Evaluation of the eigenvectors of symmetric tridiagonal matrices is one of the most basic tasks in numerical linear algebra. It is a widely known fact that, in the case of well separated eigenvalues, the eigenvectors can be evaluated with high relative accuracy. Nevertheless, in general, each coordinate of the eigenvector is evaluated with only high *absolute* accuracy. In particular, those coordinates whose magnitude is below the machine precision are not expected to be evaluated to any correct digit at all.
In this paper, we address the problem of evaluating small (e.g. 10^{-50}) coordinates of the eigenvectors of certain symmetric tridiagonal matrices with high *relative* accuracy. We propose a numerical algorithm to solve this problem, and carry out error analysis. Also, we discuss some applications in which this task is necessary. Our results are illustrated via several numerical experiments.

Evaluation of small elements of the eigenvectors of certain symmetric tridiagonal matrices with high relative accuracy

Andrei Osipov
Research Report YALEU/DCS/TR-1460
Yale University
December 13, 2012

Approved for public release: distribution is unlimited.
Keywords: *symmetric tridiagonal matrices, eigenvectors, small elements, high accuracy*

Contents

1 Introduction

The evaluation of eigenvectors of symmetric tridiagonal matrices is one of the most basic tasks in numerical linear algebra (see, for example, such classical texts as [3], [4], [5], [6], [8], [9], [16], [19], [20]). Several algorithms to perform this task have been developed; these include Power and Inverse Power methods, Jacobi Rotations, QR and QL algorithms, to mention just a few. Many of these algorithms have become standard and widely known tools.

In the case when the eigenvalues of the matrix in question are well separated, most of these algorithms will evaluate the corresponding eigenvectors to a high *relative* accuracy. More specifically, suppose that $n > 0$ is an integer, that $v \in \mathbb{R}^n$ is the vector to be evaluated, and $\hat{v} \in \mathbb{R}^n$ is its numerical approximation, produced by one of the standard algorithms. Then,

$$\frac{\|v - \hat{v}\|}{\|v\|} \approx \varepsilon, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, and ε is the machine precision (e.g. $\varepsilon \approx 10^{-16}$ for double precision calculations).

However, a closer look at (1) reveals that it only guarantees that the *coordinates* of v be evaluated to high *absolute* accuracy. This is due to the following trivial observation. Suppose that we add $\varepsilon \cdot \|v\|$ to the first coordinate $\hat{v}_1$ of $\hat{v}$. Then, the perturbed $\hat{v}$ will not violate (1). On the other hand, the relative accuracy of $\hat{v}_1$ can be as large as

$$\frac{|v_1 + \varepsilon \cdot \|v\| - v_1|}{|v_1|} = \varepsilon \cdot \frac{\|v\|}{|v_1|}. \tag{2}$$

In particular, if, say, $\|v\| = 1$ and $|v_1| < \varepsilon$, then $\hat{v}_1$ is not guaranteed to approximate v_1 to any correct digit at all!

Sometimes the poor relative accuracy of "small" coordinates is of no concern; for example, this is usually the case when v is only used to project other vectors onto it. Nevertheless, in several prominent problems, small coordinates of the eigenvector are required to be evaluated to high relative accuracy. These problems include the evaluation of Bessel functions (see Sections 3.2, 3.4.3, 6.1), and the evaluation of some quantities associated with prolate spheroidal wave functions (see Sections 3.3, 6.2, and also [14], [15]), among others.

In this paper, we propose an algorithm for the evaluation of the coordinates of eigenvectors of certain symmetric tridiagonal matrices, to high relative accuracy. As the running example, we consider the matrices whose non-zero off-diagonal elements are constant (and equal to one), and the diagonal elements constitute a monotonically increasing sequence (see Definition 1). The connection of such matrices to Bessel functions and prolate spheroidal wave functions is discussed in Sections 3.4.3, 6.2, respectively. Also, we carry out detailed

error analysis of our algorithm. While our algorithm can be viewed as a modification of already existing (and well known) algorithms, such error analysis, perhaps surprisingly, appears to be new. In addition, we conduct several numerical experiments, to both illustrate the performance of our algorithm, and to compare it to some classical algorithms.

The following is the principal analytical result of this paper.

Theorem 1. *Suppose that $a \geq 1$ and $c \geq 1$ are real numbers, and that $n > 0$ is an integer. Suppose also that $A_1, \ldots, A_n$ are defined via the formula*

$$A_j = 2 + \left(\frac{j}{c}\right)^a, \tag{3}$$

for every $j = 1, \ldots, n$, and that the n by n symmetric tridiagonal matrix A is defined via the formula

$$A = \begin{pmatrix} A_1 & 1 & & & & \\ 1 & A_2 & 1 & & & \\ & 1 & A_3 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & A_{n-1} & 1 \\ & & & & 1 & A_n \end{pmatrix}. \tag{4}$$

Suppose furthermore that λ is an eigenvalue of A, and that

$$\lambda > 2 + A_k, \tag{5}$$

for some integer $1 \leq k \leq n$. Suppose, in addition, that $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a $\lambda-$eigenvector of A. If the entries of A are defined to the relative precision $\varepsilon > 0$, then the first k coordinates $X_1, \ldots, X_k$ of X are defined to the relative precision $R(c, a)$, where

$$R(c, a) \leq \varepsilon \cdot O\left(c^{4a/(a+2)}\right), \quad c \to \infty. \tag{6}$$

Remark 1. *We observe that, according to* (6)*, the relative precision of $X_1, \ldots, X_k$ does not depend on their order of magnitude. In other words, even if, say, X_1 is significantly smaller than ε, it is still defined to fairly high relative precision.*

Remark 2. *The definition of the entries of the matrix A in Theorem 1 is motivated by particular applications (see Section 6). On the other hand, Theorem 1 and Remark 1 generalize to a much wider class of matrices; these include, for example, perturbations of A, defined via* (4)*, as well as matrices, whose diagonal entries are of a more general form than* (3)*. While such generalizations are straightforward (and are based, in part, on the results of Section 4.2), the analysis is somewhat long, and will be published at a later date.*

The proof of Theorem 1 is constructive and somewhat technical (see Section 4, in particular, Section 4.3). The resulting numerical algorithm for the evaluation of the eigenvector X is described in Sections 5.2, 5.3, 5.4, and Section 5.5 contains its error analysis.

The paper is organized as follows. Section 2 contains a brief informal overview of the principle algorithm of this paper. In Section 3, we summarize a number of well known mathematical and numerical facts to be used in the rest of this paper. In Section 4, we introduce the necessary analytical apparatus and carry out the analysis. In Section 5, we introduce the principal numerical algorithm of this paper, and perform its error analysis; we also describe several other related algorithms. In Section 6, we discuss some applications of our algorithm to other computational problems. In Section 7, we illustrate the performance of our algorithm via several numerical examples, and compare it to some related classical algorithms.

2 Overview

In this section, we overview the intuition behind the principal algorithm of this paper and the corresponding error analysis.

Suppose that A is a symmetric tridiagonal matrix, whose non-zero off-diagonal elements are constant (and equal to one), and the diagonal elements constitute a monotonically increasing sequence $A_1 < A_2 < \ldots$ (see Definition 1). Then, the coordinates of any eigenvector of A satisfy a certain three-term linear recurrence relation with non-constant coefficients, that depend on the corresponding eigenvalue λ (see Theorem 7). More specifically,

$$x_{k+1} - (\lambda - A_k) \cdot x_k + x_{k-1} = 0, \tag{7}$$

for every $k = 2, \ldots, n-1$, where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is the eigenvector corresponding to the eigenvalue λ.

It turns out that the qualitative properties of the recurrence relation (7) depend on whether $\lambda - A_k$ is greater than 2, less than -2, or between -2 and 2. Both our algorithm and the subsequent error analysis are based on the following three fairly simple observations.

Observation 1 ("growth"). Suppose that $B > 2$ and x_1, x_2, x_3 are real numbers, and that

$$x_3 - B \cdot x_2 + x_1 = 0. \tag{8}$$

If $0 < x_1 < x_2$, then the evaluation of x_3 from x_1, x_2 via (8) is stable (accurate); moreover, $x_3 > x_2$. On the other hand, the evaluation of x_1 from x_2, x_3 via (8) is unstable (inaccurate), since, loosely speaking, we attempt to evaluate a "small" number as a difference of two bigger positive numbers.

Remark 3. *This intuition is generalized and developed in Theorem 11, Corollary 3, Theorem 12, and serves a basis for the step of the principal algorithm, described in Section 5.3.1 (see also Observation 1 in Section 5.5).*

Observation 2 ("decay"). Suppose now that $B < -2$ and y_1, y_2, y_3 are real numbers, and that

$$y_3 - B \cdot y_2 + y_1 = 0. \tag{9}$$

If y_3 and y_2 have opposite signs, and $|y_2| > |y_3|$, then the evaluation of y_1 from y_2, y_3 via (9) is stable (accurate); moreover, y_1 and y_2 have opposite signs, and $|y_1| > |y_2|$. On the other

hand, the evaluation of y_3 from y_1, y_2 (9) is unstable (inaccurate), since, loosely speaking, we attempt to obtain a small number as a sum of two numbers of opposite signs of larger magnitude.

Remark 4. *This intuition is generalized and developed in Theorem 13, and serves a basis for the step of the principal algorithm, described in Section 5.3.4 (see also Observation 3 in Section 5.5).*

Observation 3 ("oscillations"). Consider the following example. Suppose that $a > 0$ is a real number, and that the real numbers $x_0, x_1, x_2, \ldots$, are defined via the formula

$$x_k = \cos(k \cdot a), \tag{10}$$

for every $k = 0, 1, \ldots,$. We recall the trigonometric identity

$$\cos\left((k+1) \cdot a\right) + \cos\left((k-1) \cdot a\right) = 2 \cdot \cos(a) \cdot \cos\left(k \cdot a\right), \tag{11}$$

that holds for all real a, k, and substitute (10) into (11) to conclude that

$$x_{k+1} = 2 \cdot \cos(a) \cdot x_k - x_{k-1}, \tag{12}$$

for every $k = 1, 2, \ldots$. Obviously, the sequence $x_0, x_1, \ldots$ contains elements of order one as well as relatively small elements - the latter ones are obtained as a difference of two larger elements, potentially resulting in a loss of accuracy, and complicating the error analysis of the recurrence relation (12).

Consider, however, the complex numbers $z_0, z_1, z_2, \ldots$, defined via the formula

$$z_k = e^{ika}, \tag{13}$$

for every $k = 0, 1, \ldots$. We combine (10) and (13) to conclude that

$$x_k = \Re\left(z_k\right), \tag{14}$$

for every $k = 0, 1, 2, \ldots$. Moreover,

$$z_{k+1} = e^{ia} \cdot z_k, \tag{15}$$

for every $k = 0, 1, \ldots$. The stability of the recurrence relation (15) is obvious (if its elements are stored in polar coordinates, each subsequent element is obtained from its predecessor via rotating the latter by the angle a; also, the absolute values of $z_0, z_1, \ldots$ are all the same, in sharp contrast to those of $x_0, x_1, \ldots$). Moreover, even at first glance, (15) seems to be easier to analyze than (12).

Remark 5. *This intuition is generalized and developed in Theorems 14, 15, 16, 17, Corollaries 4, 5, Theorems 23, 24, Corollaries 7, 8, and serves a basis for the steps of the principal algorithm, described in Sections 5.3.3, 5.3.6 (see also Observations 5-10 in Section 5.5).*

3 Mathematical and Numerical Preliminaries

In this section, we introduce notation and summarize several facts to be used in the rest of the paper.

3.1 Real Symmetric Matrices

In this subsection, we summarize several well known facts about certain symmetric tridiagonal matrices.

In the following theorem, we describe a well known characterization of the eigenvalues of a symmetric matrix.

Theorem 2 (Courant-Fisher Theorem)**.** *Suppose that $n > 0$ is a positive integer, and that A is an n by n real symmetric matrix. Suppose also that*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \tag{16}$$

are the eigenvalues of A, and $1 \leq k \leq n$ is an integer. Then,

$$\sigma_k = \max\left\{\min\left\{x^T A x : x \in U,\ \|x\| = 1\right\} :\ U \subseteq \mathbb{R}^n,\ dim(U) = k\right\}, \tag{17}$$

for every $k = 1, \ldots, n$. Also,

$$\sigma_k = \min\left\{\max\left\{x^T A x : x \in V,\ \|x\| = 1\right\} :\ V \subseteq \mathbb{R}^n,\ dim(V) = n - k + 1\right\}, \tag{18}$$

for every $k = 1, \ldots, n$.

The following theorem is an immediate consequence of Theorem 2.

Theorem 3. *Suppose that $n > 0$ is an integer, and A, B are symmetric positive definite n by n matrices. Suppose also that*

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \tag{19}$$

are the eigenvalues of A, and that

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n \tag{20}$$

are the eigenvalues of B. Suppose furthermore that $A - B$ is positive definite. Then,

$$\beta_k \leq \alpha_k, \tag{21}$$

for every integer $k = 1, \ldots, n$.

In the following theorem, we describe the eigenvalues and eigenvectors of a very simple symmetric tridiagonal matrix.

Theorem 4. *Suppose that $n > 1$ is a positive integer, and B is the n by n symmetric tridiagonal matrix, defined via the formula*

$$B = \begin{pmatrix} 2 & 1 & & & & \\ 1 & 2 & 1 & & & \\ & 1 & 2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 2 & 1 \\ & & & & 1 & 2 \end{pmatrix}. \tag{22}$$

In other words, all the diagonal entries of B are equal to 2, and all the super- and subdiagonal entries of B are equal to 1. Suppose also that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n-1} \geq \sigma_n \tag{23}$$

are the eigenvalues of B, and $v_1, \ldots, v_n \in \mathbb{R}^n$ are the corresponding unit length eigenvectors with positive first coordinate. Then,

$$\sigma_k = 2 \cdot \left(\cos\left(\frac{\pi k}{n+1} \right) + 1 \right), \tag{24}$$

for each $k = 1, \ldots, n$. Also,

$$v_k = \frac{1}{\sqrt{n}} \cdot \left(\sin\left(\frac{\pi k}{n+1} \right), \sin\left(\frac{2\pi k}{n+1} \right), \ldots, \sin\left(\frac{n\pi k}{n+1} \right) \right), \tag{25}$$

for each $k = 1, \ldots, n$.

Remark 6. *The spectral gap of the matrix B, defined via* (22)*, is*

$$\sigma_1 - \sigma_2 = 4 \cdot \sin\left(\frac{1}{2} \cdot \frac{\pi}{n+1} \right) \cdot \sin\left(\frac{3}{2} \cdot \frac{\pi}{n+1} \right) = O\left(\frac{1}{n^2} \right), \tag{26}$$

due to (24) *above. Also, the condition number of B is*

$$\kappa(B) = \frac{\sigma_1}{\sigma_n} = \frac{1 + \cos\left(\pi/(n+1)\right)}{1 - \cos\left(\pi/(n+1)\right)} = O(n^2). \tag{27}$$

3.2 Bessel Functions

In this section, we describe some well known properties of Bessel functions. All of these properties can be found, for example, in [1], [7].

Suppose that $n \geq 0$ is a non-negative integer. The Bessel function of the first kind $J_n : \mathbb{C} \to \mathbb{C}$ is defined via the formula

$$J_n(z) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \cdot (m+n)!} \cdot \left(\frac{z}{2} \right)^{2m+n}, \tag{28}$$

for all complex z. Also, the function $J_{-n} : \mathbb{C} \to \mathbb{C}$ is defined via the formula

$$J_{-n}(z) = (-1)^n \cdot J_n(z), \tag{29}$$

for all complex z.

The Bessel functions $J_0, J_{\pm 1}, J_{\pm 2}, \ldots$ satisfy the three-term recurrence relation

$$z \cdot J_{n-1}(z) + z \cdot J_{n+1}(z) = 2n \cdot J_n(z), \tag{30}$$

for any complex z and every integer n. In addition,

$$\sum_{n=-\infty}^{\infty} J_n^2(x) = 1, \tag{31}$$

for all real x. In the following theorem, we rewrite (30), (31) in the matrix form.

Theorem 5. *Suppose that $x > 0$ is a real number, and that the entries of the infinite tridiagonal symmetric matrix $A(x)$ are defined via the formulae*

$$A_{n,n-1}(x) = A_{n,n+1}(x) = 1, \tag{32}$$

for all integer n, and

$$A_{n,n}(x) = -\frac{2n}{x}, \tag{33}$$

for every integer n. Suppose also that the coordinates of the infinite vector $v(x)$ are defined via the formula

$$v_n(x) = J_n(x), \tag{34}$$

for every integer n. Then, v is a unit vector (in the l^2-sense), and, moreover,

$$A(x) \cdot v(x) = 0, \tag{35}$$

where 0 stands for the infinite dimensional zero vector.

3.3 Prolate Spheroidal Wave Functions

In this section, we summarize several well known facts about prolate spheroidal wave functions. Unless stated otherwise, all these facts can be found in [21], [17], [11], [18], [10].

Suppose that $c > 0$ is a real number. The prolate spheroidal wave functions (PSWFs) corresponding to the band limit c are the unit-norm eigenfunctions $\psi_0^{(c)}, \psi_1^{(c)}, \ldots$ of the integral operator $F_c : L^2[-1,1] \to [-1,1]$, defined via the formula

$$F_c[\varphi](x) = \int_{-1}^{1} \varphi(t) \cdot e^{icxt}\, dt. \tag{36}$$

The numerical algorithm for the evaluation of PSWFs, described in [21], is based on the following facts. Suppose that $P_0, P_1, \ldots$ are the Legendre polynomials (see, for example, [1], [7]). For every real $c > 0$ and every integer $n \geq 0$, the prolate spheroidal wave function $\psi_n^{(c)}$ can be expanded into the series of Legendre polynomials

$$\psi_n(x) = \sum_{k=0}^{\infty} \beta_k^{(n,c)} \cdot P_k(x) \cdot \sqrt{k+1/2}, \tag{37}$$

for all $-1 \leq x \leq 1$, where $\beta_0^{(n,c)}, \beta_1^{(n,c)}, \ldots$ are defined via the formula

$$\beta_k^{(n,c)} = \int_{-1}^{1} \psi_n(x) \cdot P_k(x) \cdot \sqrt{k+1/2}\, dx, \tag{38}$$

for every $k = 0, 1, 2, \ldots$. Moreover,

$$\left(\beta_0^{(n,c)}\right)^2 + \left(\beta_1^{(n,c)}\right)^2 + \left(\beta_2^{(n,c)}\right)^2 + \cdots = 1. \tag{39}$$

Suppose also that the non-zero entries of the infinite symmetric matrix $A^{(c)}$ are defined via the formulae

$$A^{(c)}_{k,k} = k(k+1) + \frac{2k(k+1)-1}{(2k+3)(2k-1)} \cdot c^2,$$
$$A^{(c)}_{k,k+2} = A^{(c)}_{k+2,k} = \frac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+1)(2k+5)}} \cdot c^2, \tag{40}$$

for every $k = 0, 1, 2, \ldots$.

The matrix $A^{(c)}$ is not tridiagonal; however, it naturally splits into two symmetric tridiagonal infinite matrices $A^{c,even}$ and $A^{c,odd}$, defined, respectively, via

$$A^{c,even} = \begin{pmatrix} A^{(c)}_{0,0} & A^{(c)}_{0,2} & & & \\ A^{(c)}_{2,0} & A^{(c)}_{2,2} & A^{(c)}_{2,4} & & \\ & A^{(c)}_{4,2} & A^{(c)}_{4,4} & A^{(c)}_{4,6} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{41}$$

and

$$A^{c,odd} = \begin{pmatrix} A^{(c)}_{1,1} & A^{(c)}_{1,3} & & & \\ A^{(c)}_{3,1} & A^{(c)}_{3,3} & A^{(c)}_{3,5} & & \\ & A^{(c)}_{5,3} & A^{(c)}_{5,5} & A^{(c)}_{5,7} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{42}$$

Suppose that $\chi^{(c)}_0 < \chi^{(c)}_2 < \ldots$ are the eigenvalues of $A^{c,even}$, and $\chi^{(c)}_1 < \chi^{(c)}_3 < \ldots$ are the eigenvalues of $A^{c,odd}$. If n is even, then

$$A^{c,even} \cdot \left(\beta^{(n,c)}_0, \beta^{(n,c)}_2, \ldots\right)^T = \chi^{(c)}_n \cdot \left(\beta^{(n,c)}_0, \beta^{(n,c)}_2, \ldots\right)^T, \tag{43}$$

and $\beta^{(n,c)}_k = 0$ for every odd k. If n is odd, then

$$A^{c,odd} \cdot \left(\beta^{(n,c)}_1, \beta^{(n,c)}_3, \ldots\right)^T = \chi^{(c)}_n \cdot \left(\beta^{(n,c)}_1, \beta^{(n,c)}_3, \ldots\right)^T, \tag{44}$$

and $\beta^{(n,c)}_k = 0$ for every even k. The eigenvectors in (43), (44) have unit length, due to (39).

The algorithm for the evaluation of $\psi^{(c)}_n$ starts with computing the corresponding eigenvector of $A^{c,even}$, if n is even, or of $A^{c,odd}$, if n of odd. In practice, these matrices are replaced with their $N \times N$ upper left square submatrices, where $N > 0$ is a sufficiently large integer (see e.g. [21] for further details). The coordinates of this eigenvector are the coefficients of the expansion of $\psi^{(c)}_n$ into the Legendre series (37) (see also (43), (44)). Once this coefficients are precomputed, we evaluate $\psi^{(c)}_n(x)$ for any given $-1 \le x \le 1$ via evaluating the sum

$$\sum_{k=0}^{2N-1} \beta^{(n,c)}_k \cdot P_k(x) \cdot \sqrt{k+1/2}, \tag{45}$$

in $O(N)$ operations (see e.g. [21], [14], [15] for further details).

3.4 Numerical Tools

In this subsection, we summarize several numerical techniques to be used in this paper.

3.4.1 Power and Inverse Power Methods

The methods described in this subsection are widely known and can be found, for example, in [3], [6]. Suppose that A is an $n \times n$ real symmetric matrix, whose eigenvalues satisfy

$$|\sigma_1| > |\sigma_2| \geq |\sigma_3| \geq \cdots \geq |\sigma_n|. \tag{46}$$

The Power Method evaluates σ_1 and the corresponding unit eigenvector in the following way.

- Set v_0 to be a random vector in $\mathbb{R}^n$ such that $\|v_0\| = \sqrt{v_0^T v_0} = 1$.
- Set $j = 1$ and $\eta_0 = 0$.
- Compute $\hat{v}_j = A v_{j-1}$.
- Set $\eta_j = v_{j-1}^T \hat{v}_j$.
- Set $v_j = \hat{v}_j / \|\hat{v}_j\|$.
- If $|\eta_j - \eta_{j-1}|$ is "sufficiently small", stop.
- Otherwise, set $j = j + 1$ and repeat the iteration.

The output value η_j approximates σ_1, and v_j approximates a unit eigenvector corresponding to σ_1. The cost of each iteration is dominated by the cost of evaluating Av_{j-1}. The rate of convergence of the algorithm is linear, and equals and the error after j iterations is of order $(|\sigma_2|/|\sigma_1|)^j$.

