MATHEMATICS OF COMPUTATION
Volume 71, Number 237, Pages 217–236
S 0025-5718(01)01325-4
Article electronically published on May 14, 2001

ACCURATE COMPUTATION OF THE SMALLEST EIGENVALUE OF A DIAGONALLY DOMINANT M-MATRIX

ATTAHIRU SULE ALFA, JUNGONG XUE, AND QIANG YE

Abstract. If each off-diagonal entry and the sum of each row of a diagonally dominant M-matrix are known to certain relative accuracy, then its smallest eigenvalue and the entries of its inverse are known to the same order relative accuracy independent of any condition numbers. In this paper, we devise algorithms that compute these quantities with relative errors in the magnitude of the machine precision. Rounding error analysis and numerical examples are presented to demonstrate the numerical behaviour of the algorithms.

1. Introduction

Diagonally dominant M-matrices form one of the most important classes of matrices in applications and have been studied extensively in the literature; see [5, Chapter 6]. Among problems of interest are solving a linear system $Ax = b$ and finding the smallest eigenvalue of A (corresponding to the Perron root of the inverse); see [2, 12, 21, 22]. There are many well established numerical methods for solving such problems and they lead to a backward stable solution, which has an error depending on the condition of the problem. For instance, applying the QR algorithm to find the smallest eigenvalue λ of A, the computed one $\widehat{\lambda}$ is the eigenvalue of a perturbed matrix $A+E$ with $\|E\|_2 \sim \epsilon\|A\|_2$ (where ϵ is the machine roundoff unit). Then, assuming λ is a simple eigenvalue, we obtain $|\widehat{\lambda} - \lambda| \sim \|E\|_2/y^*x \sim \epsilon\|A\|_2/y^*x$ and thus the relative error is given by

$$\frac{|\widehat{\lambda} - \lambda|}{\lambda} \sim \epsilon \frac{\|A\|_2}{\lambda} \frac{1}{y^*x}$$

where x and y are respectively unit right and left eigenvectors corresponding to λ. Hence there are two situations where the computed eigenvalue $\widehat{\lambda}$ has a low relative accuracy (i.e., large relative error). If y^*x is small (i.e., λ is ill-conditioned), the error will be large. On the other hand, even if λ is well-conditioned but λ is small

Received by the editor March 22, 1999 and, in revised form, March 14, 2000.

2000 *Mathematics Subject Classification.* Primary 65F18, 65F05.

Key words and phrases. Entrywise perturbation, diagonal dominant matrix, M-matrix, eigenvalue.

Research of the first author was supported by grant No. OGP0006854 from Natural Sciences and Engineering Research Council of Canada.

Research of the second author was supported by Natural Sciences Foundation of China and Alexander von Humboldt Foundation of Germany.

Research of the third author was supported by grants from University of Manitoba Research Development Fund and Natural Sciences and Engineering Research Council of Canada while this author was with University of Manitoba, Winnipeg, Manitoba, Canada.

©2001 American Mathematical Society

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

relative to $\|A\|_2$, then the error will also be large. This latter case is true even for symmetric matrices. Unfortunately, in many application problems, the smallest eigenvalue is the one of interest (see [3, 2, 12, 22]) and the computed eigenvalue by the standard algorithms may have low or even no accuracy.

The numerical difficulties mentioned above are well known and originate in the limitation of the normwise perturbation E. Namely, if the matrix A is only determined to within a normwise perturbation E, then its eigenvalues can only be determined to a low accuracy in the situations described above. Starting in a work by Demmel and Kahan [7] on computing the singular values of a bidiagonal matrix, there have been significant works in the last decade to identify special classes of problems for which the computed quantities are well determined by the matrices, usually under entrywise perturbations, and to devise algorithms for computing them to high relative accuracy. We refer to [9] and [16] for a summary of most of such classes of matrices and the literature. Some of those that are known to be determined to the machine precision are the singular values of bidiagonal matrices [7], the Perron root of a nonnegative matrix [10] and the steady state distribution of a Markov chain [18]. For the kind of problems that we are interested in here, a perturbation analysis and an algorithm have been developed in [4] for the eigenvalues of symmetric scaled diagonal dominant matrices and in [26] for the smallest eigenvalue of an M-matrix. Unfortunately, their perturbation bounds and the relative errors of the computed eigenvalues still depend on certain condition numbers that are essentially related to the diagonal dominance.

For the class of diagonally dominant M-matrices, however, we have shown in a recent work [3] that the smallest eigenvalue and the entries of inverse are determined to high relative accuracy by the off-diagonal entries and the row sums of the matrices, irrelevant of any condition number and the magnitude of the eigenvalue. Namely, if small relative errors are introduced to each off-diagonal entry of a diagonally dominant M-matrix A and to the sum of each row of A which in turn determines the corresponding diagonal entry, then the smallest eigenvalue and each entry of the inverse have relative errors of the same magnitude. We note that in many applications (such as discretized PDE, Markov chains [2], [12], [21, Chapter 3] and electric circuits [22]), the off-diagonal entries and the row sums of the matrix play the role of physical parameters, while the diagonal entries are treated as functions of them and are redundant (the importance of properly parametrizing a matrix has also been shown for some other classes of matrices; see [8]). In those cases, it is more appropriate to consider the off-diagonal entries and the row sums as the matrix data. Indeed, in this work, a diagonally dominant M-matrix will be represented by its off-diagonal entries and the sums of its rows rather than the usual representations by all entries.

Thus, the new perturbation theory suggests that it is possible to compute the smallest eigenvalue and the inverse entries to high relative accuracy. It is the purpose of the present paper to develop such algorithms. We shall show how the Gaussian elimination can be implemented to solve $Ax = b$ (with $b \geq 0$) so that each entry of x will have high relative accuracy. The idea used is an extension of the GTH-algorithm [14] for stochastic matrices and thus we call it a GTH-like algorithm. For computing the smallest eigenvalue of A, we use a shifted inverse iteration algorithm similar to the one developed in [26] and we shall carry out the iteration in such a way that the computed approximate eigenvalue converges monotonically and quadratically until its relative error is in the magnitude of the

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

machine precision. We shall also present a rigorous roundoff error analysis for the iterative algorithm using a combination of forward and backward error analysis techniques.

We remark that computing the smallest eigenvalue to high accuracy is of great interest in several applications mentioned above. One particular application we are interested in arises in computing quantity $\delta = 1 - \eta$, where η is the decay rate for queue length in GI/M/1 queuing systems. In that problem, δ is a solution to $z - \Psi(z) = 0$ with $\Psi(z)$ being the smallest eigenvalue of a parameter dependent diagonally dominant M-matrix $A(z)$. The standard method to solve this equation in engineering is by the bisection method, which requires computing $z - \Psi(z)$ for z. Near the convergence stage (when $z - \Psi(z) \ll 1$), however, the standard eigenvalue algorithm may not even compute the sign of $z - \Psi(z)$ correctly. Our algorithm will guarantee the accuracy of $z - \Psi(z)$ to the machine precision, and certainly its sign. Hence in this way, the accuracy of δ can be obtained as high as the data warrants.

The rest of this paper is organized as follows. We first give in Section 2 some definitions and preliminary results, including an entrywise perturbation theory. Section 3 presents a GTH-like algorithm and error analysis for solving $Ax = b$. Details of our algorithm for computing the smallest eigenvalue and the error analysis are presented in Section 4. Finally, some numerical examples are given in Section 5.

2. Preliminaries and notation

For $m \times n$ matrices $B = (b_{ij})$ and $C = (c_{ij})$, we denote by $|B|$ the matrix of entries $|b_{ij}|$ and by $B \geq C$ if $b_{ij} \geq c_{ij}$ for all i and j. Given an n-vector $a = (a_i)$, we define

$$\min a = \min_i a_i \quad \text{and} \quad \max a = \max_i a_i$$

and

$$D_a = \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix}.$$

For a pair of vectors $a = (a_i)$ and $b = (b_i)$ with $b_i > 0$ for all i, we let

$$\max\left(\frac{a}{b}\right) = \max_i \left(\frac{a_i}{b_i}\right) \quad \text{and} \quad \min\left(\frac{a}{b}\right) = \min_i \left(\frac{a_i}{b_i}\right).$$

Throughout this article, we let e denote the column vector of all ones, i.e.,

$$e = (1, 1, \cdots, 1)^T.$$

A matrix A is called an M-matrix if it can be expressed in the form $A = sI - B$, $B \geq 0$ with $s \geq \rho(B)$, the Perron root of B, [5]. A matrix $A = (a_{ij})$ is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$ for all i. It is a scaled (or generalized) diagonally dominant if there exists $u > 0$ such that AD_u is diagonally dominant.

