SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 565–583

A USEFUL FORM OF UNITARY MATRIX OBTAINED FROM ANY SEQUENCE OF UNIT 2-NORM n-VECTORS*

CHRISTOPHER C. PAIGE†

This is dedicated in memory of Charles Sheffield. I did not meet him, but one insight he made has contributed greatly to the interests of our community, as this paper will reveal. His friend Gene Golub encouraged this dedication shortly before Gene died.

Abstract. Charles Sheffield pointed out that the modified Gram–Schmidt (MGS) orthogonalization algorithm for the QR factorization of $B \in \mathbb{R}^{n\times k}$ is mathematically equivalent to the QR factorization applied to the matrix B augmented with a $k \times k$ matrix of zero elements on top. This is true in theory for any method of QR factorization, but for Householder's method it is true in the presence of rounding errors as well. This knowledge has been the basis for several successful but difficult rounding error analyses of algorithms which in theory produce orthogonal vectors but significantly fail to do so because of rounding errors. Here we show that the same results can be found more directly and easily without recourse to the MGS connection. It is shown that for any sequence of k unit 2-norm n-vectors there is a special $(n+k)$-square unitary matrix which we call a unitary augmentation of these vectors and that this matrix can be used in the analyses without appealing to the MGS connection. We describe the connection of this unitary matrix to Householder matrices. The new approach is applied to an earlier analysis to illustrate both the improvement in simplicity and advantages for future analyses. Some properties of this unitary matrix are derived. The main theorem on orthogonalization is then extended to cover the case of biorthogonalization.

Key words. orthogonal matrices, loss of orthogonality, rounding error analysis, elementary unitary transformations, modified Gram–Schmidt, QR factorization, bidiagonalization, singular value decomposition, biorthogonal sets of vectors, loss of biorthogonality

AMS subject classifications. 65F25, 65F30, 65F35, 65F50, 65G50, 15A12, 15A18, 15A23, 15A57

DOI. 10.1137/080725167

1. Introduction.

We are interested in any matrix or vector algorithm that in theory produces a sequence of orthonormal n-vectors from a given sequence of n-vectors by first orthogonalizing each successive given vector against some of the already computed orthonormal vectors and then normalizing the resulting orthogonal vector. An example is modified Gram–Schmidt (MGS); see, for example, [10, section 5.2.8] and [5]. With finite precision computation these algorithms produce a sequence of n-vectors which can have a severe loss of orthogonality but where each vector has a 2-norm that is almost 1. Let the columns of $V_k \equiv [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$ be such a sequence, where each vector has been normalized to have unit length (i.e., 2-norm of 1). Our ultimate aim is to obtain relatively clear and short rounding error analyses of such algorithms, but this paper is devoted to presenting a theoretical tool (a unitary matrix $Q^{(k)}$ related to V_k) which will facilitate such analyses, and to illustrate this with an analysis. We also extend this tool to two biorthogonal sets of vectors.

In Theorem 2.1 we describe how a particular $(n+k)\times(n+k)$ unitary matrix $Q^{(k)}$ can be derived from such a V_k. We show that this $Q^{(k)}$ is a product of Householder

*Received by the editors May 27, 2008; accepted for publication (in revised form) by J. L. Barlow February 9, 2009; published electronically May 20, 2009. This work was supported by NSERC of Canada grant OGP0009236.
http://www.siam.org/journals/simax/31-2/72516.html
†School of Computer Science, McGill University, Montreal, QC, H3A 2A7, Canada (paige@cs.mcgill.ca).

matrices dependent on the v_j alone. We describe the evolution of this idea and why it is so useful for rounding error analyses of algorithms of this type. In section 3 we quickly summarize a recent bidiagonalization algorithm by Barlow, Bosner, and Drmač [2] and show how the $Q^{(k)}$, V_k relationship can be used to give a simpler and shorter rounding error analysis of their algorithm. This indicates both the improvement in clarity and simplicity that the present approach provides as well as how it might be applied to other algorithms. In section 4 we provide a theorem that can be used for transforming augmented results—such as those of dimension $(n+k) \times k$—into more standard results, such as those involving the dimensions $n \times k$ of V_k. In section 5 we provide some properties of $Q^{(k)}$ for possible future use, and in section 6 we discuss the idea of optimality, and how such a $Q^{(k)}$ might be used. Finally, in section 7 we give a theorem (a biorthogonal version of Theorem 2.1) suggesting that some of the ideas here might be extended to analyze some algorithms producing two sets of biorthogonal vectors.

We will say complex nonsquare $n \times k$ Q_1 has orthonormal columns if $Q_1^H Q_1 = I$ and write $Q_1 \in \mathcal{U}^{n\times k}$, while Q_1 and Q_2 are orthogonal to each other if $Q_1^H Q_2 = 0$. For floating point arithmetic the *unit roundoff* (a measure of relative precision; see, e.g., [12]) will be denoted by ϵ. I_n denotes the $n \times n$ unit matrix (but we will sometimes use I), e_j will be the jth column of a unit matrix I, so Be_j is the jth column of B, while e will be a vector of ones of the required dimension. We will denote the absolute value of a matrix B by $|B|$, the Frobenius norm by $\|B\|_F \equiv \sqrt{\text{trace}(B^H B)}$, the vector 2-norm by $\|v\|_2 \equiv \sqrt{v^H v}$, and its subordinate matrix norm by $\|\cdot\|_2$, while $\sigma(\cdot)$ will denote a singular value, and $\kappa_2(B) \equiv \sigma_{max}(B)/\sigma_{min}(B)$.

