STABLE COMPUTATION OF THE CS DECOMPOSITION: SIMULTANEOUS BIDIAGONALIZATION

BRIAN D. SUTTON*

AMS subject classifications. Primary 65F15, 15A18, 15A23, 15B10; Secondary 65F25

Abstract. Since its discovery in 1977, the CS decomposition has resisted computation, even though it is a sibling of the well-understood eigenvalue and singular value decompositions. Several algorithms have been developed for the reduced 2-by-1 form of the decomposition, but none have been extended to the complete 2-by-2 form of the decomposition in Stewart's original paper. In this article, we present an algorithm for simultaneously bidiagonalizing the four blocks of a unitary matrix partitioned into a 2-by-2 block structure. This serves as the first, direct phase of a two-stage algorithm for the CSD, much as Golub-Kahan-Reinsch bidiagonalization serves as the first stage in computing the singular value decomposition. Backward stability is proved.

1. Introduction. The CS decomposition presents unique obstacles to computation, even though it is a sibling of the well-understood eigenvalue and singular value decompositions. In this article, we present a new algorithm for the CS decomposition (CSD) and prove its numerical stability.

The decomposition, discovered by Stewart in 1977, building on earlier work of Davis and Kahan, simultaneously diagonalizes the blocks of a unitary matrix partitioned into a 2-by-2 block structure [4, 5, 15]. In the special case when all four blocks have the same number of rows and columns, the form is the following:

$$\underset{\text{unitary}}{X} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} = \underbrace{\begin{bmatrix} U_1 & \\ & U_2 \end{bmatrix}}_{\text{unitary}} \underbrace{\begin{bmatrix} C & -S \\ S & C \end{bmatrix}}_{\substack{\text{orthogonal,} \\ C,\, S \text{ diag'l, nonneg.}}} \underbrace{\begin{bmatrix} V_1 & \\ & V_2 \end{bmatrix}^*}_{\text{unitary}}. \tag{1.1}$$

The diagonal entries of C and S must be the *cosines* and *sines* of angles in $[0, \frac{\pi}{2}]$, revealing the origin of the name *CS decomposition*. Applications include principal angles in higher-dimensional Euclidean geometry, perturbation of linear subspaces, canonical correlations in multivariate statistics, existence and computation of the generalized singular value decomposition (GSVD), and reduction of quantum computer programs to quantum logic gates. The survey of Paige and Wei is a good introduction [14].

The CSD has long resisted computation. As seen in (1.1), the decomposition is equivalent to four simultaneous but highly interrelated singular value decompositions (SVD's), $X_{ij} = U_i \Sigma_{ij} V_j^*$, with Σ_{ij} equaling C or $\pm S$ as appropriate. The extensive sharing of singular vectors explains the difficulty in computation. If X is measured imperfectly to obtain $\tilde{X} \approx X$, not exactly unitary, then the goal should be to find a nearby matrix $\tilde{X} + \Delta\tilde{X}$ that is exactly unitary and to compute the CSD of that matrix. However, if $\tilde{X}$ has two nearly equal singular values, then the corresponding singular vectors can be chosen nearly arbitrarily from a certain subspace. In computing the four SVD's of (1.1), great care must be taken to choose the singular vectors consistently across all four blocks, even if they are nearly arbitrary in any one block.

Perhaps for this reason, but also because of its connection with the generalized singular value decomposition, some articles consider the *2-by-1 CS decomposition*:

$$\begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} = \begin{bmatrix} U_1 & \\ & U_2 \end{bmatrix} \begin{bmatrix} C \\ S \end{bmatrix} V_1^*.$$

*Randolph-Macon College, Ashland, Virginia. This material is based upon work supported by the National Science Foundation under Grant No. DMS-0914559 and a Walter Williams Craigie Grant.

We call the original form (1.1) simply "the CS decomposition" or sometimes the *2-by-2 CS decomposition* for emphasis. Many articles on theory and applications consider the 2-by-2 CSD [1, 9, 11, 14, 15, 17, 22, 24, 26], while articles on computation have focused on the 2-by-1 CSD or GSVD [2, 3, 12, 16, 23, 24].

In [19], we presented a new algorithm for computing the complete 2-by-2 CSD, unifying theory and computation. The ideas built on earlier observations from random matrix theory [6, 18]. In the present article, we prove the numerical stability of the key ingredient, reduction to bidiagonal-block form. We also present a modified algorithm that is numerically equivalent but more illuminating analytically, as it collects the backward errors into a triangular matrix that is easily measured.

2. Bidiagonal-block form. The CSD is more general than (1.1) suggests [13]. Let X be any m-by-m unitary matrix, and partition it as

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}.$$

Let the top-left block be p-by-q. For convenience, we permute rows and columns and adjust signs to present the CSD as follows:

$$\left[\begin{array}{c|c} U_1 & \\ \hline & U_2 \end{array}\right]^* \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c|c} V_1 & \\ \hline & V_2 \end{array}\right] = \left[\begin{array}{ccc|ccc} C & & & & & -SF \\ & I & & & 0 & \\ & & 0 & I & & \\ \hline & & I & 0 & & \\ & 0 & & & I & \\ FS & & & & & FCF \end{array}\right].$$

C and S are r-by-r diagonal matrices, in which r is the least among p, $m-p$, q, and $m-q$. The "flip" permutation matrix F has 1's along its antidiagonal. The four I's represent square identity matrices of various sizes, while the 0's represent rectangular blocks of zeros. Some or all of these may not be present for specific values of m, p, and q. In particular, if $m = 2p = 2q$, then none of the identity or zeros blocks appear.

As an intermediate step to computing the CSD, we simultaneously bidiagonalize the four blocks of X.

THEOREM 2.1. *Let*

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

be a unitary matrix. Then there exist unitary U_1, U_2, V_1, and V_2 that simultaneously bidiagonalize the four blocks of X:

$$\left[\begin{array}{c|c} U_1 & \\ \hline & U_2 \end{array}\right]^* \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c|c} V_1 & \\ \hline & V_2 \end{array}\right] = \left[\begin{array}{ccc|ccc} B_{11} & & & & & B_{12} \\ & I & & & 0 & \\ & & 0 & I & & \\ \hline & & I & 0 & & \\ & 0 & & & I & \\ B_{21} & & & & & B_{22} \end{array}\right]. \tag{2.1}$$

The multiplication is conformable at the block level, and the orthogonal matrix $\left[\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right]$, said to be in bidiagonal-block form*, can be parameterized by $\theta_1, \ldots, \theta_r \in [0, \frac{\pi}{2}]$ and $\phi_1, \ldots, \phi_{r-1} \in [0, \frac{\pi}{2}]$ as in Figure 2.1.*

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$= \left[\begin{array}{cccccc|cccccc}
c_1 & s_1 s_1' & & & & & & & & & & -s_1 c_1' \\
 & c_2 c_1' & s_2 s_2' & & & & & & & -s_2 c_2' & c_2 s_1' & \\
 & & c_3 c_2' & \ddots & & & & & -s_3 c_3' & c_3 s_2' & & \\
 & & & \ddots & s_{r-1} s_{r-1}' & & & \iddots & \iddots & & & \\
 & & & & c_r c_{r-1}' & & -s_r & c_r s_{r-1}' & & & & \\
\hline
 & & & & s_r c_{r-1}' & & c_r & s_r s_{r-1}' & & & & \\
 & & & \iddots & -c_{r-1} s_{r-1}' & & & \ddots & \ddots & & & \\
 & & s_3 c_2' & \iddots & & & & & c_3 c_3' & s_3 s_2' & & \\
 & s_2 c_1' & -c_2 s_2' & & & & & & & c_2 c_2' & s_2 s_1' & \\
s_1 & -c_1 s_1' & & & & & & & & & & c_1 c_1'
\end{array}\right],$$

$$c_i = \cos\theta_i,\ s_i = \sin\theta_i,\ c_i' = \cos\phi_i,\ s_i' = \sin\phi_i.$$

FIGURE 2.1. *Bidiagonal-block form. Any matrix of this form is orthogonal, and any partitioned unitary matrix can be reduced to this form.*

The proof is constructive. It can be found in [6, 18, 19], and §§3, 4, and 6 of the present article follow essentially the same path.

The primary aim of this article is to prove backward stability of simultaneous bidiagonalization. The proof was first announced at the Fourteenth Leslie Fox Prize meeting, where it earned first place [20].

We add a few comments before commencing the description of the algorithm and the proof of stability.

Of course, the matrix $\left[\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right]$ of Theorem 2.1 must be unitary because it is a unitary reduction of a unitary input. However, a more direct argument shows that the form of Figure 2.1 itself guarantees orthogonality: The matrix can be expressed as

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = (G_1 \cdots G_r)\,(H_1 \cdots H_{r-1})^T, \tag{2.2}$$

in which G_i is a Givens rotation acting on indices i and $2r+1-i$ by the angle θ_i and H_i is a Givens rotation acting on indices $i+1$ and $2r+1-i$ by the angle ϕ_i.

The bidiagonal structure of (2.1) was apparently first noted by David Watkins in the context of a Lanczos-type iteration [25]. The parameterization by θ_i and ϕ_i first appeared in the present author's 2005 Ph.D. thesis [18]. It enables exactly orthogonal matrices to be manipulated on a finite-precision computer. The factorization (2.2) is new in this article and proves to be an important part of the stability analysis.

Simultaneous bidiagonalization is the first half of a two-phase procedure for the CSD. Simultaneous QR iteration can finish the job [19].

