University of Maryland College Park

Institute for Advanced Computer Studies UMIACS-TR-94-105
Department of Computer Science CS-TR-3345

Rounding Errors in Solving Block Hessenberg Systems

Urs von Matt* and G. W. Stewart†

September, 1994

Abstract. A rounding error analysis is presented for a divide-and-conquer algorithm to solve linear systems with block Hessenberg matrices. Conditions are derived under which the algorithm computes a backward stable solution. The algorithm is shown to be stable for diagonally dominant matrices and for M-matrices.

Key words. Rounding error analysis, linear systems, block Hessenberg matrices, diagonally dominant matrices, M-matrices.

*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742; e-mail: `na.vonmatt@na-net.ornl.gov`.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742; e-mail: `stewart@cs.umd.edu`.

This work was supported in part by the National Science Foundation under grant CCR 9115568.

This report is available by anonymous ftp from `cs.umd.edu` in the directory `/pub/papers/TRs`.

ROUNDING ERRORS IN SOLVING BLOCK HESSENBERG SYSTEMS

URS VON MATT AND G. W. STEWART

September, 1994

ABSTRACT. A rounding error analysis is presented for a divide-and-conquer algorithm to solve linear systems with block Hessenberg matrices. Conditions are derived under which the algorithm computes a backward stable solution. The algorithm is shown to be stable for diagonally dominant matrices and for M-matrices.

1. INTRODUCTION

In [7] an algorithm was proposed for the solution of the linear system

$$AX = B, \tag{1}$$

where A is a block Hessenberg matrix. Its development was motivated by the attempt to find the steady-state of certain Markov chains. In this paper we will present an error analysis to explain the accurate results obtained by the algorithm.

Our analysis is a backward error analysis in the style of Wilkinson [12, 13]. This means that the computed matrix X can be regarded as the exact solution of a nearby linear system. In particular we will show that the computed X satisfies

$$AX = B + \Delta B.$$

We call the matrix X a backward stable solution if

$$\|\Delta B\| \leq \eta \|A\| \, \|X\|,$$

where η denotes a small multiple of the unit roundoff ε.

A backward stable solution is not to be confused with an accurate solution. The accuracy of X is usually limited by the condition number $\kappa(A) := \|A\| \, \|A^{-1}\|$. The

1991 *Mathematics Subject Classification.* Primary 65G05; Secondary 65F05.

Key words and phrases. Rounding error analysis, linear systems, block Hessenberg matrices, diagonally dominant matrices, M-matrices.

This work was supported in part by the National Science Foundation under grant CCR 9115568.

relative error of X can be bounded by

$$\frac{\|X - A^{-1}B\|}{\|A^{-1}B\|} \leq \frac{\eta\kappa(A)}{1 - \eta\kappa(A)},$$

provided that $\eta\kappa(A) < 1$. Thus we can only compute an accurate solution X if we use a stable algorithm to solve a well-conditioned problem.

Our paper is organized as follows. In Section 2 we give a concise description of the algorithm to be analysed. This algorithm consists of a few basic building blocks for which we will cite error bounds in Section 3. Since our algorithm calls itself recursively we have to make an assumption about the structure of the errors after each invocation. This is the purpose of Section 4, where we also analyse the local errors in each stage. We combine these local errors to give a global error bound in Section 5. The structure of this global error bound reveals a potential instability of our algorithm. This is discussed in Section 6. In Sections 7 and 8 we identify two classes of matrices for which our algorithm computes a backward stable solution. We conclude our presentation with some numerical examples in Section 9.

Throughout our analysis we will use the 2-norm, except where otherwise noted. Its main advantage is that the norm of an orthogonal matrix is one.

2. Algorithm

We assume that the matrix A in (1) has the following block Hessenberg structure:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & \cdots & A_{1n} \\ A_{21} & \ddots & & & \vdots \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & A_{n-1,n} \\ & & & A_{n,n-1} & A_{nn} \end{bmatrix}.$$

The diagonal blocks A_{ii} are assumed to be square nonsingular matrices of order p_i. The total size of A is given by

$$N := \sum_{i=1}^{n} p_i.$$

If $n > 1$ we can select a tear index k with $1 \leq k < n$ and partition the matrix A as follows:

$$A = \begin{bmatrix} A_{nw} & A_{ne} \\ A_{sw} & A_{se} \end{bmatrix}.$$

The submatrix A_{nw} contains the first k diagonal blocks of A, and A_{se} contains the last $n - k$ diagonal blocks of A. Note that $A_{k+1,k}$ is the only nonzero block in A_{sw}.

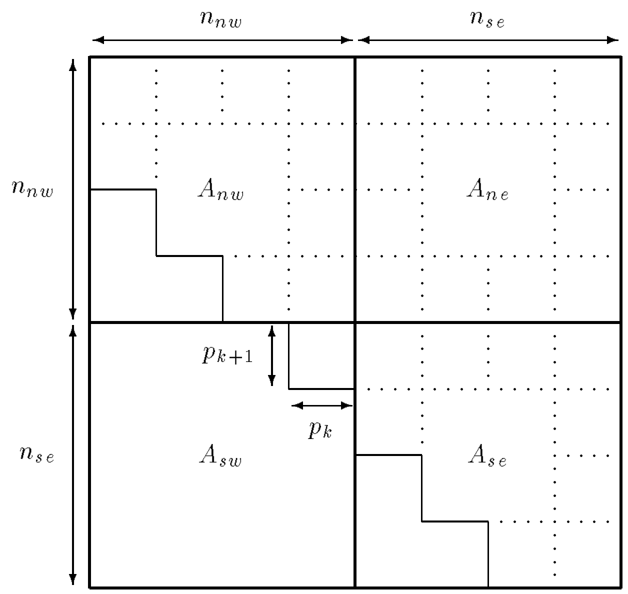

FIGURE 1. Structure of A.

This partitioning is also shown as Figure 1. The dimensions n_{nw} and n_{se} are given by

$$n_{nw} = \sum_{i=1}^{k} p_i,$$

$$n_{se} = \sum_{i=k+1}^{n} p_i.$$

Let E be the columns of the identity matrix that correspond to the rows of A spanned by $A_{k+1,k}$, and let F consist of the columns of the identity matrix corresponding to the columns of A spanned by $A_{k+1,k}$. Then we can also define

$$\hat{A} := \begin{bmatrix} A_{nw} & A_{ne} \\ 0 & A_{se} \end{bmatrix} = A - EA_{sw}F^{\mathrm{T}}.$$

In order to solve the linear system (1) for X we can compute the solution Y of

$$\hat{A}Y = B$$

and get X from Y by means of an updating formula. The well-known Sherman-Morrison-Woodbury formula (cf. [4, p. 51]) would give us

$$\begin{aligned} A^{-1} = (\hat{A} + EA_{sw}F^{\mathrm{T}})^{-1} &= \hat{A}^{-1} - \hat{A}^{-1}E(I + A_{sw}F^{\mathrm{T}}\hat{A}^{-1}E)^{-1}A_{sw}F^{\mathrm{T}}\hat{A}^{-1} \\ &= \hat{A}^{-1} - \hat{A}^{-1}EA_{sw}(I + F^{\mathrm{T}}\hat{A}^{-1}EA_{sw})^{-1}F^{\mathrm{T}}\hat{A}^{-1}. \end{aligned}$$

ALGORITHM 1. Solution of Block Hessenberg Systems.

function $X = solve(A, B)$
if not at the bottom **then**
  Compute the orthogonal URV-decomposition $A_{sw} = URV^{\mathrm{T}}$.
  $G_s := solve(A_{se}, U)$
  $G_n := solve(A_{nw}, -A_{ne}G_s)$
  $S := RV^{\mathrm{T}}G_n$
  $T := I + S$
  Solve $T\hat{R} = R$ for $\hat{R}$ by Gaussian elimination.
  $P := G\hat{R}$
  $Y_s := solve(A_{se}, B_s)$
  $Y_n := solve(A_{nw}, B_n - A_{ne}Y_s)$
  $X := Y - PV^{\mathrm{T}}Y_n$
else
  Solve $AX = B$ for X by Gaussian elimination.
end

Unfortunately, this formula does not take advantage of the many zeros in A_{sw}, requiring the solution of a large intermediate linear system. We can reduce the size of this linear system with the help of the URV-decomposition (cf. [8])

$$A_{sw} = URV^{\mathrm{T}}.$$

Let r denote the rank of A_{sw} as it is determined by the URV-decomposition. Then U will be an orthogonal n_{se}-by-r matrix with p_{k+1} nonzero rows. Also R is a square r-by-r matrix, and V is an orthogonal n_{nw}-by-r matrix with p_k nonzero rows. Now we can express the inverse of A by

$$A^{-1} = \hat{A}^{-1} - \hat{A}^{-1}EU(I + RV^{\mathrm{T}}F^{\mathrm{T}}\hat{A}^{-1}EU)^{-1}RV^{\mathrm{T}}F^{\mathrm{T}}\hat{A}^{-1}.$$

In order to avoid the multiple evaluation of the same expressions we introduce the following intermediate quantities:

$$\begin{aligned} G &:= \hat{A}^{-1}EU, \\ S &:= RV^{\mathrm{T}}F^{\mathrm{T}}G, \\ T &:= I + S, \\ \hat{R} &:= T^{-1}R, \\ P &:= G\hat{R}. \end{aligned}$$

Note that these matrices are independent of the right hand side B. The overall procedure to solve the linear system (1) is also presented as Algorithm 1.

