SIAM J. Matrix Anal. Appl.
Vol. 27, No. 2, pp. 561–581

© 2005 Society for Industrial and Applied Mathematics

PERTURBATION THEORY FOR FACTORIZATIONS OF LU TYPE THROUGH SERIES EXPANSIONS*

FROILÁN M. DOPICO† AND JUAN M. MOLERA†

Abstract. Component- and normwise perturbation bounds for the block LU factorization and block LDL* factorization of Hermitian matrices are presented. We also obtain, as a consequence, perturbation bounds for the usual pointwise LU, LDL*, and Cholesky factorizations. Some of these latter bounds are already known, but others improve previous results. All the bounds presented are easily proved by using series expansions. Given a square matrix $A = LU$ having the LU factorization, and a perturbation E, the LU factors of the matrix $A+E = \widetilde{L}\widetilde{U}$ are written as two convergent series of matrices: $\widetilde{L} = \sum_{k=0}^{\infty} L_k$ and $\widetilde{U} = \sum_{k=0}^{\infty} U_k$, where $L_k = O(\|E\|^k)$, $U_k = O(\|E\|^k)$, and $L_0 = L$, $U_0 = U$. We present expressions for the matrices L_k and U_k in terms of L, U, and E. The domain and the rate of convergence of these series are studied. Simple bounds on the remainders of any order of these series are found, which significantly improve the bounds on the second-order terms existing in the literature. This is useful when first-order perturbation analysis is used.

Key words. LU factorization, Cholesky factorization, block LU factorization, diagonal pivoting method, block LDL^T factorization, perturbation theory, series expansion

AMS subject classifications. 65F35, 15A23

DOI. 10.1137/040612142

1. Introduction.

Let A be an $n \times n$ matrix whose leading principal submatrices are nonsingular. Then there exists a unique unit lower triangular matrix L and a unique upper triangular matrix U such that $A = LU$. This is known as LU factorization of A. The LU factorization is one of the most important matrix factorizations appearing in numerical analysis and has several variants that are used in different contexts: Cholesky, LDL*, block LU, and block LDL* factorizations. As usual L^* denotes the conjugate transpose of L.

Traditionally, the LU factorization and its variants have been used to solve linear systems of equations, while in solving spectral problems orthogonal factorizations have been preferred because of their good stability properties [15]. However, in the last decade the LU factorization has been employed to solve structured spectral problems [12, 19]. For most applications related to the solution of linear systems it is the backward error and not the forward error of the LU factorization that matters; however, for the application of LU to the computation of the singular value decomposition with a high relative accuracy, what is needed instead is to compute the LU factors with small forward errors [12]. The question of how big the forward errors are may be answered by combining backward errors with an adequate perturbation theory for the LU factorization. This paper presents a new broad-scoped and unifying approach to this theory, which allows us to get, for the usual point factorizations, some bounds that are already known and, furthermore, to improve previously existing bounds. Most important, this general analysis also includes, for the first time, perturbation bounds for the block LU factorization. The block LU factorization appears

*Received by the editors July 23, 2004; accepted for publication (in revised form) by N. J. Higham May 6, 2005; published electronically November 22, 2005. This work was partially supported by the Ministerio de Educación y Ciencia of Spain through grant BFM 2003-00223.

http://www.siam.org/journals/simax/27-2/61214.html

†Departamento de Matemáticas, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain (dopico@math.uc3m.es, molera@math.uc3m.es).

in different instances [17, Chapter 13], and it is used in the most common algorithms to solve Hermitian indefinite systems of linear equations [1].

In the literature there are several papers dealing with the perturbation theory of usual, or pointwise, LU or Cholesky factorizations. In these papers, first-order perturbation bounds are frequently used. In this respect, we will say that strict or rigorous perturbation bounds are those without higher-order terms.

The first works [3, 23, 26] dealing with this problem obtained normwise perturbation bounds. Sun [27, 28] got an improvement over previously existing results by proving componentwise perturbation bounds. In [24], Stewart shed new light on the problem by giving explicit formulae for the first-order terms of the series of the perturbed factors of LU, Cholesky, and QR factorizations. Moreover, normwise bounds on the second-order terms, i.e., the remainder of the series, were also presented.

There has also been relevant work on the true normwise condition number for the Cholesky factorization by Chang, Paige, and Stewart [8, 25]. Ideas similar to those in [8] have also been used in [9] to study the normwise condition number of the LU factorization and in [10] to give a structured perturbation analysis of the Cholesky factorization. Another structured analysis of component- and normwise condition numbers of the LU factors of tridiagonal matrices has been done in [4], where easily computable expressions of the true condition numbers are presented. In these works, many different techniques have been used. One of the goals of this paper is to present a unifying perturbation theory for LU-type factorizations.

In this work, expressions for the terms of any order in the series expansions of the perturbed LU or Cholesky factors are presented. These series are easily and naturally extended to cover the block LU factorization and block LDL* factorization of Hermitian indefinite matrices. Extremely simple componentwise upper bounds for all the terms in these series will be obtained. Summing up the series for these upper bounds, reliable bounds on the remainders of any order of the series and rigorous component- and normwise perturbation bounds for the LU factors are proved. The bounds for the block factorizations are the first existing perturbation bounds for these factorizations. Some of the bounds for the usual factorizations are already known, but others improve previously existing bounds.

The paper is organized as follows. In section 2 the series of the LU factors are written as two convergent series of matrices using basic results. The domain of convergence of these series is also established. In section 3 simple componentwise bounds for the terms of these series are proved. This is a technical section which contains the crucial result on which the rest of the paper is based: Theorem 3.6, which remains valid for block factorizations. As a simple consequence, in section 4 the main theorems on the perturbation of the LU factors of a matrix are stated; they are a comprehensive summary of the component- and normwise convergence properties of the series. Moreover, it is shown how to get strict component- and normwise perturbation bounds from these series. Along this line, section 5 presents a rigorous perturbation theory for the block LU factorization. Using the results in section 5, and properties of Hermitian matrices, perturbation bounds for the block LDL* factorization are developed in section 6. The case of D being diagonal is considered and compared with previous results. Results for the Cholesky factorization are presented in section 7.

2. Series for the LU factors of a matrix.

We will deal with a nonsingular $n \times n$ complex matrix $A = [a_{ij}]$ such that its leading principal submatrices are nonsingular. Therefore, A has a unique LU factorization: $A = LU$. Throughout this paper $\| \cdot \|$ will denote a family of absolute, consistent matrix norms; i.e., if $|A|$ is the matrix with entries $(|A|)_{ij} = |a_{ij}|$, then $|A| \leq |B|$ implies $\|A\| \leq \|B\|$ and $\|AB\| \leq \|A\|\|B\|$. Norms of this type are the Frobenius norm, the 1-norm, and the ∞-norm, but not the 2-norm. Notice that if $A(1:k, 1:k)$ denotes the kth leading principal submatrix of A, then $\|A(1:k, 1:k)\| \leq \|A\|$.

Let us consider a perturbation $\widetilde{A} = A+E$ of A and assume that $\|L^{-1}EU^{-1}\| < 1$. In this case $\widetilde{A}$ has also a unique LU factorization because the matrix $I + L^{-1}EU^{-1}$ has all its leading principal submatrices nonsingular and has the following unique LU factorization:

$$I + L^{-1}EU^{-1} = \mathcal{L}\mathcal{U}. \tag{2.1}$$

Therefore,

$$\widetilde{A} = L(I + L^{-1}EU^{-1})U = (L\mathcal{L})(\mathcal{U}U) \equiv \widetilde{L}\widetilde{U} \tag{2.2}$$

is the unique LU factorization of the nonsingular matrix $\widetilde{A}$, because it is well known that if the LU factorization of a nonsingular matrix exists, then it is unique.

The set of matrices E for which we can guarantee that $I+L^{-1}EU^{-1}$ has a unique LU factorization can be enlarged by using the spectral radius of $|L^{-1}EU^{-1}|$, denoted by $\rho(|L^{-1}EU^{-1}|)$. Notice that

$$\rho(|L^{-1}EU^{-1}|) < 1 \tag{2.3}$$

implies that $I + L^{-1}EU^{-1}$ has its leading principal submatrices nonsingular. The reason is that for any $n \times n$ matrix F and for any k such that $1 \leq k \leq n$,

$$\rho(F(1:k, 1:k)) \;\leq\; \rho(|F(1:k, 1:k)|) \;\leq\; \rho(|F|) \tag{2.4}$$

holds; see [18, Theorem 8.1.18, Corollary 8.1.20]. Besides this, $\rho(|L^{-1}EU^{-1}|) \leq \|L^{-1}EU^{-1}\|$ implies that (2.3) remains valid for a wider set of perturbation matrices E than $\|L^{-1}EU^{-1}\| < 1$, which can be more easily checked in practice.

Equation (2.2) shows that for getting a Taylor-like series for the LU factors of $\widetilde{A}$ it is enough to study the series of the LU factors of perturbations $I + F$ of the identity matrix. For the sake of simplicity, let us introduce a complex variable z, write the perturbation as $I + zF$, and consider the LU factors of this matrix as complex functions of z, i.e., $I + zF = \mathcal{L}(z)\mathcal{U}(z)$ (at the end we will set $z = 1$ and $\mathcal{L} = \mathcal{L}(1)$, $\mathcal{U} = \mathcal{U}(1)$). We know that these LU factors exist and are unique in the open disk $D_F = \{z \in \mathbb{C} : |z| < \frac{1}{\rho(|F|)}\}$. Moreover, the entries of $\mathcal{L}(z)$ and $\mathcal{U}(z)$ are rational functions of z with nonzero denominator in D_F. This can be seen by examining the steps of Gaussian elimination or, directly, by using the well-known determinantal formulas for the entries of the LU factors of a matrix (see [17, p. 161]). Thus the entries of $\mathcal{L}(z)$ and $\mathcal{U}(z)$ are analytic functions of z in D_F, and the corresponding Taylor series around $z = 0$ are convergent power series in D_F. These scalar series can be compactly written in matrix notation for every $z \in D_F$ as follows:

$$\mathcal{L}(z) = \sum_{k=0}^{\infty} z^k \mathcal{L}_k \quad \text{and} \quad \mathcal{U}(z) = \sum_{k=0}^{\infty} z^k \mathcal{U}_k.$$