3. Algorithm sketch and main theorem. For simplicity, the rest of the article dispenses with complex numbers. Also, r is assumed to equal q in Theorem 2.1,

specializing the decomposition to

$$\left[\begin{array}{c|c} U_1 & \\ \hline & U_2 \end{array}\right]^* \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c|c} V_1 & \\ \hline & V_2 \end{array}\right] = \left[\begin{array}{c|cc} B_{11} & & B_{12} \\ & I & \\ \hline & I & \\ B_{21} & & B_{22} \end{array}\right]. \tag{3.1}$$

Computing this is the main goal.

The algorithm of our 2009 paper is inspired by Golub-Kahan-Reinsch bidiagonalization [7, 8]; see Figure 3.1. In the end, X is mapped to the matrix in bidiagonal-block form

$$(P_q \cdots P_1) X (Q_1 \cdots Q_{m-q}),$$

by matrices P_i and Q_i which are constructed from Householder reflectors,

$$P_i = \underbrace{I_{i-1} \oplus (I - u_1^{(i)} u_1^{(i)T})}_{p \times p} \oplus \underbrace{(I - u_2^{(i)} u_2^{(i)T}) \oplus I_{i-1}}_{(m-p)\times(m-p)}, \quad i = 1, \ldots, q, \tag{3.2}$$

$$Q_i = \underbrace{I_i \oplus (I - v_1^{(i)} v_1^{(i)T})}_{q \times q} \oplus \underbrace{(I - v_2^{(i)} v_2^{(i)T}) \oplus I_{i-1}}_{(m-q)\times(m-q)}, \quad i = 1, \ldots, q-1, \tag{3.3}$$

$$Q_i = I_q \oplus \underbrace{(I - v_2^{(i)} v_2^{(i)T}) \oplus I_{i-1}}_{(m-q)\times(m-q)}, \quad i = q, \ldots, m-q. \tag{3.4}$$

($A \oplus B$ denotes the block-diagonal matrix $\left[\begin{smallmatrix} A & \\ & B \end{smallmatrix}\right]$.) This reduction would not be possible for a general matrix, but orthogonality guarantees that certain intermediate rows and columns are colinear, allowing a single reflector to send two vectors to the same coordinate axis simultaneously. In the rest of the article, the algorithm suggested by Figure 3.1 is called the *original algorithm*.

For the stability analysis, a reorganization is helpful. The *modified algorithm* is illustrated in Figure 3.2. Matrices P_i and Q_i are constructed from Householder reflectors as before, but, surprisingly, the Givens rotations G_i, H_i violate the block structure:

$$G_i([i, m-i+1], [i, m-i+1]) = \begin{bmatrix} c_i & -s_i \\ s_i & c_i \end{bmatrix}, \quad i = 1, \ldots, q, \tag{3.5}$$

$$H_i([i+1, m-i+1], [i+1, m-i+1]) = \begin{bmatrix} c_i' & -s_i' \\ s_i' & c_i' \end{bmatrix}, \quad i = 1, \ldots, q-1. \tag{3.6}$$

These violations are later inverted "on paper" to recover the matrix in bidiagonal-block form. In fact, the modified algorithm is numerically equivalent to the original algorithm, requiring identical floating-point operations and thus experiencing identical roundoff errors.

Given an arbitrary m-by-m matrix A, the algorithm computes

$$\tilde{U}^T A \tilde{V} = R$$

in which R is upper-triangular with nonnegative diagonal and

$$\tilde{U} = (P_1 G_1)(P_2 G_2) \cdots (P_q G_q), \tag{3.7}$$

$$\tilde{V} = (Q_1 H_1)(Q_2 H_2) \cdots (Q_{q-1} H_{q-1}) Q_q \cdots Q_{m-q} \tag{3.8}$$

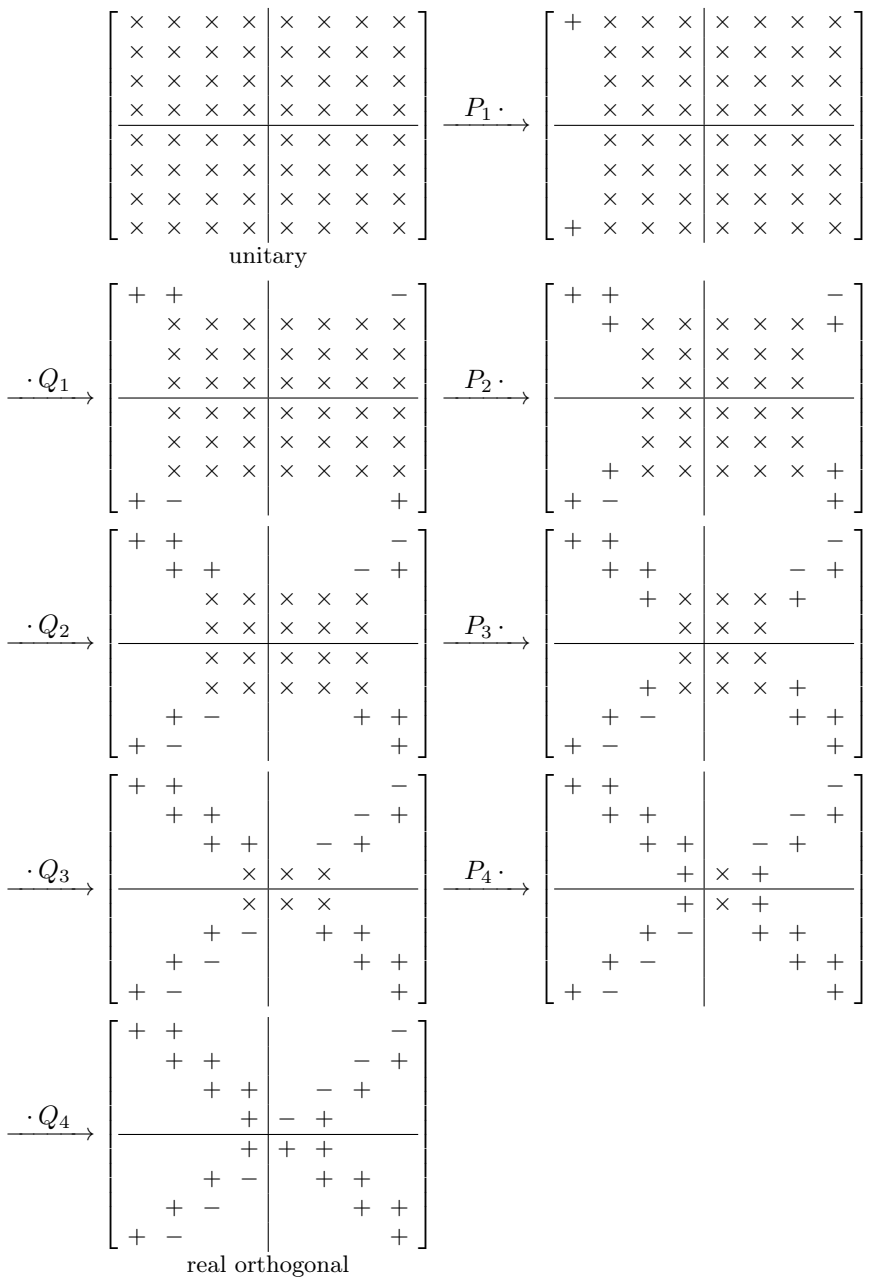

FIGURE 3.1. *The original algorithm [19]. Each P_i and Q_i is a pair of Householder reflectors. If the original matrix were exactly unitary and arithmetic computations were exact, then the entries identified by pluses and minuses would be products of cosines and sines, and the blank entries would equal zero. In practice, they would not be exact and would have to be rounded.*

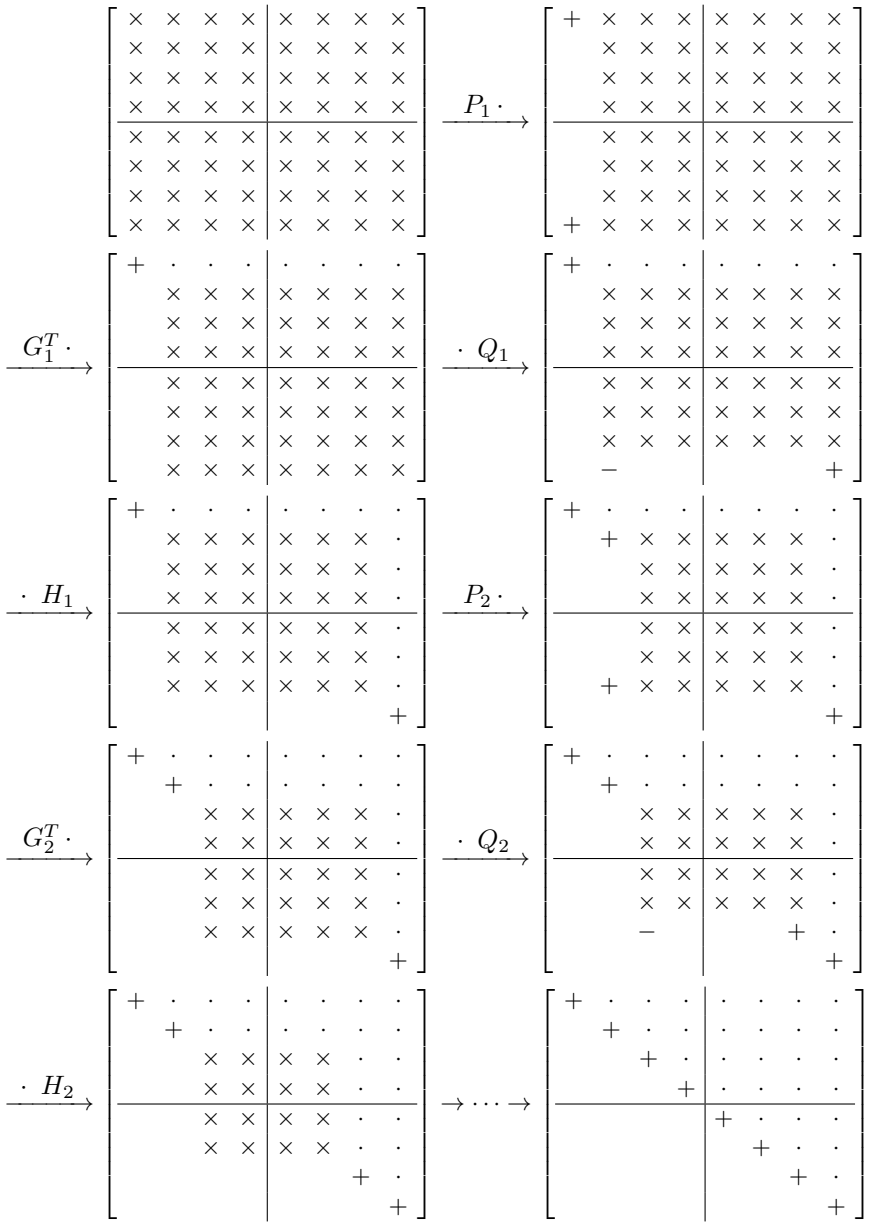

FIGURE 3.2. *The modified algorithm. For arbitrary input, the output is upper-triangular. For unitary input, the output is the identity matrix. The original algorithm of Figure 3.1 is recovered by inverting the Givens rotations G_i, H_i at the very end.*