In [7] this algorithm is refined further by introducing the auxiliary procedures “patchgen” and “topsolve”. These refinements are critical for the efficiency of the

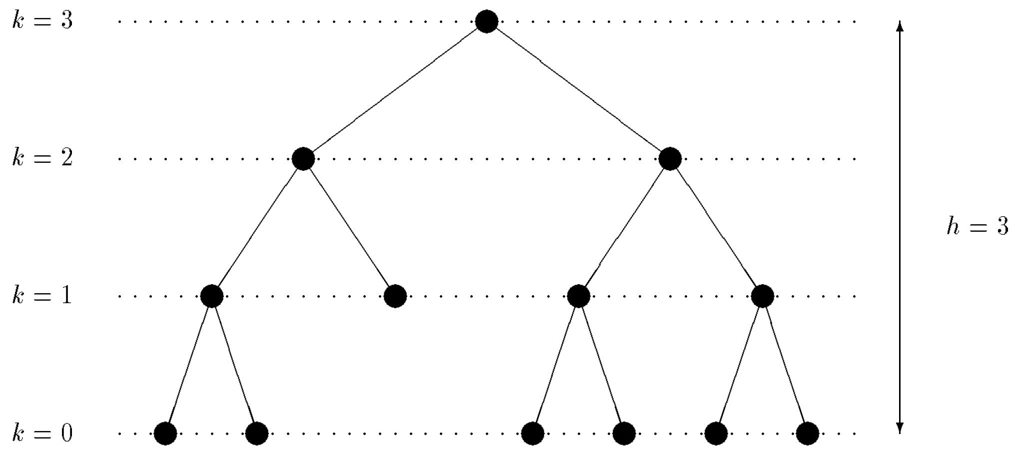

FIGURE 2. Tear Tree.

algorithm, but they are not necessary for the purpose of this error analysis. Further implementation details may be found in [9].

The solution of the linear system (1) can also be described by the tear tree of Figure 2. Each node represents a linear system to be solved. The node on the top level ($k = 3$) stands for the system (1), whereas the leaf nodes are the linear systems that are not divided any further but solved by Gaussian elimination. The number n of diagonal blocks in the matrix A, which is equal to the number of leaf nodes, and the height h of the tear tree are connected by the inequalities

$$n \leq 2^h,$$
$$h \geq \log_2 n.$$

These inequalities become equalities if the tear tree of Figure 2 is a complete binary tree.

3. Basic Operations

Algorithm 1 is composed of a few basic building blocks. These are the addition and multiplication of matrices, the calculation of a URV-decomposition, and the solution of linear systems by Gaussian elimination. We will now state bounds for the errors associated with these operations.

In the case of the addition of two matrices we have

$$\mathrm{fl}(A + B) = A + B + E,$$

where

$$\|E\| \leq \eta_1 \|A + B\|.$$

The quantity η_1 is on the order of the unit roundoff ε and slowly increases with the size of the matrices A and B. See also [13, pp. 114–115 and p. 57].

If we multiply two matrices in floating point we have

$$\mathrm{fl}(AB) = AB + E,$$

where

$$\|E\| \le \eta_2 \|A\| \, \|B\|.$$

Again, η_2 is a small multiple of the unit roundoff and slowly grows with the size of the matrices A and B (cf. [13, pp. 115–116]).

The size of the error in computing an orthogonal URV-decomposition depends on the specifics of the decomposition. One may choose a QR-decomposition [4, Chapter 5], a rank-revealing decomposition [2, 8], or the singular value decomposition [4, Section 8.3]. All of these factorizations have in common that they can be expressed as a sequence of orthogonal transformations applied from the left and the right to the initial matrix. The orthogonal transformations are also accumulated to give the matrices U and V. Wilkinson shows in [13, pp. 160–161] that there are orthogonal matrices U_0 and V_0 and an η_3 such that

$$\|R - U_0^{\mathrm{T}} A V_0\| \le 2\eta_3 \|A\|, \tag{2}$$

$$\|U - U_0\| \le \eta_3, \tag{3}$$

$$\|V - V_0\| \le \eta_3. \tag{4}$$

The quantity η_3 is on the order of the unit roundoff and slowly grows with the size of the matrix A and the number of the orthogonal transformations applied to A. From (3) and (4) it immediately follows that

$$\begin{aligned} \|U\| &\le 1 + \eta_3, \\ \|V\| &\le 1 + \eta_3. \end{aligned}$$

We can also show, by increasing η_3 slightly as necessary, that

$$\begin{aligned} \|URV^{\mathrm{T}} - A\| &\le 4\eta_3 \|A\|, \\ \|(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}\| &\le 1 + 3\eta_3, \\ \|V(V^{\mathrm{T}}V)^{-1}\| &\le 1 + 3\eta_3. \end{aligned}$$

We assume that small linear systems are solved by Gaussian elimination. In [12, p. 108] and [13, p. 252] Wilkinson shows that this process can be described by the equation

$$A\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}, \tag{5}$$

where

$$\|\Delta \mathbf{b}\| \leq \eta_4' \|A\| \, \|\mathbf{x}\|. \tag{6}$$

The value of η_4' is on the order of the unit roundoff and slowly increases with the size of A. It also depends on the pivoting strategy used. See [5] for a more recent survey.

Note that the bound (6) is only applicable if the right hand side of (5) is a vector. If we solve several linear systems with the same matrix A we get

$$AX = B + \Delta B, \tag{7}$$

where

$$\|\Delta B\| \leq \eta_4' \sqrt{r} \|A\| \, \|X\|, \tag{8}$$

and r denotes the number of columns in the matrix B.

Let $r_{\max}$ be the maximum number of right hand sides in a linear system that is solved by Gaussian elimination in Algorithm 1. If we define

$$\eta_4 := \eta_4' \sqrt{r_{\max}},$$

then we can always bound the residual ΔB in (7) by

$$\|\Delta B\| \leq \eta_4 \|A\| \, \|X\|.$$

Thanks to this convention our error bounds will become somewhat simpler.

4. Analysis of one Stage

In the following we will give expressions for the rounding errors incurred at one stage of Algorithm 1. We assume that we are not at the bottom of the tear tree, and we use the assumptions of Section 3 to bound the size of the rounding errors.

In what follows the matrix A denotes the system matrix of an arbitrary interior node of the tear tree. In an attempt to keep the notation simple we do not introduce an index to indicate the corresponding node. We also assume that the four submatrices A_{nw}, A_{ne}, A_{sw}, and A_{se} are predefined by the tearing strategy.

We make the inductive assumption that the solution X computed at level k satisfies

$$AX = B + \Delta B, \tag{9}$$

where the residual ΔB can be expressed by

$$\Delta B = \Delta L_X + \Delta M_X X. \tag{10}$$

We use the index X for the matrices ΔL_X and ΔM_X to indicate that they depend on the solution X.

We assume that at level k we always have

$$\|\Delta L_X\| \leq \xi_k \|D^{-1}X\|, \tag{11}$$

$$\|\Delta M_X D\| \leq \zeta_k, \tag{12}$$

for all matrices X. The quantity D denotes a nonsingular block diagonal matrix, which is partitioned commensurably with A. In particular we always have $D_{ne} = 0$ and $D_{sw} = 0$ for all the nodes in the tear tree. The matrix D will give us additional flexibility in bounding the norm of the residual ΔB. We will discuss this issue in more detail in Sections 7–9.

Since we solve the systems at the bottom level by Gaussian elimination, we define

$$\begin{aligned} \xi_0 &:= \eta_4 \|A\| \, \|D\|, \\ \zeta_0 &:= 0, \end{aligned}$$

The purpose of the next sections will be to compute ξ_k and ζ_k if ξ_{k-1} and ζ_{k-1} are known.

Let us also establish the following convention: If we compute a quantity, say x, the corresponding rounding error is called Δx. Consequently, Δx does not need to be read as a perturbation of x. We only break this convention if several quantities are computed simultaneously, as in the URV-decomposition, and for the residual ΔB.

Throughout our analysis we will assume that the rounding errors remain small compared to the norm of the computed quantities. This means that the computed and the exact quantities will agree to at least a few digits.

