\cos\frac{\tau_2-\tau_1}{2}$. The effect of these mappings on the unit circle is plotted step by step in Figure 4.1. Following these mappings, we obtain the statements of part (a).

To prove part (b), we use the Gaussian quadrature formula (3.8). We know that there exist positive real numbers $\beta_{1,N}, \dots, \beta_{n,N}$ such that

$$(4.13) \quad \sum_{j=1}^N w_{j,N}^2 p(\lambda_{j,N}) = \sum_{j=1}^n \beta_{j,N} p(\theta_{j,n,N})$$

for every Laurent polynomial p of degree $n-1$. Now let q be a polynomial of degree at most $n-2$ and write

$$(4.14) \quad p(z) = r_{\kappa,n,N}(z)q(z)\bar{q}(z^{-1}),$$

where \bar{q} is the polynomial whose coefficients are the complex conjugates of the coefficients of q . This p is a Laurent polynomial of degree $n-1$, so we can apply (4.13) to p . Because of part (a), we know that $r_{\kappa,n,N}(\theta_{j,n,N}) \geq 0$ for all j . Since also $q(z)\bar{q}(z^{-1}) = |q(z)|^2 \geq 0$ for all $z \in \mathbb{T}$, we see that $p(\theta_{j,n,N}) \geq 0$ for all j . So the right-hand side of (4.13) is non-negative, which implies that

$$(4.15) \quad -w_{k,N}^2 p(\lambda_{k,N}) \leq \sum_{j \neq k} w_{j,N}^2 p(\lambda_{j,N}),$$

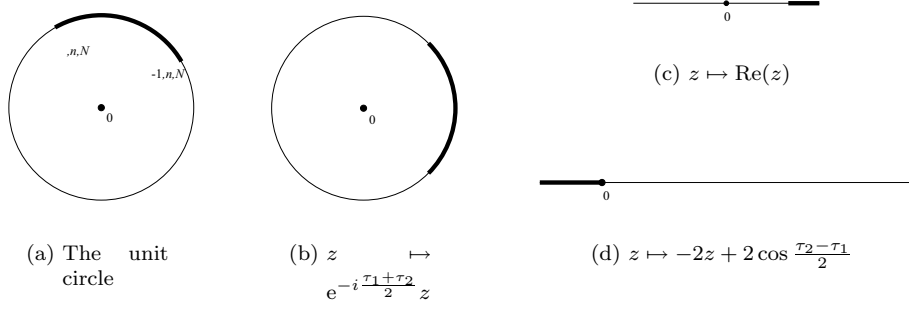


FIGURE 4.1. The image of the unit circle under the mapping $z \mapsto r_{\kappa, n, N}(z)$ step by step. Note that the arc between $\theta_{\kappa-1, n, N}$ and $\theta_{\kappa, n, N}$ is mapped to the negative real axis, and the complementary arc to the positive real axis

which gives

$$(4.16) \quad -w_{k, N}^2 r_{\kappa, n, N}(\lambda_{k, N}) |q_{\lambda_{k, N}}|^2 \leq \sum_{j \neq k} w_{j, N}^2 r_{\kappa, n, N}(\lambda_{j, N}) |q(\lambda_{j, N})|^2.$$

Now $r_{\kappa, n, N}(\lambda_{k, N}) < 0$ according to part (a) again, since $\lambda_{k, N}$ is on the arc from $\theta_{\kappa-1, n, N}$ to $\theta_{\kappa, n, N}$. Using this in (4.16) we obtain (4.11). This proves part (b).

Finally, if we use the polynomial q from (4.12) in the construction (4.14), then the right-hand side of (4.13) equals zero, since all terms vanish. This leads to equality in (4.11), so that part (c) follows. \square

For every polynomial q of degree at most $n - 2$ with $q(\lambda_{k, N}) \neq 0$, (4.11) can be rewritten as

$$(4.17) \quad |\lambda_{k, N} - \theta_{\kappa-1, n, N}| |\lambda_{k, N} - \theta_{\kappa, n, N}| \\ \leq \frac{\sum_{j \neq k} w_{j, N}^2 (\bar{\lambda}_{j, N} - \bar{\theta}_{\kappa-1, n, N}) (\lambda_{j, N} - \theta_{\kappa, n, N}) \sqrt{\theta_{\kappa-1, n, N} / \theta_{\kappa, n, N}} |q(\lambda_{j, N})|^2}{w_{k, N}^2 |q(\lambda_{k, N})|^2}.$$