are orthogonal. (The tildes are present on $\tilde{U}$ and $\tilde{V}$ to distinguish these matrices from $\left[\begin{smallmatrix} U_1 & \\ & U_2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} V_1 & \\ & V_2 \end{smallmatrix}\right]$ of (2.1).)

If the input matrix A is in fact a nearly orthogonal X, then the upper-triangular R must be nearly orthogonal as well and hence must be approximately equal to the identity matrix. We shall see that the error can be thrown onto the input matrix to achieve

$$\tilde{U}^T(X + \Delta X)\tilde{V} = I$$

for a small perturbation ΔX. Expanding the orthogonal transformations gives

$$\left(G_q^T P_q\right) \cdots \left(G_1^T P_1\right) (X + \Delta X) \left(Q_1 H_1\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q} = I.$$

Then, noting that certain orthogonal transformations commute because they operate on disjoint rows and columns, we find

$$\left(P_1 \cdots P_q\right)^T (X + \Delta X) \left(Q_1 \cdots Q_{m-q}\right) = \left(G_1 \cdots G_q\right) \left(H_1 \cdots H_{q-1}\right)^T. \tag{3.9}$$

Considering (2.2), this is a backward stable computation of (3.1).

The following is the main theorem of the article. As elaborated in Appendix A, $\mathbf{u}$ is unit roundoff, γ_n equals $n\mathbf{u}/(1 - n\mathbf{u})$, $\tilde{\gamma}_n$ equals $cn\mathbf{u}/(1 - cn\mathbf{u})$ for a small constant c, $\hat{z}$ denotes the computed approximation to a quantity z, and the notation $|\cdot|$ is used for componentwise error bounds. Also, k is a a small constant defined at the beginning of §5.

THEOREM 3.1. *Reduction to bidiagonal-block form is backward stable. Given an m-by-m nearly orthogonal matrix*

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \quad \left\|I - X^T X\right\|_2 = \varepsilon \leq \frac{1}{4},$$

the algorithm finds B_{ij}, $i, j = 1, 2$, and orthogonal U_1, U_2, V_1, and V_2 for which

$$\begin{bmatrix} U_1 & \\ & U_2 \end{bmatrix}^T (X + \Delta X) \begin{bmatrix} V_1 & \\ & V_2 \end{bmatrix} = \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline & I \\ \hline & I \\ \hline B_{21} & B_{22} \end{array}\right],$$

$$\left\|\Delta X\right\|_F \leq \sqrt{m}\left(\varepsilon + 7\gamma_{km^2}\right).$$

The bidiagonal blocks are represented implicitly by c_i, s_i, c_i', and s_i'—see Figure 2.1—which are computed to high relative precision:

$$\begin{aligned} \hat{c}_i &= c_i(1 + \eta_4), & \hat{s}_i &= s_i(1 + \eta_4'), & i &= 1, \ldots, q, \\ \hat{c}_i' &= c_i'(1 + \eta_4''), & \hat{s}_i' &= s_i'(1 + \eta_4'''), & i &= 1, \ldots, q - 1, \end{aligned}$$

all four error terms bounded by γ_4. The orthogonal transformations U_1, U_2, V_1, and V_2 are products of Householder reflectors—see (3.9) *and* (3.2)–(3.4)*—which are computed to high relative precision:*

$$\begin{aligned} \hat{u}_1^{(i)} &= u_1^{(i)} + \Delta u_1^{(i)}, & \left|\Delta u_1^{(i)}\right| &\leq \tilde{\gamma}_{p-i+1} \left|u_1^{(i)}\right|, & i &= 1, \ldots, q, \\ \hat{u}_2^{(i)} &= u_2^{(i)} + \Delta u_2^{(i)}, & \left|\Delta u_2^{(i)}\right| &\leq \tilde{\gamma}_{m-p-i+1} \left|u_2^{(i)}\right|, & i &= 1, \ldots, q, \\ \hat{v}_1^{(i)} &= v_1^{(i)} + \Delta v_1^{(i)}, & \left|\Delta v_1^{(i)}\right| &\leq \tilde{\gamma}_{q-i} \left|v_1^{(i)}\right|, & i &= 1, \ldots, q - 1, \\ \hat{v}_2^{(i)} &= v_2^{(i)} + \Delta v_2^{(i)}, & \left|\Delta v_2^{(i)}\right| &\leq \tilde{\gamma}_{m-q-i+1} \left|v_2^{(i)}\right|, & i &= 1, \ldots, m - q. \end{aligned}$$

The theorem is proved at the end of §7. The proof places a mild assumption on the size of m relative to $\mathbf{u}$ and ε.

4. Triangularization of an arbitrary matrix. As outlined in Figure 3.2, we want to triangularize an arbitrary m-by-m matrix A in a way that mostly respects but sometimes carefully violates a 2-by-2 block partitioning. Only in §6 do we use a nearly orthogonal matrix X in place of the general matrix A.

This section assumes exact arithmetic and prepares notation for the next section, which considers numerical stability. The algorithm uses routines **houseg**, **housea**, **givensg**, and **givensa** from Appendix B.

The triangularization is considered inductively.

4.1. Base case. For the base case, assume $q = 1$ and write

$$A = \left[\begin{array}{c|c} a_{11} & A_{12} \\ A_{21} & A_{22} \\ \hline A_{31} & A_{32} \\ a_{41} & A_{42} \end{array}\right].$$

(Note that A_{12} and A_{42} are row vectors and that A_{21} and A_{31} are column vectors. Capital letters are used for matrices, row vectors, and column vectors. Context clarifies.)

To begin the reduction to triangular form, compute

$$(u_1, b_{11}) = \mathbf{houseg}\left(\begin{bmatrix} a_{11} \\ A_{21} \end{bmatrix}\right), \quad (Fu_2, b_{41}) = \mathbf{houseg}\left(F\begin{bmatrix} A_{31} \\ a_{41} \end{bmatrix}\right),$$

in which F is a permutation matrix with 1's along its antidiagonal, and apply,

$$\left[\begin{array}{c} B_{12} \\ B_{22} \\ \hline B_{32} \\ B_{42} \end{array}\right] = \left[\begin{array}{c} \mathbf{housea}\left(u_1, \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}\right) \\ \hline \mathbf{housea}\left(u_2, \begin{bmatrix} A_{32} \\ A_{42} \end{bmatrix}\right) \end{array}\right].$$

In exact arithmetic, we have

$$PA = \left[\begin{array}{c|c} b_{11} & B_{12} \\ 0 & B_{22} \\ \hline 0 & B_{32} \\ b_{41} & B_{42} \end{array}\right] =: B, \tag{4.1}$$

in which $P = (I - u_1 u_1^T) \oplus (I - u_2 u_2^T)$.

Next, compute

$$(c, s, d_{11}) = \mathbf{givensg}(b_{11}, b_{41}),$$

and apply,

$$\left[\begin{array}{c} D_{12} \\ \hline D_{42} \end{array}\right] = \mathbf{givensa}\left(c, s, \left[\begin{array}{c} B_{12} \\ \hline B_{42} \end{array}\right]\right).$$

In exact arithmetic, we have

$$\left(G^T P\right) A = \left[\begin{array}{c|c} d_{11} & D_{12} \\ 0 & B_{22} \\ \hline 0 & B_{32} \\ 0 & D_{42} \end{array}\right] =: D, \tag{4.2}$$

in which G is the Givens rotation with $G(1,1) = G(m,m) = c$, $G(1,m) = -s$, and $G(m,1) = s$. (Eventually, A will be replaced by an orthogonal matrix X, forcing $d_{11} = 1$ and the rest of the first row to be zero. Computing the first row will ultimately prove to be superfluous but benign.)

Finally, compute an RQ decomposition of

$$\begin{bmatrix} B_{22} \\ B_{32} \\ D_{42} \\ D_{12} \end{bmatrix}.$$

The orthogonal factor can be expressed as a product of $m-1$ Householder reflectors. Applying the reflectors in reverse order to (4.2), we reduce the input to R,

$$\left(G^T P\right) A(Q_1 \cdots Q_{m-1}) = \left[\begin{array}{c|c} d_{11} & E_{12} \\ 0 & E_{22} \\ \hline 0 & E_{32} \\ 0 & E_{42} \end{array}\right] =: R, \tag{4.3}$$

which is square and upper-triangular.