Note that any M-matrix A is scaled diagonally dominant; i.e., there exist $u > 0$ such that $Au = v \geq 0$. In many cases, the vector u may not be explicitly known. However, if u and v are known, the M-matrix A can be defined by its off-diagonal entries and u, v as in the following.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Definition 2.1. Let $P = (p_{ij})$ be an $n \times n$ nonnegative matrix with zero diagonal entries, and let $u = (u_i)$ be a positive n-vector and $v = (v_i)$ be a nonnegative n-vector. We use (P, u, v) to represent the unique matrix A of the form $A = D - P$ that satisfies $Au = v$, where D is a diagonal matrix. We write $A = (P, u, v)$.

In the representation $A = (P, u, v)$, the off-diagonal entries of A is given by $-P$ and its diagonal entries by

$$a_{ii} = \frac{v_i + \sum_{j \neq i} p_{ij} u_j}{u_i}.$$

Clearly, A is an M-matrix. On the other hand, any M-matrix can be represented in this way with suitable u, v. If $u = e$, A is a diagonally dominant M-matrix and v is the vector of its row sums (i.e., diagonally dominant part).

By treating (P, u, v) as the parameters representing A, it turns out that several quantities such as the entries of A^{-1} and the smallest eigenvalue of A are determined to high relative accuracy. The following lemma is such a result for the special case $u = e$ (see Theorem 2.1 of [3]).

Lemma 2.2. *Let $A = (P, e, v)$, $\widetilde{A} = (\widetilde{P}, e, \widetilde{v})$ and let λ and $\widetilde{\lambda}$ be the smallest eigenvalues of A and $\widetilde{A}$ respectively. If*

$$|P - \widetilde{P}| \leq \epsilon P, \quad \text{and} \quad |v - \widetilde{v}| \leq \epsilon v,$$

then

$$\frac{(1-\epsilon)^n}{(1+\epsilon)^{n-1}} A^{-1} \leq \widetilde{A}^{-1} \leq \frac{(1+\epsilon)^n}{(1-\epsilon)^{n-1}} A^{-1} \tag{2.1}$$

and

$$\frac{(1-\epsilon)^{n-1}}{(1+\epsilon)^n} \lambda \leq \widetilde{\lambda} \leq \frac{(1+\epsilon)^{n-1}}{(1-\epsilon)^n} \lambda. \tag{2.2}$$

The proof can be found in [3] and is omitted here.

Remark 2.3. We note that there are cases where the error bound (2.2) can be strengthened to $|\lambda - \widetilde{\lambda}|/\lambda \leq 2\epsilon$. Indeed it is our conjecture that this stronger bound holds generally.

We now generalize this result to a general M-matrix $A = (P, u, v)$. We will use it repeatedly in the error analysis later.

Lemma 2.4. *Let $A = (P, u, v)$, $\widetilde{A} = (\widetilde{P}, \widetilde{u}, \widetilde{v})$ and let λ and $\widetilde{\lambda}$ be the smallest eigenvalues of A and $\widetilde{A}$ respectively. If*

$$|P - \widetilde{P}| \leq \epsilon P, \quad |u - \widetilde{u}| \leq \epsilon u, \quad |v - \widetilde{v}| \leq \epsilon v,$$

then

$$\left| \frac{\lambda - \widetilde{\lambda}}{\lambda} \right| \leq 4n\epsilon + O(\epsilon^2), \tag{2.3}$$

$$|\widetilde{A}^{-1} - A^{-1}| \leq (4n\epsilon + O(\epsilon^2)) A^{-1}. \tag{2.4}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Proof. Let

$$B = AD_u = DD_u - PD_u \quad \text{and} \quad \widetilde{B} = \widetilde{A}\widetilde{D}_u = \widetilde{D}\widetilde{D}_u - \widetilde{P}\widetilde{D}_u.$$

Obviously, B and $\widetilde{B}$ are diagonally dominant M-matrices with $Be = v$ and $\widetilde{B}e = \widetilde{v}$. We have

$$(1-\epsilon)^2 PD_u \le \widetilde{P}\widetilde{D}_u \le (1+\epsilon)^2 PD_u$$

and

$$(1-\epsilon)Be \le \widetilde{B}e \le (1+\epsilon)Be.$$

From Lemma 2.2, we have

$$|\widetilde{B}^{-1} - B^{-1}| \le ((4n-1)\epsilon + O(\epsilon^2))B^{-1},$$

from which it follows

$$|\widetilde{A}^{-1} - A^{-1}| \le (4n\epsilon + O(\epsilon^2))A^{-1}.$$

This further implies

$$|\widetilde{A}^{-1} - A^{-1}| \le (4n\epsilon + O(\epsilon^2))\widetilde{A}^{-1}.$$

Using the perturbation result for spectral radius (Theorem 1 in [10]), we obtain

$$\left|\frac{1}{\widetilde{\lambda}} - \frac{1}{\lambda}\right| \le (4n\epsilon + O(\epsilon^2))\frac{1}{\widetilde{\lambda}},$$

i.e.,

$$\left|\frac{\widetilde{\lambda} - \lambda}{\lambda}\right| \le 4n\epsilon + O(\epsilon^2).$$

□

Lemma 2.5. *Let B be an M-matrix with the smallest eigenvalue λ. If*

$$\lambda_1 e \le Be \le \lambda_2 e$$

and $\lambda_1 > 0$, then

$$\lambda_1 \le \lambda \le \lambda_2.$$

Proof. B^{-1} is a nonnegative matrix with Perron eigenvalue $1/\lambda$. Obviously

$$\frac{1}{\lambda_2}e \le B^{-1}e \le \frac{1}{\lambda_1}e.$$

From the definition of Perron root in [5, Chapter 1],

$$\frac{1}{\lambda} = \max_{u \ge 0} \min \frac{B^{-1}u}{u} = \min_{v \ge 0} \max \frac{B^{-1}v}{v};$$

thus

$$\frac{1}{\lambda_2} \le \frac{1}{\lambda} \le \frac{1}{\lambda_1}$$

which completes the proof. □

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

3. Solving linear systems

In this section, we consider solving a linear system $Ax = b$ with $b \geq 0$. Lemma 2.1 shows that if $A = (P, u, v)$, then A^{-1} and hence the solution x to $Ax = b \geq 0$ are determined to the same accuracy entrywise as in the data (P, u, v). Thus, we can expect algorithms that solve $Ax = b$ to the machine precision entrywise. It turns out that this can be achieved by modifying the standard Gaussian elimination.

We first note that if A is a diagonally dominant M-matrix, then all the submatrices produced in the process of the Gaussian elimination are still diagonally dominant M-matrices and can all be represented in the representation (P, u, v). It turns out that we can carry out the Gaussian elimination in terms of the representation (P, u, v) rather than the entries of A and the advantage of this is that there is no subtraction involved throughout the process. In this way, the final solution obtained will have high relative accuracy. This idea is an extension of the GTH-algorithm [14] for stochastic matrices and has a similar algorithm. We therefore call this algorithm a GTH-like algorithm. Since such an extension has not been considered before for M-matrices, we present the detailed derivation and analysis in this section.

3.1. GTH-like algorithm.

We now derive the algorithm for $Ax = b$ with $A = (P, u, v)$ through LU factorization, forward substitution and backward substitution. All computations are operated on P, u, v and b.