We will usually index matrices by subscripts as in V_k when the $(k+1)$st matrix can be obtained from the kth by appending a column, or a column and a row. We will use, e.g., $Q^{(k)}$ otherwise, and $Q^{(k)} = [Q_1^{(k)}, Q_2^{(k)}]$. Note that M_j in section 2.1 should really be $M^{(j)}$, but space prevented this in (2.9). P^k indicates "P to the power k."

2. Obtaining a unitary matrix from unit 2-norm n-vectors. The main theoretical results of this paper are contained in the following theorem and its corollaries. We use SUT to mean "strictly upper triangular," while "sut$(\cdot)$" gives the SUT part of the matrix in parentheses. Similarly, SLT means "strictly lower triangular."

THEOREM 2.1. *For any integers $n \geq 1$ and $k \geq 1$, and $V_k \equiv [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$ with $\|v_j\|_2 = 1, \quad j = 1, \ldots, k$, define the SUT matrix S_k as follows:*

$$S_k \equiv (I_k + U_k)^{-1} U_k \equiv U_k (I_k + U_k)^{-1} \in \mathbb{C}^{k\times k}, \qquad U_k \equiv \text{sut}(V_k^H V_k) \tag{2.1}$$

(where clearly $I_k \pm S_k$ and $I_k \pm U_k$ are always nonsingular). Then

$$U_k S_k = S_k U_k, \quad U_k = (I_k - S_k)^{-1} S_k \equiv S_k (I_k - S_k)^{-1}, \quad (I_k - S_k)^{-1} = I_k + U_k, \tag{2.2}$$

$$(I_k - S_k)^H V_k^H V_k (I_k - S_k) = I_k - S_k^H S_k, \tag{2.3}$$

$$(I_k - S_k) V_k^H V_k (I_k - S_k)^H = I_k - S_k S_k^H, \tag{2.4}$$

$$\|S_k\|_2 \leq 1; \quad V_k^H V_k = I \Leftrightarrow \|S_k\|_2 = 0; \quad V_k^H V_k \ \ \textit{singular} \Leftrightarrow \|S_k\|_2 = 1. \tag{2.5}$$

Most importantly, S_k is the unique *SUT $k \times k$ matrix such that*

$$Q^{(k)} \equiv \left[Q_1^{(k)} \;\middle|\; Q_2^{(k)} \right] \equiv \left[\begin{array}{c|c} S_k & (I_k - S_k) V_k^H \\ V_k (I_k - S_k) & I_n - V_k (I_k - S_k) V_k^H \end{array} \right] \in \mathcal{U}^{(n+k)\times(n+k)}. \tag{2.6}$$

If we write $S_{k+1} \equiv \left[\begin{smallmatrix} \widehat{S}_k & s_{k+1} \\ 0 & 0 \end{smallmatrix}\right]$, *we also have* $\widehat{S}_k = S_k$ *and*

$$s_{k+1} = (I_k - S_k)V_k^H v_{k+1}, \quad \begin{bmatrix} S_{k+1} \\ V_{k+1}(I_{k+1} - S_{k+1}) \end{bmatrix} = \left[\begin{array}{c|c} S_k & s_{k+1} \\ 0 & 0 \\ V_k(I_k - S_k) & v_{k+1} - V_k s_{k+1} \end{array}\right]. \tag{2.7}$$

Proof. We start with (2.6). For any $k \times k$ SUT matrix S, define $M \equiv Q_1^H Q_1 - I$ for $Q_1 \equiv \left[\begin{smallmatrix} S \\ V_k(I-S) \end{smallmatrix}\right]$; see (2.6). Since by definition $V_k^H V_k = I + U_k + U_k^H$, we have

$$\begin{aligned} I + M = Q_1^H Q_1 &= S^H S + (I-S)^H (I-S) + (I-S)^H (U_k + U_k^H)(I-S), \\ M &= -(I-S)^H S - S^H (I-S) + (I-S)^H (U_k + U_k^H)(I-S), \\ (I-S)^{-H} M (I-S)^{-1} &= -S(I-S)^{-1} - (I-S)^{-H} S^H + (U_k + U_k^H). \end{aligned}$$

But $U_k - S(I-S)^{-1}$ is SUT, and the rest of the right-hand side is SLT, so $M = 0$ if and only if $U_k = S(I-S)^{-1}$. But then $S = U_k(I-S) = U_k - U_k S$ and so $(I+U_k)S = U_k$, proving that S_k in (2.1) is the unique $k \times k$ SUT matrix that gives $Q_1^{(k)H} Q_1^{(k)} = I$.

From (2.1) $U_k = S_k + U_k S_k$, so $U_k(I - S_k) = S_k$, $U_k = S_k(I-S_k)^{-1} = (I-S_k)^{-1} S_k$. Also $U_k(I-S_k) + (I-S_k) = I$, so $U_k + I = (I-S_k)^{-1}$, and then $U_k S_k = S_k U_k$, proving (2.2). Note that (2.3) follows from $Q_1^{(k)H} Q_1^{(k)} = I$, and using it gives for (2.6)

$$\begin{aligned} Q_2^{(k)H} Q_2^{(k)} &= V_k(I_k - S_k)^H (I_k - S_k) V_k^H + I_n - V_k(I_k - S_k)^H V_k^H - V_k(I_k - S_k) V_k^H \\ &\quad + V_k(I_k - S_k^H S_k) V_k^H = I_n, \\ Q_1^{(k)H} Q_2^{(k)} &= S_k^H (I_k - S_k) V_k^H + (I_k - S_k)^H V_k^H - (I_k - S_k^H S_k) V_k^H = 0, \end{aligned}$$

so that (2.6) holds. Then (2.4) follows from the leading principal $k \times k$ submatrix of $Q^{(k)} Q^{(k)H}$. Next, in (2.6), S_k is a submatrix of a unitary matrix so that $\|S_k\|_2 \le 1$, while from (2.3) $V_k^H V_k$ is singular if and only if $I_k - S_k^H S_k$ is, so (2.5) follows.