4.2. Recursion step. Let A be an m-by-m matrix partitioned so that its top-left block is p-by-q. For the induction step, assume $q \geq 2$ and write

$$A = \left[\begin{array}{ccc|cc} a_{11} & a_{12} & A_{13} & A_{14} & a_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ \hline A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \\ a_{41} & a_{42} & A_{43} & A_{44} & a_{45} \end{array}\right].$$

(If $q = 2$, then $A_{i,3}$, $i = 1, \ldots, 4$, are not present. This does not affect the argument below.) The goal is to transform the matrix, using orthogonal transformations that respect the partitioning in a certain way, to upper-triangular form.

First, compute

$$(u_1, b_{11}) = \mathbf{houseg}\left(\begin{bmatrix} a_{11} \\ A_{21} \end{bmatrix}\right), \quad (Fu_2, b_{41}) = \mathbf{houseg}\left(F \begin{bmatrix} A_{31} \\ a_{41} \end{bmatrix}\right),$$

and apply,

$$\left[\begin{array}{cc|cc} b_{12} & B_{13} & B_{14} & b_{15} \\ B_{22} & B_{23} & B_{24} & B_{25} \\ \hline B_{32} & B_{33} & B_{34} & B_{35} \\ b_{42} & B_{43} & B_{44} & b_{45} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{housea}\left(u_1, \left[\begin{array}{cc|cc} a_{12} & A_{13} & A_{14} & a_{15} \\ A_{22} & A_{23} & A_{24} & A_{25} \end{array}\right]\right) \\ \hline \mathbf{housea}\left(u_2, \left[\begin{array}{cc|cc} A_{32} & A_{33} & A_{34} & A_{35} \\ a_{42} & A_{43} & A_{44} & a_{45} \end{array}\right]\right) \end{array}\right].$$

In exact arithmetic, we have

$$P_1 A = \left[\begin{array}{ccc|cc} b_{11} & b_{12} & B_{13} & B_{14} & b_{15} \\ 0 & B_{22} & B_{23} & B_{24} & B_{25} \\ \hline 0 & B_{32} & B_{33} & B_{34} & B_{35} \\ b_{41} & b_{42} & B_{43} & B_{44} & b_{45} \end{array}\right] =: B, \tag{4.4}$$

in which $P_1 = (I - u_1 u_1^T) \oplus (I - u_2 u_2^T)$.

Next, compute

$$(c, s, d_{11}) = \mathbf{givensg}(b_{11}, b_{41}),$$

and apply,

$$\left[\begin{array}{cc|cc} d_{12} & D_{13} & D_{14} & d_{15} \\ \hline d_{42} & D_{43} & D_{44} & d_{45} \end{array}\right] = \mathbf{givensa}\left(c, s, \left[\begin{array}{cc|cc} b_{12} & B_{13} & B_{14} & b_{15} \\ \hline b_{42} & B_{43} & B_{44} & b_{45} \end{array}\right]\right).$$

In exact arithmetic, we have

$$\left(G_1^T P_1\right) A = \left[\begin{array}{ccc|cc} d_{11} & d_{12} & D_{13} & D_{14} & d_{15} \\ 0 & B_{22} & B_{23} & B_{24} & B_{25} \\ \hline 0 & B_{32} & B_{33} & B_{34} & B_{35} \\ 0 & d_{42} & D_{43} & D_{44} & d_{45} \end{array}\right] := D, \tag{4.5}$$

in which G_1 is a Givens rotation with $G_1(1,1) = G_1(m,m) = c$, $G_1(1,m) = -s$, and $G_1(m,1) = s$. (Later in the article, A will be replaced by an orthogonal X, forcing $d_{11} = 1$ and the rest of the first row to be zero. For now, A is arbitrary.)

Now, working from the right, compute

$$(v_1, -e_{42}) = \mathbf{houseg}\left(-\left[\begin{array}{c} d_{42} \\ D_{43}^T \end{array}\right]\right), \quad (Fv_2, e_{45}) = \mathbf{houseg}\left(F\left[\begin{array}{c} D_{44}^T \\ d_{45} \end{array}\right]\right),$$

and apply,

$$\left[\begin{array}{cc|cc} e_{12} & E_{13} & E_{14} & e_{15} \\ E_{22} & E_{23} & E_{24} & E_{25} \\ \hline E_{32} & E_{33} & E_{34} & E_{35} \end{array}\right]$$
$$= \left[\begin{array}{c|c} \mathbf{housea}\left(v_1, \left[\begin{array}{cc} d_{12} & D_{13} \\ \hline B_{32} & B_{33} \end{array}\right]^T\right)^T & \mathbf{housea}\left(v_2, \left[\begin{array}{cc} D_{14} & d_{15} \\ \hline B_{34} & B_{35} \end{array}\right]^T\right)^T \end{array}\right].$$

In exact arithmetic, we have

$$\left(G_1^T P_1\right) A Q_1 = \left[\begin{array}{ccc|cc} d_{11} & e_{12} & E_{13} & E_{14} & e_{15} \\ 0 & E_{22} & E_{23} & E_{24} & E_{25} \\ \hline 0 & E_{32} & E_{33} & E_{34} & E_{35} \\ 0 & e_{42} & 0 & 0 & e_{45} \end{array}\right] =: E, \tag{4.6}$$

in which $Q_1 = 1 \oplus (I - v_1 v_1^T) \oplus (I - v_2 v_2^T)$

Next, compute

$$(c', -s', f_{45}) = \mathbf{givensg}(e_{45}, e_{42}),$$

and apply,

$$\left[\begin{array}{c|c} f_{12} & f_{15} \\ \hline F_{22} & F_{25} \\ \hline F_{32} & F_{35} \end{array}\right] = \mathbf{givensa}\left(c', s', \left[\begin{array}{c|c} e_{12} & e_{15} \\ \hline E_{22} & E_{25} \\ \hline E_{32} & E_{35} \end{array}\right]^T\right)^T.$$

In exact arithmetic, we have

$$\left(G_1^T P_1\right) A \left(Q_1 H_1\right) = \left[\begin{array}{ccc|cc} d_{11} & f_{12} & E_{13} & E_{14} & f_{15} \\ 0 & F_{22} & E_{23} & E_{24} & F_{25} \\ \hline 0 & F_{32} & E_{33} & E_{34} & F_{35} \\ 0 & 0 & 0 & 0 & f_{45} \end{array}\right] =: F, \tag{4.7}$$

in which H_1 is a Givens rotation with $H_1(2,2) = H_1(m,m) = c'$, $H_1(2,m) = -s'$, and $H_1(m,2) = s'$. (If A is orthogonal, then $f_{45} = 1$, and f_{15}, F_{25}, and F_{35} must be zero.)

Finally, recursively execute this procedure on the square matrix

$$\left[\begin{array}{cc|c} F_{22} & E_{23} & E_{24} \\ \hline F_{32} & E_{33} & E_{34} \end{array}\right]$$

to find upper-triangular

$$\left(\tilde{G}_q^T \tilde{P}_q\right) \cdots \left(\tilde{G}_2^T \tilde{P}_2\right) \left[\begin{array}{cc|c} F_{22} & E_{23} & E_{24} \\ \hline F_{32} & E_{33} & E_{34} \end{array}\right] \left(\tilde{Q}_2 \tilde{H}_2\right) \cdots \left(\tilde{Q}_{q-1} \tilde{H}_{q-1}\right) \tilde{Q}_q \cdots \tilde{Q}_{m-q}$$
$$= \left[\begin{array}{cc|c} K_{22} & K_{23} & K_{24} \\ \hline 0 & 0 & K_{34} \end{array}\right].$$

(Keep in mind that this matrix is square and that $\left[\begin{array}{cc} K_{22} & K_{23} \end{array}\right]$ has no more columns than K_{24}.) All together, letting Z denote $1 \oplus \tilde{Z} \oplus 1$ where appropriate, we have

$$\left(G_q^T P_q\right) \cdots \left(G_2^T P_2\right) \left(G_1^T P_1\right) A \left(Q_1 H_1\right) \left(Q_2 H_2\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q}$$
$$= \left[\begin{array}{ccc|cc} d_{11} & f_{12} & E_{13} & E_{14} & f_{15} \\ 0 & K_{22} & K_{23} & K_{24} & F_{25} \\ \hline 0 & 0 & 0 & K_{34} & F_{35} \\ 0 & 0 & 0 & 0 & f_{45} \end{array}\right] =: R,$$

which is upper-triangular (even though the irregular partitioning may suggest otherwise).

5. Numerical stability of triangularization. The preceding triangularization routine is backward stable because it is built entirely from orthogonal transformations. The rest of this section computes bounds on the backward error. The analysis is not so different from that of symmetric tridiagonalization, utilizing reflections and rotations from both sides. See, for example, [21].

The error bounds of Appendix B are used heavily. We find it convenient to introduce a constant k (already seen in the statement of Theorem 3.1) that is an upper bound on the implicit constants in (B.2), (B.4), and (B.5). The upper bounds in those inequalities become $\gamma_{km}\|A\|_F$, $\gamma_k\|A\|_F$, and $\gamma_{km^2}\|A\|_F$, respectively.

5.1. Base case. The following theory tracks §4.1 and uses the same notation. See Appendix A and the comments preceding Theorem 3.1 for other notation relating to floating-point computation.

