MATHEMATICS OF COMPUTATION
Volume 77, Number 264, October 2008, Pages 2195–2230
S 0025-5718(08)02112-1
Article electronically published on May 5, 2008

COMPUTING SINGULAR VALUES OF DIAGONALLY DOMINANT MATRICES TO HIGH RELATIVE ACCURACY

QIANG YE

Abstract. For a (row) diagonally dominant matrix, if all of its off-diagonal entries and its diagonally dominant parts (which are defined for each row as the absolute value of the diagonal entry subtracted by the sum of the absolute values of off-diagonal entries in that row) are accurately known, we develop an algorithm that computes all the singular values, including zero ones if any, with relative errors in the order of the machine precision. When the matrix is also symmetric with positive diagonals (i.e. a symmetric positive semi-definite diagonally dominant matrix), our algorithm computes all eigenvalues to high relative accuracy. Rounding error analysis will be given and numerical examples will be presented to demonstrate the high relative accuracy of the algorithm.

1. Introduction

In this paper, we are concerned with high relative accuracy algorithms for computing singular values of diagonally dominant matrices and for computing eigenvalues of symmetric diagonally dominant matrices with positive diagonals. For the matrix eigenvalue problem (or the singular value problem), it is well-known that the standard algorithms such as the QR algorithm are norm-wise backward stable, and then smaller eigenvalues may be computed with lower relative accuracy. We note that, for a general matrix, smaller eigenvalues are not determined by its entries to the same relative accuracy as in the matrix data. Then, even though an algorithm might be able to compute the smaller eigenvalues of the matrix stored in memory to high relative accuracy by, for example, using higher precision arithmetic, such high accuracy is not necessarily warranted by the data (i.e. the entries).

Starting with a work by Demmel and Kahan [10] on computing singular values of bidiagonal matrices, the research on high relative accuracy algorithms has flourished. Special matrices with certain structure or properties have been identified for which the singular values or eigenvalues are determined and can be computed to high relative accuracy. Among them are well-scalable symmetric positive definite matrices [13] and matrices that admit accurate rank-revealing factorizations [8], which will form the basis of our works. For such matrices, the Jacobi method can be used to efficiently compute, respectively, the smaller eigenvalue and singular values to high relative accuracy. Some other examples are bidiagonal and acyclic matrices [4, 7, 15], diagonally scaled totally unimodular matrices [8], and totally

Received by the editor April 9, 2007 and, in revised form, October 4, 2007.
2000 *Mathematics Subject Classification.* Primary 65F18, 65F05.
This research was supported in part by NSF under Grant DMS-0411502.

non-negative matrices [8, 14, 20] (including Cauchy and Vandermonde [6]). There have also been works on several special diagonally dominant matrices. Specifically, entrywise perturbation analysis and algorithms have been developed for the eigenvalues of so-called γ-scaled symmetric diagonally dominant matrices in [3], for the smallest eigenvalue of a diagonally dominant M-matrix in [1, 2], and for all singular values of a diagonally dominant M-matrix in [11]. We note that one idea that has played a crucial role in several recent works and in this work as well is the need to re-parameterize matrices in some cases in order to obtain high relative accuracy algorithms; see [1, 2, 6, 11, 12, 14, 21].

In this paper, we consider a (row) diagonally dominant matrix with a representation by its off-diagonal entries and its diagonally dominant parts (which are defined for each row as the absolute value of the diagonal entry subtracted by the sum of the absolute values of off-diagonal entries in that row). When the off-diagonal entries and the diagonally dominant parts are accurately known, we present an algorithm with a forward error analysis that computes all singular values with relative errors in the order of machine precision. As a byproduct, zero singular values, if any, and ranks are computed exactly. When the matrix is also symmetric with positive diagonals (i.e. a symmetric positive definite diagonally dominant matrix), the algorithm computes all eigenvalues to high relative accuracy. Our algorithm is based on computing an accurate LDU-factorization and then using the algorithm of Demmel et al. [8]. Comparing our algorithms with existing ones for diagonally dominant matrices, we note that the relative accuracy of eigenvalues computed in [3] for a γ-scaled symmetric diagonally dominant matrix still depends on a certain condition number intrinsic to the matrix, and the algorithms in [2, 11] are valid for M-matrices only.

While our forward error analysis demonstrates the ability of the new algorithm to compute singular values to the order of machine precision independent of any condition number, the error bounds are weak in that they depend on the matrix dimension exponentially. Nevertheless, such a dependence is not present in our numerical testing. It remains to be seen whether this bound can be improved. On the other hand, the forward error analysis also implies a relative perturbation bound for the singular values and singular vectors, but again the bounds are weak. So, deriving a strong relative perturbation for the singular values is an important open problem. In the case of symmetric positive definite diagonally dominant matrices, however, we have recently obtained some sharp relative perturbation bounds for their eigenvalues in [28].

We remark that diagonally dominant matrices are a class of matrices that arise in many applications. Indeed, the diagonal dominance is often a natural physical characteristic of practical problems. One example of the eigenvalue problem for diagonally dominant matrices is the finite difference discretizations of elliptic differential operators [26, p. 211]. Here, the eigenvalues that are usually of physical interest and are well approximated by the discretization are the smaller ones. However, as the mesh size decreases, the condition number of the discretization matrix increases, and so does the relative error of a smaller eigenvalue computed. Therefore, methods for computing smaller eigenvalues of such matrices to high relative accuracy would be of great interest.

The rest of this paper is organized as follows. In Section 2 we first give some definitions and notation. Section 3 presents our main algorithm for an accurate LDU

factorization and then an accurate SVD. Section 4 gives an error analysis of the LDU factorization algorithm. We separate out the technical proofs in three subsections within this section. We then present some numerical examples in Section 5, with some concluding remarks in Section 6.

2. Preliminaries and notation

Definition 1. Given an $n \times n$ matrix $M = (m_{ij})$ with zero diagonals and an n-vector $v = (v_i)$, we use $\mathcal{D}(M, v)$ to denote the matrix $A = (a_{ij})$ whose off-diagonal entries are the same as M and whose i-th diagonal entry is $a_{ii} = v_i + \sum_{j \neq i} |m_{ij}|$. Namely, we write $A = \mathcal{D}(M, v)$, if

$$a_{ij} = m_{ij} \text{ for } i \neq j; \text{ and } a_{ii} = v_i + \sum_{j \neq i} |m_{ij}|.$$

$\mathcal{D}$ can be considered as a function that maps a matrix M and a vector v to a matrix A. Now, given a matrix $A = [a_{ij}]$, we denote by A_D the matrix whose off-diagonal entries are the same as A and whose diagonal entries are zero. Then, letting $v_i = a_{ii} - \sum_{j \neq i} |a_{ij}|$ and $v = (v_1, v_2, \cdots, v_n)^T$, which will be called the diagonally dominant parts of A, we have

$$A = \mathcal{D}(A_D, v). \tag{1}$$

Therefore, the pair (A_D, v) provides a representation of the matrix A, which will be called a representation by diagonally dominant parts. In this way, the off-diagonal entries and diagonally dominant parts are the data (parameters) that defines the matrix A.

The need to represent the matrix in this way is, as we will see, that the singular values of diagonally dominant matrices can be computed to high relative accuracy from (A_D, v) but not from the entries of A. Relative perturbation bounds have been obtained in [28] to show that the eigenvalues of a symmetric diagonally dominant matrix are determined by (A_D, v) with the same relative accuracy. We also note that, when $v_i \geq 0$, the entries of A can be computed from (A_D, v) accurately, but the converse is not true. In other words, once we store the entries of A in memory which encounters roundoff errors, it will not be possible to recover v from the matrix A in memory to the machine accuracy. Therefore, (A_D, v) should be obtained from A in its exact form, and one may need to go back a few steps in the formation of A to form v accurately. In many applications, v is comprised of (physical) parameters that defines A and is given naturally.

A matrix $A = (a_{ij})$ is said to be diagonally dominant if $a_{ii} \geq \sum_{j \neq i} |a_{ij}|$ for all i. Then a matrix A, represented by $\mathcal{D}(A_D, v)$, is diagonally dominant if and only if $v_i \geq 0$. Note that this definition of diagonally dominant matrices requires the diagonal entries to be nonnegative and is more restrictive than the customary definition. However, this does not impose any restriction for computing singular values of diagonally dominant matrices with negative diagonals, as we can multiply the matrix by a diagonal matrix of ± 1 to turn the diagonals into positive numbers, which does not change the singular values of the matrix and the diagonal dominant parts. For computing eigenvalues of symmetric diagonally dominant matrices, however, this definition imposes the condition that A must be positive semi-definite.

The diagonal dominant matrices defined above are based on row dominance and we sometimes refer to them as being row diagonally dominant. If a matrix is such that $a_{jj} \geq \sum_{i \neq j} |a_{ij}|$, it is said to be column diagonally dominant.

Here is some notation that will be used throughout. Inequalities on matrices and vectors are entry-wise, sign(α) is the sign of α if $\alpha \neq 0$ and sign(0) = 1, $\mathbf{u}$ is the machine roundoff unit and $fl(z)$ is the computed result of the expression z in the floating point arithmetic.

3. Accurate SVD of diagonally dominant matrices

Our high relative accuracy algorithm for SVD of a diagonally dominant matrix is based on a general algorithm developed by Demmel et al. [8]. Two similar algorithms based on the Jacobi method [13, 22] can compute the singular values of A to high relative accuracy if a rank-revealing factorization $A = XDY$ can be computed accurately in the sense that

1. each entry of the diagonal matrix D has a relative accuracy in the order of machine precision;
2. X and Y are well-conditioned;
3. X and Y are norm-wise accurate, i.e. if $\widehat{X}$ is the computed X factor, then $\|\widehat{X} - X\|/\|X\|$ is in the order of machine precision.

Throughout this paper, we shall refer to $A = XDY$ as an accurate factorization if it satisfies the three conditions above.

For a diagonally dominant matrix given by $\mathcal{D}(A_D, v)$, we shall compute an accurate rank-revealing factorization by modifying the standard Gaussian elimination algorithm for the LDU factorization with diagonal pivoting such that each entry of D is computed to the order of machine precision and L and U have relative errors in norm bounded by the order of machine precision. Here, L is unit lower triangular, U is unit upper triangular, and D is diagonal. Since the diagonal dominance property is inherited by Schur complements in the Gaussian elimination, the diagonal pivoting, which selects the maximum entry on the diagonal for the pivot, will be equivalent to the complete pivoting. Therefore, U is (row) diagonally dominant and well-conditioned and L is usually well-conditioned as well. Thus the algorithm of Demmel et al. [8] can be used to compute SVD of A accurately.

The matrix L produced by the diagonal pivoting could potentially have a large condition number. To theoretically guarantee well-conditioning of L, we can also use a more expensive pivoting strategy[1] that produces a column diagonally dominant L. This pivoting strategy has essentially been proposed by J. Pena [25] for diagonally dominant M-matrices and Stiejies matrices, where it is called maximal absolute diagonal dominance pivoting. For a (row) diagonally dominant matrix $A = [a_{ij}]$, we have

$$\sum_{i=1}^{n} a_{ii} \geq \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Unless the matrix is entirely zero (in which case we stop the elimination and set $L = U = I$ and $D = 0$), there is at least one k such that $a_{kk} \neq 0$ and column k is column diagonally dominant, i.e.

$$a_{kk} \geq \sum_{i=1, i\neq k}^{n} |a_{ik}|.$$

Now our pivoting strategy is to permute row 1 with row k and column 1 with column k, after which the first column of the matrix is diagonally dominant, i.e. $|a_{11}| \geq$

[1] This pivoting strategy was suggested by an anonymous referee.

$\sum_{i=2}^{n} |a_{i1}|$. When there are many columns satisfying the above, we can simply pick the one with maximal a_{kk} or, in [25], the one that gives the most diagonal dominance. Applying the Gaussian elimination, the first column of $L = [l_{ij}]$ that is produced will be column diagonally dominant, i.e. $\sum_{i=2}^{n} |l_{i1}| = \sum_{i=2}^{n} |a_{i1}/a_{11}| \leq 1$. It is well known that the Schur complement after the elimination is still a (row) diagonally dominant matrix. We can apply the same pivoting strategy to the Schur complement. Repeating this strategy, at the end, we obtain a (row) diagonally dominant U as usual, but now L will be column diagonally dominant as each column of L is diagonally dominant by the pivoting. With L unit lower triangular and column diagonally dominant, its condition numbers can be bounded as

$$\kappa_\infty(L) := \|L\|_\infty \|L^{-1}\|_\infty \leq n^2 \ \text{ and } \ \kappa_1(L) := \|L\|_1 \|L^{-1}\|_1 \leq 2n;$$

see [25, Proposition 2.1]. Similarly, with U unit upper triangular and row diagonally dominant, we have

$$\kappa_\infty(U) \leq 2n \ \text{ and } \ \kappa_1(U) \leq n^2.$$

We call this pivoting strategy column diagonal dominance pivoting. It theoretically guarantees that both L and U are well-conditioned. It needs to compute the sums of off-diagonal entries for all columns at each step of the Gaussian elimination, which requires in total $\mathcal{O}(n^3)$ flops. This extra cost over the standard diagonal pivoting may be significant. However, in the context of our algorithms for computing accurate SVD (see Section 3.2), this is relatively insignificant, as the Jacobi algorithms used there will have the dominating cost overall. Nevertheless, if the extra cost is a concern, we can always use the standard diagonal pivoting as a first attempt, which does usually produce a well-conditioned L. We then check the condition number of L in $\mathcal{O}(n^2)$ flops and recompute, if necessary, the factorization with the column diagonal dominance pivoting strategy.

3.1. **Accurate LDU-factorization.** We now turn to the problem of computing L, D and U accurately. Similar to the accurate algorithm for computing LU factorization of a diagonally dominant M-matrix [2, 17], we shall proceed with the Gaussian elimination by updating the diagonals from the off-diagonals and the diagonally dominant parts without subtractions. Our key observation is that, even without the sign properties of an M-matrix, the diagonally dominant parts can still be computed without subtractions. Furthermore, when computing off-diagonals, possible subtractions and catastrophic cancellations do not affect the relative accuracy of the LDU factorization in the sense discussed at the beginning of this section.

Consider applying the Gaussian elimination to A with one of the pivoting strategies discussed earlier. Assuming that k steps of eliminations can be carried out, we let $A^{(k+1)} = [a_{ij}^{(k+1)}]$ denote the matrix obtained after the k-th Gaussian elimination, and we write $A^{(1)} = A$. For the ease of presentation, we assume that the rows and columns of A are already permuted so that no pivoting is carried out during the Gaussian elimination.

It is well known that $A^{(k)}$ is still diagonally dominant. We represent it as $A^{(k)} = \mathcal{D}(A_D^{(k)}, v^{(k)})$ where $v^{(k)} = [v_1^{(k)}, v_2^{(k)}, \cdots, v_n^{(k)}]^T$ with

$$v_i^{(k)} = a_{ii}^{(k)} - \sum_{j=k, j\neq i}^{n} |a_{ij}^{(k)}|$$

for $i \geq k$, and $v_i^{(k)} = v_i^{(k-1)}$ for $i < k$. We first establish a formula for updating $v_i^{(k)}$ for $i \geq k$.