From this we deduce

$$(4.18) \quad \min_j |\lambda_{k, N} - \theta_{j, n, N}| \leq (|\lambda_{k, N} - \theta_{\kappa-1, n, N}| |\lambda_{k, N} - \theta_{\kappa, n, N}|)^{1/2} \\ \leq \left(\frac{\sum_{j \neq k} w_{j, N}^2 |\bar{\lambda}_{j, N} - \bar{\theta}_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}| |q(\lambda_{j, N})|^2}{w_{k, N}^2 |q(\lambda_{k, N})|^2} \right)^{1/2} \\ \leq \left(\frac{\max_{j \neq k} |q(\lambda_{j, N})|}{|q(\lambda_{k, N})|} \right) \left(4 \frac{\sum_{j \neq k} w_{j, N}^2}{w_{k, N}^2} \right)^{1/2}.$$

4.3. Proof of Theorem 2.3. To prove Theorem 2.3, we use the estimate (4.18). We are going to find estimates for the numerator and denominator of the first factor in the right-hand side. To this end we will construct a suitable polynomial q .

Proof of Theorem 2.3. Let $(k_N)_N$ be a sequence of indices so that $\lim_{N \rightarrow \infty} \lambda_{k_N} = \lambda$. Since all eigenvalues and all isometric Ritz values are contained in the unit circle, there is nothing to prove if $U^{\mu_t}(\lambda) = F_t$.

So suppose $U^{\mu_t}(\lambda) < F_t$ and let $\varepsilon \in (0, -U^{\mu_t}(\lambda) + F_t)$. Define

$$K := \{z \in \mathbb{T} \mid -U^{\mu_t}(z) + F_t \geq \varepsilon\}.$$

Since U^σ is continuous, so is U^{μ_t} (see e.g. [11, Lemma 5.2]), so that K is closed and contains an η -neighborhood of λ (we take $\eta < 1$). Now $K \cap \text{supp}(\sigma - t\mu_t) = \emptyset$, so $\sigma(K) = t\mu(K)$. Without loss of generality we may suppose that $\sigma(\partial K) = 0$ (see also Remark 4.4 below). We can now obtain a sequence of sets $(Z_N)_N$ by Lemma 4.1 with $\mu = \mu_t$ and n replaced by $n - 1$.

By Condition 2.1.2) we can choose $\delta < \eta$ such that (2.2) holds for N sufficiently large and for all $k \leq N$. Note that by properties (b) and (d) of Lemma 4.1, and the definition of K , all eigenvalues $\lambda_{j,N}$ with $|\lambda_{j,N} - \lambda_{k_N,N}| < \delta$ are in Z_N , when N is large enough. We define

$$q_N(z) := \prod_{\lambda_{j,N} \in Z'_N} (z - \lambda_{j,N}),$$

where $Z'_N := Z_N \setminus \{\lambda_{k_N,N}\}$ (so q_N is a polynomial of degree $n-2$). Note that property (c) of Lemma 4.1 still holds when we replace the sets Z_N by Z'_N , i.e., the sequence of normalized zero counting measures of $(q_N)_N$ converges in weak*-sense to μ_t .

We factor q_N in two parts, one containing the zeros close to $\lambda_{k_N,N}$, and one containing the other zeros:

$$q_N^{(1)}(z) := \prod_{0 < |\lambda_{j,N} - \lambda_{k_N,N}| < \delta} (z - \lambda_{j,N}), \quad q_N^{(2)}(z) := \frac{q_N(z)}{q_N^{(1)}(z)}.$$

We also define the measures

$$\mu_N^{(1)} := \frac{1}{n-2} \sum_{q_N^{(1)}(\lambda)=0} \delta_\lambda, \quad \mu_N^{(2)} := \frac{1}{n-2} \sum_{q_N^{(2)}(\lambda)=0} \delta_\lambda.$$

Then $\chi_{q_N} = \mu_N^{(1)} + \mu_N^{(2)}$, so that

$$(4.19) \quad U^{\chi_{q_N}}(\lambda_{k_N,N}) = U^{\mu_N^{(1)}}(\lambda_{k_N,N}) + U^{\mu_N^{(2)}}(\lambda_{k_N,N}).$$