LEMMA 5.1. *In floating-point,* (4.1) *is computed backward stably,*

$$P\left(A + \Delta A^{(1)}\right) = \hat{B}, \quad \left\|\Delta A^{(1)}\right\|_F \leq \gamma_{km}\|A\|_F.$$

(*P refers to the exact orthogonal transformation $(I - u_1u_1^T) \oplus (I - u_2u_2^T)$, while $\hat{B}$ is computed from $(I - \hat{u}_1\hat{u}_1^T) \oplus (I - \hat{u}_2\hat{u}_2^T)$.) In addition,*

$$\hat{u}_1 = u_1 + \Delta u_1,\ |\Delta u_1| \le \tilde{\gamma}_p\,|u_1|\,,$$
$$\hat{u}_2 = u_2 + \Delta u_2,\ |\Delta u_2| \le \tilde{\gamma}_{m-p}\,|u_2|\,.$$

Proof. Two applications of the backward error bound (B.2) give

$$\begin{aligned}\big\|\Delta A^{(1)}\big\|_F \le \sqrt{\gamma_{kp}^2\big\|A\big\|_F^2 + \gamma_{k(m-p)}^2\big\|A\big\|_F^2} &\le \sqrt{\gamma_{kp}^2 + \gamma_{k(m-p)}^2}\big\|A\big\|_F \\ &\le (\gamma_{kp} + \gamma_{k(m-p)})\big\|A\big\|_F \le \gamma_{km}\big\|A\big\|_F.\end{aligned}$$

The bounds on $\hat{u}_1$ and $\hat{u}_2$ are from (B.1). ∎

LEMMA 5.2. *In floating-point,* (4.2) *is computed backward stably,*

$$G^TP(A + \Delta A^{(2)}) = \hat{D}, \quad \big\|\Delta A^{(2)}\big\|_F \le \gamma_{k(m+1)}\big\|A\big\|_F.$$

Here, G refers to the exact Givens rotation for the computed $\hat{b}_{11}$, $\hat{b}_{41}$. In addition,

$$\hat{c} = c(1 + \eta_4), \quad |\eta_4| \le \gamma_4, \qquad \hat{s} = s(1 + \eta_4'), \quad |\eta_4'| \le \gamma_4.$$

Proof. The backward error analysis for Givens rotations and the previous lemma provide

$$G^T(\hat{B} + \Delta\hat{B}) = \hat{D}, \quad \big\|\Delta\hat{B}\big\|_F \le \gamma_k\big\|\hat{B}\big\|_F = \gamma_k\big\|A + \Delta A^{(1)}\big\|_F.$$

The total backward error is then bounded by letting $\Delta A^{(2)} = \Delta A^{(1)} + P\Delta\hat{B}$ so that

$$G^TP(A + \Delta A^{(2)}) = \hat{D}$$

with

$$\begin{aligned}\big\|\Delta A^{(2)}\big\|_F &\le \big\|\Delta A^{(1)}\big\|_F + \big\|P\Delta\hat{B}\big\|_F \\ &\le \big\|\Delta A^{(1)}\big\|_F + \gamma_k\big\|A + \Delta A^{(1)}\big\|_F \le \gamma_k\big\|A\big\|_F + (1 + \gamma_k)\big\|\Delta A^{(1)}\big\|_F \\ &\le \gamma_k\big\|A\big\|_F + (1 + \gamma_k)\gamma_{km}\big\|A\big\|_F \le \gamma_{k(m+1)}\big\|A\big\|_F.\end{aligned}$$

The bounds on $\hat{c}$ and $\hat{s}$ come from (B.3). ∎

LEMMA 5.3. *In floating-point,* (4.3) *is computed backward stably,*

$$G^TP(A + \Delta A^{(3)})(Q_1 \cdots Q_{m-1}) = \hat{R}, \quad \big\|\Delta A^{(3)}\big\|_F \le \gamma_{km^2}\big\|A\big\|_F.$$

Each Q_i has the form $1 \oplus (I_{m-i} - v_2^{(i)}v_2^{(i)T}) \oplus I_{i-1}$ and is approximated by

$$\hat{v}_2^{(i)} = v_2^{(i)} + \Delta v_2^{(i)},\ \big|\Delta v_2^{(i)}\big| \le \tilde{\gamma}_{m-i}\big|v_2^{(i)}\big|.$$

Proof. Computing the RQ decomposition backward stably gives

$$\big(\hat{D} + \Delta\hat{D}\big)(Q_1 \cdots Q_{m-1}) = \hat{R},$$

the backward error bounded by

$$\|\Delta\hat{D}\|_F \le \gamma_{k(m-1)^2}\|\hat{D}\|_F = \gamma_{k(m-1)^2}\|A+\Delta A^{(2)}\|_F.$$

Letting $\Delta A^{(3)} = \Delta A^{(2)} + PG\Delta\hat{D}$, we find

$$G^T P(A+\Delta A^{(3)})(Q_1\cdots Q_{m-1}) = \hat{R}$$

with

$$\begin{aligned}\|\Delta A^{(3)}\|_F &\le \|\Delta A^{(2)}\|_F + \|\Delta\hat{D}\|_F \\ &\le \gamma_{k(m-1)^2}\|A\|_F + (1+\gamma_{k(m-1)^2})\|\Delta A^{(2)}\|_F \\ &\le \left(\gamma_{k(m-1)^2} + \gamma_{k(m+1)}(1+\gamma_{k(m-1)^2})\right)\|A\|_F \\ &\le \gamma_{[k(m-1)^2+k(m+1)]}\|A\|_F \le \gamma_{km^2}\|A\|_F.\end{aligned}$$

(We used the fact that $m \ge 2$.) ∎

In summary, when $q = 1$, the algorithm computes

$$G^T P(A+\Delta A^{(3)})Q_1\cdots Q_{m-1} = \hat{R}$$

backward stably, and the Householder vectors and Givens angles are computed to high relative precision.

5.2. Recursion step. The following discussion tracks §4.2 and uses the same notation.

LEMMA 5.4. *In floating-point,* (4.4) *is computed backward stably,*

$$P_1(A+\Delta A^{(1)}) = \hat{B}, \quad \|\Delta A^{(1)}\|_F \le \gamma_{km}\|A\|_F.$$

Here, P_1 refers to the direct sum of exact Householder reflectors $(I-u_1u_1^T)\oplus(I-u_2u_2^T)$. It is represented in floating-point by $\hat{u}_1$ and $\hat{u}_2$ satisfying

$$\begin{aligned}\hat{u}_1 &= u_1 + \Delta u_1,\ |\Delta u_1| \le \tilde{\gamma}_p\,|u_1|\,, \\ \hat{u}_2 &= u_2 + \Delta u_2,\ |\Delta u_2| \le \tilde{\gamma}_{m-p}\,|u_2|\,.\end{aligned}$$

The proof is identical to the proof of Lemma 5.1.

LEMMA 5.5. *In floating-point,* (4.5) *is computed backward stably,*

$$G_1^T P_1(A+\Delta A^{(2)}) = \hat{D}, \quad \|\Delta A^{(2)}\|_F \le \gamma_{k(m+1)}\|A\|_F.$$

Here, G_1 refers to the exact Givens rotation for the computed $\hat{b}_{11}$, $\hat{b}_{41}$. In addition,

$$\hat{c} = c(1+\eta_4), \quad |\eta_4| \le \gamma_4, \qquad \hat{s} = s(1+\eta_4'), \quad |\eta_4'| \le \gamma_4.$$

The proof is identical to the proof of Lemma 5.2.

LEMMA 5.6. *In floating-point,* (4.6) *is computed backward stably,*

$$G_1^T P_1(A+\Delta A^{(3)})Q_1 = \hat{E}, \quad \|\Delta A^{(3)}\|_F \le \gamma_{2km}\|A\|_F.$$

Here, Q_1 refers to the direct sum of exact Householder reflectors $1\oplus(I-v_1v_1^T)\oplus(I-v_2v_2^T)$ for the computed $\hat{D}$. The orthogonal transformation is represented in floating-point by $\hat{v}_1$ and $\hat{v}_2$ satisfying

$$\begin{aligned}\hat{v}_1 &= v_1 + \Delta v_1,\ |\Delta v_1| \le \tilde{\gamma}_{q-1}\,|v_1|\,, \\ \hat{v}_2 &= v_2 + \Delta v_2,\ |\Delta v_2| \le \tilde{\gamma}_{m-q}\,|v_2|\,.\end{aligned}$$

Proof. From the backward stability analysis for Householder reflectors, we know

$$(\hat{D} + \Delta\hat{D})Q_1 = \hat{E}$$

with

$$\begin{aligned}\left\|\Delta\hat{D}\right\|_F \leq \sqrt{\gamma_{k(q-1)}^2\left\|\hat{D}\right\|_F^2 + \gamma_{k(m-q)}^2\left\|\hat{D}\right\|_F^2} &\leq \sqrt{\gamma_{k(q-1)}^2 + \gamma_{k(m-q)}^2}\left\|\hat{D}\right\|_F \\ &\leq \gamma_{k(m-1)}\left\|\hat{D}\right\|_F = \gamma_{k(m-1)}\left\|A + \Delta A^{(2)}\right\|_F.\end{aligned}$$

Setting $\Delta A^{(3)} = \Delta A^{(2)} + P_1 G_1 \Delta\hat{D}$, we find

$$G_1^T P_1 (A + \Delta A^{(3)})Q_1 = (\hat{D} + \Delta\hat{D})Q_1 = \hat{E}$$

with

$$\begin{aligned}\left\|\Delta A^{(3)}\right\|_F \leq \left\|\Delta A^{(2)}\right\|_F + \left\|\Delta\hat{D}\right\|_F &\leq \gamma_{k(m-1)}\left\|A\right\|_F + (1+\gamma_{k(m-1)})\left\|\Delta A^{(2)}\right\|_F \\ &\leq (\gamma_{k(m-1)} + (1+\gamma_{k(m-1)})\gamma_{k(m+1)})\left\|A\right\|_F \leq \gamma_{2km}\left\|A\right\|_F.\end{aligned}$$