4.1. Calculation of the URV-Decomposition. The result of the initial URV-decomposition of Algorithm 1 can be described by

$$A_{sw} = URV^{\mathrm{T}} + \Delta A_{sw}, \tag{13}$$

where

$$\|\Delta A_{sw}\| \leq 4\eta_3 \|A_{sw}\|. \tag{14}$$

The matrices U and V are nearly orthogonal, and they satisfy

$$\|U\| \leq 1 + \eta_3, \qquad \|(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}\| \leq 1 + 3\eta_3, \tag{15}$$

$$\|V\| \leq 1 + \eta_3, \qquad \|V(V^{\mathrm{T}}V)^{-1}\| \leq 1 + 3\eta_3. \tag{16}$$

The expression RV^{T}, which we will also use later on, can be written as

$$RV^{\mathrm{T}} = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}(A_{sw} - \Delta A_{sw}). \tag{17}$$

Therefore we have the bound

$$\|RV^{\mathrm{T}}\| \leq (1 + 3\eta_3)(1 + 4\eta_3)\|A_{sw}\|. \tag{18}$$

Because of

$$R = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}(A_{sw} - \Delta A_{sw})V(V^{\mathrm{T}}V)^{-1},$$

we can also bound the norm of R by

$$\|R\| \le (1+3\eta_3)^2(1+4\eta_3)\|A_{sw}\|. \tag{19}$$

4.2. Calculation of G. The calculation of the matrix G proceeds in three steps that can be described by the equations

$$\begin{aligned} A_{se}G_s &= U + \Delta G_s, \\ U_n &= -A_{ne}G_s + \Delta U_n, \\ A_{nw}G_n &= U_n + \Delta G_n. \end{aligned}$$

The error matrix ΔU_n is bounded by

$$\|\Delta U_n\| \le \eta_2 \|A_{ne}\|\,\|G_s\| \le 1.01\eta_2\|A_{ne}\|\,\|A_{se}^{-1}\|. \tag{20}$$

The residuals ΔG_s and ΔG_n have the expansion

$$\begin{aligned} \Delta G_s &= \Delta L_{G_s} + \Delta M_{G_s}G_s, \\ \Delta G_n &= \Delta L_{G_n} + \Delta M_{G_n}G_n. \end{aligned}$$

We can also write these equations in matrix terms as

$$\hat{A}G = EU + \Delta G, \tag{21}$$

where

$$\Delta G := \begin{bmatrix} \Delta G_n + \Delta U_n \\ \Delta G_s \end{bmatrix} = \begin{bmatrix} \Delta L_{G_n} + \Delta U_n \\ \Delta L_{G_s} \end{bmatrix} + \begin{bmatrix} \Delta M_{G_n} & \\ & \Delta M_{G_s} \end{bmatrix} G. \tag{22}$$

Since the matrix G can be written as

$$G = \hat{A}^{-1}(EU + \Delta G),$$

we also have

$$\|G\| \le 1.01\|\hat{A}^{-1}\|. \tag{23}$$

4.3. Calculation of S. We can express the matrix S by

$$S = RV^{\mathrm{T}}G_n + \Delta S, \tag{24}$$

where

$$\|\Delta S\| \le (2\eta_2 + \eta_2^2)\|R\|\,\|V\|\,\|G_n\|.$$

This bound applies regardless of the sequence in which the two multiplications are performed. In view of (16,19,23) we can also bound $\|\Delta S\|$ by

$$\begin{aligned} \|\Delta S\| &\le 1.01(2\eta_2 + \eta_2^2)(1+\eta_3)(1+3\eta_3)^2(1+4\eta_3)\|A_{sw}\|\,\|\hat{A}^{-1}\| \\ &\le 2\cdot 1.01^2\eta_2\|A_{sw}\|\,\|\hat{A}^{-1}\|. \end{aligned} \tag{25}$$

By means of the equations (13,21,24) we can derive the following more explicit expression for S:

$$S = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}(A_{sw}F^{\mathrm{T}}\hat{A}^{-1}EU - \Delta A_{sw}F^{\mathrm{T}}G + A_{sw}F^{\mathrm{T}}\hat{A}^{-1}\Delta G) + \Delta S.$$

Obviously, the norm of S can be bounded by

$$\|S\| \le 1.01\|A_{sw}F^{\mathrm{T}}\hat{A}^{-1}\|.$$

4.4. Calculation of T. The matrix T satisfies the equation

$$T = I + S + \Delta T, \tag{26}$$

where

$$\|\Delta T\| \le \eta_1\|I + S\|.$$

By means of some straightforward manipulations, using (13,21,24), we can see that

$$EU(I+S) = A\hat{A}^{-1}EU - E\Delta A_{sw}F^{\mathrm{T}}G + EA_{sw}F^{\mathrm{T}}\hat{A}^{-1}\Delta G + EU\Delta S. \tag{27}$$

Consequently, we can express $I + S$ as

$$\begin{aligned} I + S = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E^{\mathrm{T}}(A\hat{A}^{-1}EU - E\Delta A_{sw}F^{\mathrm{T}}G +{} \\ EA_{sw}F^{\mathrm{T}}\hat{A}^{-1}\Delta G) + \Delta S, \end{aligned}$$

and T is given by

$$\begin{aligned} T = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E^{\mathrm{T}}(A\hat{A}^{-1}EU - E\Delta A_{sw}F^{\mathrm{T}}G +{} \\ EA_{sw}F^{\mathrm{T}}\hat{A}^{-1}\Delta G) + \Delta S + \Delta T. \end{aligned}$$

If ΔA_{sw}, ΔG, and ΔS are sufficiently small, we have

$$\|\Delta T\| \le 1.01\eta_1\|A\hat{A}^{-1}\|. \tag{28}$$

Similarly, we also have

$$\|T\| \leq 1.01\|A\hat{A}^{-1}\|. \tag{29}$$

4.5. Calculation of $\hat{R}$. If we solve the linear system for $\hat{R}$ by Gaussian elimination we get

$$T\hat{R} = R + \Delta\hat{R}, \tag{30}$$

where

$$\|\Delta\hat{R}\| \leq \eta_4\|T\|\,\|\hat{R}\|. \tag{31}$$

In order to bound $\|\Delta\hat{R}\|$ differently we need an alternative expression for $\hat{R}$. It is useful to consider the quantity $I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}$ first. By using (13,21,24,26,30) we have

$$\begin{aligned} A(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}) = \hat{A} + E\Delta A_{sw}F^{\mathrm{T}}(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}) - \\ \Delta G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}} + EU\Big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\Big)V^{\mathrm{T}}F^{\mathrm{T}}, \end{aligned} \tag{32}$$

which is equivalent to

$$\begin{aligned} I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}} = A^{-1}\hat{A} + A^{-1}E\Delta A_{sw}F^{\mathrm{T}}(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}) - \\ A^{-1}\Delta G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}} + A^{-1}EU\Big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\Big)V^{\mathrm{T}}F^{\mathrm{T}}. \end{aligned} \tag{33}$$

Consequently, we can represent $\hat{R}$ as

$$\begin{aligned} \hat{R} = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E^{\mathrm{T}}(\hat{A} - \hat{A}A^{-1}\hat{A})FV(V^{\mathrm{T}}V)^{-1} - \\ (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E^{\mathrm{T}}\Big(\hat{A}A^{-1}E\Delta A_{sw}F^{\mathrm{T}}(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}})FV(V^{\mathrm{T}}V)^{-1} + \\ EA_{sw}F^{\mathrm{T}}A^{-1}\Delta G\hat{R} + \hat{A}A^{-1}EU\Big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\Big)\Big). \end{aligned}$$

Note that we can also write $\hat{A} - \hat{A}A^{-1}\hat{A}$ as

$$\hat{A} - \hat{A}A^{-1}\hat{A} = EA_{sw}F^{\mathrm{T}}A^{-1}\hat{A} = \hat{A}A^{-1}EA_{sw}F^{\mathrm{T}}. \tag{34}$$

If we assume the rounding errors to be bounded we can show that

$$\|\hat{R}\| \leq 1.01\|\hat{A} - \hat{A}A^{-1}\hat{A}\|. \tag{35}$$

By combining this result with (29,31) we can bound $\|\Delta\hat{R}\|$ by

$$\|\Delta\hat{R}\| \leq 1.01^2\eta_4\|A\hat{A}^{-1}\|\,\|\hat{A} - \hat{A}A^{-1}\hat{A}\|. \tag{36}$$

4.6. Calculation of P. The calculation of P can be described by the equation

$$P = G\hat{R} + \Delta P, \tag{37}$$

where

$$\|\Delta P\| \leq \eta_2 \|G\| \, \|\hat{R}\|.$$

By using (23,35) we can bound $\|\Delta P\|$ by

$$\|\Delta P\| \leq 1.01^2 \eta_2 \|\hat{A}^{-1}\| \, \|\hat{A} - \hat{A}A^{-1}\hat{A}\|. \tag{38}$$

Because of (33) the following alternative expression for P applies:

$$\begin{aligned} P = {} & A^{-1}EA_{sw}V(V^{\mathrm{T}}V)^{-1} - A^{-1}E\Delta A_{sw}F^{\mathrm{T}}(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}})FV(V^{\mathrm{T}}V)^{-1} + \\ & A^{-1}\Delta G\hat{R} - A^{-1}EU\Big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\Big) + \Delta P. \end{aligned}$$