Theorem 1. *Assume that k steps of the Gaussian eliminations can be carried out for A (i.e. $a_{\ell\ell}^{(\ell)} \neq 0$ for $1 \leq \ell \leq k$). For for $k+1 \leq i, j \leq n$, let*

$$s_{ij}^{(k)} = \text{sign}\left(a_{ij}^{(k+1)}\right) \text{sign}\left(a_{ij}^{(k)}\right)$$

and

$$t_{ij}^{(k)} = \begin{cases} -\text{sign}\left(a_{ij}^{(k+1)}\right) \text{sign}\left(a_{ik}^{(k)}\right) \text{sign}\left(a_{kj}^{(k)}\right), & \text{if } i \neq j \\ \text{sign}\left(a_{ik}^{(k)}\right) \text{sign}\left(a_{ki}^{(k)}\right), & \text{if } i = j. \end{cases}$$

We have for $k+1 \leq i \leq n$,

$$v_i^{(k+1)} = v_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1 - s_{ij}^{(k)})|a_{ij}^{(k)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(v_k^{(k)} + \sum_{j=k+1}^{n} (1 - t_{ij}^{(k)})|a_{kj}^{(k)}| \right). \tag{2}$$

Proof. First, for $k+1 \leq i, j \leq n$, we have from the Gaussian elimination that

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}}.$$

Let $p_{ij}^{(k)} = \text{sign}\left(a_{ij}^{(k)}\right)$. Then, for $i \geq k+1$,

$$\begin{aligned}
v_i^{(k+1)} &= a_{ii}^{(k+1)} - \sum_{j=k+1, j\neq i}^{n} p_{ij}^{(k+1)} a_{ij}^{(k+1)} \\
&= a_{ii}^{(k)} - \frac{a_{ik}^{(k)} a_{ki}^{(k)}}{a_{kk}^{(k)}} - \sum_{j=k+1, j\neq i}^{n} p_{ij}^{(k+1)} \left(a_{ij}^{(k)} - \frac{a_{ik}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}} \right) \\
&= a_{ii}^{(k)} - \sum_{j=k+1, j\neq i}^{n} p_{ij}^{(k+1)} a_{ij}^{(k)} + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(-p_{ik}^{(k)} a_{ki}^{(k)} + \sum_{j=k+1, j\neq i}^{n} p_{ik}^{(k)} p_{ij}^{(k+1)} a_{kj}^{(k)} \right) \\
&= a_{ii}^{(k)} - |a_{ik}^{(k)}| - \sum_{j=k+1, j\neq i}^{n} s_{ij}^{(k)} |a_{ij}^{(k)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(a_{kk}^{(k)} - t_{ii}^{(k)} |a_{ki}^{(k)}| - \sum_{j=k+1, j\neq i}^{n} t_{ij}^{(k)} |a_{kj}^{(k)}| \right) \\
&= v_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1 - s_{ij}^{(k)})|a_{ij}^{(k)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(v_k^{(k)} + \sum_{j=k+1}^{n} (1 - t_{ij}^{(k)})|a_{kj}^{(k)}| \right).
\end{aligned}$$

□

Recall that we use the definition sign(0) = 1 throughout. However, the above theorem is still valid if we adopt a more conventional definition sign(0) = 0; but it will complicate the error analysis in the next section.

We note that $1 - s_{ij}^{(k)}$ and $1 - t_{ij}^{(k)}$ are either 0 or 2. Then the above formula computes $v^{(k+1)}$ without subtractions. The diagonal entries, including the pivot, can be updated with addition operations through

$$a_{ii}^{(k+1)} = v_i^{(k+1)} + \sum_{j=k+1, j\neq i}^{n} |a_{ij}^{(k+1)}|. \tag{3}$$

It turns out that computing $v^{(k+1)}$ and the diagonals $a_{ii}^{(k+1)}$ as above is sufficient for computing an accurate LDU factorization. Of course, when computing an off-diagonal entry $a_{ij}^{(k+1)}$, subtractions and catastrophic cancellations may occur, which means a large relative error in the computed $a_{ij}^{(k+1)}$. But we will show that its absolute error is actually small relative to $a_{ii}^{(k+1)}$, which is sufficient to guarantee that $a_{ii}^{(k+1)}$ as computed from (3) has high relative accuracy. Similar results hold for the computation of $v^{(k+1)}$. See the analysis of Section 4 for details. We summarize this in the following implementation of the Gaussian elimination with either the diagonal pivoting (using lines 8 and 8a below) or the column diagonal dominance pivoting (using lines 8 and 8b below).

Algorithm 1. LDU Factorization of $\mathcal{D}(A_D, v)$

1 Input: $A_D = [a_{ij}]$ and $v = [v_i] \geq 0$;
2 Initialize: $P = I$, $L = I$, $D = 0$, $U = I$.
3 For $k = 1 : (n-1)$
4 For $i = k : n$
5 $a_{ii} = v_i + \sum_{j=k, j\neq i}^{n} |a_{ij}|$;
6 End For
7 If $\max_{i\geq k} a_{ii} = 0$, stop;
8 Choose an interchange permutation P_1 s.t. $A = P_1 A P_1$ satisfies:
8a a) for diagonal pivoting: $a_{kk} = \max_{i\geq k} a_{ii}$;
8b b) for column diagonal dominance pivoting:
$0 \neq a_{kk} \geq \sum_{i=k+1}^{n} |a_{ik}|$;
9 $P = P_1 P$; $L = P_1 L P_1$; $U = P_1 U P_1$; $d_k = a_{kk}$;
10 For $i = (k+1) : n$
11 $l_{ik} = a_{ik}/a_{kk}$; $u_{ki} = a_{ki}/a_{kk}$; $a_{ik} = 0$;
12 $v_i = v_i + |l_{ik}| v_k$;
13 For $j = (k+1) : n$
14 $p = \text{sign}(a_{ij} - l_{ik} a_{kj})$;
15 $s = \text{sign}(a_{ij}) p$;
16 $t = -\text{sign}(l_{ik})\text{sign}(a_{kj}) p$;
17 If $j = i$
18 $s = 1$; $t = \text{sign}(l_{ik})\text{sign}(a_{ki})$;
19 End if
20 $v_i = v_i + (1-s)|a_{ij}| + (1-t)|l_{ik} a_{kj}|$;
21 $a_{ij} = a_{ij} - l_{ik} a_{kj}$;
22 End for
23 End For
24 End for
25 $a_{nn} = v_n$; $d_n = a_{nn}$.

For the input matrix A_D, only its off-diagonals are used by the algorithm. In the algorithm, we have created L, D, and U for the purpose of precisely identifying them in the error analysis later. Note that on line 9, $L = P_1 L P_1$ and $U = P_1 U P_1$ effectively apply the permutation P_1 on the first $(k-1)$-st columns of L and the first $(k-1)$-st rows of U, respectively, and they can be replaced by $L_{:,1:(k-1)} = P_1 L_{:,1:(k-1)}$ and $U_{1:(k-1),:} = U_{1:(k-1),:} P_1$. We have also explicitly set $a_{ik} = 0$ on

line 11 after elimination. In this way, the matrix $A = [a_{ij}]$ at the k-th loop on line 6 is precisely $A^{(k)}$, the matrix obtained after the $(k-1)$-st Gaussian elimination. In practical implementations, L, D, and U need not be created and should be stored in A by overwriting the entries of A. A corresponding algorithm should replace line 11 by $a_{ik} = a_{ik}/a_{kk}; a_{ki} = a_{ki}/a_{kk}$ and replace all subsequent l_{ik} by a_{ik}. Lines 4-6 compute all the diagonal entries. With the exception of the pivot entry, they are not used in the subsequent steps for actual computations; they are only used to determine the pivot. Then, in practical implementations, we can first use the diagonal entries as computed by the standard Gaussian elimination at line 21 to determine the pivot and a permutation, and then compute the pivot a_{kk} by formula on line 5.

In output, the algorithm produces the factorization $PAP^T = LDU$. When $a_{kk} = 0$ occurs on line 7, the remaining matrix is entirely zero (i.e. $a_{ij} = 0$ for all $k \le i, j \le n$) and no further elimination is necessary. Note that, in this case, the factorization $PAP^T = LDU$ is still valid with $d_i = 0$ for $k \le i \le n$.

3.2. **Accurate SVD via LDU-factorization.** The LDU factorization as computed by Algorithm 1 can be fed into one of the algorithms of [8] to compute SVD to high relative accuracy. For completeness, we present one of them (i.e. Algorithm 3.2 of [8]) here. The algorithm is based on the one-sided Jacobi for SVD as presented as Algorithm 4.1 of [13].

Algorithm 2. ACCURATE SVD OF $\mathcal{D}(A_D, v)$

1 Input: $A_D = [a_{ij}]$ and $v = [v_i] \ge 0$
2 Compute P, L, D, U by Algorithm 1;
3 Compute SVD of $LD = \bar{Z}\bar{\Sigma}\bar{V}^T$ by the one-sided Jacobi algorithm
4 Multiply $W = \bar{\Sigma}(\bar{V}^T U)$, respecting parentheses;
5 Compute SVD of $W = \tilde{Z}\Sigma V^T$ by the one-sided Jacobi algorithm;
6 $Z = \bar{Z}\tilde{Z}$ and $A = (P^T Z)\Sigma(P^T V)^T$ is the SVD of A.

We can change the order of computing SVD at steps 3 and 5 to take advantage of the fact that DU is already given in the Gaussian elimination. Namely, we compute the SVD of $DU = \bar{Z}\bar{\Sigma}\bar{V}^T$ first and then the SVD of $W = (L\bar{Z})\bar{\Sigma}$. Then, in the factorization algorithm, we do not explicitly compute the LDU-factorization, but rather the LU-factorization directly $A = LU_1$ with $U_1 = DU$. This will eliminate a redundant step of dividing and later multiplying by D. However, we still present Algorithm 1 as an LDU-factorization algorithm for the purpose of error analysis.

Algorithm 2 uses the one-sided Jacobi twice, once at step 3 and again at step 5. One of them can be replaced by the QR factorization with column pivoting. Namely, instead of using SVD at step 3, one can compute the QR factorization with pivoting $LD = QRP$ where P is a permutation and then compute correspondingly $W = RPU$ at step 4 followed by the same steps 5 and 6. As R is usually well-conditioned, the resulting algorithm is then as accurate as the more expensive version above; see Algorithm 3.1 of [8] for more details.

Both Algorithm 1 and Algorithm 2 require $\mathcal{O}(n^3)$ floating point operations.

If A is symmetric, a symmetric version of Algorithm 1 can be easily worked out, which computes the LDL^T factorization (with $U = L^T$). Also, there is no need to use any of the pivoting strategies since $L = U^T$ will be automatically column diagonally dominant. We might still use a diagonal permutation to ensure nonzero pivots. Then Algorithm 2 can be simplified as follows. Writing $PAP^T =$

$(LD^{\frac{1}{2}})(LD^{\frac{1}{2}})^T$, where the permutation P is used to ensure nonzero pivots during elimination, we just need to compute the SVD of $LD^{\frac{1}{2}} = \bar{V}\bar{\Sigma}\bar{Z}^T$ by the one-sided Jacobi to obtain the eigenvalue decomposition $A = V\Lambda V^T$ where $\Lambda = \bar{\Sigma}^2$ and $V = P^T\bar{V}$. We state this as the following algorithm.

Algorithm 3. Accurate Eigenvalue Decomposition of Symmetric Positive Semi-definite $\mathcal{D}(A_D, v)$

1 Input: symmetric $A_D = [a_{ij}]$ and $v = [v_i] \geq 0$
2 Compute P, L, D by Algorithm 1 (symmetric version);
3 Compute SVD of $LD^{\frac{1}{2}} = \bar{V}\bar{\Sigma}\bar{Z}^T$ by the one-sided Jacobi (no need to keep $\bar{Z}$);
4 $\Lambda = \bar{\Sigma}^2$ and $V = P^T\bar{V}$.

When we compute SVD at step 3 in the algorithm, we do not need to keep $\bar{Z}$. Then, if we use the right-handed Jacobi method (i.e. the one by applying the Jacobi rotation from the right), we do not need to accumulate the Jacobi rotations in the iterations.

4. Error analysis of the LDU algorithm

We now present a forward error analysis of Algorithm 1 in a floating point arithmetic to show that L and U are norm-wise accurate while D is entrywise accurate. Note that it is this type of forward stability that is needed for Algorithm 2 and Algorithm 3 to compute an accurate SVD. We present the idea and the main results first and leave their detailed proofs to three subsections later.

Let P be the permutation matrix constructed in a floating point arithmetic from Algorithm 1 with either the diagonal pivoting or the column diagonal dominance pivoting. For ease of presentation, we assume that the matrix A has been permuted by this P in advance so that no pivoting is carried out in the process. We also assume that the off-diagonal entries and the diagonal dominant parts v_i are machine numbers. The results that take into consideration the initial errors are similar. We shall denote the computed quantities of Algorithm 1 by adding a hat to the corresponding notation. Thus, $\widehat{A}^{(k)} = (\widehat{a}_{ij}^{(k)})$ denotes the computed matrix after the $(k-1)$-st Gaussian elimination and $\widehat{v}^{(k)} = (\widehat{v}_i^{(k)})$ denotes the computed diagonally dominant part, i.e. they are A and v on line 6 at the k-th loop of Algorithm 1 in the floating point arithmetic.

Recall that $fl(z)$ denotes the computed result of the expression z. We assume the following standard model of floating point arithmetic:

$$fl(a \text{ op } b) = (a \text{ op } b)(1+\delta), \quad |\delta| \leq \mathbf{u},$$

where op $= +, -, *$ or $/$ and $\mathbf{u}$ is the machine roundoff unit. We assume that no underflow or overflow occurs in our algorithm. At a few places, our algorithm requires computing a sum of nonnegative numbers $\sum_{i=1}^n s_i$ with $s_i \geq 0$, which has a relative error bounded by $(n-1)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$. However, the error bound can be improved to $2\mathbf{u} + \mathcal{O}(n\mathbf{u}^2)$ by using the compensated summation algorithm [19]; see also [18, p. 93]. To simplify our presentation, we shall assume that all summations of nonnegative numbers are evaluated with the compensated summation algorithm and $n\mathbf{u} << 1$. Then we have

$$fl\left(\sum_{i=1}^n s_i\right) = \left(\sum_{i=1}^n s_i\right)(1+\delta) \tag{4}$$

with $\delta \leq 2\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$. Without using the compensated summation algorithm, several constants in error bounds (e.g. Lemma 5) will be replaced by $\mathcal{O}(n)$, and the final bounds will be increased by an $\mathcal{O}(n)$ factor.

In the rest of the paper, $\mathcal{O}(\mathbf{u}^2)$ denotes a quantity bounded by $C_n\mathbf{u}^2$ with C_n a constant dependent on n only. For ease of presentation, we shall use ϵ_ℓ as a generic notation to denote a quantity bounded in magnitude by $\ell\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$. Then ϵ_ℓ from one expression to another may represent different quantities. For example, we may write $\epsilon_3(-1+\epsilon_1) = \epsilon_3$, where ϵ_3 on the left and on the right represents two different terms that are both bounded by $3\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$.

Our first result shows that zero pivots, if encountered in the process, are computed exactly. Its proof will be given in Subsection 4.1.

Theorem 2. *For k with $1 \leq k \leq n$, $a_{\ell\ell}^{(\ell)} > 0$ for $1 \leq \ell \leq k$ if and only if $\widehat{a}_{\ell\ell}^{(\ell)} > 0$ for $1 \leq \ell \leq k$. Namely, a pivot $a_{kk}^{(k)}$ is 0 in the exact arithmetic if and only if the pivot $\widehat{a}_{kk}^{(k)}$ in the floating arithmetic is 0.*

Remark. The theorem implies that exact singularity and the rank of the matrix are detected by the algorithm exactly. This is basically because a zero pivot computed by Algorithm 1 can only come from addition or multiplication operations involving zeros, all of which are computed exactly. Indeed, if a pivot becomes 0, its diagonal dominant part must be 0. For a row to have a zero diagonal dominant part after k eliminations, its diagonal dominant part must be zero to start with and never increases during each of the elimination processes. For this to happen during an elimination, it turns out that the four entries involved in updating an off-diagonal entry must have at least one zero or have an M-matrix like sign pattern[2] (i.e. the elimination operation on off-diagonal entries is an addition of two numbers of the same sign). See Lemma 2 and the proof of Lemma 1 for details.