The goal is to get simple expressions for the lower triangular matrices $\mathcal{L}_k$ and the upper triangular matrices $\mathcal{U}_k$. Taking into account that inside D_F we can sum up and multiply in the usual way the Taylor series of the entries of the matrices $\mathcal{L}(z)$ and $\mathcal{U}(z)$ to get the convergent power series, we have

$$I + zF = \mathcal{L}(z)\mathcal{U}(z) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} \mathcal{L}_j \mathcal{U}_{k-j} \right) z^k.$$

Consequently,

$$I = \mathcal{L}_0 \mathcal{U}_0, \quad F = \mathcal{L}_0 \mathcal{U}_1 + \mathcal{L}_1 \mathcal{U}_0, \tag{2.5}$$

$$0 = \mathcal{L}_0 \mathcal{U}_k + \mathcal{L}_1 \mathcal{U}_{k-1} + \cdots + \mathcal{L}_{k-1} \mathcal{U}_1 + \mathcal{L}_k \mathcal{U}_0 \quad \text{for } k \geq 2. \tag{2.6}$$

Notice that $\mathcal{L}(z)$ has diagonal entries equal to 1; therefore, $\mathcal{L}_0$ has also diagonal entries equal to 1 and $\mathcal{L}_k$ for $k \geq 1$ are strictly lower triangular matrices. Using this in (2.5) and (2.6), we get the following recurrence relation for the matrices $\mathcal{L}_k$ and $\mathcal{U}_k$:

$$\begin{aligned} &\mathcal{L}_0 = \mathcal{U}_0 = I, \quad \mathcal{L}_1 + \mathcal{U}_1 = F, \\ &\mathcal{L}_k + \mathcal{U}_k = -\mathcal{L}_1 \mathcal{U}_{k-1} - \mathcal{L}_2 \mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1} \mathcal{U}_1 \quad \text{for } k \geq 2. \end{aligned}$$

The key point to obtain $\mathcal{L}_k$ and $\mathcal{U}_k$ for $k \geq 1$ is that $\mathcal{L}_k$ is strictly lower triangular and $\mathcal{U}_k$ is upper triangular. The previous discussion can be summarized in the next theorem, where we have set $z = 1$ and the following notation is used: Given a matrix B with entries b_{ij}, we call B_L and B_U, respectively, the *strictly* lower triangular part and the upper triangular part of B, i.e.,

$$(B_L)_{ij} = \begin{cases} b_{ij} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases} \qquad (B_U)_{ij} = \begin{cases} b_{ij} & \text{if } i \leq j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

THEOREM 2.1. *Let F be an $n \times n$ complex matrix, such that $\rho(|F|) < 1$. Then*

1. *$I + F$ has a unique LU factorization: $I + F = \mathcal{L}\mathcal{U}$;*
2. $$\mathcal{L} = \sum_{k=0}^{\infty} \mathcal{L}_k \quad \textit{and} \quad \mathcal{U} = \sum_{k=0}^{\infty} \mathcal{U}_k,$$

where

$$\mathcal{L}_0 = I, \quad \mathcal{U}_0 = I; \qquad \mathcal{L}_1 = F_L, \quad \mathcal{U}_1 = F_U, \tag{2.8}$$

$$\mathcal{L}_k = (-\mathcal{L}_1 \mathcal{U}_{k-1} - \mathcal{L}_2 \mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1} \mathcal{U}_1)_L \quad \textit{for } k \geq 2, \tag{2.9}$$

$$\mathcal{U}_k = (-\mathcal{L}_1 \mathcal{U}_{k-1} - \mathcal{L}_2 \mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1} \mathcal{U}_1)_U \quad \textit{for } k \geq 2; \tag{2.10}$$

3. *$\mathcal{L}_k = O(\|F\|^k)$ and $\mathcal{U}_k = O(\|F\|^k)$.*

The last item can be easily proven by using an inductive argument. In fact it can be proven (from Theorem 2.2) that $\|\mathcal{L}_k + \mathcal{U}_k\| \leq C_{k-1} \|F\|^k$ for $k \geq 1$. The numbers C_k appear in many counting problems; they are known as the *Catalan numbers* [20, p. 315] and they grow quickly with k. We will see in section 3 that a more elaborate argument allows us to bound $|\mathcal{L}_k + \mathcal{U}_k| \leq |F|^k$.

The first-order terms were obtained in [24]; now we can easily write explicit expressions for the higher-order terms. For instance,

$$\mathcal{L}_2 + \mathcal{U}_2 = -F_L F_U, \tag{2.11}$$

$$\mathcal{L}_3 + \mathcal{U}_3 = F_L (F_L F_U)_U + (F_L F_U)_L F_U, \tag{2.12}$$

$$\begin{aligned} \mathcal{L}_4 + \mathcal{U}_4 = &-F_L (F_L (F_L F_U)_U)_U - F_L ((F_L F_U)_L F_U)_U \\ &-(F_L F_U)_L (F_L F_U)_U - (F_L (F_L F_U)_U)_L F_U - ((F_L F_U)_L F_U)_L F_U. \end{aligned} \tag{2.13}$$

For a series to be useful from the point of view of applications, its rate of convergence has to be fast enough. Taking into account that the number of terms in the recurrence relation of $\mathcal{L}_k$ and $\mathcal{U}_k$ increases very fast with k, a convenient analysis of the rate of convergence of these series requires a careful approach that will be undertaken in the following sections. We begin this by giving a more explicit description of $\mathcal{L}_k$ and $\mathcal{U}_k$. Notice that the next result holds for any square matrix F, independently of the magnitude of its spectral radius.

THEOREM 2.2. *Let F be an $n \times n$ complex matrix, and $\mathcal{L}_k$ and $\mathcal{U}_k$, $k \geq 1$, be the matrices defined by the recurrence relation* (2.8)–(2.10). *Then $\mathcal{L}_k + \mathcal{U}_k$ is $(-1)^{k+1}$ times the sum of all the matrices that can be formed with the following symbolic rules.*

1. *Specify, by using parentheses, all the possible different orders to multiply $F^k = F\,F \cdots F$.*
2. *Given an order of multiplication, every individual multiplication CD is converted to $C_L D_U$.*

Before the proof of this theorem, let us give an example showing how to use it.

EXAMPLE 1. *The product $F^3 = FFF$ can be computed in two orders, $F(FF)$ and $(FF)F$. In any of these orders there are two individual multiplications, one inside the parentheses and the other outside. The second rule of the previous theorem applies to the inner and outer multiplications to yield $F_L(F_L F_U)_U$ and $(F_L F_U)_L F_U$. Therefore, Theorem* 2.2 *gives $\mathcal{L}_3 + \mathcal{U}_3 = F_L(F_L F_U)_U + (F_L F_U)_L F_U$, which is the result obtained in* (2.12) *by using Theorem* 2.1.

Proof of Theorem 2.2. The proof is by induction on k. The result is trivially true for $k = 1, 2$. Assume that it is true for any j such that $1 \leq j \leq k$. It is known from (2.8)–(2.10) that

$$\mathcal{L}_{k+1} + \mathcal{U}_{k+1} = -\mathcal{L}_1 \mathcal{U}_k - \mathcal{L}_2 \mathcal{U}_{k-1} - \cdots - \mathcal{L}_k \mathcal{U}_1. \tag{2.14}$$

In any order of multiplication of F^{k+1}, specified with parentheses, there is one and only one multiplication operator which remains outside all parentheses. This is the final multiplication to be performed. Notice that $\mathcal{L}_j = (\mathcal{L}_j + \mathcal{U}_j)_L$ and $\mathcal{U}_j = (\mathcal{L}_j + \mathcal{U}_j)_U$; therefore, $-\mathcal{L}_j \mathcal{U}_{k+1-j}$ is $(-1)^{k+2}$ times the sum of all the matrices that can be formed as follows: (1) consider all the possible orders to multiply F^{k+1} in which the final multiplication is between the factors j and $j+1$; (2) apply to each of these orders the second rule in Theorem 2.2. Substituting this sum in (2.14) for $1 \leq j \leq k$ the theorem is proved, because the set of all the possible orders to multiply F^{k+1} is the union of the sets of all the possible orders to multiply F^{k+1} in which the final multiplication is between the factors j and $j+1$ for $j = 1, \ldots, k$. □

3. Componentwise bounds on $\mathcal{L}_k$ and $\mathcal{U}_k$.

The matrices $\mathcal{L}_k$ and $\mathcal{U}_k$ appearing in Theorem 2.1 and characterized in Theorem 2.2 are involved functions of F; however, they can be simply bounded by powers of the absolute value of F.

THEOREM 3.1. *Let F be an $n \times n$ complex matrix, and let $\mathcal{L}_k$ and $\mathcal{U}_k$, $k \geq 1$, be the matrices defined by the recurrence relation* (2.8)–(2.10)*; then*

$$|\mathcal{L}_k + \mathcal{U}_k| \leq |F|^k \quad \textit{for} \quad k \geq 1.$$

Therefore, $|\mathcal{L}_k| \leq (|F|^k)_L$ and $|\mathcal{U}_k| \leq (|F|^k)_U$.

The rest of this paper is based on this theorem, and on its generalized version Theorem 3.6. The proof of both theorems is the goal of this section. Since the proof is long, although elementary, the reader who is interested only in applications can skip this section. It is worthwhile to remark that Theorem 3.1 is valid for any square matrix F, independently of the magnitude of its spectral radius.

The next lemma shows that we can focus on the series of the LU factors of $|F|$.

LEMMA 3.2. *Let F be an $n \times n$ complex matrix, let $\mathcal{L}_k$ and $\mathcal{U}_k$, $k \geq 1$, be the matrices defined by the recurrence relation* (2.8)–(2.10)*, and let $\overline{\mathcal{L}}_k$ and $\overline{\mathcal{U}}_k$ be the matrices defined by the recurrence relation* (2.8)–(2.10) *applied to the matrix $|F|$, i.e., $\overline{\mathcal{L}}_1 = |F|_L$ and $\overline{\mathcal{U}}_1 = |F|_U$. Then*

$$|\mathcal{L}_k + \mathcal{U}_k| \;\leq\; (-1)^{k+1}\,(\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k) \;=\; |\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k| \quad \textit{for} \quad k \geq 1.$$

Proof. In Theorem 2.2, $(-1)^{k+1}\,(\mathcal{L}_k + \mathcal{U}_k)$ is expressed as a sum of certain matrices. Apply the triangular inequality of the absolute value to this sum. Notice that for any matrices C and D the inequality $|C_L D_U| \leq |C_L||D_U| = |C|_L |D|_U$ holds. This is applied to the absolute value of the terms of the sum appearing in Theorem 2.2 to get the same terms for $|F|$ instead of F. □

Before starting with the technical details of the section, we would like to show with an example that the idea of the proof of Theorem 3.1 is very simple.