Because of Condition 2.1.2),

$$(4.20) \quad U^{\mu_N^{(1)}}(\lambda_{k_N,N}) < \varepsilon$$

for N large enough. Since $\lim_{\substack{n,N \rightarrow \infty \\ n/N \rightarrow t}} \mu_N^{(2)} = \mu_t|_{\mathbb{T} \setminus B(\lambda, \delta)}$, we get

$$(4.21) \quad \lim_{\substack{n,N \rightarrow \infty \\ n/N \rightarrow t}} U^{\mu_N^{(2)}}(\lambda_{k_N,N}) = U^{\mu_t|_{\mathbb{T} \setminus B(\lambda, \delta)}}(\lambda) = U^{\mu_t}(\lambda) - U^{\mu_t|_{B(\lambda, \delta)}}(\lambda) \leq U^{\mu_t}(\lambda),$$

where the last inequality holds since $\delta < 1$. Combining the two estimates (4.20) and (4.21) with equation (4.19) we get

$$(4.22) \quad \limsup_{\substack{n,N \rightarrow \infty \\ n/N \rightarrow t}} U^{\chi_{q_N}}(\lambda_{k_N,N}) \leq U^{\mu_t}(\lambda) + \varepsilon,$$

so that

$$(4.23) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \log |q_N(\lambda_{k_N, N})|^{1/n} = \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} -U^{X_{q_N}}(\lambda_{k_N, N}) \geq -U^{\mu_t}(\lambda) - \varepsilon.$$

Now we are going to estimate the absolute value of q_N on the rest of the spectrum of U_N . By construction $q_N(\lambda_{j, N}) = 0$ for $\lambda_{j, N} \in K \setminus \{\lambda_{k_N, N}\}$, so we have

$$\max_{j \neq k_N} |q_N(\lambda_{j, N})| = \max_{\lambda_{j, N} \notin K} |q_N(\lambda_{j, N})| \leq \sup_{z \in \mathbb{T} \setminus K} |q_N(z)|.$$

Since the zero distributions of q_N converge to μ_t , we can apply the principle of descent [23, Theorem I.6.8]. Then we get

$$(4.24) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \max_{j \neq k_N} \log |q_N(\lambda_{j, N})|^{1/n} \leq \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \sup_{z \in \mathbb{T} \setminus K} \frac{1}{n} \log |q_N(z)| \\ \leq \sup_{z \in \mathbb{T} \setminus K} (-U^{\mu_t}(z)) \leq -F_t + \varepsilon,$$

where the last inequality follows from the definition of K .

If we now choose $q = q_N$ and $k = k_N$ in (4.18), we get

$$(4.25) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \\ \leq \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left[\left(\frac{\max_{j \neq k_N} |q_N(\lambda_{j, N})|}{|q_N(\lambda_{k_N, N})|} \right)^{1/n} \left(4 \frac{\sum_{j \neq k_N} w_{j, N}^2}{w_{k_N, N}^2} \right)^{1/2n} \right],$$

The second factor in the lim sup in the right-hand side of (4.25) converges to 1, because of Condition 2.1.3), while the first factor is handled by (4.23) and (4.24). The result is that

$$\limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \leq \exp(U^{\mu_t}(\lambda) - F_t + 2\varepsilon).$$

Since this holds for all $\varepsilon > 0$, (2.11) is proven. \square

Remark 4.4. If the set K is Cantor-like and the measure σ singular, we might have that $\sigma(\partial K) > 0$. However, Since

$$\partial K \subseteq \{z \in \mathbb{T} \mid -U^{\mu_t}(z) + F_t = \varepsilon\}$$

we have that $\sigma(\partial K) > 0$ can only happen for a countable number of ε 's. So if $\sigma(\partial K) > 0$, we can choose a smaller ε so that $\sigma(\partial K) = 0$ and continue with the proof of Theorem 2.3.

4.4. Proof of Theorem 2.4. We finally give the proof of Theorem 2.4. As noted before, the proof is based on the proof of [5, Theorem 2.1], but we have streamlined some of the arguments.

Proof of Theorem 2.4. Since $\lambda \in \Lambda(t, \sigma)$ we have $F_t - U^{\mu_t}(\lambda) > 0$. By continuity there is a δ -neighborhood Δ_δ of λ and an $\varepsilon > 0$ such that $F_t - U^{\mu_t}(z) > \varepsilon$ for

$z \in \Delta_\delta$. Because of Theorem 2.3 we then know that each eigenvalue $\lambda_{k,N} \in \Delta_\delta$ has an isometric Ritz value close to it if N is large. More precisely we can assure that

$$\min_j |\lambda_{k,N} - \theta_{j,n,N}| \leq e^{-n\varepsilon}$$

for all $\lambda_{k,N} \in \Delta_\delta$ if N is large enough.

Now we study the relative positions of eigenvalues and isometric Ritz values. Using (i) the separation property (see Proposition 3.4), (ii) the fact that eigenvalues are exponentially well approximated (see Theorem 2.3) and (iii) the fact that the distance between eigenvalues is *not* exponentially small (see Lemma 4.2), we can make a complete classification of these relative positions, for N large enough. The different cases were plotted in Figure 2.1. The exceptions are covered below and illustrated in Figure 4.2.