Remark 7. *In this paper, we use the modification of this algorithm, in which i and η_j are evaluated via the formulae*

$$i = \operatorname{argmax} \{|v_{j-1}(k)| \ : \ k = 1, \ldots, n\}, \quad \eta_j = \frac{\hat{v}_j(i)}{v_{j-1}(i)}. \tag{47}$$

The Inverse Power Method evaluates the eigenvalue σ_k of A and a corresponding unit eigenvector, given an initial approximation σ of σ_k that satisfies the inequality

$$|\sigma - \sigma_k| < \max \{|\sigma - \sigma_j| \ : \ j \neq k\}. \tag{48}$$

Conceptually, the Inverse Power Method is an application of the Power Method on the matrix $B = (A - \sigma I)^{-1}$. In practice, B does not have to be evaluated explicitly, and it suffices to be able to solve the linear system of equations

$$(A - \sigma I)\, \hat{v}_j = v_{j-1}, \tag{49}$$

for the unknown $\hat{v}_j$, on each iteration of the algorithm.

Remark 8. *If the matrix A is tridiagonal, the system* (49) *can be solved in $O(n)$ operations, for example, by means of Gaussian elimination or QR decomposition (see e.g. [3], [6], [20]).*

3.4.2 Sturm Sequence

The following theorem can be found, for example, in [20] (see also [2]). It provides the basis for an algorithm for evaluating the kth smallest eigenvalue of a symmetric tridiagonal matrix.

Theorem 6 (Sturm sequence)**.** *Suppose that*

$$C = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & \cdots & 0 \\ b_2 & a_2 & b_3 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & b_{n-1} & a_{n-1} & b_n \\ 0 & \cdots & \cdots & 0 & b_n & a_n \end{pmatrix} \tag{50}$$

is a symmetric tridiagonal matrix such that none of $b_2, \ldots, b_n$ is zero. Then, its n eigenvalues satisfy

$$\sigma_1(C) < \cdots < \sigma_n(C). \tag{51}$$

Suppose also that C_k is the $k \times k$ leading principal submatrix of C, for every integer $k = 1, \ldots, n$. We define the polynomials $p_{-1}, p_0, \ldots, p_n$ via the formulae

$$p_{-1}(x) = 0, \quad p_0(x) = 1 \tag{52}$$

and

$$p_k(x) = det\left(C_k - xI_k\right), \tag{53}$$

for every $k = 2, \ldots, n$. In other words, p_k is the characteristic polynomials of C_k, for every $k = 1, \ldots, n$. Then,

$$p_k(x) = (a_k - x)\, p_{k-1}(x) - b_k^2 p_{k-2}(x), \tag{54}$$

for every integer $k = 1, 2, \ldots, n$. Suppose furthermore, that, for any real number σ, the integer $A(\sigma)$ is defined to be the number of agreements of sign of consecutive elements of the sequence

$$p_0(\sigma), p_1(\sigma), \ldots, p_n(\sigma), \tag{55}$$

where the sign of $p_k(\sigma)$ is taken to be opposite to the sign of $p_{k-1}(\sigma)$ if $p_k(\sigma)$ is zero. In other words,

$$A(\sigma) = \sum_{k=1}^{n} \text{agree}\left(p_{k-1}(\sigma), p_k(\sigma)\right), \tag{56}$$

where, for any pair of real numbers a, b, the integer $\text{agree}(a, b)$ *is defined via*

$$\text{agree}(a, b) = \begin{cases} 1 & \text{if } ab > 0, \\ 1 & \text{if } a = 0,\ b \neq 0, \\ 0 & \text{if } b = 0, \\ 0 & \text{if } ab < 0. \end{cases} \tag{57}$$

Then, the number of eigenvalues of C that are strictly larger than σ is precisely $A(\sigma)$.

Corollary 1 (Sturm bisection)**.** *The eigenvalue $\sigma_k(C)$ of* (50) *can be found by means of bisection, each iteration of which costs $O(n)$ operations.*

Proof. We initialize the bisection by choosing $x_0 < \sigma_k(C) < y_0$. Then we set $j = 0$ and iterate as follows.

- Set $z_j = (x_j + y_j)/2$.
- If $y_j - x_j$ is small enough, stop and return z_j.
- Compute $A_j = A(z_j)$ using (54) and (55).
- If $A_j \geq k$, set $x_{j+1} = z_j$ and $y_{j+1} = y_j$.
- If $A_j < k$, set $x_{j+1} = x_j$ and $y_{j+1} = z_j$.
- Increase j by one and go to the first step.

In the end $|\sigma_k(C) - z_j|$ is at most $y_j - x_j$. The cost of the algorithm is due to (54) and the definition of $A(\sigma)$. ■

3.4.3 Evaluation of Bessel Functions

The following numerical algorithm for the evaluation of Bessel functions (see Section 6.1) at a given point is based on Theorem 5 (see e.g [1]).

Suppose that $x > 0$ is a real number, and that $n > 0$ is an integer. The following numerical algorithm evaluates $J_0(x), J_{\pm 1}, \ldots, J_{\pm n}(x)$.

Algorithm: evaluation of $J_0(x), J_{\pm 1}, \ldots, J_{\pm n}(x)$.

- select integer $N > n$ (such N is also greater than x).
- set $\tilde{J}_N = 1$ and $\tilde{J}_{N+1} = 0$.
- evaluate $\tilde{J}_{N-1}, \tilde{J}_{N-2}, \ldots, \tilde{J}_1$ iteratively via the recurrence relation (30), in the direction of decreasing indices. In other words,

$$\tilde{J}_{k-1} = \frac{2k}{x} \cdot \tilde{J}_k(x) - \tilde{J}_{k+1}(x), \tag{58}$$

  for every $k = N, \ldots, 2$.
- evaluate $\tilde{J}_{-1}$ from $\tilde{J}_1$ via (29).
- evaluate $\tilde{J}_0$ from $\tilde{J}_1, \tilde{J}_{-1}$ via (30).
- evaluate d via the formula

$$d = \sqrt{\tilde{J}_0^2 + 2 \cdot \sum_{k=1}^{N} \tilde{J}_k^2}. \tag{59}$$

- return $\tilde{J}_0/d, \tilde{J}_1/d, \ldots, \tilde{J}_n/d$.

Discussion of such properties of this algorithms as the accuracy to which $\tilde{J}_0/d$, $\tilde{J}_1/d$, $\ldots$, $\tilde{J}_n/d$ approximate $J_0(x), J_1(x), \ldots, J_n(x)$, or the choice of N, are beyond the scope of this paper (see, however, Section 7). Nevertheless, we make the following remark about this algorithm.

Remark 9. *Suppose that $x > 0$ is a real number, that $0 < n < N$ are integers, and that the real numbers $\tilde{J}_0, \ldots, \tilde{J}_N$ and d are those of the algorithm above (see (58), (59)). Suppose also that A is the symmetric tridiagonal $2N+1$ by $2N+1$ matrix, whose non-zero entries are defined via the formulae*

$$A_{j,j+1} = A_{j+1,j} = 1, \tag{60}$$

for every $j = 1, \ldots, 2N$, and

$$A_{j,j} = A_j = 2 + \frac{2j}{x}, \tag{61}$$

for every $j = 1, \ldots, 2N, 2N+1$. Then, in exact arithmetics, the vector $X = (X_1, \ldots, X_{2N+1}) \in \mathbb{R}^{2N+1}$, defined via

$$X = \frac{1}{d} \cdot \left(\tilde{J}_N, \ldots, \tilde{J}_1, \tilde{J}_0, -\tilde{J}_1, \ldots, (-1)^N \cdot \tilde{J}_N \right), \tag{62}$$

is the unit-length eigenvector of A, corresponding to the eigenvalue λ, defined via

$$\lambda = 2 + \frac{2 \cdot (N+1)}{x}. \tag{63}$$

In particular, $X_1 > 0$.

4 Analytical Apparatus

The purpose of this section is to provide the analytical apparatus to be used in the rest of the paper.

4.1 Properties of Certain Tridiagonal Matrices

In this subsection, we study some general properties of the matrices of the following type.

Definition 1. *Suppose that $n > 1$ is a positive integer, and $0 < f_1 < f_2 < \ldots$ is an increasing sequence of positive real numbers. We define the symmetric tridiagonal n by n matrix $A = A(f)$ via the formula*

$$A = \begin{pmatrix} A_1 & 1 & & & & \\ 1 & A_2 & 1 & & & \\ & 1 & A_3 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & A_{n-1} & 1 \\ & & & & 1 & A_n \end{pmatrix}, \tag{64}$$

where, for every integer $k > 0$, the real number A_k is defined via the formula

$$A_k = 2 + f_k. \tag{65}$$

In the following theorem, we describe some properties of the eigenvectors of the matrix A of Definition 1.

Theorem 7. *Suppose that $n > 1$ is an integer, $0 < f_1 < f_2 < \dots$ is an increasing sequence of positive real numbers, and the symmetric tridiagonal n by n matrix A is that of Definition 1. Suppose also that the real number λ is an eigenvalue of A, and $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is an eigenvector corresponding to λ. Then,*

$$x_2 = (\lambda - A_1) \cdot x_1. \tag{66}$$

Also,

$$\begin{aligned} x_3 &= (\lambda - A_2) \cdot x_2 - x_1, \\ &\dots \\ x_{k+1} &= (\lambda - A_k) \cdot x_k - x_{k-1}, \\ &\dots \\ x_n &= (\lambda - A_{n-1}) \cdot x_{n-1} - x_{n-2}. \end{aligned} \tag{67}$$

Finally,

$$x_{n-1} = (\lambda - A_n) \cdot x_n. \tag{68}$$

In particular, both x_1 and x_n differ from zero, and all eigenvalues of A are simple.

Proof. The identities (66), (67), (68) follow immediately from Definition 1 and the fact that

$$Ax = \lambda x. \tag{69}$$

To prove that the eigenvalues are simple, we observe that the coordinates $x_2, \dots, x_n$ of any eigenvector corresponding to λ are completely determined by x_1 via (66), (67) (or, alternatively, via (68), (67)). Obviously, neither x_1 nor x_n can be equal to zero, for otherwise x would be the zero vector. ■

The following theorem follows directly from Theorem 7 and Theorem 6 in Section 3.4.2.

Theorem 8. *Suppose that $n > 1$ is an integer, $0 < f_1 < f_2 < \dots$ is an increasing sequence of positive real numbers, and the symmetric tridiagonal n by n matrix A is that of Definition 1. Suppose also that $1 \le k \le n$ is an integer, that λ_k is the kth smallest eigenvalue of A, and that $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is an eigenvector corresponding to λ_k. Then,*

$$\sum_{j=1}^{n-1} \text{agree}\,(x_j, x_{j+1}) = k - 1, \tag{70}$$

where agree *is defined via* (57) *in Theorem 6.*

Proof. We define the n by n matrix C via the formula

$$C = \lambda_k \cdot I - A. \tag{71}$$

Suppose that, for any real t, the sequence $p_0(t), \ldots, p_n(t)$ is the Sturm sequence, defined via (55) in Theorem 6. Suppose also that $x = (x_1, \ldots, x_n)$ is the eigenvector of A corresponding to λ_k, and that

$$x_1 = 1. \tag{72}$$

We combine (72) with (66), (67) of Theorem 7 and (52), (54) of Theorem 6 to conclude that

$$p_0(0) = x_1, \quad p_1(0) = x_2, \quad \ldots, \quad p_{n-1}(0) = x_n. \tag{73}$$

Moreover, due to the combination of (68) and (54),

$$p_n(0) = 0. \tag{74}$$

We observe that the matrix C, defined via (71), has $k-1$ positive eigenvalues, and combine this observation with (56), (57) of Theorem 6 and (73), (74) to obtain (70). ■

Corollary 2. *Suppose that $n > 1$ is an integer, $0 < f_1 < f_2 < \ldots$ is an increasing sequence of positive real numbers, and the symmetric tridiagonal n by n matrix A is that of Definition 1. Suppose also that λ_1 and λ_n are, respectively, the minimal and maximal eigenvalues of A. Then,*

$$\lambda_1 < A_1, \tag{75}$$

and

$$\lambda_n > A_n. \tag{76}$$

Proof. Suppose that $x = (x_1, \ldots, x_n)$ is an eigenvector corresponding to λ_1. Due to the combination of Theorem 7 and Theorem 8,

$$x_1 \cdot x_2 < 0. \tag{77}$$

We combine (77) with (66) to obtain (75). On the other hand, due to the combination of Theorem 7 and Theorem 8,

$$x_{n-1} \cdot x_n > 0. \tag{78}$$

We combine (78) with (66) to obtain (76). ■

The following theorem is a direct consequence of Theorems 3, 4 in Section 3.1.

Theorem 9. *Suppose that $n > 1$ is an integer, $0 < f_1 < f_2 < \ldots$ is an increasing sequence of positive real numbers, and the symmetric tridiagonal n by n matrix A is that of Definition 1. Suppose also that*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n \tag{79}$$

are the eigenvalues of A. Then,

$$\lambda_k > 2 \cdot \left(1 - \cos\left(\frac{\pi k}{n+1}\right)\right), \tag{80}$$

for every integer $k = 1, \ldots, n$. In particular, A is positive definite. Also,

$$\lambda_k > A_k - 2, \tag{81}$$

for every $k = 1, \ldots, n$.

Proof. The inequalities (80), (81) follow from the combination of Theorems 3, 4 in Section 3.1 and Definition 1. ■

In the following theorem, we provide an upper bound on each eigenvalue of A.

Theorem 10. *Suppose that $n > 1$ is an integer, $0 < f_1 < f_2 < \ldots$ is an increasing sequence of positive real numbers, and the symmetric tridiagonal n by n matrix A is that of Definition 1. Suppose also that $1 \leq k \leq n$ is an integer, and that λ_k is the kth smallest eigenvalue of A. Then,*

$$\lambda_k \leq 2 + A_k. \tag{82}$$

Proof. Suppose, by contradiction, that

$$\lambda_k > 2 + A_k. \tag{83}$$

Suppose that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is an eigenvector of A corresponding to λ_k. Suppose also that

$$x_1 > 0. \tag{84}$$

It follows from (83) that

$$\lambda_k - A_1 > \lambda_k - A_2 > \cdots > \lambda_k - A_k > 2. \tag{85}$$

We combine (66), (67) in Theorem 7 with (84), (85) to conclude that

$$x_1 > 0, \quad x_2 > 0, \quad \ldots, \quad x_{k+1} > 0. \tag{86}$$

We combine (86) with (57) in Theorem 6 in Section 3.4.2 to conclude that

$$\sum_{j=1}^{n-1} \text{agree}\,(x_j, x_{j+1}) \geq k, \tag{87}$$

in contradiction to Theorem 8. ■

4.2 Local Properties of Eigenvectors

Throughout this subsection, $n > 0$ is an integer, and $0 < f_1 < f_2 < \dots$ is an increasing sequence of positive real numbers. We will be considering the corresponding symmetric tridiagonal n by n matrix $A = A(f)$ (see Definition 1 in Section 4.1).

In the following theorem, we assert that, under certain conditions, the first element of the eigenvectors of A must be "small".

Theorem 11. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1. Suppose also that λ is an eigenvalue of A, and that $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector whose first coordinate is positive, i.e. $x_1 > 0$. Suppose, in addition, that $1 \leq k \leq n$ is an integer, and that*

$$\lambda \geq A_k + 2. \tag{88}$$

Then,

$$0 < x_1 < x_2 < \cdots < x_k < x_{k+1}. \tag{89}$$

Also,

$$\frac{x_j}{x_{j-1}} > \frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{90}$$

for every $j = 2, \dots, k$. In addition,

$$1 < \frac{x_k}{x_{k-1}} < \cdots < \frac{x_3}{x_2} < \frac{x_2}{x_1}. \tag{91}$$

Proof. It follows from (88) that

$$\lambda_k - A_1 > \lambda_k - A_2 > \cdots > \lambda_k - A_k \geq 2. \tag{92}$$

We combine (66), (67) in Theorem 7 with (92) to obtain (89) by induction. Next, suppose that the real numbers $r_1, \dots, r_k$ are defined via

$$r_j = \frac{x_{j+1}}{x_j}, \tag{93}$$

for every $j = 1, \dots, k$. Also, we define the real numbers $\sigma_1, \dots, \sigma_k$ via the formula

$$\sigma_j = \frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{94}$$

for every $j = 1, \dots, k$. In other words, σ_j is the largest root of the quadratic equation

$$x^2 - (\lambda - A_j) \cdot x + 1 = 0. \tag{95}$$

We observe that

$$\sigma_1 > \cdots > \sigma_k \geq 1, \tag{96}$$

due to (92) and (94). Also,

$$r_1 > \sigma_1 > \sigma_2 > 1, \tag{97}$$

due to the combination of (94) and (66). Suppose now, by induction, that

$$r_{j-1} > \sigma_j > 1. \tag{98}$$

for some $2 \leq j \leq k-1$. We observe that the roots of the quadratic equation (95) are $1/\sigma_j < 1 < \sigma_j$, and combine this observation with (98) to obtain

$$r_{j-1}^2 - (\lambda - A_j) \cdot r_{j-1} + 1 > 0. \tag{99}$$

We combine (99) with (93) and (67) to obtain

$$r_j = \frac{x_{j+1}}{x_j} = \frac{(\lambda - A_j) \cdot x_j - x_{j-1}}{x_j} = \lambda - A_j - \frac{1}{r_{j-1}} < r_{j-1}. \tag{100}$$

Also, we combine (94), (98), (100) to obtain

$$r_j = \lambda - A_j - \frac{1}{r_{j-1}} > \lambda - A_j - \frac{1}{\sigma_j} = \frac{(\lambda - A_j) \cdot \sigma_j - 1}{\sigma_j} = \sigma_j > \sigma_{j+1}. \tag{101}$$

In other words, (98) implies (101), and we combine this observation with (97) to obtain

$$r_1 > \sigma_2, \quad r_2 > \sigma_3, \quad \ldots, \quad r_{k-1} > \sigma_k. \tag{102}$$

Also, due to (100),

$$r_1 > r_2 > \cdots > r_{k-1}. \tag{103}$$

We combine (93), (94), (102), (103) to obtain (90), (91). ■

Corollary 3. *Under the assumptions of Theorem 11,*

$$\frac{x_k}{x_1} > \prod_{j=2}^{k} \left(\frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2} \right)^2 - 1} \right). \tag{104}$$

Remark 10. *In [12], [13] the derivation of an upper bound on the first coordinate of an eigenvector of a certain matrix is based on analysis, similar to that of Theorem 11.*

Theorem 12. *Suppose that the integers $1 \leq k < n$, the real numbers $A_1, \ldots, A_n$ and λ, and the vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ are those of Theorem 11. Suppose also that $2 \leq j \leq k$ is an integer, and that $\varepsilon_{j-1}, \varepsilon_j$ are real numbers. Suppose furthermore that the real number $\tilde{x}_{j+1}$ is defined via the formula*

$$\tilde{x}_{j+1} = (\lambda - A_j) \cdot x_j \cdot (1 + \varepsilon_j) - x_{j-1} \cdot (1 + \varepsilon_{j-1}). \tag{105}$$

Then,

$$\left| \frac{\tilde{x}_{j+1} - x_{j+1}}{x_{j+1}} \right| \leq |\varepsilon_j| \cdot \frac{\lambda - A_j}{\lambda - A_j - 1} + |\varepsilon_{j-1}| \cdot \frac{1}{(\lambda - A_j - 1)^2}. \tag{106}$$

In particular,

$$\left| \frac{\tilde{x}_{j+1} - x_{j+1}}{x_{j+1}} \right| \leq |\varepsilon_j| \cdot \frac{2 + A_k - A_j}{1 + A_k - A_j} + |\varepsilon_{j-1}| \cdot \frac{1}{(1 + A_k - A_j)^2}. \tag{107}$$

Proof. We combine (105) with (67) of Theorem 7 to obtain

$$\tilde{x}_{j+1} - x_{j+1} = (\lambda - A_j) \cdot x_j \cdot \varepsilon_j - x_{j-1} \cdot \varepsilon_{j-1}. \tag{108}$$

Also, due to the combination of (67) and (89) of Theorem 11,

$$\frac{x_j}{x_{j+1}} < \frac{1}{\lambda - A_j - 1}. \tag{109}$$

Then, due to (91) of Theorem 11,

$$\frac{x_{j-1}}{x_{j+1}} < \frac{x_j^2}{x_{j+1}^2}. \tag{110}$$

We combine (108), (109) and (110) to obtain (106). We combine (106) with (88) to obtain (107). ■

Remark 11. *While the analysis along the lines the proof of Theorem 11 would lead to a stronger inequality that* (106)*, the latter is sufficient for the purposes of this paper.*

In the following theorem, we study the behavior of several last elements of an eigenvector of A.

Theorem 13. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1. Suppose also that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector whose last coordinate is positive, i.e. $x_n < 0$. Suppose, in addition, that $1 \leq k \leq n$ is an integer, and that*

$$\lambda \leq A_k - 2. \tag{111}$$

Then,

$$0 < |x_n| < |x_{n-1}| < \cdots < |x_k| < |x_{k-1}|. \tag{112}$$

Also,

$$-\frac{x_j}{x_{j+1}} > \frac{A_j - \lambda}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{113}$$

for every $j = k, \ldots, n-1$. *In addition,*

$$-1 > \frac{x_k}{x_{k+1}} > \cdots > \frac{x_{n-2}}{x_{n-1}} > \frac{x_{n-1}}{x_n}. \tag{114}$$

Proof. We defined the real numbers $y_1, \ldots, y_n$ via the formula

$$y_j = (-1)^{j+1} \cdot x_{n+1-j}, \tag{115}$$

for every $j = 1, \ldots, n$. Also, we define the real numbers $B_1, \ldots, B_n$ via the formula

$$B_j = -A_{n+1-j}, \tag{116}$$

for every $j = 1, \ldots, n$. In addition, we define the real number μ via the formula

$$\mu = -\lambda. \tag{117}$$

We combine (115), (116), (117) with (67), (68) of Theorem 7 to obtain

$$y_2 = (\mu - B_1) \cdot y_1 > 0 \tag{118}$$

and

$$y_{j+1} = (\mu - B_j) \cdot y_j - y_{j-1}, \tag{119}$$

for every $j = 2, \ldots, n-1$. We combine (116), (117) with (111) to obtain

$$\mu \geq B_{n+1-k} + 2. \tag{120}$$

Following the proof of Theorem 11 (with x_j, A_j, λ replaced, respectively, with y_j, B_j, μ) we use (118), (119), (120) to obtain

$$0 < y_1 < y_2 < \cdots < y_{n+1-k} < y_{n+2-k}, \tag{121}$$

as well as

$$\frac{y_j}{y_{j-1}} > \frac{\mu - B_j}{2} + \sqrt{\left(\frac{\mu - B_j}{2}\right)^2 - 1}, \tag{122}$$

for every $j = 2, \ldots, n+1-k$, and

$$1 < \frac{y_{n+1-k}}{y_{n-k}} < \cdots < \frac{y_3}{y_2} < \frac{y_2}{y_1}. \tag{123}$$

Now (112), (113), (114) follow from the combination of (115), (116), (117), (121), (122) and (123). ■

The following definition will be used to emphasize that the elements of an eigenvector of A exhibit different behaviors "in the beginning", "in the middle" and "in the end".

