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. While our algorithm can be viewed as a modification of already existing (and well known) algorithms, such error analysis appears to be new. Also, we discuss some applications in which this task is necessary. Our results are illustrated via several numerical examples.

Small coordinates of eigenvectors of certain symmetric tridiagonal matrices: numerical evaluation and error analysis

Andrei Osipov
Research Report YALEU/DCS/TR-1495
Yale University
August 25, 2014

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], [19], [21], [22]). 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, 3.4.2, 6.1), the evaluation of some quantities associated with prolate spheroidal wave functions (see Sections 3.3, 6.2, and also [18]), and the evaluation of singular values of the truncated Laplace transform (see [11]), 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. More specifically, we consider the matrices whose non-zero off-diagonal elements are constant, and whose diagonal elements constitute a monotonically increasing sequence (see, however, Remark 2 below). The connection of such matrices to Bessel functions and prolate spheroidal wave functions is discussed in Sections 3.4.2, 6.2, respectively. Also, we carry out detailed error analysis of our algorithm (see Sections 4.2, 4.3). 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 (see Section 7).

The following is one of the principal analytical results of this paper (see Theorem 22 in Section 4.3 for a more precise statement, and Theorems 16, 17, 18, Corollary 6 in Section 4.2 below for the treatment of a more general case).

Theorem 1. *Suppose that $a \geq 1$ is a real number, and that, for any real $c \geq 1$, $n = n(c) > c$ is an integer, the real numbers $A_1(c), \ldots, A_n(c)$ are defined via the formula*

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

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

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

Suppose furthermore that, for any real $c > 1$, $\lambda(c)$ is an eigenvalue of $A(c)$, that $1 < k(c) < n(c)$ is an integer, that

$$2 + A_{k(c)} < \lambda(c) \leq 2 + A_{k(c)+1}. \tag{5}$$

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

$$R(c, a) \leq \varepsilon \cdot O\left(-\log(X_1)\right) \cdot O\left(c^{4 \cdot a/(a+2)}\right), \quad c \to \infty. \tag{6}$$

Remark 1. *We observe that, according to* (6)*, the relative precision of $X_1, \ldots, X_k$ depends only logarithmically 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)*; matrices whose diagonal entries are of a more general form than* (3)*; banded (not necessarily tridiagonal) matrices, etc. While such generalizations are straightforward (and are based, in part, on the results of Section 4.1), the analysis is somewhat involved, and will be published at a later date (see, however, Theorems 16, 17 and Corollary 6 in Section 4.2 below for one such generalization).*

The proof of Theorem 1 is constructive and somewhat technical (see Sections 4.2, 4.3). The resulting numerical algorithms for the evaluation of the eigenvector X are described in Section 5.

In practice, the upper bound in (6) above seems to be overly pessimistic. In fact, the following conjecture has been verified by extensive numerical experiments (see Section 7).

Conjecture 1. *Suppose that, in addition to the hypothesis of Theorem 1 above, we evaluate the eigenvector $X(c)$ by the algorithm from Section 5.3. Then, for any real $c > 1$ and any integer $1 \leq j \leq k(c)$,*

$$\mathrm{rel}(X_j(c)) \leq 100 \cdot c^{2 \cdot a/(a+2)} \cdot \varepsilon. \tag{7}$$

In particular, $\mathrm{rel}(X_1)$ does not depend on the magnitude of X_1, for any $c > 1$.

We observe that the power of c in (7) is half the power of c in (6). In other words, Theorem 1 appears to overestimate the number of lost digits in the evaluation of the first k elements of X by roughly a factor of two.

The paper is organized as follows. Section 2 contains a brief informal overview of some principal ideas behind the algorithms of this paper and their error analysis. 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 develop the necessary analytical apparatus and perform error analysis of the algorithm, described in Section 5 (and we also describe a number of related algorithms). In Section 6, we discuss some applications of our algorithm to other computational problems. In Section 7, we illustrate the numerical stability of our algorithm and corresponding theoretical results via several numerical examples, and provide comparison to some related classical algorithms.

2 Overview

This section contains an informal discussion of several properties of eigenvectors of certain symmetric tridiagonal matrices.

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 $A_1 < A_2 < \ldots$. Then, the coordinates $x_1, \ldots, x_n$ of any eigenvector x of A corresponding to the eigenvalue λ satisfy the three-term linear recurrence relation (43) (see Theorem 5 in Section 4.1 below).

It turns out that the qualitative properties of the recurrence relation (43) 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 obvious 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 rigorously developed in Theorems 6, 13, 16 in Section 4.2 and Corollary 1 in Section 4.1 (see also Observation 1 in Section 5.4).*

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 rigorously developed in Theorem 7 in Section 4.1 (see also Observation 2 in Section 5.4).*

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

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

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

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

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

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

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).

This examples serves as a model for the relation between x_{k-1}, x_k, x_{k+1} in (43) when $-2 < \lambda - A_k < 2$. In fact, in this case (43) admits the form (12) if we allow θ to vary with k, e.g.

$$x_{k+1} = 2 \cdot \cos(\theta_{k-1}) \cdot x_k - x_{k-1}. \tag{13}$$

We say that sequences defined by the recurrence relation of the form (13) exhibit "oscillatory behavior", and analyze such sequences in Theorems 8, 10, 11, 15, 16, 17 and Corollaries 3, 4 in Sections 4.1, 4.2 (see also Observation 3 in Section 5.4).

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 the following theorem, we summarize several well known facts about certain symmetric tridiagonal matrices.

Theorem 2. *Suppose that $n > 1$ is an integer, that $2 < A_1 < \cdots < A_n$ is an increasing sequence of positive real numbers, and that the symmetric tridiagonal n by n matrix A defined via the formula*

$$A = \begin{pmatrix} A_1 & 1 & & & & \\ 1 & A_2 & 1 & & & \\ & 1 & A_3 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & A_{n-1} & 1 \\ & & & & 1 & A_n \end{pmatrix}. \tag{14}$$

Suppose also that

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n \tag{15}$$

are the eigenvalues of A. Then,

$$A_k - 2 < \lambda_k \leq A_k + 2, \tag{16}$$

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

$$\lambda_1 < A_1, \tag{17}$$

and

$$\lambda_n > A_n. \tag{18}$$

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

$$J_n(z) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \cdot (m+n)!} \cdot \left(\frac{z}{2}\right)^{2m+n}, \tag{19}$$

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

$$J_{-n}(z) = (-1)^n \cdot J_n(z), \tag{20}$$

for all complex z.

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

$$z \cdot J_{n-1}(z) + z \cdot J_{n+1}(z) = 2n \cdot J_n(z), \tag{21}$$

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

$$\sum_{n=-\infty}^{\infty} J_n^2(x) = 1, \tag{22}$$

for all real x. In the following theorem, we rewrite (21), (22) in the matrix form.

Theorem 3. *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*

$$A_{n,n-1}(x) = A_{n,n+1}(x) = 1, \tag{23}$$

for all integer n, and

$$A_{n,n}(x) = -\frac{2n}{x}, \tag{24}$$

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

$$v_n(x) = J_n(x), \tag{25}$$

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

$$A(x) \cdot v(x) = 0, \tag{26}$$

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 [23], [12], [20], [10], [15].

Suppose that $c > 0$ is a real number, and that the integral operator $F_c : L^2[-1,1] \to [-1,1]$ is defined via the formula

$$F_c[\varphi](x) = \int_{-1}^{1} \varphi(t) \cdot e^{icxt}\, dt. \tag{27}$$

Suppose also that the complex numbers $\lambda_0(c), \lambda_1(c), \dots$ are the eigenvalues of F_c (ordered such that $|\lambda_0(c)| > |\lambda_1(c)| > \dots$). The prolate spheroidal wave functions (PSWFs) corresponding to the band limit c are the unit-norm eigenfunctions $\psi_0^{(c)}, \psi_1^{(c)}, \dots$ of F_c.

Several popular numerical algorithm for the evaluation of PSWFs (see e.g. [23], [18]) are based on the following theorem (that goes back at least to [20]).

Theorem 4. *Suppose that $c > 0$ is a real number, and that $n \geq 0$ is an even integer. Suppose also that $P_0, P_1, \dots$ are the Legendre polynomials, and that the real numbers $\beta_0^{(n,c)}, \beta_2^{(n,c)}, \dots$ are defined via the formula*

$$\beta_k^{(n,c)} = \int_{-1}^{1} \psi_n^c(x) \cdot P_k(x) \cdot \sqrt{k+1/2}\, dx, \tag{28}$$

for every $k = 0, 2, 4, \dots$. Suppose, in addition, that the non-zero entries of the infinite symmetric tridiagonal matrix $A^{c,even}$ are defined via the formulae

$$\begin{aligned} A_{k,k}^{(c)} &= k(k+1) + \frac{2k(k+1)-1}{(2k+3)(2k-1)} \cdot c^2, \\ A_{k,k+2}^{(c)} &= A_{k+2,k}^{(c)} = \frac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+1)(2k+5)}} \cdot c^2, \end{aligned} \tag{29}$$

for every $k = 0, 2, 4, \dots$, that $\chi_0(c) < \chi_2(c) < \dots$ are the eigenvalues of $A^{c,even}$, and that the infinite dimensional vector $\beta^{(n,c)}$ is defined via the formula

$$\beta^{(n,c)} = \left(\beta_0^{(n,c)}, \beta_2^{(n,c)}, \dots\right)^T. \tag{30}$$

Then, $\beta^{(n,c)}$ is the unit-norm (in the l^2-sense) eigenvector of $A^{c,even}$ corresponding to the eigenvalues $\chi_n(c)$, and, moreover,

$$\psi_n^c(x) = \sum_{j=0}^{\infty} \beta_{2\cdot j}^{(n,c)} \cdot P_{2\cdot j}(x) \cdot \sqrt{2 \cdot j + 1/2}, \tag{31}$$

for all $-1 \le x \le 1$. In addition,

$$\lambda_n(c) = \frac{\beta_0^{(n,c)} \cdot \sqrt{2}}{\psi_n^c(0)}. \tag{32}$$

Remark 5. *A similar theorem for odd values of n is almost identical to Theorem 4 above and can be found, for example, in [20], [15].*

Remark 6. *While to obtain $\psi_n^c(x)$ via (31) it suffices to evaluate the coordinates of $\beta^{(n,c)}$ in (30) to high absolute accuracy, to obtain $\lambda_n(c)$ via (32) one needs to evaluate $\beta_0^{(n,c)}$ to high relative accuracy, regardless of the size of $\beta_0^{(n,c)}$.*

3.4 Numerical Tools

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

3.4.1 Shifted Inverse Power Method

Suppose that $n \ge 0$ is an integer, and that A is an n by n real symmetric matrix. Suppose also that $\sigma_1 < \sigma_2 < \cdots < \sigma_n$ are the eigenvalues of A. The Shifted Inverse Power Method iteratively finds the eigenvalue σ_k and the corresponding eigenvector $v_k \in \mathbb{R}^n$, provided that an approximation λ to σ_k is given, and that

$$|\lambda - \sigma_k| < \max\{|\lambda - \sigma_j| \ : \ j \ne k\}. \tag{33}$$

Each Shifted Inverse Power iteration solves the linear system

$$(A - \lambda_j I) \cdot x = w_j \tag{34}$$

in the unknown $x \in \mathbb{R}^n$, where λ_j and $w_j \in \mathbb{R}^n$ are the approximations to σ_k and v_k, respectively, after j iterations; the number λ_j is usually referred to as "shift". The approximations λ_{j+1} and $w_{j+1} \in \mathbb{R}^n$ (to σ_k and v_k, respectively) are evaluated from x via the formulae

$$w_{j+1} = \frac{x}{\|x\|}, \quad \lambda_{j+1} = w_{j+1}^T \cdot A \cdot w_{j+1} \tag{35}$$

(see, for example, [3], [22] for more details).

Remark 7. *For symmetric matrices, the Shifted Inverse Power Method converges cubically in the vicinity of the solution. In particular, if the matrix A is tridiagonal, and the initial approximation λ is sufficiently close to σ_k, the Shifted Inverse Power Method evaluates σ_k and v_k essentially to machine precision ε in $O(\log(-\log\varepsilon))$ iterations, and each iteration requires $O(n)$ operations (see e.g [22], [3]).*

3.4.2 Evaluation of Bessel Functions

The following numerical algorithm for the evaluation of Bessel functions (see Section 6.1) is based on Theorem 3 in Section 3.2 (see e.g [1], [13]).

Suppose that $x > 0$ is a real number, and that $m > 0$ is an integer.

Algorithm: evaluation of $J_0(x), J_1(x), \ldots, J_m(x)$.

- select integer $N > \max\{m, x\}$ (see Remark 8 below).
- 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 (21), in the direction of decreasing indices. In other words, evaluate $\tilde{J}_{k-1}$ via

$$\tilde{J}_{k-1} = \frac{2k}{x} \cdot \tilde{J}_k(x) - \tilde{J}_{k+1}(x), \tag{36}$$

  for every $k = N, \ldots, 2$.
- evaluate $\tilde{J}_{-1}$ from $\tilde{J}_1$ via (20).
- evaluate $\tilde{J}_0$ from $\tilde{J}_1, \tilde{J}_{-1}$ via (21).
- evaluate the real number d via the formula

$$d = \sqrt{\tilde{J}_0^2 + 2 \cdot \sum_{k=1}^{N} \tilde{J}_k^2}. \tag{37}$$

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

Remark 8. *In this paper, we always select sufficiently large N so that the algorithm described above, when carried out in extended precision, evaluates $J_0(x), \ldots, J_m(x)$ to at least 17 decimal digits. Further discussion of the matter is beyond the scope of this paper (see e.g. [1] for more details).*

In the following remark, we state that the algorithm described above essentially evaluates an eigenvector of a certain tridiagonal symmetric matrix (as in (14)).

Remark 9. *Suppose that $x > 0$ is a real number, that $0 < m < N$ are integers, and that the real numbers $\tilde{J}_0, \ldots, \tilde{J}_N$ and d are defined, respectively, via* (36), (37) *above. Suppose also that A is the symmetric tridiagonal $(2N+1) \times (2N+1)$ matrix, whose non-zero off-diagonal entries are defined via the formula*

$$A_{j,j+1} = A_{j+1,j} = 1, \tag{38}$$

for every $j = 1, \ldots, 2N$, and whose diagonal entries are defined via the formula

$$A_{j,j} = A_j = 2 + \frac{2j}{x}, \tag{39}$$

for every $j = 1, \ldots, 2N, 2N+1$. Suppose, in addition, that the real number λ is defined via the formula

$$\lambda = 2 + \frac{2 \cdot (N+1)}{x}, \tag{40}$$

and that the vector $X \in \mathbb{R}^{2N+1}$ is defined via the formula

$$X = \frac{1}{d} \cdot \left(\tilde{J}_N, \ldots, \tilde{J}_1, \tilde{J}_0, -\tilde{J}_1, \ldots, (-1)^N \cdot \tilde{J}_N \right). \tag{41}$$

Then, λ is an eigenvalue of A, and X is the unit-length eigenvector of A corresponding to λ.

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 Local Properties of Eigenvectors of Certain Tridiagonal Matrices

In this subsection, we develop several analytical results pertaining to the eigenvectors of certain tridiagonal symmetric matrices.

In the following theorem, we describe some obvious properties of the eigenvectors of certain tridiagonal symmetric matrices.

Theorem 5. *Suppose that $n > 1$ is an integer, that $2 < A_1 < A_2 < \ldots$ is an increasing sequence of positive real numbers, and that the symmetric tridiagonal n by n matrix A is defined via (14). Suppose also that the real number λ is an eigenvalue of A, and that $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is an eigenvector corresponding to λ. Then,*

$$x_2 = (\lambda - A_1) \cdot x_1. \tag{42}$$

Also,

$$x_{j+1} = (\lambda - A_j) \cdot x_j - x_{j-1}, \tag{43}$$

for every $j = 2, \ldots, n-1$. Finally,

$$x_{n-1} = (\lambda - A_n) \cdot x_n. \tag{44}$$

In particular, both x_1 and x_n differ from zero, and λ is simple.