From the separation property we conclude that close to an eigenvalue there can be at most two isometric Ritz values (one on either side of it on the unit circle). However it is easily seen that at most one eigenvalue $\lambda_{\ell_1,N} \in \Delta_\delta$ can be approximated by two isometric Ritz values, again because the isometric Ritz values are separated by the eigenvalues and because each eigenvalue is well approximated by at least one isometric Ritz value. In this case we define the exceptional index as $k_N^*(\lambda) := \ell_1$.

Another possibility is that an eigenvalue $\lambda_{\ell_2,N}$ and an isometric Ritz value coincide. In similar fashion one can see that this happens at most once, and that this case is not compatible with the previous one. Then we define the exceptional index as $k_N^*(\lambda) := \ell_2$.

It is also possible that there are two consecutive eigenvalues $\lambda_{\ell_3-1,N}$ and $\lambda_{\ell_3,N}$ in Δ_δ that do not have an isometric Ritz value on the arc between them. Again, it is easily seen that this can happen only once in Δ_δ and that this excludes the previous two possibilities. In this case the exceptional index is either $\ell_3 - 1$ or ℓ_3 , depending on the proximity of the nearest isometric Ritz value. More precisely, let $\theta_{\kappa,n,N}$ be the first isometric Ritz value after $\lambda_{\ell_3,N}$. We define the exceptional index as

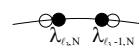
$$(4.26) \quad k_N^*(\lambda) := \begin{cases} \ell_3, & \text{if } |\theta_{\kappa,n,N} - \lambda_{\ell_3,N}| \leq |\theta_{\kappa-1,n,N} - \lambda_{\ell_3-1,N}|, \\ \ell_3 - 1, & \text{otherwise.} \end{cases}$$



(a) case 2: The exceptional eigenvalue is approximated by two isometric Ritz values.



(b) case 5: The exceptional eigenvalue is coincides with an isometric Ritz value.



(c) case 6: The exceptional eigenvalue is one of the two eigenvalues indicated, see (4.26).

FIGURE 4.2. The definition of the exceptional indices in the different cases (see Figure 2.1). In cases 1, 3 and 4 no exceptions need to be made.

Now if $\lambda_{k_N,N} \rightarrow \lambda$, then for N large enough $\lambda_{k_N,N} \in \Delta_\delta$. Furthermore, if $k_N \neq k_N^*(\lambda)$, there is exactly one isometric Ritz value $\theta_{j,n,N}$ close to $\lambda_{k_N,N}$ (case 2 is the only exception to this). All other isometric Ritz values are at a distance whose

n^{th} root limit is 1. It then follows that (4.18) can be sharpened to

$$\min_j |\lambda_{k_N, N} - \theta_{j, n, N}| \leq c_{n, N} \left(\frac{\max_{j \neq k} |q(\lambda_{j, N})|}{|q(\lambda_{k, N})|} \right)^2,$$

with constants $c_{n, N}$ such that $\lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} c_{n, N}^{1/n} = 1$. Examining the proof of Theorem 2.3, we see that this leads to

$$(4.27) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \leq \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right).$$

Next we prove the lower bound for $\min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n}$ when $k_N \neq k_N^*(\lambda)$. Choose κ such that $\lambda_{k_N, N}$ is on the arc from $\theta_{\kappa-1, n, N}$ to $\theta_{\kappa, n, N}$. From Remark 3.5 it follows that

$$\begin{aligned} \sum_{\lambda_{j, N} \in [\theta_{\kappa-1, n, N}, \theta_{\kappa, n, N}]} w_{j, N}^2 \frac{|\psi_{n, N}(\lambda_{j, N})|^2}{|\lambda_{j, N} - \theta_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}|} \\ = \sum_{\lambda_{j, N} \notin [\theta_{\kappa-1, n, N}, \theta_{\kappa, n, N}]} w_{j, N}^2 \frac{|\psi_{n, N}(\lambda_{j, N})|^2}{|\lambda_{j, N} - \theta_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}|}, \end{aligned}$$

where $[\theta_{\kappa-1, n, N}, \theta_{\kappa, n, N}]$ denotes the circular arc going from $\theta_{\kappa-1, n, N}$ to $\theta_{\kappa, n, N}$. Thus

$$\begin{aligned} \sum_{\lambda_{j, N} \in [\theta_{\kappa-1, n, N}, \theta_{\kappa, n, N}]} w_{j, N}^2 \frac{|\psi_{n, N}(\lambda_{j, N})|^2}{|\lambda_{j, N} - \theta_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}|} \\ = \frac{1}{2} \sum_{j=1}^N w_{j, N}^2 \frac{|\psi_{n, N}(\lambda_{j, N})|^2}{|\lambda_{j, N} - \theta_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}|} \\ \geq \frac{1}{8} \sum_{j=1}^N w_{j, N}^2 |\psi_{n, N}(\lambda_{j, N})|^2 = \frac{1}{8} \|\psi_{n, N}(U_N) b_N\|^2. \end{aligned}$$