EXAMPLE 2. *Consider the case $k = 4$ and denote for simplicity $G \equiv |F|$:*

$$\begin{aligned}
|F|^4 &= (G_L + G_U)^4 \\
&\geq G_L G_L G_L G_U + G_L G_L G_U G_U + G_L G_U G_L G_U + G_L G_U G_U G_U \\
&= G_L G_L G_L G_U + G_L (G_L G_U)_L G_U + G_L (G_L G_U)_U G_U \\
&\quad + G_L G_U G_L G_U + G_L G_U G_U G_U \\
&\geq G_L (G_L (G_L G_U)_U)_U + G_L ((G_L G_U)_L G_U)_U + (G_L (G_L G_U)_U)_L G_U \\
&\quad + (G_L G_U)_L (G_L G_U)_U + ((G_L G_U)_L G_U)_L G_U \\
&= -(\overline{\mathcal{L}}_4 + \overline{\mathcal{U}}_4) \geq |\mathcal{L}_4 + \mathcal{U}_4|.
\end{aligned}$$

3.1. Tree terminology and auxiliary results. To prove Theorem 3.1 it is useful to introduce some tree terminology. Among the terminologies commonly used in tree theory we will follow that used in [20, section 8.1]. Since Theorem 3.1 is trivial for $k = 1$, we will consider only the case $k \geq 2$.

DEFINITION 3.3. *Let the matrix Π be one of the summands in the expression of $(-1)^{k+1}(\mathcal{L}_k + \mathcal{U}_k)$, $k \geq 2$, as the sum defined in Theorem 2.2. $\mathcal{T}(\Pi)$ is the ordered full binary tree defined as follows.*

1. *The root vertex represents $\Pi = C_L D_U$. The left child of the root represents C_L, and the right child of the root represents D_U.*
2. *Any vertex ν of $\mathcal{T}(\Pi)$ different from the root represents a matrix B_L or B_U. If $B = P_L Q_U$, then the left child of ν represents P_L, and the right child of ν represents Q_U. If $B = F$, then ν is a leaf of $\mathcal{T}(\Pi)$.*

In the previous definition the word "full binary" appears because the internal vertices, i.e., those different from the leaves, have exactly two children. The word "ordered" appears because the children of the internal vertices are ordered from left to right. Moreover, notice that the left children are strictly lower triangular matrices, while the right children are upper triangular matrices. These meanings will be kept in the rest of the paper.

EXAMPLE 3. *Figure* 3.1 *shows the tree representing the summand*

$$(F_L F_U)_L\,((F_L (F_L F_U)_U)_L F_U)_U$$

of $-(\mathcal{L}_6 + \mathcal{U}_6)$.

Based on Definition 3.3, the next definition introduces other trees and some matrices necessary to prove Theorem 3.1. The matrices are important; the trees are

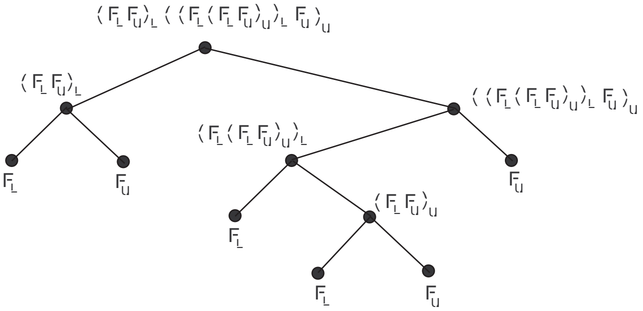
$\langle F_L F_U \rangle_L \langle \langle F_L \langle F_L F_U \rangle_U \rangle_L F_U \rangle_U$
$\langle F_L F_U \rangle_L$
$\langle \langle F_L \langle F_L F_U \rangle_U \rangle_L F_U \rangle_U$
$\langle F_L \langle F_L F_U \rangle_U \rangle_L$
F_L
F_U
F_U
$\langle F_L F_U \rangle_U$
F_L
F_L
F_U

FIG. 3.1. *The tree associated with* $(F_LF_U)_L\,((F_L(F_LF_U)_U)_LF_U)_U$.

used just to define the matrices. In this context, let us remember that the height of a rooted tree is the length of the longest path from the root to any vertex.

DEFINITION 3.4. *Let the matrix* Π *be one of the summands in the expression of* $(-1)^{k+1}(\mathcal{L}_k+\mathcal{U}_k)$, $k \geq 2$, *as the sum defined in Theorem* 2.2. *Let* h *be the height of the tree* $\mathcal{T}(\Pi)$. *Let us define the following trees and matrices.*

1. $\mathcal{T}_h(\Pi) := \mathcal{T}(\Pi)$. *The matrix* Π_h *is the product from left to right of the matrices represented by the leaves of* $\mathcal{T}_h(\Pi)$.

2. *For* $1 \leq j \leq (h-1)$, *the tree* $\mathcal{T}_{h-j}(\Pi)$ *is obtained from the tree* $\mathcal{T}_{h-(j-1)}(\Pi)$ *by removing all the pairs of leaves that are siblings and the edges incident with the removed leaves. The matrix* Π_{h-j} *is the product from left to right of the matrices represented by the leaves of* $\mathcal{T}_{h-j}(\Pi)$.

The order from left to right for the leaves of the trees in the previous definition is well defined independently of the graphic representation of the tree: given any parent vertex ν, its left child and all the descendants of this left child have to be considered as being to the left of the right child of ν and all the descendants of this right child.

EXAMPLE 4. *In the example in Figure* 3.1, *the matrices in Definition* 3.4 *are*

$$\Pi_h = \Pi_4 = F_LF_UF_LF_LF_UF_U, \tag{3.1}$$

$$\Pi_{h-1} = \Pi_3 = (F_LF_U)_LF_L(F_LF_U)_UF_U, \tag{3.2}$$

$$\Pi_{h-2} = \Pi_2 = (F_LF_U)_L(F_L(F_LF_U)_U)_LF_U, \tag{3.3}$$

$$\Pi_{h-3} = \Pi_1 = (F_LF_U)_L((F_L(F_LF_U)_U)_LF_U)_U. \tag{3.4}$$

Notice that Π_1 *is equal to* Π. *Moreover,* Π_4 *is the product of six factors,* Π_3 *is the product of four factors,* Π_2 *is the product of three factors, and* Π_1 *is the product of just two factors.*

For the trees and matrices defined in Definition 3.4 the following lemma holds.

LEMMA 3.5. *Let* $\mathcal{T}_{h-j}(\Pi)$ *and* Π_{h-j} *with* $0 \leq j \leq (h-1)$ *be, respectively, the trees and matrices defined in Definition* 3.4. *Then the following results hold.*

1. $\mathcal{T}_{h-j}(\Pi)$ *is an ordered full binary tree with height* $(h-j)$.

2. $\Pi = \Pi_1$.

3. *Let* $B_1, B_2, \ldots, B_q$ *be the matrices represented by the leaves of* $\mathcal{T}_{h-j}(\Pi)$, *i.e.,* $\Pi_{h-j} = B_1\,B_2\cdots B_q$. *Then* B_i *and* B_{i+1}, $1 \leq i \leq (q-1)$, *are represented by two leaves that are siblings if and only if* B_i *is strictly lower triangular and* B_{i+1} *is upper triangular.*

4. *Let us use the same notation as in the previous item:* $\Pi_{h-j} = B_1\,B_2\cdots B_q$ *with* $0 \leq j \leq (h-2)$. *Consider all the products* $B_i\,B_{i+1}$, $1 \leq i \leq (q-1)$, *such*

that B_i is strictly lower triangular and B_{i+1} is upper triangular; assume that there are p of these products. Write all these products as the sum of two terms, $B_iB_{i+1} = (B_iB_{i+1})_L + (B_iB_{i+1})_U$, and substitute and expand them in the expression for Π_{h-j}. Thus, Π_{h-j} is written as a sum of 2^p terms. Then one of these terms is $\Pi_{h-(j+1)}$.

Proof. 1. It is true by definition for $j = 0$. Assume that it is true for $(j - 1)$. $\mathcal{T}_{h-j}(\Pi)$ is obtained from $\mathcal{T}_{h-(j-1)}(\Pi)$ without changing the order between left and right children, and by removing pairs of leaves that are siblings. Thus, $\mathcal{T}_{h-j}(\Pi)$ is again an ordered full binary tree. Concerning the height, let ν be any leaf of $\mathcal{T}_{h-(j-1)}(\Pi)$ such that the path from the root to ν has length $(h - j + 1)$. Then the length of the path from the root to the sibling of ν is also $(h - j + 1)$. This implies that the sibling of ν is also a leaf; otherwise the height of $\mathcal{T}_{h-(j-1)}(\Pi)$ will be greater than $(h - j + 1)$. Therefore, ν and its sibling are not vertices of $\mathcal{T}_{h-j}(\Pi)$, but their parent is.

2. $\mathcal{T}_1(\Pi)$ is an ordered full binary tree with height 1, and it is also a subgraph of $\mathcal{T}(\Pi)$ containing the root of $\mathcal{T}(\Pi)$. This means that $\mathcal{T}_1(\Pi)$ is the tree formed by the root of $\mathcal{T}(\Pi)$ and its two children. From Definition 3.3 it is trivial that $\Pi = \Pi_1$.

3. If B_i and B_{i+1} are represented by two leaves that are siblings, the ordered property of the tree $\mathcal{T}_{h-j}(\Pi)$ implies that B_i is strictly lower triangular and B_{i+1} is upper triangular.

Assume now that B_i is strictly lower triangular and B_{i+1} is upper triangular. Let β and ν be the leaves of $\mathcal{T}_{h-j}(\Pi)$ representing, respectively, B_i and B_{i+1}. The ordering property of $\mathcal{T}_{h-j}(\Pi)$ implies that β is the left child of a parent β' and that ν is the right child of a parent ν'. The right child of β' and the left child of ν' are vertices of $\mathcal{T}_{h-j}(\Pi)$ because this is a full binary tree. Hence, if $\beta' \neq \nu'$, the matrices represented by the leaves descending from the right child of β' and from the left child of ν' should be between B_i and B_{i+1} in Π_{h-j}, but there is nothing between B_i and B_{i+1}. Therefore, $\beta' = \nu'$, and β and ν are siblings.