Definition 2. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1. Suppose also that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector whose first coordinate is positive, i.e. $x_1 > 0$. The eigenvalue λ induces the division of the set $\{1, \ldots, n\}$ into three disjoint parts $I_G(\lambda), I_O(\lambda), I_D(\lambda)$, defined via the formulae*

$$I_G(\lambda) = \{1 \leq j \leq n \ : \ \lambda - A_j \geq 2\}, \tag{124}$$
$$I_O(\lambda) = \{1 \leq j \leq n \ : \ -2 > \lambda - A_j > 2\}, \tag{125}$$
$$I_D(\lambda) = \{1 \leq j \leq n \ : \ -2 \geq \lambda - A_j\}. \tag{126}$$

In other words, if

$$\lambda - A_k \geq 2 > \lambda - A_{k+1} > \cdots > \lambda - A_m > -2 \geq \lambda - A_{m+1} \tag{127}$$

for some integer k, m, then

$$\begin{aligned} I_G(\lambda) &= \{1, \ldots, k\}, \\ I_O(\lambda) &= \{k+1, \ldots, m\}, \\ I_D(\lambda) &= \{m+1, \ldots, n\}. \end{aligned} \tag{128}$$

The sets $I_G(\lambda), I_O(\lambda), I_D(\lambda)$ are referred to as the "region of growth", "oscillatory region" and "region of decay", respectively.

Remark 12. *Obviously, depending on λ, some of $I_G(\lambda), I_O(\lambda), I_D(\lambda)$ may be empty. For example, for the matrix A of Theorem 4 in Section 3.1, both $I_G(\lambda)$ and $I_D(\lambda)$ are empty for any eigenvalue λ.*

Remark 13. *The terms "region of growth" and "region of decay" are justified by Theorems 11, 13, respectively. Loosely speaking, in the region of growth, the elements of x have the same sign, and their absolute values grow as the indices increase. On the other hand, in the region of decay, the elements of x have alternating signs, and their absolute values decay as the indices increase.*

In the rest of this subsection, we investigate the behavior of the elements of an eigenvector corresponding to λ in the oscillatory region $I_O(\lambda)$ (see Definition 2). In the following theorem, we partially justify the term "oscillatory region" of Definition 2.

Theorem 14. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector. Suppose also that the set $I_O(\lambda)$, defined via (125) in Definition 2, is*

$$I_O(\lambda) = \{k+1, \ldots, m\}, \tag{129}$$

for some integer $1 \leq k < m < n$. Suppose furthermore that the real numbers $\theta_k, \ldots, \theta_{m-1}$ are defined via the formula

$$\theta_j = \arccos\left(\frac{\lambda - A_{j+1}}{2}\right), \tag{130}$$

for every $j = k, \ldots, m-1$, that the two dimensional vectors $v_k, \ldots, v_m$ are defined via the formula

$$v_j = \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix}, \tag{131}$$

for every integer $j = k, \ldots, m$, and that the 2 by 2 matrices $Q_k, \ldots, Q_{m-1}$ are defined via the formula

$$Q_j = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix}, \tag{132}$$

for every integer $j = k, \ldots, m-1$. Then,

$$v_{j+1} = Q_j \cdot v_j, \tag{133}$$

for every integer $j = k, \ldots, m-1$.

Proof. We observe that

$$0 < \theta_k < \theta_{k+1} < \cdots < \theta_{m-1} < \pi, \tag{134}$$

due to the combination of (125) and (130). We also observe that the recurrence relation (133) follows immediately from the combination of Theorem 7 and (130), (131), (132). ■

In the following theorem, we describe some elementary properties of a certain 2 by 2 matrix.

Theorem 15. *Suppose that $0 < \theta < \pi$ is a real number. Then,*

$$\frac{i}{2\sin(\theta)} \cdot \begin{pmatrix} e^{-i\theta} & -1 \\ -1 & e^{-i\theta} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix} = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}. \tag{135}$$

Suppose also that x, y are real numbers. Then,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ e^{i\theta} \end{pmatrix} + \beta \begin{pmatrix} e^{i\theta} \\ 1 \end{pmatrix}, \tag{136}$$

where the complex numbers α, β are defined via

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{i}{2\sin(\theta)} \cdot \begin{pmatrix} e^{-i\theta} & -1 \\ -1 & e^{-i\theta} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{137}$$

moreover,

$$\beta = e^{-i\theta} \cdot \overline{\alpha}, \tag{138}$$

where $\overline{\alpha}$ denotes the complex conjugate of α. In particular,

$$x = 2\Re(\alpha), \tag{139}$$

$$y = 2\Re(e^{i\theta} \cdot \alpha), \tag{140}$$

where $\Re(z)$ denotes the real part of any complex number z.

Proof. The proof is straightforward, elementary, and is based on the observation that

$$\begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix}^{-1} = \frac{i}{2\sin(\theta)} \cdot \begin{pmatrix} e^{-i\theta} & -1 \\ -1 & e^{-i\theta} \end{pmatrix}. \tag{141}$$

■

In the following theorem, we evaluate the condition number of the 2 by 2 matrix of Theorem 15.

Theorem 16. *Suppose that $0 < \theta < \pi$ is a real number, and that $\kappa(\theta)$ is the condition number of the matrix*

$$U(\theta) = \begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix}. \tag{142}$$

In other words, $\kappa(\theta)$ is defined via the formula

$$\kappa(\theta) = \frac{\sigma_1(\theta)}{\sigma_2(\theta)}, \tag{143}$$

where $\sigma_1(\theta), \sigma_2(\theta)$ are, respectively, the largest and smallest singular values of $U(\theta)$. Then,

$$\kappa(\theta) = \frac{1 + |\cos(\theta)|}{1 - |\cos(\theta)|} = \max\left\{ \tan\left(\frac{\theta}{2}\right), \cot\left(\frac{\theta}{2}\right) \right\}. \tag{144}$$

Proof. The proof is elementary, straightforward, and is based on the fact that the eigenvalues of the matrix $U(\theta) \cdot U(\theta)^*$ are the roots of the quadratic equation

$$x^2 - 2x + \sin^2(\theta) = 0, \tag{145}$$

in the unknown x. ■

In the following theorem, we introduce a complexification of the coordinates of an eigenvector of A in the oscillatory region (see Definitions 1, 2).

Theorem 17. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector. Suppose also that the set $I_O(\lambda)$, defined via (125) in Definition 2, is*

$$I_O(\lambda) = \{k+1, \ldots, m\}, \tag{146}$$

for some integer $1 \leq k < m < n$. Suppose furthermore that the real numbers $\theta_k, \ldots, \theta_{m-1}$ are defined via (130) *of Theorem 14, and that the complex numbers $\alpha_k, \ldots, \alpha_{m-1}$ and $\beta_k, \ldots, \beta_{m-1}$ are defined via the equation*

$$\begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix}, \tag{147}$$

in the unknowns α_j, β_j, for $j = k, \ldots, m-1$. Then,

$$x_j = 2\Re\left(\alpha_j\right), \tag{148}$$

$$x_{j+1} = 2\Re\left(\alpha_j \cdot e^{i\theta_j}\right), \tag{149}$$

for every $j = k, \ldots, m-1$. Moreover,

$$\alpha_{j+1} = \\ \alpha_j \cdot e^{i\theta_j} - i \cdot \Im\left[\frac{\cos(\theta_j) - \cos(\theta_{j+1})}{\sin(\theta_{j+1}) \cdot \sin\left(\frac{\theta_j + \theta_{j+1}}{2}\right)} \cdot \alpha_j \cdot e^{i\theta_j} \cdot e^{\frac{i}{2}\cdot(\theta_j + \theta_{j+1})}\right], \tag{150}$$

for every $j = k, \ldots, m-2$.

Proof. The identities (148), (149) follow from the combination of Theorems 14, 15. To establish (150), we proceed as follows. Suppose that $k \leq j \leq m-2$ is an integer. We combine (131), (132), (133) of Theorem 14 with (147) to obtain

$$\begin{pmatrix} x_{j+1} \\ x_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \tag{151}$$

We substitute (135), (141) of Theorem 15 into (151) to obtain

$$\begin{aligned} \begin{pmatrix} x_{j+1} \\ x_{j+2} \end{pmatrix} &= \begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} e^{i\theta_j} & 0 \\ 0 & e^{-i\theta_j} \end{pmatrix} \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \\ &= \begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \cdot e^{i\theta_j} \\ \beta_j \cdot e^{-i\theta_j} \end{pmatrix}. \end{aligned} \tag{152}$$

Next, we combine (138), (147) and (152) to obtain

$$\begin{aligned} \begin{pmatrix} \alpha_{j+1} \\ \beta_{j+1} \end{pmatrix} &= \begin{pmatrix} 1 & e^{i\theta_{j+1}} \\ e^{i\theta_{j+1}} & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \cdot e^{i\theta_j} \\ \beta_j \cdot e^{-i\theta_j} \end{pmatrix} \\ &= \frac{1}{1 - e^{2i\theta_{j+1}}} \begin{pmatrix} 1 & -e^{i\theta_{j+1}} \\ -e^{i\theta_{j+1}} & 1 \end{pmatrix} \begin{pmatrix} 1 & e^{i\theta_j} \\ e^{i\theta_j} & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \cdot e^{i\theta_j} \\ \overline{\alpha_j} \cdot e^{-2i\theta_j} \end{pmatrix} \\ &= \frac{1}{1 - e^{2i\theta_{j+1}}} \begin{pmatrix} 1 - e^{i(\theta_{j+1} + \theta_j)} & e^{i\theta_j} - e^{i\theta_{j+1}} \\ e^{i\theta_j} - e^{i\theta_{j+1}} & 1 - e^{i(\theta_{j+1} + \theta_j)} \end{pmatrix} \begin{pmatrix} \alpha_j \cdot e^{i\theta_j} \\ \overline{\alpha_j} \cdot e^{-2i\theta_j} \end{pmatrix}. \end{aligned} \tag{153}$$

It follows from (153) that

$$\begin{aligned}\alpha_{j+1} &= \frac{\left(1-e^{i(\theta_{j+1}+\theta_j)}\right)\cdot\alpha_j\cdot e^{i\theta_j}+\left(e^{i\theta_j}-e^{i\theta_{j+1}}\right)\cdot\overline{\alpha_j}\cdot e^{-2i\theta_j}}{1-e^{2i\theta_{j+1}}}\\ &= \alpha_j\cdot e^{i\theta_j}+\frac{\alpha_j\cdot e^{i(\theta_j+\theta_{j+1})}\cdot\left(e^{i\theta_{j+1}}-e^{i\theta_j}\right)+\overline{\alpha_j}\cdot e^{-i(\theta_j+\theta_{j+1})}\cdot\left(e^{i\theta_{j+1}}-e^{i(2\theta_{j+1}-\theta_j)}\right)}{1-e^{2i\theta_{j+1}}}\\ &= \alpha_j\cdot e^{i\theta_j}+\frac{\alpha_j\cdot e^{i(\theta_j+\theta_{j+1})}\cdot\left(e^{i\theta_{j+1}}-e^{i\theta_j}\right)}{1-e^{2i\theta_{j+1}}}-\frac{\overline{\alpha_j}\cdot e^{-i(\theta_j+\theta_{j+1})}\cdot\left(e^{-i\theta_{j+1}}-e^{-i\theta_j}\right)}{1-e^{-2i\theta_{j+1}}}\\ &= \alpha_j\cdot e^{i\theta_j}+2i\cdot\Im\left(\frac{\alpha_j\cdot e^{i(\theta_j+\theta_{j+1})}\cdot\left(e^{i\theta_{j+1}}-e^{i\theta_j}\right)}{1-e^{2i\theta_{j+1}}}\right),\end{aligned}\tag{154}$$

where $\Im(z)$ denotes the imaginary part of a complex number z. We observe that, for any real numbers t, h,

$$\frac{1-e^{it}}{1-e^{ih}}=\frac{\sin\left(t/2\right)}{\sin\left(h/2\right)}\cdot\exp\left[\frac{i}{2}\cdot(t-h)\right],\tag{155}$$

and use (155) to obtain

$$\frac{1-e^{i(\theta_j-\theta_{j+1})}}{1-e^{2i\theta_{j+1}}}=\frac{\sin\left((\theta_j-\theta_{j+1})/2\right)}{\sin\left(\theta_{j+1}\right)}\cdot\exp\left[\frac{i}{2}\cdot(\theta_j-3\cdot\theta_{j+1})\right].\tag{156}$$

Due to (156),

$$\begin{aligned}&\frac{e^{i(\theta_j+\theta_{j+1})}\cdot\left(e^{i\theta_{j+1}}-e^{i\theta_j}\right)}{1-e^{2i\theta_{j+1}}}=\\ &\frac{\sin\left((\theta_j-\theta_{j+1})/2\right)}{\sin\left(\theta_{j+1}\right)}\cdot\exp\left[i\cdot\left(\theta_j+\frac{\theta_j+\theta_{j+1}}{2}\right)\right].\end{aligned}\tag{157}$$

However,

$$\sin\left(\frac{\theta_j-\theta_{j+1}}{2}\right)=-\frac{\cos(\theta_j)-\cos(\theta_{j+1})}{2\cdot\sin\left((\theta_j+\theta_{j+1})/2\right)}\tag{158}$$

Finally, we substitute (157), (158) into (154) to obtain (150). ■

Corollary 4. *Under the assumptions of Theorem 17,*

$$\left|\alpha_{j+1}-\alpha_j\cdot e^{i\theta_j}\right|\le|\alpha_j|\cdot\frac{\cos(\theta_j)-\cos(\theta_{j+1})}{\sin(\theta_{j+1})\cdot\sin\left(\frac{\theta_j+\theta_{j+1}}{2}\right)},\tag{159}$$

for every $j=k,\ldots,m-2$.

Corollary 5. *Under the assumptions of Theorem 17,*

$$x_j^2+x_{j+1}^2=4R_j^2\cdot\left(1+\cos(\theta_j)\cdot\cos(\theta_j+2\xi_j)\right),\tag{160}$$

for every $j=k,\ldots,m-2$, *where the real numbers* $R_j>0$ *and* ξ_j *are defined via*

$$\alpha_j=R_j\cdot e^{i\xi_j},\tag{161}$$

for every $j=k,\ldots,m-2$.

Proof. We combine (147) of Theorem 17 and (138) of Theorem 15 to obtain

$$\begin{aligned} x_j^2 + x_{j+1}^2 &= 2 \cdot \left(|\alpha_j|^2 + |\beta_j|^2 + \cos(\theta_j) \cdot \left(\alpha_j \cdot \overline{\beta_j} + \overline{\alpha_j} \cdot \beta_j \right) \right) \\ &= 2 \cdot \left(2 \cdot |\alpha_j|^2 + 2 \cdot \cos(\theta_j) \cdot \Re \left(\alpha_j \cdot \overline{\beta_j} \right) \right) \\ &= 4 \cdot \left(|\alpha_j|^2 + \cos(\theta_j) \cdot \Re \left(\alpha_j^2 \cdot e^{i\theta_j} \right) \right). \end{aligned} \tag{162}$$

We substitute (161) into (162) to obtain (160). ∎

In the following theorem, we discuss the "zero-crossing" of certain sequences.

Theorem 18. *Suppose that $k > 0$ is an integer, and that*

$$0 < \theta_1 < \theta_2 < \cdots < \theta_k < \pi \tag{163}$$

are real numbers. Suppose also that $\varepsilon > 0$ is a real number, that the sequence $y_1, \ldots, y_{k+2}$ is defined via the formulae

$$y_1 = y_2 = 1 \tag{164}$$

and

$$\begin{pmatrix} y_{j+1} \\ y_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} y_j \\ y_{j+1} \end{pmatrix}, \tag{165}$$

for every $j = 1, \ldots, k$, and that the sequence $z_1, \ldots, z_{k+2}$ is defined via the formulae

$$z_1 = 1, \quad z_2 = 1 + \varepsilon \tag{166}$$

and

$$\begin{pmatrix} z_{j+1} \\ z_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} z_j \\ z_{j+1} \end{pmatrix}, \tag{167}$$

for every $j = 1, \ldots, k$. Suppose furthermore that

$$y_1, \ldots, y_{k+1} > 0. \tag{168}$$

Then,

$$z_1, \ldots, z_{k+1} > 0, \tag{169}$$

In other words, the sequence $z_1, \ldots, z_{k+2}$ cannot "cross zero" before the sequence $y_1, \ldots, y_{k+2}$ does.

Proof. It follows from (168) that

$$\frac{y_2}{y_1}, \ldots, \frac{y_{k+1}}{y_k} > 0. \tag{170}$$

Also, we combine (164), (165), (166), (167) to obtain

$$\frac{y_2}{y_1} = 1 < 1 + \varepsilon = \frac{z_2}{z_1}, \tag{171}$$

and the recurrence relations

$$\frac{y_{j+2}}{y_{j+1}} = 2\cos(\theta_j) - \frac{y_j}{y_{j+1}}, \tag{172}$$

for every $j = 1, \ldots, m$, and

$$\frac{z_{j+2}}{z_{j+1}} = 2\cos(\theta_j) - \frac{z_j}{z_{j+1}}, \tag{173}$$

for every $j = 1, \ldots, m$. Suppose now that $1 \leq l < k$, and that

$$\frac{y_{l+1}}{y_l} < \frac{z_{l+1}}{z_l}. \tag{174}$$

We combine (172), (173), (174) to obtain

$$\frac{y_{l+2}}{y_{l+1}} = 2\cos(\theta_l) - \frac{y_l}{y_{l+1}} < 2\cos(\theta_l) - \frac{z_l}{z_{l+1}} = \frac{z_{l+2}}{z_{l+1}}. \tag{175}$$

In other words, (174) implies (175), and we combine this observations with (171) to obtain, by induction on l,

$$\frac{y_{l+1}}{y_l} < \frac{z_{l+1}}{z_l}, \tag{176}$$

for all $l = 1, \ldots, k$. We combine (170) with (176) to obtain (169). ■

The following theorem is similar to Theorem 18 above.

Theorem 19. *Suppose that $k > 0$ is an integer, and that*

$$0 < \theta_1 < \theta_2 < \cdots < \theta_k \leq \theta < \pi \tag{177}$$

are real numbers. Suppose also that the sequence $y_1, \ldots, y_{k+2}$ is defined via the formulae

$$y_1 = y_2 = 1 \tag{178}$$

and

$$\begin{pmatrix} y_{j+1} \\ y_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} y_j \\ y_{j+1} \end{pmatrix}, \tag{179}$$

for every $j = 1, \ldots, k$, and that the sequence $t_1, \ldots, t_{k+2}$ is defined via the formulae

$$t_1 = t_2 = 1, \tag{180}$$

and

$$\begin{pmatrix} t_{j+1} \\ t_{t+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta) \end{pmatrix} \begin{pmatrix} t_j \\ t_{j+1} \end{pmatrix}, \tag{181}$$

for every $j = 1, \ldots, k$. In other words, as opposed to (179)*, the matrix is* (181) *is the same for every j. Suppose furthermore that*

$$t_1, \ldots, t_{k+1} > 0. \tag{182}$$

Then,

$$y_1, \ldots, y_{k+1} > 0, \tag{183}$$

In other words, the sequence $y_1, \ldots, y_{k+2}$ cannot "cross zero" before the sequence $t_1, \ldots, t_{k+2}$ does.

Proof. The proof is almost identical to the proof of Theorem 18, and is based on the observation that

$$2 \cdot \cos(\theta_j) - \frac{1}{\rho} \geq 2 \cdot \cos(\theta) - \frac{1}{r}, \tag{184}$$

for any $j = 1, \ldots, k$, and any real numbers $0 < r \leq \rho$. ■

In the following theorem, we address the question how long it takes for a certain sequence to cross zero.

Theorem 20. *Suppose that $k > 0$ and $l > 0$ are integers, and that*

$$0 < \theta_k < \theta_{k+1} < \cdots < \theta_{k+l} < \pi \tag{185}$$

are real numbers. Suppose also that $\varepsilon > 0$, and that the sequence $x_k, \ldots, x_{k+l+2}$ is defined via the formulae

$$x_k = 1, \quad x_{k+1} = 1 + \varepsilon, \tag{186}$$

and

$$\begin{pmatrix} x_{j+1} \\ x_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix}, \tag{187}$$

for every $j = k, \ldots, k + l$. Suppose furthermore that

$$\left(l - \frac{1}{2} \right) \cdot \theta_{k+l} < \frac{\pi}{2}. \tag{188}$$

Then,

$$x_k, x_{k+1}, \ldots, x_{k+l} > 0. \tag{189}$$

Proof. Suppose that the sequence $t_k, \ldots, t_{k+l+2}$ is defined via the formulae

$$t_k = t_{k+1} = 1 \tag{190}$$

and

$$\begin{pmatrix} t_{j+1} \\ t_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_{k+l}) \end{pmatrix} \begin{pmatrix} t_j \\ t_{j+1} \end{pmatrix}, \tag{191}$$

for every $j = k, \ldots, k+l$ (note that the matrix in (191) is the same for every j). It follows from (190), (191) that

$$\begin{pmatrix} t_{k+m} \\ t_{k+m+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_{k+l}) \end{pmatrix}^m \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{192}$$

for every $m = 0, \ldots, l+1$. We observe that

$$\begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta) \end{pmatrix}^m = \frac{1}{1-e^{2i\theta}} \begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix} \begin{pmatrix} e^{im\theta} & 0 \\ 0 & e^{-im\theta} \end{pmatrix} \begin{pmatrix} 1 & -e^{i\theta} \\ -e^{i\theta} & 1 \end{pmatrix}, \tag{193}$$

for every real θ and every integer $m = 0, 1, \ldots$ (see also Theorem 15). We combine (192), (193) to obtain

$$\begin{aligned} \begin{pmatrix} t_{k+m} \\ t_{k+m+1} \end{pmatrix} &= \frac{1}{1+e^{i\theta_{k+l}}} \begin{pmatrix} 1 & e^{i\theta_{k+l}} \\ e^{i\theta_{k+l}} & 1 \end{pmatrix} \begin{pmatrix} e^{im\theta_{k+l}} \\ e^{-im\theta_{k+l}} \end{pmatrix} \\ &= \frac{e^{-i\theta_{k+l}/2}}{2\cos\left(\theta_{k+l}/2\right)} \begin{pmatrix} e^{im\theta_{k+l}} + e^{-i(m-1)\theta_{k+l}} \\ e^{i(m+1)\theta_{k+l}} + e^{-im\theta_{k+l}} \end{pmatrix} \\ &= \frac{1}{\cos\left(\theta_{k+l}/2\right)} \begin{pmatrix} \cos\left((m-1/2)\cdot\theta_{k+l}\right) \\ \cos\left((m+1/2)\cdot\theta_{k+l}\right) \end{pmatrix}. \end{aligned} \tag{194}$$

In other words,

$$t_{k+m} = \frac{\cos\left((m-1/2)\cdot\theta_{k+l}\right)}{\cos\left(\theta_{k+l}/2\right)}, \tag{195}$$

for every $m = 0, \ldots, l+2$. We combine (195) with (188) to conclude that

$$t_{k+1}, \ldots, t_{k+l} > 0. \tag{196}$$

We combine (190), (191), (196) with Theorems 18, 19 to obtain (189). ■

Corollary 6. *Suppose, in addition, that $l \geq 3$, and that $k+1 \leq j \leq k+l-2$ is an integer. Then,*

$$x_{j+1} > \frac{x_j}{2} > \frac{x_{j-1}}{3}. \tag{197}$$

Proof. We combine (185), (187) and (189) to obtain

$$x_{j+2} = 2x_{j+1}\cos(\theta_j) - x_j > 0, \tag{198}$$

and hence

$$\frac{x_j}{2} < \frac{x_j}{2\cos(\theta_j)} < x_{j+1}. \tag{199}$$

We combine (187), (189) and (199) to obtain

$$\frac{x_j}{2} < x_{j+1} = 2x_j\cos(\theta_{j-1}) - x_{j-1}, \tag{200}$$

and hence

$$x_{j-1} < x_j \cdot \left(2\cos(\theta_{j-1}) - \frac{1}{2}\right) < \frac{3x_j}{2}. \tag{201}$$

We combine (199), (201) to obtain (197). ■

The following theorem follows from the combination of Theorems 11, 17, 20 and Corollary 6.