Because of the limit (2.7) in Theorem 2.2, it then follows that

$$(4.28) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(\sum_{\lambda_{j, N} \in [\theta_{\kappa-1, n, N}, \theta_{\kappa, n, N}]} w_{j, N}^2 \frac{|\psi_{n, N}(\lambda_{j, N})|^2}{|\lambda_{j, N} - \theta_{\kappa-1, n, N}| |\lambda_{j, N} - \theta_{\kappa, n, N}|} \right)^{1/n} \geq \exp(-2F_t).$$

The sum in the left-hand side has at most two terms, one of them being for $\lambda_{k_N, N}$.

If there is only one term in the sum in the left-hand side of (4.28) (cases 1, 2, 3 and 4) or if one of the terms is 0 (case 5), then (4.28) says

$$(4.29) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(w_{k_N, N}^2 \frac{|\psi_{n, N}(\lambda_{k_N, N})|^2}{|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|} \right)^{1/n} \geq \exp(-2F_t).$$

Note that $\frac{\psi_{n, N}(z)}{(z - \theta_{\kappa-1, n, N})(z - \theta_{\kappa, n, N})}$ is a monic polynomial of degree $n - 2$ with roots $\theta_{j, n, N}$, $j \neq \kappa - 1, \kappa$. From (2.6) it follows that μ_t is the weak*-limit of the normalized

zero counting measures of these polynomials, and from this it follows that, by the principle of descent [23, Theorem I.6.8]),

$$(4.30) \quad \limsup_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(\frac{|\psi_{n, N}(\lambda_{k_N, N})|}{|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|} \right)^{1/n} \leq \exp(-U^{\mu_t}(\lambda)).$$

Using $\lim_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} w_{k_N, N}^{1/n} = 1$, we obtain from (4.29) that

$$\begin{aligned} & \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} (|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|)^{1/n} \\ & \geq \exp(-2F_t) \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(\frac{|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|}{|\psi_{n, N}(\lambda_{k_N, N})|} \right)^{2/n}, \end{aligned}$$

and together with (4.30) this gives us

$$(4.31) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} (|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|)^{1/n} \geq \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right).$$

Now we can conclude

$$\begin{aligned} (4.32) \quad & \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k_N, N} - \theta_{j, n, N}|^{1/n} \\ & \geq \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(\frac{|\lambda_{k_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k_N, N} - \theta_{\kappa, n, N}|}{2} \right)^{1/n} \\ & \geq \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right). \end{aligned}$$

The other possibility is that there are two terms in the sum in the left-hand side of (4.28). Then we are in case 6. Let k'_N be the index j giving the largest term in the sum. Then

$$\liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \left(w_{k'_N, N}^2 \frac{|\psi_{n, N}(\lambda_{k'_N, N})|^2}{|\lambda_{k'_N, N} - \theta_{\kappa-1, n, N}| |\lambda_{k'_N, N} - \theta_{\kappa, n, N}|} \right)^{1/n} \geq \exp(-2F_t)$$

and from this it follows as before that

$$(4.33) \quad \liminf_{\substack{n, N \rightarrow \infty \\ n/N \rightarrow t}} \min_j |\lambda_{k'_N, N} - \theta_{j, n, N}|^{1/n} \geq \exp\left(2(U^{\mu_t}(\lambda) - F_t)\right).$$

Since $\min_j |\lambda_{k_N, N} - \theta_{j, n, N}| \geq \min_j |\lambda_{k'_N, N} - \theta_{j, n, N}|$ (by the definition of k'_N in (4.26)), we also obtain (4.32) in this case.

So we have (4.32) in both cases. Together with (4.27) this proves (2.12). \square

5. Numerical experiments. For the numerical experiments, we take a large unitary matrix U of size $N \times N$, and we execute the IAP for every $n \leq N$ (so we let $t = n/N$ vary from 0 to 1).