We shall present in Subsection 4.2 a related result (Theorem 4) that states that a singular A can be permuted into a block upper triangular with a singular M-matrix on its diagonal block after a sign scaling. Then applying the Gaussian elimination (Algorithm 1) to A effectively carries out the Gaussian elimination on the singular M-matrix, which involves no subtraction operations. So, even in a floating arithmetic, the zero pivots are computed exactly. That provides additional insights into why the singularity and rank can be computed exactly. However, we note that Theorem 2 additionally says that encountering a zero pivot in a floating arithmetic also implies a zero pivot in the exact arithmetic.

We now turn to the elimination steps before encountering zero pivots, if any. As mentioned before, we assume that A is already permuted by the permutation matrix determined under one of the pivoting schemes in the floating point arithmetic. Let N be the maximal integer such that

$$N \leq n-1 \quad \text{and} \quad \widehat{a}_{\ell\ell}^{(\ell)} > 0 \quad \text{for} \quad 1 \leq \ell \leq N.$$

Equivalently, $\widehat{A}^{(N)}$ is the last matrix that we can apply the elimination to, which results in $\widehat{A}^{(N+1)}$. By Theorem 2, $A^{(N)}$ is also the last matrix we can apply the elimination to in the exact arithmetic. If $N < n-1$, then $\widehat{a}_{N+1,N+1}^{(N+1)} = 0$, and Algorithm 1 terminates early at the $(N+1)$-st iteration with no need of further

[2]This sign property was observed by J. Demmel and communicated to the author.

elimination, as it follows from the pivoting that $a_{ij}^{(N+1)} = \widehat{a}_{ij}^{(N+1)} = 0$ for all $N+1 \leq i, j \leq n$.

Consider the elimination steps k with $1 \leq k \leq N$. For $k \leq i, j \leq n$, let

$$\delta_{ij}^{(k)} = \widehat{a}_{ij}^{(k)} - a_{ij}^{(k)} \quad \text{and} \quad \delta_i^{(k)} = \widehat{v}_i^{(k)} - v_i^{(k)}. \tag{5}$$

The first step in our forward error analysis is to bound the errors after the k-th elimination $\delta_{ij}^{(k+1)}$ and $\delta_i^{(k+1)}$ in terms of the errors in the previous steps $\delta_{ij}^{(k)}$ and $\delta_i^{(k)}$. This is fairly easy to do for $\delta_{ij}^{(k+1)}$, which corresponds to taking a differential on the corresponding formula, but it is extremely difficult for $\delta_i^{(k+1)}$, where we have to account for the possibility that some signs contained in $s_{ij}^{(k)}$ and $t_{ij}^{(k)}$ are computed wrong. When a computed sign differs from its corresponding sign in exact arithmetic, $\delta_i^{(k+1)}$ contains nondifferential terms that are not readily bounded by $\mathcal{O}(\mathbf{u})$ (assuming the errors in the previous steps are of $\mathcal{O}(\mathbf{u})$). The key idea to dealing with this situation is that, when a certain quantity is computed with a wrong sign, it must be small. By examining various cases and analyzing the terms involved carefully, it turns out that we can bound these nondifferential terms, sometimes in combination, by the quantity whose computed counterpart has the wrong sign (see Lemma 3 in Subsection 4.3).

The second step in our error analysis is to show that these errors are "relatively" small. Namely, we show that $|\delta_{ij}^{(k)}|$ for $j \neq i$ (i.e. the error for an off-diagonal entry) is of $\mathcal{O}(\mathbf{u})$ relative to $v_i^{(k)} + |a_{ij}^{(k)}|$, and the relative errors for $\hat{v}_i^{(k)}$ and $\hat{a}_{ii}^{(k)}$ are of $\mathcal{O}(\mathbf{u})$. This is established inductively using the bounds we have obtained for $\delta_{ij}^{(k+1)}$ and $\delta_i^{(k+1)}$. Assuming they are true for $\delta_{ij}^{(k)}$ and $\delta_i^{(k)}$, then our bounds for $\delta_{ij}^{(k+1)}$ and $\delta_i^{(k+1)}$ will be in terms of related quantities in $A^{(k)}$ (e.g. $a_{ij}^{(k)}$ and $v_i^{(k)}$). Then the key in establishing our relative bounds is the inequality

$$v_i^{(k)} + |a_{ij}^{(k)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(v_k^{(k)} + |a_{kj}^{(k)}| \right) \leq v_i^{(k+1)} + |a_{ij}^{(k+1)}|.$$

which bounds those quantities in $A^{(k)}$ in terms of those quantities in $A^{(k+1)}$, leading to relative bounds (see Lemma 4 in Subsection 4.3).

Applying the relative bounds to the LDU factors leads to our main result, a complete proof of which is presented in Subsection 4.3.

Theorem 3. *Let $\widehat{L} = [\widehat{l}_{ik}]$, $\widehat{D} = diag\{\widehat{d}_i\}$ and $\widehat{U} = [\widehat{u}_{ik}]$ be the computed factors of the LDU-factorization of $\mathcal{D}(A_D, v)$ by Algorithm 1 and let $L = [l_{ik}]$, $D = diag\{d_i\}$ and let $U = [u_{ik}]$ be the corresponding factors computed exactly. We have*

$$\begin{aligned} \|\widehat{L} - L\|_\infty &\leq \left(n\nu_{n-1}\mathbf{u} + \mathcal{O}(\mathbf{u}^2)\right) \|L\|_\infty, \\ |\widehat{d}_i - d_i| &\leq \left(\xi_{n-1}\mathbf{u} + \mathcal{O}(\mathbf{u}^2)\right) d_i, \quad \textit{for } 1 \leq i \leq n, \\ \|\widehat{U} - U\|_\infty &\leq \left(\nu_{n-1}\mathbf{u} + \mathcal{O}(\mathbf{u}^2)\right) \|U\|_\infty, \end{aligned}$$

where $\nu_{n-1} \leq 6 \cdot 8^{n-1} - 2$ and $\xi_{n-1} \leq 5 \cdot 8^{n-1} - \frac{5}{2}$.

The theorem shows that Algorithm 1 computes an accurate LDU factorization with the appropriate entrywise and norm-wise relative errors of order $\mathcal{O}(\mathbf{u})$. The main point of this analysis is to demonstrate that these relative errors are independent of the intermediate matrices in the Gaussian elimination process and the

condition of the matrix. In particular, possible large roundoff errors caused by cancellations in computing off-diagonal entries do not affect the final accuracy. However, our bounds are weak in that the constant coefficients ν_{n-1} and ξ_{n-1} in $\mathcal{O}(\mathbf{u})$ depend exponentially on n; a slightly better but still exponential bound can be derived (see the remarks after the proof of the theorem). This exponential bound is inherent in the forward error analysis where the worst cases of roundoff error accumulations have to be taken into account. In practice, the bounds on the constants are most likely pessimistic. Indeed, our numerical tests show that these constants appear more like $\mathcal{O}(n)$; see the examples in the next section. It will be interesting to study whether a better theoretical bound can be obtained. We note that a corresponding forward error analysis for diagonally dominant M-matrices have a bound with the constants of order $\mathcal{O}(n^3)$; see [2, 11, 24].

The forward error analysis can also provide a perturbation bound on the LDU factors. Specifically, let $A = \mathcal{D}(A_D, v)$ and $\widetilde{A} = \mathcal{D}(\widetilde{A}_D, \widetilde{v})$ with $A = [a_{ij}]$, $\widetilde{A} = [\widetilde{a}_{ij}]$, $v = [v_i]$ and $\widetilde{v} = [\widetilde{v}_i]$ be such that

$$|a_{ij} - \widetilde{a}_{ij}| \le \delta |a_{ij}|, \quad \text{and } |v_i - \widetilde{v}_i| \le \delta v_i, \tag{6}$$

for all $i \neq j$. If $\widetilde{A}^{(k)} = \mathcal{D}(\widetilde{A}_D^{(k)}, \widetilde{v}^{(k)})$ denotes the matrix obtained from $\widetilde{A}$ after the $(k-1)$-st Gaussian elimination in the exact arithmetic with $\widetilde{A}^{(k)} = (\widetilde{a}_{ij}^{(k)})$ and $\widetilde{v}^{(k)} = (\widetilde{v}_i^{(k)})$, then Lemma 3 holds with all $\epsilon_\ell = 0$ (exact arithmetic). Using that, a result corresponding to Lemma 5 with ϵ_1 replaced by δ can be worked out similarly and, with $\phi(1) = 1$ and $\psi(1) = 1$, we can obtain

$$\begin{aligned} \|\widetilde{L} - L\|_\infty &\le \left(\alpha_n \delta + \mathcal{O}(\delta^2)\right) \|L\|_\infty \\ |\widetilde{d}_i - d_i| &\le \left(\beta_n \delta + \mathcal{O}(\delta^2)\right) d_i \\ \|\widetilde{U} - U\|_\infty &\le \left(\gamma_n \delta + \mathcal{O}(\delta^2)\right) \|U\|_\infty \end{aligned}$$

where $\alpha_n, \beta_n, \gamma_n$ are some constants dependent on n exponentially. Applying the results of [8], we have corresponding relative perturbation bounds on singular values and singular vectors. Again, it will be interesting to see if a better bound can be obtained from a direct perturbation analysis.

Finally, we remark that in the situation that the diagonally dominant part v is not accurately known (or given), v has to be first computed from the entries of A in order to apply Algorithm 1. In that case, we can only guarantee that each computed v_i, denoted by $\widehat{v}_i$, is computed in a backward stable way while its relative error could be large. Yet, it could still be beneficial to use our forward stable algorithm, which computes the singular values of $\mathcal{D}(A_D, \widehat{v})$ accurately. Thus the algorithm has a mixed forward-backward stability in the sense that it accurately computes an (entrywise) nearby problem. To be more precise, we write

$$\widehat{v}_i = (a_{ii} - (1+\epsilon_2) \sum_{j \neq i} |a_{ij}|)(1+\epsilon_1)$$

where we assume that the off-diagonals are summed first in computing v_i. Let

$$\tilde{v}_i := \frac{\widehat{v}_i}{1+\epsilon_3} = \frac{a_{ii}}{1+\epsilon_2} - \sum_{j \neq i} |a_{ij}|.$$

Then, applying our algorithm to $\mathcal{D}(A_D, \widehat{v})$, the computed singular values are accurate approximations of those of $\mathcal{D}(A_D, \widehat{v})$ and hence of those of $\mathcal{D}(A_D, \tilde{v})$ (with

$\tilde{v} = [\tilde{v}_i]$). Letting $\widetilde{A} = [\widetilde{a}_{ij}] := \mathcal{D}(A_D, \tilde{v})$, we have $\widetilde{a}_{ii} = \frac{a_{ii}}{1+\epsilon_2}$ and $\widetilde{a}_{ij} = a_{ij}$ for $i \neq j$. Thus, the computed singular values are entrywise backward stable, which is the best one can hope for when only the entries of A are given.

We now present the detailed proofs of the results presented in this section.

4.1. **Proof of Theorem 2.** We first present a lemma. We note that the diagonal entries in both exact and finite precision arithmetic must be nonnegative.

Lemma 1. *Let $1 \le k \le n$ and assume that $a_{\ell\ell}^{(\ell)} > 0$ and $\widehat{a}_{\ell\ell}^{(\ell)} > 0$ for $1 \le \ell \le k-1$. If either $v_i^{(k)} = 0$ or $\widehat{v}_i^{(k)} = 0$ for some $i \ge k$, then $v_i^{(k)} = \widehat{v}_i^{(k)} = 0$, and for a fixed $j \ge k$ with $j \neq i$, we have that*

$$\widehat{a}_{ij}^{(k)} \text{ and } a_{ij}^{(k)} \text{ are either both zero or both nonzero with the same sign,}$$

i.e. $\widehat{a}_{ij}^{(k)} = c_{ij}^{(k)} a_{ij}^{(k)}$ for some $c_{ij}^{(k)} > 0$.

Proof. We prove by induction in k. We prove for the case $\widehat{v}_i^{(k)} = 0$ only; the proof for the case $v_i^{(k)} = 0$ is similar. When $k = 1$, $\widehat{a}_{ij}^{(k)} = a_{ij}^{(k)}$ $(j \neq i)$ and $\widehat{v}_i^{(k)} = v_i^{(k)}$. The lemma is obviously true. Assuming that it is true for some k, we prove it for $k+1$, namely, supposing $a_{\ell\ell}^{(\ell)} > 0$ and $\widehat{a}_{\ell\ell}^{(\ell)} > 0$ for $1 \le \ell \le k$ and $\widehat{v}_i^{(k+1)} = 0$ for some $i \ge k+1$, we prove $v_i^{(k+1)} = 0$ and that $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are either both zero or both nonzero with the same sign.

First, we have

(7)
$$fl\left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1 - \widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1 - \widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right)\right) = 0.$$

Since all terms in the summation are nonnegative, they must all be zero. In particular, $\widehat{v}_i^{(k)} = 0$. It follows from the induction assumption that for any $j \ge k$ with $j \neq i$,

$$\widehat{a}_{ij}^{(k)} = c_{ij}^{(k)} a_{ij}^{(k)} \quad \text{for some } c_{ij}^{(k)} > 0. \tag{8}$$

Furthermore $v_i^{(k)} = 0$. We now consider two cases.

Case 1. $\widehat{a}_{ik}^{(k)} = 0$. It follows from this and (8) that $a_{ik}^{(k)} = 0$. We therefore have for $j \neq i$,

$$\widehat{a}_{ij}^{(k+1)} = fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right) = \widehat{a}_{ij}^{(k)} \quad \text{and} \quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}a_{kj}^{(k)}}{a_{kk}^{(k)}} = a_{ij}^{(k)}.$$

Thus $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are either both zero or both nonzero with the same sign. Furthermore, we have that $\widehat{s}_{ij}^{(k)} = \text{sign}(\widehat{a}_{ij}^{(k+1)})\text{sign}(\widehat{a}_{ij}^{(k)}) = \text{sign}(a_{ij}^{(k+1)})\text{sign}(a_{ij}^{(k)}) = s_{ij}^{(k)}$. Now, it follows from (7) that $(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| = 0$ and hence $(1-s_{ij}^{(k)})|a_{ij}^{(k)}| = 0$, which together with $a_{ik}^{(k)} = 0$ and (2) lead to $v_i^{(k+1)} = 0$. The lemma is proved in this case.