Proof. The identities (42), (43), (44) follow immediately from (14) and the fact that

$$A \cdot x = \lambda \cdot x. \tag{45}$$

We observe that the coordinates $x_2, \ldots, x_n$ are completely determined by x_1 and λ via (42), (43), and hence the eigenvalue λ is simple. Obviously, neither x_1 nor x_n can be equal to zero, for otherwise x would be the zero vector. ■

In the following theorem, we assert that, under certain conditions, the first element of the eigenvectors of the matrix A from Theorem 5 must be "small".

Theorem 6. *Suppose that the n by n symmetric tridiagonal matrix A is defined via (14) in Section 3.1. 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. $x_1 > 0$. Suppose, in addition, that $1 \leq k \leq n$ is an integer, and that*

$$\lambda \geq A_k + 2. \tag{46}$$

Then,

$$0 < x_1 < x_2 < \cdots < x_k < x_{k+1}. \tag{47}$$

Also,

$$\frac{x_j}{x_{j-1}} > \frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{48}$$

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

$$1 < \frac{x_k}{x_{k-1}} < \cdots < \frac{x_3}{x_2} < \frac{x_2}{x_1}. \tag{49}$$

Proof. It follows from (46) that

$$\lambda_k - A_1 > \lambda_k - A_2 > \cdots > \lambda_k - A_k \geq 2. \tag{50}$$

We combine (42), (43) in Theorem 5 with (50) to obtain (47) by induction. Suppose now that the real numbers $r_1, \ldots, r_k$ are defined via the formula

$$r_j = \frac{x_{j+1}}{x_j}, \tag{51}$$

for every $j = 1, \ldots, k$, and that the real numbers $\sigma_1, \ldots, \sigma_k$ are defined via the formula

$$\sigma_j = \frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{52}$$

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

$$x^2 - (\lambda - A_j) \cdot x + 1 = 0. \tag{53}$$

We observe that

$$\sigma_1 > \cdots > \sigma_k \geq 1, \tag{54}$$

due to (50) and (52). Also,

$$r_1 > \sigma_1 > \sigma_2 > 1, \tag{55}$$

due to the combination of (52) and (42). Suppose now, by induction, that

$$r_{j-1} > \sigma_j > 1. \tag{56}$$

for some $2 \leq j \leq k-1$. We observe that the roots of the quadratic equation (53) are $1/\sigma_j < 1 < \sigma_j$, and combine this observation with (56) to obtain

$$r_{j-1}^2 - (\lambda - A_j) \cdot r_{j-1} + 1 > 0. \tag{57}$$

We combine (57) with (51) and (43) to obtain

$$r_j = \frac{x_{j+1}}{x_j} = \frac{(\lambda - A_j) \cdot x_j - x_{j-1}}{x_j} = \lambda - A_j - \frac{1}{r_{j-1}} < r_{j-1}. \tag{58}$$

Also, we combine (52), (56), (58) to obtain

$$r_j = \lambda - A_j - \frac{1}{r_{j-1}} > \lambda - A_j - \frac{1}{\sigma_j} = \frac{(\lambda - A_j) \cdot \sigma_j - 1}{\sigma_j} = \sigma_j > \sigma_{j+1}. \tag{59}$$

In other words, (56) implies (59), and we combine this observation with (55) to obtain

$$r_1 > \sigma_2, \quad r_2 > \sigma_3, \quad \ldots, \quad r_{k-1} > \sigma_k. \tag{60}$$

Also, due to (58),

$$r_1 > r_2 > \cdots > r_{k-1}. \tag{61}$$

We combine (51), (52), (60), (61) to obtain (48), (49). ■

Corollary 1. *Under the assumptions of Theorem 6,*

$$\frac{x_k}{x_1} > \prod_{j=2}^{k} \left(\frac{\lambda - A_j}{2} + \sqrt{\left(\frac{\lambda - A_j}{2} \right)^2 - 1} \right). \tag{62}$$

Remark 10. *In [17], the derivation of an upper bound on the first coordinate of an eigenvector of a certain matrix is based on a generalization of Theorem 6.*

In the following theorem, we study the behavior of several last elements of an eigenvector of the matrix A from Theorem 5 above.

Theorem 7. *Suppose that the n by n symmetric tridiagonal matrix A is defined via* (14) *in Section 3.1. 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*

$$\lambda \leq A_k - 2. \tag{63}$$

Then,

$$0 < |x_n| < |x_{n-1}| < \cdots < |x_k| < |x_{k-1}|. \tag{64}$$

Also,

$$-\frac{x_j}{x_{j+1}} > \frac{A_j - \lambda}{2} + \sqrt{\left(\frac{\lambda - A_j}{2}\right)^2 - 1}, \tag{65}$$

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

$$-1 > \frac{x_k}{x_{k+1}} > \cdots > \frac{x_{n-2}}{x_{n-1}} > \frac{x_{n-1}}{x_n}. \tag{66}$$

Proof. The proof is essentially identical to that of Theorem 6 above and will be omitted. ■

In the rest of this subsection, we investigate the behavior of the "middle" elements of an eigenvector of the matrix A from Theorems 5, 6, 7 above. We start with the following theorem.

Theorem 8. *Suppose that* $k, m > 0$ *are integers, that* $x_k, \ldots, x_{k+m+2}$ *are real numbers, that* $B_{k+1}, \ldots, B_{k+m+1}$ *are real numbers, that*

$$2 > B_{k+1} > \cdots > B_{k+m+1} \geq 0, \tag{67}$$

and that

$$x_{j+1} = B_j \cdot x_j - x_{j-1}, \tag{68}$$

for every $j = k+1, \ldots, k+m+1$. *Suppose also that, for any real number* $0 < \theta \leq \pi/2$, *the real* 2×2 *matrix* $A(\theta)$ *is defined via the formula*

$$A(\theta) = \begin{pmatrix} 0 & 1 \\ -1 & 2 \cdot \cos(\theta) \end{pmatrix}. \tag{69}$$

Then,

$$\begin{pmatrix} x_{j+1} \\ x_{j+2} \end{pmatrix} = A\left(\arccos\left(\frac{B_{j+1}}{2}\right)\right) \cdot \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix}, \tag{70}$$

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

Proof. The identity (70) follows from the combination of (68) and (69). ■

Theorem 9. *Suppose that* $k > 0$ *and* $l > 0$ *are integers, and that*

$$0 < \theta_k < \theta_{k+1} < \cdots < \theta_{k+l-1} \leq \frac{\pi}{4} \cdot \frac{1}{l + 3/2} \tag{71}$$

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

$$x_k = 1, \quad x_{k+1} = 1 + \varepsilon, \tag{72}$$

and

$$\begin{pmatrix} x_{j+1} \\ x_{j+2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta_j) \end{pmatrix} \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix}, \tag{73}$$

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

$$x_k, x_{k+1}, \ldots, x_{k+l}, x_{k+l+1} > 0. \tag{74}$$

In addition,

$$\frac{m+1}{m} > \frac{x_{k+m+1}}{x_{k+m}} > \frac{\cos\left((m+1/2)\cdot\theta_{k+l-1}\right)}{\cos\left((m-1/2)\cdot\theta_{k+l-1}\right)}, \tag{75}$$

for every integer $m = 1, 2, \ldots, l$; *in particular,*

$$1 + \frac{1}{l} > \frac{x_{k+l+1}}{x_{k+l}} > 1 - \frac{1}{l+3/2}. \tag{76}$$

Proof. We observe that

$$\frac{x_{k+m+1}}{x_{k+m}} = 2\cdot\cos(\theta_{k+m-1}) - \frac{x_{k+m-1}}{x_{k+m}}, \tag{77}$$

for every $m = 1, \ldots, l$. We use (77) to prove (75) by induction on m. For $m = 1$,

$$\frac{x_{k+2}}{x_{k+1}} = 2\cdot\cos(\theta_k) - \frac{1}{1+\varepsilon} < 2, \tag{78}$$

and also

$$\begin{aligned} \frac{\cos(3\cdot\theta_{k+l-1}/2)}{\cos(\theta_{k+l-1}/2)} &= 4\cdot\cos(\theta_{k+l-1}/2) - 3 = 2\cdot\cos(\theta_{k+l-1}) - 1 \\ &< 2\cdot\cos(\theta_k) - 1 < \frac{x_{k+2}}{x_{k+1}}. \end{aligned} \tag{79}$$

By induction, for $2 \le m \le l$,

$$\frac{x_{k+m+1}}{x_{k+m}} < 2\cdot\cos(\theta_{k+m-1}) - \frac{m-1}{m} < 2 - \frac{m-1}{m} = \frac{m+1}{m}, \tag{80}$$

which proves the left-hand side of (75), and also

$$\frac{x_{k+m+1}}{x_{k+m}} > 2\cdot\cos(\theta_{k+m-1}) - \frac{\cos(\theta_{k+l-1}\cdot(m-3/2))}{\cos(\theta_{k+l-1}\cdot(m-1/2))}. \tag{81}$$

However, for any real θ,

$$\frac{\cos(\theta\cdot(m-3/2))}{\cos(\theta\cdot(m-1/2))} + \frac{\cos(\theta\cdot(m+1/2))}{\cos(\theta\cdot(m-1/2))} = 2\cdot\cos(\theta), \tag{82}$$

and we combine (81), (82) to conclude the right-hand side of (75). The inequality (75) implies (74). Next, we observe that

$$\cos(x) - \sin(x) \geq 1 - \frac{4x}{\pi}, \tag{83}$$

for all real $0 \leq x \leq \pi/4$, and combine (83) with (71) to obtain

$$\begin{aligned} \frac{\cos\left((l+1/2)\cdot\theta_{k+l-1}\right)}{\cos\left((l-1/2)\cdot\theta_{k+l-1}\right)} &= \cos\left(\theta_{k+l-1}\right) - \sin\left(\theta_{k+l-1}\right)\cdot\tan\left(\theta_{k+l-1}\cdot(l-1/2)\right) \\ &> \cos\left(\theta_{k+l-1}\right) - \sin\left(\theta_{k+l-1}\right) \\ &> 1 - \frac{4}{\pi}\cdot\frac{\pi}{4}\cdot\frac{1}{l+3/2}. \end{aligned} \tag{84}$$

Finally, we combine (84) with (75) to obtain (76). ■

Corollary 2. *If, in addition to* (71)*,*

$$\left(m + \frac{3}{2}\right)\cdot\theta_{k+m-1} < \frac{\pi}{4} \tag{85}$$

for every $m = 1, \ldots, l$*, then*

$$1 + \frac{1}{m} > \frac{x_{k+m+1}}{x_{k+m}} > 1 - \frac{1}{m+3/2}, \tag{86}$$

for every $m = 1, \ldots, l$*.*

Remark 11. *One can prove (along the lines of Theorem 9) that* $x_{j+1} > x_j$ *for every* $j = k, \ldots, k+l$*, provided that* $l < k$ *and that* $\varepsilon > k^{-1}$*.*

Theorem 10. *Suppose that* $m > 0$ *is an integer, and* $\theta_1, \ldots, \theta_m$ *are real numbers such that*

$$0 < \theta_1 < \cdots < \theta_m \leq \frac{\pi}{2}. \tag{87}$$

Suppose also that, for any real number $0 < \theta \leq \pi/2$*, the real* 2×2 *matrix* $A(\theta)$ *is defined via* (69)*, and the complex* 2×2 *matrices* $D(\theta), \Lambda(\theta)$ *are defined, respectively, via the formulae*

$$D(\theta) = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}, \tag{88}$$

$$\Lambda(\theta) = \begin{pmatrix} -2\cdot i\cdot\sin(\theta/2) & 0 \\ 0 & 2\cos(\theta/2) \end{pmatrix}. \tag{89}$$

Suppose furthermore that, for any real numbers $0 < \eta_1, \eta_2 \leq \pi/2$*, the complex* 2×2 *matrix* $D(\eta_1, \eta_2)$ *is defined via the formula*

$$D(\eta_1, \eta_2) = \begin{pmatrix} \sin(\eta_1/2)/\sin(\eta_2/2) & 0 \\ 0 & \cos(\eta_1/2)/\cos(\eta_2/2) \end{pmatrix}, \tag{90}$$

and that the unitary complex 2×2 *matrix* V *is defined via the formula*

$$V = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{91}$$

Then,

$$\begin{aligned} & A(\theta_m) \cdot \cdots \cdot A(\theta_1) = \\ & V \cdot \Lambda(\theta_m) \cdot V \cdot \\ & \quad D(\theta_m) \cdot V \cdot D(\theta_{m-1}, \theta_m) \cdot V \cdot \\ & \quad D(\theta_{m-1}) \cdot V \cdot D(\theta_{m-2}, \theta_{m-1}) \cdot V \cdot \\ & \quad \cdots \\ & \quad D(\theta_2) \cdot V \cdot D(\theta_1, \theta_2) \cdot V \cdot \\ & \quad D(\theta_1) \cdot V \cdot \Lambda^{-1}(\theta_1) \cdot V. \end{aligned} \tag{92}$$

Proof. Suppose that, for any real number $0 < \theta \leq \pi/2$, the complex 2×2 matrix $U(\theta)$ is defined via the formula

$$U(\theta) = \begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix}. \tag{93}$$

Obviously, $U(\theta)$ admits the decomposition

$$U(\theta) = e^{i \cdot \theta/2} \cdot V \cdot \Lambda(\theta) \cdot V. \tag{94}$$

Due to (94), the inverse of $U(\theta)$ admits the decomposition

$$U(\theta)^{-1} = e^{-i \cdot \theta/2} \cdot V \cdot \Lambda^{-1}(\theta) \cdot V. \tag{95}$$

Due to the combination of (94), (95),

$$\begin{aligned} U(\theta_2)^{-1} \cdot U(\theta_1) &= e^{i(\theta_1 - \theta_2)/2} \cdot V \cdot \Lambda^{-1}(\theta_2) \cdot \Lambda(\theta_1) \cdot V \\ &= e^{i(\theta_1 - \theta_2)/2} \cdot V \cdot D(\theta_1, \theta_2) \cdot V. \end{aligned} \tag{96}$$

We observe that, for any $0 < \theta < \pi$,

$$\frac{i}{2\sin(\theta)} \cdot \begin{pmatrix} e^{-i\theta} & -1 \\ -1 & e^{-i\theta} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 2\cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & e^{i\theta} \\ e^{i\theta} & 1 \end{pmatrix} = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}, \tag{97}$$

and combine (97) with (94), (88), (69) to conclude that

$$A(\theta) = U(\theta) \cdot D(\theta) \cdot U^{-1}(\theta). \tag{98}$$

Subsequently, due to the combination of (94), (95), (96), (98),

$$\begin{aligned} A(\theta_2) \cdot A(\theta_1) &= U(\theta_2) \cdot D(\theta_2) \cdot U^{-1}(\theta_2) \cdot U(\theta_1) \cdot D(\theta_1) \cdot U^{-1}(\theta_1) \\ &= V \cdot \Lambda(\theta_2) \cdot V \cdot D(\theta_2) \cdot V \cdot D(\theta_1, \theta_2) \cdot V \cdot D(\theta_1) \cdot V \cdot \Lambda^{-1}(\theta_1) \cdot V. \end{aligned} \tag{99}$$

Now (92) follows from (99). ■

Corollary 3. *Suppose that, for any complex square matrix A, we denote by $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$, respectively, the minimal and maximal singular values of A. Then, under the assumptions of Theorem 10 above,*

$$\sigma_{\min}(A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot V \cdot \Lambda(\theta_1)) \geq 2 \cdot \sin\left(\frac{\theta_1}{2}\right), \tag{100}$$

$$\sigma_{\max}(A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot V \cdot \Lambda(\theta_1)) \leq 2 \cdot \cos\left(\frac{\theta_1}{2}\right), \tag{101}$$

and also

$$\sigma_{\min}(A(\theta_m) \cdot \cdots \cdot A(\theta_1)) \geq \tan\left(\frac{\theta_1}{2}\right), \tag{102}$$

$$\sigma_{\max}(A(\theta_m) \cdot \cdots \cdot A(\theta_1)) \leq \cot\left(\frac{\theta_1}{2}\right). \tag{103}$$