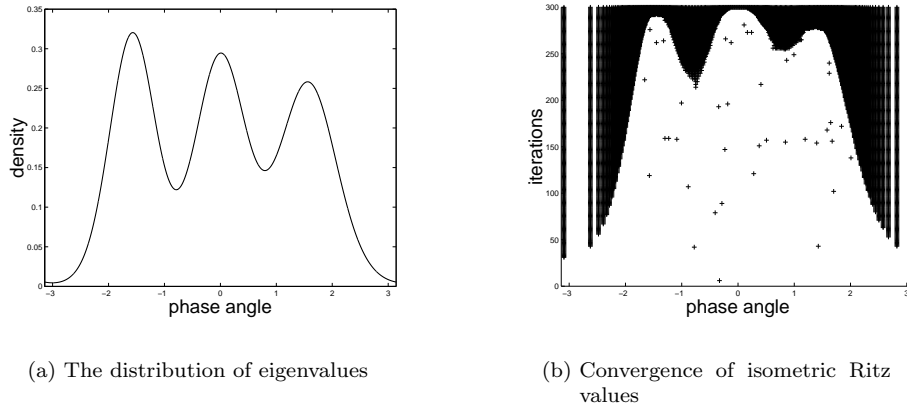


FIGURE 5.1. Convergence result for the IAP applied on a 300×300 matrix U with eigenvalues distributed as in (a). In (b) the iteration step (dimension of the modified Hessenberg submatrix) is plotted on the Y-axis, and the phase angle of the eigenvalues on the X-axis. If an isometric Ritz value is closer to an eigenvalue than 10^{-5} , a “+” is plotted.

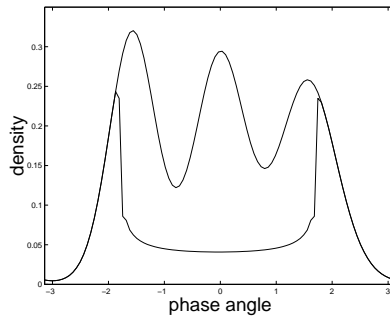


FIGURE 5.2. A simulation of $t\mu_t$ when σ is as in Figure 5.1(a) for $t = 0.4$.

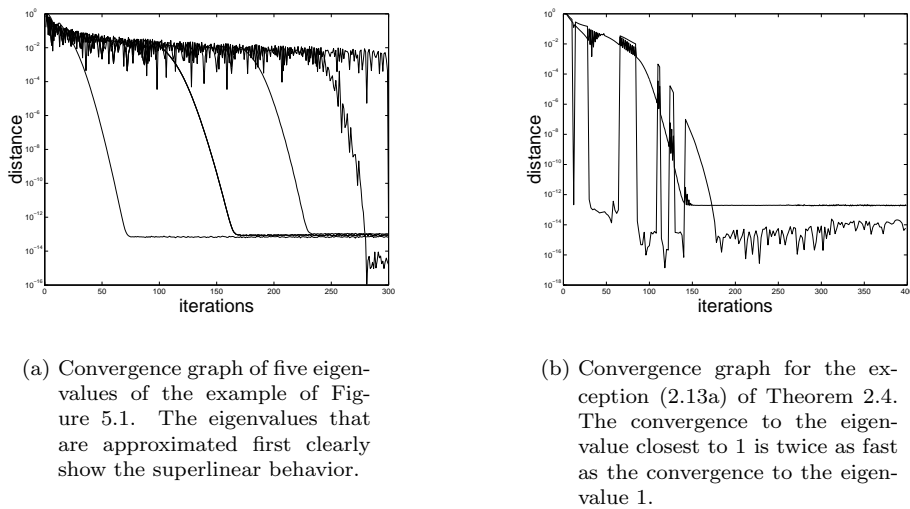


FIGURE 5.3. Convergence graphs of individual eigenvalues.

Our theoretical results are independent of the choice of the parameters $\rho_{n,N}$, but for the experiments we have to make a choice. We choose $\rho_{n,N} = \gamma_{n,N}/|\gamma_{n,N}|$, as this choice assures the modified submatrix stays as close as possible to the original submatrix [16, Lemma 2.1].

The experiments were done on matrices U whose eigenvalues are distributed according to a combination of von Mises distributions. A von Mises distribution is a continuous distribution on \mathbb{T} with density

$$P(e^{i\theta}) = \frac{1}{2\pi I_0(\alpha)} e^{\alpha \cos(\theta - \theta_0)},$$

where I_0 is the modified Bessel function of the first kind and order 0. We have that θ_0 is the *mean direction* and α is the *concentration parameter*. Von Mises distributions appear in directional statistics [20].