Case 2. $\widehat{a}_{ik}^{(k)} \neq 0$. In this case, (7) implies that $\widehat{v}_k^{(k)} = 0$. By the induction assumption, we have $v_k^{(k)} = 0$ and

$$\widehat{a}_{kj}^{(k)} = c_{kj}^{(k)} a_{kj}^{(k)} \quad \text{for some } c_{kj}^{(k)} > 0. \tag{9}$$

We now prove that $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are either both zero or both nonzero with the same sign for $j \geq k+1$ and $j \neq i$ in three subcases.

a) $\widehat{a}_{ij}^{(k)} = 0$. By (8), we have $a_{ij}^{(k)} = 0$. Then,

$$\widehat{a}_{ij}^{(k+1)} = fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right) = -\frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_2) = -\frac{a_{ik}^{(k)}a_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}c_{ik}^{(k)}c_{kj}^{(k)}(1+\epsilon_2)$$

where we have used (8) and (9). Also, $a_{ij}^{(k+1)} = -\frac{a_{ik}^{(k)}a_{kj}^{(k)}}{a_{kk}^{(k)}}$. Therefore $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are either both zero or both nonzero with the same sign.

b) $\widehat{a}_{ij}^{(k)} \neq 0$ and $\widehat{a}_{kj}^{(k)} = 0$. By (9), we have $a_{kj}^{(k)} = 0$. Then $\widehat{a}_{ij}^{(k+1)} = fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right) = \widehat{a}_{ij}^{(k)}$ and $a_{ij}^{(k+1)} = a_{ij}^{(k)}$. It therefore follows from (8) that $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are either both zero or both nonzero with the same sign.

c) $\widehat{a}_{ij}^{(k)} \neq 0$ and $\widehat{a}_{kj}^{(k)} \neq 0$. From (7), we have $(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| = 0$ and $(1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| = 0$. Therefore, $\widehat{s}_{ij}^{(k)} = \widehat{t}_{ij}^{(k)} = 1$. Then $\mathrm{sign}(\widehat{a}_{ij}^{(k+1)}) = \mathrm{sign}(\widehat{a}_{ij}^{(k)})$ and $\mathrm{sign}(\widehat{a}_{ij}^{(k+1)}) = -\mathrm{sign}(\widehat{a}_{ik}^{(k)})\mathrm{sign}(\widehat{a}_{kj}^{(k)})$. Thus $\mathrm{sign}(\widehat{a}_{ij}^{(k)}) = -\mathrm{sign}(\widehat{a}_{ik}^{(k)})\mathrm{sign}(\widehat{a}_{kj}^{(k)})$. Therefore $\widehat{a}_{ij}^{(k+1)} = fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right)$ has the same sign as $\widehat{a}_{ij}^{(k)}$. On the other hand, $\widehat{a}_{ij}^{(k)}, \widehat{a}_{ik}^{(k)}, \widehat{a}_{kj}^{(k)}$ are all nonzero; they have the same signs as $a_{ij}^{(k)}, a_{ik}^{(k)}, a_{kj}^{(k)}$, respectively, by (8) and (9). Thus $\mathrm{sign}(a_{ij}^{(k)}) = -\mathrm{sign}(a_{ik}^{(k)})\mathrm{sign}(a_{kj}^{(k)})$ as well. Therefore $a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}a_{kj}^{(k)}}{a_{kk}^{(k)}}$ has the same sign as $a_{ij}^{(k)}$ as well, which implies that $\widehat{a}_{ij}^{(k+1)}$ and $a_{ij}^{(k+1)}$ are nonzero and have the same sign.

To finish the proof for Case 2, we need to show that $v_i^{(k+1)} = 0$. From what we have proved and (8), it follows that

$$\widehat{s}_{ij}^{(k)} = \mathrm{sign}(\widehat{a}_{ij}^{(k+1)})\mathrm{sign}(\widehat{a}_{ij}^{(k)}) = \mathrm{sign}(a_{ij}^{(k+1)})\mathrm{sign}(a_{ij}^{(k)}) = s_{ij}^{(k)}.$$

In addition, from (9), we have $\widehat{t}_{ij}^{(k)} = t_{ij}^{(k)}$. On the other hand, (7) leads to $(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| = 0$ and $(1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| = 0$, from which it follows that $(1-s_{ij}^{(k)})|a_{ij}^{(k)}| = 0$ and $(1-t_{ij}^{(k)})|a_{kj}^{(k)}| = 0$. Thus, we have $v_i^{(k+1)} = 0$ by (2). □

Proof of Theorem 2. We prove by induction on k. For $k = 1$,

$$\widehat{a}_{11}^{(1)} = fl\left(v_1^{(1)} + \sum_{j=2}^{n}|a_{1j}^{(1)}|\right)$$

which is 0 if and only if each term in the sum is 0, i.e. if and only if $a_{11}^{(1)}$ is 0. Now assuming that the theorem is true for $k-1$, we prove that the *if* part is true for k; the *only if* part is proved similarly. Let $\widehat{a}_{\ell\ell}^{(\ell)} > 0$ for $1 \leq \ell \leq k$ and we show $a_{\ell\ell}^{(\ell)} > 0$.

By the induction assumption, we have $a_{ll}^{(l)} > 0$ for $1 \le l \le k-1$. Furthermore,

$$\widehat{a}_{kk}^{(k)} = fl\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} |\widehat{a}_{kj}^{(k)}|\right) = 0$$

implies that $\widehat{v}_k^{(k)} = 0$ and $\widehat{a}_{kj}^{(k)} = 0$ for $k+1 \le j \le n$. By Lemma 1, we have $v_k^{(k)} = 0$ and $a_{kj}^{(k)} = 0$ for $k+1 \le j \le n$, which implies that $a_{kk}^{(k)} = 0$. Therefore $a_{\ell\ell}^{(\ell)} > 0$ for $1 \le \ell \le k$. □

4.2. **Characterization of singularity.** We present a result that further helps to understand why the algorithm computes singularity exactly. This result is not required for the proofs of other results.

Lemma 2. *For a fixed k, assume $v_i^{(k+1)} = 0$ for some i with $k+1 \le i \le n$. Then for $1 \le \ell \le k$, $v_i^{(\ell)} = 0$. Furthermore, if $a_{ij}^{(\ell)} \ne 0$ for some $j \ge \ell+1$ and $j \ne i$, then $a_{ij}^{(\ell)} a_{i\ell}^{(\ell)} a_{\ell j}^{(\ell)} \le 0$ and $a_{ij}^{(\ell+1)} \ne 0$ has the same sign as $a_{ij}^{(\ell)}$. In particular, if $a_{ij}^{(k+1)} = 0$ for some $j \ge k+2$, then $a_{ij}^{(\ell)} = 0$ for $1 \le \ell \le k$.*

Proof. The argument is essentially contained in the proof of Lemma 1, but we present one here for completeness. We just need to show it for the case $\ell = k$ only, i.e. $v_i^{(k)} = 0$, and if $a_{ij}^{(k)} \ne 0$, then $a_{ij}^{(k)} a_{ik}^{(k)} a_{kj}^{(k)} \le 0$ and $a_{ij}^{(k+1)} \ne 0$ has the same sign as $a_{ij}^{(k)}$.

By noting that each term in (2) is nonnegative, $v_i^{(k+1)} = 0$ immediately implies that $v_i^{(k)} = 0$ and

$$(1 - s_{ij}^{(k)}) a_{ij}^{(k)} = 0 \quad \text{and} \quad (1 - t_{ij}^{(k)}) \frac{a_{ik}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}} = 0,$$

where $j \ge k+1$ and $j \ne i$. Since $a_{ij}^{(k)} \ne 0$, then $s_{ij}^{(k)} = 1$, i.e.

$$\operatorname{sign}(a_{ij}^{(k)}) = \operatorname{sign}(a_{ij}^{(k+1)}). \tag{10}$$

Now, if $a_{ik}^{(k)} a_{kj}^{(k)} = 0$, then $a_{ij}^{(k+1)} = a_{ij}^{(k)}$, and they have the same sign. In this case $a_{ij}^{(k)} a_{ik}^{(k)} a_{kj}^{(k)} = 0$. If $a_{ik}^{(k)} a_{kj}^{(k)} \ne 0$, then $1 - t_{ij}^{(k)} = 0$, which together with (10) implies

$$\operatorname{sign}(a_{ik}^{(k)}) \operatorname{sign}(a_{kj}^{(k)}) = -\operatorname{sign}(a_{ij}^{(k)}).$$

Therefore, $a_{ij}^{(k)} a_{ik}^{(k)} a_{kj}^{(k)} \le 0$ and $a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k)} / a_{kk}^{(k)}$ must have the same sign as $a_{ij}^{(k)}$. The proof is complete. □

Theorem 4. *A diagonally dominant matrix A is singular if and only if there is a permutation matrix P such that*

$$P^T A P = \left(\begin{array}{c|ccc} A_{00} & A_{01} & \cdots & A_{0s} \\ \hline & A_{11} & & \\ & & \ddots & \\ & & & A_{ss} \end{array}\right), \tag{11}$$

where the diagonal blocks are square with the A_{00} block possibly empty and, for $1 \le i \le s$, $A_{ii} = D_i M_{ii} D_i$ with $D_i = \operatorname{diag}(\pm 1)$ and M_{ii} an irreducible diagonally

dominant M-matrix (i.e. with nonnegative diagonals and nonpositive off-diagonals) with zero diagonally dominant parts. Also, A_{00}, if not empty, is nonsingular.

Proof. The *if* part is trivial, as M_{ii} must be singular. We prove the *only if* part. First, for a diagonally dominant matrix A, there is a permutation P_0 such that

$$P_0^T A P_0 = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1q} \\ & X_{22} & \cdots & X_{2q} \\ & & \ddots & \vdots \\ & & & X_{qq} \end{pmatrix} \tag{12}$$

where X_{ii} is square and irreducible. Since A is singular, there is at least one singular diagonal block. If a diagonal block X_{ii} (for $1 \le i \le q-1$) is singular, then $X_{ij} = 0$ for $i < j \le q$, as otherwise X_{ii} will have at least one row that is strictly diagonally dominant and will then be nonsingular (noting that an irreducible diagonally dominant matrix with at least one row that is strictly diagonally dominant is nonsingular; see [26, p. 23]). Therefore there is a permutation P such that $P^T AP$ has all singular blocks in (12) moved to the lower-right corner on the diagonal. Namely, rewriting the singular blocks in (12) as $A_{11}, \cdots, A_{ss}$, and putting together all remaining nonsingular blocks on the upper-left corner, if any, and writing them as a single block A_{00}, $P^T AP$ has the form of (11). Here, A_{00} is possibly empty but, if not, it is nonsingular. A_{ii} is singular and irreducible for $1 \le i \le s$.

It remains to show that each A_{ii} has the desired form. For this, we need to show that if an $n \times n$ diagonally dominant matrix A is singular and irreducible, then there exists $D = \mathrm{diag}(\pm 1)$ such that DAD is an M-matrix. We next prove this by induction in n.

The case $n = 1$ is trivial. Assume it is true for $(n-1)\times(n-1)$ matrices. Consider an $n \times n$ singular and irreducible matrix A. Since A is irreducible and $n \ge 2$, there must be some $a_{1j} \ne 0$ and hence $a_{11} \ne 0$. Also, all diagonally dominant parts v_i must be zero, as otherwise A will be nonsingular (again using [26, p. 23]). Apply one step of the Gaussian elimination to A and denote as before the resulting matrix as

$$A^{(2)} = \begin{matrix} \\ 1 \\ n-1 \end{matrix} \begin{matrix} \begin{matrix} 1 & n-1 \end{matrix} \\ \begin{pmatrix} a_{11} & \mathtt{x} \\ 0 & B \end{pmatrix} \end{matrix}, \tag{13}$$

where $\mathtt{x}$ denotes a vector or matrix of appropriate dimension. B is then singular and still diagonally dominant. We show that B must be irreducible.

Suppose B is reducible. Using the proof at the beginning, we have a permutation matrix P_1 such that $P_1^T B P_1$ takes the form of (11). In particular, we have

$$P_1^T B P_1 = \begin{matrix} \\ n-1-p \\ p \end{matrix} \begin{matrix} \begin{matrix} n-1-p & p \end{matrix} \\ \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \end{matrix} \tag{14}$$

with B_{22} irreducible and singular. Without loss of generality, assume $P_1 = I$, and we write

$$A^{(2)} = \begin{matrix} \\ 1 \\ n-1-p \\ p \end{matrix} \begin{matrix} \begin{matrix} 1 & n-1-p & p \end{matrix} \\ \begin{pmatrix} a_{11} & \mathtt{x} & \mathtt{x} \\ 0 & B_{11} & B_{12} \\ 0 & 0 & B_{22} \end{pmatrix} \end{matrix}. \tag{15}$$

B_{22} being singular implies that its diagonally dominant parts are 0, i.e. $v_i^{(2)} = 0$ for $n-p+1 \le i \le n$. Since the $(3,2)$ block of $A^{(2)}$ is zero, it follows from Lemma 2 that the $(3,2)$ block of A must be zero. If $a_{i1} = 0$ for all $n-p+1 \le i \le n$, we have

$$A = \begin{array}{c} \\ 1 \\ n-1-p \\ p \end{array} \begin{array}{c} \begin{array}{ccc} 1 & n-1-p & p \end{array} \\ \begin{pmatrix} a_{11} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ 0 & 0 & \mathtt{x} \end{pmatrix} \end{array}.$$

So, A is reducible, which is a contradiction. If $a_{i1} \neq 0$ for some $n-p+1 \le i \le n$, we must have $a_{1j} = 0$ for $2 \le j \le n-p$, as otherwise the $(3,2)$ block of $A^{(2)}$, which is obtained from a zero block after the Gaussian elimination, cannot be zero. Therefore we have

$$A = \begin{array}{c} \\ 1 \\ n-1-p \\ p \end{array} \begin{array}{c} \begin{array}{ccc} 1 & n-1-p & p \end{array} \\ \begin{pmatrix} a_{11} & 0 & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & 0 & \mathtt{x} \end{pmatrix} \end{array}.$$

Now, by permuting the first two blocks in both rows and columns, A becomes

$$\begin{array}{c} \\ n-1-p \\ 1 \\ p \end{array} \begin{array}{c} \begin{array}{ccc} n-1-p & 1 & p \end{array} \\ \begin{pmatrix} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ 0 & a_{11} & \mathtt{x} \\ 0 & \mathtt{x} & \mathtt{x} \end{pmatrix} \end{array}.$$

This implies that A is reducible, which is a contradiction. Thus, we have shown that B must be irreducible.

Applying the induction assumption to B, there exists $D_1 = \mathrm{diag}(\pm 1)$, such that $D_1 B D_1$ is an M-matrix with the diagonally dominant part $v_i^{(2)} = 0$ for $2 \le i \le n$. For ease of notation, we can assume without loss of generality that $D_1 = I$, i.e. B is an M-matrix. For $2 \le i \le n$, it follows from $v_i^{(2)} = 0$ and Lemma 2 that $a_{ij} \le 0$ (for $2 \le j \ne i$) since, if a_{ij} is nonzero, it must have the same sign as the corresponding off-diagonal entry of B which is nonpositive. Lemma 2 implies in addition that $a_{i1}a_{1j} \ge 0$. On the other hand, $v_i^{(2)} = 0$ also leads to $(1 - t_{ii}^{(1)})|a_{1i}| = 0$. Then $a_{i1}a_{1i} \ge 0$. Therefore, $a_{i1}a_{1j} \ge 0$ for all $j \ge 2$. Noting that A is irreducible, there is one index $i_0 \ge 2$ such that $a_{i_0 1} \neq 0$. Let $d_1 = \pm 1$ be such that $d_1 a_{i_0 1} < 0$. Then $d_1 a_{1j} = \frac{d_1}{a_{i_0 1}} a_{i_0 1} a_{1j} \le 0$ for all $j \ge 2$. But there is at least one j_0 with $a_{1j_0} \neq 0$. Therefore, $d_1 a_{i1} = \frac{d_1}{a_{1j_0}} a_{i1} a_{1j_0} \le 0$ for all i. Now, with $D = \mathrm{diag}\{d_1, I\}$, DAD is an M-matrix with zero diagonally dominant parts. The induction is completed. □

When we apply the Gaussian elimination to a singular A, it effectively applies the elimination to each of the diagonal blocks A_{ii}. Examining the elimination process on A_{ii}, it is obvious that it is the same as the one applied to the M-matrix M_{ii}. For the singular M-matrix M_{ii}, our algorithm (or the standard GTH algorithm [17]) does not involve any subtraction operation throughout, and any result that is exactly zero will be computed exactly.

4.3. **Proof of Theorem 3.** We first prove some lemmas.