Theorem 21. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector with positive first coordinate, i.e. $x_1 > 0$. Suppose also that*

$$\lambda - A_1 > \cdots > \lambda - A_k \geq 2 > \lambda - A_{k+1} > \cdots > \lambda - A_{k+l+1} > 0, \tag{202}$$

for some integer $1 \leq k < n$ and $1 \leq l \leq n - k - 1$. Suppose furthermore that

$$\frac{\lambda - A_{k+l+1}}{2} > \cos\left(\frac{\pi}{2l-1}\right). \tag{203}$$

Then,

$$0 < x_1 < x_2 < \cdots < x_k < x_{k+1}. \tag{204}$$

Also,

$$x_k, x_{k+1}, \ldots, x_{k+l} > 0. \tag{205}$$

If, in addition, $l \geq 3$, then

$$x_{k+1} > \frac{2 \cdot x_k}{3}, \quad x_{k+2} > \frac{2 \cdot x_{k+1}}{3}, \quad \ldots, \quad x_{k+l-2} > \frac{2 \cdot x_{k+l-3}}{3}, \tag{206}$$

and

$$x_{k+l-1} > \frac{x_{k+l-2}}{2}. \tag{207}$$

Proof. We obtain (204) from the combination of (202) and Theorem 11. Also, we obtain (205) from the combination of Theorems 17, 20. Finally, we combine Theorems 17, 20 and Corollary 6 to obtain (206), (207). ■

Theorem 22. *Suppose that the integers $1 \leq k < n$ and $l \geq 3$, the real numbers $A_1, \ldots, A_n$ and λ, and the vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ are those of Theorem 21. Suppose also that $k+1 \leq j \leq k+l-3$ is an integer, and that $\varepsilon_{j-1}, \varepsilon_j$ are real numbers. Suppose furthermore that the real number $\tilde{x}_{j+1}$ is defined via the formula*

$$\tilde{x}_{j+1} = (\lambda - A_j) \cdot x_j \cdot (1 + \varepsilon_j) - x_{j-1} \cdot (1 + \varepsilon_{j-1}) \,. \tag{208}$$

Then,

$$\left| \frac{\tilde{x}_{j+1} - x_{j+1}}{x_{j+1}} \right| \leq \frac{3}{2} \cdot |\varepsilon_j| \cdot (\lambda - A_j) + \frac{9}{4} \cdot |\varepsilon_{j-1}|. \tag{209}$$

Proof. We combine (105) with (67) of Theorem 7 to obtain

$$\tilde{x}_{j+1} - x_{j+1} = (\lambda - A_j) \cdot x_j \cdot \varepsilon_j - x_{j-1} \cdot \varepsilon_{j-1}. \tag{210}$$

Also, due to the combination of (67) and (206) of Theorem 21,

$$\frac{x_{j-1}}{x_{j+1}} < \frac{9}{4}. \tag{211}$$

We combine (210), (211) to obtain (209). ■

Remark 14. *While a more careful analysis would lead to a stronger inequality that* (209)*, the latter is sufficient for the purposes of this paper.*

The following theorem is almost identical to Theorem 14, with the only difference that the recurrence relation is reversed.

Theorem 23. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector. Suppose also that the set $I_O(\lambda)$, defined via* (125) *in Definition 2, is*

$$I_O(\lambda) = \{k+1, \ldots, m\} \,, \tag{212}$$

for some integer $1 \leq k < m < n$. Suppose furthermore that the real numbers $\varphi_k, \ldots, \varphi_{m-1}$ are defined via the formula

$$\varphi_j = \arccos\left(-\frac{\lambda - A_{j+1}}{2} \right), \tag{213}$$

for every $j = k, \ldots, m-1$, that the two dimensional vectors $w_k, \ldots, w_{m+1}$ are defined via the formula

$$w_j = \begin{pmatrix} x_j \cdot (-1)^{m-j} \\ x_{j-1} \cdot (-1)^{m-j+1} \end{pmatrix}, \tag{214}$$

for every integer $j = k, \ldots, m+1$, and that the 2 by 2 matrices $P_k, \ldots, P_{m-1}$ are defined via the formula

$$P_j = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\varphi_j) \end{pmatrix}, \tag{215}$$

for every integer $j = k, \ldots, m-1$. Then,

$$w_j = P_{j-1} \cdot w_{j+1}, \tag{216}$$

for every integer $j = k+1, \ldots, m$.

Proof. The proof is essentially identical to that of Theorem 14. ■

The following theorem is similar to Theorem 17, with the only difference that the recurrence relation is reversed.

Theorem 24. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector. Suppose also that the set $I_O(\lambda)$, defined via (125) in Definition 2, is*

$$I_O(\lambda) = \{k+1, \ldots, m\}, \tag{217}$$

for some integer $1 \leq k < m < n$. Suppose furthermore that the real numbers $\varphi_k, \ldots, \varphi_{m-1}$ are defined via (213) of Theorem 23, and that the complex numbers $\gamma_k, \ldots, \gamma_{m-1}$ and $\delta_k, \ldots, \delta_{m-1}$ are defined via the equation

$$\begin{pmatrix} 1 & e^{i\varphi_j} \\ e^{i\varphi_j} & 1 \end{pmatrix} \begin{pmatrix} \gamma_j \\ \delta_j \end{pmatrix} = \begin{pmatrix} x_{j+2} \cdot (-1)^{m-j-2} \\ x_{j+1} \cdot (-1)^{m-j-1} \end{pmatrix}, \tag{218}$$

in the unknowns γ_j, δ_j, for $j = k, \ldots, m-1$. Then,

$$x_{j+2} = 2\Re\left(\gamma_j\right) \cdot (-1)^{m-j-2}, \tag{219}$$

$$x_{j+1} = 2\Re\left(\gamma_j \cdot e^{i\varphi_j}\right) \cdot (-1)^{m-j-1}, \tag{220}$$

for every $j = k, \ldots, m-1$. Moreover,

$$\gamma_{j-1} = \gamma_j \cdot e^{i\varphi_j} - i \cdot \Im\left[\frac{\cos(\varphi_j) - \cos(\varphi_{j-1})}{\sin(\varphi_{j-1}) \cdot \sin\left(\frac{\varphi_j + \varphi_{j-1}}{2}\right)} \cdot \gamma_j \cdot e^{i\varphi_j} \cdot e^{\frac{i}{2}\cdot(\varphi_j + \varphi_{j-1})}\right], \tag{221}$$

for every $j = k+1, \ldots, m-1$.

Proof. The proof is essentially identical to that of Theorem 17, and will be omitted. ■

Corollary 7. *Under the assumptions of Theorem 24,*

$$\left|\gamma_{j-1} - \gamma_j \cdot e^{i\varphi_j}\right| \leq |\gamma_j| \cdot \frac{\cos(\varphi_j) - \cos(\varphi_{j-1})}{\sin(\varphi_{j-1}) \cdot \sin\left(\frac{\varphi_j + \varphi_{j-1}}{2}\right)}, \tag{222}$$

for every $j = k+1, \ldots, m-1$.

Corollary 8. *Under the assumptions of Theorem 24,*

$$x_{j+1}^2 + x_{j+2}^2 = 4R_j^2 \cdot (1 + \cos(\varphi_j) \cdot \cos(\varphi_j + 2\xi_j)), \tag{223}$$

for every $j = k, \ldots, m-1$, where the real numbers $R_j > 0$ and ξ_j are defined via

$$\gamma_j = R_j \cdot e^{i\xi_j}, \tag{224}$$

for every $j = k, \ldots, m-1$.

Proof. We combine (218) of Theorem 24 and (138) of Theorem 15 to obtain

$$\begin{aligned} x_{j+1}^2 + x_{j+2}^2 &= 2 \cdot \left(|\gamma_j|^2 + |\delta_j|^2 + \cos(\varphi_j) \cdot \left(\gamma_j \cdot \overline{\delta_j} + \overline{\gamma_j} \cdot \delta_j\right)\right) \\ &= 2 \cdot \left(2 \cdot |\gamma_j|^2 + 2 \cdot \cos(\varphi_j) \cdot \Re\left(\gamma_j \cdot \overline{\delta_j}\right)\right) \\ &= 4 \cdot \left(|\gamma_j|^2 + \cos(\varphi_j) \cdot \Re\left(\gamma_j^2 \cdot e^{i\varphi_j}\right)\right). \end{aligned} \tag{225}$$

We substitute (224) into (225) to obtain (223). ■

The following theorem is analogous to Theorem 21.

Theorem 25. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, that λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a corresponding eigenvector. Suppose also that*

$$0 > \lambda - A_{m-r} > \cdots > \lambda - A_m > -2 \geq \lambda - A_{m+1}, \tag{226}$$

for some integer $1 \leq m < n$ and $1 \leq r \leq m-1$. Suppose furthermore that

$$\frac{A_{m-r} - \lambda}{2} > \cos\left(\frac{\pi}{2r-1}\right), \tag{227}$$

and that $x_m > 0$. Then,

$$0 < x_n \cdot (-1)^{n-m} < x_{n-1} \cdot (-1)^{n-1-m} < \cdots < -x_{m+1} < x_m. \tag{228}$$

Also,

$$-x_{m+1}, x_m, -x_{m-1}, \ldots, x_{m-(r-1)} \cdot (-1)^{r-1} > 0. \tag{229}$$

If, in addition, $r \geq 3$, then

$$\begin{aligned} &x_m > -\frac{2 \cdot x_{m+1}}{3}, \quad -x_{m-1} > \frac{2 \cdot x_m}{3}, \quad \ldots, \\ &(-1)^{r-3} \cdot x_{m-(r-3)} > (-1)^{r-4} \cdot \frac{2 \cdot x_{m-(r-4)}}{3}, \end{aligned} \tag{230}$$

and

$$(-1)^{r-2} \cdot x_{m-(r-2)} > (-1)^{r-3} \cdot \frac{x_{m-(r-3)}}{2}. \tag{231}$$

Proof. We obtain (228) from the combination of (226) and Theorem 13. Also, we obtain (229) from the combination of Theorems 24, 20 (see also the proof of Theorem 21). Finally, we combine Theorems 24, 20 and Corollary 6 to obtain (230), (231) (see also the proof of Theorem 21). ■

In the following theorem, we estimate the accuracy to which the complex numbers α_j of Theorem 17 are evaluated.

Theorem 26. *Suppose that the n by n symmetric tridiagonal matrix A is that of Definition 1 in Section 4.1, and that λ is an eigenvalue of A. Suppose also that $1 \leq k < n$ and $1 \leq l < n - k - 1$ are integers, and that*

$$\lambda - A_1 > \cdots > \lambda - A_k \geq 2 > \lambda - A_{k+1} > \cdots > \lambda - A_{k+l+1} \geq 0. \tag{232}$$

Suppose furthermore that the real numbers $\theta_k, \ldots, \theta_{k+l}$ are defined via (130) *in Theorem 14. Then,*

$$\frac{\cos(\theta_j) - \cos(\theta_{j+1})}{\sin(\theta_{j+1}) \cdot \sin\left(\frac{\theta_j + \theta_{j+1}}{2}\right)} < \frac{A_{j+2} - A_{j+1}}{A_{j+1} - A_{k+1}}, \tag{233}$$

for every integer $j = k+1, \ldots, k+l-1$. Also,

$$\frac{1}{1 - \cos(\theta_j)} < \frac{2}{A_{j+1} - A_{k+1}}, \tag{234}$$

for every integer $j = k+1, \ldots, k+l-1$.

Proof. We combine (130) of Theorem 14 with (232) to obtain

$$\begin{aligned} \frac{\cos(\theta_j) - \cos(\theta_{j+1})}{\sin(\theta_{j+1}) \cdot \sin\left(\frac{\theta_j + \theta_{j+1}}{2}\right)} &< \frac{\cos(\theta_j) - \cos(\theta_{j+1})}{(1 - \cos(\theta_j)) \cdot (1 + \cos(\theta_j))} < \frac{\cos(\theta_j) - \cos(\theta_{j+1})}{1 - \cos(\theta_j)} \\ &= \frac{A_{j+2} - A_{j+1}}{2 - \lambda + A_{j+1}} < \frac{A_{j+2} - A_{j+1}}{A_{j+1} - A_{k+1}}, \end{aligned} \tag{235}$$

which implies (233). By the same token,

$$\frac{1}{1 - \cos(\theta_j)} = \frac{2}{2 - \lambda + A_{j+1}} < \frac{2}{A_{j+1} - A_{k+1}}, \tag{236}$$

which implies (234). ■

Corollary 9. *Under the assumptions of Theorem 26, suppose, in addition, that $k+1 \leq j \leq k+l-1$ is an integer, and that*

$$\frac{A_{j+2} - A_{j+1}}{A_{j+1} - A_{k+1}} < \frac{1}{2}. \tag{237}$$

Then, the evaluation of α_{j+1} from α_j via (150) *in Theorem 17 is stable.*

Proof. The statement follows from the combination of Theorem 26 and Corollary 4. ■

Corollary 10. *Under the assumptions of Theorem 26, suppose, in addition, that* $k+1 \leq j \leq k+l-1$ *is an integer, and that the vector* $(x_j, x_{j+1}) \in \mathbb{R}^2$ *is given to relative accuracy* $\varepsilon > 0$. *Then, the vector* $(\alpha_j, \beta_j) \in \mathbb{C}^2$, *defined via* (147) *in Theorem 17, has relative accuracy* $r(\alpha_j)$, *where*

$$r(\alpha_j) \leq \varepsilon \cdot \frac{4}{A_{j+1} - A_{k+1}}. \tag{238}$$

Proof. The statement follows from the combination of Theorem 26 and Theorem 16. ■

4.3 Error Analysis for Certain Matrices

In this section, we consider a certain subclass of tridiagonal matrices, described in Definition 1. For the matrices of this class, we will present a detailed analysis of some of the quantities, introduced in Section 4.2.

Due to Theorem 26 and Corollaries 9, 10, the precision of the complex parameters α_j of Theorem 17 is determined by the magnitude of the right-hand side of (233), (234) in Theorem 26. Obviously, the latter depend on the choice of the sequence $0 < f_1 < f_2 < \ldots$ in Definition 1. In the following two theorems, we will estimate these quantities for several special cases.

Theorem 27. *Suppose that* $a, c, K \geq 1$ *are real numbers, that* $n \leq K^{1/a} \cdot c$ *is an integer, and that* $A_1, \ldots, A_n$ *are defined via*

$$A_j = 2 + \left(\frac{j}{c}\right)^a, \tag{239}$$

for every $j = 1, \ldots, n$. *Suppose also that the* n *by* n *symmetric tridiagonal matrix* A *is that of Definition 1, that* λ *is an eigenvalue of* A, *and that*

$$\lambda = 4 + K \cdot c^{-b}, \tag{240}$$

for some real $0 \leq b < 1$. *Suppose also that*

$$b < \frac{2a}{a+2}, \tag{241}$$

that $1 < k < n$ *and* $l \geq 3$ *are integers, that*

$$\lambda - A_1 > \cdots > \lambda - A_k \geq 2 > \lambda - A_{k+1} > \cdots > \lambda - A_{k+l+2} \geq 0, \tag{242}$$

and that

$$\frac{\lambda - A_{k+l+1}}{2} > \cos\left(\frac{\pi}{2l-1}\right) \geq \frac{\lambda - A_{k+l+2}}{2}. \tag{243}$$

Then,

$$\frac{2}{A_{k+l} - A_k} = o\left(c^{2a/(a+2)}\right), \quad c \to \infty, \tag{244}$$

and also

$$\frac{A_{k+l+1}-A_{k+l}}{A_{k+l}-A_k}=O\left(c^{-(1+b-b/a)/3}\right)=o\left(c^{-1/3}\right),\quad c\to\infty. \tag{245}$$

Proof. First, we combine (239) with (240) to conclude that

$$k^a=K\cdot c^{a-b}\cdot(1+o(1)),\quad c\to\infty. \tag{246}$$

Next, we assume that

$$l=D\cdot c^d\cdot(1+o(1)),\quad c\to\infty, \tag{247}$$

for some real numbers $D,d>0$, and that

$$\frac{l}{k}=o(1),\quad c\to\infty. \tag{248}$$

To validate these assumptions, we combine (246), (247), (248) with (243) to obtain

$$\frac{\lambda-A_{k+l}}{2}=\frac{\lambda-A_k}{2}-\frac{A_{k+l}-A_k}{2}=\cos\left(\frac{\pi}{2l}\right)\cdot(1+o(1)),\quad c\to\infty. \tag{249}$$

We combine (239), (246) and (249) to obtain

$$\frac{(k+l)^a-k^a}{c^a}=\frac{\pi^2}{4l^2}\cdot(1+o(1)),\quad c\to\infty. \tag{250}$$

We combine (246), (248) and (250) to obtain

$$ak^{a-1}l=\frac{\pi^2}{4}\cdot\frac{c^a}{l^2}\cdot(1+o(1)),\quad c\to\infty, \tag{251}$$

and substitute (246), (247) into (251) to obtain

$$\begin{aligned}l^3&=\frac{\pi^2}{4a}\cdot c^a\cdot k^{1-a}\cdot(1+o(1))\\&=\frac{\pi^2}{4a}\cdot c^a\cdot\left(K\cdot c^{a-b}\right)^{\frac{1-a}{a}}\cdot(1+o(1))\\&=\frac{\pi^2\cdot K^{\frac{1-a}{a}}}{4a}\cdot c^{1+b-b/a}\cdot(1+o(1)),\quad c\to\infty.\end{aligned} \tag{252}$$

We combine (247), (252) to obtain

$$D=\left(\frac{\pi^2\cdot K^{\frac{1-a}{a}}}{4a}\right)^{\frac{1}{3}} \tag{253}$$

and

$$d=\frac{1+b-b/a}{3}. \tag{254}$$

Due to the combination of (241) and (254), the assumptions (247), (248) were consistent. We combine (247), (249), (253), (254) to conclude that

$$\begin{aligned}\frac{2}{A_{k+l}-A_k} &= \frac{8\cdot l^2}{\pi^2}\cdot(1+o(1))\\ &= O(1)\cdot c^{\frac{2}{3}\cdot(1+b(1-1/a))}\cdot(1+o(1))\,,\quad c\to\infty.\end{aligned}\tag{255}$$

We use (241) to obtain

$$\frac{2}{3}\cdot\left(1+b\cdot\left(1-\frac{1}{a}\right)\right)<\frac{2}{3}\cdot\frac{3a}{a+2}=\frac{2a}{a+2},\tag{256}$$

and substitute (256) into (255) to obtain (244). Also, we combine (246), (247), (248) with (239), (243) to obtain

$$\begin{aligned}A_{k+l+1}-A_{k+l} &= ak^{a-1}\cdot c^{-a}\cdot(1+o(c))\\ &= O(1)\cdot c^{-a+(a-b)\cdot(a-1)/a}\cdot(1+o(c))\\ &= O(1)\cdot c^{-(1+b-b/a)}\cdot(1+o(c))\,,\quad c\to\infty.\end{aligned}\tag{257}$$

Finally, we combine (255), (256) and (257) to obtain (245). ■

Theorem 28. *Suppose that $a,c,K\geq 1$ and are real numbers, that $n\leq K^{1/a}\cdot c$ is an integer, and that $A_1,\ldots,A_n$ are defined via*

$$A_j=2+\left(\frac{j}{c}\right)^a,\tag{258}$$

for every $j=1,\ldots,n$. Suppose also that the n by n symmetric tridiagonal matrix A is that of Definition 1, that λ is an eigenvalue of A, and that

$$\lambda=4+K\cdot c^{-b},\tag{259}$$

for some real $0\leq b<1$. Suppose also that, as opposed to (241) *in Theorem 27,*

$$b\geq\frac{2a}{a+2},\tag{260}$$

that $1<k<n$ and $l\geq 3$ are integers, that

$$\lambda-A_1>\cdots>\lambda-A_k\geq 2>\lambda-A_{k+1}>\cdots>\lambda-A_{k+l+2}\geq 0,\tag{261}$$

and that

$$\frac{\lambda-A_{k+l+1}}{2}>\cos\left(\frac{\pi}{2l-1}\right)\geq\frac{\lambda-A_{k+l+2}}{2}.\tag{262}$$

Then,

$$\frac{2}{A_{k+l}-A_k}=O\left(c^{2a/(a+2)}\right),\quad c\to\infty,\tag{263}$$

and also

$$\frac{A_{k+l+1}-A_{k+l}}{A_{k+l}-A_k}=O\left(c^{-a/(a+2)}\right)=O\left(c^{-1/3}\right),\quad c\to\infty.\tag{264}$$

Proof. First, we combine (258) with (259) to conclude that

$$k^a = K \cdot c^{a-b} \cdot (1 + o(1)), \quad c \to \infty. \tag{265}$$

Next, we assume that

$$l = D \cdot c^d \cdot (1 + o(1)), \quad c \to \infty, \tag{266}$$

for some real numbers $D, d > 0$, and (as opposed to (248)) that

$$\frac{k}{l} = O(1), \quad c \to \infty. \tag{267}$$

To validate these assumptions, we combine (265), (266), (267) with (262) to obtain

$$\frac{\lambda - A_{k+l}}{2} = \frac{\lambda - A_k}{2} - \frac{A_{k+l} - A_k}{2} = \cos\left(\frac{\pi}{2l}\right) \cdot (1 + o(1)), \quad c \to \infty. \tag{268}$$

We combine (258), (265) and (268) to obtain

$$\frac{(k+l)^a - k^a}{c^a} = \frac{\pi^2}{4l^2} \cdot (1 + o(1)), \quad c \to \infty. \tag{269}$$

We combine (265), (267) and (269) to obtain

$$l^a = O(1) \cdot \frac{c^a}{l^2} \cdot (1 + o(1)), \quad c \to \infty, \tag{270}$$

and substitute (265), (266) into (270) to obtain

$$l = O(1) \cdot c^{a/(a+2)} \cdot (1 + o(1)), \quad c \to \infty. \tag{271}$$

Due to the combination of (260) and (271), the assumptions (266), (267) were consistent. We combine (266), (268), (271) to conclude that

$$\frac{2}{A_{k+l} - A_k} = \frac{8 \cdot l^2}{\pi^2} \cdot (1 + o(1)) = O\left(c^{2a/(a+2)}\right), \quad c \to \infty, \tag{272}$$

which implies (263). Also, we combine (265), (266), (267) with (258), (262) to obtain

$$\begin{aligned} A_{k+l+1} - A_{k+l} &= a \cdot l^{a-1} \cdot c^{-a} \cdot (1 + o(c)) \\ &= O\left(c^{-a+a(a-1)/(a+2)}\right) = O\left(c^{-3a/(a+2)}\right), \quad c \to \infty. \end{aligned} \tag{273}$$

Finally, we combine (272) and (273) to obtain (264). ∎

Remark 15. *Suppose that $a, k > 0$ are real numbers. Suppose also that the function $g : (0, \infty) \to \mathbb{R}$ is defined via the formula*

$$g(x) = \frac{(k + x + 1)^a - (k + x)^a}{(k + x)^a - k^a}, \tag{274}$$

for all real $x > 0$. Then, g is monotonically decreasing.