We first consider the LU factorization of A, carried out without pivoting. This produces a series of matrices of decreasing order $A = A^{(1)}, A^{(2)}, A^{(3)}, \cdots$, where $A^{(k)}$ denotes the matrix to the southeast of the k-th pivot entry (and including that pivot entry), just before the k-th Gaussian elimination is applied. It is easily verified that $A^{(k)}$ inherits the property of being an M-matrix. We shall find out its representation $A^{(k)} = (P^{(k)}, u^{(k)}, v^{(k)})$. In the following, we let $p_{ij}^{(k)}$ be the $(i-k+1, j-k+1)$-th entry of $P^{(k)}$, and $u_i^{(k)}$ and $v_i^{(k)}$ be the $(i-k+1)$-th entries of $u^{(k)}$ and $v^{(k)}$ respectively. To seek the relation between $(P^{(k)}, u^{(k)}, v^{(k)})$ and $(P^{(k+1)}, u^{(k+1)}, v^{(k+1)})$, we partition $A^{(k)}$ as

$$A^{(k)} = \begin{pmatrix} \alpha_k & -w^T \\ -z & B^{(k)} \end{pmatrix},$$

where $B^{(k)}$ is of order $n-k$. We have

$$A^{(k+1)} = B^{(k)} - \frac{zw^T}{\alpha_k}. \tag{3.1}$$

For $i, j > k$ and $i \neq j$, $p_{ij}^{(k)}$ is the $(i-k, j-k)$-th entry of $B^{(k)}$, $p_{jk}^{(k)}$ and $p_{kj}^{(k)}$ are the $(j-k)$-th entries of z and w respectively. From the first row of the equation $A^{(k)}u^{(k)} = v^{(k)}$, we can get

$$\alpha_k = \frac{v_k^{(k)} + \sum_{j=k+1}^{n} p_{kj}^{(k)} u_j^{(k)}}{u_k^{(k)}}. \tag{3.2}$$

From (3.1), we can compute $P^{(k+1)}$ according to the relation

$$p_{ij}^{(k+1)} = p_{ij}^{(k)} + \frac{p_{ik}^{(k)} p_{kj}^{(k)}}{\alpha_k}, \quad \text{for} \quad i, j > k,\ i \neq j. \tag{3.3}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Now we show that $A^{(k+1)}$ is still a diagonally dominant M-matrix by finding $u^{(k+1)}$ and $v^{(k+1)}$. Let $\overline{u}^{(k)}$ and $\overline{v}^{(k)}$ be the respective subvectors of $u^{(k)}$ and $v^{(k)}$ with the first entries deleted. From

$$\begin{pmatrix} \alpha_k & -w^T \\ -z & B^{(k)} \end{pmatrix} \begin{pmatrix} u_k^{(k)} \\ \overline{u}^{(k)} \end{pmatrix} = \begin{pmatrix} v_k^{(k)} \\ \overline{v}^{(k)} \end{pmatrix},$$

we have

$$-w^T \overline{u}^{(k)} = -\alpha_k u_k^{(k)} + v_k^{(k)}$$

and

$$B^{(k)} \overline{u}^{(k)} = u_k^{(k)} z + \overline{v}^{(k)}.$$

Thus

$$A^{(k+1)} \overline{u}^{(k)} = B^{(k)} \overline{u}^{(k)} - \frac{w^T \overline{u}^{(k)}}{\alpha_k} z = \overline{v}^{(k)} + \frac{v_k^{(k)}}{\alpha_k} z.$$

We can choose

$$u^{(k+1)} = \overline{u}^{(k)} \quad \text{and} \quad v^{(k+1)} = \overline{v}^{(k)} + \frac{v_k^{(k)}}{\alpha_k} z,$$

i.e.,

$$u_j^{(k+1)} = u_j^{(k)} \quad \text{and} \quad v_j^{(k+1)} = v_j^{(k)} + \frac{v_k^{(k)}}{\alpha_k} p_{jk}^{(k)}, \quad j > k. \tag{3.4}$$

After computing α_n, we in fact have calculated the LU factors of A, which is stored in terms of α_k, $p_{kj}^{(k)}$ and $p_{jk}^{(k)}$ for $j > k$. Finally we perform forward and backward substitution to get the solution.

The following algorithm summarizes this new Gaussian elimination process.

Algorithm 1

Step 1: LU factorization:

For $k = 1, 2, \cdots, n-1$,

1. Calculate α_k according to (3.2)
2. Calculate $P^{(k+1)}$ according to (3.3)
3. Calculate $u^{(k+1)}$ and $v^{(k+1)}$ according to (3.4)

Step 2: Solving Ly=b:

$y_1 = b_1/\alpha_1$

For $k = 2, 3, \cdots n$,

1. $y_k = b_k + \sum_{j=1}^{k-1} p_{k,j}^{(j)} y_j$,
2. $y_k = y_k/\alpha_k$

Step 3: Solving Ux=y:

$x_n = y_n$

For $k = n-1, n-2, \cdots, 1$

1. $x_k = y_k + (\sum_{j=k+1}^{n} p_{kj}^{(k)} x_j)/\alpha_k$

3.2. **Error analysis.** Clearly, there is no subtraction involved in Algorithm 1. In this subsection we perform a priori rounding error analysis for the Algorithm 1 to demonstrate the computed solution x will have small relative error entrywise. Our analysis is parallel to the error analysis for the GTH algorithm performed by O'Cinneide [18].

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

We assume the following model for the floating point arithmetic [6, p. 9]:

$$fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \quad |\delta| \le \epsilon,$$

where op $= +, -, *$ or $/$ and ϵ is the machine roundoff unit. For the ease of notation, we shall use ϵ_s with subscripts to denote quantities bounded in magnitude by ϵ.

In the following, a "hat" is added to the value computed in floating-point arithmetic.

Theorem 3.1. *Suppose Algorithm 1 is carried out in a floating-point arithmetic to solve the linear system $Ax = b$, where $A = (P, u, v)$, $b \ge 0$ and the data p_{ij}, u_i, v_i, b_i $(i, j = 1, 2, \cdots, n)$ are floating-point numbers. Then the computed solution $\widehat{x}$ satisfies*

$$|x - \widehat{x}| \le (\phi(n)\epsilon + O(\epsilon^2))x, \tag{3.5}$$

where

$$\phi(n) = \frac{2(n+2)(n+3)(2n+5)}{3}.$$

Proof. Our proof is by induction on n. It is trivial to show bound (3.5) holds for $n = 1$. Suppose the theorem is true for linear systems of size $n-1$. We partition A as

$$A = \begin{pmatrix} \alpha_1 & -w^T \\ -z & B^{(1)} \end{pmatrix},$$

where $B^{(1)}$ is of order $n-1$. We have

$$A^{(2)} = B^{(1)} - \frac{zw^T}{\alpha_1} = (P^{(2)}, u^{(2)}, v^{(2)}).$$

The diagonal entry α_1 is computed as

$$\begin{aligned} \widehat{\alpha}_1 &= fl\left(\frac{v_1 + p_{12}u_2 + p_{13}u_3 + \cdots + p_{1n}u_n}{u_1}\right) \\ &= \alpha_1(1+\eta_1), \quad |\eta_1| \le (n+1)\epsilon + O(\epsilon^2). \end{aligned}$$

The off-diagonal entries of $P^{(2)}$ are computed with relative errors characterized by

$$\begin{aligned} \widehat{p}_{ij}^{(2)} &= fl\left(p_{ij} + \frac{p_{i1}p_{1j}}{\widehat{\alpha}_1}\right) \\ &= p_{ij}(1+\epsilon_3) + \frac{p_{i1}p_{1j}}{\widehat{\alpha}_1}(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3), \\ &= p_{ij}^{(2)}(1+\eta_2), \quad |\eta_2| \le (n+4)\epsilon + O(\epsilon^2). \end{aligned} \tag{3.6}$$

Similarly, for $i = 2, 3, \cdots, n$

$$\widehat{v}_i^{(2)} = v_i^{(2)}(1+\eta_3), \quad |\eta_3| \le (n+4)\epsilon + O(\epsilon^2). \tag{3.7}$$