Provided that the rounding errors remain bounded we certainly have

$$\|P\| \leq 1.01 \|A^{-1}EA_{sw}\|. \tag{39}$$

4.7. Calculation of Y. The matrix Y is computed in three steps as follows:

$$A_{se}Y_s = B_s + \Delta Y_s, \tag{40}$$

$$\hat{B}_n = B_n - A_{ne}Y_s + \Delta\hat{B}_n, \tag{41}$$

$$A_{nw}Y_n = \hat{B}_n + \Delta Y_n. \tag{42}$$

The error in computing $\hat{B}_n$ can be bounded by

$$\begin{aligned} \|\Delta\hat{B}_n\| &\leq \eta_1 \|B_n - A_{ne}Y_s\| + (1 + \eta_1)\eta_2 \|A_{ne}\| \, \|Y_s\| \\ &\leq \eta_1 \|B_n\| + (\eta_1 + \eta_2 + \eta_1\eta_2) \|A_{ne}\| \, \|Y_s\|. \end{aligned} \tag{43}$$

On the other hand we can assume the following expressions for the residuals ΔY_s and ΔY_n:

$$\Delta Y_s = \Delta L_{Y_s} + \Delta M_{Y_s} Y_s, \tag{44}$$

$$\Delta Y_n = \Delta L_{Y_n} + \Delta M_{Y_n} Y_n. \tag{45}$$

The equations (40,41,42,44,45) can also be written in matrix terms as

$$\hat{A}Y = B + \Delta Y, \tag{46}$$

where

$$\Delta Y := \begin{bmatrix} \Delta Y_n + \Delta\hat{B}_n \\ \Delta Y_s \end{bmatrix} = \begin{bmatrix} \Delta L_{Y_n} + \Delta\hat{B}_n \\ \Delta L_{Y_s} \end{bmatrix} + \begin{bmatrix} \Delta M_{Y_n} & \\ & \Delta M_{Y_s} \end{bmatrix} Y. \tag{47}$$

It will turn out to be useful to eliminate B_n from the right hand side of (43). Observe that

$$\|B_n\| \leq \|B\| \leq \|\hat{A}\|\,\|Y\| + \|\Delta Y\| \leq 1.01\|\hat{A}\|\,\|Y\|.$$

Consequently, we can also bound $\|\Delta\hat{B}_n\|$ by

$$\|\Delta\hat{B}_n\| \leq (2.01\eta_1 + \eta_2 + \eta_1\eta_2)\|\hat{A}\|\,\|Y\| \leq 1.01(2\eta_1 + \eta_2)\|\hat{A}\|\,\|Y\|. \tag{48}$$

4.8. Calculation of X. The computed value of X satisfies

$$X = Y - PV^{\mathrm{T}}Y_n + \Delta X, \tag{49}$$

where

$$\|\Delta X\| \leq \eta_1\|Y - PV^{\mathrm{T}}Y_n\| + (1+\eta_1)(2+\eta_2)\eta_2\|P\|\,\|V\|\,\|Y_n\|.$$

This bound can be processed further by using (16,39), and we get

$$\|\Delta X\| \leq \Big(\eta_1 + 1.01(1+\eta_3)\big(\eta_1 + (1+\eta_1)(2+\eta_2)\eta_2\big)\|A^{-1}EA_{sw}\|\Big)\|Y\|.$$

A simpler bound is given by

$$\|\Delta X\| \leq \big(\eta_1 + 1.01^2(\eta_1 + 2\eta_2)\|A^{-1}EA_{sw}\|\big)\|Y\|. \tag{50}$$

5. Backward Error

We are now ready to give an expression for the residual ΔB in (9). By using the equations (32,37,46,49) we get

$$\begin{aligned}\Delta B = \Big(&E\Delta A_{sw}F^{\mathrm{T}}(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}) - \Delta G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}} + \\ &EU\big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\big)V^{\mathrm{T}}F^{\mathrm{T}} - \\ &A\Delta PV^{\mathrm{T}}F^{\mathrm{T}}\Big)Y + \Delta Y + A\Delta X.\end{aligned} \tag{51}$$

The value of ΔB can be processed further. By combining (37) and (49) we get

$$(I - G\hat{R}V^{\mathrm{T}}F^{\mathrm{T}})Y = X + \Delta PV^{\mathrm{T}}F^{\mathrm{T}}Y - \Delta X. \tag{52}$$

If we use (9,13,21,33,34,46,51,52) we can also derive the following alternative expression for $\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}Y$:

$$\begin{aligned}\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}Y = {} & RV^{\mathrm{T}}F^{\mathrm{T}}X - \big((\Delta S + \Delta T)\hat{R} - \Delta\hat{R}\big)V^{\mathrm{T}}F^{\mathrm{T}}Y + \\ & RV^{\mathrm{T}}F^{\mathrm{T}}\Delta PV^{\mathrm{T}}F^{\mathrm{T}}Y - RV^{\mathrm{T}}F^{\mathrm{T}}\Delta X.\end{aligned} \tag{53}$$

Because of (22,47,52,53) we are led to the following expression for ΔB:

$$
\begin{aligned}
\Delta B = {} & E\Delta A_{sw}F^{\mathrm{T}}X - \\
& \begin{bmatrix} \Delta L_{G_n} + \Delta U_n \\ \Delta L_{G_s} \end{bmatrix} RV^{\mathrm{T}}F^{\mathrm{T}}X - \begin{bmatrix} \Delta M_{G_n} & \\ & \Delta M_{G_s} \end{bmatrix} GRV^{\mathrm{T}}F^{\mathrm{T}}X + \\
& EU(\Delta S + \Delta T)RV^{\mathrm{T}}F^{\mathrm{T}}X - EU\Delta\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}Y - A\Delta PV^{\mathrm{T}}F^{\mathrm{T}}Y + \\
& \begin{bmatrix} \Delta L_{Y_n} + \Delta\hat{B}_n \\ \Delta L_{Y_s} \end{bmatrix} + \begin{bmatrix} \Delta M_{Y_n} & \\ & \Delta M_{Y_s} \end{bmatrix} Y + A\Delta X + O(\varepsilon^2).
\end{aligned} \tag{54}
$$

The matrix $O(\varepsilon^2)$ represents second order terms. They tend to zero as the unit roundoff ε tends to zero. Specifically, we have

$$
\lim_{\varepsilon \downarrow 0} \frac{\|O(\varepsilon^2)\|}{\varepsilon} = 0.
$$

A key expression in (54) is $GRV^{\mathrm{T}}F^{\mathrm{T}}$. With the help of (13,21) we can transform this matrix into

$$
GRV^{\mathrm{T}}F^{\mathrm{T}} = \hat{A}^{-1}EA_{sw}F^{\mathrm{T}} - \hat{A}^{-1}E\Delta A_{sw}F^{\mathrm{T}} + \hat{A}^{-1}\Delta GRV^{\mathrm{T}}F^{\mathrm{T}}.
$$

Similarly, we use (9,46) to write Y as

$$
\begin{aligned}
Y &= \hat{A}^{-1}AX - \hat{A}^{-1}(\Delta B - \Delta Y) \\
&= D(D^{-1}\hat{A}^{-1}AD)D^{-1}X - \hat{A}^{-1}(\Delta B - \Delta Y).
\end{aligned} \tag{55}
$$

Now it is straightforward to write the residual ΔB as

$$
\Delta B = \Delta L_X + \Delta M_X X,
$$

where

$$
\Delta L_X = \begin{bmatrix} \Delta L_{Y_n} + \Delta\hat{B}_n \\ \Delta L_{Y_s} \end{bmatrix} + A\Delta X + O(\varepsilon^2), \tag{56}
$$

$$
\begin{aligned}
\Delta M_X = {} & E\Delta A_{sw}F^{\mathrm{T}} - \begin{bmatrix} \Delta L_{G_n} + \Delta U_n \\ \Delta L_{G_s} \end{bmatrix} RV^{\mathrm{T}}F^{\mathrm{T}} - \\
& \begin{bmatrix} \Delta M_{G_n} & \\ & \Delta M_{G_s} \end{bmatrix} \hat{A}^{-1}EA_{sw}F^{\mathrm{T}} + \\
& EU(\Delta S + \Delta T)RV^{\mathrm{T}}F^{\mathrm{T}} - EU\Delta\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}\hat{A}^{-1}A - \\
& A\Delta PV^{\mathrm{T}}F^{\mathrm{T}}\hat{A}^{-1}A + \begin{bmatrix} \Delta M_{Y_n} & \\ & \Delta M_{Y_s} \end{bmatrix} \hat{A}^{-1}A.
\end{aligned} \tag{57}
$$