□

LEMMA 5.7. *In floating-point,* (4.7) *is computed backward stably,*

$$G_1^T P_1 (A + \Delta A^{(4)})Q_1 H_1 = \hat{F}, \quad \left\|\Delta A^{(4)}\right\|_F \leq \gamma_{k(2m+1)}\left\|A\right\|_F.$$

Here, H_1 refers to the exact Givens rotation for the computed $\hat{e}_{45}$, $\hat{e}_{42}$. In addition,

$$\hat{c}' = c'(1+\eta_4), \quad |\eta_4| \leq \gamma_4, \qquad \hat{s}' = s'(1+\eta_4'), \quad |\eta_4'| \leq \gamma_4.$$

Proof. First,

$$(\hat{E} + \Delta\hat{E})H_1 = \hat{F}, \quad \left\|\Delta\hat{E}\right\|_F \leq \gamma_k\left\|\hat{E}\right\|_F = \gamma_k\left\|A + \Delta A^{(3)}\right\|_F.$$

Setting $\Delta A^{(4)} = \Delta A^{(3)} + P_1 G_1 \Delta\hat{E} Q_1$, we find

$$G_1^T P_1 (A + \Delta A^{(4)})Q_1 H_1 = (\hat{E} + \Delta\hat{E})H_1 = \hat{F}$$

with

$$\begin{aligned}\left\|\Delta A^{(4)}\right\|_F \leq \left\|\Delta A^{(3)}\right\|_F + \left\|\Delta\hat{E}\right\|_F &\leq \gamma_k\left\|A\right\|_F + (1+\gamma_k)\left\|\Delta A^{(3)}\right\|_F \\ &\leq (\gamma_k + (1+\gamma_k)\gamma_{2km})\left\|A\right\|_F \leq \gamma_{k(2m+1)}\left\|A\right\|_F.\end{aligned}$$

The bounds on $\hat{c}'$ and $\hat{s}'$ are direct from (B.3). □

THEOREM 5.8. *The triangularization procedure of Figure 3.2 is backward stable. Given a square matrix A, the algorithm computes upper-triangular $\hat{R}$ satisfying*

$$\tilde{U}^T (A + \Delta A)\tilde{V} = \hat{R}, \quad \left\|\Delta A\right\|_F \leq \gamma_{km^2}\left\|A\right\|_F,$$

for $\tilde{U}$ and $\tilde{V}$ as described in (3.7)–(3.8) *and* (3.2)–(3.6). *The error bounds from Theorem 3.1 on $\hat{c}_i$, $\hat{s}_i$, $\hat{c}_i'$, and $\hat{s}_i'$ and also on $\hat{u}_1^{(i)}$, $\hat{u}_2^{(i)}$, $\hat{v}_1^{(i)}$, and $\hat{v}_2^{(i)}$ hold.*

Proof. The proof of the backward error bound is by induction on q.

The base case $q = 1$ is Lemma 5.3.

For the induction step, assume that $q \geq 2$ and that the theorem has already been proved for $q-1$. From Lemmas 5.4 through 5.7, we know

$$G_1^T P_1 (A + \Delta A^{(4)}) Q_1 H_1 = \hat{F}, \quad \|\Delta A^{(4)}\|_F \leq \gamma_{k(2m+1)} \|A\|_F.$$

Let $\tilde{F}$ be the $(m-2)$-by-$(m-2)$ principal submatrix of $\hat{F}$ obtained by deleting indices 1 and m. Triangularize $\tilde{F}$ inductively. By the induction hypothesis,

$$\begin{aligned}\left(\tilde{G}_q^T \tilde{P}_q\right) \cdots \left(\tilde{G}_2^T \tilde{P}_2\right) \left(\tilde{F} + \Delta\tilde{F}\right) \left(\tilde{Q}_2 \tilde{H}_2\right) \cdots \left(\tilde{Q}_{q-1} \tilde{H}_{q-1}\right) \tilde{Q}_q \cdots \tilde{Q}_{m-q} \\ = \left[\begin{array}{cc|c} \hat{K}_{22} & \hat{K}_{23} & \hat{K}_{24} \\ \hline 0 & 0 & \hat{K}_{34} \end{array}\right]\end{aligned}$$

for some $\Delta\tilde{F}$ satisfying

$$\|\Delta\tilde{F}\|_F \leq \gamma_{k(m-2)^2} \|\tilde{F}\|_F \leq \gamma_{k(m-2)^2} \|\hat{F}\|_F = \gamma_{k(m-2)^2} \|A + \Delta A^{(4)}\|_F.$$

(Remember that K is square and upper-triangular, despite the irregular partitioning.) Border $\Delta\tilde{F}$ by zeros on all four sides to construct the m-by-m matrix $\Delta\hat{F}$. Then

$$\begin{aligned}\left(G_q^T P_q\right) \cdots \left(G_2^T P_2\right) \times \left(G_1^T P_1 \left(A + \Delta A^{(4)}\right) Q_1 H_1 + \Delta\hat{F}\right) \\ \times \left(Q_2 H_2\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q} = \hat{R},\end{aligned}$$

that is,

$$\left(G_q^T P_q\right) \cdots \left(G_1^T P_1\right) \left(A + \Delta A^{(5)}\right) \left(Q_1 H_1\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q} = \hat{R},$$

with $\Delta A^{(5)} = \Delta A^{(4)} + P_1 G_1(\Delta\hat{F}) H_1^T Q_1$. Hence, the backward error is bounded by

$$\begin{aligned}\|\Delta A^{(5)}\|_F &\leq \|\Delta A^{(4)}\|_F + \|\Delta\hat{F}\|_F \\ &\leq \|\Delta A^{(4)}\|_F + \gamma_{k(m-2)^2} \|A + \Delta A^{(4)}\|_F \\ &\leq \gamma_{k(m-2)^2} \|A\|_F + (1 + \gamma_{k(m-2)^2}) \|\Delta A^{(4)}\|_F \\ &\leq (\gamma_{k(m-2)^2} + (1 + \gamma_{k(m-2)^2}) \gamma_{k(2m+1)}) \|A\|_F \\ &\leq \gamma_{[k(m-2)^2 + k(2m+1)]} \|A\|_F \leq \gamma_{k(m^2-2m+5)} \|A\|_F.\end{aligned}$$

We are assuming $r = q \geq 2$, which implies $m = q + (m-q) \geq 2 + 2 = 4$. Thus, the subscript is bounded above by km^2. Renaming $\Delta A^{(5)}$ to ΔA, we find the backward error bound

$$\|\Delta A\|_F \leq \gamma_{km^2} \|A\|_F.$$

Error bounds on $\hat{c}_i$, $\hat{s}_i$, $\hat{c}'_i$, $\hat{s}'_i$ and on $\hat{u}_1^{(i)}$, $\hat{u}_2^{(i)}$, $\hat{v}_1^{(i)}$, and $\hat{v}_2^{(i)}$ have already been proved in Lemmas 5.1 through 5.7. □

6. Simultaneous bidiagonalization of an orthogonal matrix. Now, suppose the matrix A is an orthogonal matrix X. In exact arithmetic,

$$\tilde{U}^T X \tilde{V} = R.$$

R is designed to be upper-triangular with positive diagonal, and it must be orthogonal. There is only one possibility: $R = I$. Unrolling U and V, we have

$$\left(G_q^T P_q\right) \cdots \left(G_1^T P_1\right) X \left(Q_1 H_1\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q} = I.$$

Because G_i and P_j operate on distinct rows when $j > i$, the matrices commute: $P_j G_i^T = G_i^T P_j$. Similarly, $H_i Q_j = Q_j H_i$ for $j > i$. Therefore,

$$\left(P_q \cdots P_1\right) X \left(Q_1 \cdots Q_{m-q}\right) = \left(G_1 \cdots G_q\right) \left(H_1 \cdots H_{q-1}\right)^T.$$

This is the sought equation (2.1), with the right-hand side given in the form of (2.2).

The bidiagonalization algorithm is complete. Given a partitioned orthogonal matrix X, the decomposition

$$X = \left[\begin{array}{cc} U_1 & \\ & U_2 \end{array}\right] \left[\begin{array}{cc|cc} B_{11} & & & B_{12} \\ & I & & \\ \hline & & I & \\ B_{21} & & & B_{22} \end{array}\right] \left[\begin{array}{cc} V_1 & \\ & V_2 \end{array}\right]^T$$

is computed as follows:

(i) Reduce X to the identity matrix using the orthogonal transformations illustrated by Figure 3.2 and defined in §4,

$$\left(G_q^T P_q\right) \cdots \left(G_1^T P_1\right) X \left(Q_1 H_1\right) \cdots \left(Q_{q-1} H_{q-1}\right) Q_q \cdots Q_{m-q} = I.$$

(ii) Generate the matrix in bidiagonal-block form from the Givens rotations,

$$\left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \end{array}\right] = \left(G_1 \cdots G_q\right) \left(H_1 \cdots H_{q-1}\right)^T.$$

(iii) Generate $\left[\begin{smallmatrix} U_1 & \\ & U_2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} V_1 & \\ & V_2 \end{smallmatrix}\right]$ from the Householder reflectors,

$$\left[\begin{array}{cc} U_1 & \\ & U_2 \end{array}\right] = P_1 \cdots P_q, \quad \left[\begin{array}{cc} V_1 & \\ & V_2 \end{array}\right] = Q_1 \cdots Q_{m-q}.$$

7. Numerical stability of simultaneous bidiagonalization. In practice, X is not exactly orthogonal, and roundoff error is unavoidable. Therefore, the triangular matrix $\hat{R} = \tilde{U}^T (X + \Delta X^{(1)}) \tilde{V}$ should *approximately* equal the identity matrix. In theory, we round $\hat{R}$ to the identity matrix:

$$\begin{aligned} &\tilde{U}^T (X + \Delta X^{(1)}) \tilde{V} = \hat{R} \\ \Longrightarrow\ &\tilde{U}^T (X + \Delta X^{(2)}) \tilde{V} = I, \quad \Delta X^{(2)} = \Delta X^{(1)} + \tilde{U}(I - \hat{R})\tilde{V}^T. \end{aligned} \tag{7.1}$$

(To save time in practice, we avoid computing quantities that are immediately rounded to zero or one. However, for the stability analysis, it is useful to imagine that all entries of $\hat{R}$ are computed.) The backward error is bounded by

$$\left\|\Delta X^{(2)}\right\|_F \le \left\|\Delta X^{(1)}\right\|_F + \left\|I - \hat{R}\right\|_F.$$

How big is this? The following three lemmas are helpful. They assume that m is small enough, relative to $\mathbf{u}$ and $\varepsilon := \left\|I - X^T X\right\|_2$, that

$$\sqrt{8m}\,\varepsilon \le 1/4, \quad \sqrt{m}(1+\varepsilon)\gamma_{km^2} < 1. \tag{7.2}$$

(This is the "mild assumption" mentioned at the end of §3.)