Now the computed $A^{(2)}$ is $\widehat{A}^{(2)} = (\widehat{P}^{(2)}, u^{(2)}, \widehat{v}^{(2)})$. Let q and $\widehat{q}$ be the respective subvectors of x and $\widehat{x}$ from the second entry to the last one. It is easy to verify that q is the solution to the linear system

$$A^{(2)}q = a,$$

where the $(i-1)$-th entry a_i of a is

$$a_i = b_i + \frac{b_1 p_{i1}}{\alpha_1}.$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Now we are in a position to explain $\widehat{q}$. After LU factors of A are computed, we perform the forward substitution. Considering Step 2 in Algorithm 1, we have

$$\widehat{y}_1 = fl(b_1/\widehat{\alpha}_1) = y_1(1+\eta_4), \qquad |\eta_4| \le (n+2)\epsilon + O(\epsilon^2)$$

and for $j = 2, 3, \cdots, n$,

$$\begin{aligned} \widehat{y}_j &= fl\left(\frac{b_j + p_{j1}\widehat{y}_1 + \widehat{p}_{j2}^{(2)}\widehat{y}_2 + \cdots + \widehat{p}_{j,j-1}^{(j-1)}\widehat{y}_{j-1}}{\widehat{\alpha}_j}\right) \\ &= fl\left(\frac{\widehat{a}_j + \widehat{p}_{j2}^{(2)}\widehat{y}_2 + \cdots + \widehat{p}_{j,j-1}^{(j-1)}\widehat{y}_{j-1}}{\widehat{\alpha}_j}\right), \end{aligned}$$

where

$$\widehat{a}_j = fl(b_j + p_{j1}\widehat{y}_1) = a_j(1+\eta_5), \quad |\eta_5| \le (n+4)\epsilon + O(\epsilon^2).$$

Let $\widehat{a} = (\widehat{a}_i)$; then $\widehat{q}$ can be viewed as the computed solution to the linear system

$$\widehat{A}^{(2)} p = \widehat{a}$$

via Algorithm 1. From the induction hypothesis,

$$|\widehat{q} - p| \le (\phi(n-1)\epsilon + O(\epsilon^2))p.$$

It follows from Lemma 2.3 that

$$\left|(\widetilde{A}^{(2)})^{-1} - (A^{(2)})^{-1}\right| \le (4(n-1)(n+4)\epsilon + O(\epsilon^2))(A^{(2)})^{-1},$$

and thus

$$|p - q| \le ((4n-3)(n+4)\epsilon + O(\epsilon^2))q.$$

Therefore

$$|q - \widehat{q}| \le (\phi(n-1)\epsilon + (4n-3)(n+4)\epsilon + O(\epsilon^2))q,$$

i.e., for $2 \le i \le n$

$$\left|\frac{x_i - \widehat{x}_i}{x_i}\right| \le \phi(n-1)\epsilon + (4n-3)(n+4)\epsilon + O(\epsilon^2).$$

The first entry of x can be computed as

$$\begin{aligned} \widehat{x}_1 &= fl\left(b_1 + \frac{p_{12}\widehat{x}_2 + p_{13}\widehat{x}_3 + \cdots + p_{1n}\widehat{x}_n}{\widehat{\alpha}_1}\right) \\ &= x_1(1+\eta_6), \end{aligned}$$

where $|\eta_6| \le (\phi(n-1) + 4n^2 + 15n - 10)\epsilon + O(\epsilon^2)$. Noting that $\phi(n) \ge \phi(n-1) + 4n^2 + 15n - 10$, we obtain inequality (3.5) and complete the proof. □

Remark 3.2. We note that $\phi(n) \sim O(n^3)$ in this worst case bound seems to be pessimistic in some important aspects. A similar observation was made by O'Cinneide [18, 19] for the analysis of the GTH-algorithm. Here we note that, based on our floating point arithmetic model, the relative error in computing $\widetilde{\alpha}_k$ through an inner product is of order $O(n-k)\epsilon$, but in many implementations, the accumulation of inner product can be carried out in registers with longer digits and thus will have relative errors on the order of $O(1)\epsilon$. Hence, $\phi(n)$ can be reduced to $O(n^2)$ in such cases. Furthermore, the structure of matrix can also affect the bound. For example, if A is a banded matrix with bandwidth k, then $\phi(n)$ can be reduced to $O(kn^2)$. In our numerical test, $\phi(n)$ behaves like $O(n)$.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

4. Computing the smallest eigenvalue

In this section, we consider how to compute the smallest eigenvalue of a diagonally dominant irreducible M-matrix A, which is given in the representation $A = (P, u, v)$.

4.1. **The inverse iteration algorithm.** The algorithm to be developed here is based on the following inverse iteration shifted by a Rayleigh quotient like approximation of the eigenvalue.

Shifted Inverse Iteration:

- For a given $u^{(0)} > 0$, iteratively define

$$\begin{aligned} \lambda_s &= \min\left(\frac{Au^{(s)}}{u^{(s)}}\right), \\ w^{(s+1)} &= (A - \lambda_s I)^{-1} u^{(s)}, \\ u^{(s+1)} &= \frac{w^{(s+1)}}{\|w^{(s+1)}\|_\infty}. \end{aligned}$$

This inverse iteration algorithm was presented by Xue in [26] for M-matrices. It stems from the algorithm by Noda in [17] for computing the Perron root of an irreducible nonnegative matrix. Elsner [10] has shown that Noda's algorithm is quadratically convergent. Thus the above eigenvalue is increasing and quadratically convergent, i.e.,

$$\lambda_s \le \lambda_{s+1} \le \lambda \quad \text{and} \quad \lambda - \lambda_{s+1} \le \beta(\lambda - \lambda_s)^2,$$

where β is a constant depending on $u^{(0)}$ and A. It is noted in [26] that λ_{s+1} can be computed from λ_s without subtractions following the relation

$$\lambda_{s+1} = \min\left(\frac{Aw^{(s+1)}}{w^{(s+1)}}\right) = \min\left(\frac{u^{(s)} + \lambda_s w^{(s+1)}}{w^{(s+1)}}\right) = \lambda_s + \min\left(\frac{u^{(s)}}{w^{(s+1)}}\right).$$

The main task at each iterative step is to solve the linear system

$$(A - \lambda_s I) w^{(s+1)} = u^{(s)}, \tag{4.1}$$

where $A - \lambda_s I$ is an M-matrix. Indeed, the accuracy in forming and solving this system directly affects the final accuracy of the computed eigenvalue of the above algorithm. It is suggested in [11] to use Ahac-Olesky algorithm [1] followed by one step of iterative refinement to solve this linear system. Under some conditions, this method can produce an entrywise backward stable solution (see [23]). However the accuracy of the computed smallest eigenvalue still depends on its magnitude and certain condition number (see [26]). Here, we consider forming and solving (4.1) accurately through the GTH-like algorithm in section 3.

We note that the M-matrix $A - \lambda_s I$ is (scaled) diagonally dominant since $(A - \lambda_s I)u^{(s)} \ge 0$ by the definition $\lambda_s = \min\left(\frac{Au^{(s)}}{u^{(s)}}\right) > 0$ (this property is also observed and used by O'Cinneide [20]). Thus, the key idea is that $A - \lambda_s I$ can be represented without forming the diagonals as

$$A - \lambda_s I = (P, u^{(s)}, v^{(s)}),$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

where

$$\begin{aligned} v^{(s)} &= (A-\lambda_s I)u^{(s)} \\ &= \frac{1}{\|w^{(s)}\|_\infty}(Aw^{(s)}-\lambda_s w^{(s)}) \\ &= \frac{1}{\|w^{(s)}\|_\infty}(\lambda_{s-1}w^{(s)}+u^{(s-1)}-\lambda_s w^{(s)}) \\ &= \frac{1}{\|w^{(s)}\|_\infty}\left(u^{(s-1)}-\min\left(\frac{u^{(s-1)}}{w^{(s)}}\right)w^{(s)}\right)\geq 0. \end{aligned}$$

Hence, we shall form $A-\lambda_s I$ by the representation $(P,u^{(s)},v^{(s)})$ and then solve (4.1) with Algorithm 1. In this process, subtraction is encountered only in the computation of $v^{(s)}$. On account of possible cancellation, we cannot expect $w^{(s+1)}$ and $u^{(s+1)}$ are computed with small entrywise relative error. Fortunately, however, this will not affect the accuracy of the computed eigenvalue, which will be shown in the later error analysis.