Theorem 11. *Suppose, in addition to the hypothesis of Theorem 10, that $\delta > 0$ is a real number, and that the vector $x \in \mathbb{R}^2$ is defined via the formula*

$$x = \begin{pmatrix} 1 \\ 1+\delta \end{pmatrix}. \tag{104}$$

Then,

$$\frac{\min\left\{|A(\theta_j) \cdot \cdots \cdot A(\theta_1) \cdot x| \ : \ 1 \leq j \leq m\right\}}{\max\left\{|A(\theta_j) \cdot \cdots \cdot A(\theta_1) \cdot x| \ : \ 1 \leq j \leq m\right\}} \geq \frac{\theta_1}{2}, \tag{105}$$

and also,

$$\frac{|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x|}{|x|} \ \leq 1 + \frac{1}{2} \cdot \frac{(4/\theta_1^2 - 1) \cdot \delta^2}{(2+\delta)^2 + \delta^2}. \tag{106}$$

Proof. Due to the combination of (89), (91) and (104),

$$\Lambda^{-1}(\theta_1) \cdot V \cdot x = \frac{1}{2\sqrt{2}} \cdot \begin{pmatrix} i \cdot \delta / \sin(\theta_1/2) \\ (2+\delta) / \cos(\theta_1/2) \end{pmatrix}. \tag{107}$$

We combine (107) with (92) and (101) to conclude that

$$|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x| \leq \frac{1}{\sqrt{2}} \left| \begin{pmatrix} \delta \cdot \cot(\theta_1/2) \\ 2+\delta \end{pmatrix} \right| \leq \frac{1}{\sqrt{2}} \left| \begin{pmatrix} 2 \cdot \delta/\theta_1 \\ 2+\delta \end{pmatrix} \right|. \tag{108}$$

and

$$|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x| \geq \frac{1}{\sqrt{2}} \left| \begin{pmatrix} \delta \\ (2+\delta) \cdot \tan(\theta_1/2) \end{pmatrix} \right| \geq \frac{1}{\sqrt{2}} \left| \begin{pmatrix} \delta \\ (2+\delta) \cdot \theta_1/2 \end{pmatrix} \right|. \tag{109}$$

It follows from (108) that

$$|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x|^2 \leq \frac{1}{2} \cdot \left((2+\delta)^2 + \left(\frac{2 \cdot \delta}{\theta_1} \right)^2 \right). \tag{110}$$

Also, it follows from (109) that

$$|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x|^2 \geq \frac{1}{2} \cdot \left((2+\delta)^2 + \left(\frac{2 \cdot \delta}{\theta_1} \right)^2 \right) \cdot \frac{\theta_1^2}{4}. \tag{111}$$

Now (105) follows from the combination of (110) and (111). Next we observe that, due to (104),

$$|x|^2 = (1+\delta)^2 + 1 = \frac{1}{2} \cdot \left((2+\delta)^2 + \delta^2 \right). \tag{112}$$

We combine (110) with (112) to conclude that

$$\frac{|A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x|}{|x|} \leq \sqrt{1 + \frac{(4/\theta_1^2 - 1) \cdot \delta^2}{(2+\delta)^2 + \delta^2}}, \tag{113}$$

which implies (106). ■

Corollary 4. *Suppose, in addition to the hypotheses of Theorem 11, that* $l \geq 1$ *is an integer, that*

$$\theta_1 \cdot \left(l + \frac{5}{2} \right) \geq \frac{\pi}{4} \tag{114}$$

(compare to (71)*), and that*

$$-\frac{1}{l + 3/2} < \delta < \frac{1}{l} \tag{115}$$

(see (76)*). Then,*

$$\frac{|x|}{9 \cdot l} < |A(\theta_m) \cdot \cdots \cdot A(\theta_1) \cdot x| < 4 \cdot |x|. \tag{116}$$

Proof. The right inequality in (116) follows from the combination of (106), (114), (115). The left inequality in (116) follows from the combination of (114) and (105). ■

4.2 Error Analysis

In Section 4.1 above, we investigated various analytical properties of eigenvectors of certain tridiagonal symmetric matrices. This section deals with stability issues pertaining to the numerical evaluation of such eigenvectors.

The following theorem is closely related to Theorem 6 in Section 4.1.

Theorem 12. *Suppose that $k > 2$ is an integer, and that*

$$B_1 > B_2 > \cdots > B_k \geq 2 \tag{117}$$

are real numbers. Suppose also that $x_1, \ldots, x_{k+1}$ are real numbers defined via the recurrence relation

$$\begin{aligned} x_1 &= 1, \\ x_2 &= B_1, \\ x_{j+1} &= B_j \cdot x_j - x_{j-1}, \end{aligned} \tag{118}$$

for $j \geq 2$, and that the real numbers $r_1, \ldots, r_k$ are defined via the formula

$$r_j = \frac{x_{j+1}}{x_j}, \tag{119}$$

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

$$r_j = B_j - \frac{1}{r_{j-1}}, \tag{120}$$

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

Proof. The recurrence relation (120) follows from the combination of (118), (119). ■

Theorem 13. *Suppose that $k > 2$ is an integer, and that the real numbers $B_1, \ldots, B_k$, $x_1, \ldots, x_{k+1}$, $r_1, \ldots, r_k$ are those of Theorem 12 above. Suppose also that $\varepsilon > 0$ is the machine precision, that $B_1, \ldots, B_k$ are defined to machine precision, and that $x_1, \ldots, x_{k+1}$, $r_1, \ldots, r_k$ are calculated, respectively, via* (118), (120). *Then,*

$$\mathrm{rel}(r_j) \leq (2 \cdot j - 1) \cdot \varepsilon, \tag{121}$$

for every $j = 1, \ldots, k$,

$$\mathrm{rel}(x_{j+1}) \leq \varepsilon \cdot j^2, \tag{122}$$

for every $j = 1, \ldots, k$, and also

$$\mathrm{rel}(x_1^2 + x_2^2 + \cdots + x_k^2 + x_{k+1}^2) \leq \varepsilon \cdot 2 \cdot k^2. \tag{123}$$

Proof. First, suppose that $\varepsilon_1, \ldots, \varepsilon_k$ and $\delta_1, \ldots, \delta_k$ are real numbers, that

$$|\delta_{j-1}| \leq \varepsilon, \tag{124}$$

for every $j = 2, \ldots, k$, that

$$\hat{r}_{j-1} = r_{j-1} \cdot (1 + \varepsilon_{j-1}), \tag{125}$$

for every $j = 2, \ldots, k$, that

$$\hat{B}_{j-1} = B_{j-1} \cdot (1 + \delta_{j-1}), \tag{126}$$

for every $j = 2, \ldots, k$, and that

$$\hat{r}_j = \hat{B}_j - \frac{1}{\hat{r}_{j-1}}, \tag{127}$$

for every $j = 2, \ldots, k$. Then, due to the combination of (125), (127), (120),

$$\begin{aligned} \hat{r}_j &= \hat{B}_j - \frac{1}{r_{j-1}} + \frac{1}{r_{j-1}} - \frac{1}{\hat{r}_{j-1}} \\ &= r_j \cdot \left(1 + \frac{\varepsilon_{j-1}}{r_{j-1} \cdot r_j \cdot (1 + \varepsilon_{j-1})} + \frac{B_j \cdot \delta_j}{r_j}\right). \end{aligned} \tag{128}$$

Also, due to Theorem 6 in Section 4.1,

$$B_1 = r_1 > r_2 > \cdots > r_k > 1, \tag{129}$$

and, moreover, for every $j = 1, \ldots, k$,

$$\frac{B_j}{r_j} < 2. \tag{130}$$

We combine (124), (128), (129), (130) to conclude (121). Next, due to (119),

$$x_{j+1} = r_1 \cdot r_2 \cdot r_3 \cdot \cdots \cdot r_j, \tag{131}$$

and we combine (131) with (121) to obtain

$$\mathrm{rel}(x_{j+1}) \leq \varepsilon \cdot (1 + 3 + \cdots + 2j - 1), \tag{132}$$

for every $j = 1, \ldots, k-1$, which implies (122). Finally, due to (122),

$$\begin{aligned} \mathrm{rel}(x_1^2 + \cdots + x_{k+1}^2) &\leq \frac{\sum_{j=1}^{k} x_{j+1}^2 \cdot (1 + \varepsilon \cdot j^2)^2 - (x_1^2 + \cdots + x_{k+1}^2)}{x_1^2 + \cdots + x_{k+1}^2} \\ &= \varepsilon \cdot \frac{\sum_{j=1}^{k} x_{j+1}^2 \cdot (2 \cdot j^2 + \varepsilon \cdot j^4)}{x_1^2 + \cdots + x_{k+1}^2}, \end{aligned} \tag{133}$$

which implies (123). ■

Theorem 14. *Suppose that $k > 0$ and $l > 0$ are integers, that*

$$0 < \theta_k < \theta_{k+1} < \cdots < \theta_{k+l-1} < \frac{\pi}{4} \cdot \frac{1}{l + 3/2} \tag{134}$$

are real numbers, and that the real numbers $B_{k+1}, \ldots, B_{k+l}$ are defined via the formula

$$B_{j+1} = 2 \cdot \cos(\theta_j), \tag{135}$$

for every $j = k, \ldots, k + l - 1$. Suppose also that $\varepsilon > 0$, that the real numbers x_k, x_{k+1} are those of Theorem 12 above, that the sequence $x_{k+2}, \ldots, x_{k+l+1}$ is defined via the formula

$$x_{j+2} = B_{j+1} \cdot x_{j+1} - x_j, \tag{136}$$

for every $j = k, \ldots, k + l - 1$, and that the real numbers $r_k, \ldots, r_{k+l}$ are defined via (119) *for every $j = k, \ldots, k + l$. Suppose furthermore that $\varepsilon > 0$ is the machine precision, that $B_{k+1}, \ldots, B_{k+l}$ are defined to precision ε, and that the precision of r_k, x_k, x_{k+1} is described in* (121), (122) *of Theorem 13 above. Then,*

$$\mathrm{rel}(x_{k+m+1}) < \varepsilon \cdot (k + 2 \cdot m)^2, \tag{137}$$

for every $m = 1, \ldots, l$. Also,

$$\mathrm{rel}(x_{k+2}^2 + \cdots + x_{k+l}^2 + x_{k+l+1}^2) < 2 \cdot \varepsilon \cdot (k + 2 \cdot l)^2. \tag{138}$$

In addition,

$$\mathrm{rel}(r_{k+l}) < 4 \cdot (k + l) \cdot \varepsilon. \tag{139}$$

Proof. Suppose that the real numbers $C_1, \ldots, C_l$ are defined via the formula

$$C_j = \frac{\cos\left((j - 1/2) \cdot \theta_{k+l-1}\right)}{\cos\left((j + 1/2) \cdot \theta_{k+l-1}\right)}, \tag{140}$$

for every $j = 1, \ldots, l$. Then, due to (75),

$$\frac{1}{r_{k+j}} = \frac{x_{k+j}}{x_{k+j+1}} < C_j, \tag{141}$$

for every $j = 1, \ldots, l$. It follows from (141) that

$$\frac{1}{r_{k+1} \cdot \cdots \cdot r_{k+m-1}} < \frac{1}{\cos\left((m - 1/2) \cdot \theta_{k+l-1}\right)}, \tag{142}$$

for every $m = 2, \ldots, l$. Therefore,

$$\begin{aligned} \frac{1}{r_k \cdot r_{k+1}^2 \cdot \cdots \cdot r_{k+m-1}^2 \cdot r_{k+m}} &< \frac{C_m}{\cos^2\left((m - 1/2) \cdot \theta_{k+l-1}\right)} \\ &< \frac{1}{\cos^2\left((m + 1/2) \cdot \theta_{k+l-1}\right)}, \end{aligned} \tag{143}$$

for every $m = 2, \ldots, l$. We observe that, similar to (120),

$$\frac{B_{k+m}}{r_{k+m}} = 1 + \frac{1}{r_{k+m} \cdot r_{k+m-1}}, \tag{144}$$

for every $m = 1, \ldots, l$. Suppose that for every $j = k, k+1, \ldots, k+l$ the relative errors of r_j, B_j are denoted, respectively, by ε_j, δ_j (similar to (125), (126)). Due to the combination of (47), (120), (144),

$$\hat{r}_{k+m} = r_{k+m} \cdot (1 + \frac{\varepsilon_{k+m-1}}{r_{k+m-1} \cdot r_{k+m} \cdot (1 + \varepsilon_{k+m-1})} + \delta_{k+m} \cdot \left(1 + \frac{1}{r_{k+m-1} \cdot r_{k+m}}\right)), \tag{145}$$

for every $m = 1, \ldots, l$. In particular, using (141),

$$\varepsilon_{k+1} \leq \varepsilon_k \cdot C_1 + \varepsilon \cdot (1 + C_1) \leq (\varepsilon_k + 2\varepsilon) \cdot C_1, \tag{146}$$

and, more generally,

$$\varepsilon_{k+m} < (\varepsilon_k + 2 \cdot m \cdot \varepsilon) \cdot C_1^2 \cdot C_2^2 \cdot \ldots C_{k+m-1}^2 \cdot C_{k+m}, \tag{147}$$

for every $m = 1, \ldots, l$. Next, we combine (147) with (143) and Theorem 6 in Section 4.1 to conclude that

$$\varepsilon_{k+m} < (\varepsilon_k + 2 \cdot m \cdot \varepsilon) \cdot \cos^{-2}((m + 1/2) \cdot \theta_{k+l-1}), \tag{148}$$

for every $m = 1, \ldots, l$. We substitute (134) into (148) to obtain the inequality

$$\varepsilon_{k+m} < (\varepsilon_k + 2 \cdot m \cdot \varepsilon) \cdot \cos^{-2}\left(\frac{\pi}{4} \cdot \frac{2m+1}{2l+3}\right) < 2 \cdot (\varepsilon_k + 2 \cdot m \cdot \varepsilon), \tag{149}$$

for every $m = 1, \ldots, l$. In particular, for $m = l$,

$$\varepsilon_{k+l} < 2 \cdot (\varepsilon_k + 2 \cdot l \cdot \varepsilon). \tag{150}$$

It follows from (149) that

$$\varepsilon_{k+1} + \cdots + \varepsilon_{k+m} < 2 \cdot m \cdot \varepsilon_k + 2 \cdot m \cdot (m+1) \cdot \varepsilon, \tag{151}$$

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

$$x_{k+m+1} = x_{k+1} \cdot r_{k+1} \cdot \cdots \cdot r_{k+m}, \tag{152}$$

for every $m > 1$, and hence (ignoring the $O(\varepsilon^2)$ terms)

$$\mathrm{rel}(x_{k+m+1}) < \mathrm{rel}(x_{k+1}) + 2 \cdot m \cdot \varepsilon_k + 2 \cdot m \cdot (m+1) \cdot \varepsilon, \tag{153}$$

for every $m = 1, 2, \ldots, l$. We combine (153) with Theorem 13 above to obtain (137), (138), and combine Theorem 13 with (150) to obtain (139). ■

Theorem 15. *Suppose that $k > 0$ and $0 < l < m$ are integers, that $\theta_{k+l}, \ldots, \theta_{k+m}$ are real numbers such that*

$$0 < \frac{\pi}{2 \cdot (2 \cdot l + 5)} \leq \theta_{k+l} < \cdots < \theta_{k+m} \leq \frac{\pi}{2}, \tag{154}$$

that $x_{k+l}, \ldots, x_{k+m+2}$ *are real numbers, that* x_{k+l}, x_{k+l+1} *satisfy* (76)*, and that* $v_{k+l}, \ldots, v_{k+m+1}$ *are vectors in* $\mathbb{R}^2$ *defined via the formula*

$$v_j = \begin{pmatrix} x_j \\ x_{j+1} \end{pmatrix} \tag{155}$$

for every $j = k+l, \ldots, k+m+1$*. Suppose also that the real* 2×2 *matrices* $A(\theta_{k+l}), \ldots, A(\theta_{k+m+1})$ *are defined via* (69)*, and that*

$$v_{j+1} = A(\theta_j) \cdot v_j \tag{156}$$

for every $j = k+l, \ldots, k+m$*. Suppose, in addition, that* $\varepsilon > 0$ *is the machine precision, that* $\cos(\theta_{k+j})$ *are defined to relative precision* ε *for every* $j = l, \ldots, m$*, and that* $v_{k+l+1}, \ldots, v_{k+m}$ *are evaluated recursively via* (156)*. Then,*

$$\operatorname{rel}(v_j) \leq 9 \cdot l \cdot \operatorname{rel}(v_{k+l}) \cdot \frac{\|v_{k+l}\|}{\|v_j\|}, \tag{157}$$

for every $j = k+l+1, \ldots, k+m+1$*. Also,*

$$\operatorname{rel}(v_j) \leq 81 \cdot l^2 \cdot \operatorname{rel}(v_{k+l}), \tag{158}$$

for every $j = k+l+1, \ldots, k+m+1$*. Finally,*

$$\operatorname{rel}\left(x_{k+l}^2 + 2 \cdot \left(x_{k+l+1}^2 + \cdots + x_{k+m+1}^2\right) + x_{k+m+2}^2\right) \leq 162 \cdot l^2 \cdot \operatorname{rel}(v_{k+l}). \tag{159}$$

Proof. Due to the combination of (154), (156) with (103) and (76),

$$\begin{aligned} \operatorname{rel}(v_j) \cdot \|v_j\| \ &\leq \cot\left(\frac{\theta_{k+l}}{2}\right) \cdot \|v_{k+l}\| \cdot \operatorname{rel}(v_{k+l}) \leq \frac{2}{\theta_{k+l}} \cdot \|v_{k+l}\| \cdot \operatorname{rel}(v_{k+l}) \\ &\leq 9 \cdot l \cdot \|v_{k+l}\| \cdot \operatorname{rel}(v_{k+l}), \end{aligned} \tag{160}$$

for every $j = k+l+1, \ldots, k+m+1$, which implies (157). The combination of (157) and (116) implies (158).