Let us give bounds for the norms of ΔL_X and ΔM_X. From (48,50,55) we have

$$\begin{aligned}\|\Delta L_X\| \le 1.01\Big(&\sqrt{2}\xi_{k-1} + 1.01(2\eta_1+\eta_2)\|\hat{A}\|\,\|D\| + \\ &\qquad\Big(\eta_1 + 1.01^2(\eta_1+2\eta_2)\|A^{-1}EA_{sw}\|\Big)\|A\|\,\|D\|\Big)\cdot \\ &\|D^{-1}\hat{A}^{-1}AD\|\,\|D^{-1}X\|.\end{aligned} \tag{58}$$

In order to bound $\|\Delta M_X D\|$ we need to multiply (57) by D from the right:

$$\begin{aligned}\Delta M_X D = {} & E\Delta A_{sw}F^{\mathrm{T}}D - \begin{bmatrix}\Delta L_{G_n}+\Delta U_n \\ \Delta L_{G_s}\end{bmatrix} RV^{\mathrm{T}}F^{\mathrm{T}}D - \\ & \begin{bmatrix}\Delta M_{G_n}D_{nw} & \\ & \Delta M_{G_s}D_{se}\end{bmatrix} D^{-1}\hat{A}^{-1}EA_{sw}F^{\mathrm{T}}D + \\ & EU(\Delta S+\Delta T)RV^{\mathrm{T}}F^{\mathrm{T}}D - EU\Delta\hat{R}V^{\mathrm{T}}F^{\mathrm{T}}DD^{-1}\hat{A}^{-1}AD - \\ & A\Delta PV^{\mathrm{T}}F^{\mathrm{T}}DD^{-1}\hat{A}^{-1}AD + \\ & \begin{bmatrix}\Delta M_{Y_n}D_{nw} & \\ & \Delta M_{Y_s}D_{se}\end{bmatrix} D^{-1}\hat{A}^{-1}AD.\end{aligned}$$

Because of (17) we can write the expression $RV^{\mathrm{T}}F^{\mathrm{T}}D$ also as

$$RV^{\mathrm{T}}F^{\mathrm{T}}D = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}(A_{sw}-\Delta A_{sw})F^{\mathrm{T}}D.$$

Consequently, its norm is certainly bounded by

$$\|RV^{\mathrm{T}}F^{\mathrm{T}}D\| \le 1.01\|A_{sw}D_{nw}\|. \tag{59}$$

If we use (11,12,14,15,16,18,20,25,28,36,38,59) we can bound $\|\Delta M_X D\|$ as follows:

$$\begin{aligned}\|\Delta M_X D\| \le {} & 4\eta_3\|A_{sw}\|\,\|D_{nw}\| + \\ & 1.01^2(\sqrt{2}\|\hat{A}^{-1}\|\xi_{k-1} + \eta_2\|A_{ne}\|\,\|A_{se}^{-1}\|)\|A_{sw}D_{nw}\| + \\ & (1+\eta_3)1.01\Big(2\cdot 1.01^2\eta_2\|A_{sw}\|\,\|\hat{A}^{-1}\| + \\ & \qquad 1.01\eta_1\|A\hat{A}^{-1}\|\Big)\|A_{sw}D_{nw}\| + \\ & (1+\eta_3)^2 1.01^2\eta_4\|A\hat{A}^{-1}\|\,\|\hat{A}-\hat{A}A^{-1}\hat{A}\|\,\|D_{nw}\|\,\|D^{-1}\hat{A}^{-1}AD\| + \\ & 1.01^2(1+\eta_3)\eta_2\|A\|\,\|\hat{A}^{-1}\|\,\|\hat{A}-\hat{A}A^{-1}\hat{A}\|\,\|D_{nw}\|\,\|D^{-1}\hat{A}^{-1}AD\| + \\ & (\|D^{-1}\hat{A}^{-1}EA_{sw}D_{nw}\| + \|D^{-1}\hat{A}^{-1}AD\|)\zeta_{k-1}.\end{aligned}$$

It is useful to simplify this bound somewhat:

$$
\begin{aligned}
\|\Delta M_X D\| \le\ & 4\eta_3\|A_{sw}\|\,\|D_{nw}\| + 1.01^4(\eta_1+3\eta_2)\|A_{sw}D_{nw}\|\,\|A\|\,\|\hat{A}^{-1}\| + \\
& 1.01^3(\eta_2+\eta_4)\|A\|\,\|\hat{A}^{-1}\|\,\|\hat{A}-\hat{A}A^{-1}\hat{A}\|\,\|D_{nw}\|\,\|D^{-1}\hat{A}^{-1}AD\| + \\
& 1.01^2\sqrt{2}\|A_{sw}D_{nw}\|\,\|\hat{A}^{-1}\|\xi_{k-1} + \\
& (\|D^{-1}\hat{A}^{-1}EA_{sw}D_{nw}\| + \|D^{-1}\hat{A}^{-1}AD\|)\zeta_{k-1}.
\end{aligned} \tag{60}
$$

The inequalities (58) and (60) contain the quantities ξ_{k-1} and ζ_{k-1} which are used to bound the errors at the previous level of the tear tree. Consequently, we can use (58) and (60) to give recurrence relationships for ξ_k and ζ_k. Let us define the quantities

$$f_\xi := \max 1.01\sqrt{2}\|D^{-1}\hat{A}^{-1}AD\|, \tag{61}$$

$$f_\zeta := \max \|D^{-1}\hat{A}^{-1}EA_{sw}D_{nw}\| + \|D^{-1}\hat{A}^{-1}AD\|, \tag{62}$$

$$g := \max 1.01^2\sqrt{2}\|A_{sw}D_{nw}\|\,\|\hat{A}^{-1}\|, \tag{63}$$

$$
\begin{aligned}
c_\xi := \max 1.01\Big(& 1.01(2\eta_1+\eta_2)\|\hat{A}\|\,\|D\| + \\
& \Big(\eta_1 + 1.01^2(\eta_1+2\eta_2)\|A^{-1}EA_{sw}\|\Big)\|A\|\,\|D\|\Big)\cdot \\
& \|D^{-1}\hat{A}^{-1}AD\|,
\end{aligned} \tag{64}
$$

$$
\begin{aligned}
c_\zeta := \max\ & 4\eta_3\|A_{sw}\|\,\|D_{nw}\| + \\
& 1.01^4(\eta_1+3\eta_2)\|A_{sw}D_{nw}\|\,\|A\|\,\|\hat{A}^{-1}\| + \\
& 1.01^3(\eta_2+\eta_4)\|A\|\,\|\hat{A}^{-1}\|\,\|\hat{A}-\hat{A}A^{-1}\hat{A}\|\cdot \\
& \|D_{nw}\|\,\|D^{-1}\hat{A}^{-1}AD\|,
\end{aligned} \tag{65}
$$

where the maximum is to be taken over all the nodes in the tear tree. The sequences $\{\xi_k\}$ and $\{\zeta_k\}$ thus satisfy the following recurrence relationships:

$$
\begin{aligned}
\xi_0 &= \eta_4\|A\|\,\|D\|, \\
\xi_k &= f_\xi\xi_{k-1} + c_\xi, && k \ge 1, \\
\zeta_0 &= 0, \\
\zeta_k &= f_\zeta\zeta_{k-1} + g\xi_{k-1} + c_\zeta, && k \ge 1.
\end{aligned}
$$

Their explicit solutions are given by[1]

$$\xi_k = \xi_0 f_\xi^k + c_\xi \frac{f_\xi^k - 1}{f_\xi - 1}, \tag{66}$$

$$\begin{aligned}\zeta_k = {}& \xi_0 g \frac{f_\xi^k - f_\zeta^k}{f_\xi - f_\zeta} + c_\zeta \frac{f_\zeta^k - 1}{f_\zeta - 1} + \\ & c_\xi g \Big(\frac{1}{(f_\xi - 1)(f_\zeta - 1)} + \frac{f_\xi^k}{(f_\xi - 1)(f_\xi - f_\zeta)} - \frac{f_\zeta^k}{(f_\zeta - 1)(f_\xi - f_\zeta)} \Big).\end{aligned} \tag{67}$$

These explicit expressions for ξ_k and ζ_k are only valid if $f_\xi \neq 1$, $f_\zeta \neq 1$, and $f_\xi \neq f_\zeta$. It would be possible to give similar expressions for these special cases, too. However these formulas would not give us more insight than (66) and (67).

At the top level of the tear tree the residual ΔB is given by

$$\Delta B = \Delta L_X + \Delta M_X X,$$

where

$$\begin{aligned}\|\Delta L_X\| &\leq \xi_h \|D^{-1}X\|, \\ \|\Delta M_X D\| &\leq \zeta_h.\end{aligned}$$

Consequently, $\|\Delta B\|$ can be bounded by

$$\|\Delta B\| \leq (\xi_h + \zeta_h)\|D^{-1}X\|. \tag{68}$$

The expression $\|D^{-1}X\|$ is nothing else than the size of the solution X expressed in another norm.