To prove (2.7) note that $V_{k+1} \equiv [V_k, v_{k+1}]$ and write $u_{k+1} \equiv V_k^H v_{k+1}$ so that $U_{k+1} e_{k+1} = \left[\begin{smallmatrix} u_{k+1} \\ 0 \end{smallmatrix}\right]$. Now from (2.2) $S_{k+1} = (I_{k+1} - S_{k+1}) U_{k+1}$. The proof follows by using this and the fact that U_{k+1} and S_{k+1} are SUT and noting that if $\widehat{S}_k = (I - \widehat{S}_k) U_k$, then $\widehat{S}_k (I + U_k) = U_k$, so that $\widehat{S}_k = U_k (I + U_k)^{-1} = S_k$ from (2.1):

$$\begin{aligned} \left[\frac{\widehat{S}_k}{0}\right] &= S_{k+1} \left[\frac{I_k}{0}\right] = (I_{k+1} - S_{k+1}) \left[\frac{U_k}{0}\right] = \left[\frac{(I_k - \widehat{S}_k) U_k}{0}\right] = \left[\frac{S_k}{0}\right], \\ \left[\frac{s_{k+1}}{0}\right] &= S_{k+1} e_{k+1} = (I_{k+1} - S_{k+1}) \left[\frac{u_{k+1}}{0}\right] = \left[\frac{(I_k - S_k) u_{k+1}}{0}\right]. \end{aligned}$$

∎

We can call the construction in Theorem 2.1 a *unitary or orthonormal augmentation of an array or sequence of unit length vectors* (the "augmentation" from V_k to $Q^{(k)}$ in (2.6)). It can also be thought of as a kind of reorthogonalization from V_k to $Q_1^{(k)}$, but moving to a higher dimension with the inclusion of each additional vector v_{k+1}. Note from (2.1) that $V_k^H V_k = I \;\Leftrightarrow\; U_k = 0 \;\Leftrightarrow\; S_k = 0$, and $V_k^H V_k = I$ corresponds to no "reorthogonalization." Also forming $V_k(I - S_k) = V_k - V_k S_k$ subtracts multiples of previous columns from each column, and so this looks a bit like Gram–Schmidt orthogonalization. In fact, if $S_k = 0$, then for *any* v_{k+1} the next step $v_{k+1} - V_k s_{k+1}$ in (2.7) gives

$$V_k^H (v_{k+1} - V_k s_{k+1}) = V_k^H (v_{k+1} - V_k V_k^H v_{k+1}) = 0,$$

which *is* ordinary (re)orthogonalization of v_{k+1}.

Note that $0 \le \|S_k\|_2 \le 1$ is a beautiful measure of the loss of orthogonality in V_k.

2.1. The relationship to Householder transformations. It is not obvious from Theorem 2.1, but the unitary matrix $Q^{(k)}$ in (2.6) is the product of k Householder matrices (unitary elementary Hermitian matrices) $P^{(1)}, \ldots, P^{(k)}$, where each $P^{(j)}$ depends on v_j alone. To prove this we first restate a theorem from [5, Theorem 4.1].

THEOREM 2.2. *Let $V_k = [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$, and for $j = 1, \ldots, k$, define*

$$M_j = I_n - v_j v_j^H, \qquad p_j = \begin{bmatrix} -e_j \\ v_j \end{bmatrix} \in \mathbb{C}^{n+k}, \qquad P^{(j)} = I_{n+k} - p_j p_j^H. \tag{2.8}$$

Then with the partitioning we use throughout this theorem

$$P \equiv P^{(1)}P^{(2)}\cdots P^{(k)} \equiv \begin{array}{c} \\ k \\ n \end{array} \left[\begin{array}{c|c} \overset{k}{P_{11}} & \overset{n}{P_{12}} \\ \hline P_{21} & P_{22} \end{array}\right]$$

$$= \left[\begin{array}{ccccc|c} 0 & v_1^H v_2 & v_1^H M_2 v_3 & \cdots & v_1^H M_2 M_3 \cdots M_{k-1} v_k & v_1^H M_2 M_3 \cdots M_k \\ 0 & 0 & v_2^H v_3 & \cdots & v_2^H M_3 M_4 \cdots M_{k-1} v_k & v_2^H M_3 M_4 \cdots M_k \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & v_{k-1}^H v_k & v_{k-1}^H M_k \\ 0 & 0 & 0 & \cdots & 0 & v_k^H \\ \hline v_1 & M_1 v_2 & M_1 M_2 v_3 & \cdots & M_1 M_2 \cdots M_{k-1} v_k & M_1 M_2 \cdots M_k \end{array}\right] \tag{2.9}$$

$$= \left[\begin{array}{c|c} P_{11} & (I_k - P_{11})V_k^H \\ \hline V_k(I_k - P_{11}) & I_n - V_k(I_k - P_{11})V_k^H \end{array}\right]. \tag{2.10}$$

P is a unitary matrix if and only if $\|v_j\|_2 = 1$ for $j = 1, \ldots, k$; and $P_{11} = 0$ if and only if $V_k^H V_k$ is diagonal.