4. Let $\{B_{i_1}\,B_{i_1+1}, B_{i_2}\,B_{i_2+1}, \dots, B_{i_p}\,B_{i_p+1}\}$, $1 \le i_1 < i_2 < \cdots < i_p \le (q-1)$, be the set of all the products $B_i\,B_{i+1}$ with B_i strictly lower triangular, and B_{i+1} upper triangular. After the proposed expansion the summands of Π_{h-j} are all the 2^p possible products $B_1\,B_2 \cdots B_q$ in which every $B_{i_t}\,B_{i_t+1}$, $1 \le t \le p$, is replaced by $(B_{i_t}\,B_{i_t+1})_L$ or $(B_{i_t}\,B_{i_t+1})_U$. The previous item has shown that the leaves of $\mathcal{T}_{h-j}(\Pi)$ representing $\{B_{i_1}, B_{i_1+1}, B_{i_2}, B_{i_2+1}, \dots, B_{i_p}, B_{i_p+1}\}$ are precisely all the pairs of leaves that are siblings. These are the leaves to be removed to get $\mathcal{T}_{h-(j+1)}(\Pi)$. Therefore, $\mathcal{T}_{h-(j+1)}(\Pi)$ has as leaves those of $\mathcal{T}_{h-j}(\Pi)$ which have not been removed, and the parents of those which have been removed. But by Definition 3.3 these parents represent $(B_{i_t}\,B_{i_t+1})_L$ or $(B_{i_t}\,B_{i_t+1})_U$, $1 \le t \le p$. □

3.2. Proof of Theorem 3.1. The case $k = 1$ is trivial. We will focus on the case $k \ge 2$. By Lemma 3.2 it is enough to prove the result for the series of the LU factors of $I + |F|$. Remember that the terms of these series have been denoted by $\overline{\mathcal{L}}_k$ and $\overline{\mathcal{U}}_k$ according to Lemma 3.2. Let $\Pi^{(1)}, \dots, \Pi^{(r_k)}$ be the summands of $(-1)^{k+1}(\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k)$ in the sum defined in Theorem 2.2 applied to $|F|$. Let $h_1, \dots, h_{r_k}$ be the heights of the trees $\mathcal{T}(\Pi^{(1)}), \dots, \mathcal{T}(\Pi^{(r_k)})$ defined in Definition 3.3. We write $|F|^k$ as the following sums of matrices of nonnegative elements.

1. Expand $|F|^k = (|F|_L + |F|_U)^k = \sum_{i=1}^{2^k} S_i^{(0)}$ as the sum of all the 2^k different products of k factors $|F|_L$ or $|F|_U$.

2. For $j \ge 1$,

$$|F|^k = \sum_{i=1}^{p_j} S_i^{(j)} \tag{3.5}$$

is obtained from

$$|F|^k = \sum_{i=1}^{p_{j-1}} S_i^{(j-1)} \tag{3.6}$$

as follows: for every summand $S_i^{(j-1)}$, $1 \leq i \leq p_{j-1}$, which is the product of more than two factors, write every pair of consecutive factors of the type $C_L D_U$ as $C_L D_U = (C_L D_U)_L + (C_L D_U)_U$. Substitute and expand these expressions in $S_i^{(j-1)}$, writing the result as a sum of terms. Substitute the new expressions for $S_i^{(j-1)}$ in (3.6) and reenumerate the summands to get (3.5).

In the first place, notice that the set of matrices $\{\Pi_{h_1}^{(1)}, \ldots, \Pi_{h_{r_k}}^{(r_k)}\}$ is a subset of the set of summands $\{S_1^{(0)}, \ldots, S_{2^k}^{(0)}\}$, because by Definitions 3.3 and 3.4 the matrices $\Pi_{h_i}^{(i)}$, $1 \leq i \leq r_k$, are products of k factors $|F|_L$ or $|F|_U$.[1]

Now, we prove the following inductive step: if, for some index i, $\Pi_{h_i - j}^{(i)}$ with $(h_i - j) > 1$ is an element of the set $\{S_1^{(j)}, \ldots, S_{p_j}^{(j)}\}$, then $\Pi_{h_i-(j+1)}^{(i)}$ is an element of the set $\{S_1^{(j+1)}, \ldots, S_{p_{j+1}}^{(j+1)}\}$. This follows directly from item 4 in Lemma 3.5 and the way in which (3.5) is obtained from (3.6).

Finally, it is obvious that if, for some index i, $\Pi_1^{(i)}$ is an element of the set $\{S_1^{(j)}, \ldots, S_{p_j}^{(j)}\}$, then $\Pi_1^{(i)}$ is also an element of the set $\{S_1^{(j+1)}, \ldots, S_{p_{j+1}}^{(j+1)}\}$, because by the second item in Lemma 3.5, $\Pi_1^{(i)} = \Pi^{(i)}$, and it is the product of just two factors $C_L D_U$. Therefore, this summand remains unchanged when the set $\{S_1^{(j+1)}, \ldots, S_{p_{j+1}}^{(j+1)}\}$ is obtained from $\{S_1^{(j)}, \ldots, S_{p_j}^{(j)}\}$.

As a consequence of the previous argument, the set of summands of $(-1)^{k+1}(\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k)$, i.e., $\{\Pi^{(1)}, \ldots, \Pi^{(r_k)}\}$, is a subset of the set of summands $\{S_1^{(m)}, \ldots, S_{p_m}^{(m)}\}$, where $m = \max\{h_1 - 1, \ldots, h_{r_k} - 1\}$. Taking into account that all the matrices $\{\Pi^{(1)}, \ldots, \Pi^{(r_k)}\}$ are different and that the matrices $S_1^{(m)}, \ldots, S_{p_m}^{(m)}$ have nonnegative entries, then from (3.5), with $j = m$, $(-1)^{k+1}(\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k) = |\overline{\mathcal{L}}_k + \overline{\mathcal{U}}_k| \leq |F|^k$. Theorem 3.1 follows from Lemma 3.2.

3.3. A general version of Theorem 3.1.

To study the series and perturbation results corresponding to the Cholesky factorization, the factorization coming from the diagonal pivoting method, and block LU factorizations, a more general version of Theorem 3.1 is needed. This is presented in this subsection.

Consider a given $n \times n$ real matrix $[\lambda_{ij}]$ such that $0 \leq \lambda_{ij} \leq 1$ for any (i, j). For any $n \times n$ matrix $B = [b_{ij}]$, let us define the two $n \times n$ matrices $B_\Lambda = [(B_\Lambda)_{ij}]$ and $B_\Upsilon = [(B_\Upsilon)_{ij}]$ as follows:

$$(B_\Lambda)_{ij} = \lambda_{ij}\, b_{ij}, \quad (B_\Upsilon)_{ij} = (1 - \lambda_{ij})\, b_{ij}. \tag{3.7}$$

Notice that when $\lambda_{ij} = 1$ if $i > j$ and $\lambda_{ij} = 0$ otherwise, the usual strictly lower and upper triangular parts appearing in (2.7) are obtained. Another interesting case that we will use in the context of the Cholesky factorization, when B is Hermitian, is

[1] It is interesting to note that it may happen that $\Pi_{h_i}^{(i)} = \Pi_{h_j}^{(j)}$ with $i \neq j$. This happens, for instance, in the second and fourth summands in the right-hand side of (2.13). However, all the summands $S_i^{(0)}$, $1 \leq i \leq 2^k$, are different. This is consistent with the fact that we are speaking about *sets*, because for sets, it does not matter if an element is listed more than once. Our arguments are independent of these details, but they are cumbersome precisely due to the fact that equalities of the kind $\Pi_{h_i}^{(i)} = \Pi_{h_j}^{(j)}$ with $i \neq j$ may appear.

$\lambda_{ij} = 1$ if $i > j$, $\lambda_{ii} = 1/2$, and $\lambda_{ij} = 0$ otherwise. Notice that in this case $B_\Upsilon = B_\Lambda^*$. Finally, consider that B is *partitioned into blocks* as follows: $B = [B_{lm}]$, $1 \leq l, m \leq p$, where B_{lm} is an $n_l \times n_m$ matrix, and $\sum_{l=1}^p n_l = n$. If we choose $\lambda_{ij} = 1$ in the positions corresponding to the entries of the blocks B_{lm} with $l > m$ and $\lambda_{ij} = 0$ otherwise, then B_Λ and B_Υ are, respectively, the *block* strictly lower and block upper triangular parts of B.

For any matrix B, the matrices B_Λ and B_Υ satisfy the properties (1) $B = B_\Lambda + B_\Upsilon$; (2) $|B_\Lambda| = |B|_\Lambda$ and $|B_\Upsilon| = |B|_\Upsilon$; (3) if $|A| \leq |B|$, then $|A|_\Lambda \leq |B|_\Lambda$ and $|A|_\Upsilon \leq |B|_\Upsilon$; (4) $|B_\Lambda| \leq |B|$ and $|B_\Upsilon| \leq |B|$. With all these we get the following.

THEOREM 3.6. *Let F be an $n \times n$ complex matrix and let us define recursively the following matrices:*

$$\mathcal{L}_1 = F_\Lambda,\ \mathcal{U}_1 = F_\Upsilon, \tag{3.8}$$

$$\mathcal{L}_k = (-\mathcal{L}_1\mathcal{U}_{k-1} - \mathcal{L}_2\mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1}\mathcal{U}_1)_\Lambda \quad \textit{for} \ \ k \geq 2, \tag{3.9}$$

$$\mathcal{U}_k = (-\mathcal{L}_1\mathcal{U}_{k-1} - \mathcal{L}_2\mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1}\mathcal{U}_1)_\Upsilon \quad \textit{for} \ \ k \geq 2. \tag{3.10}$$

Then

$$|\mathcal{L}_k + \mathcal{U}_k| \leq |F|^k \quad \textit{for} \ \ k \geq 1,$$

and, therefore, $|\mathcal{L}_k| \leq (|F|^k)_\Lambda$ and $|\mathcal{U}_k| \leq (|F|^k)_\Upsilon$.