Lemma 3. *For $1 \le k \le N$ and $k \le i, j \le n$, let $\delta_{ij}^{(k)} = \widehat{a}_{ij}^{(k)} - a_{ij}^{(k)}$ and $\delta_i^{(k)} = \widehat{v}_i^{(k)} - v_i^{(k)}$. Then, for $k+1 \le i, j \le n$, we have*

$$\begin{aligned} \delta_{ij}^{(k+1)} &= a_{ij}^{(k+1)}\epsilon_1 - \frac{a_{ik}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}}\epsilon_2 + \delta_{ij}^{(k)}(1+\epsilon_1) - \frac{\delta_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}} a_{kj}^{(k)}(1+\epsilon_3) \\ &\quad - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\delta_{kj}^{(k)}(1+\epsilon_3) + \frac{a_{ik}^{(k)}\delta_{kk}^{(k)}}{a_{kk}^{(k)}\widehat{a}_{kk}^{(k)}} a_{kj}^{(k)}(1+\epsilon_3), \end{aligned} \tag{16}$$

if $j \ne i$, and

$$|\delta_{ii}^{(k)}| \le a_{ii}^{(k)}|\epsilon_2| + \left(|\delta_i^{(k)}| + \sum_{j=k, j\ne i}^{n} |\delta_{ij}^{(k)}|\right)(1+\epsilon_2), \tag{17}$$

$$\begin{aligned} |\delta_i^{(k+1)}| &\le |\delta_i^{(k)}| + \sum_{j=k+1, j\ne i}^{n} (1 - s_{ij}^{(k)})|\delta_{ij}^{(k)}| \\ &\quad + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(|\delta_k^{(k)}| + \sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|\delta_{kj}^{(k)}|\right) \\ &\quad + \frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(v_k^{(k)} + \sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right) \\ &\quad + \frac{|a_{ik}^{(k)}\delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)}a_{kk}^{(k)}}\left(v_k^{(k)} + \sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right) \\ &\quad + 2\sum_{j\in K_1^{(i)}} |\delta_{ij}^{(k)}| + \Theta_i + 2\sum_{j\in K_3}\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| \\ &\quad + 2\sum_{j\in K_4^{(i)}} |\delta_{ij}^{(k+1)}|(1+\epsilon_3) + \epsilon_7\widehat{v}_i^{(k+1)} \end{aligned} \tag{18}$$

where

$$\begin{aligned} K_1^{(i)} &= \{j \ne i :\ k+1 \le j \le n \text{ and } |a_{ij}^{(k)}| \le |\delta_{ij}^{(k)}|\}, \\ K_2 &= \{i :\ k+1 \le i \le n \text{ and } |a_{ik}^{(k)}| \le |\delta_{ik}^{(k)}|\}, \\ K_3 &= \{j :\ k+1 \le j \le n \text{ and } |a_{kj}^{(k)}| \le |\delta_{kj}^{(k)}|\}, \\ K_4^{(i)} &= \{j \ne i :\ k+1 \le j \le n \text{ and } |a_{ij}^{(k+1)}| \le |\delta_{ij}^{(k+1)}|\} \end{aligned}$$

and

$$\Theta_i = \begin{cases} 2\sum_{j=k+1}^{n} \frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|, & \text{if } i \in K_2, \\ 0, & \text{otherwise.} \end{cases}$$

Also recall that ϵ_ℓ is a generic notation for a quantity bounded by $\ell\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$.

Proof. For $i, j \geq k+1$ and $i \neq j$, we have

$$\begin{aligned}
\widehat{a}_{ij}^{(k+1)} &= fl\left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right) = \left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_2)\right)(1+\epsilon_1) \\
&= \left(a_{ij}^{(k)} + \delta_{ij}^{(k)} - \frac{(a_{ik}^{(k)} + \delta_{ik}^{(k)})(a_{kj}^{(k)} + \delta_{kj}^{(k)})}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_2)\right)(1+\epsilon_1) \\
&= \left(a_{ij}^{(k)} - \frac{a_{ik}^{(k)}a_{kj}^{(k)}}{a_{kk}^{(k)}}(1+\epsilon_2)\right)(1+\epsilon_1) + \delta_{ij}^{(k)}(1+\epsilon_1) - \frac{a_{kj}^{(k)}\delta_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_3) \\
&\quad - \frac{(a_{ik}^{(k)} + \delta_{ik}^{(k)})\delta_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_3) + \frac{a_{ik}^{(k)}a_{kj}^{(k)}\delta_{kk}^{(k)}}{a_{kk}^{(k)}\widehat{a}_{kk}^{(k)}}(1+\epsilon_3) \\
&= a_{ij}^{(k+1)} + a_{ij}^{(k+1)}\epsilon_1 - \frac{a_{ik}^{(k)}a_{kj}^{(k)}}{a_{kk}^{(k)}}\epsilon_2 + \delta_{ij}^{(k)}(1+\epsilon_1) - \frac{a_{kj}^{(k)}\delta_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}(1+\epsilon_3) \\
&\quad - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\delta_{kj}^{(k)}(1+\epsilon_3) + \frac{a_{ik}^{(k)}a_{kj}^{(k)}\delta_{kk}^{(k)}}{a_{kk}^{(k)}\widehat{a}_{kk}^{(k)}}(1+\epsilon_3).
\end{aligned}$$

This proves (16). Similarly, for $i \geq k+1$, using (4), we have

$$\widehat{a}_{ii}^{(k)} = fl\left(\widehat{v}_i^{(k)} + \sum_{j=k, j\neq i}^{n} |\widehat{a}_{ij}^{(k)}|\right) = \left(\widehat{v}_i^{(k)} + \sum_{j=k, j\neq i}^{n} |\widehat{a}_{ij}^{(k)}|\right)(1+\epsilon_2).$$

Then

$$|\delta_{ii}^{(k)}| \leq a_{ii}^{(k)}|\epsilon_2| + (|\delta_i^{(k)}| + \sum_{j=k, j\neq i}^{n} |\delta_{ij}^{(k)}|)(1+\epsilon_2).$$

On computing $v_i^{(k+1)}$, we assume that the two in the formula summations are computed first by the compensated summation algorithm, and we have by (4)

$$fl\left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|\right) = \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|\right)(1+\epsilon_2)$$

and

$$fl\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right) = \left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right)(1+\epsilon_2),$$

where we assume that multiplications by $1-\widehat{s}_{ij}^{(k)}$ or $1-\widehat{t}_{ij}^{(k)}$, which are either 0 or 2, encounter no roundoff errors. Then

$$\begin{aligned}\widehat{v}_i^{(k+1)} &= fl\left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n}(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| \right.\\ &\quad \left. + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n}(1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right)\right)\\ &= \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n}(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|\right)(1+\epsilon_3)\\ &\quad + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n}(1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right)(1+\epsilon_5).\end{aligned}$$

Furthermore, replacing ϵ_3 and ϵ_5 by $\min\{\epsilon_3,\ \epsilon_5\}$ and noting that $|\min\{\epsilon_3,\ \epsilon_5\}| \leq 5\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, we have

$$\begin{aligned}\widehat{v}_i^{(k+1)} \geq \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n}(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| \right.\\ \left. + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n}(1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}|\right)\right)\times(1+\epsilon_5)\end{aligned} \tag{19}$$

Let i be fixed ($k+1 \leq i \leq n$). In considering the difference $\widehat{v}_i^{(k+1)} - v_i^{(k+1)}$, we examine the difference in each term of the summation and for $k+1 \leq j \leq n$ let

$$\begin{aligned}\Omega_j &= \begin{cases}(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| - (1-s_{ij}^{(k)})|a_{ij}^{(k)}|, & \text{if } j\neq i,\\ 0, & \text{if } j = i,\end{cases}\\ \Gamma_j &= (1-\widehat{t}_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| - (1-t_{ij}^{(k)})\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}|.\end{aligned}$$

Then, we can write

$$
\begin{aligned}
\delta_i^{(k+1)} &= \widehat{v}_i^{(k+1)} - v_i^{(k+1)} \\
&= \widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| \right) \\
&\quad + \epsilon_3 \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| \right) \\
&\quad + \epsilon_5 \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| \right) \\
&\quad - v_i^{(k)} - \sum_{j=k+1, j\neq i}^{n} (1-s_{ij}^{(k)})|a_{ij}^{(k)}| - \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(v_k^{(k)} + \sum_{j=k+1}^{n} (1-t_{ij}^{(k)})|a_{kj}^{(k)}| \right) \\
&= \delta_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} \Omega_j + \sum_{j=k+1}^{n} \Gamma_j + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \widehat{v}_k^{(k)} - \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} v_k^{(k)} \\
&\quad + \epsilon_3 \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| \right) \\
&\quad + \epsilon_5 \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| \right) \\
&= \delta_i^{(k)} + \sum_{j=k+1}^{n} (\Omega_j + \Gamma_j) + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \delta_k^{(k)} + \left(\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} - \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \right) v_k^{(k)} \\
&\quad + \epsilon_3 \left(\widehat{v}_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| \right) \\
&\quad + \epsilon_5 \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \left(\widehat{v}_k^{(k)} + \sum_{j=k+1}^{n} (1-\widehat{t}_{ij}^{(k)})|\widehat{a}_{kj}^{(k)}| \right).
\end{aligned}
$$

Hence

$$
\begin{aligned}
|\delta_i^{(k+1)}| \leq |\delta_i^{(k)}| + \sum_{j=k+1}^{n} |\Omega_j + \Gamma_j| + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} |\delta_k^{(k)}| \\
+ \left(\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} + \frac{|a_{ik}^{(k)} \delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)} a_{kk}^{(k)}} \right) v_k^{(k)} + \epsilon_5 \widehat{v}_i^{(k+1)}
\end{aligned} \tag{20}
$$

where we have used (19). To bound $\Omega_j + \Gamma_j$, we need to consider the situations where the sign of a certain term is computed wrong such that $\widehat{s}_{ij}^{(k)} \neq s_{ij}^{(k)}$ or $\widehat{t}_{ij}^{(k)} \neq t_{ij}^{(k)}$. In that case, $\Omega_j + \Gamma_j$ is not a differential and may not appear to be small. However, we shall bound them using the fact that a sign can only be computed wrong when the corresponding quantity is small. We first separate out the part that can be

written as a differential and let

$$\Delta_j := (1-t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| - (1-t_{ij}^{(k)})\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}|.$$

Then

$$|\Delta_j| \le (1-t_{ij}^{(k)})\left(\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| + \frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|a_{kj}^{(k)}| + \frac{|a_{ik}^{(k)}\delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)}a_{kk}^{(k)}}|a_{kj}^{(k)}|\right). \tag{21}$$

We now prove that

$$\begin{aligned}|\Omega_j+\Gamma_j| \quad \le \quad & (1-s_{ij}^{(k)})|\delta_{ij}^{(k)}| + |\Delta_j| + 2|\delta_{ij}^{(k)}| + 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| + 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| \\ & +2|\delta_{ij}^{(k+1)}|(1+\epsilon_3) + \epsilon_2\left((1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| + (1-\widehat{t}_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|\right), \end{aligned} \tag{22}$$

where the first term only appears if $j \neq i$, the third term only appears if $j \in K_1^{(i)}$, the fourth term only appears if $i \in K_2$, the fifth term only appears if $j \in K_3$, and the last two terms only appear if $j \in K_4^{(i)}$. The proof is divided into two cases and several subcases that address all possible situations in the computed signs $\widehat{s}_{ij}^{(k)} = \text{sign}(\widehat{a}_{ij}^{(k+1)})\text{sign}(\widehat{a}_{ij}^{(k)})$, $\widehat{t}_{ij}^{(k)} = -\text{sign}(\widehat{a}_{ij}^{(k+1)})\text{sign}(\widehat{a}_{ik}^{(k)})\text{sign}(\widehat{a}_{kj}^{(k)})$ (for $i \neq j$) and $\widehat{t}_{ii}^{(k)} = \text{sign}(\widehat{a}_{ik}^{(k)})\text{sign}(\widehat{a}_{ki}^{(k)})$. Recall the convention $\text{sign}(0) = 1$.

Case 1. Either $j = i$ or $\text{sign}(\widehat{a}_{ij}^{(k+1)}) = \text{sign}(a_{ij}^{(k+1)})$. We bound Ω_j and Γ_j separately. For Ω_j, if $j = i$, we have $\Omega_j = 0$. If $j \neq i$, we write

$$\Omega_j = (1-s_{ij}^{(k)})(|\widehat{a}_{ij}^{(k)}| - |a_{ij}^{(k)}|) - (\widehat{s}_{ij}^{(k)} - s_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|.$$

We then consider two subcases with $j \neq i$:

a) If $\widehat{s}_{ij}^{(k)} = s_{ij}^{(k)}$, we have $|\Omega_j| \le (1-s_{ij}^{(k)})|\delta_{ij}^{(k)}|$.
b) If $\widehat{s}_{ij}^{(k)} \neq s_{ij}^{(k)}$, we have $\text{sign}(\widehat{a}_{ij}^{(k)}) \neq \text{sign}(a_{ij}^{(k)})$ and hence $|\widehat{a}_{ij}^{(k)}| + |a_{ij}^{(k)}| = |\delta_{ij}^{(k)}|$. Therefore, we have $|\widehat{a}_{ij}^{(k)}| \le |\delta_{ij}^{(k)}|$. Hence

$$j \in K_1^{(i)} \quad \text{and} \quad |\Omega_j| \le (1-s_{ij}^{(k)})|\delta_{ij}^{(k)}| + 2|\delta_{ij}^{(k)}|.$$

For Γ_j, we write for all j

$$\Gamma_j = \Delta_j - (\widehat{t}_{ij}^{(k)} - t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|.$$

We similarly consider two subcases:

a) If $\widehat{t}_{ij}^{(k)} = t_{ij}^{(k)}$, we have $|\Gamma_j| = |\Delta_j|$.
b) If $\widehat{t}_{ij}^{(k)} \neq t_{ij}^{(k)}$, we have either $\text{sign}(\widehat{a}_{ik}^{(k)}) \neq \text{sign}(a_{ik}^{(k)})$ or $\text{sign}(\widehat{a}_{kj}^{(k)}) \neq \text{sign}(a_{kj}^{(k)})$. Then either $|\widehat{a}_{ik}^{(k)}| + |a_{ik}^{(k)}| = |\delta_{ik}^{(k)}|$ or $|\widehat{a}_{kj}^{(k)}| + |a_{kj}^{(k)}| = |\delta_{kj}^{(k)}|$. In the former case, we have $|\widehat{a}_{ik}^{(k)}| \le |\delta_{ik}^{(k)}|$ and hence

$$i \in K_2 \quad \text{and} \quad |\Gamma_j| \le |\Delta_j| + 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|.$$

In the latter case, we have $|\widehat{a}_{kj}^{(k)}| \le |\delta_{kj}^{(k)}|$ and hence

$$j \in K_3 \quad \text{and} \quad |\Gamma_j| \le |\Delta_j| + 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}|.$$

Combining the bounds for Ω_j and Γ_j together, we have proved (22) for Case 1.