In Theorem 2.1 we assumed that $\|v_j\|_2 = 1$ for $j = 1, \ldots, k$. In that case P here is unitary, just as $Q^{(k)}$ is. Also $Q^{(k)}$ in (2.6) and P in (2.10) have the same partitioning and form and are identical if $S_k = P_{11}$. But P_{11} is seen to be SUT from (2.9), and from Theorem 2.1 SUT S_k is unique, and therefore $S_k = P_{11}$, proving that in Theorems 2.1 and 2.2 with $\|v_j\|_2 = 1$ for $j = 1, \ldots, k$

$$\text{unitary } \; Q^{(k)} = P \equiv P^{(1)}P^{(2)}\cdots P^{(k)}. \tag{2.11}$$

It can be seen that using the Householder matrices $P^{(j)}$ is another way of producing this "unitary augmentation" of these unit length columns of V_k. Theorem 2.1 also gives the important relationship of S_k to $U_k = \text{sut}(V_k^H V_k)$ (see (2.1) and (2.2)) as well as other useful properties of S_k, while Theorem 2.2 gives the interesting (2.9). It is difficult to remember, but it appears that we were not aware of the S_k, U_k relationship in [5]. This relationship was given by Giraud, Gratton, and Langou in [7, (3.1)], where it was a key idea in their method, enabling a saving of computations. The relationship of S_k to U_k was also spelled out in [18, (5.5)].

Note from (2.9) that when a new unit length vector v_{k+1} is added, the last column of $Q_1^{(k+1)}$ in (2.6) is $Q_2^{(k)} v_{k+1}$ with a zero inserted immediately after the kth element. This can also be derived from (2.7). And $Q_2^{(k+1)}$ is $Q_2^{(k)} M_{k+1}$ with v_{k+1}^H inserted immediately after the kth row. Of course, for unit length v_j the M_j are orthogonal projection matrices, so multiplication by each cannot increase the 2-norm.

Theorem 2.2 allows P in (2.9) with $\|v_j\|_2 \neq 1$, but the subset of those matrices P with all unit length v_j can be described and used more easily via Theorem 2.1. This is an important advantage since we can always phrase our analyses in terms of all unit length v_j. Just to emphasize that the $P^{(j)}$ in (2.8) need not appear in any analysis that makes use of $Q^{(k)}$ in (2.6), we restate the following obvious part of Theorem 2.1.

COROLLARY 2.3. *Given any $n \times k$ matrix V_k with unit length columns, if $U_k \equiv \mathrm{sut}(V_k^H V_k)$ and $S_k \equiv (I + U_k)^{-1} U_k$, then $Q^{(k)}$ defined in (2.6) is a unitary matrix.*

2.2. The evolution of this idea. Charles Sheffield [24] pointed out that the modified Gram–Schmidt (MGS) orthogonalization algorithm for the QR factorization of $B \in \mathbb{R}^{n \times k}$ is mathematically equivalent to the QR factorization applied to the matrix B augmented with a $k \times k$ matrix of zero elements on top. This is true in theory for any method of QR factorization, but for Householder's method it is true in the presence of rounding errors as well. That is, the Householder QR factorization of the matrix $\left[\begin{smallmatrix} 0 \\ B \end{smallmatrix}\right] \in \mathbb{R}^{(n+k) \times k}$ leads to the identical *computed* upper triangular matrix $\tilde{R}$ in MGS as well as all intermediate vectors produced by MGS; see [5, section 2]. However, Langou [15, pp. 88–89] argues that a very minor change has to be made to either MGS or the Householder method so that the two algorithms perform *exactly* the same operations. Such a minor change would not affect any rounding error bounds.

Björck and Paige [5, 6] applied Sheffield's observation in their stability analyses of MGS and some applications in order to prove some new results that are additional to those derived by Björck in [3]. Following some ideas in [5], Barlow, Bosner, and Drmač [2] used Sheffield's insight to prove numerical stability properties of their recent bidiagonalization algorithm in [2]. Together with the insight of Giraud and Langou [8] that under very mild conditions MGS produces a well-conditioned set of vectors (see Corollary 5.2 here), Sheffield's insight allowed Paige, Rozložník, and Strakoš [18] to prove the backward stability of the MGS-GMRES algorithm of Saad and Schultz [23] under very mild conditions. Sheffield's insight has thus been of great value in the understanding of some widely used numerical algorithms.

For MGS it was shown that the resulting $S_k = P_{11}$ in (2.6) and (2.10) satisfied

$$S_k = P_{11} = E_1 \tilde{R}^{-1}, \qquad \|E_1\|_2 \le O(\epsilon)\|B\|_2, \tag{2.12}$$

where $\tilde{R}$ was the computed version of R from the MGS QR factorization of B; see [5, section 3]. It was realized there that not only could this insight be used to analyze the numerical stability of algorithms as in [5, 6, 18], but also these ideas could be extended to derive new algorithms; see, for example, [5, section 5]. This insight was also used to derive a new algorithm in [7]. In [5, 6, 18] it appeared that the structure of P in Theorem 2.2 was relevant only to the MGS algorithm. But Barlow, Bosner, and Drmač [2] showed that this approach has wider applications, and this understanding has been transformed here into the simple and generally applicable Theorem 2.1. This theorem now appears to be useful for analyzing *any* algorithm which in theory produces orthonormal vectors but in practice, because of rounding errors, can fail to do so to a significant extent. For any analysis of an algorithm that computes a matrix V_k of supposedly orthonormal columns, the important ancilliary result will be an expression that can be used to take account of S_k in (2.6) (or U_k; see (2.2))—for example, to bound it as in (2.12), or to use it in some different form, such as $\|S_k \widetilde{J}\|_F \le O(\epsilon)\|X\|_F$; see (3.17) here for such a use.

The development of the theory in Theorem 2.1 and its use in section 3.1 make no reference to or use of the MGS connection. In fact, since the ideas can be applied to any sequence of unit length n-vectors, MGS is just a particular, but remarkable, case (remarkable because of the numerical equivalence with the Householder QR factorization of $\left[\begin{smallmatrix} 0 \\ B \end{smallmatrix}\right]$). But the connection with the special Householder matrices in section 2.1 is always present, and Theorems 2.1 and 2.2 can be thought of as two ways of pro-

ducing the same object $Q^{(k)}$ from the unit length vectors $v_1, \ldots, v_k$. Theorem 2.1 appears more simple and useful.