Thus, ignoring the $O(\varepsilon^2)$ terms,

$$\operatorname{rel}(\|v_j\|^2) = \operatorname{rel}(v_j \cdot v_j) \leq 2 \cdot \operatorname{rel}(v_j) \leq 18 \cdot l \cdot \operatorname{rel}(v_{k+l}) \cdot \frac{\|v_{k+l}\|}{\|v_j\|}, \tag{161}$$

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

$$\begin{aligned} &\operatorname{rel}(\|v_{k+l}\|^2 + \cdots + \|v_{k+m+1}\|^2) \leq \\ &18 \cdot l \cdot \operatorname{rel}(v_{k+l}) \cdot \|v_{k+l}\| \cdot \frac{\|v_{k+l}\| + \cdots + \|v_{k+m+1}\|}{\|v_{k+l}\|^2 + \cdots + \|v_{k+m+1}\|^2}. \end{aligned} \tag{162}$$

We substitute (116) into (162) to obtain

$$\operatorname{rel}(\|v_{k+l}\|^2 + \cdots + \|v_{k+m+1}\|^2) \leq 18 \cdot 9 \cdot l^2 \cdot \operatorname{rel}(v_{k+l}) \cdot \|v_{k+l}\|, \tag{163}$$

and substitute (155) into (163) to obtain (159). ∎

Corollary 5. *Suppose, in addition to the hypothesis of Theorem 15, that the relative accuracy of x_{k+l} satisfies* (137) *in Theorem 14. Then,*

$$\operatorname{rel}\left(x_{k+l}^2 + \cdots + x_{k+m+2}^2\right) \leq 162 \cdot l^2 \cdot (k + 2 \cdot l)^2 \cdot \varepsilon. \tag{164}$$

Proof. We observe that

$$\operatorname{rel}(x_{k+m+1}^2 + 2 \cdot x_{k+m+2}^2) \leq 2 \cdot \operatorname{rel}\left((x_{k+m+1}, x_{k+m+2})^T\right), \tag{165}$$

and combine this observation with (137), (159) to obtain (164). ■

In the following two theorems, we summarize Theorems 12, 13, 14, 15 and Corollary 5 above.

Theorem 16. *Suppose that $k > 0$, $l > 0$ and $r > k + l$ are integers, that $B_1, \ldots, B_r$ is a sequence of real numbers, that*

$$B_1 > B_2 > \cdots > B_k \geq 2 > B_{k+1} > \cdots > B_{k+l} > 2 \cdot \cos\left(\frac{\pi}{4} \cdot \frac{1}{l + 3/2}\right) \tag{166}$$

and that

$$2 \cdot \cos\left(\frac{\pi}{4} \cdot \frac{1}{l + 5/2}\right) \geq B_{k+l+1} > \cdots > B_r \geq 0. \tag{167}$$

Suppose also that $\varepsilon > 0$ is the machine precision, that $B_1, \ldots, B_r$ are defined to precision ε, and that the real numbers $x_1, x_2, \ldots, x_{r+1}$ are evaluated from $B_1, \ldots, B_r$ via the recurrence relation (118). *Then,*

$$\operatorname{rel}(x_j) \leq (j - 1)^2 \cdot \varepsilon, \tag{168}$$

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

$$\operatorname{rel}(x_{k+1+j}) \leq (k + 2 \cdot j)^2 \cdot \varepsilon, \tag{169}$$

for every $j = 1, \ldots, l$. In addition,

$$\operatorname{rel}\binom{x_{j-1}}{x_j} \leq 81 \cdot l^2 \cdot (k + 2 \cdot l)^2 \cdot \varepsilon, \tag{170}$$

$$\operatorname{rel}(x_j) \leq 18 \cdot l \cdot (k + 2 \cdot l)^2 \cdot \left|\frac{x_{k+l}}{x_j}\right| \cdot \varepsilon \tag{171}$$

for every $j = k + l + 1, \ldots, r + 1$. Finally,

$$\operatorname{rel}\left(x_1^2 + \cdots + x_r^2 + x_{r+1}^2\right) \leq 162 \cdot l^2 \cdot (k + 2 \cdot l)^2 \cdot \varepsilon. \tag{172}$$

Proof. The combination of Theorems 13, 14, 15 and Corollary 5 above. ■

Theorem 17. *Suppose that $n > 0$ and $r, p, q > 0$ are integers, that*

$$r + p + q + 1 \leq n, \tag{173}$$

that $B_{r+1}, \ldots, B_n$ is a sequence of real numbers, that

$$B_n < \cdots < B_{n+1-q} \leq -2 < B_{n-q} < \cdots < B_{n+1-p-q} < -2 \cdot \cos\left(\frac{\pi}{4} \cdot \frac{1}{p + 3/2}\right) \tag{174}$$

and that

$$-2 \cdot \cos\left(\frac{\pi}{4} \cdot \frac{1}{p + 5/2}\right) \leq B_{n-q-p} < \cdots < B_{r+1} < 0. \tag{175}$$

Suppose also that $\varepsilon > 0$ is the machine precision, that $B_{r+1}, \ldots, B_n$ are defined to precision ε, and that the real numbers $y_n, y_{n-1}, \ldots, y_{r+1}, y_r$ are evaluated from $B_{r+1}, \ldots, B_n$ via the recurrence relation

$$\begin{aligned} y_n &= 1, \\ y_{n-1} &= B_n, \\ y_{j-1} &= B_j \cdot y_j - y_{j+1}, \end{aligned} \tag{176}$$

for $j < n$ (similar to (118), but the direction is reversed). Then,

$$\operatorname{rel}(y_{n-j}) \leq j^2 \cdot \varepsilon, \tag{177}$$

for every $j = 1, \ldots, q$. Also,

$$\operatorname{rel}(y_{n-q-j}) \leq (q + 2 \cdot j)^2 \cdot \varepsilon, \tag{178}$$

for every $j = 1, \ldots, p$. In addition,

$$\operatorname{rel}\binom{y_{j+1}}{y_j} \leq 81 \cdot p^2 \cdot (q + 2 \cdot p)^2 \cdot \varepsilon, \tag{179}$$

$$\operatorname{rel}(y_j) \leq 18 \cdot l \cdot (k + 2 \cdot l)^2 \cdot \left|\frac{y_{n-p-q}}{y_j}\right| \cdot \varepsilon, \tag{180}$$

for every $j = r, \ldots, n - q - p - 1$. Finally,

$$\operatorname{rel}\left(y_n^2 + \cdots + y_{r+2}^2\right) \leq 162 \cdot p^2 \cdot (q + 2 \cdot p)^2 \cdot \varepsilon. \tag{181}$$

Proof. We define $\tilde{B}_1, \ldots$ and $\tilde{x}_1, \ldots$ via the formula

$$\tilde{B}_j = -B_{n+1-j} \tag{182}$$

and

$$\tilde{x}_j = (-1)^{j+1} \cdot y_{n+1-j}, \tag{183}$$

for $j \geq 1$. Then, due to the combination of (182), (183) with (176),

$$\begin{aligned} \tilde{x_{j+1}} &= (-1)^j \cdot y_{n-j} \\ &= (-1)^j \cdot \left(B_{n-(j-1)} \cdot y_{n-(j-1)} - y_{n-(j-2)}\right) \\ &= (-1)^(j+1) \cdot \left(\tilde{B}_j \cdot \tilde{x}_j \cdot (-1)^{j+1} + \tilde{x}_{j-1} \cdot (-1)^j\right) \\ &= \tilde{B}_j \cdot \tilde{x}_j - \tilde{x}_{j-1}, \end{aligned} \tag{184}$$

for $j \geq 2$. We conclude by combining (184) with (173), (174) and Theorem 16 above. ■

Theorem 18. *Suppose, in addition to hypotheses of Theorems 16, 17, that the $n \times n$ matrix B, defined via the formula*

$$B = \begin{pmatrix} B_1 & 1 & & & & \\ 1 & B_2 & 1 & & & \\ & 1 & B_3 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & B_{n-1} & 1 \\ & & & & 1 & B_n \end{pmatrix}, \tag{185}$$

is singular, that $x_1, \ldots, x_r, x_{r+1}$ are those of Theorem 16, that $y_r, y_{r+1}, \ldots, y_n$ are those of Theorem 17, that the real number s is defined via the formula

$$s = \frac{x_r \cdot y_r + x_{r+1} \cdot y_{r+1}}{|x_r \cdot y_r + x_{r+1} \cdot y_{r+1}|} \cdot \sqrt{\frac{x_r^2 + x_{r+1}^2}{y_r^2 + y_{r+1}^2}}, \tag{186}$$

and that the vector $z = (z_1, \ldots, z_n)^T$ in $\mathbb{R}^n$ is defined via the formula

$$z = \left(x_1, \ldots, x_r, x_{r+1}, s \cdot y_{r+2}, \ldots, s \cdot y_n\right)^T. \tag{187}$$

Then, z is an eigenvector of B corresponding to the zero eigenvalue. Moreover,

$$\mathrm{rel}(s) \leq 81 \cdot \left((q + 2 \cdot p)^2 \cdot p^2 + (k + 2 \cdot l)^2 \cdot l^2\right) \cdot \varepsilon, \tag{188}$$

and

$$\mathrm{rel}(z_1^2 + \cdots + z_n^2) \leq 243 \cdot \left(p^2 \cdot (q + 2 \cdot p)^2 + (k + 2 \cdot l)^2 \cdot l^2\right) \cdot \varepsilon. \tag{189}$$

Proof. Due to Theorem 5 in Section 4.1, $x_1, \ldots, x_{r+1}$ are the first $r+1$ coordinates of an eigenvector of B corresponding to the zero eigenvalue; also, $y_r, \ldots, y_n$ are the last $n+1-r$ coordinates of an eigenvector of B in the same eigenspace. We combine this observation with Theorem 5 and (186) to conclude that z is the eigenvector in the null-space of B whose first coordinate is equal to 1. The inequality (188) follows from the combination of (170) and (179) (in particular, s in (186) is well defined). We combine (188) with (181) to obtain

$$\begin{aligned} &\mathrm{rel}\left(s^2 \cdot \left(y_n^2 + \cdots + y_{r+2}^2\right)\right) \leq \\ &\left(3 \cdot 81 \cdot p^2 \cdot (q + 2 \cdot p)^2 + 81 \cdot (k + 2 \cdot l)^2 \cdot l^2\right) \cdot \varepsilon. \end{aligned} \tag{190}$$

Finally, we combine (190) with (172) to obtain (189). ■

Corollary 6. *Suppose that, in addition to the hypothesis of Theorem 18, the vector $X \in \mathbb{R}^n$ is evaluated from z in* (187) *via the formula*

$$X = (X_1, \ldots, X_n)^T = \frac{z}{\|z\|}. \tag{191}$$

Then,

$$\mathrm{rel}(X_1) \leq 243 \cdot \left(p^2 \cdot (q + 2 \cdot p)^2 + (k + 2 \cdot l)^2 \cdot l^2\right) \cdot \varepsilon, \tag{192}$$

where k, l, p, q, r are those of Theorems 16, 17. More generally,

$$\mathrm{rel}(X_j) \leq \mathrm{rel}(X_1) + \mathrm{rel}(x_j), \tag{193}$$

for every $2 \leq j \leq r + 1$, and

$$\mathrm{rel}(X_j) \leq \mathrm{rel}(X_1) + \mathrm{rel}(y_j) + \mathrm{rel}(s), \tag{194}$$

for every $j = r + 2, \ldots, n$, where the sequences $\{x_j\}$, $\{y_j\}$ and the real number s are those from Theorems 16, 17, 18.

4.3 Asymptotic Error Analysis of a Special Case

The analysis of Section 4.2 (e.g. Theorems 16, 17, 18 and Corollary 6) is carried out for a fairly general class of sequences $\{B_j\}$ (and related matrices B defined via (185)). The resulting upper bounds on relative errors of coordinates of the null-space eigenvector of B depend on the parameters k, l, p, q determined from $\{B_j\}$ via (166), (167), (174), (175) (see e.g. the bounds in (189), (192)).

Despite the fact that these bounds are explicitly defined by B, the relation between the *relative error* of, say, the first coordinate X_1 of an eigenvector of unit norm and the *magnitude* of X_1 is not immediately obvious (see (192)). In this section, this relation is investigated in some detail for a special, but still fairly broad class of matrices B (that also appear in various applications; see e.g. Section 6). First, we need a technical theorem.

Theorem 19. *Suppose that $a \geq 1$ is a real number, that $\delta > 1$ is a real number, that the real number D_a is defined via the formula*

$$D_a = \sqrt{2} \cdot \int_0^{\pi/2} (\sin(\theta))^{1+2/a} \, d\theta, \tag{195}$$

and that the real number $\alpha(a, \delta)$ is the solution of the equation

$$\alpha^2 \cdot ((1 + \alpha)^a - 1) \cdot \delta^2 = \frac{\pi^2}{32} \tag{196}$$

in the unknown α. Then,

$$\frac{2 \cdot \sqrt{2}}{3} \leq D_a = \sqrt{\frac{\pi}{2}} \cdot \frac{\Gamma(1 + 1/a)}{\Gamma(3/2 + 1/a)}, \leq \sqrt{2}, \tag{197}$$

where Γ is the standard Gamma function, and also

$$\alpha(a, \delta) \leq \left(\frac{\pi^2}{32 \cdot a \cdot \delta^2}\right)^{1/3}. \tag{198}$$

Proof. The proof is straightforward, elementary, and will be omitted. ■

The rest of this section is dedicated to asymptotic error analysis pertaining to a certain class of symmetric tridiagonal matrices.