Remark 16. *Due to the combination of Theorem 10 with* (258), (259), *any eigenvalue* λ *of* A *satisfies the inequality*

$$\lambda < A_n + 2 = 4 + \left(\frac{n}{c}\right)^a = 4 + K = 4 + K \cdot c^0. \tag{275}$$

The following theorem is in the spirit of Theorems 27, 28.

Theorem 29. *Suppose that* $a, c, K \geq 1$ *are real numbers, that* $n \leq K^{1/a} \cdot c$ *is an integer, and that* $A_1, \ldots, A_n$ *are defined via*

$$A_j = 2 + \left(\frac{j}{c}\right)^a, \tag{276}$$

for every $j = 1, \ldots, n$. *Suppose also that the* n *by* n *symmetric tridiagonal matrix* A *is that of Definition 1, that* λ *is an eigenvalue of* A, *and that*

$$\lambda = 4 + K \cdot c^{-b}, \tag{277}$$

for some real $0 \leq b < 1$. *Suppose also that* $1 < m < n$ *and* $r \geq 3$ *are integers, that*

$$0 > \lambda - A_{m-r} > \cdots > \lambda - A_m > -2 \geq \lambda - A_{m+1}, \tag{278}$$

and that

$$\frac{A_{m-r} - \lambda}{2} > \cos\left(\frac{\pi}{2r-1}\right) \geq \frac{A_{m-r-1} - \lambda}{2}. \tag{279}$$

Then,

$$\frac{2}{A_m - A_{m-r}} = O\left(c^{2/3}\right), \quad c \to \infty, \tag{280}$$

and also

$$\frac{A_{m-r} - A_{m-r-1}}{A_m - A_{m-r}} = O\left(c^{-1/3}\right), \quad c \to \infty. \tag{281}$$

Proof. We combine (276) and (277) to obtain

$$m = 4^{1/a} \cdot c \cdot (1 + o(c)), \quad c \to \infty. \tag{282}$$

Next, we assume that

$$r = D \cdot c^d \cdot (1 + o(1)), \quad c \to \infty, \tag{283}$$

for some real numbers $D, d > 0$, and that

$$\frac{r}{m} = o(1), \quad c \to \infty. \tag{284}$$

To validate these assumptions, we combine (282), (283), (284) with (279) to obtain

$$\frac{A_{m-r}-\lambda}{2} = \frac{A_m-\lambda}{2} - \frac{A_m - A_{m-r}}{2} = \cos\left(\frac{\pi}{2r}\right)\cdot(1+o(1)), \quad c\to\infty. \tag{285}$$

We combine (276), (282) and (285) to obtain

$$\frac{m^a-(m-r)^a}{c^a} = \frac{\pi^2}{4r^2}\cdot(1+o(1)), \quad c\to\infty. \tag{286}$$

We combine (282), (284) and (286) to obtain

$$r\cdot m^{a-1} = O(1)\cdot\frac{c^a}{r^2}\cdot(1+o(1)), \quad c\to\infty, \tag{287}$$

and substitute (282), (283) into (287) to obtain

$$r = O(c^{1/3}), \quad c\to\infty. \tag{288}$$

Due to the combination of (282) and (288), the assumptions (283), (284) were consistent. We combine (283), (285), (288) to conclude that

$$\frac{2}{A_m - A_{m-r}} = \frac{8\cdot r^2}{\pi^2}\cdot(1+o(1)) = O\left(c^{2/3}\right), \quad c\to\infty, \tag{289}$$

which implies (280). Also, we combine (282), (283), (284) with (276), (279) to obtain

$$\begin{aligned} A_{m-r} - A_{m-r-1} &= a\cdot m^{a-1}\cdot c^{-a}\cdot(1+o(c)) \\ &= O\left(\frac{1}{c}\right), \quad c\to\infty. \end{aligned} \tag{290}$$

Finally, we combine (289) and (290) to obtain (281). ■

In the following theorem, we summarize Theorems 27, 28, 29.

Theorem 30. *Suppose that $a\geq 1$ and $c\geq 1$ are real numbers, and that $n>0$ is an integer such that*

$$n = O(c), \quad c\to\infty. \tag{291}$$

Suppose also that $A_1,\ldots,A_n$ are defined via

$$A_j = 2+\left(\frac{j}{c}\right)^a, \tag{292}$$

for every $j=1,\ldots,n$. Suppose also that the n by n symmetric tridiagonal matrix A is that of Definition 1, that λ is an eigenvalue of A, and that

$$\lambda > 4. \tag{293}$$

If

$$\lambda - A_1 > \cdots > \lambda - A_k \geq 2 > \lambda - A_{k+1} > \cdots > \lambda - A_{k+l+2} \geq 0, \tag{294}$$

and

$$\frac{\lambda - A_{k+l+1}}{2} > \cos\left(\frac{\pi}{2l-1}\right) \geq \frac{\lambda - A_{k+l+2}}{2} \tag{295}$$

for some integer $1 < k < n$ *and* $l \geq 3$*, then*

$$\frac{2}{A_{k+l} - A_k} = O\left(c^{2a/(a+2)}\right), \quad c \to \infty, \tag{296}$$

and also

$$\frac{A_{k+l+1} - A_{k+l}}{A_{k+l} - A_k} = O\left(c^{-1/3}\right), \quad c \to \infty. \tag{297}$$

If

$$0 > \lambda - A_{m-r} > \cdots > \lambda - A_m > -2 \geq \lambda - A_{m+1}, \tag{298}$$

and

$$\frac{A_{m-r} - \lambda}{2} > \cos\left(\frac{\pi}{2r-1}\right) \geq \frac{A_{m-r-1} - \lambda}{2} \tag{299}$$

for some integer $1 < m < n$ *and* $r \geq 3$*, then*

$$\frac{2}{A_m - A_{m-r}} = O\left(c^{2/3}\right), \quad c \to \infty, \tag{300}$$

and also

$$\frac{A_{m-r} - A_{m-r-1}}{A_m - A_{m-r}} = O\left(c^{-1/3}\right), \quad c \to \infty. \tag{301}$$

Remark 17. *While a more careful analysis allows to replace the asymptotic estimates* (296)*,* (297)*,* (300)*,* (301) *with explicit non-asymptotic inequalities, the proofs, albeit straightforward, are somewhat long, and will be published at a later date.*

5 Numerical Algorithms

In this section, we describe several numerical algorithms for the evaluation of the eigenvectors of certain symmetric tridiagonal matrices.

5.1 Problem Settings

In this subsection, we describe the problem we want to solve.

Assumptions. Suppose that $n > 0$ is an integer, that $0 < f_1 < f_2 < \dots$ is monotone sequence of positive real numbers, and that the numbers $A_1, \dots, A_n$ are defined via (65) of Definition 1. Suppose also that the n by n symmetric tridiagonal matrix A is defined via (64) of Definition 1, and that the real number λ is an eigenvalue of A.

Goal. Our goal is to evaluate the unit-length eigenvector

$$X = (X_1, \dots, X_n) \in \mathbb{R}^n \tag{302}$$

of A corresponding to λ, whose first coordinate is positive (i.e. $X_1 > 0$).

Observation. We observe that $\lambda > 0$ (see Theorem 9), and that $\lambda \leq 2 + A_n$ (see Theorem 10).

Equivalent formulations of the goal. Obviously, the problem of evaluating x given A and λ is equivalent to computing the solution of the linear system

$$B \cdot X = 0, \tag{303}$$

where $B = A - \lambda \cdot I$. Also, due to Theorem 7, this problem is equivalent to evaluating the solution to a certain three-terms recurrence relation (namely, the one specified via (66), (67), (68) of Theorem 7).

Desired numerical properties of the solution. We want our solution to possess the following property: each coordinate X_j of the vector X should be evaluated with high relative precision. In other words, suppose that $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_n) \in \mathbb{R}^n$ is our numerical approximation to X. For every $j = 1, \dots, n$, we want the absolute value of the quantity ε_j, defined via

$$\varepsilon_j = \frac{\tilde{X}_j - X_j}{X_j}, \tag{304}$$

to be small.

Observation. This task is potentially difficult if $|X_j|$ is small compared to $\|X\| = 1$. For example, if $|X_1| < \varepsilon$, where ε is the machine precision (e.g. $\varepsilon \approx 10^{-16}$ for double-precision calculations), it is not obvious why $|\varepsilon_1|$ should be less than 1, let alone be "small" (see also Section 1).

Relation between X_{j-1}, X_j and X_{j+1}. For every $j = 2, \dots, n-1$, the relation between three consecutive coordinates X_{j-1}, X_j and X_{j+1} of the vector X is expressed via (67) of Theorem 7; more specifically,

$$X_{j-1} + (A_j - \lambda) \cdot X_j + X_{j+1} = 0, \tag{305}$$

for every $j = 2, \dots, n-1$. It turns out that the qualitative behavior of X_{j-1}, X_j and X_{j+1} relative to each other depends on $\lambda - A_j$ in the following way. If $\lambda - A_j \geq 2$, then all the three coordinates have the same sign, and $|X_{j-1}| < |X_j| < |X_{j+1}|$ (see Theorem 11). If $\lambda - A_j \leq -2$, then the signs of X_{j-1}, X_{j+1} are opposite to the sign of X_j, and $|X_{j-1}| > |X_j| > |X_{j+1}|$ (see Theorem 13). Finally, if $-2 < \lambda - A_j < 2$, then the relation is somewhat more complicated (see, for example, Theorems 17, 21, 24, 25).

Assumption on λ. In the view of the latter observation, we will consider the case in which the coordinates of X exhibit all the behaviors described above (that is, in this sense, the most general case). This is achieved by making the following assumption on λ. Suppose that

$$1 \leq k < k+l+1 \leq p < p+1 \leq m-r < m < n \tag{306}$$

are integers. Suppose also that

$$\begin{aligned} \lambda - A_1 > \cdots > \lambda - A_k \geq & \quad 2 > \\ \lambda - A_{k+1} > \cdots > \lambda - A_{k+l+1} \geq \cdots \geq \lambda - A_p \geq & \quad 0 > \\ \lambda - A_{p+1} \geq \cdots \geq \lambda - A_{m-r} > \cdots > \lambda - A_m > & \quad -2 \geq \\ \lambda - A_{m+1} > \cdots > \lambda - A_n. & \end{aligned} \tag{307}$$

Suppose furthermore that

$$\frac{\lambda - A_{k+l+1}}{2} > \cos\left(\frac{\pi}{2l-1}\right) \geq \frac{\lambda - A_{k+l+2}}{2}, \tag{308}$$

and that

$$\frac{A_{m-r} - \lambda}{2} > \cos\left(\frac{\pi}{2r-1}\right) \geq \frac{A_{m-r-1} - \lambda}{2}. \tag{309}$$

Observation. We combine (307) with Definition 1 to conclude that

$$4 + f_1 < \lambda < f_n. \tag{310}$$

If the left-hand side inequality in (310) did not hold, then the region of growth would be empty. If the right-hand side inequality in (310) did not hold, then the region of decay would be empty. While the obvious simplification of the algorithm of the next section will handle such cases, we discuss only the general case for the sake of simplicity of presentation.

5.2 Informal Description of the Algorithm

This section contains an informal description of an algorithm for the evaluation of $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ (see (302)). In Section 5.3, the algorithm is described in a more precise and detailed way. Section 5.4 contains an outline of the steps of the algorithm.

For any $\lambda-$eigenvector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ of A and every $j = 2, \ldots, n-1$, the three consecutive coordinates x_{j-1}, x_j, x_{j+1} satisfy the identity (67) of Theorem 7 (see also (305) above). If $\lambda - A_j > 2$, we say that j is in the region of growth. If $\lambda - A_j < -2$, we say that j is in the region of decay. Otherwise, we say that j is in the oscillatory region (see Definition 2).

First, we evaluate the coordinates of x in the region of growth by setting the first coordinate to be equal to 1 and "going forward" (i.e., x_{j+1} is computed from x_j and x_{j-1} via (67)). We use the same recurrence relation to evaluate the first few coordinate in the leftmost part of the oscillatory region (when $\lambda - A_j$ is still close to 2). All of these coordinates

are positive, and this computation is numerically stable (see Sections 5.3.1, 5.3.2 below for more details).

Then, up to the middle of the oscillatory region (i.e. up to $j = p$ with $\lambda - A_p \approx 0$), we "complexify" the coordinates of x by relating to each pair (x_j, x_{j+1}) a complex number α_j, whose real part is $x_j/2$. These α_j's are evaluated as follows. The "leftmost" α_j is evaluated from the corresponding x_j, x_{j+1} via solving a certain two by two linear system; in this computation we lose several digits (see, for example, Corollary 10). However, the evaluation of each subsequent α_j from its predecessor is a stable procedure (see, for example, Corollary 9 and Section 5.3.3 for more details). Once we have all α_j's, each x_j in the left half of the oscillatory region is evaluated as $2 \cdot \Re(\alpha_j)$ (see Section 5.3.3 for more details).

At this point, we have evaluated the left coordinates $x_1, \ldots, x_p, x_{p+1}$ of the $\lambda-$eigenvector x of A, whose first coordinate is $x_1 = 1$.

Suppose now that $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n) \in \mathbb{R}^n$ is the $\lambda-$eigenvector $\hat{x}$ of A whose last coordinate is $\hat{x}_n = 1$. The evaluation of the right coordinates of $\hat{x}$ mirrors the evaluation of the left coordinates of x, with the main difference being that the direction is reversed. More specifically, we set $\hat{x}_n = 1$, and evaluate the coordinates of $\hat{x}$ in the region of decay via the recurrence relation (67) in the backwards direction (i.e. $\hat{x}_{j-1}$ is evaluated from $\hat{x}_j$ and $\hat{x}_{j+1}$). We use the same recurrence relation to evaluate the last few coordinate in the rightmost part of the oscillatory region (when $\lambda - A_j$ is still close to -2). All of these coordinates have alternating signs, and this computation is numerically stable (see Sections 5.3.4, 5.3.5 below for more details).

The evaluation of the rest of the coordinates of $\hat{x}$ in the right part of the oscillatory region (i.e. for $0 > \lambda - A_j$) is similar to the evaluation of the coordinates of x in the left part of the oscillatory region (see also Section 5.3.6 below for more details). By the end of this computation, we have the coordinates $\hat{x}_p, \hat{x}_{p+1}, \ldots, \hat{x}_n$ of $\hat{x}$.

Since the $\lambda-$eigenspace of A is one-dimensional (see Theorem 7), x and $\hat{x}$ must be, up to a scaling, the coordinates of the same eigenvector (in exact arithmetics). We use this observation to rescale $\hat{x}_p, \ldots, \hat{x}_n$ in such a way that $(x_p, x_{p+1}) = (\hat{x}_p, \hat{x}_{p+1})$. By doing so, we obtain all the coordinates of the $\lambda-$eigenvector x of A whose first coordinate is $x_1 = 1$. We normalize x (i.e. divide it by $\|x\|$) to obtain the unit-length eigenvector X whose first coordinate is positive (see Sections 5.3.7, 5.3.8 below for more details).

5.3 Detailed Description of the Algorithm

We evaluate $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ (see (302)) as follows.

5.3.1 The region of growth: $x_1, \ldots, x_{k+1}$

First, we set $x_1 = 1$, and evaluate $x_2, \ldots, x_{k+1}$ iteratively via (66), (67) of Theorem 7. These coordinates are all positive, and grow as the indices increase (see discussion in Section 5.1, and also Theorem 11 for more details). Also, this computation is stable numerically (see Theorem 12 for more details). Intuitively, it says that three-term recurrences are stable in the growing direction.

5.3.2 The leftmost part of the oscillatory region: $x_{k+2}, \ldots, x_{k+l-1}$

We observe that $0 < \lambda - A_{k+1} < 2$, due to (307). In other words, we have reached the oscillatory region (see Definition 2). In particular, the coordinates are not expected to have the same sign, the monotonicity is not guaranteed, and, moreover, some coordinates might be significantly smaller than the other (for an extreme example, see Theorem 4). In other words, the use of the recurrence relation (67) to evaluate the coordinates in the oscillatory region can potentially result in loss of accuracy.

On the other hand, the first few coordinates in the oscillatory regime do not change too fast, and are all positive (see Theorem 21 for a more precise statement). We compute the coordinates $x_{k+2}, \ldots, x_{k+l-1}$ iteratively via (67), moreover, this computation is numerically stable (see Theorem 22). We note that l is determined from (308); in particular, $\lambda - A_{k+l+1}$ is still fairly close to 2.

5.3.3 Up to the middle of the oscillatory region: $x_{k+l}, \ldots, x_{p+1}$

Up to the middle of the oscillatory region ($\lambda - A_p \approx 0$), the coordinates of x can vary in both sign and order of magnitude (see, for example, the eigenvectors of the matrix in Theorem 4). Therefore, if we still use (67) in this region to evaluate the coordinates one by one, "small" coordinates might be computed as a sum of two "larger" coordinates, resulting in a loss of accuracy; moreover, the stability properties of such computation are not obvious.

To bypass these obstacles, we complexify the coordinates in the following way. First, we compute the complex number α_{k+l-2} from x_{k+l-2}, x_{k+l-1} via the two by two linear system (147) of Theorem 17 (the solution is obtained via (137) of Theorem 15). The condition matrix of the corresponding two by two linear system is roughly l^2, where l is defined via (308) (see Theorem 16, (130) of Theorem 14, and (308)). In other words, in the evaluation of α_{k+l-2} from x_{k+l-2} and x_{k+l-1} we expect to lose $2 \cdot \log_{10}(l)$ decimal digits (see e.g. Corollary 10; also, for an example for this condition number for a certain class of matrices A, see (296) of Theorem 30).

We then evaluate $\alpha_{k+l-1}, \ldots, \alpha_{p-1}, \alpha_p$ iteratively via (150) of Theorem 17. In other words, each α_{j+1} is obtained from α_j via rotation by the angle θ_j and adding a pure imaginary correction term. This computation is stable provided that the correction term is small compared to $|\alpha_j|$ (see Corollary 4, Theorem 26 and Corollary 9). In Section 4.3, we demonstrate that, for a certain class of matrices, the correction term is indeed small, and thus all $\alpha_{k+l-2}, \ldots, \alpha_p$ are evaluated with the same relative accuracy (see Theorem 26, Remark 15 and Theorem 30).

Next, we evaluate $x_{k+l}, \ldots, x_{p-1}, x_p$ via (148) of Theorem 17. In other words, x_j is the real part of $\alpha_j/2$, for each $j = k+l, \ldots, p$. In particular, each x_j is computed with the same absolute accuracy as α_j. On the other hand, each x_j is expected be evaluated to roughly $\log_{10}(|\alpha_j/x_j|)$ correct decimal digits less than α_j.

Finally, we evaluate x_{p+1} from α_p via (149) of Theorem 17.

5.3.4 The region of decay: $\hat{x}_m, \ldots, \hat{x}_n$

The rest of the coordinates of x are computed in the reverse direction, i.e. we iterate backwards from n to p. We will denote these coordinates by $\hat{x}_p, \hat{x}_{p+1}, \ldots, \hat{x}_n$, to distinguish

them from $x_1, \ldots, x_{p+1}$, whose evaluation is described in Sections 5.3.1, 5.3.2, 5.3.3.

First, we set $\hat{x}_n = 1$, and evaluate $\hat{x}_{n-1}, \ldots, \hat{x}_m$ iteratively, in the backward direction, via (68), (67) of Theorem 7. All $\hat{x}_m, \ldots, \hat{x}_n$ have alternating signs, and their absolute values decay as the indices increase (see and Theorem 13 for more details). Also, this computation is stable numerically (see Theorems 13, 12 for more details). Intuitively, it says that three-term recurrences are stable in the growing direction.

We note that the evaluation of $\hat{x}_m, \ldots, \hat{x}_n$ mirrors the evaluation of $x_1, \ldots, x_{k+1}$, described in Section 5.3.1 above. We also observe that, depending on the parity of $n - m$, the coordinate $\hat{x}_m$ can be either positive or negative. If $\hat{x}_m < 0$ (that is, $n - m$ is odd), we flip the signs of all $\hat{x}_m, \ldots, \hat{x}_n$. Obviously, the recurrence relation (67), (68) is still satisfied by $\hat{x}_m, \ldots, \hat{x}_n$; on the other hand, we have made sure that $\hat{x}_m$ is positive.

5.3.5 The rightmost part of the oscillatory region: $\hat{x}_{m-(r-2)}, \ldots, \hat{x}_{m-1}$

As we enter the oscillatory region from the right, the magnitudes of the first few $\hat{x}_j$'s do not change too fast, and they all have alternating signs (see Theorem 25 for a more precise statement). We compute the coordinates $\hat{x}_{m-(r-2)}, \ldots, \hat{x}_{m-1}$ iteratively via (67) of Theorem 7, in the backward direction. Moreover, this computation is numerically stable (see Theorem 22). We note that r is determined from (309); in particular, $A_{m-r} - \lambda$ is still fairly close to 2.

We note that the evaluation of $\hat{x}_{m-(r-2)}, \ldots, \hat{x}_{m-1}$ mirrors the evaluation of $x_{k+2}, \ldots, x_{k+l-1}$, described in Section 5.3.2 above.

5.3.6 Down to the middle of the oscillatory region: $\hat{x}_p, \ldots, \hat{x}_{m-(r-1)}$

The evaluation of $\hat{x}_p, \ldots, \hat{x}_{m-(r-1)}$ mirrors the evaluation of $x_{k+l+1}, \ldots, x_{p+1}$, described in Section 5.3.3 above. In other words, we complexify $\hat{x}_j$'s by introducing γ_j's (the real part of each γ_j is twice $\hat{x}_j$), and evaluate γ_j's iteratively, in the backward direction.

More specifically, we first compute $\gamma_{m-(r-1)}$ from $\hat{x}_{m-(r-2)}, \hat{x}_{m-(r-3)}$ by solving the two by two linear system (218) of Theorem 24 (the solution is obtained via (137) of Theorem 15). The condition matrix of the corresponding two by two linear system is roughly r^2, where r is defined via (309) (see Theorem 16, (213) of Theorem 23, and (309)). In other words, in the evaluation of $\gamma_{m-(r-1)}$ from $\hat{x}_{m-(r-2)}$ and $\hat{x}_{m-(r-3)}$ we expect to lose $2 \cdot \log_{10}(r)$ decimal digits (see e.g. Corollary 10; also, for an example for this condition number for a certain class of matrices A, see (296) of Theorem 30).

Next, we evaluate $\gamma_{m-r}, \ldots, \gamma_p, \gamma_{p-1}$ iteratively, in the backwards direction, via (221) of Theorem 24. This computation is similar to the evaluation of $\alpha_{k+l-1}, \ldots, \alpha_{p-1}, \alpha_p$, described in Section 5.3.3, and has similar properties in terms of accuracy and stability (see also Theorem 26, Corollary 9, Remark 15, and Theorem 30).

Then, we evaluate $\hat{x}_{m-(r-1)}, \hat{x}_{m-r}, \ldots, \hat{x}_{p+1}$ from $\gamma_{m-(r+1)}, \ldots, \gamma_{p-1}$, respectively, via (219) of Theorem 24.

Finally, we evaluate $\hat{x}_p$ from γ_{p-1} via (220) of Theorem 24.