First, $\hat{R}$ is nearly orthogonal.

LEMMA 7.1. *If* $\|I - X^T X\|_2 = \varepsilon \le \frac{1}{4}$, *then* $\left\|I - \hat{R}^T \hat{R}\right\|_F \le \sqrt{m}\,(\varepsilon + 5\gamma_{km^2})$.

Proof. Let $\sigma_1, \ldots, \sigma_m$ be the singular values of X. We know $|1 - \sigma_i| \le |1 - \sigma_i|\,|1 + \sigma_i| = \left|1 - \sigma_i^2\right| \le \varepsilon$, so all singular values are in $1 \pm \varepsilon$, in particular, $\|X\|_2 \le 1 + \varepsilon$ and $\|X\|_F \le \sqrt{m}\|X\|_2 \le \sqrt{m}(1 + \varepsilon)$. Theorem 5.8 shows $\|\Delta X^{(1)}\|_F \le \gamma_{km^2}\|X\|_F \le \sqrt{m}(1 + \varepsilon)\gamma_{km^2}$. Finally,

$$\begin{aligned}
\left\|I - \hat{R}^T \hat{R}\right\|_F &= \left\|I - \tilde{V}^T (X + \Delta X^{(1)})^T \tilde{U}\tilde{U}^T (X + \Delta X^{(1)})\tilde{V}\right\|_F \\
&= \left\|I - (X + \Delta X^{(1)})^T (X + \Delta X^{(1)})\right\|_F \\
&\le \|I - X^T X\|_F + 2\left\|X^T \Delta X^{(1)}\right\|_F + \left\|\Delta X^{(1)}\right\|_F^2 \\
&\le \|I - X^T X\|_F + \left(2\|X\|_2 + 1\right)\left\|\Delta X^{(1)}\right\|_F \\
&\le \varepsilon\sqrt{m} + (2(1 + \varepsilon) + 1)\,\sqrt{m}(1 + \varepsilon)\gamma_{km^2} \\
&\le \sqrt{m}\,(\varepsilon + (3 + 2\varepsilon)\,(1 + \varepsilon)\gamma_{km^2}) \\
&\le \sqrt{m}\,(\varepsilon + 5\gamma_{km^2})\,.
\end{aligned}$$

(We used assumption (7.2) to bound $\|\Delta X^{(1)}\|_F < 1$.) ∎

The next two lemmas will show that $\hat{R}$ is not just nearly orthogonal but close to the identity matrix in particular. The first lemma provides a somewhat rough estimate, and the second refines it.

LEMMA 7.2. *If an* m*-by-*m *upper-triangular matrix* R *has positive diagonal and satisfies* $\|I - R^T R\|_F = \varepsilon \le \frac{1}{4}$, *then* $\|I - R\|_F \le \sqrt{8m}\,\varepsilon$ *for all* m *satisfying* $\sqrt{8m}\,\varepsilon \le \frac{1}{4}$.

Proof. The proof is by induction on m.

If $m = 1$, then $\left|1 - r_{11}^2\right| \le \varepsilon$, which implies $|1 - r_{11}| < |1 - r_{11}|\,|1 + r_{11}| = \left|1 - r_{11}^2\right| \le \varepsilon$.

Now suppose the result is true for all $(m-1)$-by-$(m-1)$ matrices, and let R be an m-by-m matrix satisfying $\|I - R^T R\|_F \le \varepsilon$. Write

$$R = \begin{bmatrix} S & x \\ & \alpha \end{bmatrix}.$$

Then

$$I - R^T R = \begin{bmatrix} I - S^T S & -S^T x \\ -x^T S & 1 - x^T x - \alpha^2 \end{bmatrix}.$$

Because

$$\|I - S^T S\|_F \le \|I - R^T R\|_F \le \varepsilon,$$

the induction hypothesis provides $\|I - S\|_F \le \sqrt{8(m-1)}\,\varepsilon$. Given any v, by the triangle inequality,

$$\begin{aligned}
\|S^T v\|_2 \ge \|v\|_2 - \|(I - S)^T v\|_2 &\ge (1 - \|I - S\|_2)\|v\|_2 \\
&\ge (1 - \|I - S\|_F)\|v\|_2 \ge (1 - \sqrt{8(m-1)}\,\varepsilon)\|v\|_2.
\end{aligned}$$

Considering the last column of $I - R^T R$, we find

$$\left\|S^T x\right\|_2 \leq \varepsilon,$$

so

$$\left\|x\right\|_2 \leq \frac{\left\|S^T x\right\|_2}{1 - \sqrt{8(m-1)}\,\varepsilon} \leq \frac{\varepsilon}{1 - \sqrt{8(m-1)}\,\varepsilon} \leq 2\varepsilon.$$

Finally, $\left|1 - x^T x - \alpha^2\right| \leq \varepsilon$, so

$$\left|1 - \alpha^2\right| \leq \left|1 - \alpha^2 - x^T x\right| + \left|x^T x\right| \leq \varepsilon + 4\varepsilon^2 \leq 2\varepsilon,$$

which implies

$$|1 - \alpha| = \frac{\left|1 - \alpha^2\right|}{1 + \alpha} \leq \frac{2\varepsilon}{1} = 2\varepsilon.$$

All together,

$$\begin{aligned}\left\|I - R\right\|_F = \sqrt{\left\|I - S\right\|_F^2 + \left\|x\right\|_2^2 + (1-\alpha)^2} \\ \leq \sqrt{8(m-1)\varepsilon^2 + 4\varepsilon^2 + 4\varepsilon^2} = \sqrt{8m}\,\varepsilon.\end{aligned}$$

□

LEMMA 7.3. *If an m-by-m upper-triangular matrix R has positive diagonal and satisfies* $\left\|I - R^T R\right\|_F = \varepsilon \leq \frac{1}{4}$, *then* $\left\|I - R\right\|_F \leq \varepsilon$ *for all m satisfying* $\sqrt{8m}\,\varepsilon \leq \frac{1}{4}$.

Proof. Let $R = I + E$. From the previous lemma, $\left\|E\right\|_F \leq \sqrt{8m}\,\varepsilon$. To refine this estimate, consider

$$\left\|I - R^T R\right\|_F = \left\|I - (I+E)^T (I+E)\right\|_F = \left\|E + E^T + E^T E\right\|_F.$$

Since E is upper-triangular, there can be no cancellation in $E + E^T$ and so $\left\|E + E^T\right\|_F \geq \sqrt{2}\left\|E\right\|_F$. Using this observation and the triangle inequality,

$$\begin{aligned}\left\|I - R^T R\right\|_F &= \left\|E + E^T + E^T E\right\|_F \\ &\geq \left\|E + E^T\right\|_F - \left\|E^T E\right\|_F \geq \left\|E + E^T\right\|_F - \left\|E\right\|_F^2 \\ &\geq \left\|E\right\|_F \left(\sqrt{2} - \left\|E\right\|_F\right) \geq \left\|E\right\|_F \left(\sqrt{2} - \sqrt{8m}\,\varepsilon\right) \\ &\geq \left\|E\right\|_F \left(\sqrt{2} - \frac{1}{4}\right) \geq \left\|E\right\|_F = \left\|I - R\right\|_F.\end{aligned}$$

□

Now we are ready to prove the main theorem.

Proof of Theorem (3.1). $\Delta X^{(2)}$ of (7.1) has been renamed to ΔX. We know from that equation that

$$\tilde{U}^T (X + \Delta X)\tilde{V} = I, \quad \left\|\Delta X\right\|_F \leq \left\|\Delta X^{(1)}\right\|_F + \left\|I - \hat{R}\right\|_F.$$

Using Theorem 5.8 and Lemma 7.3,

$$\left\|\Delta X\right\|_F \leq \gamma_{km^2}\left\|X\right\|_F + \left\|I - \hat{R}^T \hat{R}\right\|_F.$$

Using the proof and conclusion of Lemma 7.1,

$$\begin{aligned}\left\|\Delta X\right\|_F &\le \sqrt{m}(1+\varepsilon)\gamma_{km^2} + \sqrt{m}(\varepsilon + 5\gamma_{km^2}) \\ &\le \sqrt{m}\left(\varepsilon + \gamma_{km^2}(6+\varepsilon)\right) \le \sqrt{m}\left(\varepsilon + 7\gamma_{km^2}\right).\end{aligned}$$

Error bounds on $\hat{c}_i$, $\hat{s}_i$, $\hat{c}'_i$, $\hat{s}'_i$ and on $\hat{u}_1^{(i)}$, $\hat{u}_2^{(i)}$, $\hat{v}_1^{(i)}$, and $\hat{v}_2^{(i)}$ were already stated in Theorem 5.8. □

Acknowledgment. I thank the anonymous reviewers for their helpful comments and Alan Edelman for introducing me to the CS decomposition.