Theorem 20. *Suppose that $a \geq 1$ is a real number, that $\delta > 1$ is a real number, that the real numbers $D_a, \alpha(a,\delta)$ are those of Theorem 19 above. Suppose also that, for any real number $c \geq 1$, the real number $\kappa(c)$ is defined via the formula*

$$\kappa(c) = \delta^{2/(a+2)} \cdot c^{a/(a+2)}, \tag{199}$$

and the sequence $B_1(c), B_2(c), \ldots$ is defined via the formula

$$B_j(c) = 2 + 2 \cdot \left(\frac{\kappa(c)}{c}\right)^a - 2 \cdot \left(\frac{j}{c}\right)^a, \tag{200}$$

for every $j = 1, 2, \ldots$. Suppose also that, for any real number $c \geq 1$, the sequence $x_1(c), x_2(c), \ldots$ is defined from $\{B_j(c)\}$ via (118), and the integers $k = k(c), l = l(c)$ are defined from $\{B_j(c)\}$ via (166), (167). Then,

$$k = k(c) = \kappa(c) \cdot (1 + o(c)), \quad c \to \infty, \tag{201}$$

$$l = l(c) = \alpha(a,\delta) \cdot \kappa(c) \cdot (1 + o(c)), \quad c \to \infty, \tag{202}$$

and also

$$x_1(c) \leq x_k(c) \cdot \exp\left(-D_a \cdot \delta \cdot (1 + o(1))\right), \quad c \to \infty. \tag{203}$$

Proof. In this proof, we omit the dependence of various parameters on c whenever it causes no confusion. First, (201) follows from the combination of (200), (199) and (166). We substitute (200), (201) into (62) to obtain

$$\begin{aligned} \frac{x_k}{x_1} &\geq \prod_{j=2}^{k} \left(1 + \left(\frac{\kappa}{c}\right)^a - \left(\frac{j}{c}\right)^a + \sqrt{\left(1 + \left(\frac{\kappa}{c}\right)^a - \left(\frac{j}{c}\right)^a\right)^2 - 1}\right) \\ &= \prod_{j=2}^{k} \left(1 + \sqrt{2 \cdot \left(\left(\frac{k}{c}\right)^a - \left(\frac{j}{c}\right)^a\right)}\right) \cdot (1 + o(1)), \quad c \to \infty. \end{aligned} \tag{204}$$

We define the real-valued function g via the formula

$$g(x) = 1 + \sqrt{2 \cdot \left(\left(\frac{k}{c}\right)^a - \left(\frac{x}{c}\right)^a\right)}, \tag{205}$$

for real $0 \leq x \leq k$, and combine (200), (199), (204), (205) to obtain

$$\prod_{j=2}^{k} \left(\frac{B_j}{2} + \sqrt{\left(\frac{B_j}{2}\right)^2 - 1}\right) = \exp\left((1 + o(1)) \cdot \int_0^k \log(g(x))\, dx\right), \quad c \to \infty. \tag{206}$$

Since $\log(g(k)) = 0$ due to (205),

$$\int_0^k \log(g(x)) = -\int_0^k x \cdot \frac{d}{dx} \log(g(x))\, dx = -\int_0^k \frac{x \cdot g'(x)}{g(x)}\, dx. \tag{207}$$

We combine (205) and (207) to obtain

$$\int_0^k \log(g(x)) = \frac{a}{\sqrt{2}} \int_0^k \frac{x^a\, dx}{\sqrt{2} \cdot (k^a - x^a) + \sqrt{c^a} \cdot \sqrt{k^a - x^2}}. \tag{208}$$

We perform the changes of variable

$$x^a = k^a \cdot \sin^2(\theta), \tag{209}$$

and substitute (209) into (208) to obtain

$$\int_0^k \log(g(x)) = k \cdot \int_0^{\pi/2} \frac{(\sin(\theta))^{1+2/a}\, d\theta}{\cos(\theta) + \sqrt{c^a/(2 \cdot k^a)}}. \tag{210}$$

Due to the combination of (210) and (195), (199), (201),

$$\int_0^k \log(g(x)) = D_a \cdot \sqrt{\frac{k^{a+2}}{c^a}} \cdot (1 + o(1)), \quad c \to \infty, \tag{211}$$

and we substitute (211) into (206) to obtain

$$\prod_{j=2}^{k} \left(\frac{B_j}{2} + \sqrt{\left(\frac{B_j}{2}\right)^2 - 1} \right) = \exp\left(D_a \cdot \sqrt{\frac{k^{a+2}}{c^a}} \cdot (1 + o(1)) \right), \quad c \to \infty. \tag{212}$$

We combine (212) with (199), (201) to obtain (203). Next, we combine (166), (167), (199), (200) to obtain

$$\frac{(k+l)^a - k^a}{c^a} = \frac{\pi^2}{32 \cdot l^2} \cdot (1 + o(1)), \quad c \to \infty. \tag{213}$$

If

$$k(c) \ll l(c), \quad c \to \infty, \tag{214}$$

then due to (213)

$$l^{a+2} = c^a \cdot \frac{\pi^2}{32} \cdot (1 + o(1)), \quad c \to \infty, \tag{215}$$

in contradiction to the combination of (214) and (166), (167). If, on the other hand,

$$l \ll k, \quad c \to \infty, \tag{216}$$

then due to (213), (201)

$$l^3 = \frac{c^a}{k^{a-1}} \cdot (1 + o(c)) = O\left(c^{a-(a-1)\cdot a/(a+2)}\right) = O\left(c^{3a/(a+2)}\right), \quad c \to \infty, \tag{217}$$

in contradiction to the combination of (216) and (166), (167). Therefore,

$$l(c) = O(k(c)), \quad c \to \infty, \tag{218}$$

and we combine (218) with (199), (201), (213) to obtain (202). ■

The following theorem compliments Theorem 20 above.

Theorem 21. *Suppose that $a \geq 1$ and $\varepsilon > 0$ are real numbers. Suppose also that, for any real number $c \geq 1$, the real numbers $\mu(c), \nu(c), \rho(c)$ are defined via the formulae*

$$\mu(c) = \left(\frac{2^{1/a} \cdot c}{a}\right)^{1/3} \cdot \left(-\frac{3}{4} \cdot \log(\varepsilon)\right)^{2/3}, \tag{219}$$

$$\nu(c) = 2^{1/a} \cdot c + \mu(c), \tag{220}$$

$$\rho(c) = \left(\frac{\pi^2 \cdot 2^{1/a}}{64 \cdot a}\right)^{1/3} \cdot c^{1/3}, \tag{221}$$

and that the integer $n(c)$ is defined via the formula

$$n(c) = \text{floor}(\nu(c)) + 1. \tag{222}$$

Suppose furthermore that, for any real $c \geq 1$, the sequence $B_1(c), B_2(c), \ldots,$ is defined via (200)*, that the integers $q = q(c)$ and $p = p(c)$ are defined from $\{B_j(c)\}$ via* (174)*,* (175)*, and that the sequence $y_1(c), \ldots, y_n(c)$ is defined via* (176)*. Then,*

$$q(c) = \mu(c) \cdot (1 + o(1)), \quad c \to \infty, \tag{223}$$

$$p(c) = \rho(c) \cdot (1 + o(1)), \quad c \to \infty, \tag{224}$$

and also

$$|y_{n(c)}(c)| \leq \varepsilon \cdot |y_{n(c)+1-q(c)}(c)| \cdot (1 + o(1)), \quad c \to \infty. \tag{225}$$

Proof. We observe that, due to (199), (174),

$$2 + 2 \cdot \left(\frac{\kappa(c)}{c}\right)^a - 2 \cdot \left(\frac{n(c) - q(c)}{c}\right)^a = -2 + o(1), \quad c \to \infty, \tag{226}$$

and combine (226), (199), (201), (219), (220), (222) to obtain (223). We combine (221), (222), (223), (174), (175) to obtain

$$\begin{aligned} \frac{(n-q)^a - (n-q-p)^a}{c^a} &= 2\left(1 - \left(1 - \frac{p}{c \cdot 2^{1/a}}\right)^a\right) \cdot (1 + o(1)) \\ &= \frac{\pi^2}{32 \cdot p^2} \cdot (1 + o(1)), \quad c \to \infty. \end{aligned} \tag{227}$$

We combine (227) with (221) to obtain (224). Next, for $j = 1, \ldots, q(c)$,

$$\begin{aligned} B_{n-q+j} &= -2 \cdot \left(1 + \left(\frac{n(c) - q(c)}{c}\right)^a \cdot \left(\left(1 + \frac{j}{n(c) - q(c)}\right)^a - 1\right)\right) \\ &= -2 \cdot \left(1 + \frac{2 \cdot a \cdot j}{2^{1/a} \cdot c}\right) \cdot (1 + o(1)), \quad c \to \infty, \end{aligned} \tag{228}$$

and hence, similar to (206),

$$\prod_{j=1}^{q}\left(\frac{B_{n-q+j}}{2}+\sqrt{\left(\frac{B_{n-q+j}}{2}\right)^2-1}\right)= \\ \exp\left((1+o(1))\cdot\int_0^q\log\left(1+\sqrt{\frac{4\cdot a\cdot x}{2^{1/a}\cdot c}}\right)\,dx\right),\quad c\to\infty. \tag{229}$$

We observe that

$$\int_0^1\log(1+Z\cdot\sqrt{s})\,ds=\frac{2\cdot Z}{3}\cdot(1+o(1)),\quad Z\to 0, \tag{230}$$

and combine (229), (229) and Theorem 7 in Section 4.1 to obtain

$$|y_n|\le|y_{n-q+1}|\cdot\exp\left(-\frac{4}{3}\cdot\sqrt{\frac{a\cdot q^3}{2^{1/a}\cdot c}}\cdot(1+o(1))\right),\quad c\to\infty, \tag{231}$$

and combine (219), (223), (231) to obtain (225). ■

The following theorem is a consequence of Theorems 20, 21 above.

Theorem 22. *Suppose that $\varepsilon>0$ is the machine precision, and that $a\ge 1$ and $1\le\tilde{\delta}<\delta$ are real numbers. Suppose also that, for any real $c\ge 1$, we define $\mu(c)$ via* (219)*, that $n(c)$ is an integer, that*

$$2^{1/a}\cdot c<n(c)<2^{1/a}\cdot c+\mu(c)+1, \tag{232}$$

that the sequence $A_1(c),\ldots,A_{n(c)}(c)$ is defined via the formula

$$A_j(c)=2+2\cdot\left(\frac{j}{c}\right)^a, \tag{233}$$

for every $j=1,\ldots,n(c)$, and the $n(c)\times n(c)$ matrix $A(c)$ is defined from $\{A_j(c)\}$ via (14)*. Suppose also that, for any $c\ge 1$, the real number $\lambda(c)$ is an eigenvalue of $A(c)$, that $\delta(c)$ is a real number, that*

$$1<\tilde{\delta}<\delta(c)<\delta, \tag{234}$$

that

$$\lambda(c)=4+2\cdot\left(\frac{\delta(c)}{c}\right)^{2a/(a+2)}, \tag{235}$$

and that $X(c)=(X_1(c),\ldots,X_n(c))^T$ is the unit-norm $\lambda(c)$-eigenvector of $A(c)$. Suppose furthermore that, for any $c\ge 1$, the quantities $A_j(c)-\lambda(c)$ are defined to precision ε, for any $c\ge 1$ and every $j=1,\ldots,n(c)$. Then,

$$|X_1(c)|<\exp\left(-\tilde{\delta}\cdot D_a\right)\cdot(1+o(1)),\quad c\to\infty, \tag{236}$$

where D_a is defined via (195). *Also, if $a > 1$, then*

$$\mathrm{rel}(X_1(c)) \le 620 \cdot \delta^{(16-4a)/(3a+6)} \cdot c^{4a/(a+2)} \cdot \varepsilon \cdot (1 + o(1)), \quad c \to \infty. \tag{237}$$

If $a = 1$, then

$$\mathrm{rel}(X_1(c)) \le 960 \cdot \left(\frac{\delta^{4/3}}{4} + (-\log \varepsilon)^{4/3} + 1 \right) \cdot c^{4/3} \cdot \varepsilon \cdot (1 + o(1)), \quad c \to \infty. \tag{238}$$

Proof. Suppose that $c \ge 1$, and that k, l, p, q are defined from $A(c)$ via (166), (167), (174), (175), respectively. If $a > 1$, we combine (232), (233), (234), (235) with Theorems 20, 21 above to obtain

$$\begin{aligned}
&243 \cdot l^2 \cdot (k + 2 \cdot l)^2 = \\
&243 \cdot k^4 \cdot \alpha^2 \cdot (1 + 2 \cdot \alpha)^2 < \\
&243 \cdot \left(\frac{\pi^2}{32 \cdot a} \right)^{2/3} \cdot \left(1 + 2 \cdot \left(\frac{\pi^2}{32 \cdot a} \right) \right) \cdot \delta^{8/(a+2)-4/3} \cdot c^{4a/(a+2)} < \\
&620 \cdot \delta^{(4/3)\cdot(4-a)/(a+2)} \cdot c^{4a/(a+2)}.
\end{aligned} \tag{239}$$

and combine (239) with Corollary 6 in Section 4.2 to obtain (237). If $a = 1$, then we combine (232), (233), (234), (235) with Theorems 20, 21 above to obtain

$$\begin{aligned}
&243 \cdot \left(l^2 \cdot (k + 2 \cdot l)^2 + p^2 \cdot (q + 2 \cdot p)^2 \right) \le \\
&243 \cdot c^{4/3} \cdot \left(\frac{\pi^2}{32} \right)^{2/3} \cdot \left(\left(\delta^{2/3} + 2 \cdot \left(\frac{\pi^2}{32} \right)^{1/3} \right)^2 + \left((-3 \cdot \log \varepsilon)^{2/3} + 2 \cdot \left(\frac{\pi^2}{32} \right)^{1/3} \right)^2 \right) \\
&960 \cdot c^{4/3} \cdot \left(\frac{\delta^{4/3}}{4} + (-\log \varepsilon)^{4/3} + 1 \right),
\end{aligned} \tag{240}$$

and combine (240) with Corollary 6 in Section 4.2 to obtain (238). For any $a \ge 1$, the inequality (236) follows now from (203). ■

Remark 12. *The conclusions of Theorem 22 above hold even under a milder assumption that each of $A_j(c)$ and $\lambda(c)$ separately is defined to relative precision ε for every j (and not necessarily their difference). The related analysis (beyond the scope of this paper) is based on Theorems 16, 17 in Section 4.2, and on the observation that when $\lambda(c) \approx A_j(c)$ what matters is the absolute (and not relative) accuracy of $\lambda(c) - A_j(c)$.*

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

Suppose that $n > 0$ is an integer, that $2 < A_1 < A_2 < \dots$ is a sequence of positive real numbers, that the n by n symmetric tridiagonal matrix A is defined via (14), and that the real number λ is an eigenvalue of A.

Task. Evaluate the unit-length eigenvector

$$X = (X_1, \dots, X_n) \in \mathbb{R}^n \tag{241}$$

of A corresponding to λ.

Observation. Due to Theorem 5 in Section 4.1, this problem is equivalent to evaluating the solution to the three-terms recurrence relation (42), (43), (44).

Desired accuracy of the solution. We want the coordinates X_j of X to be evaluated to high *relative* accuracy (as opposed to *absolute* accuracy; see also Section 1).

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 X_1 should be evaluated to any correct digit at all (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 (43) of Theorem 5; more specifically,

$$X_{j-1} + (A_j - \lambda) \cdot X_j + X_{j+1} = 0, \tag{242}$$

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 6 in Section 4.1). 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 7 in Section 4.1). Finally, if $-2 < \lambda - A_j < 2$, then the relation is somewhat more complicated (see, for example, Theorems 9, 10, 11 in Section 4.1).

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 assuming (166), (167), (174), (175).

Observation. We combine (166), (174) with (14) to conclude that

$$2 + A_1 < \lambda < A_n - 2. \tag{243}$$

While the obvious simplification of the algorithm described below will handle any eigenvalue λ of A, in the rest of this section we will assume (243) for the sake of clarity of presentation.

5.2 Informal Description of the Algorithm

This section contains an informal description of an algorithm for the evaluation of $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ (see (241)). On the other hand, Section 5.3 below contains a complete outline of the steps of the algorithm.