5.3.7 Gluing $x_1, \ldots, x_p, x_{p+1}$ and $\hat{x}_p, \hat{x}_{p+1}, \ldots, \hat{x}_n$ together

We have evaluated the first $p+1$ coordinates of the $\lambda-$eigenvector x of A whose first coordinate is $x_1 = 1$. Also, we have evaluated the last $n-p+1$ coordinates of the $\lambda-$eigenvector $\hat{x}$ of A whose last coordinate is $\hat{x}_n = (-1)^{n-m}$. However, due to Theorem 7, the $\lambda-$eigenspace of A is one-dimensional. We use this observation to conclude that, up to scaling, $\hat{x}_p, \ldots, \hat{x}_n$ would be, in exact arithmetics, the last $n-p+1$ coordinates of the eigenvector x. In other words, in exact arithmetics, if the real number s is defined via

$$s = \begin{cases} x_p/\hat{x}_p, & \text{if } |\hat{x}_p| > |\hat{x}_{p+1}| \neq 0, \\ x_{p+1}/\hat{x}_{p+1}, & \text{otherwise ,} \end{cases} \tag{311}$$

then the vector

$$(x_1, \ldots, x_p, x_{p+1}, s \cdot \hat{x}_{p+2}, \ldots, s \cdot \hat{x}_n) \tag{312}$$

is the $\lambda-$eigenvector of A whose first coordinate is $x_1 = 1$. We observe that $s \neq 0$, since $\hat{x}_p^2 + \hat{x}_{p+1}^2$ is strictly positive, due to Theorem 7. Thus, we evaluate $x_{p+2}, \ldots, x_n$ via the formula

$$x_j = s \cdot \hat{x}_j, \tag{313}$$

for every $j = p+2, \ldots, n$, where s is defined via (311).

We observe that s is evaluated from $x_p, x_{p+1}, \hat{x}_p, \hat{x}_{p+1}$, and each individual x_j in the oscillatory region might be evaluated less accurately than α_j (see Section 5.3.3). Nevertheless, we observe that $\theta_p \approx \pi/2$ and $\varphi_{p-1} \approx \pi/2$, due to the combination of Theorems 14, 23 with (307). We combine this observation with (148), (149) of Theorem 17, (219), (220) of Theorem 24, Corollary 5 and Corollary 8 to conclude that

$$\begin{aligned} &|x_p^2| + |x_{p+1}^2| \approx 4|\alpha_p|^2, \\ &|\hat{x}_p|^2 + |\hat{x}_{p+1}|^2 \approx 4|\gamma_{p-1}^2|. \end{aligned} \tag{314}$$

It follows from the combination of (311) and (314) that s is evalauted to roughly the same relative accuracy as α_p/γ_{p-1}.

5.3.8 Normalization

We define the real number d via the formula

$$d = \sqrt{x_1^2 + \cdots + x_n^2}. \tag{315}$$

In exact arithmetics, the vector

$$X = (X_1, \ldots, X_n) = \left(\frac{x_1}{d}, \ldots, \frac{x_n}{d}\right) \tag{316}$$

is the unit-length $\lambda-$eigenvector of A, whose first coordinate is positive.

In Section 5.3.3, we mentioned that, in the oscillatory region, the coordinates x_j might have a significantly lower relative accuracy than the corresponding α_j. Motivated by this

observation, we evaluate d without using the coordinates x_j in the oscillatory region explicitly, in the following way.

First, we evaluate the real numbers d_l via the formula

$$\begin{aligned} d_l &= 4 \sum_{j=k+l-1}^{p} \left(|\alpha_j|^2 + \cos(\theta_j) \cdot \Re \left(\alpha_j^2 e^{i\theta_j} \right) \right) \\ &= 4 \sum_{j=k+l-1}^{p} |\alpha_j|^2 \cdot (1 + \cos(\theta_j) \cdot \cos\left(\theta_j + 2\mathrm{Arg}(\alpha_j)\right)), \end{aligned} \tag{317}$$

where $\mathrm{Arg}(z) = \xi$ for any complex number $z = R \cdot e^{i\xi}$. Due to Corollary 5,

$$d_l = |x_{k+l-1}|^2 + |x_{p+1}|^2 + 2 \sum_{j=k+l}^{p} |x_j|^2. \tag{318}$$

Also, we evaluate the real number d_r via the formula

$$\begin{aligned} d_r &= 4 \sum_{j=p}^{m-(r-1)} \left(|\gamma_j|^2 + \cos(\varphi_j) \cdot \Re \left(\gamma_j^2 e^{i\varphi_j} \right) \right) \\ &= 4 \sum_{j=p}^{m-(r-1)} |\gamma_j|^2 \cdot (1 + \cos(\varphi_j) \cdot \cos\left(\varphi_j + 2\mathrm{Arg}(\gamma_j)\right)). \end{aligned} \tag{319}$$

Due to Corollary 8,

$$d_r = |\hat{x}_{p+1}|^2 + |\hat{x}_{m-(r-3)}|^2 + 2 \sum_{j=k+l}^{m-(r-2)} |\hat{x}_j|^2. \tag{320}$$

We combine (311), (318), (320) to obtain

$$d_l + s \cdot d_r = |x_{k+l-1}|^2 + |x_{m-(r-3)}|^2 + +2 \sum_{j=k+l}^{m-(r-3)} |x_j|^2. \tag{321}$$

Therefore, in exact arithmetics,

$$\begin{aligned} &d^2 = \\ &\frac{|x_{k+l-1}|^2 + |x_{m-(r-3)}|^2 + d_l + s \cdot d_r}{2} + \sum_{j=1}^{k+l-2} |x_j|^2 + \sum_{j=m-r+4}^{n} |x_j|^2. \end{aligned} \tag{322}$$

We compute d by evaluating the right-hand side of (322) numerically.

The relative accuracy to which d is evaluated will be dominated by the relative accuracy of the summands in (317), (319) (each one of these summands is obviously positive).

Finally, we obtain the unit-length eigenvector X by normalizing the eigenvector x. This is done by dividing every coordinate of x by d (see (316)).

5.4 Short Description of the Algorithm

Step A: evaluation of the left coordinates.

1. Set $x_1 = 1$.
2. Compute x_2 via (66) of Theorem 7.
3. Compute $x_3, \ldots, x_k, x_{k+1}$ iteratively via (67) of Theorem 7.
4. Compute $x_{k+2}, \ldots, x_{k+l-2}, x_{k+l-1}$ iteratively via (67) of Theorem 7.
5. Compute α_{k+l-2} from x_{k+l-2}, x_{k+l-1} via (147) of Theorem 17.
6. Compute $\alpha_{k+l-1}, \ldots, \alpha_{p-1}, \alpha_p$ iteratively via (150) of Theorem 17.
7. Compute $x_{k+l}, \ldots, x_{p-1}, x_p$ via (148) of Theorem 17.
8. Compute x_{p+1} from α_p via (149) of Theorem 17.

Step B: evaluation of the right coordinates.

1. Set $\hat{x}_n = 1$.
2. Compute $\hat{x}_{n-1}$ via (68) of Theorem 7.
3. Compute $\hat{x}_{n-2}, \ldots, \hat{x}_{m+1}, \hat{x}_m$ iteratively via (67) of Theorem 7.
If $\hat{x}_m < 0$, flip the signs of all $\hat{x}_m, \ldots, \hat{x}_n$.
4. Compute $\hat{x}_{m-1}, \ldots, \hat{x}_{m-(r-2)}$ iteratively via (67) of Theorem 7.
5. Compute $\gamma_{m-(r-1)}$ from $\hat{x}_{m-(r-2)}, \hat{x}_{m-(r-3)}$ via (218) of Theorem 24.
6. Compute $\gamma_{m-r}, \ldots, \gamma_p, \gamma_{p-1}$ iteratively via (221) of Theorem 24.
7. Compute $\hat{x}_{m-(r-1)}, \hat{x}_{m-r}, \ldots, \hat{x}_{p+1}$ via (219) of Theorem 24.
8. Compute $\hat{x}_p$ from γ_{p-1} via (220) of Theorem 24.

Step C: glue them together.

1. If $x_p \cdot \hat{x}_p + x_{p+1} \cdot \hat{x}_{p+1} < 0$, flip the signs of all $\hat{x}_p, \ldots, \hat{x}_n$.
2. Set $s = x_p/\hat{x}_p$.
3. If $|x_{p+1}| > |x_p|$, set $s = x_{p+1}/\hat{x}_{p+1}$.
4. For every $j = p+2, \ldots, n$, compute x_j via the formula $x_j = s \cdot \hat{x}_j$.
5. Compute $d = \sqrt{x_1^2 + \cdots + x_n^2}$, via evaluating the right-hand side of (322).
6. For every $j = 1, \ldots, n$, compute X_j via $X_j = x_j/d$.

5.5 Error Analysis

In Sections 5.3, 5.4, we described an algorithm for the evaluation of the unit length $\lambda-$eigenvector $X = (X_1, \ldots, X_n)$ of A, whose first coordinate is positive. In this section, we analyze the relative accuracy to which each coordinate is evaluated. In this analysis, we use the following notation. Suppose that y is a quantity to be evaluated, and $eval(y)$ is a numerical approximation to this quantity. We define the absolute error $abs(y)$ via the formula

$$abs(y) = |y - eval(y)|. \tag{323}$$

Also, we define the relative error $rel(y)$ via the formula

$$rel(y) = \left| \frac{y - eval(y)}{y} \right|. \tag{324}$$

Loosely speaking, $-\log_{10}(rel(y))$ is the number of decimal digits in which y and $eval(y)$ coincide.

Our analysis is based on the following observations.

Observation 1. All of $x_1, \ldots, x_k, x_{k+1}$ have roughly the same relative accuracy (see Section 5.3.1). This is despite the fact that x_1 can be significantly smaller than x_k (see Theorem 11).

Observation 2. Also, all of $x_k, x_{k+1}, \ldots, x_{k+l-1}$ have roughly the same relative accuracy (see Section 5.3.2).

Observation 3. All of $x_m, x_{m+1}, \ldots, x_n$ have roughly the same relative accuracy (see Section 5.3.4).

Observation 4. Also, all of $x_{m-r+2}, \ldots, x_m, x_{m+1}$ have roughly the same relative accuracy (see Section 5.3.5).

Observation 5. The relative accuracy to which α_{k+l-2} is evaluated is roughly

$$rel(\alpha_{k+l-2}) \approx rel(x_{k+l-2}) \cdot \frac{4}{A_{k+l} - A_k} \approx rel(x_{k+l-2}) \cdot l^2 \tag{325}$$

(see Section 5.3.3, and also Theorems 16, 26, Corollary 10).

Observation 6. The relative accuracy to which $\alpha_{k+l-1}, \alpha_{k+l}, \ldots, \alpha_p$ are evaluated is roughly the same, provided that, for example,

$$\frac{A_{j+2} - A_{j+1}}{A_{j+1} - A_{k+1}} < \frac{1}{2}, \tag{326}$$

for every $j = k+l-1, \ldots, p$ (see Section 5.3.3, and also Theorems 17, 26 and Corollaries 4, 9). In other words,

$$rel(\alpha_j) \approx rel(\alpha_{k+l-2}), \tag{327}$$

for every $j = k+l-1, \ldots, p$, provided that (326) holds.

Observation 7. The relative accuracy to which $\gamma_{m-(r-1)}$ is evaluated is roughly

$$rel(\gamma_{m-(r-1)}) \approx rel(\hat{x}_{m-(r-3)}) \cdot \frac{4}{A_{m-(r-1)} - A_m} \approx rel(\hat{x}_{m-(r-3)}) \cdot r^2 \tag{328}$$

(see Section 5.3.6, and also Theorems 16, 26, Corollary 10).

Observation 8. The relative accuracy to which $\gamma_{p-1}, \gamma_p, \ldots, \gamma_{m-r}$ are evaluated is roughly the same, provided that, for example,

$$\frac{A_{j+2} - A_{j+1}}{A_m - A_{j+1}} < \frac{1}{2}, \tag{329}$$

for every $j = p-1, \ldots, m-(r-1)$ (see Section 5.3.6, and also Theorems 24, 26 and Corollaries 4, 9). In other words,

$$rel(\gamma_j) \approx rel(\gamma_{m-r}), \tag{330}$$

for every $j = p-1, \ldots, m-r$, provided that (329) holds.

Observation 9. The scaling parameter s, defined via (311) in Section 5.3.7, is evaluated with relative accuracy

$$rel(s) \approx rel(\alpha_p) + rel(\gamma_{p-1}) \tag{331}$$

(see, for example, (311), (314) in Section 5.3.7).

Observation 10. The relative error of each summand in (317) is roughly bounded by

$$rel\left(|\alpha_j|^2 \cdot (\cos(\theta_j) \cdot \cos\left(2 \cdot \mathrm{Arg}(\alpha_j) + \theta_j\right) + 1)\right) \lesssim \frac{rel(\alpha_j)}{1 - \cos(\theta_j)}, \tag{332}$$

for every $j = k + l - 2, \ldots, p$. We observe that, for any positive numbers $p_1, \ldots, p_n$ and real numbers $\varepsilon_1, \ldots, \varepsilon_n \in (-1, 1)$,

$$\left| \frac{p_1 \cdot (1 + \varepsilon_1) + \cdots + p_n \cdot (1 + \varepsilon_n)}{p_1 + \cdots + p_n} - 1 \right| \leq \max\left\{|\varepsilon_1|, \ldots, |\varepsilon_n|\right\}, \tag{333}$$

and combine (333), (317), (325), (327) and (332) to conclude that the relative error of d_l is roughly bounded by

$$rel(d_l) \lesssim rel(\alpha_{k+l-2}) \cdot \frac{4}{A_{k+l-2} - A_k} \approx rel(x_{k+l-2}) \cdot l^4. \tag{334}$$

Observation 11. By the same token as in Observation 10, we conclude that the relative error of d_r, defined via (319), is roughly bounded by

$$rel(d_r) \lesssim \frac{rel(\gamma_{m-(r-1)})}{1 - \cos(\varphi_{m-(r-1)})} \approx rel(\hat{x}_{m-(r-2)}) \cdot r^4. \tag{335}$$

Observation 12. Suppose that ε is the relative accuracy to which the quantities $\lambda - A_1, \ldots, \lambda - A_n$ are given. In other words, suppose that

$$rel(\lambda - A_j) \approx \varepsilon, \tag{336}$$

for every $j = 1, \ldots, n$. (For example, if $\lambda - A_j$ are given to full precision, $\varepsilon \approx 10^{-16}$ for double precision calculations and $\varepsilon \approx 10^{-35}$ for extended precision calculations.) We combine (322), Observations 1-4 above, (334), (335) and (336) to conclude that the real number d, evaluated via the right-hand side of (322), has relative accuracy roughly

$$rel(d) \lesssim \varepsilon \cdot \left(l^4 + r^4\right). \tag{337}$$

We make the following conclusions from Observations 1-12.

Conclusion 1. Suppose that the real number $\varepsilon > 0$ is that of (336), and the integers k, l, p, r, m, n are those of (307), (308), (309). We combine (316), (322), Observations 1-4 and (337) to conclude that

$$rel(X_j) \leq \varepsilon \cdot \left(l^4 + r^4\right), \tag{338}$$

for every $j = 1, \ldots, k + l - 1$ and every $j = m - r + 2, \ldots, n$. In other word, the coordinates in the region of growth and the coordinates in the region of decay are evaluated to the

relative accuracy, which is larger than the machine accuracy by only a factor of at most $l^4 + r^4$.

Conclusion 2. We combine (325), (327), (337), (148) of Theorem 17 and (316) to obtain

$$abs(X_j) \leq \varepsilon \cdot |\alpha_j| \cdot (rel(\alpha_j) + rel(d)) \lesssim \varepsilon \cdot |\alpha_j| \cdot \left(l^4 + r^4\right), \tag{339}$$

for every $j = k + l, \ldots, p$. We combine (148) with (339) to obtain

$$rel(X_j) \lesssim \varepsilon \cdot \frac{|\alpha_j|}{|\Re(\alpha_j)|} \cdot \left(l^4 + r^4\right), \tag{340}$$

for every $j = k + l, \ldots, p$.

Conclusion 3. Also, We combine (328), (330), (337), (219) of Theorem 24 and (316) to obtain

$$abs(X_j) \leq \varepsilon \cdot |\gamma_{j-2}| \cdot (rel(\gamma_{j-2}) + rel(d)) \lesssim \varepsilon \cdot |\gamma_{j-2}| \cdot \left(l^4 + r^4\right), \tag{341}$$

for every $j = p + 1, \ldots, r - m + 1$. We combine (219) with (341) to obtain

$$rel(X_j) \lesssim \varepsilon \cdot \frac{|\gamma_{j-2}|}{|\Re(\gamma_{j-2})|} \cdot \left(l^4 + r^4\right), \tag{342}$$

for every $j = p + 1, \ldots, r - m + 1$.

Example. Suppose that $a, c \geq 1$ are real numbers, and that the n by n matrix A is that of Theorem 30. In other words, the diagonal entries of A are defined via (292). Suppose also that $\lambda > 4$, and that (307) holds. We combine (338) with Theorem 30 to obtain

$$rel(X_j) \leq \varepsilon \cdot O\left(c^{4a/(a+2)}\right), \quad c \to \infty. \tag{343}$$

for every $j = 1, \ldots, k + l - 1$ and every $j = m - r + 2, \ldots, n$. In other words, if, for example, $a = 2$, then the first several coordinates $X_1, X_2, \ldots$ of the eigenvector are expected to be evaluated correctly to all but the last $2 \cdot \log_{10}(c)$ digits (see also Theorem 1 in Section 1).

Remark 18. *Extensive numerical experiments seem to indicate that the estimates* (338)*,* (339)*,* (340)*,* (341)*,* (342) *are somewhat pessimistic. In other words, in practice the relative error tends to be smaller than our estimates suggest (see also Section 7).*

Remark 19. *It is somewhat surprising that, according to* (338)*, the relative error of, say,* X_1 *seems to be independent of the order of magnitude of* X_1*. In particular, while* X_1 *can be fairly small (see e.g. Theorem 11 and Corollary 3), it still will be evaluated to reasonable relative precision.*

5.6 Related Algorithms

In Section 5.3, 5.4, we presented an algorithm to solve the problem, described in Section 5.1. In this section, we describe several related algorithms to solve the same problem, and briefly discuss their performance.

5.6.1 A Simplified Algorithm

The following algorithm is based on Theorem 7, and avoids complexification of the coordinates of the eigenvector (see Sections 5.3.3, 5.3.6).

First, we set $x_1 = 1$, and evaluate $x_2, \ldots, x_p, x_{p+1}$ iteratively via (66), (67) of Theorem 7. In other words, we compute all the coordinates in the region of growth and up to the middle of the oscillatory region via the three-term recurrence relation (67), in the forward direction (see Section 5.3.1).

Then, we set $\hat{x}_n = 1$, and evaluate $\hat{x}_{n-1}, \hat{x}_{n-2}, \ldots, \hat{x}_{p+1}, \hat{x}_p$ via (68), (67) of Theorem 7, in the backward direction. In other words, we compute all the coordinates in the region of decay and down to the middle of the oscillatory region via the three-term recurrence relation (67), in the backward direction (see Section 5.3.4).

Next, we glue $x_1, \ldots, x_p, x_{p+1}$ with $\hat{x}_p, \hat{x}_{p+1}, \ldots, \hat{x}_n$ via the procedure, described in Section 5.3.7. In other words, we defined the real number s via (311), and defined x_j via (313), for every $j = p+2, \ldots, n$. As in Section 5.3.7, we observe that (at least in exact arithmetics), the vector

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n \tag{344}$$

is the $\lambda-$eigenvector of A, whose first coordinate is $x_1 = 1$.

Finally, we evaluate the unit-length eigenvector $X = (X_1, \ldots, X_n)$ via the formula

$$(X_1, \ldots, X_n) = \frac{1}{\sqrt{x_1^2 + \ldots x_n^2}} \cdot (x_1, \ldots, x_n)\,. \tag{345}$$

In other words, the normalization factor is computed by summing x_j's directly, as opposed to using (322) of Section 5.3.8.

Remark 20. *Up to our best knowledge, the error analysis (similar to that of Section 5.5) for the simplified algorithm, described in this section, is not available in the literature. Nevertheless, extensive numerical experiments seem to indicate that, in practice, the accuracy of the simplified algorithm is similar to that described in Sections 5.3, 5.4 (see also Section 7).*

Remark 21. *When the coordinates of the eigenvector are evaluated via the three-terms recurrence* (67)*, the choice of direction plays a crucial role. Roughly speaking, this recurrence is unstable in the backward direction in the region of growth, and is unstable in the forward direction in the region of decay (see also Sections 5.3.1, 5.3.4). The use of this recurrence relation in the "wrong" direction leads to a disastrous loss of accuracy.*

5.6.2 Inverse Power

The unit-length $\lambda-$eigenvector X of A can be obtained via Inverse Power Method (see Section 3.4.1 for more details). This method is iterative, and, on each iteration, the approximation $x^{(k+1)}$ of X is obtained from $x^{(k)}$ via solving the linear system

$$(\lambda \cdot I - A) \cdot x^{(k+1)} = x^{(k)}, \tag{346}$$

and normalizing the solution (i.e. we divide it by its norm). We observe that this method also evaluates λ, but we have assumed that λ has already been evaluated, by Inverse Power Method, Sturm Bisection (see Section 3.4.2), or some other means. On each iteration, we solve the linear system (346) by Gaussian elimination (we recall that A is tridiagonal, and thus one iteration costs $O(n)$ operations; see Remark 8).

The following conjecture about the accuracy of Inverse Power Method is substantiated by extensive numerical experiments (see Section 7).

Conjecture 1. *Suppose that $\varepsilon > 0$ is the machine precision (e.g. $\varepsilon \approx 10^{-16}$ for double-precision calculations), and that the eigenvalue λ of A is given up to ε, i.e.*

$$rel(\lambda) \approx \varepsilon, \tag{347}$$

where rel is the relative error, defined via (324) *in Section 5.5. Suppose also that $\lambda - A_1 > 2$ (see* (307)*). Suppose furthermore that $K > 0$ is an integer, and that*

$$K > \frac{\log(|X_1|)}{\log(\varepsilon)} + 1, \tag{348}$$

where $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is the unit-length $\lambda-$eigenvector of A. Then, after K iterations of Inverse Power Method, X_1 is evaluated with high relative accuracy. More specifically, this relative accuracy is of the same order of magnitude as for the algorithms, described in Sections 5.2, 5.3, 5.4, 5.6.1.

Remark 22. *The statement of Conjecture 1 also holds for all the coordinates of X in the region of growth and the region of decay. In other words, suppose that k, m are those of* (306), (307)*. Then, each of $X_1, \ldots, X_{k+1}$ and $X_m, \ldots, X_n$ is evaluated with high relative accuracy by K iterations of Inverse Power Method.*

Remark 23. *The inequality* (348) *reflects on the fact that each iteration of Inverse Power Method can reduce the coordinates of the approximation $x^{(k)}$ by the factor at most ε^{-1}. In other words, if $X_1 \approx 10^{-50}$, and, in the initial approximation, $x_1^{(1)} = O(1)$, then $x_1^{(4)}$ will already be of the same order of magnitude as X_1, and $x^{(5)}$ will approximate X_1 to a high relative precision.*

5.6.3 Jacobi Rotations

In the view of Sections 5.6.1, 5.6.2, one might suspect that virtually any standard algorithm would accurately solve the problem, introduced in Section 5.1. In other words, one might suspect that the small coordinates of X in the region of growth and the region of decay will be evaluated with high relative accuracy by any reasonable algorithm that computes eigenvectors.