Theorem 2.1 can also simplify some previously daunting analyses. For example, the 50-page paper of Barlow, Bosner, and Drmač [2] is filled largely with a formidable rounding error analysis of their recent bidiagonalization algorithm. They essentially used Sheffield's idea via [5] at the core of their analysis; see [2, section 3.3]. But the new Theorem 2.1 allows us to analyze their algorithm more quickly and clearly; see section 3 here. The analysis here will serve two purposes—to illustrate the power of this approach and to indicate how this approach might be applied in general. Theorem 2.1 also offers hope for the successful analyses of other important algorithms based on orthogonalization, such as the eigenvalue algorithms of Arnoldi [1] and Lanczos [13], the method of conjugate gradients [11], and similar algorithms that are particularly suitable for large sparse matrix problems; see, for example, [14, 19, 20, 22]. Theorem 7.1 should have similar advantages for biorthogonalization algorithms.

3. Application to bidiagonalization. Barlow, Bosner, and Drmač [2] suggested a new method for orthogonally transforming $X \in \mathbb{R}^{n\times m}$, $n \geq m$, to give the $n \times m$ upper bidiagonal form (in order to compute the singular value decomposition (SVD)):

$$V^T X W = \begin{bmatrix} J \\ 0 \end{bmatrix}, \quad J \equiv \begin{bmatrix} \gamma_1 & \phi_2 & & \\ & \gamma_2 & \phi_3 & \\ & & \ddots & \ddots \\ & & & \gamma_m \end{bmatrix} \geq 0, \quad V^{-1} = V^T, \quad W^{-1} = W^T. \tag{3.1}$$

The equivalent notation in [2] was $U^T X V = B$. We use the present notation in order that the analysis here will match the notation in sections 2 and 5. This bidiagonalization is not unique but can be made so by choosing $w_1 \equiv W e_1$.

Some insight can be gained by presenting the method in [2] as intermediate between the two bidiagonalization algorithms proposed by Golub and Kahan [9]—the one based on Householder transformations (which we refer to as the "direct" algorithm) and the one which follows from the Lanczos process [13] applied to $\left[\begin{smallmatrix} 0 & X^T \\ X & 0 \end{smallmatrix}\right]$ with initial vector $\left[\begin{smallmatrix} w_1 \\ 0 \end{smallmatrix}\right]$ (which we refer to as the "iterative" algorithm). The direct one stops in m steps and is backward stable. It is ideal for problems of small to moderately large dimensions. The iterative one tends to lose orthogonality and because of this usually will not stop. Thus it can take many more than m steps to obtain, for example, full information on all the singular value-vector triplets. It is useful for problems with sparse X of very large dimensions.

The method in [2] uses Householder transformations ($W^{(j)}$ in our notation) to produce the effect of the orthogonal matrix W of smaller dimensions in (3.1), and this forces m-step termination. But it uses the "iterative" process to find the *columns* of V in (3.1). This leads to a saving in floating point operations (flops), often at the cost of significant loss of orthogonality in the columns of V. It is useful for moderately large problems where we do not require orthogonality in all the "left" singular vectors of X (in particular, those found from V corresponding to the smaller singular values). This last conclusion follows from the rounding error analysis of the algorithm.

The direct approach in [9] applied $W^{(1)}, V^{(1)}, W^{(2)}, V^{(2)}, \ldots$, in order to give

$$V \equiv V^{(1)} \cdots V^{(n)} \equiv [V_m, v_{m+1}, \ldots, v_n], \quad V_m \equiv [v_1, \ldots, v_m], \quad W \equiv W^{(1)} \cdots W^{(m)},$$

$$W^{(j)} \equiv \begin{bmatrix} I_{j-1} & 0 \\ 0 & W_H^{(j)} \end{bmatrix} \Big\}_{m-j+1}, \quad V^{(j)} \equiv \begin{bmatrix} I_{j-1} & 0 \\ 0 & V_H^{(j)} \end{bmatrix} \Big\}_{n-j+1}, \tag{3.2}$$

where the $W_H^{(j)}$ and $V_H^{(j)}$ are symmetric Householder transformations. $W^{(m)}$ just makes $\phi_m \geq 0$. $W^{(1)}$ is arbitrary, but $W^{(1)} = I$ was chosen in [2]. We now roughly sketch the approach in [2]. It does not use the Householder transformations $V^{(j)}$, but these are needed in our description. Note that for $j = k+1, \ldots, m$, $V^{(j)}e_k = e_k$, $W^{(j)}e_k = e_k$, so from (3.2),

$$v_k \equiv Ve_k = V^{(1)} \cdots V^{(k)} e_k, \quad w_k \equiv We_k = W^{(1)} \cdots W^{(k)} e_k, \quad k = 1, \ldots, m. \tag{3.3}$$

In [2] $W^{(k+1)}$ is designed to produce the last $m-k$ elements in the kth row of (3.1), i.e., $\phi_{k+1}, 0, \ldots, 0$,

$$\begin{aligned} e_k^T V^{(k)} \cdots V^{(1)} X W^{(1)} \cdots W^{(k+1)} &= (v_k^T X W^{(1)} \cdots W^{(k)}) W^{(k+1)} \\ &= \begin{bmatrix} \otimes & \ldots & \otimes & \gamma_k & \phi_{k+1} & 0 & \ldots & 0 \end{bmatrix} = v_k^T X W^{(1)} \cdots W^{(m)} = v_k^T X W \end{aligned} \tag{3.4}$$