6. Stability Criterion

Let us now analyse the conditions under which Algorithm 1 computes a backward stable solution. We assume that all the matrices A, $\hat{A}$, and D in the tear tree are nonsingular and only moderately ill-conditioned. Without this assumption the quantities c_ξ and c_ζ could become arbitrarily large, like in the case of the matrices

$$A := \begin{bmatrix} \epsilon & 1 \\ 1 & \epsilon \end{bmatrix}, \qquad \hat{A} := \begin{bmatrix} \epsilon & 1 \\ 0 & \epsilon \end{bmatrix},$$

with $\epsilon \downarrow 0$.

If c_ξ and c_ζ are only small multiples of $\varepsilon\|A\|$ the norm of the residual ΔB in (68) will be on the order of $\varepsilon\|A\|\,\|D^{-1}X\|$ for $f_\xi \approx 1$ and $f_\zeta \approx 1$. This condition is equivalent to the requirement that there exists a nonsingular block diagonal matrix D, partitioned commensurably with A, such that

$$\|D^{-1}\hat{A}^{-1}AD\| \approx 1 \tag{69}$$

[1]We used Maple [3] to derive this result.

for all the matrices in the tear tree. If this stability criterion is met Algorithm 1 is guaranteed to compute a backward stable solution provided that all the matrices A, $\hat{A}$, and D in the tear tree are only moderately ill-conditioned.

It should be noted that (69) is a sufficient but not a necessary stability condition. Since we use the quantities ξ_k and ζ_k to bound $\|\Delta L_X\|$ and $\|\Delta M_X D\|$ at each level in the tear tree, these bounds may grow even if ΔL_X and $\Delta M_X D$ remain bounded. On the positive side, we get a manageable error analysis and a simple stability criterion.

In the next two sections we will identify two classes of matrices for which the criterion (69) is always satisfied.

7. Diagonally Dominant Matrices

An important class of matrices, for which the condition (69) is always satisfied, is given by the set of nonsingular diagonally dominant matrices. In order to see this we need the following theorem.

Theorem 7.1. *Let A be an m-by-n matrix with $m < n$, and let A be diagonally dominant, i.e.*

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \leq |a_{ii}|, \qquad i = 1, \ldots, m.$$

We partition A into

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix},$$

where A_1 is a square m-by-m matrix, and A_2 is an m-by-$(n-m)$ matrix. We also assume that A_1 is nonsingular. Under these assumptions the inequality

$$\|A_1^{-1} A_2\|_\infty \leq 1$$

applies.

Proof. The proof is based on the observation that Gaussian elimination without pivoting can be applied to the matrix A, and that A remains diagonally dominant during this process (cf. [11, pp. 288–289]). Specifically there are elementary matrices L_k, R_k, and a diagonal matrix D such that

$$R_{m-1} \cdots R_1 D L_{m-1} \cdots L_1 A_1 = I.$$

If we apply these transformations to A we get

$$R_{m-1} \cdots R_1 D L_{m-1} \cdots L_1 A = \begin{bmatrix} I & A_1^{-1} A_2 \end{bmatrix}.$$

Since these transformations preserve the diagonal dominance of the matrix A we conclude that $\|A_1^{-1} A_2\|_\infty \leq 1$. $\square$

As a straightforward application of Theorem 7.1 we consider the norm of $\hat{A}^{-1}A$ when A is diagonally dominant. It is easily verified that

$$\hat{A}^{-1}A = \begin{bmatrix} I - A_{nw}^{-1}A_{ne}A_{se}^{-1}A_{sw} & 0 \\ A_{se}^{-1}A_{sw} & I \end{bmatrix}.$$

The two matrices $\begin{bmatrix} A_{nw} & A_{ne} \end{bmatrix}$ and $\begin{bmatrix} A_{se} & A_{sw} \end{bmatrix}$ satisfy the assumptions of Theorem 7.1, and we have

$$\|A_{nw}^{-1}A_{ne}\|_\infty \leq 1,$$
$$\|A_{se}^{-1}A_{sw}\|_\infty \leq 1.$$

It is now easy to see that

$$\|\hat{A}^{-1}A\|_\infty \leq 2.$$

This bound is tight. It is attained, for instance, by the matrices

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \hat{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

If a matrix A is diagonally dominant then so are all its submatrices in the tear tree. Consequently, if we set $D = I$, the condition (69) is satisfied, and Algorithm 1 will compute a stable solution of the linear system (1), provided that all the matrices A and $\hat{A}$ in the tear tree are only moderately ill-conditioned.

8. M-Matrices

A nonsingular n-by-n matrix A is called an M-matrix if $a_{ij} \leq 0$ for $i \neq j$ and all the entries in A^{-1} are nonnegative. Many alternative characterizations of an M-matrix may be found in [1, Chapter 6] and [6, Section 6.4]. For the purpose of this error analysis the following definition is the most useful: A nonsingular n-by-n matrix A is called an M-matrix if $a_{ij} \leq 0$ for $i \neq j$, $a_{ii} > 0$, and there exists a positive diagonal matrix D such that AD is strictly diagonally dominant, i.e.

$$a_{ii}d_{ii} > \sum_{j \neq i} |a_{ij}|d_{jj}, \qquad i = 1, \ldots, n.$$

Without loss of generality we may assume that $\|D\| = 1$. Note that this definition is equivalent to the condition (M_{35}) in [1, p. 137].

In view of the preceding discussion in Section 7 the stability criterion (69) is satisfied for this particular choice of the matrix D. Therefore, Algorithm 1 will compute a stable solution for linear systems with M-matrices.

9. Numerical Results

In this section we will present numerical results for three classes of test matrices to illustrate our error bounds. As our first example we choose the matrix

$$A := \left[\begin{array}{cc|cc|cc|cc} \epsilon & & 1 & & & & & \\ & 1 & 1 & & & & & \\ \hline & 1 & \epsilon & & 1 & & & \\ & & & 1 & 1 & & & \\ \hline & & & 1 & \epsilon & & 1 & \\ & & & & & 1 & 1 & \\ \hline & & & & & 1 & \epsilon & \\ & & & & & & & 1 \end{array}\right].$$

We will see in a moment that the condition (69) is not satisfied for this matrix. The basic building block of A is given by

$$M := \begin{bmatrix} 1 & 1 \\ 1 & \epsilon \end{bmatrix},$$

which is a well-conditioned matrix for $0 < \epsilon < 1/2$. Specifically we have

$$\kappa_\infty(M) = \frac{4}{1-\epsilon}.$$

On the other hand the matrix

$$\hat{M} := \begin{bmatrix} 1 & 1 \\ 0 & \epsilon \end{bmatrix}$$

becomes increasingly ill-conditioned as ϵ tends to zero:

$$\kappa_\infty(\hat{M}) = 2\frac{1+\epsilon}{\epsilon}.$$

Consequently, if we partition the matrix A as indicated we can expect large values of $\|D^{-1}\hat{A}^{-1}AD\|$ for all nodes in the tear tree.

In order to avoid large matrices we choose an unusual tearing strategy: If the size N of the matrix A is two we solve the linear system by Gaussian elimination, otherwise we set $n_{nw} = N - 2$ and $n_{se} = 2$ (cf. Figure 1). This strategy leads to a degenerate tear tree with a height of $h = N/2 - 1$. The point of our example, however, does not depend on a particular tearing strategy. For any given strategy we can construct a matrix A that exhibits the same problems.

TABLE 1. Properties of the First Class of Test Matrices.

N	$\|A\|$	$\|A^{-1}\hat{A}\|$	$\|\hat{A}^{-1}\|$	$\|D^{-1}\hat{A}^{-1}AD\|$	$\|\hat{A}-\hat{A}A^{-1}\hat{A}\|$
4	1.735	$1.010\cdot 10^{2}$	$1.000\cdot 10^{4}$	$9.901\cdot 10^{1}$	$1.010\cdot 10^{-2}$
6	1.906	$1.021\cdot 10^{2}$	$1.010\cdot 10^{4}$	$1.414\cdot 10^{2}$	$1.010\cdot 10^{-2}$
8	1.954	$1.031\cdot 10^{2}$	$1.021\cdot 10^{4}$	$1.744\cdot 10^{2}$	$1.010\cdot 10^{-2}$
10	1.974	$1.041\cdot 10^{2}$	$1.031\cdot 10^{4}$	$2.026\cdot 10^{2}$	$1.010\cdot 10^{-2}$
12	1.984	$1.052\cdot 10^{2}$	$1.042\cdot 10^{4}$	$2.278\cdot 10^{2}$	$1.010\cdot 10^{-2}$
14	1.990	$1.063\cdot 10^{2}$	$1.053\cdot 10^{4}$	$2.509\cdot 10^{2}$	$1.010\cdot 10^{-2}$
16	1.994	$1.074\cdot 10^{2}$	$1.064\cdot 10^{4}$	$2.724\cdot 10^{2}$	$1.010\cdot 10^{-2}$
18	1.996	$1.085\cdot 10^{2}$	$1.074\cdot 10^{4}$	$2.928\cdot 10^{2}$	$1.010\cdot 10^{-2}$
20	1.998	$1.096\cdot 10^{2}$	$1.085\cdot 10^{4}$	$3.122\cdot 10^{2}$	$1.010\cdot 10^{-2}$
22	1.999	$1.107\cdot 10^{2}$	$1.097\cdot 10^{4}$	$3.308\cdot 10^{2}$	$1.010\cdot 10^{-2}$

For our numerical example we choose $\epsilon = 0.01$. If we set $D = I$ the value of the quantity $\|D^{-1}\hat{A}^{-1}AD\|$ is on the order of $\epsilon^{-2} = 10^4$. However, if we choose

$$D := \left[\begin{array}{cc|cc|cc|cc} 1 & & & & & & & \\ & c & & & & & & \\ \hline & & 1 & & & & & \\ & & & c & & & & \\ \hline & & & & 1 & & & \\ & & & & & c & & \\ \hline & & & & & & 1 & \\ & & & & & & & c \end{array}\right],$$

where $c := \epsilon^2 = 10^{-4}$, this norm can be reduced to the order of $\epsilon^{-1} = 10^2$.