Unfortunately, this is emphatically not the case, and the accuracy of the result strongly depends on the choice of the algorithm. Consider, for example, the popular Jacobi Rotations algorithm for the evaluation of the eigenvalues and eigenvectors of A (see, for example, [3], [6], [19], [20]). This algorithm is known for its simplicity and stability, and, indeed, it

typically evaluates all the eigenvalues of A fairly accurately. Moreover, the corresponding eigenvectors are evaluated to high relative accuracy, in the sense that

$$\frac{\|X - eval(X)\|}{\|X\|} = \|X - eval(X)\| \approx \varepsilon, \tag{349}$$

where X is the unit-length eigenvector (see (302)), $eval(X)$ is its numerical approximation produced by Jacobi Rotations, and ε is the machine precision. However, the *coordinates* of X are typically evaluated only to high *absolute* accuracy, e.g.

$$abs(X_1) \approx \varepsilon, \tag{350}$$

where *abs* is defined via (323). On the other hands, the relative accuracy of small coordinates will typically be poor. In particular, if, for example, $X_1 \approx 10^{-50}$, its numerical approximation, produced by Jacobi Rotations, will usually have no correct digits at all!

These observations about Jacobi Rotations applied to the problem of Section 5.1 are supported by extensive numerical evidence.

5.6.4 Gaussian Elimination

Another possible method to evaluate X would be to solve the linear system

$$(\lambda \cdot I - A) \cdot X = 0, \tag{351}$$

by means of Gaussian Elimination (see, for example, [3], [6], [19], [20]). Unfortunately, this method, in general, fails to evaluate the small coordinates of X with high relative accuracy (see, however, Section 5.6.2, where Gaussian Elimination is used several times, as a step of Inverse Power Method).

6 Applications

In this section, we will describe some applications of the algorithm, introduced in Section 5, to other computational problems.

6.1 Bessel Functions

The Bessel functions of the first kind $J_0, J_{\pm 1}, J_{\pm 2}, \ldots$, are defined via (28), (29) in Section 3.2. A numerical algorithm for the evaluation of $J_0(x), \ldots, J_n(x)$ for a given real number $x > 0$ and a given integer $n > 0$ is described in Section 3.4.3. In particular, according to Remark 9 in Section 3.4.3, the values $J_0(x), \ldots J_n(x)$ can be obtained as coordinates of the unit length $\lambda-$eigenvector of a certain symmetric tridiagonal matrix $A(x)$ (see (60), (61), (62), (63)).

The matrix $A(x)$ belongs to the class of matrices, introduced in Definition 1. More specifically, the diagonal entries of $A(x)$ are those of the matrix A of Theorem 30, with $a = 1$ and $c = x/2$ (see (292)).

In other words, the principal algorithm of this paper (see Sections 5.2, 5.3, 5.4) can be used to evaluate the Bessel functions $J_0, \ldots, J_n$ at a given point. Even more so, the

accuracy of this evaluation is analyzed in Theorem 30 and Section 5.5. Finally, in this case, the simplified algorithm (see Section 5.6.1) coincides with the well-known algorithm, described in Section 3.4.3.

See Section 7 for the related numerical experiments.

6.2 Prolate Spheroidal Wave Functions

For any real number $c > 0$, the prolate spheroidal wave functions (PSWFs) of band limit c are defined in Section 3.3. The numerical algorithm for the evaluation of PSWFs, briefly described in Section 3.3, is based on computing the unit-length eigenvectors of certain symmetric tridiagonal matrices (namely, truncated versions of the matrices $A^{c,even}$, $A^{c,odd}$, defined, respectively, via (41), (42) in Section 3.3).

Strictly speaking, these matrices do not belong to the class of matrices, introduced in Definition 1, since their non-zero off-diagonal entries are not equal to one (see (40)). Nevertheless, in the notation of (40),

$$A^{(c)}_{k,k+2} = A^{(c)}_{k+2,k} = \frac{c^2}{4} \cdot \left(1 + O\left(\frac{1}{k^2}\right)\right), \tag{352}$$

for every $k = 0, 1, 2, \ldots$, and

$$A^{(c)}_{k,k} = \frac{c^2}{4} \cdot \left(2 + \left(\frac{2k}{c}\right)^2 \cdot \left(1 + O\left(\frac{1}{k}, \frac{c^2}{k^4}\right)\right)\right), \tag{353}$$

for every $k = 0, 1, 2, \ldots$. In other words, for example, the matrix $A^{c,odd}$ can be viewed as a small perturbation of the symmetric tridiagonal matrix A, defined via the formula

$$A = \frac{c^2}{4} \cdot \begin{pmatrix} A_0 & 1 & & & \\ 1 & A_1 & 1 & & \\ & 1 & A_2 & 1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{354}$$

where $A_0, A_1, \ldots$ are defined via the formula

$$A_j = 2 + \left(\frac{4j}{c}\right)^2, \tag{355}$$

for every $j = 0, 1, \ldots$.

In particular, the algorithm of Sections 5.2, 5.3, 5.4, with obvious minor modifications, is applicable to the task of evaluating PSWFs numerically. Moreover, the error analysis of such evaluation has been carried out in Theorem 30 and Section 5.5.

See Section 7 for numerical examples, involving the matrices similar to (354). The results of some numerical experiments, where slightly modified versions of the algorithms of this paper (see Section 5) are used to evaluate PSWFs, see, for example, [14], [15].

7 Numerical Results

In this section, we illustrate the analysis of Section 4 via several numerical experiments. All the calculations were implemented in FORTRAN (the Lahey 95 LINUX version), and were carried out in double precision. Occasionally, extended precision calculations were used to estimate the accuracy of double precision calculations.

c	100	1,000	10,000	100,000
n	250	2,100	20,215	200,500
λ	0.51665E+01	0.41168E+01	0.40117E+01	0.40012E+01
X_1	0.37636E-39	0.29308E-41	0.23304E-42	0.37902E-43
$rel(X_1)$	0.16471E-13	0.36528E-13	0.64311E-11	0.12414E-09
$rel(Y_1)$	0.10619E-13	0.58053E-13	0.67080E-12	0.94020E-11
$\max(abs(X))$	0.23315E-14	0.12386E-14	0.27500E-12	0.27399E-11
$\max(abs(Y))$	0.58981E-15	0.13891E-14	0.94943E-14	0.64567E-13
$\max(rel(X))$	0.27903E-12	0.26192E-10	0.77438E-07	0.43275E-03
$\max(rel(Y))$	0.57672E-13	0.32331E-10	0.15776E-07	0.54285E-04
$\text{avg}(abs(X))$	0.63819E-15	0.48374E-15	0.40466E-13	0.31112E-12
$\text{avg}(abs(Y))$	0.74922E-16	0.61229E-15	0.43697E-14	0.25992E-13
$\text{avg}(rel(X))$	0.19247E-13	0.11704E-12	0.27750E-10	0.30753E-08
$\text{avg}(rel(Y))$	0.53559E-14	0.15563E-12	0.59087E-11	0.39322E-09

Table 1: *Inverse Power (30 iterations) vs Principal Algorithm.*

7.1 Experiment 1.

In this experiment, we illustrate the performance of the algorithm on certain matrices.

Description. We proceed as follows. First, we choose, more or less arbitrarily, a real number $c > 1$. Then, we choose an integer number $n \approx 2c + 10\sqrt[3]{c}$, and construct a symmetric tridiagonal real n by n matrix A (see Definition 1), whose diagonal entries $A_1, \ldots, A_n$ are defined via (292) of Theorem 30 with $a = 2$. In other words,

$$A_j = 2 + \frac{j^2}{c^2}, \tag{356}$$

for every $j = 1, \ldots, n$. Then, we define the real number $\tilde{\lambda}$ via the formula

$$\tilde{\lambda} = 4 + \frac{4 \cdot 40}{\pi} \cdot \log(10), \tag{357}$$

and use $\tilde{\lambda}$ as an initial approximation to an eigenvalue of A for Inverse Power Method (see Section 3.4.1). We use 30 iterations of Inverse Power Method to evaluate the eigenvalue λ of A (λ is closer to $\tilde{\lambda}$ than any other eigenvalue of A), and the corresponding unit length eigenvector $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, whose first coordinate is positive. We repeat the calculation in extended precision to obtain the vector Y^{ext}.

Next, we run the principal algorithm (see Sections 5.2, 5.3, 5.4) to evaluate the unit length $\lambda-$eigenvector $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ of A, whose first coordinate is positive (obviously, in exact arithmetics, we would have $X = Y$). We repeat the calculation in extended precision to obtain the vector X^{ext}.

We repeat the experiment four times, with $c = 10^2, 10^3, 10^4, 10^5$. First, we make the following observations.

Observation 1. For every $c = 10^2, 10^3, 10^4, 10^5$, the eigenvalue λ is evaluated to at least 15 correct decimal digits (in double precision). In other words, the evaluated value of λ is exact up to the machine precision.

Observation 2. For every $c = 10^2, 10^3, 10^4, 10^5$, every coordinate of Y^{ext} coincides with the corresponding coordinate of X^{ext} in at least 16 decimal digits. Moreover, coordinate-wise verification confirms that both X^{ext} and Y^{ext} approximate the corresponding eigenvector of A to at least 16 decimal digits.

It follows from Observation 2 that any of X^{ext}, Y^{ext} can be used as a "reference eigenvector", to evaluate the accuracy of each coordinate of X, Y.

Table structure. We display the results of the experiment in Table 1. This table has the following structure. The first column contains the description of the evaluated quantity, and the subsequent four columns correspond, respectively, to $c = 10^2, 10^3, 10^4, 10^5$. The first four rows contain c, the dimensionality n, the eigenvalue λ and the first coordinate of the eigenvalue X_1. The next two rows contain the relative errors of X_1 and Y_1 (see (324)). The subsequent two rows contain the maximal absolute errors (see (323)) among all the coordinates of X, Y, respectively, i.e.

$$\max(abs(X)) = \max_{1\le j\le n} \{abs(X_j)\}, \quad \max(abs(Y)) = \max_{1\le j\le n} \{abs(Y_j)\}. \tag{358}$$

The next two rows contain the maximal relative errors (see (324)) among all the coordinates of X, Y, respectively, i.e.

$$\max(rel(X)) = \max_{1\le j\le n} \{rel(X_j)\}, \quad \max(rel(Y)) = \max_{1\le j\le n} \{rel(Y_j)\}. \tag{359}$$

The subsequent two rows contain the average absolute errors among all the coordinates of X, Y, respectively, i.e.

$$avg(abs(X)) = \frac{1}{n}\sum_{j=1}^{n} abs(X_j), \quad avg(abs(Y)) = \frac{1}{n}\sum_{j=1}^{n} abs(Y_j). \tag{360}$$

The last two rows contain the average relative errors among all the coordinates of X, Y, respectively, i.e.

$$avg(rel(X)) = \frac{1}{n}\sum_{j=1}^{n} rel(X_j), \quad avg(rel(Y)) = \frac{1}{n}\sum_{j=1}^{n} rel(Y_j). \tag{361}$$

Figures. Also, we display the results of the experiment in Figures 1, 2, 3, 4. In each of the figures, the $x-$axis corresponds to the indices of the eigenvector, i.e. $1 \le j \le n$. Thus, each figure contains a plot of a certain function of the indices.

The indices are divided into several groups (see also Definition 2). The color of the plot varies between different groups. The groups are defined via (306), (307). The integers $k < p < m$ of (307) are marked on the $x-$axis of each plot. Also, each group is described by how the corresponding coordinates have been computed by the algorithm of Sections 5.2, 5.3, 5.4. The description of different groups of indices is described in Table 2.

indices	$\lambda - A_j$	based on	described in	color
$1 \le j \le k$	$\lambda - A_j \ge 2$	Theorem 7	Section 5.3.1	blue
$k < j \le k + l$	$2 > \lambda - A_j$	Theorem 7	Section 5.3.2	black
$k + l < j \le p$	$2 > \lambda - A_j \ge 0$	Theorem 17	Section 5.3.3	red
$p < j \le m - r$	$0 > \lambda - A_j > -2$	Theorem 24	Section 5.3.6	magenta
$m - r < j \le m$	$\lambda - A_j > -2$	Theorem 7	Section 5.3.5	black
$m < j \le n$	$-2 \ge \lambda - A_j$	Theorem 7	Section 5.3.4	blue

Table 2: *Division of the indices. The integers* $1 < k < k + l < p < m - r < m < n$ *are those of* (306), (307).

Figures 1, 2, 3, 4 correspond, respectively, to $c = 10^2, 10^3, 10^4, 10^5$. Each of the four figures contains six plots.

On Figures 1(a), 2(a), 3(a), 4(a), we plot the coordinates X_j of X, on the linear scale (left), and on the logarithmic scale (right).

On Figures 1(b), 2(b), 3(b), 4(b), we plot the relative and absolute errors of X_j (see (324), (323)), on the logarithmic scale: relative error (left), and absolute error (right).

On Figures 1(c), 2(c), 3(c), 4(c), we plot the relative and absolute errors of Y_j (see (324), (323)), on the logarithmic scale: relative error (left), and absolute error (right).

Remark. On Figure 3, roughly every 5th point is plotted; On Figure 4, roughly every 50th point is plotted (for the sake of visual enhancement).

In addition, in Figure 5 we plot the following two functions of $k+l \le j \le p$ (see Table 2). First, we plot $|\alpha_j|$, where, α_j is that of Theorem 17 (see also Section 5.3.3). Also, we plot $X_j/2$, where X_j are the coordinates of the eigenvector. We recall that, due to (148) of Theorem 17, $X_j/2 = \Re(\alpha_j)$, for every $j = k + l + 1, \ldots, p$. Figure 5 contains two plots, corresponding, respectively, to $c = 100, n = 250$ and $c = 1000, n = 2100$.

The following observations can be made from Table 1, Figures 1, 2, 3, 4, 5, and additional experiments by the authors.

Observation 1. For every c, the coordinates X_j of X behave (as a function of j) as predicted by Theorems 11, 13, 17, 24 (see also Definition 2). In other words, X_j grow fast up until $j = k$, i.e. in the region of growth; then, oscillate between $j = k$ and $j = m$, i.e. in the oscillatory region; and, finally, decay fast (in absolute value) as j increases from m to n, i.e. in the region of decay (see Figures 1(a), 2(a), 3(a), 4(a)).

Observation 2. Relative errors of both X_j and Y_j exhibit fairly regular behavior in the regions of growth and decay, as expected (see Sections 5.3.1, 5.3.4, 5.5, in particular (338)).

Observation 3. As opposed to Observation 2, the relative errors of both X_j and Y_j oscillate in the oscillatory region, varying by several orders of magnitude. This observation is not surprising (see Sections 5.3.3, 5.3.6, 5.5, in particular (340), (342)).

Observation 4. While, for all four values of c, the first coordinate X_1 is rather small (of order 10^{-40}), it is evaluated fairly accurately by both the principal algorithm and Inverse Power Method (see Section 5.5).

Observation 5. According to the combination of Theorem 30, (343) in Section 5.5, and (356), we expect to lose roughly 2 decimal digits in the evaluation of X_1, as we multiply c by the factor of 10. This prediction is confirmed by the data in Table 1 (the fifth row): as c changes from 10^3 to 10^4 to 10^5, the relative error increases from 0.4E-13 to 0.6E-11 to 0.1E-9. Still, even for as large c as 10^5 and as small X_1 as $\approx 10^{-40}$, the latter is evaluated to 10 decimal digits.

Observation 6. While the coordinates X_j oscillate rather rapidly in the oscillatory regime, the absolute value of the complex parameter α_j (see Theorem 17) changes fairly slowly (see Figure 5). This observation provides an intuitive justification for the steps of the principal algorithm, described in Sections 5.3.3, 5.3.6. Also, it agrees with the error analysis of Section 5.5.

Observation 7. The coordinates of the vector Y are usually evaluated somewhat more accurately than those of X (by roughly one decimal digit). In other words, 30 iterations of Inverse Power Method, while being more expensive by a constant factor in terms of operations, slightly outperform the principal algorithm of this paper. At the bottom line, however, the two algorithms are rather similar in terms of both computational cost and accuracy.

7.2 Experiment 2.

In this experiment, we illustrate the performance of the algorithm on the matrices similar to those of Experiment 1.

Description. We proceed as follows. We choose the real number $c = 1000$ and integer $n = 2100$. Then, we construct the symmetric tridiagonal n by n matrix A, whose non-zero off-diagonal entries are all one, and whose diagonal entries $A_1, \ldots, A_n$ are defined via (356) in Section 7.1.

We use Inverse Power Method (see Section 3.4.1) to evaluate several eigenvalues of A, and, for each eigenvalue λ, we evaluate the corresponding unit length eigenvector $Y = Y(\lambda) = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, whose first coordinate is positive (using 30 iterations of Inverse Power Method for each λ). For each such eigenvalue λ, we repeat the calculation in extended precision to obtain the vectors $Y^{ext} = Y^{ext}(\lambda)$.

Next, for each such λ, we run the principal algorithm (see Sections 5.2, 5.3, 5.4) to evaluate the unit length $\lambda-$eigenvector $X = X(\lambda) = (X_1, \ldots, X_n) \in \mathbb{R}^n$ of A, whose first coordinate is positive (obviously, in exact arithmetics, we would have $X = Y$). For each λ, we repeat the calculation in extended precision to obtain the vectors $X^{ext} = X^{ext}(\lambda)$.

First, we make the following observations.

Observation 1. In double precision, all the evaluated eigenvalues λ are correct to tat least 15 decimal digits. In other words, these eigenvalues are exact up to the machine precision.

Observation 2. For every λ, every coordinate of $Y^{ext}(\lambda)$ coincides with the corresponding coordinate of $X^{ext}(\lambda)$ in at least 16 decimal digits. Moreover, coordinate-wise

verification confirms that, for every λ, both $X^{ext}(\lambda)$ and $Y^{ext}(\lambda)$ approximate the corresponding eigenvector of A to at least 16 decimal digits.

It follows from Observation 2 that, for every λ, any of $X^{ext}(\lambda), Y^{ext}(\lambda)$ can be used as a "reference eigenvector", to evaluate the accuracy of each coordinate of $X(\lambda), Y(\lambda)$.

λ	X_1	$rel(X_1)$	max($rel(X)$)	avg($rel(X)$)
0.40351E+01	0.10809E-13	0.58646E-12	0.29176E-09	0.13114E-11
0.40471E+01	0.99452E-18	0.66326E-12	0.28637E-09	0.17473E-11
0.40595E+01	0.63720E-22	0.86132E-12	0.99517E-09	0.23739E-11
0.40705E+01	0.12641E-25	0.28375E-11	0.30593E-08	0.81554E-11
0.40836E+01	0.46754E-30	0.10490E-12	0.13829E-09	0.63507E-12
0.40932E+01	0.27309E-33	0.70170E-12	0.13838E-09	0.16682E-11
0.41069E+01	0.66341E-38	0.35095E-12	0.59918E-09	0.14458E-11
0.41168E+01	0.29308E-41	0.36528E-13	0.26192E-10	0.11704E-12
0.41289E+01	0.24013E-45	0.42306E-12	0.29842E-08	0.27627E-11
0.41392E+01	0.84484E-49	0.87444E-13	0.15693E-09	0.60358E-12
0.41537E+01	0.10509E-53	0.70798E-12	0.25694E-09	0.12596E-11
0.41643E+01	0.29507E-57	0.17023E-13	0.62182E-09	0.54695E-12
0.41728E+01	0.39899E-60	0.29785E-11	0.76945E-10	0.39780E-11
0.42665E+01	0.13675E-91	0.34374E-13	0.20497E-09	0.60583E-12

Table 3: *Principal Algorithm. Parameters:* $c = 1000$, $n = 2100$.

Table structure. The results of the experiment are displayed in Tables 3, 4. These tables have the following structures. The first column contains some eigenvalues λ of the matrix A. The second column contains the first coordinate of the eigenvector X (Table 3) or Y (Table 4). The third column contains the relative error of this coordinate (see (324)). The fourth column contains the maximal relative error of X, Y, respectively (see (359) in Section 7.1). The last column contains the average relative error of X, Y, respectively (see (361) in Section 7.2).

The following observations can be made from Tables 3, 4.

Observation 1. Usually, the vector $Y(\lambda)$ is more accurate than the vector $X(\lambda)$ by roughly one decimal digit.

Observation 2. The accuracy of $X(\lambda)$ is roughly the same throughout all of Table 3, in terms of relative error. In particular, the first coordinate X_1 of the eigenvector X is evaluated by the principal algorithm with roughly the same accuracy, for every λ, independent of the order of magnitude of X_1. In other words, even though X_1 varies between $\approx 10^{-14}$ (first row) and 10^{-92} (last row), it is always evaluated to roughly 12 decimal digits (see also Section 5.3.1).

Observation 3. The accuracy of $Y(\lambda)$ is roughly the same throughout all of Table 4, in terms of relative error. In particular, the first coordinate Y_1 of the eigenvector Y is evaluated by the principal algorithm with roughly the same accuracy, for every λ (to roughly 13 decimal digits), independent of the order of magnitude of Y_1.

λ	Y_1	$rel(Y_1)$	$\max(rel(Y))$	$\text{avg}(rel(Y))$
0.40351E+01	0.10809E-13	0.24959E-13	0.13232E-09	0.23343E-12
0.40471E+01	0.99452E-18	0.38924E-13	0.22990E-10	0.13062E-12
0.40595E+01	0.63720E-22	0.54421E-13	0.15798E-09	0.26474E-12
0.40705E+01	0.12641E-25	0.49833E-13	0.12447E-09	0.26264E-12
0.40836E+01	0.46754E-30	0.47955E-13	0.30739E-10	0.16290E-12
0.40932E+01	0.27309E-33	0.63108E-13	0.35663E-10	0.15882E-12
0.41069E+01	0.66341E-38	0.40131E-13	0.60993E-10	0.19122E-12
0.41168E+01	0.29308E-41	0.58053E-13	0.32331E-10	0.15563E-12
0.41289E+01	0.24013E-45	0.61224E-13	0.53569E-09	0.40129E-12
0.41392E+01	0.84484E-49	0.48218E-13	0.50575E-10	0.21127E-12
0.41537E+01	0.10509E-53	0.64663E-13	0.29615E-10	0.13552E-12
0.41643E+01	0.29507E-57	0.49151E-13	0.27351E-09	0.25378E-12
0.41728E+01	0.39899E-60	0.48311E-13	0.83383E-11	0.11069E-12
0.42665E+01	0.13675E-91	0.46828E-13	0.33287E-10	0.74736E-13

Table 4: *Inverse Power (30 iterations). Parameters:* $c = 1000$, $n = 2100$.

7.3 Experiment 3.

In this experiment, we illustrate the numerical algorithms of Section 5 via evaluation of Bessel functions (see Sections 3.2, 3.4.3, 6.1).

Description. We choose, more or less arbitrarily, the real number $c > 0$ and the integer $n > 0$. Then, we choose the integer $N > \max\{n, c\}$. We construct the symmetric tridiagonal $2N + 1$ by $2N + 1$ matrix A, whose non-zero off-diagonal entries are all one (see (60)), and whose diagonal entries $A_1, \ldots, A_{2N+1}$ are defined via (61) of Remark 9 in Section 3.4.3. We define the real number λ via (63) (λ is an eigenvalue of A, due to Remark 9).