Suppose that $1 < r < n$ is an integer, and that

$$A_r \leq \lambda < A_{r+1} \tag{244}$$

(see (167), (175)). 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 recurrence relation (43) of Theorem 5 (see also (242) above).

We set $x_1 = 1$ and use (43) to iteratively evaluate $x_2, \ldots, x_{r+1}$ (e.g. "going forward"). Obviously, we have evaluated the first $r+1$ coordinates of X up to a scaling constant. Next, we set $y_n = 1$ and use (43) to iteratively evaluate $y_{n-1}, y_{n-2}, \ldots, y_r$ (e.g. "going backward"). Again, this gives the last $n-r+1$ coordinates of X up to a *different* scaling constant. The accuracy of both evaluations is investigated in detail in Section 4.2.

The indices of the two sequences overlap at $j = r, r+1$. In exact arithmetic, the planar vectors (x_r, x_{r+1}) and (y_r, y_{r+1}) are linearly dependent (see Theorem 18 in Section 4.2). We "glue the two sequences together" by multiplying $y_r, \ldots, y_n$ through by the correct scaling factor s; in particular, $x_j = s \cdot y_j$ for $j = r, r+1$. The resulting vector z in $\mathbb{R}^n$ is a $\lambda-$eigenvector of A (see Theorem 18). We then normalize it to obtain X.

5.3 Short Description of the Algorithm

Suppose that $n > 0$ is an integer, that the n by n matrix A is that from Section 5.1, that λ is an eigenvalue of A, and that the integer $1 < r < n$ is defined via (244) above.

Step A: evaluation of the left coordinates of X (see (241)).

1. Set $x_1 = 1$.
2. Compute x_2 via (42) of Theorem 5.
3. Compute $x_3, \ldots, x_r, x_{r+1}$ iteratively via (43) of Theorem 5.

Step B: evaluation of the right coordinates of X.

1. Set $y_n = 1$.
2. Compute y_{n-1} via (44) of Theorem 5.
3. Compute $y_{n-2}, \ldots, y_{r+1}, y_r$ iteratively via (43) of Theorem 5.

Step C: glue them together.

1. Compute the real number s via (186) in Theorem 18.
2. Compute the vector $z = (z_1, \ldots, z_n)$ via (187) in Theorem 18.
3. Compute the vector $X = (X_1, \ldots, X_n)$ from z via (191) in Corollary 6.

Observation. The vector $X \in \mathbb{R}^n$ is the unit-norm $\lambda-$eigenvector of A whose first coordinate is positive (see Corollary 6 in Section 4.2).

Running time. Obviously, the running time of this algorithm is $O(n)$ operations, where n is the dimensionality of the matrix.

5.4 Accuracy

In Sections 5.2, 5.3, 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. The accuracy of this procedure is investigated in some detail in Section 4.2 for a general tridiagonal matrix with constant off-diagonal elements and monotone diagonal. More specifically, the *relative* accuracy of various coordinates is described in Theorems 16, 17, 18 and Corollary 6 in Section 4.2. For example, (192) provides a bound on $\mathrm{rel}(X_1)$ in terms of the integers $1 < k, l, p, q < n$ (defined via (166), (167), (174), (175)) and the relative accuracy ε of $\lambda - A_j$ for $j = 1, \ldots, n$ (see also Remark 12 in Section 4.3). We summarize the results of Section 4.2 qualitatively in the following observations (see also Section 7 for related numerical experiments).

Observation 1. For all j such that $\lambda - A_j \geq 2$ (e.g. for $1 \leq j \leq k$ in the notation of Theorem 16 in Section 4.2), the coordinates X_j are evaluated to roughly the same *relative* accuracy, *independent* of how small they are (see e.g. Theorem 13 in Section 4.2 and (168) in Theorem 16). These coordinates form a monotonically increasing sequence (see Theorem 6 in Section 4.1 for an estimate on its growth).

Observation 2. For all j such that $\lambda - A_j \leq -2$ (e.g. for $n - q \leq j \leq n$ in the notation of Theorem 17 in Section 4.2), the coordinates X_j are evaluated to roughly the same *relative* accuracy, *independent* of how small they are (see e.g. (177) in Theorem 17). These coordinates form an alternating sequence, and their absolute values form a monotonically decreasing sequence (see Theorem 7 in Section 4.1 for an estimate on its decay).

Observation 3. For all j such that $\lambda - 2 \leq A_j \leq \lambda + 2$ (e.g. for $k < j < n - q$ in the notation of Theorems 16, 17) in Section 4.2, the coordinates X_j are evaluated to roughly the same *absolute* accuracy (see e.g. (157) in Theorem 15, (170), (171) in Theorem 16, (179), (180) in Theorem 17). These coordinates vary in magnitude in a fairly moderate way and exhibit an oscillatory behavior (see e.g. Theorems 10, 11 and Corollaries 3, 4 in Section 4.1, and also Section 7).

Remark 13. *Extensive numerical experiments seem to indicate that the estimates from Section 4.2 are somewhat pessimistic. In other words, in practice the relative error tends to be smaller than our estimates suggest (see also Section 7).*

Remark 14. *It is somewhat surprising that, according to* (193) *in Corollary 6, 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 6 and Corollary 1 in Section 4.1), it still will be evaluated to reasonable relative precision.*

Remark 15. *When the coordinates of the eigenvector are evaluated via the three-terms recurrence* (43)*, 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 Section 4.2). As expected, the use of this recurrence relation in a "wrong" direction leads to a disastrous loss of accuracy.*

5.5 Related Algorithms

In Section 5.2, 5.3, we presented an algorithm for accurate evaluation of the coordinates of the eigenvector X (see (241) in Section 5.1). In this section, we briefly discuss the accuracy of several classical algorithms for the solution of the same problem.

5.5.1 Inverse Power

The unit-length $\lambda-$eigenvector X of A can be obtained via Inverse Power Method with Shifts (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

$$(\lambda \cdot I - A) \cdot x^{(k+1)} = x^{(k)}, \tag{245}$$

and normalizing the solution. We observe that this method also evaluates λ (even though in Section 5.1 we assume that λ has already been evaluated). On each iteration, we solve the linear system (245) by Gaussian elimination (since A is tridiagonal, each iteration costs $O(n)$ operations; moreover, $O(1)$ iterations are required: see Remark 7 in Section 3.4.1).

The following conjecture about the accuracy of Inverse Power Method is substantiated by extensive numerical experiments (see Section 7).

Conjecture 2. *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 defined to accuracy ε. Suppose also that $\lambda - A_1 > 2$. Suppose furthermore that $K > 0$ is an integer, and that*

$$K > \frac{\log(|X_1|)}{\log(\varepsilon)} + 1, \tag{246}$$

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 to high relative accuracy. More specifically, this relative accuracy is roughly of the same order of magnitude as for the algorithm described in Sections 5.2, 5.3 (see also (256), (262) *below).*

Remark 16. *The inequality* (246) *reflects on the fact that each iteration of Inverse Power Method can reduce the coordinates of the approximation $x^{(k)}$ by a factor of 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.*

5.5.2 Jacobi Rotations

In the view of Section 5.5.1, 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 to high relative precision 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], [21], [22]). 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

$$\frac{\|X - \hat{X}\|}{\|X\|} = \|X - \hat{X}\| \approx \varepsilon, \tag{247}$$

where X is the unit-length eigenvector (see (241)), $\hat{X}$ 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.

$$|X_1 - \hat{X}_1| \approx \varepsilon. \tag{248}$$

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 typically have no correct digits at all (the latest statement is supported by extensive numerical evidence).

5.5.3 Gaussian Elimination

Another possible method to evaluate X would be to solve the linear system

$$(\lambda \cdot I - A) \cdot X = 0, \tag{249}$$

by means of Gaussian Elimination (see, for example, [3], [6], [21], [22]). Unfortunately, this method, in general, fails to evaluate the small coordinates of X with high relative accuracy (see, however, Section 5.5.1, where Gaussian Elimination is used several times, as a step of Inverse Power Method).

6 Applications

In this section, we describe some applications of the algorithm from 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 (19), (20) in Section 3.2. A numerical algorithm for the evaluation of $J_0(x), \ldots, J_m(x)$ for a given real number $x > 0$ and a given integer $m > 0$ is described in Section 3.4.2. In particular, according to Remark 9 in Section 3.4.2, the values $J_0(x), \ldots J_m(x)$ can be obtained as coordinates of the unit length $\lambda-$eigenvector of a certain symmetric tridiagonal matrix $A(x)$ (see (38), (39), (40), (41)).

The matrix $A(x)$ belongs to the class of matrices defined via (14). More specifically, the diagonal entries of $A(x)$ are those of the matrix A of Theorem 22 in Section 4.3, with $a = 1$ and $c = x$ (see (233)).

In other words, the principal algorithm of this paper (see Sections 5.2, 5.3) 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 Theorems 20, 21 in Section 4.3. Obviously, in this case, the algorithm of Sections 5.2, 5.3 is essentially identical to the well-known algorithm described in Section 3.4.2; on the other hand, the analysis of Section 4.2, 4.3 appears to be new (see (238) in Theorem 22 in Section 4.3 and Conjecture 1 in Section 1, as well as Section 7.3 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. A popular numerical algorithm for the evaluation of PSWFs and some associated quantities is based on computing unit-length eigenvectors of certain symmetric tridiagonal matrices (see e.g. Theorem 4, Remark 6 in Section 3.3 and [15]).

Strictly speaking, the matrices from Theorem 4 (as well as their truncated versions) do not belong to the class of matrices described via (14), since their non-zero off-diagonal entries are not equal to one (see (29)). Nevertheless, in the notation of (29),

$$A^{(c)}_{k,k+2} = A^{(c)}_{k+2,k} = \frac{c^2}{4} \cdot \left(1 + O\left(\frac{1}{k^2}\right)\right), \tag{250}$$

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

$$A^{(c)}_{k,k} = \frac{c^2}{4} \cdot \left(2 + \left(\frac{2k}{c}\right)^2 \cdot \left(1 + O\left(\frac{1}{k}, \frac{c^2}{k^4}\right)\right)\right), \tag{251}$$

for every $k = 0, 1, 2, \ldots$. In other words, the matrix $A^{c,even}$ from Theorem 4 in Section 3.3 can be viewed as a small perturbation of the symmetric tridiagonal matrix A, defined via the formula

$$A = \frac{c^2}{4} \cdot \begin{pmatrix} A_0 & 1 & & & \\ 1 & A_1 & 1 & & \\ & 1 & A_2 & 1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{252}$$

where $A_0, A_1, \ldots$ are defined via the formula

$$A_j = 2 + \left(\frac{4j}{c}\right)^2, \tag{253}$$

for every $j = 0, 1, \ldots$ (compare to (233) in Theorem 22).

In particular, the algorithm of Sections 5.2, 5.3, with obvious minor modifications, is applicable to the task of evaluating eigenvectors of $A^{c,even}$ numerically. Moreover, the error analysis of such evaluation, in a somewhat more general form, has been carried out in Theorems 20, 21, 22 in Section 4.3 (see also Corollary 6 in Section 4.2).

In Section 7, we present several numerical examples involving matrices similar to (252). For the results of additional numerical experiments, where slightly modified versions of the algorithms of this paper (see Section 5) are used to evaluate PSWFs, see, for example, [18].

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. In addition, extended precision calculations were used to estimate the accuracy of double precision calculations.

7.1 Experiment 1.

In this experiment, we illustrate the performance of the algorithm on certain matrices.

Description. We first choose, more or less arbitrarily, the real numbers $a, \delta \geq 0$. Then, for each choice of five different values $c = 10^2, 10^3, 10^4, 10^5, 10^6$, we proceed as follows. We define the integer $n = n(c)$ via (222) in Theorem 21, define $A_1, \ldots, A_n$ via (233) in Theorem 22, and then define the symmetric tridiagonal $n \times n$ matrix $A = A(c)$ via (14). Then, we define the real number $\tilde{\lambda}$ via the formula

$$\tilde{\lambda} = 4 + 2 \cdot \left(\frac{\delta}{c}\right)^{2a/(a+2)}, \tag{254}$$

(see (235) in Theorem 22), and find the closest eigenvalue $\lambda(c)$ of $A(c)$ by Shifted Inverse Power method, using $\tilde{\lambda}$ as the initial approximation to $\lambda(c)$ (see Section 3.4.1). We then compute $\delta(c)$ from $\lambda(c)$ via (235). We also compute the real number R via the formula

$$R = \exp\left(-D_a \cdot \delta\right), \tag{255}$$

where D_a is defined via (195) in Theorem 19.

Next, we obtain the unit-length $\lambda(c)$-eigenvector of A by four different methods:

1. $Y = (Y_1, \ldots, Y_n)$ via 30 iterations of Shifted Inverse Power, in double precision.

2. $X = (X_1, \ldots, X_n)$ via the algorithm from Section 5.3, in double precision.

3. $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ via 30 iterations of Shifted Inverse Power, in extended precision (we also recompute the eigenvalue $\hat{\lambda}(c)$ in extended precision).

4. $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_n)$ via the algorithm from Section 5.3, in extended precision.

We verify that each of $\hat{X}$ and $\hat{Y}$ satisfies the definition of an eigenvector coordinate-wise to at least 17 decimal digits, and also that $\hat{X} = \hat{Y}$ to at least 17 decimal digits. In other words, each of $\hat{X}, \hat{Y}$ is the unit-length $\lambda(c)-$eigenvector of A defined to full double precision. We use this observation to evaluate the relative and absolute errors of X_j, Y_j, for every $j = 1, \ldots, n$.

Tables and Figures. The results of the experiment are displayed in Tables 1–6. Each of these tables corresponds to a particular choice of a and δ, and has the following structure. Each of five columns corresponds to a different value of c, between 10^2 and 10^6. The first three rows contain c, the matrix size n, and the index k (such that $A_k \approx \lambda(c) - 2$: see (166) in Theorem 16 for the precise definition). The next four rows contain the eigenvalue $\lambda(c)$, its accuracy, its cardinal number $1 \leq i \leq n$, and $\delta(c)$ (see (254)). The next four rows contain the coordinates X_1 and X_k, their ratio, and an asymptotic estimate R of this ratio (see (255)). The next two rows contain the relative accuracy of X_1 and Y_1. The last two rows contain the maximal absolute accuracy among all coordinates of X, Y, respectively.