In Table 1 we give the values of some key quantities from (61,62,63,64,65). For any A with $N \geq 4$ we obtain the corresponding $\hat{A}$ by setting $a_{N-1,N-2}$ to zero. This is consistent with our special tearing strategy. We can see from the values of Table 1 that all the subproblems in the tear tree are only moderately ill-conditioned.

The entries of the right hand side $\mathbf{b}$ are given by $b_i = i$. We computed our results on a DECstation 3100 using a MATLAB [10] implementation of Algorithm 1. The unit roundoff is given by $\varepsilon = 2^{-52} \approx 2.2204 \cdot 10^{-16}$. We used the singular value decomposition as our URV-decomposition.

In Table 2 we present the absolute and relative residuals for increasing matrix sizes N. Due to our special tearing strategy the height h of the tear tree increases linearly with N. We also compare the residuals obtained from Algorithm 1 with those from Gaussian elimination (backslash operator in MATLAB). The reader should observe the exponential error growth of the residual $\|A\mathbf{x} - \mathbf{b}\|$, which is due to the large value of $\|D^{-1}\hat{A}^{-1}AD\|$.

TABLE 2. Residuals for the First Class of Test Matrices.

		$\|\|A\mathbf{x}-\mathbf{b}\|\|$		$\dfrac{\|\|A\mathbf{x}-\mathbf{b}\|\|}{\|\|A\|\|\,\|\|\mathbf{x}\|\|}$	
N	h	Algorithm 1	Gaussian elimination	Algorithm 1	Gaussian elimination
4	1	$3.634 \cdot 10^{-14}$	$4.965 \cdot 10^{-16}$	$1.042 \cdot 10^{-16}$	$1.423 \cdot 10^{-18}$
6	2	$5.298 \cdot 10^{-12}$	$4.441 \cdot 10^{-16}$	$9.169 \cdot 10^{-15}$	$7.687 \cdot 10^{-19}$
8	3	$1.523 \cdot 10^{-10}$	$8.882 \cdot 10^{-16}$	$1.918 \cdot 10^{-13}$	$1.119 \cdot 10^{-18}$
10	4	$3.020 \cdot 10^{-8}$	$6.280 \cdot 10^{-16}$	$2.996 \cdot 10^{-11}$	$6.231 \cdot 10^{-19}$
12	5	$7.708 \cdot 10^{-6}$	$8.882 \cdot 10^{-16}$	$6.307 \cdot 10^{-9}$	$7.268 \cdot 10^{-19}$
14	6	$2.044 \cdot 10^{-3}$	$1.538 \cdot 10^{-15}$	$1.422 \cdot 10^{-6}$	$1.070 \cdot 10^{-18}$
16	7	$9.639 \cdot 10^{-2}$	$2.176 \cdot 10^{-15}$	$5.754 \cdot 10^{-5}$	$1.315 \cdot 10^{-18}$
18	8	$6.382 \cdot 10^{0}$	$1.776 \cdot 10^{-15}$	$2.538 \cdot 10^{-3}$	$9.483 \cdot 10^{-19}$
20	9	$1.462 \cdot 10^{3}$	$1.776 \cdot 10^{-15}$	$4.403 \cdot 10^{-3}$	$8.483 \cdot 10^{-19}$
22	10	$2.708 \cdot 10^{5}$	$1.986 \cdot 10^{-15}$	$1.958 \cdot 10^{-2}$	$8.572 \cdot 10^{-19}$

TABLE 3. Properties of the Class of Diagonally Dominant Matrices.

N	$\|\|A\|\|$	$\|\|A^{-1}\hat{A}\|\|$	$\|\|\hat{A}^{-1}\|\|$	$\|\|D^{-1}\hat{A}^{-1}AD\|\|$	$\|\|\hat{A}-\hat{A}A^{-1}\hat{A}\|\|$
4	1.990	$7.143 \cdot 10^{1}$	$1.611 \cdot 10^{0}$	1.407	$5.024 \cdot 10^{1}$
6	1.995	$1.060 \cdot 10^{2}$	$1.010 \cdot 10^{2}$	1.572	$6.677 \cdot 10^{1}$
8	1.997	$1.328 \cdot 10^{2}$	$2.025 \cdot 10^{2}$	1.752	$7.505 \cdot 10^{1}$
10	1.998	$1.554 \cdot 10^{2}$	$3.458 \cdot 10^{2}$	1.922	$8.003 \cdot 10^{1}$
12	1.999	$1.752 \cdot 10^{2}$	$5.303 \cdot 10^{2}$	2.081	$8.335 \cdot 10^{1}$
14	1.999	$1.931 \cdot 10^{2}$	$7.558 \cdot 10^{2}$	2.230	$8.573 \cdot 10^{1}$
16	1.999	$2.095 \cdot 10^{2}$	$1.022 \cdot 10^{3}$	2.371	$8.751 \cdot 10^{1}$
18	1.999	$2.248 \cdot 10^{2}$	$1.330 \cdot 10^{3}$	2.503	$8.889 \cdot 10^{1}$
20	2.000	$2.391 \cdot 10^{2}$	$1.678 \cdot 10^{3}$	2.630	$9.000 \cdot 10^{1}$
22	2.000	$2.526 \cdot 10^{2}$	$2.067 \cdot 10^{3}$	2.751	$9.091 \cdot 10^{1}$

As our second test matrix we choose the diagonally dominant matrix

$$A := \left[\begin{array}{cc|cc|cc|cc} 1 & a & & & & & & \\ a & 1 & b & & & & & \\ \hline & b & 1 & a & & & & \\ & & a & 1 & b & & & \\ \hline & & & b & 1 & a & & \\ & & & & a & 1 & b & \\ \hline & & & & & b & 1 & a \\ & & & & & & a & 1 \end{array}\right],$$

where we set $a := 0.01$ and $b := 0.99$. All the subproblems in the tear tree are relatively well-conditioned. To illustrate this we present in Table 3 the values of some key quantities in the tear tree.

TABLE 4. Residuals for the Class of Diagonally Dominant Matrices.

		$\|\|A\mathbf{x}-\mathbf{b}\|\|$		$\dfrac{\|\|A\mathbf{x}-\mathbf{b}\|\|}{\|\|A\|\|\,\|\|\mathbf{x}\|\|}$	
N	h	Algorithm 1	Gaussian elimination	Algorithm 1	Gaussian elimination
4	1	$8.379 \cdot 10^{-15}$	$2.665 \cdot 10^{-15}$	$6.066 \cdot 10^{-17}$	$1.929 \cdot 10^{-17}$
6	2	$1.801 \cdot 10^{-14}$	$1.779 \cdot 10^{-14}$	$4.593 \cdot 10^{-17}$	$4.535 \cdot 10^{-17}$
8	3	$7.121 \cdot 10^{-14}$	$3.815 \cdot 10^{-14}$	$8.786 \cdot 10^{-17}$	$4.708 \cdot 10^{-17}$
10	4	$8.175 \cdot 10^{-14}$	$1.113 \cdot 10^{-13}$	$5.759 \cdot 10^{-17}$	$7.842 \cdot 10^{-17}$
12	5	$2.283 \cdot 10^{-13}$	$1.088 \cdot 10^{-13}$	$1.018 \cdot 10^{-16}$	$4.851 \cdot 10^{-17}$
14	6	$2.863 \cdot 10^{-13}$	$1.632 \cdot 10^{-13}$	$8.677 \cdot 10^{-17}$	$4.947 \cdot 10^{-17}$
16	7	$4.768 \cdot 10^{-13}$	$1.835 \cdot 10^{-13}$	$1.034 \cdot 10^{-16}$	$3.981 \cdot 10^{-17}$
18	8	$5.265 \cdot 10^{-13}$	$3.521 \cdot 10^{-13}$	$8.504 \cdot 10^{-17}$	$5.687 \cdot 10^{-17}$
20	9	$1.019 \cdot 10^{-12}$	$5.198 \cdot 10^{-13}$	$1.265 \cdot 10^{-16}$	$6.450 \cdot 10^{-17}$
22	10	$8.965 \cdot 10^{-13}$	$4.299 \cdot 10^{-13}$	$8.763 \cdot 10^{-17}$	$4.201 \cdot 10^{-17}$

We use the same tearing strategy as in the first example. Again, the vector $\mathbf{b}$ is given by $b_i = i$. The norms of the residuals for different matrix sizes are presented as Table 4.