Case 2. $j \neq i$ and $\operatorname{sign}(\widehat{a}_{ij}^{(k+1)}) \neq \operatorname{sign}(a_{ij}^{(k+1)})$. In this case, $|\widehat{a}_{ij}^{(k+1)}| + |a_{ij}^{(k+1)}| = |\delta_{ij}^{(k+1)}|$ and therefore $j \in K_4^{(i)}$. We write

$$\begin{aligned}
\Omega_j &= (1 - s_{ij}^{(k)})(|\widehat{a}_{ij}^{(k)}| - |a_{ij}^{(k)}|) + (\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| - 2\widehat{s}_{ij}^{(k)}|\widehat{a}_{ij}^{(k)}|, \\
\Gamma_j &= \Delta_j + (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| - 2\widehat{t}_{ij}^{(k)}\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|.
\end{aligned}$$

We bound the second term in Ω_j above as follows. If $\widehat{s}_{ij}^{(k)} \neq s_{ij}^{(k)}$, then $(\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| = 0$. If $\widehat{s}_{ij}^{(k)} = s_{ij}^{(k)}$, we have $\operatorname{sign}(\widehat{a}_{ij}^{(k)}) \neq \operatorname{sign}(a_{ij}^{(k)})$ and $|\widehat{a}_{ij}^{(k)}| + |a_{ij}^{(k)}| = |\delta_{ij}^{(k)}|$. Therefore

$$j \in K_1^{(i)} \quad \text{and} \quad |(\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|| \le 2|\delta_{ij}^{(k)}|.$$

Similarly, we bound the second term in Γ_j above. If $\widehat{t}_{ij}^{(k)} \neq t_{ij}^{(k)}$, then $(\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| = 0$. But if $\widehat{t}_{ij}^{(k)} = t_{ij}^{(k)}$, we have either $\operatorname{sign}(\widehat{a}_{ik}^{(k)}) \neq \operatorname{sign}(a_{ik}^{(k)})$ or $\operatorname{sign}(\widehat{a}_{kj}^{(k)}) \neq \operatorname{sign}(a_{kj}^{(k)})$. Then either

$$i \in K_2 \quad \text{and} \quad (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| \le 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|$$

or

$$j \in K_3 \quad \text{and} \quad (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| \le 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}|.$$

Thus, we have shown

$$(\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}| + (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| \le 2|\delta_{ij}^{(k)}| + 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| + 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}|, \tag{23}$$

which leads to

$$|\Omega_j + \Gamma_j| \le (1 - s_{ij}^{(k)})|\delta_{ij}^{(k)}| + |\Delta_j| + 2|\delta_{ij}^{(k)}| + 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| + 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| + 2\Lambda, \tag{24}$$

where the third, fourth and fifth terms only appear if $j \in K_1^{(i)}$, $i \in K_2$, or $j \in K_3$, respectively, and

$$\begin{aligned}
\Lambda &:= \left| \widehat{s}_{ij}^{(k)}|\widehat{a}_{ij}^{(k)}| + \widehat{t}_{ij}^{(k)}\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| \right| \\
&= \left| \operatorname{sign}(\widehat{a}_{ij}^{(k+1)})\widehat{a}_{ij}^{(k)} - \operatorname{sign}(\widehat{a}_{ij}^{(k+1)})\frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)} \right| \\
&= \left| \widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)} \right|.
\end{aligned}$$

It follows from

$$\widehat{a}_{ij}^{(k+1)} = \left(\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)}(1+\epsilon_2)\right)(1+\epsilon_1) = \left(\frac{\widehat{a}_{ij}^{(k)}}{1+\epsilon_2} - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)}\right)(1+\epsilon_3)$$

that

$$\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)} = \widehat{a}_{ij}^{(k+1)}(1+\epsilon_1) - \epsilon_2\frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)}$$

and

$$\widehat{a}_{ij}^{(k)} - \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\widehat{a}_{kj}^{(k)} = \widehat{a}_{ij}^{(k+1)}(1+\epsilon_3) - \epsilon_2\widehat{a}_{ij}^{(k)}.$$

We now consider three subcases. In the first two subcases, we bound $\Omega_j + \Gamma_j$ through bounding Λ, but in the third, we bound $\Omega_j + \Gamma_j$ directly.

a) If $\widehat{s}_{ij}^{(k)} = -1$, we have

$$\Lambda \le |\widehat{a}_{ij}^{(k+1)}|(1+\epsilon_3) + \epsilon_2|\widehat{a}_{ij}^{(k)}| \le |\delta_{ij}^{(k+1)}|(1+\epsilon_3) + \epsilon_1(1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|.$$

Substituting this into (24) leads to (22).

b) If $\widehat{t}_{ij}^{(k)} = -1$, we have

$$\Lambda \le |\widehat{a}_{ij}^{(k+1)}|(1+\epsilon_1) + \epsilon_2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| \le |\delta_{ij}^{(k+1)}|(1+\epsilon_1) + \epsilon_1(1-\widehat{t}_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|.$$

Again, substituting this into (24) leads to (22).

c) If $\widehat{s}_{ij}^{(k)} = \widehat{t}_{ij}^{(k)} = 1$, then we go back to the definitions of Ω_j and Γ_j and have

$$\begin{aligned}
\Omega_j + \Gamma_j &= -(1-s_{ij}^{(k)})|a_{ij}^{(k)}| - (1-t_{ij}^{(k)})\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}| \\
&= -(\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|a_{ij}^{(k)}| - (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}| + 2s_{ij}^{(k)}|a_{ij}^{(k)}| \\
&\quad + 2t_{ij}^{(k)}\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}| \\
&= -(\widehat{s}_{ij}^{(k)} + s_{ij}^{(k)})|a_{ij}^{(k)}| - (\widehat{t}_{ij}^{(k)} + t_{ij}^{(k)})\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}|a_{kj}^{(k)}| + 2|a_{ij}^{(k+1)}|.
\end{aligned}$$

Bounding the first two terms as in (23), we obtain

$$|\Omega_j + \Gamma_j| \le 2|\delta_{ij}^{(k)}| + 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}| + 2\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| + 2|\delta_{ij}^{(k+1)}|$$

where the first, the second and the third terms appear only if $j \in K_1^{(i)}$, $i \in K_2$, or $j \in K_3$, respectively. So, (22) is true in this subcase, too.

Thus, we have proved (22) for Case 2.

Finally, noting $\sum_{j\in K_4^{(i)}}\left((1-\widehat{s}_{ij}^{(k)})|\widehat{a}_{ij}^{(k)}|+(1-\widehat{t}_{ij}^{(k)})\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\widehat{a}_{kj}^{(k)}|\right)\le \widehat{v}_i^{(k+1)}/(1+\epsilon_5)$ by (19), we have

$$\begin{aligned}\sum_{j=k+1}^{n}|\Omega_j+\Gamma_j| \quad\le\quad & \sum_{j=k+1,j\neq i}^{n}(1-s_{ij}^{(k)})|\delta_{ij}^{(k)}|+\sum_{j=k+1}^{n}|\Delta_j|+2\sum_{j\in K_1^{(i)}}|\delta_{ij}^{(k)}|\\ & +\Theta_i+2\sum_{j\in K_3}\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}|+2\sum_{j\in K_4^{(i)}}|\delta_{ij}^{(k+1)}|(1+\epsilon_3)+\epsilon_2\widehat{v}_i^{(k+1)}.\end{aligned}$$

Substitute this and (21) into (20), we have (18). □

We now present an inequality that allows bounding (16), (17) and (18) in terms of the quantities in $A^{(k+1)}$.

Lemma 4. *Let $1\le k\le N$. For any $k+1\le i\le n$ and any $J\subset\{k+1,k+2,\cdots,n\}\setminus\{i\}$, we have*

$$v_i^{(k+1)}+\sum_{j\in J}|a_{ij}^{(k+1)}|\ge v_i^{(k)}+\sum_{j\in J}|a_{ij}^{(k)}|+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right). \tag{25}$$

Proof. Let $K=\{k+1,k+2,\cdots,n\}\setminus(\{i\}\cup J)$. Using the same equation in the proof of Theorem 1, we have

$$\begin{aligned}v_i^{(k+1)} \quad+\quad & \sum_{j\in J}|a_{ij}^{(k+1)}|=a_{ii}^{(k+1)}-\sum_{j\in K}|a_{ij}^{(k+1)}|\\ =\quad & a_{ii}^{(k)}-|a_{ik}^{(k)}|-\sum_{j\in K}s_{ij}^{(k)}|a_{ij}^{(k)}|+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(a_{kk}^{(k)}-t_{ii}^{(k)}|a_{ki}^{(k)}|-\sum_{j\in K}t_{ij}^{(k)}|a_{kj}^{(k)}|\right)\\ \ge\quad & v_i^{(k)}+\sum_{j\in J}|a_{ij}^{(k)}|+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right).\end{aligned}$$

□

Next lemma gives relative error bounds.

Lemma 5. *For some fixed k ($1\le k\le N$), assume that for any $i\ge k$ and for any $J\subset\{k,k+1,\cdots,n\}\setminus\{i\}$ that*

$$\sum_{j\in J}|\delta_{ij}^{(k)}|\le\phi(k)\epsilon_1\left(v_i^{(k)}+\sum_{j\in J}|a_{ij}^{(k)}|\right) \tag{26}$$

and

$$|\delta_i^{(k)}|\le\psi(k)\epsilon_1 v_i^{(k)}, \tag{27}$$

where $\phi(k)\ge 0$ and $\psi(k)\ge 0$ are some functions of k. Then,

$$|\delta_{ii}^{(k)}|\le\xi(k)\epsilon_1 a_{ii}^{(k)} \tag{28}$$

with $\xi(k)=\phi(k)+\psi(k)+2$. Furthermore, for any $i\ge k+1$ and for any $J\subset\{k+1,k+2,\cdots,n\}\setminus\{i\}$, we have

$$\sum_{j\in J}|\delta_{ij}^{(k+1)}|\le(3\phi(k)+\psi(k)+5)\epsilon_1\left(v_i^{(k+1)}+\sum_{j\in J}|a_{ij}^{(k+1)}|\right), \tag{29}$$

and

$$|\delta_i^{(k+1)}| \le (14\phi(k) + 4\psi(k) + 19)\epsilon_1 v_i^{(k+1)}, \tag{30}$$

where we assume $\xi(k)\epsilon_1 < 1$. *Recall that* $\epsilon_1 \le \mathbf{u} + \mathcal{O}(\mathbf{u}^2)$.

Proof. From (17), we obtain

$$\begin{aligned} |\delta_{ii}^{(k)}| &\le a_{ii}^{(k)}|\epsilon_2| + \psi(k)\epsilon_1 v_i^{(k)}(1+\epsilon_2) + \phi(k)\epsilon_1 \left(v_i^{(k)} + \sum_{j=k, j\neq i}^{n} |a_{ij}^{(k)}| \right)(1+\epsilon_2) \\ &\le (\phi(k) + \psi(k) + 2)\epsilon_1 a_{ii}^{(k)}. \end{aligned}$$

Now, for $i \ge k+1$ and for any $J \subset \{k+1, k+2, \cdots, n\}/\{i\}$, we use (16) to prove (29). We note first that the terms ϵ_ℓ in (16) are dependent on j (for a fixed i). Therefore, we use a superscript j on all ϵ_ℓ appearing in the equation for $\delta_{ij}^{(k+1)}$ to distinguish them in the inequalities below. Summing (16) over $j \in J$, we have

$$\begin{aligned} \sum_{j\in J} |\delta_{ij}^{(k+1)}| \le & \sum_{j\in J} |a_{ij}^{(k+1)}\epsilon_1^{(j)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \sum_{j\in J} |a_{kj}^{(k)}\epsilon_2^{(j)}| + \sum_{j\in J} |\delta_{ij}^{(k)}|(1+\epsilon_1^{(j)}) \\ & + \frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \sum_{j\in J} |a_{kj}^{(k)}(1+\epsilon_3^{(j)})| + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \sum_{j\in J} |\delta_{kj}^{(k)}|(1+\epsilon_3^{(j)}) \\ & + \frac{|a_{ik}^{(k)}\delta_{kk}^{(k)}|}{a_{kk}^{(k)}\widehat{a}_{kk}^{(k)}} \sum_{j\in J} |a_{kj}^{(k)}(1+\epsilon_3^{(j)})|. \end{aligned}$$

It is easy to show that $\max_{j\in J} |\epsilon_\ell^{(j)}| = \epsilon_\ell$. Then using

$$\begin{aligned} |\widehat{a}_{ij}^{(k)}| &\le |a_{ij}^{(k)}| + |\delta_{ij}^{(k)}| \le |a_{ij}^{(k)}| + \phi(k)\epsilon_1(v_i^{(k)} + |a_{ij}^{(k)}|) \\ &\le |a_{ij}^{(k)}|(1+\phi(k)\epsilon_1) + \phi(k)\epsilon_1 v_i^{(k)} \end{aligned} \tag{31}$$

and

$$\widehat{a}_{ii}^{(k)} \ge a_{ii}^{(k)} - |\delta_{ii}^{(k)}| \ge a_{ii}^{(k)}(1 - \xi(k)\epsilon_1), \tag{32}$$

we bound each of the terms above as follows:

$$\sum_{j\in J}|a_{ij}^{(k+1)}\epsilon_1^{(j)}| \le \sum_{j\in J}|a_{ij}^{(k+1)}|\max_{j\in J}|\epsilon_1^{(j)}| = \epsilon_1\sum_{j\in J}|a_{ij}^{(k+1)}|$$

$$\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}\epsilon_2^{(j)}| \le \max_{j\in J}|\epsilon_2^{(j)}|\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}| = \epsilon_2\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}|$$

$$\sum_{j\in J}|\delta_{ij}^{(k)}|(1+\epsilon_1^{(j)}) \le (1+\max_{j\in J}|\epsilon_1^{(j)}|)\sum_{j\in J}|\delta_{ij}^{(k)}|$$

$$\le (1+\epsilon_1)\phi(k)\epsilon_1\left(v_i^{(k)}+\sum_{j\in J}|a_{ij}^{(k)}|\right)$$

$$= \phi(k)\epsilon_1\left(v_i^{(k)}+\sum_{j\in J}|a_{ij}^{(k)}|\right)$$

$$\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}(1+\epsilon_3^{(j)})| \le \frac{\phi(k)\epsilon_1(v_i^{(k)}+|a_{ik}^{(k)}|)}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)}\sum_{j\in J}|a_{kj}^{(k)}|(1+\epsilon_3)$$

$$\le \phi(k)\frac{\epsilon_1(1+\epsilon_3)}{1-\xi(k)\epsilon_1}\left(\frac{v_i^{(k)}}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}|+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}|\right)$$

$$\le \phi(k)\epsilon_1\left(v_i^{(k)}+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}|\right)$$

$$\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\sum_{j\in J}|\delta_{kj}^{(k)}|(1+\epsilon_3^{(j)}) \le \frac{|a_{ik}^{(k)}|(1+\phi(k)\epsilon_1)+\phi(k)\epsilon_1 v_i^{(k)}}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)}$$

$$\times\,\phi(k)\epsilon_1\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right)(1+\epsilon_3)$$

$$= \frac{|a_{ik}^{(k)}|\phi(k)\epsilon_1+\phi(k)^2\epsilon_1^2 v_i^{(k)}}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right)$$

$$\le \phi(k)\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right)$$

$$+\,\phi(k)^2\epsilon_1^2 v_i^{(k)}\frac{v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|}{a_{kk}^{(k)}}$$

$$\le \phi(k)\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j\in J}|a_{kj}^{(k)}|\right)+\phi(k)^2\epsilon_1^2 v_i^{(k)}$$

$$\frac{|a_{ik}^{(k)}\delta_{kk}^{(k)}|}{a_{kk}^{(k)}\widehat{a}_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}(1+\epsilon_3^{(j)})| \le \frac{|a_{ik}^{(k)}|\xi(k)\epsilon_1 a_{kk}^{(k)}}{a_{kk}^{(k)}a_{kk}^{(k)}(1-\xi(k)\epsilon_1)}\sum_{j\in J}|a_{kj}^{(k)}|(1+\epsilon_3)$$

$$\le (\phi(k)+\psi(k)+2)\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\sum_{j\in J}|a_{kj}^{(k)}|.$$

Adding them together and noting that $\phi(k)^2\epsilon_1^2 v_i^{(k)}$ in the fifth term can be combined into $\phi(k)\epsilon_1 v_i^{(k)}$ in the third term, i.e. $\phi(k)\epsilon_1 v_i^{(k)} + \phi(k)^2\epsilon_1^2 v_i^{(k)} = \phi(k)\epsilon_1 v_i^{(k)}$, we have

$$\begin{aligned}
\sum_{j\in J}|\delta_{ij}^{(k+1)}| &\le \epsilon_1 \sum_{j\in J}|a_{ij}^{(k+1)}| + \epsilon_2 \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \sum_{j\in J}|a_{kj}^{(k)}| \\
&\quad + \phi(k)\epsilon_1 \left(2v_i^{(k)} + \sum_{j\in J}|a_{ij}^{(k)}| + \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(v_k^{(k)} + 3\sum_{j\in J}|a_{kj}^{(k)}| \right) \right) \\
&\quad + (\psi(k)+2)\epsilon_1 \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \sum_{j\in J}|a_{kj}^{(k)}| \\
&\le (3\phi(k)+\psi(k)+5)\epsilon_1 \left(v_i^{(k+1)} + \sum_{j\in J}|a_{ij}^{(k+1)}| \right),
\end{aligned}$$

where we have used Lemma 4.