Finally, for the stopping criterion, we adopt

$$\frac{\max\left(\frac{u^{(s)}}{w^{(s+1)}}\right)-\min\left(\frac{u^{(s)}}{w^{(s+1)}}\right)}{\lambda_{s+1}}\leq tol,$$

where *tol* is a small threshold. We will prove in the next subsection that the relative error of the approximate eigenvalue, when the above stopping criterion is satisfied, is no more than *tol*.

Our algorithm can be formulated as follows.

Algorithm 2

- Given $A=(P,u,v)$ and *tol*;
- set $u^{(0)}=u$, $\lambda_0=\min\left(\frac{v}{u}\right)$, $v^{(0)}=v-\lambda_0 u$.
- For $s=0,1,2,\cdots$
  1. Use the GTH-like algorithm to solve
  $$(P,u^{(s)},v^{(s)})w^{(s+1)}=u^{(s)}$$
  2.
  $$\lambda_{s+1}=\lambda_s+\min\left(\frac{u^{(s)}}{w^{(s+1)}}\right)$$
  3.
  $$u^{(s+1)}=\frac{w^{(s+1)}}{\|w^{(s+1)}\|_\infty}$$
  4. Calculate $v^{(s+1)}$ according to
  $$v_i^{(s+1)}=u_i^{(s+1)}\left(\frac{u_i^{(s)}}{w_i^{(s+1)}}-\min\left(\frac{u^{(s)}}{w^{(s+1)}}\right)\right)$$
  5. Proceed until
  $$\frac{\max\left(\frac{u^{(s)}}{w^{(s+1)}}\right)-\min\left(\frac{u^{(s)}}{w^{(s+1)}}\right)}{\lambda_{s+1}}\leq tol.$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Remark 4.1. This algorithm can be adapted to compute the smallest eigenvalue of an arbitrary M-matrix A by first finding a representation, i.e., finding a positive vector u such that $Au > 0$. After calculating $v = Au$, we apply Algorithm 2 to compute the smallest eigenvalue of (P, u, v), where $-P$ is the off-diagonal part of A. As for u, it can be obtained by solving the linear system $Au = e$. With a small residual, we can expect $A\widehat{u}$, where $\widehat{u}$ is the computed solution, to be positive.

4.2. **Error analysis.** In this section, we present a detailed error analysis for Algorithm 2. Again, we add "hats" to the computed intermediate quantities.

First note that theoretically (in the exact arithmetic), from $(A-\lambda_s I)u^{(s)} = v^{(s)}$, we have the representation of A in the s-th iteration step as

$$A = (P, u^{(s)}, \lambda_s u^{(s)} + v^{(s)}).$$

In the floating point arithmetic, let $\widehat{\lambda}_s$, $\widehat{u}^{(s)}$ and $\widehat{v}^{(s)}$ be the computed quantities at the s-th iteration. Then

$$A_s = (P, \widehat{u}^{(s)}, \widehat{\lambda}_s \widehat{u}^{(s)} + \widehat{v}^{(s)})$$

is an approximation to A. Because of possible cancellations in the computation of $v^{(s-1)}$, $\widehat{u}^{(s)}$ can be a bad approximation of $u^{(s)}$, and for this reason it cannot be expected that A_s approximate A with small entrywise relative error. What makes our algorithm work is that, no matter whether such cancellation occurs or not, the relative error between γ_s, the smallest eigenvalue of A_s, and λ, the smallest eigenvalue of A, is always small. To show this, we first investigate the relative error between γ_s and γ_{s+1}, which is caused by one step of iteration of Algorithm 2.

Lemma 4.2. *Let $\widehat{u}^{(s)}$, $\widehat{\lambda}_s$, $\widehat{v}^{(s)}$ and $\widehat{u}^{(s+1)}$, $\widehat{\lambda}_{s+1}$ $\widehat{v}^{(s+1)}$ be the computed quantities at the s-th and $(s+1)$-th iteration of Algorithm 2 respectively, and let γ_s and γ_{s+1} be the smallest eigenvalues of*

$$A_s = (P, \widehat{u}^{(s)}, \widehat{\lambda}_s \widehat{u}^{(s)} + \widehat{v}^{(s)}) \quad \textit{and} \quad A_{s+1} = (P, \widehat{u}^{(s+1)}, \widehat{\lambda}_{s+1} \widehat{u}^{(s+1)} + \widehat{v}^{(s+1)})$$

respectively. Then

$$\frac{|\gamma_s - \gamma_{s+1}|}{\gamma_{s+1}} \le \varphi(n)\epsilon + O(\epsilon^2),$$

where $\varphi(n) = 12n(\phi(n) + 3)$.

Proof. At the s-th step of finite precision iteration, let $u^{(s+1)}$, λ_{s+1} and $v^{(s+1)}$ be the quantities that are computed in the exact arithmetic from $\widehat{u}^{(s)}$, $\widehat{\lambda}_s$, $\widehat{v}^{(s)}$ for the $(s+1)$-th iteration step of Algorithm 2. Then, it can be checked that

$$A_s = (P, u^{(s+1)}, \lambda_{s+1} u^{(s+1)} + v^{(s+1)}). \tag{4.2}$$

To bound the relative error between γ_s and γ_{s+1}, it follows from Lemma 2.1 that it is sufficient to bound the entrywise relative errors between $u^{(s+1)}$ and $\widehat{u}^{(s+1)}$, and between $q^{(s+1)}$ and $\widehat{q}^{(s+1)}$, where

$$q^{(s+1)} = \lambda_{s+1} u^{(s+1)} + v^{(s+1)} \quad \text{and} \quad \widehat{q}^{(s+1)} = \widehat{\lambda}_{s+1} \widehat{u}^{(s+1)} + \widehat{v}^{(s+1)}.$$

Let $w^{(s+1)}$ be the solution to the linear system

$$(P, \widehat{u}^{(s)}, \widehat{v}^{(s)}) w^{(s+1)} = \widehat{u}^{(s)}.$$

From Theorem 3.1, the computed solution $\widehat{w}^{(s+1)}$ via Algorithm 1 satisfies

$$|\widehat{w}^{(s+1)} - w^{(s+1)}| \le (\phi(n)\epsilon + O(\epsilon^2)) w^{(s+1)}. \tag{4.3}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Noting that

$$u^{(s+1)} = \frac{w^{(s+1)}}{\|w^{(s+1)}\|_\infty} \quad \text{and} \quad \widehat{u}^{(s+1)} = fl\left(\frac{\widehat{w}^{(s+1)}}{\|\widehat{w}^{(s+1)}\|_\infty}\right),$$

we obtain

$$|u^{(s+1)} - \widehat{u}^{(s+1)}| \le ((2\phi(n)+1)\epsilon + O(\epsilon^2))u^{(s+1)}. \tag{4.4}$$

The respective i-th entries of $q^{(s+1)}$ and $\widehat{q}^{(s+1)}$ are

$$\begin{aligned} q_i^{(s+1)} &= \left(\widehat{\lambda}_s + \min\frac{\widehat{u}^{(s)}}{w^{(s+1)}}\right)u_i^{(s+1)} + \left(\frac{\widehat{u}_i^{(s)}}{w_i^{(s+1)}} - \min\frac{\widehat{u}^{(s)}}{w^{(s+1)}}\right)u_i^{(s+1)} \\ &= \left(\widehat{\lambda}_s + \frac{\widehat{u}_i^{(s)}}{w_i^{(s+1)}}\right)u_i^{(s+1)} \qquad (4.5) \end{aligned}$$

and

$$\begin{aligned} \widehat{q}_i^{(s+1)} &= fl\left(\widehat{\lambda}_s + \min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right)\widehat{u}_i^{s+1} + fl\left(\widehat{u}_i^{(s+1)}\left(\frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} - \min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right)\right) \\ &= (1+\epsilon_2)\left(\widehat{\lambda}_s + (1+\epsilon_1)\min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right)\widehat{u}_i^{(s+1)} \\ &\quad +(1+\epsilon_5)(1+\epsilon_6)\widehat{u}_i^{(s+1)}\left((1+\epsilon_3)\frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} - (1+\epsilon_4)\min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right) \\ &= \widehat{u}_i^{(s+1)}\left(\widehat{\lambda}_s + \frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} + \epsilon_2\widehat{\lambda}_s + \eta_1\frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} + \eta_2\min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right), \qquad (4.6) \end{aligned}$$

where

$$\eta_1 = \epsilon_3 + \epsilon_5 + \epsilon_6 + O(\epsilon^2)$$

and

$$\eta_2 = \epsilon_1 + \epsilon_2 - \epsilon_4 - \epsilon_5 - \epsilon_6 + O(\epsilon^2).$$