The proof of Theorem 3.6 follows step by step the proof of Theorem 3.1, just by replacing $(\cdot)_L$ by $(\cdot)_\Lambda$ and $(\cdot)_U$ by $(\cdot)_\Upsilon$. With these replacements, an analogue of Theorem 2.2 also holds for the matrices defined by the recurrence relation (3.8)–(3.10).

4. Convergence properties for the series of the LU factors and strict perturbation bounds. In this section, the most relevant properties on the convergence of the series of the LU factors are shown. As an application, previously existing strict norm- and componentwise perturbation bounds are proven. The results appearing in this section are direct consequences of Theorem 3.1. We begin with a result for the LU factorization of a perturbation of the identity matrix. The first two items in this theorem also appear in Theorem 2.1. They are reproduced here for the sake of clarity and completeness. Additionally, we present here a new proof of item 2 using only matrix techniques and without any reference to Taylor series. We think that it is interesting from a mathematical point of view.

THEOREM 4.1. *Let F be an $n \times n$ complex matrix, such that $\rho(|F|) < 1$, let $\mathcal{L}_k$ and $\mathcal{U}_k$, $k \geq 1$, be the matrices defined by the recurrence relation* (2.8)–(2.10)*, and let $\mathcal{L}_0 = \mathcal{U}_0 = I$. Then*

1. *$I + F$ has a unique LU factorization: $I + F = \mathcal{L}\mathcal{U}$;*
2.

$$\mathcal{L} = I + \sum_{k=1}^{\infty} \mathcal{L}_k \quad \textit{and} \quad \mathcal{U} = I + \sum_{k=1}^{\infty} \mathcal{U}_k, \tag{4.1}$$

and these series converge absolutely componentwise; i.e., the numerical series corresponding to each entry of $\mathcal{L}$ and $\mathcal{U}$ converges absolutely;

3.

$$\left| \mathcal{L} - \sum_{k=0}^{N} \mathcal{L}_k \right| \leq \left(|F|^{N+1} (I - |F|)^{-1} \right)_L, \tag{4.2}$$

$$\left| \mathcal{U} - \sum_{k=0}^{N} \mathcal{U}_k \right| \leq \left(|F|^{N+1} (I - |F|)^{-1} \right)_U; \tag{4.3}$$

4. *if moreover* $\|\cdot\|$ *is an absolute and consistent matrix norm and* $\|F\| < 1$*, then*

$$\max\left\{\left\|\mathcal{L}-\sum_{k=0}^{N}\mathcal{L}_k\right\|, \left\|\mathcal{U}-\sum_{k=0}^{N}\mathcal{U}_k\right\|\right\} \le \left\|\mathcal{L}+\mathcal{U}-\sum_{k=0}^{N}(\mathcal{L}_k+\mathcal{U}_k)\right\| \le \frac{\|F\|^{N+1}}{1-\|F\|}. \tag{4.4}$$

In [24, Theorem 2.1] Stewart presented a normwise bound for $\|\mathcal{L}+\mathcal{U}-I-\mathcal{L}_1-I-\mathcal{U}_1\|$, i.e., just in the case $N=1$. Theorem 4.1 improves the result in [24] in several ways: the remainders of *any order* of the series are also bounded *componentwise.* The normwise bound (4.4) holds for $\|F\| < 1$ while the bound in [24] holds only if $\|F\| \le 1/4$. Finally the bound (4.4) is smaller than the one in [24, Theorem 2.1].[2]

Proof of Theorem 4.1. 1. This is a consequence of $\rho(|F|) < 1$ and (2.4), which guarantee that all the leading principal submatrices of $I+F$ are nonsingular.

2. We consider the series of the entrywise absolute values; the convergence follows trivially from Theorem 3.1:

$$I+\sum_{k=1}^{\infty}|\mathcal{L}_k| \le I+\sum_{k=1}^{\infty}(|F|^k)_L = I+\left(\sum_{k=1}^{\infty}|F|^k\right)_L = I+\left(|F|(I-|F|)^{-1}\right)_L.$$

The last step is a consequence of [18, Lemma 5.6.10, Corollary 5.6.16]. A similar argument holds for the series of $\mathcal{U}$. We have proven that the series in (4.1) converge, but not that they converge to $\mathcal{L}$ and $\mathcal{U}$. To prove this, it suffices to prove that

$$\lim_{N\to\infty}\left(\sum_{k=0}^{N}\mathcal{L}_k\right)\left(\sum_{j=0}^{N}\mathcal{U}_j\right) = I+F, \tag{4.5}$$

and to use the fact that the LU factorization of $I+F$ is unique. To do this notice that

$$\left(\sum_{k=0}^{N}\mathcal{L}_k\right)\left(\sum_{j=0}^{N}\mathcal{U}_j\right) = \sum_{j=0}^{N}\sum_{k=0}^{j}\mathcal{L}_k\mathcal{U}_{j-k} + \sum_{j=N+1}^{2N}\sum_{\substack{k=0\\ k,(j-k)\le N}}^{j}\mathcal{L}_k\mathcal{U}_{j-k}. \tag{4.6}$$

By using (2.8)–(2.10), the first summand in the right-hand side is just

$$\sum_{j=0}^{N}\sum_{k=0}^{j}\mathcal{L}_k\mathcal{U}_{j-k} = I+F. \tag{4.7}$$

By using Lemma 3.2, (2.8)–(2.10), and Theorem 3.1, the second summand in the right-hand side of (4.6) can be bounded as follows:

$$\begin{aligned}\left|\sum_{j=N+1}^{2N}\sum_{\substack{k=0\\ k,(j-k)\le N}}^{j}\mathcal{L}_k\mathcal{U}_{j-k}\right| &= \left|\sum_{j=N+1}^{2N}\sum_{\substack{k=1\\ k,(j-k)\le N}}^{j-1}\mathcal{L}_k\mathcal{U}_{j-k}\right| \le \sum_{j=N+1}^{2N}\sum_{k=1}^{j-1}|\mathcal{L}_k||\mathcal{U}_{j-k}| \\ &\le \sum_{j=N+1}^{2N}(-1)^j\sum_{k=1}^{j-1}\overline{\mathcal{L}}_k\overline{\mathcal{U}}_{j-k} = \sum_{j=N+1}^{2N}(-1)^{j+1}(\overline{\mathcal{L}}_j+\overline{\mathcal{U}}_j) \\ &\le \sum_{j=N+1}^{2N}|F|^j \le |F|^{N+1}\sum_{j=0}^{\infty}|F|^j = |F|^{N+1}(I-|F|)^{-1}.\end{aligned}$$

[2] It should be remarked that there is a minor misprint in the statement of Theorem 2.1 in [24]: the bound appearing there has to be multiplied by 2, as follows from the proof of this theorem. This is necessary to see that the bound (4.4) for $N=1$ is always smaller than the bound in [24].

This means that the second summand in (4.6) goes to zero as N goes to infinity. This and (4.7) finally prove (4.5).

3. We prove only the result for $\mathcal{L}$, i.e., the bound (4.2). The proof for (4.3) is similar. Again, these results are trivial consequences of the bound in Theorem 3.1. Notice that

$$\left|\mathcal{L}-\sum_{k=0}^{N}\mathcal{L}_k\right| \leq \sum_{k=N+1}^{\infty}|\mathcal{L}_k| \leq \sum_{k=N+1}^{\infty}(|F|^k)_L = \left(\sum_{k=N+1}^{\infty}|F|^k\right)_L = \left(|F|^{N+1}\sum_{k=0}^{\infty}|F|^k\right)_L,$$

and by summing the series, (4.2) is obtained.

4. By using Theorem 3.1 and the fact that the norm $\|\cdot\|$ is absolute,

$$\left\|\mathcal{L}+\mathcal{U}-\sum_{k=0}^{N}(\mathcal{L}_k+\mathcal{U}_k)\right\| \leq \sum_{k=N+1}^{\infty}\|F\|^k,$$

and by summing the series, (4.4) is obtained. □

As a direct consequence of Theorem 4.1 we get the following theorem for the series and perturbation bounds of the LU factors of a general matrix A.

THEOREM 4.2. *Let the nonsingular $n \times n$ complex matrix A have the LU factorization $A = LU$. Let us consider the matrix $A + E$ and define $F = L^{-1}EU^{-1}$. If $\rho(|F|) < 1$, then*

1. *the matrix $A + E$ is nonsingular and has a unique LU factorization*

$$A+E = L\,(I+F)\,U = L\,\mathcal{L}\mathcal{U}\,U = \widetilde{L}\,\widetilde{U};$$

2. *if $\mathcal{L}_k$ and $\mathcal{U}_k$ are the terms of the series appearing in* (4.1)*, then*

$$\widetilde{L} = L\,\mathcal{L} = L\left(I+\sum_{k=1}^{\infty}\mathcal{L}_k\right) \quad \textit{and} \quad \widetilde{U} = \mathcal{U}\,U = \left(I+\sum_{k=1}^{\infty}\mathcal{U}_k\right)U; \tag{4.8}$$

3.

$$|\widetilde{L}-L| \leq |L|\,(|F|(I-|F|)^{-1})_L, \tag{4.9}$$

$$|\widetilde{U}-U| \leq (|F|(I-|F|)^{-1})_U|U|; \tag{4.10}$$

4. *if moreover $\|\cdot\|$ is an absolute and consistent matrix norm and $\|F\| < 1$, then*

$$\max\left\{\frac{\|\widetilde{L}-L\|}{\|L\|}, \frac{\|\widetilde{U}-U\|}{\|U\|}\right\} \leq \frac{\|F\|}{1-\|F\|}. \tag{4.11}$$

The componentwise bounds (4.9) and (4.10) are essentially the same as those presented in [28, Theorem 5.1], and the normwise bound (4.11) is of the same kind as the one presented in [3, Theorem 3.1]. However, the proofs appearing in [3, 28] are based on different approaches. For the sake of brevity, we do not present bounds on the remainders of any order for the series (4.8). They can be easily deduced from Theorem 4.1. The results in Theorem 4.2 directly imply weaker bounds, under more restrictive conditions, that may be more useful in practice; see [28, Theorem 5.2].

The normwise bounds (4.11) can systematically overestimate the actual errors [9]; however, some inner diagonal scalings can be introduced in the normwise perturbation theory of the LU factorization to improve the bounds and to estimate the true normwise condition numbers of the LU factors. We will further discuss these questions after Theorem 5.1.