(see (3.2)), just as in the direct method in [9] (where the $\otimes$ denote zero elements, but for [2] these are unknowns to be proven zero). But then v_{k+1} in [2] was obtained from an *iterative* version in [9]. From the $(k+1)$st column of $XW = V\begin{bmatrix} J \\ 0 \end{bmatrix} = V_m J$ they effectively compute

$$v_{k+1}, \quad \gamma_{k+1}, \quad \text{where} \quad v_{k+1}\gamma_{k+1} = Xw_{k+1} - v_k\phi_{k+1}, \quad \|v_{k+1}\|_2 = 1, \ (\phi_1 \equiv 0). \tag{3.5}$$

They do not compute W but overwrite X in order by $XW^{(1)}$, $XW^{(1)}W^{(2)}$, etc., and in the kth step use this to compute $v_k^T(XW^{(1)} \cdots W^{(k)})$ and design $W^{(k+1)}$ on the resulting vector. A rough summary (of our very simplified version of [2]) is then

$$t := Xe_1, \quad \gamma_1 := \|t\|_2, \quad v_1 := t/\gamma_1; \quad \text{see (3.5) with } \ W^{(1)} \equiv I; \quad XW^{(1)} = X. \tag{3.6}$$

For $k = 1, \ldots, m-1$,

$$\text{form } v_k^T(XW^{(1)} \cdots W^{(k)}) = \begin{bmatrix} \otimes & \ldots & \otimes & \gamma_k & \times & \times & \ldots & \times \end{bmatrix}; \tag{3.7}$$

design ϕ_{k+1} and the Householder matrix $W^{(k+1)}$ as in (3.4);

overwrite $(XW^{(1)} \cdots W^{(k)})$ by $(XW^{(1)} \cdots W^{(k)})W^{(k+1)}$;

$$\text{take } t := (XW^{(1)} \cdots W^{(k+1)})e_{k+1} = Xw_{k+1}; \text{ see (3.3)}; \tag{3.8}$$

$$t := t - v_k\phi_{k+1}, \quad \gamma_{k+1} := \|t\|_2, \quad v_{k+1} := t/\gamma_{k+1}, \text{ giving (3.5)}. \tag{3.9}$$

For simplicity we have assumed that $\gamma_1, \gamma_2, \ldots$ are all nonzero. The proof that the elements marked $\otimes$ are actually zero, and that the bidiagonalization (3.1) is obtained, follows from [2] or [9], but for completeness we give a short proof here.

We see from (3.5) that $XW = V_m J$, where $W = [w_1, \ldots, w_m]$, $V_m = [v_1, \ldots, v_m]$, with J as in (3.1). $W = W^{(1)} \cdots W^{(m)}$ is orthogonal since the $W^{(j)}$ are Householder transformations. We first seek to prove by induction that V_m has orthonormal columns. All $\|v_j\|_2 = 1$ from (3.6) and (3.9), so now

$$\text{assume } [v_1, \ldots, v_k] \in \mathbb{R}^{n \times k} \text{ has orthonormal columns for some } 1 \leq k < m \tag{3.10}$$

(which is true for $k = 1$). Then from (3.5), and (3.4) with (3.3),

$$v_k^T v_{k+1}\gamma_{k+1} = v_k^T X w_{k+1} - \phi_{k+1} = \phi_{k+1} - \phi_{k+1} = 0.$$

For $j < k$, (3.5), (3.10), and then (3.3) show that $v_j^T v_{k+1}\gamma_{k+1} = v_j^T X w_{k+1} = v_j^T XWe_{k+1}$. But from (3.4) we see that $v_j^T XW^{(1)} \cdots W^{(j+1)} = v_j^T XW$, and since for $j < k$, $v_j^T XW^{(1)} \cdots W^{(j+1)}e_{k+1} = 0$, we have for $j = 1, \ldots, k-1$

$$v_j^T v_{k+1}\gamma_{k+1} = v_j^T X w_{k+1} = v_j^T XWe_{k+1} = v_j^T XW^{(1)} \cdots W^{(j+1)}e_{k+1} = 0,$$

so that $[v_1, \ldots, v_{k+1}]$ has orthonormal columns, and so does V_m by induction. Then $V_m^T X W = V_m^T V_m J = J$, and each element marked $\otimes$ above is actually zero.

It might be thought that there is another useful algorithm for upper bidiagonalization—one which derives the w_k from the *iterative* algorithm in [9] and the v_k from the *direct* algorithm, also in [9]. But this would essentially be applying the algorithm in [2] to X^T and so is not new. What is more, it would be applying Householder transformations to the side of X with higher dimension and would be costly.

3.1. Rounding error analysis. We now give a relatively quick rounding error analysis of our version, (3.6)–(3.9), of the algorithm in [2] using a computer with unit roundoff ϵ. We will use $\mathrm{fl}(\cdot)$ to denote the result of a floating point computation. We will sometimes use a tilde above a symbol to denote a computed quantity, so if V_k is an ideal mathematical quantity, $\widetilde{V}_k$ will denote its actual computed value or something very close. Matrices E and F, and matrices and vectors whose first symbol is Δ, such as ΔV_k, will denote rounding error terms. To save space we will simply state some of the more obvious results, and to make life easy for the reader we will just write $O(\epsilon)$ and hide the dependence on the dimensions or step number. For those interested, the full detailed results can be found in [2, section 3] or derived via [12].