It is straightforward to show that

$$\left|\eta_1\frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} + \eta_2\min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right| \le (8\epsilon + O(\epsilon^2))\frac{\widehat{u}_i^{(s)}}{w_i^{(s+1)}}.$$

Plugging the bounds (4.3) and (4.5) into (4.6), we have

$$\widehat{q}_i^{(s+1)} = (1+\eta_3)q_i^{(s+1)}, \quad |\eta_3| \le 3(\phi(n)+3)\epsilon + O(\epsilon^2).$$

Thus

$$|q^{(s+1)} - \widehat{q}^{(s+1)}| \le (3(\phi(n)+3)\epsilon + O(\epsilon^2))q^{(s+1)}. \tag{4.7}$$

Associating (4.4) and (4.7) with Lemma 2.1 yields

$$0 < \frac{\gamma_{s+1} - \gamma_s}{\gamma_s} \le 12n(\phi(n)+3)\epsilon + O(\epsilon^2).$$

□

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Remark 4.3. The bound is based on those established in Lemma 2.2 and Theorem 3.1 that may be pessimistic (see the remarks there). If our conjecture for Lemma 2.2 is true and Algorithm 1 is implemented with more accurate inner product accumulations, we would have $\varphi(n) \sim O(n^2)$. In our numerical tests, there is no excessive growth in n observed.

By applying a similar proof to

$$A = (P, u^{(0)}, \lambda_0 u^{(0)} + v^{(0)}) \quad \text{and} \quad A_0 = (P, \widehat{u}^{(0)}, \widehat{\lambda}_0 \widehat{u}^{(0)} + \widehat{v}^{(0)}),$$

it can also be shown that the relative error between λ and γ_0 is bounded by $12n(\phi(n) + 3)\epsilon + O(\epsilon^2)$. Thus applying Lemma 4.1 for s steps of the iteration, the relative error between γ_s and λ is bounded as

$$\left| \frac{\gamma_s - \lambda}{\lambda} \right| \leq 12(s+1)\varphi(n)\epsilon + O(\epsilon^2). \tag{4.8}$$

Combining this lemma with the stopping criterion, we can evaluate the accuracy of the computed eigenvalue.

Theorem 4.4. *Assume that Algorithm 2 terminates after s steps of iteration and let $\widehat{\lambda}_{s+1}$ be the computed approximate eigenvalue obtained. Then we have*

$$\left| \frac{\widehat{\lambda}_{s+1} - \lambda}{\lambda} \right| \leq tol + \psi(n, s)\epsilon + O(\epsilon^2), \tag{4.9}$$

where $\psi(n, s) = 12(s+1)n\varphi(n) + 2\phi(n) + 36(s+1)n + 4$.

Proof. We first bound the relative error between $\widehat{\lambda}_{s+1}$ and γ_s. In finite precision, the stopping criterion is

$$fl\left(\frac{\max \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} - \min \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}}{\widehat{\lambda}_{s+1}} \right) \leq tol$$

and this gives

$$\frac{(1+\epsilon_3)(1+\epsilon_4)\left((1+\epsilon_1) \max \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} - (1+\epsilon_2) \min \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} \right)}{\widehat{\lambda}_{s+1}} \leq tol$$

which can be written as

$$\max \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} \leq (1+\eta_4) \min \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} + \eta_5 \widehat{\lambda}_{s+1} \tag{4.10}$$

where $|\eta_4| \leq 2\epsilon + O(\epsilon^2)$ and $|\eta_5| \leq tol + O(\epsilon)$. Letting $w^{(s+1)}$ be the solution to the linear system

$$(P, \widehat{u}^{(s)}, \widehat{v}^{(s)}) w^{(s+1)} = \widehat{u}^{(s)}, \quad \text{i.e.,} \quad (A_s - \widehat{\lambda}_s I) w^{(s+1)} = \widehat{u}^{(s)},$$

from Theorem 3.1 we have

$$|w^{(s+1)} - \widehat{w}^{(s+1)}| \leq (\phi(n)\epsilon + O(\epsilon^2)) w^{(s+1)}.$$

Substituting this into (4.10) yields

$$\max \frac{\widehat{u}^{(s)}}{w^{(s+1)}} \leq (1+\eta_6) \min \frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} + \eta_7 \widehat{\lambda}_{s+1},$$

where $|\eta_6| \leq (\phi(n) + 2)\epsilon + O(\epsilon^2)$ and $|\eta_7| \leq tol + \phi(n)\epsilon + O(\epsilon^2)$. Let

$$B_s = D_{w^{(s+1)}}^{-1} A_s D_{w^{(s+1)}}.$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Then γ_s is the smallest eigenvalue of B_s and

$$(B_s - \widehat{\lambda}_s I)e = D^{-1}_{w^{(s+1)}}\widehat{u}^{(s)}.$$

Since

$$\widehat{\lambda}_{s+1} = fl\left(\widehat{\lambda}_s + \min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right) = (1+\eta_8)\left(\widehat{\lambda}_s + \min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right),$$

where $|\eta_8| \le 2\epsilon + O(\epsilon^2)$, we have

$$\begin{aligned} B_s e &\le \left(\widehat{\lambda}_s + \max\frac{\widehat{u}^{(s)}}{w^{(s+1)}}\right)e \\ &\le \left(\widehat{\lambda}_s + (1+\eta_6)\min\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}} + \eta_7\widehat{\lambda}_{s+1}\right)e \\ &= (1+\eta_9)\widehat{\lambda}_{s+1}e, \end{aligned}$$

where $|\eta_9| \le |\eta_6| + |\eta_7| + |\eta_8| \le tol + 2(\phi(n)+2)\epsilon + O(\epsilon^2)$. Similarly we can prove that

$$B_s e \ge (1+\eta_{10})\widehat{\lambda}_{s+1}e, \quad |\eta_{10}| \le (\phi(n)+2)\epsilon + O(\epsilon^2).$$

It follows from Lemma 2.2 that

$$(1+\eta_{10})\widehat{\lambda}_{s+1} \le \gamma_s \le (1+\eta_9)\widehat{\lambda}_{s+1},$$

and hence

$$\left|\frac{\widehat{\lambda}_{s+1} - \gamma_s}{\gamma_s}\right| \le tol + 2(\phi(n)+2)\epsilon + O(\epsilon^2).$$

Combining this with (4.8) completes the proof. □

Next, we show that the quadratic convergence behaviour is still valid for $\widehat{\lambda}_s$ in finite precision.

Theorem 4.5. *Let $\widehat{u}^{(s)}$, $\widehat{\lambda}_s$, $\widehat{v}^{(s)}$ and $\widehat{u}^{(s+1)}$, $\widehat{\lambda}_{s+1}$ $\widehat{v}^{(s+1)}$ be the computed quantities at the s-th and $(s+1)$-th iteration of Algorithm 2, respectively. Then*

$$\left|\frac{\widehat{\lambda}_{s+1} - \lambda}{\lambda}\right| \le (\beta_s\lambda)\left(\frac{\widehat{\lambda}_s - \lambda}{\lambda}\right)^2 + (\phi(n) + s\varphi(n))\epsilon + O(\epsilon^2), \tag{4.11}$$

where β_s is a constant depending on A_s and $\widehat{u}^{(s)}$.