Proof of Theorem 4.2. 1. The proof of this item follows from the first item in Theorem 4.1, and (2.2). See also the comments following (2.2).

2. This follows from (2.2) and the second item in Theorem 4.1. Obviously, if a convergent series of matrices is multiplied by a constant matrix, then we get a series which converges to the product of the former series times the constant matrix.

3. This is a consequence of (4.8), (4.2), and (4.3) for the case $N = 0$.

4. This follows from (4.8) and (4.4) in the case $N = 0$. □

This theorem can be used in practice in combination with the backward error analysis for the LU factorization. The classical backward error result (see, for instance, [17, Theorem 9.3]) states that the computed LU factors, $\widehat{L}$, $\widehat{U}$, are the exact LU factors of a matrix $A + E$, $A + E = \widehat{L}\widehat{U}$ with $|E| \leq (n\epsilon/(1 - n\epsilon))|\widehat{L}||\widehat{U}|$, where ϵ is the unit roundoff of the computer. Combining this backward error bound with the perturbation bounds appearing in (4.9)–(4.11) (interchanging the roles of A and $A + E$, and the roles of L, U and $\widehat{L}$, $\widehat{U}$), bounds on the forward error bounds of the computed LU factors can be obtained [7].

5. Perturbation theory for block LU factorization.

5. Perturbation theory for block LU factorization. In this section, we deal with the block LU factorization of a nonsingular square matrix A. The main result in this section is that the perturbation theory for the block LU factorization is completely analogous to the perturbation theory of the usual LU factorization, just by replacing pointwise by block lower and upper triangular matrices.

The stability properties and applications of block LU factorizations appear in [17, Chapter 13] and the references therein. A block LU factorization of an $n \times n$ complex matrix A is as follows:

$$\underbrace{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p}\\ A_{21} & A_{22} & \cdots & A_{2p}\\ \vdots & \vdots & \vdots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp}\end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} I & & & \\ L_{21} & I & & \\ \vdots & & \ddots & \\ L_{p1} & \cdots & L_{p,p-1} & I\end{bmatrix}}_{L} \underbrace{\begin{bmatrix} U_{11} & U_{12} & \cdots & U_{1p}\\ & U_{22} & & \vdots \\ & & \ddots & U_{p-1,p}\\ & & & U_{pp}\end{bmatrix}}_{U}, \tag{5.1}$$

where the matrices L_{ij}, U_{ij}, and A_{ij} have dimensions $n_i \times n_j$, and $\sum_{i=1}^{p} n_i = n$. These dimensions are a priori given numbers which define the exact meaning of the *block* LU factorization. It is assumed that these dimensions remain unchanged throughout this section. We have chosen, as usual, to keep the same notation for block and pointwise LU factorizations. Therefore, here L and U have a different meaning than that in the previous sections. The reader should be careful about this question.

The block LU factorization and block triangular matrices have analogous properties to those of the usual LU factorization and triangular matrices: if the p leading block principal submatrices of A are nonsingular, then A has a unique block LU factorization; if a block LU factorization of a nonsingular matrix exists, then it is unique (see [17, Theorem 13.2] for a more general result covering the case of singular matrices). Notice that according to the previous properties, it may happen that a matrix has a unique block LU factorization with blocks of given dimensions, but not a usual pointwise LU factorization. However, if a matrix has a unique pointwise LU factorization, then it also has a unique block LU factorization for any possible block dimensions.

The product of two upper (lower) block triangular matrices is an upper (lower) block triangular matrix with the same block dimensions. Finally, the inverse of an upper (lower) block triangular matrix is also an upper (lower) block triangular matrix with the same block dimensions. Moreover, the diagonal blocks of the inverse are the inverse matrices of the diagonal blocks of the matrix. All these properties can easily be proven.

Now, we can proceed with a similar argument to that used to study the series and perturbation theory of the LU factorization. Let $A = LU$ be the unique block LU factorization of the $n \times n$ nonsingular matrix A. Let us consider a perturbation $\widetilde{A} = A + E$ of A, such that $\rho(|L^{-1}EU^{-1}|) < 1$. Then

$$\widetilde{A} = L(I + L^{-1}EU^{-1})U = (L\mathcal{L})(\mathcal{U}U) \equiv \widetilde{L}\widetilde{U}, \tag{5.2}$$

where $(I + L^{-1}EU^{-1}) = \mathcal{L}\mathcal{U}$ is the unique block LU factorization of $(I + L^{-1}EU^{-1})$ with the same block dimensions as those of $A = LU$. Therefore, $\widetilde{A} = (L\mathcal{L})(\mathcal{U}U) \equiv \widetilde{L}\widetilde{U}$ is the unique block LU factorization of the nonsingular matrix $\widetilde{A}$ with the same block dimensions as those of $A = LU$. The condition $\rho(|L^{-1}EU^{-1}|) < 1$, along with (2.4), guarantees that all these block factorizations exist and are unique. These results are formally the same as those appearing in (2.2)–(2.4), but here $L, U, \mathcal{L}$, and $\mathcal{U}$ are block triangular matrices; thus the results and the condition $\rho(|L^{-1}EU^{-1}|) < 1$ depend on the block structure we are considering.

Given a block partitioned matrix A, such as the one appearing in (5.1), let us define its *block* strictly lower triangular and *block* upper triangular parts, A_L and A_U, respectively, in the obvious way:

(5.3)

$$\underbrace{\begin{bmatrix} A_{11} & \cdots & A_{1p} \\ A_{21} & \cdots & A_{2p} \\ \vdots & \vdots & \vdots \\ A_{p1} & \cdots & A_{pp} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & & & \\ A_{21} & 0 & & \\ \vdots & & \ddots & \\ A_{p1} & \cdots & A_{p,p-1} & 0 \end{bmatrix}}_{A_L} + \underbrace{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ & A_{22} & & \vdots \\ & & \ddots & A_{p-1,p} \\ & & & A_{pp} \end{bmatrix}}_{A_U},$$

where for the sake of simplicity the same notation as in the case of pointwise strictly lower and upper triangular parts is used. Let us remember that for block strictly lower and block upper triangular parts a block analogue of Theorem 3.1 holds, as we have discussed in the general Theorem 3.6. With this, a block version of Theorem 4.1 can be proved in an analogous way. To keep the paper concise, we will not state this block theorem. This theorem allows us to prove the main result of this section.

THEOREM 5.1. *Let the nonsingular $n \times n$ complex matrix A have the block LU factorization $A = LU$ appearing in* (5.1). *For any partitioned matrix B, let us denote by B_L and B_U, respectively, its block strictly lower triangular and block upper triangular parts, as defined in* (5.3). *Let us consider the matrix $A + E$ and define $F = L^{-1}EU^{-1}$. If $\rho(|F|) < 1$, then*

1. *the matrix $A + E$ is nonsingular and has a unique block LU factorization, $A + E = \widetilde{L}\widetilde{U}$, with the same block dimensions as those in $A = LU$;*
2. *let $\mathcal{L}_k$ and $\mathcal{U}_k$ be the matrices defined recursively by*

$$\begin{aligned} &\mathcal{L}_1 = F_L,\ \mathcal{U}_1 = F_U, \\ &\mathcal{L}_k = (-\mathcal{L}_1\mathcal{U}_{k-1} - \mathcal{L}_2\mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1}\mathcal{U}_1)_L \quad \text{for } k \geq 2, \\ &\mathcal{U}_k = (-\mathcal{L}_1\mathcal{U}_{k-1} - \mathcal{L}_2\mathcal{U}_{k-2} - \cdots - \mathcal{L}_{k-1}\mathcal{U}_1)_U \quad \text{for } k \geq 2; \end{aligned}$$

then with $\mathcal{L}$ *and* $\mathcal{U}$ *defined below, giving* $\mathcal{L}\mathcal{U} = I + F$,

$$\widetilde{L} = L\,\mathcal{L} = L\left(I + \sum_{k=1}^{\infty} \mathcal{L}_k\right) \quad \textit{and} \quad \widetilde{U} = \mathcal{U}\,U = \left(I + \sum_{k=1}^{\infty} \mathcal{U}_k\right) U; \tag{5.4}$$

3.

$$\left|\widetilde{L} - L\right| \leq |L| \left(|F|\,(I - |F|)^{-1} \right)_L, \tag{5.5}$$

$$\left|\widetilde{U} - U\right| \leq \left(|F|\,(I - |F|)^{-1} \right)_U |U|; \tag{5.6}$$

4. *if moreover* $\|\cdot\|$ *is an absolute and consistent matrix norm and* $\|F\| < 1$, *then*

$$\max\left\{ \frac{\|\widetilde{L} - L\|}{\|L\|}, \frac{\|\widetilde{U} - U\|}{\|U\|} \right\} \leq \frac{\|F\|}{1 - \|F\|}. \tag{5.7}$$

This theorem and its proof are completely analogous to the proof of Theorem 4.2.