As our last example we choose the M-matrix

$$A := \left[\begin{array}{cc|cc|cc|cc}
1 & b & & & & & & \\
a & 1 & a & & & & & \\
\hline
 & b & 1 & b & & & & \\
 & & a & 1 & a & & & \\
\hline
 & & & b & 1 & b & & \\
 & & & & a & 1 & a & \\
\hline
 & & & & & b & 1 & b \\
 & & & & & & a & 1
\end{array}\right],$$

where we set $a := -0.00049$ and $b := -490$. The matrix AD is diagonally dominant for

$$D := \left[\begin{array}{cc|cc|cc|cc}
c & & & & & & & \\
 & 1 & & & & & & \\
\hline
 & & c & & & & & \\
 & & & 1 & & & & \\
\hline
 & & & & c & & & \\
 & & & & & 1 & & \\
\hline
 & & & & & & c & \\
 & & & & & & & 1
\end{array}\right],$$

where $c := 1000$. The properties of A are listed as Table 5. We choose the same tearing strategy and the same right hand side $\mathbf{b}$ as before, and the resulting residuals

TABLE 5. Properties of the Class of M-Matrices.

N	$\|A\|$	$\|A^{-1}\hat{A}\|$	$\|\hat{A}^{-1}\|$	$\|D^{-1}\hat{A}^{-1}AD\|$	$\|\hat{A}-\hat{A}A^{-1}\hat{A}\|$
4	$7.928 \cdot 10^{2}$	$1.158 \cdot 10^{3}$	$7.547 \cdot 10^{2}$	1.303	$8.387 \cdot 10^{2}$
6	$8.830 \cdot 10^{2}$	$1.445 \cdot 10^{3}$	$2.167 \cdot 10^{3}$	1.349	$9.549 \cdot 10^{2}$
8	$9.209 \cdot 10^{2}$	$1.614 \cdot 10^{3}$	$4.028 \cdot 10^{3}$	1.396	$1.001 \cdot 10^{3}$
10	$9.403 \cdot 10^{2}$	$1.712 \cdot 10^{3}$	$6.077 \cdot 10^{3}$	1.431	$1.021 \cdot 10^{3}$
12	$9.515 \cdot 10^{2}$	$1.768 \cdot 10^{3}$	$8.130 \cdot 10^{3}$	1.453	$1.029 \cdot 10^{3}$
14	$9.586 \cdot 10^{2}$	$1.799 \cdot 10^{3}$	$1.007 \cdot 10^{4}$	1.466	$1.033 \cdot 10^{3}$
16	$9.633 \cdot 10^{2}$	$1.815 \cdot 10^{3}$	$1.182 \cdot 10^{4}$	1.474	$1.035 \cdot 10^{3}$
18	$9.666 \cdot 10^{2}$	$1.823 \cdot 10^{3}$	$1.338 \cdot 10^{4}$	1.478	$1.035 \cdot 10^{3}$
20	$9.691 \cdot 10^{2}$	$1.828 \cdot 10^{3}$	$1.474 \cdot 10^{4}$	1.480	$1.036 \cdot 10^{3}$
22	$9.709 \cdot 10^{2}$	$1.830 \cdot 10^{3}$	$1.591 \cdot 10^{4}$	1.481	$1.036 \cdot 10^{3}$

TABLE 6. Residuals for the Class of M-Matrices.

		$\|A\mathbf{x}-\mathbf{b}\|$		$\frac{\|A\mathbf{x}-\mathbf{b}\|}{\|A\|\,\|\mathbf{x}\|}$	
N	h	Algorithm 1	Gaussian elimination	Algorithm 1	Gaussian elimination
4	1	$1.017 \cdot 10^{-12}$	$4.441 \cdot 10^{-16}$	$1.570 \cdot 10^{-19}$	$6.859 \cdot 10^{-23}$
6	2	$4.547 \cdot 10^{-12}$	$9.095 \cdot 10^{-13}$	$2.087 \cdot 10^{-19}$	$4.173 \cdot 10^{-20}$
8	3	$5.457 \cdot 10^{-12}$	$1.819 \cdot 10^{-12}$	$1.100 \cdot 10^{-19}$	$3.667 \cdot 10^{-20}$
10	4	$7.500 \cdot 10^{-12}$	$1.819 \cdot 10^{-12}$	$8.214 \cdot 10^{-20}$	$1.992 \cdot 10^{-20}$
12	5	$1.373 \cdot 10^{-11}$	$8.702 \cdot 10^{-15}$	$9.347 \cdot 10^{-20}$	$5.923 \cdot 10^{-23}$
14	6	$1.925 \cdot 10^{-11}$	$3.638 \cdot 10^{-12}$	$8.923 \cdot 10^{-20}$	$1.686 \cdot 10^{-20}$
16	7	$4.002 \cdot 10^{-11}$	$3.638 \cdot 10^{-12}$	$1.349 \cdot 10^{-19}$	$1.226 \cdot 10^{-20}$
18	8	$4.426 \cdot 10^{-11}$	$5.087 \cdot 10^{-14}$	$1.139 \cdot 10^{-19}$	$1.309 \cdot 10^{-22}$
20	9	$7.570 \cdot 10^{-11}$	$7.276 \cdot 10^{-12}$	$1.544 \cdot 10^{-19}$	$1.484 \cdot 10^{-20}$
22	10	$8.678 \cdot 10^{-11}$	$2.183 \cdot 10^{-11}$	$1.444 \cdot 10^{-19}$	$3.632 \cdot 10^{-20}$

are presented as Table 6.

For the second and third class of test problems the condition (69) always holds. For these cases, the results in Table 4 and 6 confirm that Algorithm 1 can compute a solution with a residual whose norm is on the same order as the norm of the residual from Gaussian elimination.

10. CONCLUSIONS

We have presented an error analysis for a divide-and-conquer algorithm to solve linear systems with block Hessenberg matrices. Our error analysis corresponds closely to the recursive nature of this algorithm. The key to our analysis is equation (10) which gives a representation for the residuals ΔB in the tear tree. Another important equation is (51) which shows the structure of the residuals in terms of local errors and errors from previous nodes in the tear tree. By combining (10) and (51) we can derive the linear recurrence relationships for ξ_k and ζ_k which lead to the final error bound (68).

The precise value of the bound (68) is of minor importance. Rather we can show that Algorithm 1 computes a backward stable solution if the condition (69) is satisfied for all the matrices in the tear tree. This condition ensures that all the quantities in Algorithm 1 remain bounded. In particular equation (55) shows that the matrix $D^{-1}Y$ can never be much larger than $D^{-1}X$. On the other hand, if $\|D^{-1}\hat{A}^{-1}AD\|$ is significantly larger than one for all the nodes in the tear tree, we may encounter very large matrices $D^{-1}Y$ during the execution of Algorithm 1, even if the final result $D^{-1}X$ is small. This is exactly what happens in our first test case in Section 9, leading to a large residual ΔB.

We have also shown that the condition (69) is always satisfied in the case of diagonally dominant matrices and M-matrices. This explains the accurate results of the algorithm for this type of problems.

References

1. A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Academic Press, New York, 1979.
2. T. F. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl. **88/89** (1987), 67–82.
3. B. Char, K. Geddes, G. Gonnet, B. Leong, M. Monagan, and S. Watt, *Maple V language reference manual*, Springer, New York, 1991.
4. G. H. Golub and C. F. Van Loan, *Matrix computations*, second ed., The Johns Hopkins University Press, Baltimore, 1989.
5. N. J. Higham, *How accurate is Gaussian elimination?*, Numerical Analysis 1989, Proceedings of the 13th Dundee Conference (D. F. Griffiths and G. A. Watson, eds.), Longman Scientific and Technical, 1990, pp. 137–154.
6. H. Minc, *Nonnegative matrices*, John Wiley & Sons, New York, 1988.
7. G. W. Stewart, *On the solution of block Hessenberg systems*, Tech. Report CS-TR-2973, Department of Computer Science, University of Maryland, October 1992, to appear in Numerical Linear Algebra and Applications.
8. ———, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Processing **40** (1992), 1535–1541.
9. ———, *Implementing an algorithm for solving block Hessenberg systems*, Tech. Report CS-TR-3295, Department of Computer Science, University of Maryland, June 1994.
10. The MathWorks Inc., *MATLAB, high-performance numeric computation and visualization software*, Natick, Massachusetts, 1992.
11. J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach. **8** (1961), 281–330.
12. ———, *Rounding errors in algebraic processes*, Prentice-Hall, New Jersey, 1963.
13. ———, *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965.

Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

E-mail address: na.vonmatt@na-net.ornl.gov

Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

E-mail address: stewart@cs.umd.edu