For the experiments we used MATLAB¹. Codes for Unitary Hessenberg QR were kindly provided to us by William B. Gragg and Michael Stewart. Random numbers from the von Mises distributions were generated using the R environment². We generated a very large sample of size mN , we ordered it, and then selected every m th point from it. A typical value of m we used was $m = 4000$. The points were then used as the eigenvalues of an $N \times N$ unitary matrix to which we applied the IAP. We followed this procedure in order to obtain eigenvalues that follow the limiting distribution pretty well. For the matrix sizes we used (namely $N = 300$), a fully random sample does not follow the limiting distribution very well, and our asymptotical results do not apply.

5.1. Distribution of isometric Ritz values. To improve the understanding of the experiments, we recall the minimizing property of μ_t , see Theorem 2.2. If we minimize $I(\mu)$ among *all* Borel probability measures supported on \mathbb{T} then the solution is the normalized Lebesgue measure on \mathbb{T} [23, p. 25], which we denote here by λ . Now if t is so small that $t\lambda < \sigma$, then μ_t is equal to λ , because of their respective minimizing properties. So then everywhere $t\mu_t < \sigma$, so that no convergence can be expected (see the discussion after Theorem 2.3). If t grows, $t\lambda$ will also increase, until for a certain critical t_{cr} it hits σ at the point (or points) where the density of σ is minimal. For slightly larger $t > t_{cr}$, eigenvalues will be found in a neighborhood of that point (or those points), since eigenvalues are found where $t\mu_t = \sigma$. If we let t increase further, the region of good convergence also increases.

Continuing this line of thought, one might think that convergence will be slowest in the region where the eigenvalue density has its maximum. However, this is not necessarily true (although in many cases it is), since μ_t might be very different from the normalized Lebesgue measure when t is not small.

In Figure 5.1 we present an example. The eigenvalues of the 300×300 matrix U are distributed according to a combination of three von Mises distributions, with respective parameter pairs

$$(\theta_0, \alpha) = (-\pi/2, 6), (0, 5) \text{ and } (\pi/2, 4).$$

The density is shown in part (a) of Figure 5.1. The distribution has three local maxima near the values $-i$, 0 , and i , that is, the points with angles $-\pi/2$, 0 , and $\pi/2$. Part (b) shows the convergence plot for the IAP. A $+$ is plotted for every isometric Ritz value whose distance to its nearest eigenvalue is less than 10^{-5} .

¹MATLAB is a registered trademark of The Mathworks, inc.

²The R project for statistical computing, <http://www.r-project.org>.

It can be seen that the shape of the convergence plot resembles the eigenvalue density. For the regions of low eigenvalue density, this follows from the preceding discussion. If we look at higher values of t (i.e., more iterations), then we see a difference between the two plots. The eigenvalue density has a maximum near $-\frac{\pi}{2}$, but the eigenvalues near that maximum are approximated earlier than eigenvalues near 0, where the peak is lower. To explain this phenomenon, we plotted a simulation of $t\mu_t$ for $t = 0.4$ in Figure 5.2. In the region where the constraint σ is active, μ_t does not look like λ at all, which is rather obvious (it is prohibited to do so by the constraint σ). Figure 5.2 shows that eigenvalues in the peaks around $-\frac{\pi}{2}$ and $\pi/2$, are indeed found earlier than eigenvalues in the peak around 0.

5.2. Convergence speed. Now we will check the assertions of Theorem 2.4. If we assume the right-hand side of (2.12) is constant as a function of t , we expect linear convergence. In fact, that right-hand side is slightly decreasing, so we should be able to observe a superlinear convergence (this superlinearity is of the same nature as the one discussed in [6, 8]). In Figure 5.3(a) the convergence graphs are plotted and indeed the (super)linearity appears.

We also tried to generate the exception (2.13a) of Theorem 2.4, see Figure 5.3(b). We created a real orthogonal matrix, with 1 as an eigenvalue. Since for a real matrix eigenvalues appear in complex conjugate pairs, in the even steps there are two options. Either two isometric Ritz values are close to 1, or one is equal to 1 and another equal to -1 . (We took $\rho_{n,N} = 1$.) This explains the erratic behavior: if only one isometric Ritz value approximates 1, the distance is much smaller than if two of them are close.

We only plotted the even steps, because in the odd steps one isometric Ritz value needs to be 1, again because of orthogonality. We compare this even-step convergence with the convergence of the eigenvalue closest to 1. If the erratic behavior is discarded, the difference between the convergence rates is approximately a factor 2: the convergence starts after some 90 steps and while the eigenvalue at 1 takes some hundred more steps to be well approximated, the next one only needs fifty. The superlinearity also appears clearly in both convergence graphs.

Acknowledgements. We thank Bernhard Beckermann for allowing us to use the material from [5]. We thank William B. Gragg and Michael Stewart for providing us with codes for UHQR. We thank Bernhard Beckermann and Walter Van Assche for useful discussions.