Let $\widetilde{J}$, $\tilde{\gamma}_k$, $\tilde{\phi}_k$, $\underline{\tilde{v}}_k$, and $\widetilde{X}^{(k)}$ be the computed values of the ideal J, γ_k, ϕ_k, v_k, and $X^{(k)} \equiv XW^{(1)} \cdots W^{(k)}$, and, special to this approach, let $\tilde{v}_k$ be $\underline{\tilde{v}}_k$ normalized to unity so that with $\widetilde{V}_m \equiv [\tilde{v}_1, \ldots, \tilde{v}_m]$ and $\underline{\widetilde{V}}_m \equiv [\underline{\tilde{v}}_1, \ldots, \underline{\tilde{v}}_m]$ we have $\|\widetilde{V}_m - \underline{\widetilde{V}}_m\|_2 \leq O(\epsilon)$. Let $\widehat{W}^{(k+1)}$ be the exact Householder transformation required in step k of the *computed* process (see (3.4)), so that $\widehat{W} \equiv \widehat{W}^{(1)} \cdots \widehat{W}^{(m)}$ is an orthogonal matrix, even though it might differ significantly from the ideal $W \equiv W^{(1)} \cdots W^{(m)}$.

We first derive the key expressions involving the computed values. It follows from Higham [12, Lemma 19.3] (see [2, (3.24), (3.25)]) that for $k = 1, \ldots, m$

$$\widetilde{X}^{(k)} = (X + \Delta X^{(k)})\widehat{W}^{(1)} \cdots \widehat{W}^{(k)}, \tag{3.11}$$
$$\|\Delta X^{(k)}\|_F \leq O(\epsilon)\|X\|_F, \quad \|\widetilde{X}^{(k)}\|_{2,F} = \|X\|_{2,F}[1+O(\epsilon)]; \qquad k = 1, \ldots, m.$$

If $\widetilde{W} \equiv \mathrm{fl}(\mathrm{fl}(\cdots \mathrm{fl}(\widehat{W}^{(1)}\widehat{W}^{(2)})\cdots)\widehat{W}^{(m)})$ was computed, it would be almost orthogonal:

$$\widetilde{W} = \widehat{W} + E_0, \quad \widehat{W}^T = \widehat{W}^{-1}, \quad \|E_0\|_F \leq O(\epsilon) \quad \text{(see, for example, [2, (3.23)])}, \tag{3.12}$$

since $\widehat{W}$ is a product of Householder transformations. Such results follow from the work of Wilkinson [25] on backward stable algorithms, but the analysis of an algorithm like this which is not stable in that beautiful sense is more demanding. For example, $\widetilde{V}_m$ here can be far from orthonormal because of cancellation followed by magnification of rounding errors in (3.9), and this has to be quantified. As a first step toward this, it is straightforward to show from (3.11), (3.6), (3.8), and (3.9) (see (3.5)) that

$$X\widehat{W} = \widetilde{X}^{(m)} + F_1' = \widetilde{V}_m \widetilde{J} + F_1; \quad \|F_1\|_F, \|F_1'\|_F \leq O(\epsilon)\|X\|_F; \tag{3.13}$$

see also [2, (3.24)–(3.27)]. Then $X = \widetilde{V}_m \widetilde{J}\widehat{W}^T + F_1\widehat{W}^T$ (see [2, (3.28)–(3.29)]), and these two results closely mimic the corresponding ideal behavior obtained from (3.1).

At this point we diverge from the analysis in [2] and model the first expression in (3.1). Since $\tilde{\gamma}_k$, $\tilde{\phi}_{k+1}$, and the last $m-k-1$ zeros of (3.4) are not altered by the later application of $\widehat{W}^{(k+2)} \cdots \widehat{W}^{(m)}$, it will be shown from the development of (3.7) (see (3.4)) that for some F_2 and SLT L (possibly large)

$$\widetilde{V}_m^T X \widehat{W} = L + \widetilde{J} + F_2, \quad \|F_2\|_F \leq O(\epsilon)\|X\|_F, \tag{3.14}$$

where we now prove this rigorously. With (3.11) the computation in (3.7) will give (see, for example, [12, section 3.5])

$$\tilde{g}_k^T \equiv \mathrm{fl}(\tilde{v}_k^T \widetilde{X}^{(k)}) = \tilde{v}_k^T(\widetilde{X}^{(k)} + \Delta X^{(k)'}), \qquad \|\Delta X^{(k)'}\|_F \le O(\epsilon)\|X\|_F,$$
$$\|\tilde{g}_k\|_2 \le \|X\|_2[1 + O(\epsilon)].$$

For the computed γ_k and ϕ_{k+1} in (3.7) and (3.4) (see [12, Theorem 19.4] with (3.11)), we have for some $\lambda_{k,1}, \ldots, \lambda_{k,k-1}$ and $\Delta\tilde{g}_k$ with $\|\Delta\tilde{g}_k\|_2 \le O(\epsilon)\|\tilde{g}_k\|_2$

$$(\lambda_{k,1}, \ldots, \lambda_{k,k-1}, \tilde{\gamma}_k, \tilde{\phi}_{k+1}, 0, \ldots, 0) = \mathrm{fl}(\tilde{g}_k^T \widehat{W}^{(k+1)}) = (\tilde{g}_k + \Delta\tilde{g}_k)^T \widehat{W}^{(k+1)}$$
$$= (\tilde{g}_k + \Delta\tilde{g}_k)^T \widehat{W}^{(k+1)} \widehat{W}^{(k+2)} \cdots \widehat{W}^{(m)}$$
$$= [\tilde{v}_k^T(\widetilde{X}^{(k)} + \Delta X^{(k)'}) + \Delta\tilde{g}_k^T]\widehat{W}^{(k+1)} \cdots \widehat{W}^{(m)} = \tilde{v}_k^T X\widehat{W} + \Delta g_k^T,$$
$$\Delta g_k^T \equiv \tilde{v}_k^T(\Delta X^{(k)}\widehat{W} + \Delta X^{(k)'}\widehat{W}^{(k+1)} \cdots \widehat{W}^{(m)}) + \Delta\tilde{g}_k^T \widehat{W}^{(k+1)} \cdots \widehat{W}^{(m)},$$

so that $\|\Delta g_k\|_2 \le \|\Delta X^{(k)}\|_2 + \|\Delta X^{(k)'}\|_2 + \|\Delta\tilde{g}_k\|_2 \le O(\epsilon)\|X\|_F$, proving (3.14).