We now prove (30). Using (31) and (32) again, we bound the terms of (18) as follows.

$$\begin{aligned}
\sum_{j=k+1, j\neq i}^{n} (1-s_{ij}^{(k)})|\delta_{ij}^{(k)}| &= \sum_{j=k+1, j\neq i, s_{ij}^{(k)}=-1}^{n} (1-s_{ij}^{(k)})|\delta_{ij}^{(k)}| \\
&= 2 \sum_{j=k+1, j\neq i, s_{ij}^{(k)}=-1}^{n} |\delta_{ij}^{(k)}| \\
&\le 2\phi(k)\epsilon_1 \left(v_i^{(k)} + \sum_{j=k+1, j\neq i, s_{ij}^{(k)}=-1}^{n} |a_{ij}^{(k)}| \right) \\
&= \phi(k)\epsilon_1 \left(2v_i^{(k)} + \sum_{j=k+1, j\neq i}^{n} (1-s_{ij}^{(k)})|a_{ij}^{(k)}| \right)
\end{aligned}$$

$$\begin{aligned}
\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_k^{(k)}| &\le \frac{|a_{ik}^{(k)}|(1+\phi(k)\epsilon_1) + \phi(k)\epsilon_1 v_i^{(k)}}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)} \psi(k)\epsilon_1 v_k^{(k)} \\
&\le \psi(k)\epsilon_1 \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} v_k^{(k)} + \psi(k)\phi(k)\epsilon_1^2 v_i^{(k)}
\end{aligned}$$

$$\begin{aligned}
\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}} \sum_{j=k+1}^{n} (1-t_{ij}^{(k)})|\delta_{kj}^{(k)}| &\le \frac{|a_{ik}^{(k)}|(1+\phi(k)\epsilon_1) + \phi(k)\epsilon_1 v_i^{(k)}}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)} \\
&\quad \times \sum_{j=k+1}^{n} (1-t_{ij}^{(k)})|\delta_{kj}^{(k)}| \\
&\le \phi(k)\epsilon_1 \frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}} \left(2v_k^{(k)} + \sum_{j=k+1}^{n} (1-t_{ij}^{(k)})|a_{kj}^{(k)}| \right) + 2\phi(k)^2\epsilon_1^2 v_i^{(k)}
\end{aligned}$$

$$\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|)\le\frac{\phi(k)\epsilon_1(v_i^{(k)}+|a_{ik}^{(k)}|)}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)}(v_k^{(k)}$$
$$+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|)$$
$$\le\phi(k)\epsilon_1\left(2v_i^{(k)}+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right)\right)$$
$$\frac{|a_{ik}^{(k)}\delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)}a_{kk}^{(k)}}(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|)\le\frac{|a_{ik}^{(k)}|\xi(k)\epsilon_1 a_{kk}^{(k)}}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)a_{kk}^{(k)}}$$
$$\times(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|)$$
$$\le(\phi(k)+\psi(k)+2)\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|).$$

Using the definition of $K_1^{(i)}$, we have

$$\sum_{j\in K_1^{(i)}}|\delta_{ij}^{(k)}|\le\phi(k)\epsilon_1\left(v_i^{(k)}+\sum_{j\in K_1^{(i)}}|a_{ij}^{(k)}|\right)\le\phi(k)\epsilon_1\left(v_i^{(k)}+\sum_{j\in K_1^{(i)}}|\delta_{ij}^{(k)}|\right).$$

Then

$$\sum_{j\in K_1^{(i)}}|\delta_{ij}^{(k)}|\le\frac{\phi(k)\epsilon_1 v_i^{(k)}}{1-\phi(k)\epsilon_1}=\phi(k)\epsilon_1 v_i^{(k)}.$$

Similarly, if $i\in K_2$, then $|\delta_{ik}^{(k)}|\le\phi(k)\epsilon_1(v_i^{(k)}+|a_{ik}^{(k)}|)\le\phi(k)\epsilon_1(v_i^{(k)}+|\delta_{ik}^{(k)}|)$ which implies $|\delta_{ik}^{(k)}|\le\phi(k)\epsilon_1 v_i^{(k)}$. Then using

$$\widehat{a}_{kk}^{(k)}=\left(\widehat{v}_k^{(k)}+\sum_{j=k+1}^{n}|\widehat{a}_{kj}^{(k)}|\right)(1+\epsilon_2)\ge\sum_{j=k+1}^{n}|\widehat{a}_{kj}^{(k)}|(1+\epsilon_2)\tag{33}$$

we have

$$\begin{aligned}\Theta_i &\le 2\frac{|\delta_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\sum_{j=k+1}^{n}|\widehat{a}_{kj}^{(k)}|\le 2\phi(k)\epsilon_1 v_i^{(k)}\frac{\sum_{j=k+1}^{n}|\widehat{a}_{kj}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\\ &\le 2\phi(k)\epsilon_1 v_i^{(k)}/(1+\epsilon_2)=2\phi(k)\epsilon_1 v_i^{(k)}.\end{aligned}$$

Similarly, using the definition of K_3, (31) and (32), we have

$$\begin{aligned}\sum_{j\in K_3}\frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}|\delta_{kj}^{(k)}| &\le \frac{|a_{ik}^{(k)}|(1+\phi(k)\epsilon_1)+\phi(k)\epsilon_1 v_i^{(k)}}{a_{kk}^{(k)}(1-\xi(k)\epsilon_1)}\phi(k)\epsilon_1 v_k^{(k)}\\ &\le \phi(k)\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}v_k^{(k)}+\phi(k)^2\epsilon_1^2 v_i^{(k)}.\end{aligned}$$

Using (29) that is already proved, we have

$$\begin{aligned}\sum_{j\in K_4^{(i)}}|\delta_{ij}^{(k+1)}| &\leq (3\phi(k)+\psi(k)+5)\epsilon_1\left(v_i^{(k+1)}+\sum_{j\in K_4^{(i)}}|a_{ij}^{(k+1)}|\right)\\ &\leq (3\phi(k)+\psi(k)+5)\epsilon_1\left(v_i^{(k+1)}+\sum_{j\in K_4^{(i)}}|\delta_{ij}^{(k+1)}|\right)\end{aligned}$$

which implies

$$\sum_{j\in K_4^{(i)}}|\delta_{ij}^{(k+1)}|\leq(3\phi(k)+\psi(k)+5)\epsilon_1 v_i^{(k+1)}.$$

Summing all the terms in (18) as we have bounded above and noting that higher order terms like $\epsilon_1^2 v_i^{(k)}$ can be combined into $\phi(k)\epsilon_1 v_i^{(k)}$ in the second term, we have

$$\begin{aligned}|\delta_i^{(k+1)}|\leq{}&\psi(k)\epsilon_1\left(v_i^{(k)}+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(2v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right)\right)\\ &+\phi(k)\epsilon_1\left(8v_i^{(k)}+\sum_{j=k+1,j\neq i}^{n}(1-s_{ij}^{(k)})|a_{ij}^{(k)}|\right.\\ &\qquad\left.+\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(6v_k^{(k)}+3\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right)\right)\\ &+2\epsilon_1\frac{|a_{ik}^{(k)}|}{a_{kk}^{(k)}}\left(v_k^{(k)}+\sum_{j=k+1}^{n}(1-t_{ij}^{(k)})|a_{kj}^{(k)}|\right)\\ &+2(3\phi(k)+\psi(k)+5)\epsilon_1 v_i^{(k+1)}+\epsilon_7\widehat{v}_i^{(k+1)}\\ \leq{}&(14\phi(k)+4\psi(k)+19)\epsilon_1 v_i^{(k+1)}+\epsilon_7|\delta_i^{(k+1)}|.\end{aligned}$$

This leads to

$$|\delta_i^{(k+1)}|\leq\frac{(14\phi(k)+4\psi(k)+19)\epsilon_1 v_i^{(k+1)}}{1-\epsilon_7}=(14\phi(k)+4\psi(k)+19)\epsilon_1 v_i^{(k+1)}.$$

The proof is complete. □

Applying the above lemma inductively, we see that (26), (27) and (28) hold for $1\leq k\leq N+1$ with $\phi(k)$ and $\psi(k)$ defined by

$$\phi(k+1) = 3\phi(k)+\psi(k)+5,\quad \phi(1)=0; \tag{34}$$

$$\psi(k+1) = 14\phi(k)+4\psi(k)+19,\quad \psi(1)=0. \tag{35}$$

Clearly, $\phi(k)$ and $\psi(k)$ are increasing sequences.

Proof of Theorem 3. For $1\leq k\leq N$, the L and U factors are obtained from the Gaussian elimination matrices through

$$\widehat{l}_{ik}=fl\left(\frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right)\quad\text{and}\quad\widehat{u}_{kj}=fl\left(\frac{\widehat{a}_{kj}^{(k)}}{\widehat{a}_{kk}^{(k)}}\right)$$

where $i, j = k+1, \cdots, n$. For $i \geq k+1$, we have $|\widehat{a}_{ik}^{(k)}| \leq \widehat{a}_{ii}^{(k)} \leq \widehat{a}_{kk}^{(k)}$ if using the diagonal pivoting and $|\widehat{a}_{ik}^{(k)}| \leq \widehat{a}_{kk}^{(k)}$ if using the column diagonal dominance pivoting. Therefore, $|\widehat{a}_{ik}^{(k)}|/\widehat{a}_{kk}^{(k)} \leq 1$, and we have

$$\begin{aligned} |\widehat{l}_{ik} - l_{ik}| &\leq \left| \frac{\widehat{a}_{ik}^{(k)}}{\widehat{a}_{kk}^{(k)}} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \right| + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\mathbf{u} \\ &\leq \frac{|\delta_{ik}^{(k)}|}{a_{kk}^{(k)}} + \frac{|\widehat{a}_{ik}^{(k)}\delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)}a_{kk}^{(k)}} + \frac{|\widehat{a}_{ik}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\mathbf{u} \\ &\leq \phi(k)\epsilon_1 + \xi(k)\epsilon_1 + \epsilon_1 \\ &\leq \nu_{n-1}\epsilon_1, \end{aligned}$$

where $\nu_{n-1} = 2\phi(n-1) + \psi(n-1) + 3$. Similarly, for $j > k$, we have

$$\begin{aligned} \sum_{j=k+1}^{n} |\widehat{u}_{kj} - u_{kj}| &\leq \frac{\sum_{j=k+1}^{n} |\delta_{kj}^{(k)}|}{a_{kk}^{(k)}} + \sum_{j=k+1}^{n} \frac{|\widehat{a}_{kj}^{(k)}\delta_{kk}^{(k)}|}{\widehat{a}_{kk}^{(k)}a_{kk}^{(k)}} + \frac{\sum_{j=k+1}^{n} |\widehat{a}_{kj}^{(k)}|}{\widehat{a}_{kk}^{(k)}}\mathbf{u} \\ &\leq (\phi(k) + \xi(k) + 1)\epsilon_1 \\ &\leq \nu_{n-1}\epsilon_1 \end{aligned}$$

where we have used (33). Now, if $N+1 < n$, the elimination is terminated at step $N+1$, and we have $a_{ij}^{(N+1)} = \widehat{a}_{ij}^{(N+1)} = 0$ for all $N+1 \leq i, j \leq n$. Thus for $N+1 \leq i, j \leq n$, $\widehat{l}_{ik} = l_{ik} = 0$ and $\widehat{u}_{kj} = u_{kj} = 0$. Putting all of these together, we have shown that

$$\|\widehat{L} - L\|_\infty \leq (n-1)\nu_{n-1}\epsilon_1 = (n-1)\nu_{n-1}\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$$

and

$$\|\widehat{U} - U\|_\infty \leq \nu_{n-1}\epsilon_1 = \nu_{n-1}\mathbf{u} + \mathcal{O}(\mathbf{u}^2).$$

On the other hand, $\widehat{d}_i = \widehat{a}_{ii}^{(i)}$ for $1 \leq i \leq N+1$. If $N+1 < n$, we also have that $a_{ii}^{(N+1)} = \widehat{a}_{ii}^{(N+1)} = 0$ and hence $\widehat{d}_i = d_i = 0$ for $N+1 \leq i \leq n$. So, letting $\xi_{n-1} = \xi(n-1) = \phi(n-1) + \psi(n-1) + 2$, the bound on $\widehat{d}_i - d_i$ follows from (28).

Finally, using (34) and (35), it can be proved by induction that

$$\phi(k) \leq 8^{k-1} - \frac{1}{2} \quad \text{and} \quad \psi(k) \leq 4 \cdot 8^{k-1} - 4.$$

Thus,

$$\nu_{n-1} \leq 6 \cdot 8^{n-1} - 2 \quad \text{and} \quad \xi_{n-1} \leq 5 \cdot 8^{n-1} - \frac{5}{2}.$$

Now, the theorem is proved by noting that $\|L\|_\infty \geq 1$ and $\|U\|_\infty \geq 1$. □

We note that slightly better bounds for ν_{n-1}, ξ_{n-1} can be obtained by solving

$$\begin{pmatrix} \Phi(k+1) \\ \Psi(k+1) \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 14 & 4 \end{pmatrix} \begin{pmatrix} \Phi(k) \\ \Psi(k) \end{pmatrix} + \begin{pmatrix} 5 \\ 19 \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} \Phi(1) \\ \Psi(1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{36}$$

Its solution is

$$\begin{pmatrix} \Phi(k) \\ \Psi(k) \end{pmatrix} = c_1\lambda_1^{k-1}v_1 + c_2\lambda_2^{k-1}v_2 + \begin{pmatrix} -1/2 \\ -4 \end{pmatrix},$$

where (λ_1, v_1) and (λ_2, v_2) are two eigenpairs of the coefficient matrix of (36) and c_1, c_2 are chosen to satisfy the initial conditions. Since $\lambda_1 \approx -0.275$ and $\lambda_2 \approx$

7.275, this will lead to slightly smaller bounds for ν_{n-1}, ξ_{n-1}, but they are more complicated.

5. Numerical examples

In this section, we present three numerical examples of diagonally dominant matrices (that are not M-matrices) to demonstrate the high relative accuracy achieved by the new algorithm. We compare it with the Jacobi algorithms [13] as well as the `svd` (or `eig` if the matrix is symmetric) of MATLAB which is based on the QR algorithm. We also present one example of singular matrix that is essentially an M-matrix to demonstrate the ability of the algorithm to compute the rank exactly. All tests were performed in MATLAB on an Intel Pentium PC.

We have implemented Algorithm 1, Algorithm 2 and Algorithm 3 as presented, except that A is overwritten by L, D and U in our implementation of Algorithm 1. Our implementations of the Jacobi algorithms are also straight out of Algorithm 3.1 of [13] (two-sided Jacobi for symmetric positive definite eigenvalue problems) and Algorithm 4.1 of [13] (one-sided right-handed Jacobi for singular value problems) except that, in the computation of $t = \text{sign}(\zeta)/(|\zeta| + \sqrt{1+\zeta^2})$ there, we scale the numerator and denominator by the exponential part of ζ to prevent overflow in ζ^2. The termination threshold for the Jacobi method is set as $n * \texttt{eps}$.