The perturbation theorem above has to be combined with backward error results to estimate the forward errors produced by the algorithms that compute the block LU factorization. If this factorization is computed using conventional matrix operations, then it is possible to prove component- and normwise backward error bounds. However, in the case of block algorithms, fast matrix multiplication techniques are frequently used to improve the performance, and then the backward error bounds are necessarily normwise [11] (see also [17, Theorem 13.6]). In this case, the bound (5.7) can systematically overestimate the actual errors, but, at the cost of discarding second-order terms, this bound can be improved. From (5.4), using again $F = L^{-1}EU^{-1}$,

$$\widetilde{L} - L = L(L^{-1}EU^{-1})_L + \|L\|\,O(\|F\|^2) = (LS)((LS)^{-1}EU^{-1})_L + \|L\|\,O(\|F\|^2),$$

where S is any block diagonal matrix whose blocks have the same dimensions as the diagonal blocks of A. Let us denote by $\mathcal{S}$ the set of block diagonal matrices with blocks of these dimensions and by $\kappa(B) = \|B\|\|B^{-1}\|$ the condition number, in the norm we are considering, of the matrix B. Then, the previous equation implies

$$\frac{\|\widetilde{L} - L\|}{\|L\|} \leq \min_{S\in\mathcal{S}} \kappa(LS) \frac{\|U^{-1}\|\|E\|}{\|L\|} + O(\|F\|^2) \leq \left(\min_{S\in\mathcal{S}} \kappa(LS) \right) \kappa(U) \frac{\|E\|}{\|A\|} + O(\|F\|^2).$$

A similar argument for U implies that

$$\frac{\|\widetilde{U} - U\|}{\|U\|} \leq \min_{S\in\mathcal{S}} \kappa(SU) \frac{\|L^{-1}\|\|E\|}{\|U\|} + O(\|F\|^2) \leq \left(\min_{S\in\mathcal{S}} \kappa(SU) \right) \kappa(L) \frac{\|E\|}{\|A\|} + O(\|F\|^2).$$

For the usual LU factorization the set $\mathcal{S}$ is just the set of diagonal matrices, and $\min_{S\in\mathcal{S}} \kappa(LS)$ (resp., $\min_{S\in\mathcal{S}} \kappa(SU)$) can be reliably estimated, for p norms, by choosing S in a way such that all the columns (resp., rows) of LS (resp., SU) have unit norms [29]. Moreover, $\min_{S\in\mathcal{S}} \kappa(LS)\kappa(U)$ and $\min_{S\in\mathcal{S}} \kappa(SU)\kappa(L)$ are, respectively, good approximations of the true normwise condition numbers of the L and U factors [9, 25]. The true normwise condition numbers were obtained in [9], but it is difficult to estimate them efficiently. In the case of block LU factorization, the set $\mathcal{S}$ contains all the diagonal matrices; then $\min_{S\in\mathcal{S}} \kappa(LS)$ and $\min_{S\in\mathcal{S}} \kappa(SU)$ are smaller than for the usual pointwise LU factorization. But unfortunately there is no easy way to estimate these optimally block scaled condition numbers [21].

6. Perturbation theory for block LDL* factorization of general Hermitian matrices.

The block LU factorization plays a relevant role in the solution of linear systems of equations with a Hermitian indefinite coefficient matrix B [17, section 11.1]. In this case, the most widely used approach is to perform a block factorization as follows:

$$PBP^T = LDL^*,$$

where P is a permutation matrix, L is unit lower triangular, and D is block diagonal with diagonal blocks of dimension 1×1 or 2×2. The 2×2 diagonal blocks of D are Hermitian indefinite matrices. Moreover, the diagonal blocks of L corresponding to the 2×2 blocks of D are 2×2 identity matrices. This factorization method is usually called the *diagonal pivoting method*. Three well-known pivoting strategies to choose the permutation matrix P are available: complete pivoting [6], partial pivoting [5], and rook pivoting [2]. The partial pivoting strategy is available in LAPACK [1], and it is used in the LAPACK symmetric indefinite linear equation solver.

The block LDL* factorization of a nonsingular Hermitian matrix is a special case of the block LU factorization, as the following lemma shows.[3]

LEMMA 6.1. *Let A be a nonsingular Hermitian matrix having a block LU factorization given by* (5.1)*, $A = LU$; then the block diagonal matrix $\Delta = \operatorname{diag}(U_{11}, U_{22}, \ldots, U_{pp})$ is Hermitian and $L^* = \Delta^{-1}U$. Therefore, $A = L\Delta L^*$ and this factorization is unique once the block dimensions are fixed.*

Proof. Let us write $A = LU = L\Delta(\Delta^{-1}U)$ and—using the fact that $A = A^*$—$A = (\Delta^{-1}U)^*(L\Delta)^*$. The block LU factorization of a nonsingular matrix is unique; therefore, $L = (\Delta^{-1}U)^*$ and $A = L\Delta L^*$, which implies that Δ is Hermitian. The uniqueness of $A = L\Delta L^*$ follows from the uniqueness of the block LU factorization. □

Therefore, the conditions for the existence of an LDL* factorization of a nonsingular Hermitian matrix are again that all the corresponding leading principal block submatrices of A are nonsingular. Moreover, to guarantee that the 2×2 diagonal blocks of D are indefinite matrices, the inertias of the leading principal block submatrices of A have to follow a special pattern with respect to the 2×2 blocks. This is the case whenever $A = PBP^T$, and P is obtained by applying some of the previously mentioned pivoting strategies.

The following shorthand notation will be frequently used: let $B = [B_{ij}]$, $1 \leq i, j \leq p$, be a block partitioned matrix, where B_{ij} is an $n_i \times n_j$ matrix. B_D will denote the block diagonal part of B, i.e., $B_D = \operatorname{diag}(B_{11}, \ldots, B_{pp})$.

Since the LDL* factorization is a block LU factorization,[4] its perturbation theory can be obtained from Theorem 5.1 by using that in this case $U = DL^*$.

THEOREM 6.2. *Let the nonsingular $n \times n$ Hermitian matrix A have the block LDL* factorization $A = LDL^*$. Let us denote for any partitioned matrix B by B_L and B_D its block strictly lower and block diagonal parts, respectively. Let us consider the Hermitian matrix $A + E$ and define $F = L^{-1}EL^{-*}D^{-1}$. If $\rho(|F|) < 1$, then the following results hold.*

[3] In Lemma 6.1, the letter Δ is used instead of D because the dimensions of the blocks in Δ are not necessarily 1×1 or 2×2.

[4] In this section, we restrict ourselves to LDL* factorizations with blocks of dimensions 1×1 or 2×2, because this is the most useful situation in numerical analysis. However, the results we present can be extended without effort to general block dimensions.

1. *The matrix $A+E$ is nonsingular and has a unique block LDL^* factorization, $A+E=\widetilde{L}\widetilde{D}\widetilde{L}^*$, with the same block dimensions as those in $A=LDL^*$. Moreover, let us denote the block diagonal matrices D and $\widetilde{D}$ as follows: $D=\operatorname{diag}(D_{11},\ldots,D_{pp})$ and $\widetilde{D}=\operatorname{diag}(\widetilde{D}_{11},\ldots,\widetilde{D}_{pp})$. Then, if D_{ii} and $\widetilde{D}_{ii}$ are 1×1 blocks, both have the same sign, and if D_{ii} and $\widetilde{D}_{ii}$ are 2×2 blocks, both are Hermitian indefinite matrices.*

2.

$$|\widetilde{L}-L| \leq |L|\,(|F|\,(I-|F|)^{-1})_L, \tag{6.1}$$

$$|\widetilde{D}-D| \leq (|F|\,(I-|F|)^{-1})_D\,|D|. \tag{6.2}$$

3. *If moreover $\|\cdot\|$ is an absolute and consistent matrix norm and $\|F\|<1$, then*

$$\max\left\{\frac{\|\widetilde{L}-L\|}{\|L\|},\frac{\|\widetilde{D}-D\|}{\|D\|}\right\} \leq \frac{\|F\|}{1-\|F\|}. \tag{6.3}$$

Proof. 1. We only have to prove the properties of the diagonal blocks of D and $\widetilde{D}$; the rest of this item follows from the first item of Theorem 5.1. We will use a continuity argument. Let us consider the one-parametric family of matrices $A(t)=A+tE$, where $0\leq t\leq 1$. Notice that $\rho(|L^{-1}(tE)L^{-*}D^{-1}|)=t\rho(|L^{-1}EL^{-*}D^{-1}|)<1$, and therefore, $A(t)$ is nonsingular and has a unique LDL* factorization for each t with the same block dimensions as those in $A=LDL^*$. Let us denote this factorization $A(t)=L(t)D(t)L(t)^*$ for $0\leq t\leq 1$. Clearly $L(t)$ and $D(t)$ are continuous functions of $t\in[0,1]$. This follows, for instance, from (5.5) and (5.6), taking into account that D is just the block diagonal part of U. On the other hand, $\det A(t)=\det D_{11}(t)\cdots\det D_{pp}(t)\neq 0$ for all $t\in[0,1]$. Therefore, if $D_{ii}(t)$ is a 1×1 block, then its sign remains constant for $t\in[0,1]$, and if $D_{ii}(t)$ is a 2×2 block, then the sign of $\det D_{ii}(t)$ also remains constant for $t\in[0,1]$. This implies that if the two eigenvalues of $D_{ii}=D_{ii}(0)$ have opposite signs, the two eigenvalues of $\widetilde{D}_{ii}=D_{ii}(1)$ also have opposite signs.

2. The bound (6.1) is just (5.5), and the bound (6.2) follows from considering the block diagonal part of (5.6).

3. The bound for L is the same as in (5.7). The bound for D follows directly from the second equation in (5.4) and, from Theorem 3.6, that $|(\mathcal{U}_k)_D|\leq|\mathcal{U}_k|\leq|F|^k$. □

The normwise bound (6.3) can be improved using inner block diagonal scalings in the same way as was done for the block LU factorization after Theorem 5.1.

Notice that the componentwise bound (6.2) is not a symmetric matrix. This is a drawback of the previous theorem because $|\widetilde{D}-D|$ is symmetric. The following theorem partially solves this drawback by splitting the bound on $|\widetilde{D}-D|$ into two terms: a symmetric first-order term, by far the most relevant in applications, and a second-order term that is not symmetric.

THEOREM 6.3. *Using the same assumptions and notation as those of Theorem* 6.2, *we obtain*

$$|\widetilde{D}-D| \leq |L^{-1}EL^{-*}|_D + (|F|^2(I-|F|)^{-1})_D\,|D|.$$

Proof. From (5.4), $\widetilde{U}-U=\mathcal{U}_1U+\sum_{k=2}^{\infty}\mathcal{U}_kU$. Its block diagonal part is

$$\widetilde{D}-D=F_DD+\sum_{k=2}^{\infty}(\mathcal{U}_k)_DD, \tag{6.4}$$

where we have used the fact that $\mathcal{U}_1 = F_U$, $U_D = D$, and $(\mathcal{U}_k U)_D = (\mathcal{U}_k)_D D$, because $\mathcal{U}_k$ and U are both upper block triangular matrices. Notice that $F_D D = (FD)_D = (L^{-1}EL^{-*}D^{-1}D)_D = (L^{-1}EL^{-*})_D$. Therefore, (6.4) and Theorem 3.6 imply

$$\begin{aligned}|\widetilde{D} - D| &\leq |L^{-1}EL^{-*}|_D + \sum_{k=2}^{\infty} |\mathcal{U}_k|_D |D| \leq |L^{-1}EL^{-*}|_D + \sum_{k=2}^{\infty} (|F|^k)_D |D| \\ &= |L^{-1}EL^{-*}|_D + \left(\sum_{k=2}^{\infty} |F|^k \right)_D |D| = |L^{-1}EL^{-*}|_D + (|F|^2 (I - |F|)^{-1})_D |D|. \qquad \square\end{aligned}$$

Theorems 6.2 and 6.3 can be used in connection with the backward error results for the usual algorithm to compute the LDL^T factorization [16] (see also [17, Theorem 11.3]) to get bounds for the forward errors of the computed L and D. In this context, it is convenient to interchange the roles of A and $A + E$ in the bounds, because the computed LDL* factorization is the exact factorization of $A + E$.