Also, in Figures 1(a), 1(b) we plot the relative errors of X_1, Y_1, respectively, on a logarithmic scale as functions of $\log_{10}(c)$. More specifically, each of Figures 1(a), 1(b) contains five plots of such errors, corresponding to $a = 1, 2, 3, 4, 6$, respectively. Each point on such plot is the geometric mean of ten relative errors (corresponding to ten different values of δ between 50 and 200). For example, to generate plots corresponding to $a = 2$ in Figure 1(a), we use the data from Tables 1–3 (as well as the data corresponding to seven other values of δ).

c	10^2	10^3	10^4	10^5	10^6
n	180	1,497	14,320	141,803	1,415,035
k	71	226	706	2,244	7,109
λ	0.50164E+01	0.41021E+01	0.40099E+01	0.40010E+01	0.40000E+01
rel(λ)	0.35409E-15	0.00000E+00	-.22149E-15	0.00000E+00	0.00000E+00
i	119	943	9,058	90,100	900,398
$\delta(c)$	0.50826E+02	0.51086E+02	0.49906E+02	0.50379E+02	0.50551E+02
X_1	0.19744E-24	0.46025E-26	0.21813E-26	0.20152E-27	0.26903E-28
X_k	0.12621E+00	0.60020E-01	0.28690E-01	0.14439E-01	0.73972E-02
X_1/X_k	0.15642E-23	0.76683E-25	0.76030E-25	0.13957E-25	0.36368E-26
R	0.30331E-24	0.22762E-24	0.84367E-24	0.49922E-24	0.41222E-24
rel(X_1)	0.19302E-13	0.26421E-12	0.43114E-11	0.13247E-10	0.11171E-09
rel(Y_1)	0.55816E-14	0.24161E-13	0.55651E-12	0.68590E-11	0.30212E-10
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.17885E-14	0.10874E-13	0.81497E-13	0.11156E-12	0.48541E-12
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.47183E-15	0.62991E-15	0.86371E-14	0.56234E-13	0.13395E-12

Table 1: *Experiment 1. Parameters:* $a = 2$, $\delta = 50$.

c	10^2	10^3	10^4	10^5	10^6
n	180	1,497	14,320	141,803	1,415,035
k	101	315	1,004	3,180	9,992
λ	0.60503E+01	0.41993E+01	0.40201E+01	0.40019E+01	0.40002E+01
rel(λ)	0.00000E+00	0.21150E-15	0.00000E+00	0.22193E-15	0.00000E+00
i	140	976	9,106	90,161	900,470
$\delta(c)$	0.10251E+03	0.99703E+02	0.10087E+03	0.10118E+03	0.99842E+02
X_1	0.29706E-47	0.33654E-49	0.73691E-51	0.77717E-52	0.54741E-52
X_k	0.13199E+00	0.56585E-01	0.28214E-01	0.14026E-01	0.71872E-02
X_1/X_k	0.22505E-46	0.59474E-48	0.26118E-49	0.55410E-50	0.76165E-50
R	0.35338E-49	0.80393E-48	0.21716E-48	0.15569E-48	0.68891E-48
rel(X_1)	0.14729E-13	0.20625E-12	0.14676E-11	0.40918E-10	0.46459E-10
rel(Y_1)	0.47051E-14	0.39500E-13	0.57239E-12	0.68254E-11	0.32697E-10
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.11519E-14	0.79096E-14	0.21711E-13	0.30486E-12	0.17340E-12
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.78063E-15	0.75123E-15	0.77475E-14	0.52657E-13	0.12385E-12

Table 2: *Experiment 1. Parameters:* $a = 2$, $\delta = 100$.

c	10^2	10^3	10^4	10^5	10^6
n	180	1,497	14,320	141,803	1,415,035
k	123	389	1,227	3,875	12,296
λ	0.70491E+01	0.43029E+01	0.40301E+01	0.40030E+01	0.40003E+01
rel(λ)	0.00000E+00	0.20641E-15	0.00000E+00	-.22187E-15	0.00000E+00
i	156	1,008	9,150	90,217	900,542
$\delta(c)$	0.15244E+03	0.15146E+03	0.15076E+03	0.15021E+03	0.15121E+03
X_1	0.24360E-68	0.10108E-73	0.79506E-75	0.19809E-75	0.10325E-76
X_k	0.14129E+00	0.59531E-01	0.27646E-01	0.13861E-01	0.70498E-02
X_1/X_k	0.17240E-67	0.16979E-72	0.28757E-73	0.14290E-73	0.14647E-74
R	0.28826E-73	0.86951E-73	0.18921E-72	0.34954E-72	0.11510E-72
rel(X_1)	0.57053E-14	0.39666E-12	0.31336E-13	0.28896E-10	0.16840E-09
rel(Y_1)	0.29582E-14	0.44484E-13	0.58518E-12	0.69078E-11	0.32768E-10
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.64401E-15	0.14322E-13	0.88880E-14	0.19695E-12	0.58894E-12
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.30530E-15	0.81206E-15	0.75181E-14	0.50491E-13	0.11008E-12

Table 3: *Experiment 1. Parameters:* $a = 2$, $\delta = 150$.

c	10^2	10^3	10^4	10^5	10^6
n	148	1,251	12,025	119,207	1,189,823
k	80	371	1,725	8,052	37,584
λ	0.48307E+01	0.40378E+01	0.40018E+01	0.40000E+01	0.40000E+01
rel(λ)	0.00000E+00	0.21996E-15	0.00000E+00	0.00000E+00	0.00000E+00
i	105	925	9,090	90,728	907,100
$\delta(c)$	0.51745E+02	0.51066E+02	0.51375E+02	0.52214E+02	0.53092E+02
X_1	0.15657E-27	0.56925E-29	0.34307E-30	0.11988E-31	0.40486E-33
X_k	0.16156E+00	0.70686E-01	0.31217E-01	0.14289E-01	0.65824E-02
X_1/X_k	0.96908E-27	0.80532E-28	0.10990E-28	0.83895E-30	0.61506E-31
R	0.16701E-27	0.38677E-27	0.26394E-27	0.93576E-28	0.31612E-28
rel(X_1)	0.24916E-13	0.17710E-12	0.82978E-11	0.45445E-09	0.39497E-08
rel(Y_1)	0.50118E-14	0.40247E-13	0.15850E-11	0.24346E-10	0.10033E-09
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.14710E-14	0.50368E-14	0.85255E-13	0.21667E-11	0.85706E-11
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.53949E-15	0.14180E-14	0.15365E-13	0.11264E-12	0.27496E-12

Table 4: *Experiment 1. Parameters:* $a = 4$, $\delta = 50$.

c	10^2	10^3	10^4	10^5	10^6
n	149	1,251	12,025	119,207	1,189,823
k	99	468	2,160	10,074	46,353
λ	0.59504E+01	0.40964E+01	0.40044E+01	0.40002E+01	0.40000E+01
rel(λ)	0.14926E-15	0.00000E+00	0.00000E+00	-.44406E-15	0.22204E-15
i	117	942	9,108	90,747	907,118
$\delta(c)$	0.98136E+02	0.10293E+03	0.10085E+03	0.10226E+03	0.99596E+02
X_1	0.65592E-50	0.16890E-56	0.13441E-56	0.22928E-58	0.60367E-58
X_k	0.16663E+00	0.69229E-01	0.31413E-01	0.14323E-01	0.65935E-02
X_1/X_k	0.39364E-49	0.24397E-55	0.42789E-55	0.16007E-56	0.91554E-56
R	0.20868E-52	0.55400E-55	0.71969E-54	0.12721E-54	0.34337E-53
rel(X_1)	0.36733E-13	0.11611E-12	0.64777E-11	0.56745E-10	0.58871E-08
rel(Y_1)	0.17734E-13	0.71711E-13	0.15602E-11	0.24704E-10	0.12500E-09
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.20053E-14	0.38650E-14	0.58993E-13	0.24506E-12	0.11594E-10
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.88124E-15	0.11261E-14	0.14018E-13	0.10435E-12	0.30450E-12

Table 5: *Experiment 1. Parameters:* $a = 4$, $\delta = 100$.

c	10^2	10^3	10^4	10^5	10^6
n	148	1,251	12,025	119,207	1,189,823
k	115	535	2,472	11,446	53,300
λ	0.75092E+01	0.41649E+01	0.40075E+01	0.40003E+01	0.40000E+01
rel(λ)	0.11827E-15	0.00000E+00	0.00000E+00	0.22202E-15	0.00000E+00
i	128	958	9,126	90,765	907,138
$\delta(c)$	0.15244E+03	0.15386E+03	0.15112E+03	0.14999E+03	0.15141E+03
X_1	0.28839E-74	0.19053E-83	0.17930E-83	0.66661E-84	0.11235E-85
X_k	0.19676E+00	0.68972E-01	0.31725E-01	0.14354E-01	0.66008E-02
X_1/X_k	0.14657E-73	0.27623E-82	0.56519E-82	0.46442E-82	0.17022E-83
R	0.14492E-81	0.25640E-82	0.75552E-81	0.30408E-80	0.51747E-81
rel(X_1)	0.76598E-14	0.18105E-12	0.89870E-11	0.47429E-09	0.44710E-08
rel(Y_1)	0.76598E-14	0.67667E-13	0.13524E-11	0.23468E-10	0.14141E-09
$\max_j \lvert X_j - \hat{X}_j \rvert$	0.25396E-14	0.53898E-14	0.80525E-13	0.19056E-11	0.81957E-11
$\max_j \lvert Y_j - \hat{Y}_j \rvert$	0.24146E-14	0.10780E-14	0.12499E-13	0.99000E-13	0.22676E-12

Table 6: *Experiment 1. Parameters:* $a = 4$, $\delta = 150$.

a	1	2	3	4	6
$\beta_Y(a)$	0.791E+00	0.104E+01	0.103E+01	0.109E+01	0.110E+01
$\beta_X(a)$	0.586E+00	0.101E+01	0.115E+01	0.131E+01	0.146E+01
$\beta(a)$	0.666E+00	0.100E+01	0.119E+01	0.133E+01	0.150E+01
$4a/(a+2)$	0.133E+01	0.200E+01	0.239E+01	0.266E+01	0.300E+01

Table 7: *Experiment 1. Best fit slopes of* $\log_{10}(\mathrm{rel}(Y_1))$, $\log_{10}(\mathrm{rel}(X_1))$ *as functions of* $\log_{10}(c)$.

c	10^2	10^3	10^4	10^5	10^6
m	162	1,135	10,292	100,629	1,001,357
N	192	1,175	10,392	100,829	1,001,757
$J_m(c)$	0.13298E-20	0.11471E-21	0.32071E-22	0.14301E-22	0.59576E-23
$J_c(c)$	0.96366E-01	0.44730E-01	0.20762E-01	0.96369E-02	0.44730E-02
$J_m(c)/J_c(c)$	0.13800E-19	0.25644E-20	0.15447E-20	0.14840E-20	0.13319E-20
R	0.33515E-20	0.29104E-20	0.29127E-20	0.33515E-20	0.32303E-20
$\lvert 1 - X_m/J_m(c)\rvert$	0.33801E-13	0.15085E-12	0.24630E-12	0.22284E-11	0.77524E-11
$\lvert 1 - Y_m/J_m(c)\rvert$	0.36770E-14	0.22545E-13	0.14788E-12	0.98237E-12	0.24681E-11

Table 8: *Experiment 3. Parameters:* $\delta = 50$.

c	10^2	10^3	10^4	10^5	10^6
m	200	1,215	10,464	101,000	1,002,154
N	230	1,255	10,564	101,200	1,002,554
$J_m(c)$	0.20593E-40	0.61117E-42	0.10612E-42	0.39770E-43	0.18323E-43
$J_c(c)$	0.96366E-01	0.44730E-01	0.20762E-01	0.96369E-02	0.44730E-02
$J_m(c)/J_c(c)$	0.21370E-39	0.13663E-40	0.51112E-41	0.41268E-41	0.40965E-41
R	0.27453E-41	0.78625E-41	0.87694E-41	0.98375E-41	0.10920E-40
$\lvert 1 - X_m/J_m(c)\rvert$	0.38368E-14	0.13658E-12	0.20091E-11	0.10091E-11	0.56160E-11
$\lvert 1 - Y_m/J_m(c)\rvert$	0.28466E-14	0.93836E-14	0.14805E-12	0.11176E-11	0.29720E-11

Table 9: *Experiment 3. Parameters:* $\delta = 100$.

c	10^2	10^3	10^4	10^5	10^6
m	231	1,282	10,608	101,310	1,002,823
N	261	1,322	10,708	101,510	1,003,223
$J_m(c)$	0.25898E-59	0.45624E-62	0.42252E-63	0.13902E-63	0.57054E-64
$J_c(c)$	0.96366E-01	0.44730E-01	0.20762E-01	0.96369E-02	0.44730E-02
$J_m(c)/J_c(c)$	0.26875E-58	0.10199E-60	0.20350E-61	0.14425E-61	0.12754E-61
R	0.80027E-62	0.22722E-61	0.29062E-61	0.34453E-61	0.35676E-61
$\lvert 1 - X_m/J_m(c)\rvert$	0.72561E-14	0.28169E-12	0.13717E-12	0.72506E-12	0.25122E-10
$\lvert 1 - Y_m/J_m(c)\rvert$	0.64024E-15	0.28275E-13	0.12375E-12	0.13185E-11	0.38545E-11

Table 10: *Experiment 3. Parameters:* $\delta = 150$.

To each plot in Figures 1(a), 1(b), one can fit a line (in the least square sense). The slopes of such lines are displayed in Table 7. This table has the following structure. Each column corresponds to a different value of a. Second row contains the slopes corresponding to $\mathrm{rel}(Y_1)$ (see Figure 1(b)). Third row contains the slopes corresponding to $\mathrm{rel}(X_1)$ (see Figure 1(a)). Fourth row contains $\beta(a)$, where $\beta(a)$ is defined via (257) below (the values in third and fourth rows would be identical if $\mathrm{rel}(X_1)$ were proportional to $c^{\beta(a)}$). Last row contains the number $4 \cdot a/(a+2)$ (the power of c in (237) of Theorem 22).

Observations. Several observations can be made from Tables 1–6, Figure 1, Table 7, and some additional numerical experiments by the author.

Observation 1. For every choice of parameters in Experiment 1, the coordinate X_1 is fairly small compared to X_k, as predicted by Theorem 6 and Corollary 1 in Section 4.1. In addition, $X_k = O(c^{-1/3})$, and the ratio X_1/X_k is roughly of the same order of magnitude as the estimate R defined via (255) above. More specifically, $\log(R)$ never deviates from $\log(X_1) - \log(X_k)$ by more than 10%, and also $R > X_1/X_k$ for large c, as expected from (203) in Theorem 20.

Observation 2. Despite the fact that X_1/X_k is much smaller than machine zero in all experiments (for all c, $X_1/X_k \approx 10^{-25}, 10^{-50}, 10^{-75}$ for $\delta = 50, 100, 150$, respectively), both X_1 and Y_1 are still evaluated to fairly high *relative* accuracy, in all cases.

Observation 3. For any c and a, the relative accuracy of both X_1 and Y_1 seems to be essentially independent of their magnitude. For example, for $a = 4$ and $c = 10^6$, the relative accuracy of X_1 is 0.4E-8, 0.6E-8, 0.4E-8 for $\delta = 50, 100, 150$, respectively (despite the fact that X_1 itself is equal to 0.4E-33, 0.6E-58, 0.1E-85, respectively). In other words, the δ-dependent factor in (237) of Theorem 22 seems to be an artifact of the analysis.

Observation 4. On the other hand, the relative accuracy of both X_1 and Y_1 does depend on c (as Theorem 22 suggests). In particular, for any fixed a, the relative error of Y_1 seems to be roughly proportional to c, e.g.

$$\mathrm{rel}(Y_1) = O(c) \cdot \varepsilon, \tag{256}$$

where ε is the machine precision (see second row in Table 7).

Observation 5. For any fixed a, the relative error of X_1 seems to be roughly proportional to c^β, where $\beta = \beta(a)$ defined via the formula

$$\beta(a) = \frac{2 \cdot a}{a+2} \tag{257}$$

(see third and fourth rows in Table 7, and also Conjecture 1). On the other hand, in Theorem 22 in Section 4.3 we derived a certain upper bound on the relative error of X_1 (see (237) and last row in Table 7); this bound is proportional to $c^{4a/(a+2)}$. In other words, numerical experiments seem to indicate that Theorem 22 overestimates the number of lost digits roughly by a factor of two. For example, for $a = 4$, $\delta = 150$ and $c = 10^6$ (see last column in Table 6) we lose almost $\beta(a) \cdot 6 = 8$ decimal digits, while the pessimistic estimate from Theorem 22 suggest that we will lose 16 decimal digits. In other words, the estimate from Theorem 22 is overly cautious.

7.2 Experiment 2.

In Experiment 1, we took a rather detailed look at relative errors to which the first coordinate of an eigenvector of certain tridiagonal matrices is evaluated. The purpose of this section is to illustrate the analysis of Section 4 in a more qualitative way.

To that end, we carry out the experiment described in Section 7.1 with the following parameters: $a = 2$, $c = 1000$, $n = 1497$, $\delta = 50$. We obtain the four unit-length vectors $X, Y, \hat{X}, \hat{Y}$ in $\mathbb{R}^n$, as described in Section 7.1.

Figures. We display the results of this experiment in Figures 2(a)–2(c). In each figure, the abscissa corresponds to the indices of the eigenvector, i.e. $1 \leq j \leq n$; thus, we plot certain functions of the indices of the eigenvector.

In Figure 2(a), we plot the coordinates X_j of X, on the linear scale (left) and on the logarithmic scale (right).

In Figure 2(b), we plot the relative (left) and absolute (right) errors of X_j on the logarithmic scale.