We use the principal algorithm (see Sections 5.2, 5.3, 5.4) to evaluate the $\lambda-$eigenvector $X \in \mathbb{R}^{2N+1}$ of A whose first coordinate is positive. We repeat the computation in extended precision, to obtain $X^{ext} \in \mathbb{R}^{2N+1}$. Then, we run 30 iterations of Inverse Power Method (see Section 3.4.1) to evaluate the unit length $\lambda-$eigenvector $Y \in \mathbb{R}^{2N+1}$ of A whose first coordinate is positive (obviously, in exact arithmetics, X and Y would be the same vector). We repeat the computation in extended precision to obtain $Y^{ext} \in \mathbb{R}^{2N+1}$.

Based on Remark 9, we evaluate $J_0(c), \ldots, J_n(c)$ from X (or Y) as follows. Suppose, for example, that

$$X = (X_N, X_{N-1}, \ldots, X_1, X_0, X_{-1}, \ldots, X_{-(N-1)}, X_{-N}). \tag{362}$$

We evaluate $J_0(c)$ as X_0, $J_1(c)$ as X_1, up to $J_n(c)$ as X_n.

We repeat the experiments for all the choices of parameters c, n, N, listed in Table 5.

First, we make the following observations.

Observation 1. For $c = 100$, we ran the experiment in extended precision with $N = 215$ and $N = 235$, to obtain $X^{ext} \in \mathbb{R}^{2N+1}$ and $Y^{ext} \in \mathbb{R}^{2N+1}$, for each N. These four vectors coincide in the middle $2n+1 = 401$ coordinates up to at least 15 digits. In other words, using

c	n	N
100	200	215
100	200	235
1,000	1,200	1,250
1,000	1,200	1,270
10,000	10,490	10,550
10,000	10,490	10,570
100,000	101,000	101,150
100,000	101,000	101,200

Table 5: *Choice of parameters for Experiment 3.*

the notation (362), these four vectors coincide in the coordinates with indices $0, \pm 1, \ldots, \pm n$ up to at least 15 digits.

Observation 2. Similar to Observation 1, for each $c = 10^3, 10^4, 10^5$ and each corresponding pair of N from Table 5, the four vectors coincide in the middle $2n+1$ coordinates up to at least 15 digits (in extended precision).

We conclude from Observations 1, 2 that, for each pair of c, n from Table 5, for either of the two corresponding values of N, both X^{ext} and Y^{ext} can be used as "reference" vectors to evaluate the accuracy of the evaluation of $J_0(c), \ldots, J_n(c)$. In particular, in extended precision, each of $J_0(c), \ldots, J_n(c)$ is evaluated to at least 15 decimal digits, by either method.

c	N	k	$J_k(c)$	$\lvert X_k/J_k(c) - 1\rvert$	$\lvert Y_k/J_k(c) - 1\rvert$
100	215	0	0.19986E-01	0.48607E-14	0.14408E-13
		1	-.77145E-01	0.69438E-13	0.17989E-14
		2	-.21529E-01	0.83800E-14	0.13053E-13
		200	0.20594E-40	0.48517E-13	0.29705E-14
1,000	1,250	0	0.24787E-01	0.27994E-14	0.83983E-15
		1	0.47283E-02	0.10944E-11	0.84382E-13
		2	-.24777E-01	0.32206E-14	0.00000E+00
		1,200	0.83509E-38	0.11268E-12	0.53135E-14
10,000	10,550	0	-.70962E-02	0.17552E-12	0.25069E-12
		1	0.36475E-02	0.13038E-10	0.93087E-12
		2	0.70969E-02	0.17648E-12	0.25042E-12
		10,490	0.35152E-46	0.14137E-11	0.16278E-12
100,000	101,150	0	-.17192E-02	0.76153E-11	0.88710E-11
		1	0.18468E-02	0.48246E-10	0.76843E-11
		2	0.17192E-02	0.76138E-11	0.88708E-11
		101,000	0.39770E-43	0.28153E-10	0.11461E-11

Table 6: *Evaluation of $J_k(c)$. Corresponds to Experiment 3.*

Next, we run the experiment in double precision, with $c = 10^2, 10^3, 10^4, 10^5$, and $N = 215, 1250, 10550, 101150$, respectively.

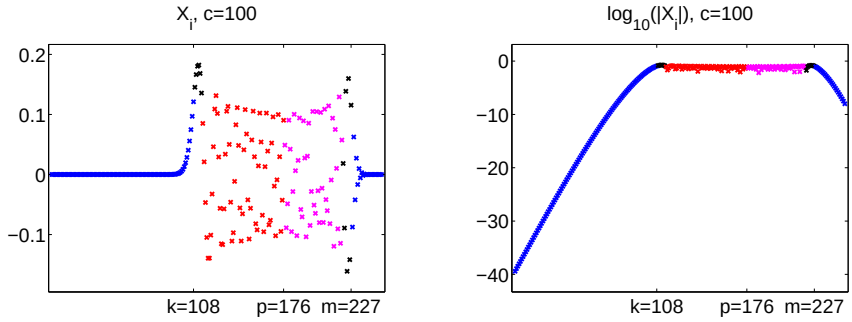
Table structure. The results of this experiment are displayed in Table 6. This table has the following structure. The first two columns contain the parameters c and N. The third column contains the integer k between 0 and n, where $n < N$ is that of Table 5. The fourth column contains $J_k(c)$ (evaluated in extended precision). The fifth column contains the relative error to which X_k approximates $J_k(c)$. The last column contains the relative error to which Y_k approximates $J_k(c)$.

We make the following observations from Table 6 and some additional experiments.

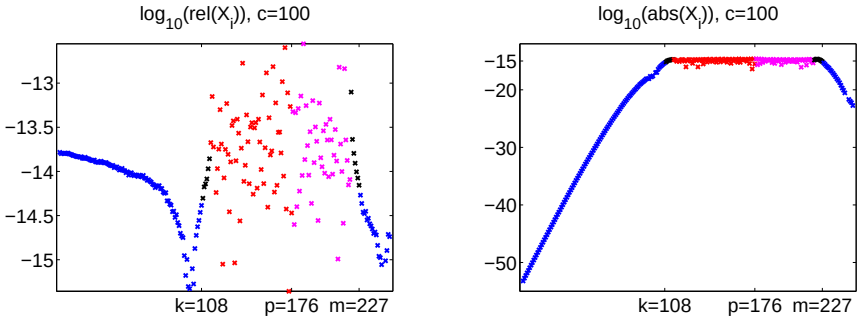
Observation 1. For every c, the relative accuracy of X_n is similar to that of X_0, X_1, X_2. This is despite the fact that $|X_k| \approx 10^{-2}$ for $k = 0, 1, 2$, while $|X_n| \approx 10^{-40}$. This phenomenon, however, confirms the analysis of Section 5.5.

Observation 2. According to (343) of Section 5.5 (see also Section 3.4.3), when we evaluate $J_n(c)$ via computing X_n, we expect to lose roughly 4/3 decimal digits every time we increase c by the factor of 10. This prediction is confirmed by the data in Table 6 (see rows 4,8,12,16 in the fifth column).

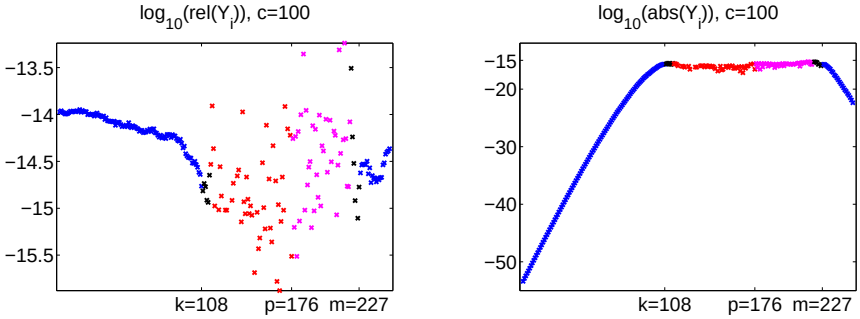
Observation 3. The coordinate of Y, evaluated via 30 iterations of Inverse Power Method, are usually more accurate than those of X (by roughly one decimal digits). However, sometimes the opposite is true (see rows 1,5,9,13).



(a) coordinates: linear and logarithmic scales

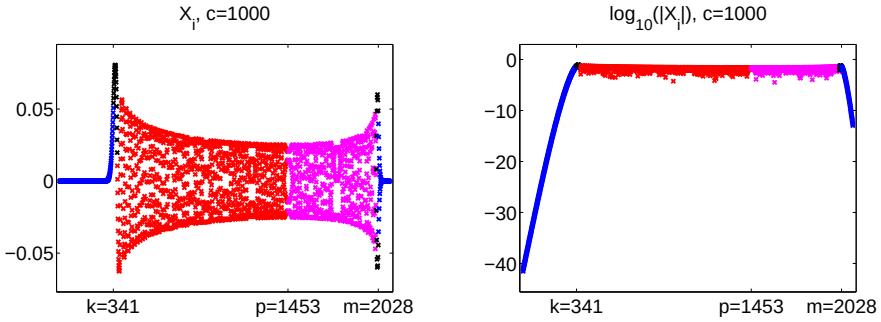


(b) principal algorithm: relative and absolute errors

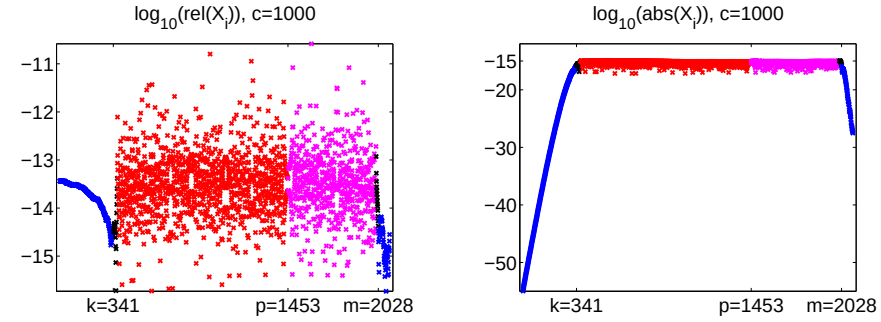


(c) inverse power: relative and absolute errors

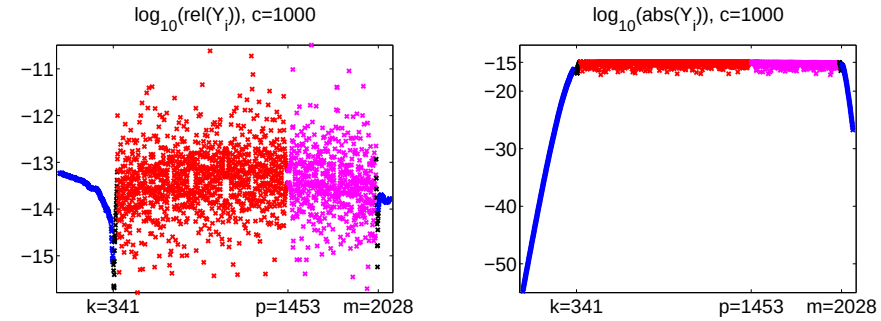
Figure 1: The coordinates of X (principal algorithm) and Y (30 iterations of Inverse Power). Parameters: $c = 100$, $n = 250$, $\lambda = 0.51665\text{E+}01$.



(a) coordinates: linear and logarithmic scales

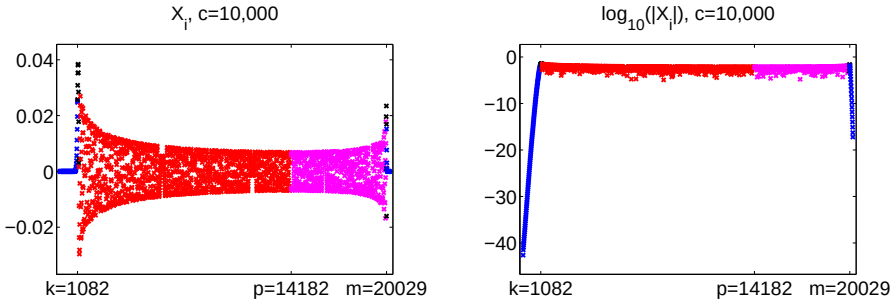


(b) principal algorithm: relative and absolute errors

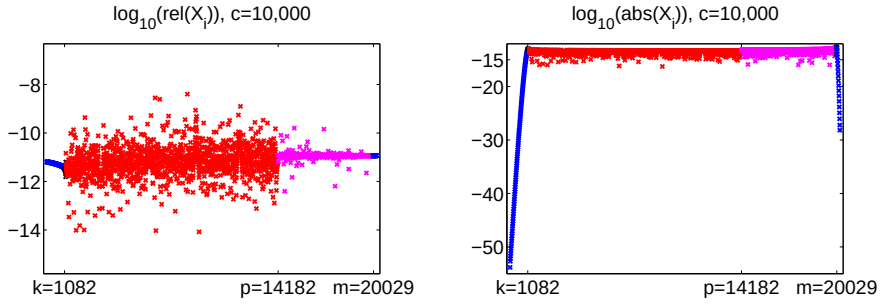


(c) inverse power: relative and absolute errors

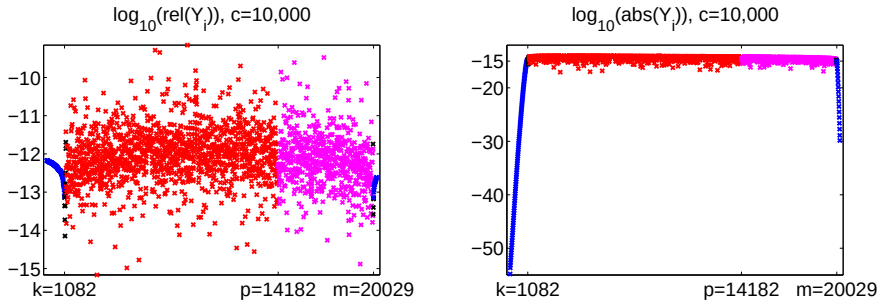
Figure 2: The coordinates of X (principal algorithm) and Y (30 iterations of Inverse Power). Parameters: $c = 1000$, $n = 2100$, $\lambda = 0.41168\text{E+}01$.



(a) coordinates: linear and logarithmic scales

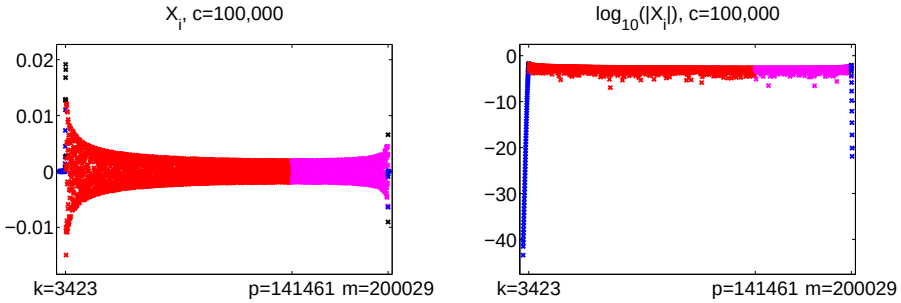


(b) principal algorithm: relative and absolute errors

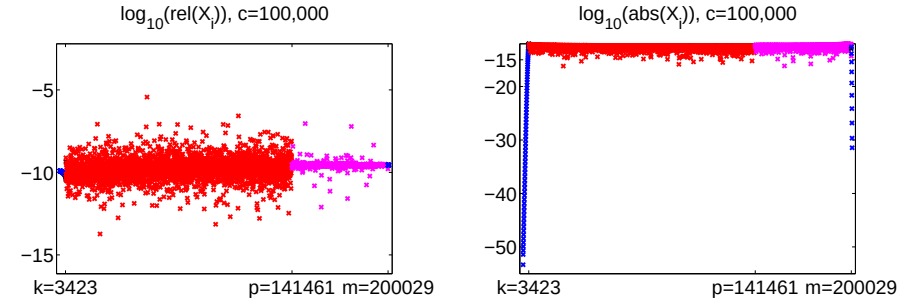


(c) inverse power: relative and absolute errors

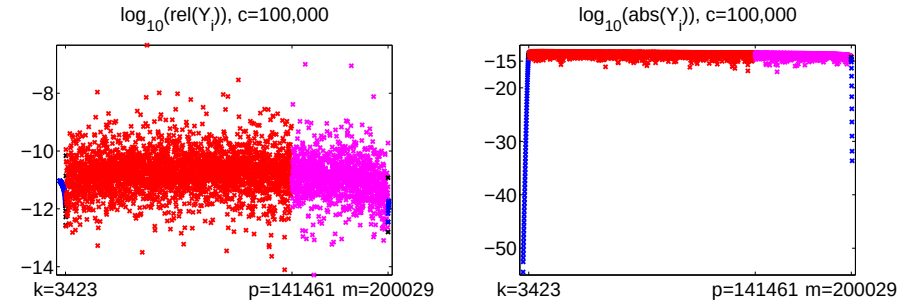
Figure 3: The coordinates of X (principal algorithm) and Y (30 iterations of Inverse Power). Parameters: $c = 10,000$, $n = 20,215$, $\lambda = 0.40117\text{E+}01$.



(a) coordinates: linear and logarithmic scales



(b) principal algorithm: relative and absolute errors



(c) inverse power: relative and absolute errors

Figure 4: The coordinates of X (principal algorithm) and Y (30 iterations of Inverse Power). Parameters: $c = 100,000$, $n = 200,500$, $\lambda = 0.40012\text{E+}01$.

References

[1] M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover Publications, 1964.

[2] W. Barth, R. S. Martin, J. H. Wilkinson, *Calculation of the Eigenvalues of a Symmetric Tridiagonal Matrix by the Method of Bisection*, Numerische Mathematik 9, 386-393, 1967.

[3] G. Dahlquist, A. Björk, *Numerical Methods*, Prentice-Hall Inc., 1974.

[4] G. J. F. Francis *The QR transformation, parts I and II*, Computer J. 4, 265-271, 332-345. 1961-2.

[5] W. J. Givens *Numerical computation of the characteristic values of a real symmetric matrix*, Technical Report ORNL-1574, Oak Ridge National Laboratory, TX. 1954.

[6] G. Golub, C. V. Loan, *Matrix Computations*, Second Edition, Johns Hopkins University Press, Baltimore, 1989.

[7] I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products*, Seventh Edition, Elsevier Inc., 2007.

[8] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, New York: Wiley, 1966.

[9] V. N. Kublanovskaya *On some algorithms for the solution of the complete eigenvalue problem*, Zh. Vych. Mat. 1, pp. 555-570. 1961.

[10] H. J. Landau, H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty - II*, Bell Syst. Tech. J. January 65-94, 1961.

[11] H. J. Landau, H. Widom, *Eigenvalue distribution of time and frequency limiting*, J. Math. Anal. Appl. 77, 469-81, 1980.

[12] A. Osipov, *Explicit upper bounds on the eigenvalues associated with prolate spheroidal wave functions*, Yale CS Technical Report #1450, 2012.

[13] A. Osipov, *Certain upper bounds on the eigenvalues associated with prolate spheroidal wave functions*, arXiv:1206.4541v1, 2012.

[14] A. Osipov, V. Rokhlin, *Detailed analysis of prolate quadratures and interpolation formulas*, Yale CS Technical Report #1458, 2012.

[15] A. Osipov, V. Rokhlin, *Detailed analysis of prolate quadratures and interpolation formulas*, arXiv:1208.4816v1, 2012.

[16] B. N. Parlett *The symmetric eigenvalue problem*, Prentice Hall, Inc. 1980.

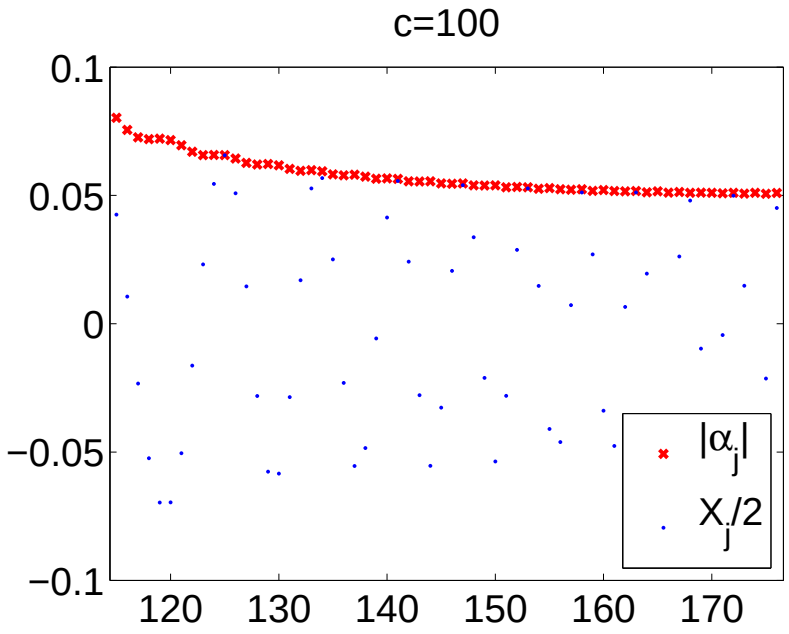
[17] Vladimir Rokhlin, Hong Xiao, *Approximate Formulae for Certain Prolate Spheroidal Wave Functions Valid for Large Value of Both Order and Band Limit.*

[18] D. Slepian, H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty - I*, Bell Syst. Tech. J. January 43-63, 1961.

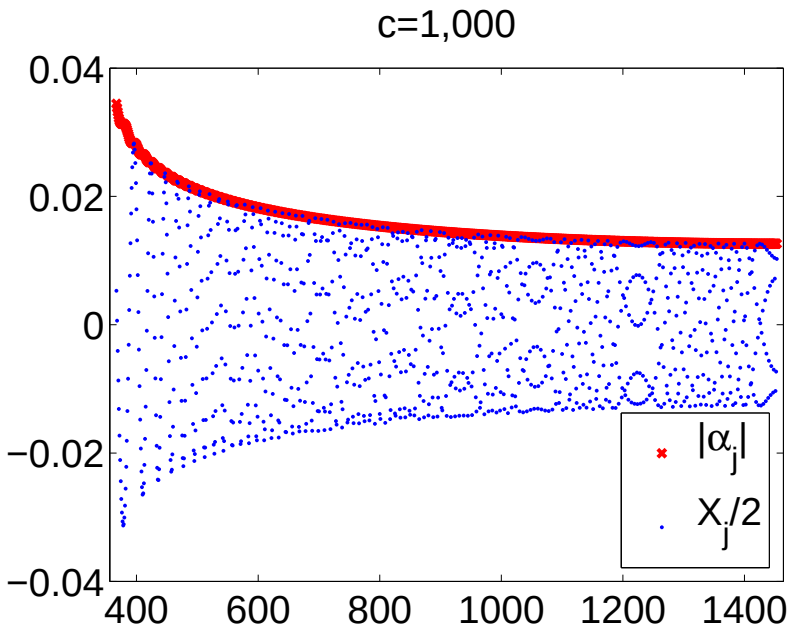
[19] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Second Edition, Springer-Verlag, 1993.

[20] J. H. Wilkinson, *Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

[21] H. Xiao, V. Rokhlin, N. Yarvin, *Prolate spheroidal wavefunctions, quadrature and interpolation*, Inverse Problems, 17(4):805-828, 2001.



(a) $c = 100$, $n = 250$, $\lambda = 0.51665\text{E+}01$



(b) $c = 1,000$, $n = 2,100$, $\lambda = 0.41168\text{E+}01$

Figure 5: The coordinates X_j vs. $|\alpha_j|$. Corresponds to Experiment 1.