REFERENCES

- [1] G.S. AMMAR AND C. HE, *On an inverse eigenvalue problem for unitary Hessenberg matrices*, Linear. Algebra Appl., 218 (1995), pp. 263–271.
- [2] J. BAIK, T. KRIECHERBAUER, K.T.-R. MCLAUGHLIN, AND P.D. MILLER, *Uniform asymptotics for polynomials orthogonal with respect to a general class of discrete weights and universality results for associated ensembles: Announcement of results*, Int. Math. Res. Not., 2003 (2003), pp. 821–858.
- [3] ———, *Uniform asymptotics for polynomials orthogonal with respect to a general class of discrete weights and universality results for associated ensembles*. preprint math.CA/0310278.
- [4] C. BEATTIE, M. EMBREE, AND J. ROSSI, *Convergence of restarted Krylov subspaces to invariant subspaces*. Oxford University Computing Laboratory Numerical Analysis Report 01/21, 2001.
- [5] B. BECKERMANN, *A note on the convergence of Ritz values for sequences of matrices*. Publication ANO 408, Université de Lille, 2000.
- [6] B. BECKERMANN AND A.B.J. KUIJLAARS, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal., 39 (2001), pp. 300–329.

- [7] ———, *On the sharpness of an asymptotic error estimate for conjugate gradients*, BIT, 41 (2001), pp. 856–867.
- [8] ———, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19.
- [9] A. BUNSE-GERSTNER AND H. FASSBENDER, *Error bounds in the isometric Arnoldi process*, J. Comp. Appl. Math., 86 (1997), pp. 53–72.
- [10] A. BUNSE-GERSTNER AND C. HE, *On a Sturm sequence of polynomials for unitary Hessenberg matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1043–1055.
- [11] P.D. DRAGNEV AND E.B. SAFF, *Constrained energy problems with applications to orthogonal polynomials of a discrete variable*, J. Anal. Math., 72 (1997), pp. 223–259.
- [12] T.A. DRISCOLL, K.-C. TOH, AND L.N. TREFETHEN, *From potential theory to matrix iterations in six steps*, SIAM Review, 40 (1998), pp. 547–578.
- [13] H. FASSBENDER, *Inverse unitary eigenproblems and related orthogonal functions*, Numer. Math., 77 (1997), pp. 323–345.
- [14] W.B. GRAGG, *The QR algorithm for unitary Hessenberg matrices*, J. Comp. Appl. Math., 16 (1986), pp. 1–8.
- [15] ———, *Positive definite Toeplitz matrices, the Arnoldi method for isometric operators, and Gaussian quadrature on the unit circle*, J. Comp. Appl. Math., 46 (1993), pp. 183–198.
- [16] C. JAGELS AND L. REICHEL, *The isometric Arnoldi process and an application to the iterative solution of large linear systems*, in Iterative Methods in Linear Algebra, R. Beauwens and P. de Groen, eds., North Holland, Amsterdam, 1992, pp. 361–369.
- [17] W.B. JONES, O. NJÅSTAD, AND W.J. THRON, *Moment theory, orthogonal polynomials, quadrature, and continued fractions associated with the unit circle*, Bull. London Math. Soc., 21 (1989), pp. 113–152.
- [18] A.B.J. KUIJLAARS, *Which eigenvalues are found by the Lanczos method?*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 306–321.
- [19] A.B.J. KUIJLAARS AND E.A. RAKHMANOV, *Zero distributions for discrete orthogonal polynomials*, J. Comp. Appl. Math., 99 (1998), pp. 255–274.
- [20] K.V. MARDIA AND P.E. JUPP, *Directional Statistics*, Wiley, Chichester, 2000.
- [21] E.A. RAKHMANOV, *Equilibrium measure and the distribution of zeros of the extremal polynomials of a discrete variable*, Sb. Math., 187 (1996), pp. 1213–1228.
- [22] T. RANSFORD, *Potential Theory in the Complex Plane*, Cambridge University Press, Cambridge, 1995.
- [23] E.B. SAFF AND V. TOTIK, *Logarithmic Potentials with External Fields*, Springer, Berlin, 1997.
- [24] J. SHEN, G. STRANG, AND A.J. WATHEN, *The potential theory of several intervals and its applications*, Appl. Math. Optim., 44 (2001) 67–85.
- [25] M. STEWART, *Stability properties of several variants of the unitary Hessenberg QR algorithm*, in Structured Matrices in Mathematics, Computer Science and Engineering, II, V. Olshevsky, ed., vol. 281 of Contemporary Mathematics, AMS, 2001.
- [26] L.N. TREFETHEN AND D. BAU III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.