Example 1. We consider a 100×100 symmetric matrix $\mathcal{D}(A_D, v)$ whose off-diagonals are -1 except at the anti-diagonals ($i + j = n + 1$), where they are $\eta = 10^{-16}$. We let the diagonals of A be such that the row sums of A are all $\lambda = 10^{-15}$. Then λ is an eigenvalue of A with $e = [1, \cdots, 1]^T$ as an eigenvector.

The diagonally dominant parts are then $v_i = \lambda - 2\eta$, which are computed with high relative accuracy for the given λ and η. We list and compare the smallest eigenvalue $\widehat{\lambda}$ as computed by Algorithm 3, by the Jacobi algorithm (one-sided for singular values) and by MATLAB's `eig` in Table 1 below. Since the matrix is symmetric, no pivoting scheme is used in our LDL^T factorization algorithm, and the L and L^T obtained are well-conditioned with $\kappa_\infty(L) = 617.0$ and $\kappa_\infty(L^T) = 12.4$. With the matrix extremely close to being singular (or being indefinite), the two-sided Jacobi for this matrix fails as it encounters a negative diagonal. The result listed is based on computing singular values by the one-sided Jacobi.

We see from the table that our algorithm computes $\lambda = 1e-15$ to the order of machine precision, while both the one-sided Jacobi and `eig` lost all significant digits. All other eigenvalues are computed to the order of machine precision by all three algorithms.

Table 1. Computed $\widehat{\lambda}$ and relative error for approximating $\lambda = 1e-15$

	$\widehat{\lambda}$	$rel.error = \frac{\lvert\lambda-\widehat{\lambda}\rvert}{\lambda}$
Algorithm 3	$1.00000000000000070e-15$	$5.9e-16$
one-sided Jacobi	$1.88038238345742960e-14$	$1.8e1$
`eig`	$7.14271307031277840e-14$	$7.0e1$

More generally we can construct a set of random test matrices like this with a known tiny eigenvalue as follows. We first construct an $n \times n$ random sparse matrix with normally distributed random entries and then take negative absolute value

(`B=-abs(sprandn(n,n,1))` in MATLAB); we then take the strict lower triangular part and symmetrize it (`B=tril(B, -1); A=B+B'`); we then assign to every zero off-diagonal entry of B a random number uniformly distributed in $[0,1]$ multiplied by η. We now let the diagonals of A be such that the row sum of A is λ. Then λ is an eigenvalue of A with $e=[1,\cdots,1]^T$ as an eigenvector. The diagonally dominant part can be computed as

$$v_i = a_{ii} - \sum_{j\neq i} |a_{ij}| = \lambda - \sum_{j\neq i, b_{ij}=0} 2|a_{ij}|$$

which is computed to high relative accuracy if η is sufficiently small so that $v_i \approx \lambda$. We have tested Algorithm 3 on such random matrices with n up to 500, and we have obtained a similar result as in Table 1.

Example 2. This matrix is modified from the one in Example 1 so that there are two close small eigenvalues. A is a 20×20 symmetric matrix whose off-diagonals are -1 except the last column and the last row, where they are $\eta = 10^{-16}$. We let the diagonals of A be such that the row sums of A are all $\lambda = 10^{-13}$. Then λ is an eigenvalue, but there is a smaller eigenvalue. We use MATLAB's Symbolic Toolbox with 200-digits arithmetic to compute all eigenvalues. The two smallest eigenvalues computed are $\lambda_1 = 9.8000000000000012e-14$ and $\lambda_2 = 1.0000000000000000e-13$.

We list and compare the relative errors for the two smallest eigenvalues as computed by Algorithm 3, by the Jacobi algorithm (two-sided for eigenvalues) and by MATLAB's `eig` in Table 2. Again, no pivoting scheme is used in our LDL^T factorization algorithm, and the L and L^T obtained are well-conditioned with $\kappa_\infty(L) = 85.4$ and $\kappa_\infty(L^T) = 8.9$. Our algorithm computes both eigenvalues to the order of machine precision while the two-sided Jacobi and `eig` can obtain about 2 to 4 significant digits.

TABLE 2. Relative errors for approximating $\lambda_1 = 9.8000000000000012e-14$ and $\lambda_2 = 1.0000000000000000e-13$

	$\frac{\lvert\lambda_1-\widehat{\lambda_1}\rvert}{\lambda_1}$	$\frac{\lvert\lambda_2-\widehat{\lambda_2}\rvert}{\lambda_2}$
Algorithm 3	$3.9e-16$	$1.3e-16$
two-sided Jacobi	$1.1e-4$	$2.3e-3$
`eig`	$2.2e-2$	$1.8e-2$

Our third example is a randomly generated extremely ill-conditioned nonsymmetric matrix taken from [11] and slightly modified so that A is not an M-matrix.

Example 3. We construct a matrix $\mathcal{D}(A_D, v)$ as follows.

1. We choose a 20×20 random sparse matrix A with nonzero offdiagonal entries random numbers uniformly distributed in $[-1,0]$ (`B=-sprand(n,n,1)` in MATLAB); we then replace the zero off-diagonal entries by a random number uniformly distributed in $[0,1]$ multiplied by $\eta = 1e-15$.
2. We choose 20 random numbers of the form $r\cdot 10^k$ for the diagonally dominant parts v_i, where r is a uniform random number in $[0,1]$ and k is a uniform random integer in $[-40,-20]$.
3. We multiply the i-th row of A and v_i by another random number of the form $r\cdot 10^j$, where r is a uniform random number in $[0,1]$ and j is a uniform random integer in $[-100,100]$.

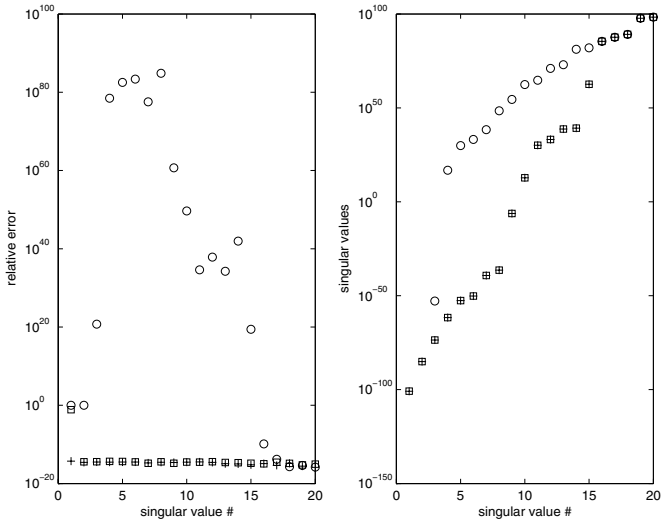

FIGURE 1. Left: Relative errors of singular values; Right: Singular values. +-sign: Algorithm 2 (or symbolic toolbox on the right); square box: one-sided Jacobi; *o*-sign: `svd` of MATLAB.

We compute the singular values of A using MATLAB's Symbolic Toolbox with 400-digits arithmetic and consider the result as "exact"" for comparison. (The results obtained using 600-digits arithmetic are the same.) The singular values are plotted in + in Figure 1 (right). We then compute the singular values by Algorithm 2, by the one-sided Jacobi and by `svd` of MATLAB, and we compare them with the ones obtained in 400-digits arithmetic. The relative errors are plotted in Figure 1 (left) and the singular values are plotted in Figure 1 (right).

We see that Algorithm 2 computes all singular values to the order of machine precision (the relative errors are all less than $7 \cdot 10^{-15}$), while `svd` only computes the largest four correctly (the smallest two were 0 and were not shown on the plot). The one-sided Jacobi did not compute the smallest singular value accurately with relative error $7.2e-2$, while all other singular values are accurate to the order of machine precision. This can be explained by the fact that all tiny singular values except one is due to extremely bad scaling, which the Jacobi algorithm can effectively overcome.

In our implementation of Algorithm 1 here, we have used both diagonal pivoting and the column diagonal dominance pivoting which return the same result. The condition numbers for the L and U factors are $\kappa_\infty(L) = 1.03$ and $\kappa_\infty(U) = 9.26$. We also note that all one-sided Jacobi we have mentioned are the right-handed version, but we have also tested the left-handed Jacobi for this problem, which give a similar result.

Finally, we present an example to demonstrate that the algorithm can detect singularity and compute the rank exactly.

Example 4. Let B be the 4×4 matrix whose off-diagonals are -1 and whose diagonally dominant parts are all zero. For $D_1 = \text{diag}\{1, -1, 1, -1\}$ and $D_2 = \text{diag}\{1, 1, -1, -1\}$, let $A = \text{diag}\{D_1 B D_1, D_2 B D_2\}$. We show the diagonal blocks below:

```
 3     1    -1     1                    3    -1     1     1
 1     3     1    -1                   -1     3     1     1
-1     1     3     1                    1     1     3    -1
 1    -1     1     3                    1     1    -1     3
```

A has exactly two zero pivots, one from each block. Applying Algorithm 1, we obtained the following for L and d ($D = \mathrm{diag}\{d\}$) with the permutation [1 5 3 6 2 7 4 8].

```
L =
    1.0000       0       0       0       0       0       0       0
         0  1.0000       0       0       0       0       0       0
   -0.3333       0  1.0000       0       0       0       0       0
         0 -0.3333       0  1.0000       0       0       0       0
    0.3333       0  0.5000       0  1.0000       0       0       0
         0  0.3333       0  0.5000       0  1.0000       0       0
    0.3333       0  0.5000       0 -1.0000       0  1.0000       0
         0  0.3333       0  0.5000       0 -1.0000       0  1.0000

d =
    3.0000  3.0000  2.6667  2.6667  2.0000  2.0000       0       0
```

We see that the algorithm computes two zero pivots exactly. Further applying Algorithm 3, we obtain two zero eigenvalues exactly.

6. Concluding remarks

We have obtained an algorithm that computes all singular values of a diagonally dominant matrix to the order of machine precision. A forward error analysis is given to demonstrate the high relative accuracy of the algorithm. As a byproduct, ranks and zero singular values are computed exactly. It will be interesting to see if the constants in our error bounds can be improved. Although there are substantial difficulties, it might be possible to obtain better bounds along the analysis for the GTH-algorithm by O'Cinneide [23, 24]. It is also possible that a backward error analysis in combination with the perturbation bounds [28] could lead to better bounds. We shall study these issues in our future works.

Acknowledgement

Professor Jim Demmel suggested to me the possibility of computing an accurate LDU factorization of diagonally dominant matrices. I am grateful to him for posing the problem and for numerous suggestions on early drafts of the paper, including one that led to the elimination of a significant condition in the error analysis. I am also grateful to two anonymous referees for their very careful reading and numerous constructive suggestions that have improved the paper.

References

[1] S. Alfa, J. Xue and Q. Ye, *Entrywise perturbation theory for diagonally dominant M-matrices with applications,* Numer. Math. 90(2002):401-414. MR1884223 (2002j:65049)

[2] S. Alfa, J. Xue and Q. Ye, *Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix,* Math. Comp. 71(2002):217-236. MR1862996 (2002h:65054)

[3] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices,* SIAM J. Num. Anal. 27(1990):762-791. MR1041262 (91g:65071)

[4] P. Deift, J. Demmel, L.-C. Li, C. Tomei, *The bidiagonal singular value decomposition and Hamiltonian mechanics,* SIAM J. Num. Anal. 28(1991):1463-1516. MR1119279 (92i:65071)

[5] J. Demmel, *Applied Numerical Linear Algebra,* SIAM, Philadelphia, 1997. MR1463942 (98m:65001)
[6] J. Demmel, *Accurate SVDs of structured matrices,* SIAM J. Matrix Anal. Appl. 21(1999):562-580. MR1742810 (2001h:65036)
[7] J. Demmel and W. Gragg, *On computing accurate singular values and eigenvalues of matrices with acyclic graphs*, Linear Algebra Appl. 185 (1993):203-217. MR1213179 (94h:65044)
[8] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić and Z. Drmač, *Computing the singular value decomposition with high relative accuracy,* Linear Alg. Appl. 299(1999):21-80. MR1723709 (2000j:65044)
[9] J. Demmel and Y. Hida, *Accurate and Efficient Floating Point Summation*, SIAM J. Sci. Comput. 25(4):1214–1248, 2003. MR2045054 (2005b:65055)
[10] J. Demmel and W. Kahan, *Accurate singular values of bidiagonaly matrices,* SIAM J. Sci. Stat. Comput. 11(1990):873-912. MR1057146 (91i:65072)
[11] J. Demmel, P. Koev, *Accurate SVDs of weakly diagonally dominant M-matrices,* Numer. Math. 98(2004):99-104. MR2076055 (2005g:65069)
[12] J. Demmel, P. Koev, *Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials*, Linear Algebra Appl. 417(2006):382-396. MR2250320 (2007e:65038)
[13] J. Demmel and K. Veselic, *Jacobi's method is more accurate than QR,* SIAM J. Matrix Anal. Appl. 13(1992):1204-1246. MR1182723 (93e:65057)
[14] F. Dopico and P. Koev, *Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices*, SIAM J. Matrix Anal. Appl., 28 (2006): 1126-1156. MR2276557
[15] K. Fernando, B. Parlett, *Accurate singular values and differential qd algorithms,* Numerische Mathematik 67 (1994):191-229. MR1262781 (95a:65071)
[16] G. Golub and C. Van Loan, *Matrix Computations,* 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989. MR1002570 (90d:65055)
[17] W.K. Grassmann, M.J. Taksar and D.P. Heyman, *Regenerative analysis and steady-state distributions for Markov chains,* Operations Research 33(1985):1107-1116. MR806921 (86k:60125)
[18] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996. MR1368629 (97a:65047)
[19] W. Kahan, *A survey of error analysis,* In Proc. IFIP Congress, Ljubljana, Information Processing 71, North-Holland, Amsterdam, 1972, pp 1214-1239. MR0458845 (56:17045)
[20] P. Koev, *Accurate Eigenvalues and SVDs of Totally Nonnegative Matrices,* SIAM J. Matrix Anal. Appl. 27(2005):1-23. MR2176803 (2006j:15067)
[21] P. Koev and F. Dopico, *Accurate eigenvalues of certain sign regular matrices,* Linear Algebra Appl. 424 (2007):435-447. MR2329485
[22] R. Mathias, *Accurate eigensystem computations by Jacobi methods,* SIAM J. Matrix Anal. Appl. 16 (1995):977-1003. MR1337657 (96f:65045)
[23] C. O'Cinneide, *Entrywise perturbation theory and error analysis for Markov chains,* Numer. Math. 65(1993):109-120. MR1217442 (94k:60101)
[24] C. O'Cinneide, *Relative-error bounds for the LU decomposition via the GTH algorithm,* Numer. Math. 73(1996):507-519. MR1393178 (97h:65033)
[25] J. M. Peña, *LDU decompositions with L and U well conditioned,* Electr. Trans. Numer. Anal. 18 (2004): 198-208. MR2150769 (2006b:65039)
[26] R.S. Varga, *Matrix Iterative Analysis,* 2nd ed. Springer-Berlag, Berlin, 2000. MR1753713 (2001g:65002)
[27] J. Wilkinson, *The algebraic eigenvalue problem,* Oxford University Press, 1965. MR0184422 (32:1894)
[28] Q. Ye, *Relative Perturbation Bounds for Eigenvalues of Symmetric Positive Definite Diagonally Dominant Matrices*, SIAM J. Matrix Anal. Appl., to appear.

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

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