Appendix A. Computational model.

Our computational model is the now-standard one, and our presentation follows Higham's reference [10]. Unit roundoff is denoted by $\mathbf{u}$. A computed approximation to a quantity z is denoted $\mathrm{fl}(z)$ or $\hat{z}$.

If x and y are exactly representable and normalized, then the four basic arithmetic operations are relatively accurate, $\mathrm{fl}(x \text{ op } y) = (x \text{ op } y)(1+\delta)$, $|\delta| \le \mathbf{u}$, in which op can be any of the four basic arithmetic operations, $+, -, *, /$.

Certain bounds are more directly expressed in terms of $\gamma_k = k\mathbf{u}/(1-k\mathbf{u})$ rather than $\mathbf{u}$ itself. One particularly useful property is $\gamma_k + \gamma_l + \gamma_k\gamma_l \le \gamma_{k+l}$. "Whenever we write γ_n there is an implicit assumption that $n\mathbf{u} < 1$" [10]. To avoid chasing constants, the notation $\tilde{\gamma}_k := ck\mathbf{u}/(1-ck\mathbf{u})$ can be used, in which c is a small implicit constant.

An algorithm for computing $y = f(x)$ is *backward stable* if the computed $\hat{y}$ is the exact solution to a perturbation of the input, $\hat{y} = f(x + \Delta x)$, with Δx small under an appropriate norm.

Sometimes componentwise error bounds are preferable. If v is an m-by-1 vector, then $|\Delta v| \le c\,|v|$ is shorthand for $|\Delta v_i| \le c\,|v_i|$, $i = 1, \ldots, m$.

Appendix B. Householder reflectors and Givens rotations.

The bounds in this appendix are proved in Higham's reference [10].

The routine **houseg** computes a Householder reflector. (The name is short for "Householder generate.") If x is a given m-by-1 vector, then $(v, r) = \mathbf{houseg}(x)$ returns an m-by-1 vector v and a scalar $r \ge 0$ for which $P := (I - vv^T)$ is orthogonal and satisfies $Px = (I - vv^T)x = re_1$, in which e_1 is the first standard basis vector. Note that $\|v\|_2 = \sqrt{2}$. In floating-point,

$$\hat{v} = v + \Delta v, \quad |\Delta v| \le \tilde{\gamma}_m |v|, \qquad \hat{r} = r(1+\tilde{\eta}_m), \quad |\tilde{\eta}_m| \le \tilde{\gamma}_m. \tag{B.1}$$

Note also the backward error bound

$$P(x + \Delta x) = \hat{r}e_1, \quad \left\|\Delta x\right\|_2 \le \tilde{\gamma}_m \left\|x\right\|_2.$$

(Set $\Delta x = \tilde{\eta}_m r P e_1$.) In the backward error bound, P refers to the exact Householder reflector $I - vv^T$ for x rather than the reflector $I - 2\hat{v}\hat{v}^T/\|\hat{v}\|_2^2$ defined by the computed $\hat{v}$.

The routine **housea** can then apply the Householder reflector. (The name is short for "Householder apply.") If A is a given m-by-n matrix and v is the Householder vector defined above, then $B = \mathbf{housea}(v, A)$ computes $B = (I - vv^T)A$. The following backward error bound is achievable in floating-point:

$$P(A + \Delta A) = \hat{B}, \quad \left\|\Delta A\right\|_F \le \tilde{\gamma}_m \left\|A\right\|_F. \tag{B.2}$$

Again, P refers to the exact Householder reflector for x. If some column a_j of A equals x, then it is more natural to set $\hat{b}_j := (\hat{r}, 0, \ldots, 0)$ than to compute $\hat{b}_j := (I - \hat{v}\hat{v}^T)x = x - (\hat{v}^T x)\hat{v}$. The columnwise error bound $P(a_j + \Delta a_j) = \hat{b}_j$, $\|\Delta a_j\|_2 \leq \tilde{\gamma}_m \|a_j\|_2$, holds in either case, so (B.2) holds regardless of whether any column of A equals x.

The routine **givensg** computes a Givens rotation. If x and y are given scalars, then $(c, s, r) = \mathbf{givensg}(x, y)$ computes c and s such that $c^2 + s^2 = 1$ and

$$G^T \begin{bmatrix} x \\ y \end{bmatrix} := \begin{bmatrix} c & -s \\ s & c \end{bmatrix}^T \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}.$$

In floating-point,

$$\hat{c} = c(1 + \eta_4), \quad |\eta_4| \leq \gamma_4, \qquad \hat{s} = s(1 + \eta_4'), \quad |\eta_4'| \leq \gamma_4, \tag{B.3}$$

and $\hat{r} = r(1 + \eta_3)$, $|\eta_3| \leq \gamma_3$. Note also the backward error bound

$$G^T \left(\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \right) = \begin{bmatrix} \hat{r} \\ 0 \end{bmatrix}, \quad \left\| \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \right\|_2 \leq \eta_3 \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2.$$

Here, G refers to the exact Givens rotation, not the computed one.

The routine **givensa** can then apply the Givens rotation. If A is a given 2-by-n matrix and c, s, and G are defined as above, then $B = \mathbf{givensa}(c, s, A)$ computes $B = G^T A$. In floating-point, we find

$$G^T(A + \Delta A) = \hat{B}, \quad \|\Delta A\|_F \leq \tilde{\gamma}_1 \|A\|_F. \tag{B.4}$$

In this equation, c and s refer to the exact cosine and sine for (x, y) rather than the computed $\hat{c}$, $\hat{s}$.

Finally, we note the availability of stable algorithms for the RQ decomposition. It is possible to reduce an m-by-n matrix A ($m \geq n$) to an upper-trapezoidal $\hat{R}$,

$$(A + \Delta A)Q_1 \cdots Q_n = \hat{R}, \quad \|\Delta A\|_F \leq \tilde{\gamma}_{n^2} \|A\|_F, \tag{B.5}$$

using Householder reflectors $Q_1, \ldots, Q_n$. The Householder reflectors can be generated from vectors $v^{(1)}, \ldots, v^{(n)}$ satisfying bounds of the form (B.1).

REFERENCES

[1] Z. Bai, *The CSD, GSVD, their applications and computations*, Tech. Rep. 958, IMA Preprint Series, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1992.

[2] Z. Bai and J. W. Demmel, *Computing the generalized singular value decomposition*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1464–1486.

[3] Z. Bai and H. Zha, *A new preprocessing algorithm for the computation of the generalized singular value decomposition*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1007–1012.

[4] C. Davis and W. M. Kahan, *Some new bounds on perturbation of subspaces*, Bulletin of the American Mathematical Society, 75 (1969), pp. 863–868.

[5] ———, *The rotation of eigenvectors by a perturbation. III*, SIAM Journal on Numerical Analysis, 7 (1970), pp. 1–46.

[6] A. Edelman and B. D. Sutton, *The beta-Jacobi matrix model, the CS decomposition, and generalized singular value problems*, Found. Comput. Math., 8 (2008), pp. 259–285.

[7] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 205–224.

[8] G. H. Golub and C. Reinsch, *Handbook series linear algebra: Singular value decomposition and least squares solutions*, Numerische Mathematik, 14 (1970), pp. 403–420.
[9] V. Hari, *Accelerating the SVD block-Jacobi method*, Computing, 75 (2005), pp. 27–53.
[10] N. J. Higham, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002.
[11] M. Möttönen, J. J. Vartiainen, V. Bergholm, and M. M. Salomaa, *Quantum circuits for general multiqubit gates*, Physical Review Letters, 93 (2004), p. 130502.
[12] C. C. Paige, *Computing the generalized singular value decomposition*, Society for Industrial and Applied Mathematics. Journal on Scientific and Statistical Computing, 7 (1986), pp. 1126–1146.
[13] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM Journal on Numerical Analysis, 18 (1981), pp. 398–405.
[14] C. C. Paige and M. Wei, *History and generality of the CS decomposition*, Linear Algebra and its Applications, 208/209 (1994), pp. 303–326.
[15] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Review, 19 (1977), pp. 634–662.
[16] ———, *Computing the CS decomposition of a partitioned orthonormal matrix*, Numerische Mathematik, 40 (1982), pp. 297–306.
[17] G. W. Stewart, *Matrix Algorithms: Basic decompositions*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.
[18] B. D. Sutton, *The stochastic operator approach to random matrix theory*, PhD thesis, Massachusetts Institute of Technology, 2005.
[19] ———, *Computing the complete CS decomposition*, Numer. Algorithms, 50 (2009), pp. 33–65.
[20] ———, *Computing the complete CS decomposition*, Fourteenth Leslie Fox Prize Meeting, University of Warwick, Warwick, UK, June 2009.
[21] F. Tisseur, *Backward stability of the QR algorithm*, TR 239, UMR 5585 Lyon Saint-Etienne, Oct. 1996.
[22] R. R. Tucci, *A rudimentary quantum compiler (2nd ed.)*, quant-ph/9902062, (1999).
[23] C. Van Loan, *Computing the CS and the generalized singular value decompositions*, Numerische Mathematik, 46 (1985), pp. 479–491.
[24] C. Van Loan and J. Speiser, *Computation of the C-S decomposition, with application to signal processing*, in SPIE Proceedings, vol. 696, 1986.
[25] D. S. Watkins, *Some perspectives on the eigenvalue problem*, SIAM Review. A Publication of the Society for Industrial and Applied Mathematics, 35 (1993), pp. 430–471.
[26] D. S. Watkins, *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.