Proof. In the computation of $\widehat{v}^{(s)}$, we have $\widehat{v}_i^{(s)} = 0$ if

$$\frac{\widehat{u}_i^{(s)}}{\widehat{w}_i^{(s+1)}} = \min\left(\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right).$$

Noting that $A_s\widehat{u}^{(s)} = \widehat{\lambda}_s\widehat{u}^{(s)} + \widehat{v}^{(s)}$ and at least one entry of $\widehat{v}^{(s)}$ is zero, we have

$$\widehat{\lambda}_s = \min\frac{A_s\widehat{u}^{(s)}}{\widehat{u}^{(s)}}.$$

Let λ_{s+1} be the quantity that is computed in the exact arithmetic from $\widehat{u}^{(s)}$, $\widehat{\lambda}_s$, $\widehat{v}^{(s)}$ by the $(s+1)$-th iteration step of Algorithm 2. We have

$$\lambda_{s+1} = \widehat{\lambda}_s + \min\left(\frac{\widehat{u}^{(s)}}{w^{(s+1)}}\right) \quad \text{and} \quad \widehat{\lambda}_{s+1} = fl\left(\widehat{\lambda}_s + \min\left(\frac{\widehat{u}^{(s)}}{\widehat{w}^{(s+1)}}\right)\right).$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

From Theorem 3.1,

$$\left|\frac{\widehat{\lambda}_{s+1}-\lambda_{s+1}}{\lambda_{s+1}}\right| \le (\phi(n)+2)\epsilon + O(\epsilon^2). \tag{4.12}$$

Because Algorithm 2 is quadratically convergent in the exact arithmetic, we have

$$|\gamma_s - \lambda_{s+1}| \le \beta_s(\gamma_s - \widehat{\lambda}_s)^2, \tag{4.13}$$

where β_s depends on A_s and $\widehat{u}^{(s)}$ (see [10]). Noting the fact that γ_s, λ_{s+1} and $\widehat{\lambda}_s$ are less than $(1+O(\epsilon))\lambda$ and substituting (4.8), (4.13) and (4.12) into the inequality

$$|\widehat{\lambda}_{s+1}-\lambda| \le |\widehat{\lambda}_{s+1}-\lambda_{s+1}| + |\lambda_{s+1}-\gamma_s| + |\gamma_s-\lambda|,$$

we obtain the bound. □

Remark 4.6. Theorem 4.3 essentially shows that the relative error of the computed approximation $\widehat{\lambda}_s$ converges quadratically until it reaches the level of $(\phi(n)+s\varphi(n))\epsilon$. On the other hand, by Theorem 4.2, the occurrence of convergence can be detected by the stopping criterion; i.e., when terminated, the relative error is approximately $\max\{tol, \psi(n,s)\epsilon\}$. Note that this accuracy is independent of λ or any condition number. Thus, to efficiently terminate the iterations, *tol* should be chosen according to the accuracy desired but not smaller than $(\phi(n)+s\varphi(n))\epsilon$. In practice, however, one can expect the factor $\phi(n)+s\varphi(n)$ in the bound to be pessimistic and we found choosing $tol = 100\epsilon$ works well in all of our tests.

Remark 4.7. For a general M-matrix A, for which the diagonal dominance is not explicitly given, we can first find a representation $A = (P, u, v)$, as in the remark at the end of Section 4.1, and then apply Algorithm 2. In this computed representation, $v = fl(Au) = (A+\delta A)u$, with $|\delta A| \le (n\epsilon + O(\epsilon^2))|A|$. Then we have computed the smallest eigenvalue of (P, u, v) and hence that of $A+\delta A$ with small relative errors. Therefore, the smallest eigenvalue of A computed in this way has a mixed stability in the sense that it produces an eigenvalue which approximates the smallest eigenvalue of a slightly entrywise perturbed matrix with tiny relative error.

5. Numerical examples

In this section, we present the results of numerical experiments. Our experiments were done using Matlab with $\epsilon \approx 2.2\times 10^{-16}$. We have extensively tested Algorithm 2 with $tol = 100\epsilon$ on various diagonally dominant M-matrices and compared the results with those produced by the standard QR algorithm of Matlab. In all of our tests, Algorithm 2 converges with the termination criterion satisfied. For those diagonally dominant M-matrix whose smallest eigenvalues are well conditioned and not tiny, both algorithms compute the smallest eigenvalue to high relative accuracy. However, if the smallest eigenvalues are ill conditioned or tiny, the standard QR algorithm has low relative accuracy or even no accuracy at all, while Algorithm 2 still produces high accuracy approximations. The following are two sets of such examples.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

TABLE 1. Case 1: λ is ill-conditioned

δ	λ	$y^T x$	$\frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lambda}$	$\frac{\lvert\lambda-\lambda_{QR}\rvert}{\lambda}$
1.0×10^{-3}	6.67×10^{-2}	1.1×10^{-3}	4.2×10^{-16}	4.7×10^{-13}
1.0×10^{-6}	1.29×10^{-1}	1.15×10^{-6}	4.3×10^{-16}	2.06×10^{-11}
1.0×10^{-9}	1.87×10^{-1}	1.2×10^{-9}	5.9×10^{-16}	4.7×10^{-10}
1.0×10^{-12}	2.41×10^{-1}	1.3×10^{-12}	0	4.9×10^{-9}
1.0×10^{-18}	3.39×10^{-1}	1.5×10^{-18}	0	3.5×10^{-5}
1.0×10^{-24}	4.25×10^{-1}	1.7×10^{-24}	1.8×10^{-15}	8.0×10^{-2}
1.0×10^{-30}	4.99×10^{-1}	2.0×10^{-30}	0	1.7×10^{-1}

Example 1. Consider the $n\times n$ M-matrix $A=(P,e,v)$, where

$$P=\begin{pmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ \delta & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{5.1}$$

and

$$v=(0,0,\cdots,0,1-\delta), \tag{5.2}$$

i.e. $A=I-P$. Let x and y be the unit right and left eigenvectors corresponding to λ, the smallest eigenvalue of A. It is shown in [11] that $\lambda=1-\delta^{\frac{1}{n}}$ and $y^Tx\approx\delta^{\frac{n-1}{n}}$. If δ is tiny, then λ is ill-conditioned since y^Tx is very close to zero. If δ tends to 1, y^Tx is not small and thus the smallest eigenvalue is well-conditioned, but it is tiny.

We test our algorithm and the QR algorithm on both cases of such M-matrices. In the following, we let $\widehat{\lambda}$ and λ_{QR} denote the smallest eigenvalues computed by Algorithm 2 and the QR algorithm respectively.

Case 1. $n=100$ and $\delta\sim0$. Table 1 presents the results. As it shows, Algorithm 2 computes the smallest eigenvalues almost to full precision no matter how small δ is. On the other hand, the QR algorithm loses significant figures as δ decreases. If $\delta\le10^{-24}$, only one figure of the computed eigenvalue is correct.

For this example, we also plot convergence history of $|\widetilde{\lambda}_s-\lambda|/\lambda$ against s in Figure 1 (in solid line for $\delta=10^{-3}$ and in dotted line for $\delta=10^{-9}$). It clearly shows the quadratic convergence property in finite precision as demonstrated by Theorem 4.3.

Case 2. $n=20$ and $\delta\sim1$. We report the results in Table 2. Again, our algorithm can compute λ to full precision no matter how tiny it is, while QR algorithm has low accuracy as δ decreases.