In some cases, the D factor of a block LDL* factorization is just a diagonal matrix, i.e., when the chosen pivoting strategy does not find any 2×2 pivot. One trivial instance of this situation is when the Hermitian matrix A is positive definite, but it may also occur for indefinite matrices. The componentwise bound for $|\widetilde{D} - D|$ appearing in Theorem 6.2 can be simplified. If D is diagonal, then $|FD| = |F|\,|D|$ and, therefore,

$$|\widetilde{D} - D| \leq \operatorname{diag}((I - |F|)^{-1} |L^{-1}EL^{-*}|), \tag{6.5}$$

where diag(B) is the diagonal matrix whose entries are those of the diagonal of B.

Although Theorem 6.2 is the first perturbation result for the *block LDL** factorization, the case when D is diagonal has been studied by other authors. In [28, Theorem 3.1] componentwise perturbation bounds for the LDL* factorization of positive definite Hermitian matrices are presented. Theorem 6.2, along with (6.5), represents an improvement over Theorem 3.1 in [28], because Theorem 6.2 is valid also for nonpositive diagonal matrices D, it requires less restrictive assumptions concerning the magnitude of the perturbation E, and, moreover, the bounds for L are smaller and simpler. In [22, Theorem 4.1, p. 269] some of the results in [27] for positive definite matrices have been extended to indefinite matrices with factorization LDL* with D still diagonal, i.e., without 2×2 blocks.

7. Perturbation bounds for the Cholesky factorization.

In this section, we present a series expansion and component- and normwise perturbation bounds for the Cholesky factorization of a Hermitian positive definite matrix. The perturbation bounds we present are essentially the same as the bounds in [26, 28] with some improvements. The main goal of this section is to see how the series expansions we have developed allow us to get perturbation bounds also for the Cholesky factorization.

Let A be a Hermitian positive definite matrix with the Cholesky factorization $A = R^*R$, and let E be Hermitian such that $\rho(|R^{-*}ER^{-1}|) < 1$. Then

$$\widetilde{A} = A + E = R^*(I + R^{-*}ER^{-1})R$$

is positive definite, because $(I + R^{-*}ER^{-1})$ is positive definite. In addition, if $(I + R^{-*}ER^{-1}) = \mathcal{R}^*\mathcal{R}$ is the corresponding Cholesky factorization, then

$$\widetilde{A} = (R^*\mathcal{R}^*)(\mathcal{R}R) \equiv \widetilde{R}^*\widetilde{R}$$

is the Cholesky factorization of $\widetilde{A}$. Thus, let us focus on the Cholesky factorization of $(I + R^{-*}ER^{-1}) \equiv I + F$.

Given a Hermitian $n \times n$ matrix $B = [b_{ij}]$, we define B_{Λ_H} and $B_{\Upsilon_H} = B^*_{\Lambda_H}$:

$$(B_{\Lambda_H})_{ij} = \begin{cases} b_{ij} & \text{if } i > j, \\ b_{ii}/2 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad (B_{\Upsilon_H})_{ij} = \begin{cases} b_{ij} & \text{if } i < j, \\ b_{ii}/2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

If Theorem 3.6 is applied to a Hermitian matrix F, for instance, $F = R^{-*}ER^{-1}$, with $(\cdot)_\Lambda = (\cdot)_{\Lambda_H}$ and $(\cdot)_\Upsilon = (\cdot)_{\Upsilon_H}$, then a simple argument shows that $\mathcal{L}_k + \mathcal{U}_k$ is Hermitian and that $\mathcal{L}_k = \mathcal{U}_k^*$. Using these facts and introducing the more convenient notation $\mathcal{R}_k$ instead of $\mathcal{U}_k$, we get the following Hermitian version of Theorem 3.6.

Theorem 7.1. *Let F be an $n \times n$ Hermitian matrix and let us define recursively the following upper triangular matrices:*

$$\mathcal{R}_1 = F_{\Upsilon_H}; \quad \mathcal{R}_k = (-\mathcal{R}_1^*\mathcal{R}_{k-1} - \mathcal{R}_2^*\mathcal{R}_{k-2} - \cdots - \mathcal{R}_{k-1}^*\mathcal{R}_1)_{\Upsilon_H} \quad \textit{for } k \geq 2. \tag{7.2}$$

Then $|\mathcal{R}_k^* + \mathcal{R}_k| \leq |F|^k$ for $k \geq 1$, and therefore, $|\mathcal{R}_k| \leq (|F|^k)_{\Upsilon_H}$.

The result above allows us to prove the next theorem very easily.

Theorem 7.2. *Let the Hermitian positive definite matrix A have the Cholesky factorization $A = R^*R$. Let us consider the Hermitian matrix $A + E$ and define $F = R^{-*}ER^{-1}$. If $\rho(|F|) < 1$, then*

1. *the matrix $A + E$ is positive definite and has a unique Cholesky factorization $A + E = \widetilde{R}^*\widetilde{R}$;*
2. *if $\mathcal{R}_k$ are the matrices defined in* (7.2)*, then*

$$\widetilde{R} = \left(I + \sum_{k=1}^{\infty} \mathcal{R}_k \right) R; \tag{7.3}$$

3.

$$|\widetilde{R} - R| \leq (|F|(I - |F|)^{-1})_{\Upsilon_H} |R|; \tag{7.4}$$

4. *let us denote by $\|\cdot\|_F$ and by $\|\cdot\|_2$ the Frobenius and the spectral norms, respectively; if $\|F\|_F < 1$, then*

$$\frac{\|\widetilde{R} - R\|_F}{\|R\|_2} \leq \frac{1}{\sqrt{2}} \frac{\|F\|_F}{1 - \|F\|_F}; \tag{7.5}$$

if, moreover, $\|A^{-1}\|_2 \|E\|_F < 1$ and $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$, then

$$\frac{\|\widetilde{R} - R\|_F}{\|R\|_2} \leq \frac{1}{\sqrt{2}} \frac{\kappa_2(A)\frac{\|E\|_F}{\|A\|_2}}{1 - \kappa_2(A)\frac{\|E\|_F}{\|A\|_2}}. \tag{7.6}$$

The componentwise bound (7.4) is slightly better than the one in [28, Theorem 2.1], because there the upper triangular part $(|F|\,(I - |F|)^{-1})_U$ appears, which is greater than $(|F|\,(I - |F|)^{-1})_{\Upsilon_H}$. However, the normwise bound in [26, Theorem 1.4] is slightly better than (7.6), because the bound in Theorem 7.1 involves the entrywise absolute value, and the spectral norm is not absolute. The bound (7.5) is new. The normwise bound (7.5) can be improved using inner diagonal scalings as was done for the block LU factorization after Theorem 5.1. Bounds of the type (7.5) involving only the spectral norm can be found in [13, 14] for small perturbations.

Proof of Theorem 7.2. With the exception of the last item, the proof is similar to the proof of Theorem 4.2. From (7.3), (7.5) is easily proved by realizing that for any Hermitian matrix B, $\|B_{\Upsilon_H}\|_F \leq \frac{1}{\sqrt{2}}\|B\|_F$. □

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 513–561.

[3] A. Barrlund, *Perturbation bounds for the LDL^H and LU decompositions*, BIT, 31 (1991), pp. 358–363.

[4] M. I. Bueno and F. M. Dopico, *Stability and sensitivity of tridiagonal LU factorization without pivoting*, BIT, 44 (2004), pp. 651–673.

[5] J. R. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 163–179.

[6] J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

[7] X.-W. Chang, *Some features of Gaussian elimination with rook pivoting*, BIT, 42 (2002), pp. 66–83.

[8] X.-W. Chang, C. C. Paige, and G. W. Stewart, *New perturbation analyses for the Cholesky factorization*, IMA J. Numer. Anal., 16 (1996), pp. 457–484.

[9] X.-W. Chang and C. C. Paige, *On the sensitivity of the LU factorization*, BIT, 38 (1998), pp. 486–501.

[10] X.-W. Chang and C. C. Paige, *Sensitivity analyses for factorizations of sparse or structured matrices*, Linear Algebra Appl., 284 (1998), pp. 53–71.

[11] J. W. Demmel, N. J. Higham, and R. S. Schreiber, *Stability of block LU factorization*, Numer. Linear Algebra Appl., 2 (1995), pp. 173–190.

[12] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[13] Z. Drmač, M. Omladič, and K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

[14] A. Edelman and W. F. Mascarenhas, *On Parlett's matrix norm inequality for the Cholesky decomposition*, Numer. Linear Algebra Appl., 2 (1995), pp. 243–250.

[15] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[16] N. J. Higham, *Stability of the diagonal pivoting method with partial pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 52–65.

[17] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.

[19] B. N. Parlett, *The new qd algorithms*, in Acta Numerica 1995, Acta. Numer., Cambridge University Press, Cambridge, UK, 1995, pp. 459–491.

[20] K. H. Rosen, *Discrete Mathematics and Its Applications*, 4th ed., McGraw–Hill, Boston, 1999.

[21] A. Shapiro, *Optimal block diagonal l_2-scaling of matrices*, SIAM J. Numer. Anal., 22 (1985), pp. 81–94.

[22] I. Slapničar, *Componentwise analysis of direct factorization of real symmetric and Hermitian matrices*, Linear Algebra Appl., 272 (1998), pp. 227–275.

[23] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[24] G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1141–1145.

[25] G. W. Stewart, *On the perturbation of LU and Cholesky factors*, IMA J. Numer. Anal., 17 (1997), pp. 1–6.

[26] J.-G. Sun, *Perturbation bounds for the Cholesky and QR factorizations*, BIT, 31 (1991), pp. 341–352.

[27] J.-G. Sun, *Rounding-error and perturbation bounds for the Cholesky and LDL^T factorizations*, Linear Algebra Appl., 173 (1992), pp. 77–97.

[28] J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.

[29] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.