In Figure 2(c), we plot the relative (left) and absolute (right) errors of Y_j on the logarithmic scale.

Observations. Several observations can be made from Figures 2(a)–2(c).

The following three observations pertain to the behavior of the coordinates of X (see Figure 2(a)).

Observation 1. In the beginning, the coordinates of X grow rapidly from $\approx 10^{-26}$ to $\approx 10^{-1}$ up to the index k such that $\lambda \approx A_k + 2$ (in agreement with Theorem 6 in Section 4.1). We refer to the corresponding indices as the "region of growth".

Observation 2. At the other end, they decay rapidly (while changing signs) from ≈ 0.05 to $\approx 10^{-14}$, starting from the index $n - q$ such that $\lambda \approx A_{n-q} - 2$ (in agreement with Theorem 7 in Section 4.1). We refer to the corresponding indices as the "region of decay".

Observation 3. In the middle (i.e. for indices j such that $\lambda - 2 \leq A_j \leq \lambda + 2$), the coordinates behave in an "oscillatory way", in the sense of Section 2. Such behavior is expected from Theorems 10, 11 and Corollaries 3, 4 in Section 4.1. We refer to the corresponding indices as the "oscillatory region" (see also [16] for an alternative approach to the evaluation of X_j in the oscillatory region that, *inter alia*, further justifies this term).

The following observations pertain to the behavior of relative and absolute errors to which the coordinates of the eigenvector are evaluated, by either Inverse Power or the algorithm from Section 5.3.

Observation 4. Qualitatively, the behavior of relative errors of X_j is similar to that of Y_j and depends of whether j is in the region of growth, in the region of decay, or in the oscillatory region.

Observation 5. In the region of growth, the relative errors of X_j change monotonically with j and always stays "small" (below 10^{-12}), in agreement with Theorems 13, 16, Corollary 6 in Section 4.2 and Theorem 22 in Section 4.3. In the region of decay, the relative errors of X_j display a similar behavior, in agreement with Theorem 17, Corrolary 6 in Section 4.2, and Theorem 22 in Section 4.3. In particular, both in the regions of growth and in the region of decay the relative errors of X_j essentially do not depend on the magnitude of X_j.

Observation 6. In the oscillatory region, the relative errors of X_j oscillate between

10^{-16} and 10^{-10}. On the other hand, the *absolute* errors of X_j always stay below roughly 10^{-14}. In other words, the relative errors of X_j in the oscillatory region depend on the magnitude of X_j, in agreement with Theorems 15, 16 in Section 4.2.

7.3 Experiment 3.

In this experiment, we illustrate the numerical algorithms of Section 5 via evaluation of Bessel functions (see Sections 3.2, 3.4.2, 6.1).

Description. We first choose, more or less arbitrarily, the real number $\delta \geq 0$. Then, for each choice of five different values $c = 10^2, 10^3, 10^4, 10^5, 10^6$, we do the following. We define the integer $m = m(\delta, c)$ via the formula

$$m = c + \delta^{2/3} \cdot c^{1/3} \tag{258}$$

(see (199) in Theorem 20 and (222) in Theorem 21), select the integer $N > m$ (according to Remark 8 in Section 3.4.2), define the integer n via the formula

$$n = 2 \cdot N + 1, \tag{259}$$

define $A_1, \ldots, A_n$ via (233) with $a = 1$ in Theorem 22 (see also (39)), and then define the symmetric tridiagonal $n \times n$ matrix $A = A(c)$ via (14). Then, we define the real number $\lambda(c)$ via the formula

$$\lambda(c) = 2 + \frac{n+1}{c}. \tag{260}$$

(We observe that $\lambda(c)$ is an eigenvalue of A, according to (40) in Section 3.4.2.) We also compute the real number R via the formula

$$R = \exp\left(-\delta \cdot \sqrt{\frac{8}{9}}\right) \tag{261}$$

(see (255) above and (195) in Theorem 19).

Next, we obtain the unit-length $\lambda(c)$-eigenvector of A by four different methods:

1. $Y = (Y_N, \ldots, Y_0, \ldots, Y_{-N})$ via 30 iterations of Shifted Inverse Power, in double precision (observe that the indices vary between N and $-N$, as in (41)).

2. $X = (X_N, \ldots, X_0, \ldots X_{-N})$ via the algorithm from Section 5.3, in double precision.

3. $\hat{Y} = (\hat{Y}_N, \ldots, \hat{Y}_0, \ldots, \hat{Y}_{-N})$ via 30 iterations of Shifted Inverse Power, in extended precision.

4. $\hat{X} = (\hat{X}_N, \ldots, \hat{X}_0, \ldots, \hat{X}_{-N})$ via the algorithm from Section 5.3, in extended precision.

The experiment is conducted for each pair of values δ, c, where $\delta = 50, 100, 150$ and $c = 10^2, 10^3, 10^4, 10^5, 10^6$. In each case, we verify that each of $\hat{X}$ and $\hat{Y}$ satisfies the definition of an eigenvector coordinate-wise to at least 17 decimal digits, and also that $\hat{X} = \hat{Y}$ to at least 17 decimal digits. In other words, each of $\hat{X}, \hat{Y}$ is the unit-length $\lambda(c)-$eigenvector of A defined to full double precision. Also, we verify that the middle $2 \cdot m + 1$ coordinates of both $\hat{X}$ and $\hat{Y}$ are equal to $J_m(c), \ldots, J_0(c), \ldots, J_{-m}(c)$ to at least 17 decimal digits (see

Remarks 8, 9 in Section 3.4.2). We use these observations to compute the accuracy to which the coordinates $X_m, \ldots, X_0$ of X and $Y_m, \ldots, Y_0$ of Y approximate $J_m(c), \ldots, J_0(c)$.

The results of the experiment are displayed in Tables 8–10. Each of these tables corresponds to a particular choice δ in (258), and has the following structure. Each of five columns corresponds to a different value of c, between 10^2 and 10^6. The first three rows contain c, the integer m defined via (258), and the integer $N > m$ (see Remark 8 in Section 3.4.2). The next four rows contain $J_m(c)$ and $J_c(c)$, their ratio, and the asymptotic estimate R of this ratio (see (261)). The last two rows contain the relative accuracy to which X_m and Y_m, respectively, approximate $J_m(c)$.

Observations. Several observations can be made from Tables 8–10.

Observation 1. For every choice of parameters in Experiment 3, $J_m(c)$ is fairly small compared to $J_c(c)$, as predicted by Theorem 6 and Corollary 1 in Section 4.1. In addition, for all $c \geq 10^3$, the ratio $J_m(c)/J_c(c)$ is within a factor of three from the estimate R defined via (261) above (see (203) in Theorem 20).

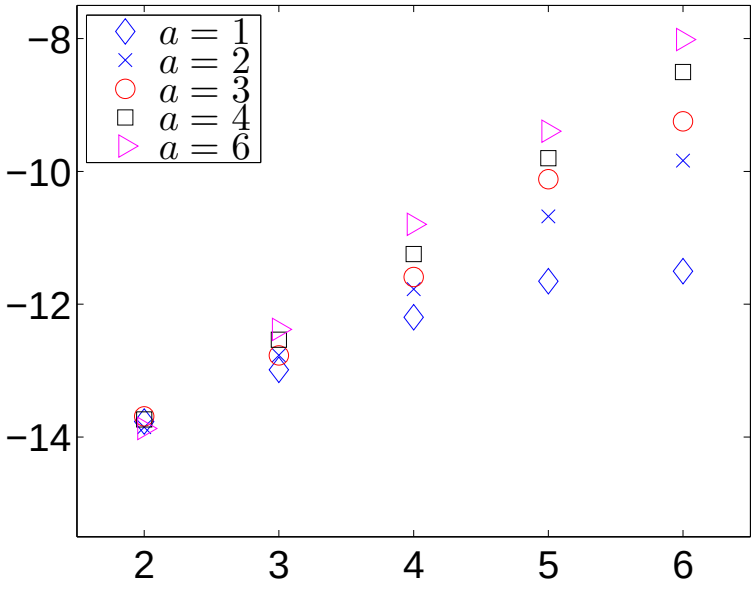
Remark 17. *Observation 1 above is obviously related to the well known Debye's expansions of Bessel functions (see e.g. [14]).*

Observation 2. Despite the fact that $J_m(c)$ is much smaller than machine zero in all experiments (for all c, $J_c(c) \leq 10^{-20}, 10^{-40}, 10^{-59}$ for $\delta = 50, 100, 150$, respectively), both X_m and Y_m approximate $J_m(c)$ to a fairly high *relative* accuracy, in all cases. Moreover, for any c, this accuracy seems to be independent of the magnitude of $J_m(c)$ (compare to (237) of Theorem 22; see also Conjecture 1).

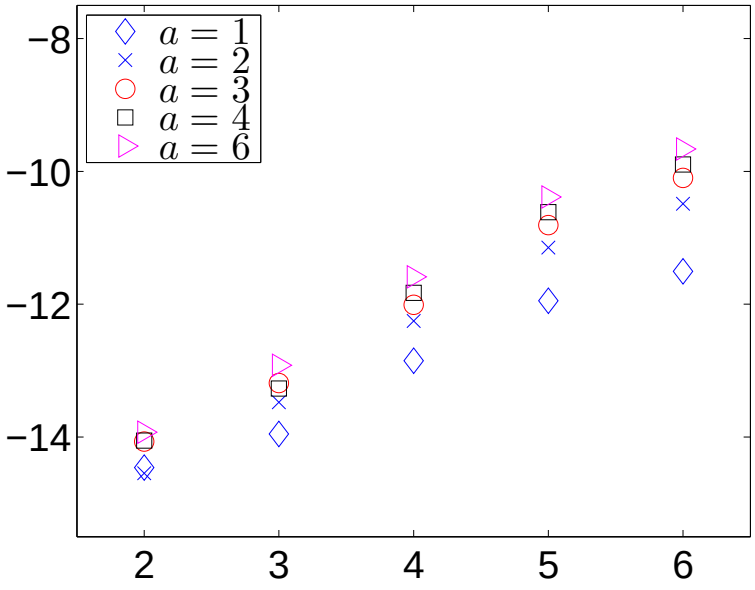
Observation 4. On the other hand, the relative accuracy of both X_1 and Y_1 does depend on c (as Theorem 22 in Section 4.3 suggests). In particular, for any fixed a, the relative error of Y_1 seems to be roughly proportional to $c^{0.8}$, e.g.

$$\mathrm{rel}(Y_1) = O(c^{0.8}) \cdot \varepsilon, \tag{262}$$

where ε is the machine precision (see second column in Table 7). Also, the relative error of X_1 seems to be roughly proportional to $c^{2/3}$ (see Table 7), in agreement with Conjecture 1 above (compare to (237) of Theorem 22).



(a) $\log_{10}(\mathrm{rel}(X_1))$ as a function of $\log_{10}(c)$.



(b) $\log_{10}(\mathrm{rel}(Y_1))$ as a function of $\log_{10}(c)$.

Figure 1: Relative errors of X_1, Y_1, on a logarithmic scale, as a function of $\log_{10}(c)$, for $a = 1, 2, 3, 4, 6$. Corresponds to Experiment 1.

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] R. Lederman, *On the Analytical and Numerical Properties of the Truncated Laplace Transform*, Yale CS Technical Report #1490, 2014.

[12] H. J. Landau, H. Widom, *Eigenvalue distribution of time and frequency limiting*, J. Math. Anal. Appl. 77, 469-81, 1980.

[13] J. C. P. Miller, *Bessel Functions. Part II, Functions of Positive Integer Order*, Cambridge University Press, Cambridge, 1952.

[14] F. W. J. Olver, *Some new asymptotic expansions for Bessel functions of large orders*, Proc. Cambridge Philos. Soc. 48 (3), pp. 414427 (1952).

[15] A. Osipov, V. Rokhlin, H. Xiao, *Prolate Spheroidal Wave Functions of Order Zero*, Springer, Applied Mathematical Sciences, Vol. 187 (2013).

[16] A. Osipov, *Evaluation of small elements of the eigenvectors of certain symmetric tridiagonal matrices with high relative accuracy*, Yale CS Technical Report #1460, 2012.

[17] A. Osipov, *Certain upper bounds on the eigenvalues associated with prolate spheroidal wave functions*, Appl. Comput. Harmon. Anal. (2013), http://dx.doi.org/10.1016/j.acha.2013.03.002.

[18] A. Osipov, V. Rokhlin, *On the evaluation of prolate spheroidal wave functions and associated quadrature rules*, Appl. Comput. Harmon. Anal. (2013), http://dx.doi.org/10.1016/j.acha.2013.04.002.

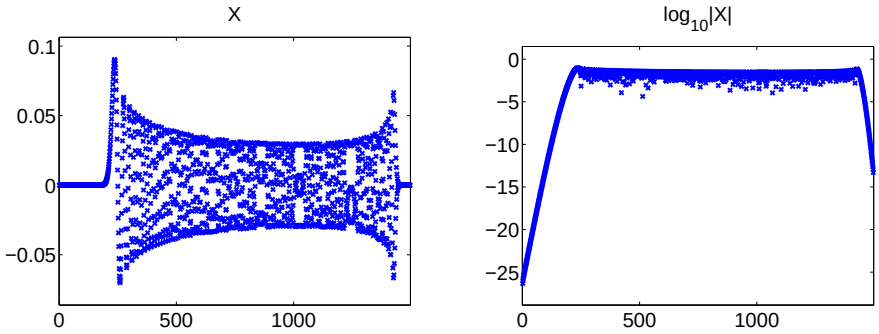
[19] B. N. Parlett, *The symmetric eigenvalue problem*, Prentice Hall, Inc. 1980.

[20] D. Slepian, H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty - I*, Bell Syst. Tech. J. January 43-63, 1961.

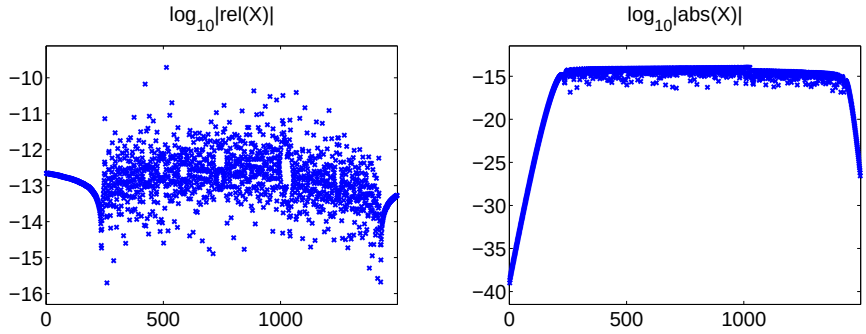
[21] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis*, Second Edition, Springer-Verlag, 1993.

[22] J. H. Wilkinson, *Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

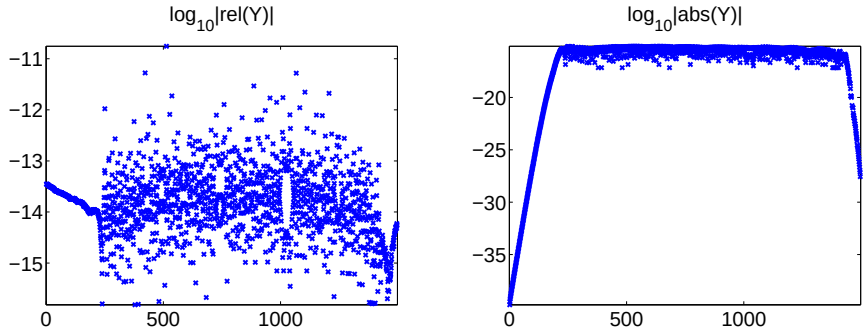
[23] H. Xiao, V. Rokhlin, N. Yarvin, *Prolate spheroidal wavefunctions, quadrature and interpolation*, Inverse Problems, 17(4):805-828, 2001.



(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: $c = 1000$, $n = 1500$, $\lambda = 0.41022\text{E+}01$, $k = 226$, $q = 65$. Corresponds to Experiment 2.