The matrices in Example 1 are very sparse. Next we consider testing on dense matrices.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

TABLE 2. Case 2: λ is tiny relative to $\|A\|$

δ	λ	$\frac{\lvert\lambda-\hat{\lambda}\rvert}{\lambda}$	$\frac{\lvert\lambda-\lambda_{QR}\rvert}{\lambda}$
$(1-10^{-3})^{20}$	10^{-3}	2.2×10^{-16}	3.3×10^{-13}
$(1-10^{-6})^{20}$	10^{-6}	4.2×10^{-16}	8.1×10^{-10}
$(1-10^{-9})^{20}$	10^{-9}	2.1×10^{-16}	8.6×10^{-7}
$(1-10^{-12})^{20}$	10^{-12}	0	2.4×10^{-4}
$(1-10^{-15})^{20}$	10^{-15}	2.0×10^{-16}	3.1

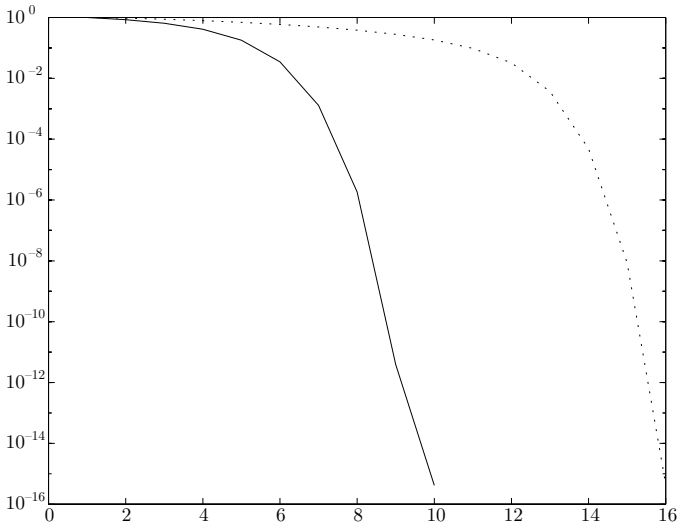

FIGURE 1. Example 1: Quadratic Convergence solid - $\delta = 10^{-3}$; dotted - $\delta = 10^{-9}$

Example 2. Consider the $n \times n$ M-matrix $A = (P, e, v)$ defined by

$$v = (\delta, \cdots, \delta, 65\delta/128, 191\delta/128)^T$$

and

$$P = \begin{pmatrix} P_1 & w \\ u^T & 0 \end{pmatrix}.$$

where P_1 is of order $n-1$ with all the off-diagonal entries equal to 1,

$$u = (0, \cdots, 0, \delta/128)^T \quad w = (0, \cdots, 0, \delta/2)^T.$$

The smallest eigenvalue of this matrix is δ and the corresponding eigenvector is $(1, \cdots, 1, 1/64)^T$.

We test our algorithm with various n and δ. Tables 3 and 4 reports the numerical results for $n = 100$ and $n = 1000$. We observe that, for these dense matrices, increase of n does not affect the accuracy of computed eigenvalue $\widetilde{\lambda}$ either.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

TABLE 3. 100×100 dense matrix

δ	λ	$\frac{\lvert\lambda-\hat{\lambda}\rvert}{\lambda}$	$\frac{\lvert\lambda-\lambda_{QR}\rvert}{\lambda}$
10^{-3}	10^{-3}	2.2×10^{-16}	4.9×10^{-12}
10^{-6}	10^{-6}	4.2×10^{-16}	2.5×10^{-8}
10^{-9}	10^{-9}	6.2×10^{-16}	2.3×10^{-6}
10^{-12}	10^{-12}	0	5.5×10^{-2}
10^{-15}	10^{-15}	3.9×10^{-16}	1

TABLE 4. 1000×1000 dense matrix

δ	λ	$\frac{\lvert\lambda-\hat{\lambda}\rvert}{\lambda}$	$\frac{\lvert\lambda-\lambda_{QR}\rvert}{\lambda}$
10^{-3}	10^{-3}	2.2×10^{-16}	2.5×10^{-10}
10^{-6}	10^{-6}	8.5×10^{-16}	5.8×10^{-8}
10^{-9}	10^{-9}	8.3×10^{-16}	1.3×10^{-4}
10^{-12}	10^{-12}	2.0×10^{-16}	2.1×10^{-2}
10^{-15}	10^{-15}	5.9×10^{-16}	2.2×10^{-1}

REFERENCES

[1] A. Ahac and D. Olesky, *A stable method for the LU factorization of M-matrices,* SIAM J. Alg. Disc. Meth., 7(1986):368-378. MR **87i:**65035

[2] J. Abate, L. Choudhury and W. Whitt, *Asymptotics for steady-state tail probabilities in structured Markov queueing models,* Commun. Statist-Stochastic Models, 1(1994):99-143. MR **95f:**60106

[3] A.S. Alfa, J. Xue and Q. Ye, *Entrywise perturbation theory for diagonally dominant M-matrices with applications,* to appear in Numer. Math.

[4] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices,* SIAM J. Num. Anal., 27(3), (1990):762-791. MR **91g:**65071

[5] A. Berman and R. Plemmons, *Nonnegative matrices in mathematical science,* Academic Press, New York, 1979. MR **82b:**15013

[6] S. Conte and C. de Boor, *Elementary numerical analysis* Third edition, New York-Tokyo:McGraw-Hill 1980.

[7] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices,* SIAM J. Sci. Stat. Comput, 11(5)(1990):873-912. MR **91i:**65072

[8] J. Demmel, *Accurate SVDs of structured matrices,* LAPACK Working Note #130, Dept. of Computer Science, Univ. of Tennessee, Oct. 1977.

[9] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić and Z. Dramč, *Computing the singular value decomposition with high relative accuracy,* Linear Algebra Appl. 299(1999):21-80. MR **2000j:**65044

[10] L. Elsner, *Inverse iteration for calculating the spectral radius of a nonnegative irreducible matrix,* Linear Algebra Appl., 15(1976):235-242. MR **56:**12026

[11] L. Elsner, I. Koltracht, M.Neumann and D.Xiao, *On accurate computations of the Perron root,* SIAM J. Matrix Anal. Appl., 14(1993):456-467. MR **93m:**65057

[12] E. Falkenberg, *On the asymptotic behaviour of the stationary distribution of Markov chains of M/G/1 type,* Commun. Statist.-Stochastic Models, 10(1994):75-97. MR **94i:**60111

[13] G. Golub and C. Van Loan, *Matrix Computations,* 2nd ed., The Johns Hopkins University Press, Baltimore, MD,1989. MR **90d:**65055

[14] W.K. Grassmann, M.J. Taksar and D.P. Heyman, *Regenerative analysis and steady-state distributions for Markov chains,* Oper. Res., 33(1985):1107-1116. MR **82k:**60125

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

[15] R. Gregory and D. Karney, *A collection of matrices for testing computational Algorithm,* Wiley, New York, 1969. MR **40:**6752

[16] N.J. Higham *A survey of componentwise perturbation theory in numerical linear algebra,* Proc. Sympos. Appl. Math., Amer. Math. Soc., Providence, R.I., 48:49-77, 1994. MR **96a:**65065

[17] T. Noda, *Note on the computation of maximal eigenvalue of a nonnegative irreducible matrix,* Numer. Math., 17(1971):382-386. MR **45:**4620

[18] C. O'Cinneide, *Entrywise perturbation theory and error analysis for Markov chains,* Numer. Math., 65(1993):109-120.

[19] C. O'Cinneide, *Relative-error bounds for the LU decomposition via the GTH algorithm,* Numer. Math., 73(1996):507-519. MR **97h:**65033

[20] C. O'Cinneide, Private communication.

[21] G. Robertazzi *Computer networks and systems: queueing theory and performance evaluation*, Springer-Verlag, 1990.

[22] P. Shivakumar, J. Williams, Q. Ye, C. Marinov, *On two-sided bounds related to weakly diagonally dominant M-matrices with applications to digital circuit dynamics*, SIAM J. Matrix Anal. Appl. 17(1996):298-312. MR **97c:**15030

[23] R. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination,* Math. Comp. 35:817-832, 1980. MR **83e:**65058

[24] J.H. Wilkinson, *The algebraic eigenvalue problem,* Oxford University Press, 1965. MR **32:**1894

[25] J.Xue and E.Jiang, *Entrywise relative perturbation theory for nonsingular M-matrices and applications,* BIT, 35(1995):417-427. MR **97k:**65070

[26] J.Xue *Computing the smallest eigenvalue of an M-matrix,* SIAM J. Matrix Anal. Appl., 17(1996):748-762. MR **97g:**65081

Department of Industrial and Manufacturing Systems Engineering, University of Windsor, Windsor, Ontario, Canada N9B 3P4

E-mail address: alfa@uwindsor.ca

Fakultaet fuer Mathematik, Technishe Universitaet Chemnitz, Reichenhainer Str. 41, 09126 Chemnitz, Germany

Current address: Department of Industrial and Manufacturing Systems Engineering, University of Windsor, Windsor, Ontario, Canada N9B 3P4

E-mail address: jxue@server.uwindsor.ca

Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506-0027

E-mail address: qye@ms.uky.edu

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use