We are now in a position to make use of Theorem 2.1. Note that (3.13) and (3.14) implicitly give information on $\widetilde{V}_m^T\widetilde{V}_m$. Define $U \equiv \mathrm{sut}(\widetilde{V}_m^T\widetilde{V}_m)$ so that $\widetilde{V}_m^T\widetilde{V}_m = U^T + I_m + U$, and SUT $S \equiv (I+U)^{-1}U$ so that $U = (I-S)^{-1}S$; see Corollary 2.3 and (2.2). We will obtain an expression for U, use this to obtain an expression for S, and use S to produce Q_1 such that $Q_1^TQ_1 = I_m$. Combining (3.13) and (3.14)

$$\widetilde{V}_m^T X\widehat{W} = L + \widetilde{J} + F_2 = \widetilde{V}_m^T(\widetilde{V}_m\widetilde{J} + F_1) = U^T\widetilde{J} + \widetilde{J} + U\widetilde{J} + \widetilde{V}_m^T F_1. \tag{3.15}$$

Canceling the lone $\widetilde{J}$ on each side and equating the SUT parts

$$U\widetilde{J} = F_3 \equiv \mathrm{sut}(F_2 - \widetilde{V}_m^T F_1), \qquad \|F_3\|_F \le O(\epsilon)\|X\|_F. \tag{3.16}$$

This is the desired expression for U. Now, for S, since $\|S\|_2 \le 1$ from (2.5),

$$F_3 = U\widetilde{J} = (I-S)^{-1}S\widetilde{J}, \qquad S\widetilde{J} = (I-S)F_3, \qquad \|(I-S)F_3\|_F \le O(\epsilon)\|X\|_F. \tag{3.17}$$

But from (3.13) $\widetilde{V}_m(I-S)\widetilde{J} = X\widehat{W} - F_1 - \widetilde{V}_m S\widetilde{J} = X\widehat{W} - F_1 - \widetilde{V}_m(I-S)F_3$, so

$$Q_1\widetilde{J} \equiv \begin{bmatrix} S \\ \widetilde{V}_m(I-S) \end{bmatrix}\widetilde{J} = \begin{bmatrix} (I-S)F_3\widehat{W}^T \\ X - [F_1 + \widetilde{V}_m(I-S)F_3]\widehat{W}^T \end{bmatrix}\widehat{W} \equiv \begin{bmatrix} E_1 \\ X + E_2 \end{bmatrix}\widehat{W}, \tag{3.18}$$

where $\|E_1\|_F, \|E_2\|_F \le O(\epsilon)\|X\|_F$; $\widehat{W}^{-T} = \widehat{W}$; and $Q_1^TQ_1 = I_m$; see Theorem 2.1. This shows that the computation for $\widetilde{J}$ is essentially *backward stable*—$\widetilde{J}$ is the *exact* bidiagonal matrix produced by orthogonally transforming the "nearby" data matrix $\left[\begin{smallmatrix} E_1 \\ X+E_2 \end{smallmatrix}\right]$. This with (3.12) shows that the computation of $\widetilde{W}$ is also essentially backward stable. However, the computation of $\widetilde{V}_m$ is not. From (3.16) we see that $\|U\|_2 = \|\mathrm{sut}(\widetilde{V}_m^T\widetilde{V}_m)\|_2$, one measure of the loss of orthogonality in $\widetilde{V}_m$, can be large if $\widetilde{J}$ is ill conditioned. But here $\widetilde{V}_m$ *is* part of the orthonormal Q_1 in (3.18).

With the true SVD of $\widetilde{J}$ we have for $\Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \ge 0$ (see (3.18))

$$\widetilde{J} = P\Sigma Z^T; \quad \begin{bmatrix} E_1 \\ X + E_2 \end{bmatrix}(\widehat{W}Z) = (Q_1P)\Sigma; \quad (Q_1P)^T\begin{bmatrix} E_1 \\ X + E_2 \end{bmatrix} = \Sigma(\widehat{W}Z)^T; \tag{3.19}$$
$$(\widehat{W}Z)^T = (\widehat{W}Z)^{-1}; \qquad (Q_1P)^T(Q_1P) = I; \qquad \|E_1\|_F, \|E_2\|_F \le O(\epsilon)\|X\|_F.$$

If $\widetilde{P}$, $\widetilde{\Sigma}$, and $\widetilde{Z}$ are the computed values of P, Σ, and Z by a numerically stable algorithm, the computed value of $\widetilde{W}\widetilde{Z} = (\widehat{W} + E_0)\widetilde{Z}$ (see (3.12)) will be within $O(\epsilon)$ of the exact matrix of right-hand singular vectors for a matrix very close to $\left[\begin{smallmatrix} 0 \\ X \end{smallmatrix}\right]$, while $\widetilde{\Sigma}$ gives the singular values of such a matrix, too, and we clearly have backward stability for computing these two objects. It follows that the ordered singular values of $\widetilde{J}$ will be within $O(\epsilon)\|X\|_F$ of those of X, so as long as $\sigma_{\min}(X) \gg \epsilon$, from (3.16) (see [2, (3.85)])

$$\|U\|_F = \|\mathrm{sut}(\widetilde{V}_m^T \widetilde{V}_m)\|_F \le O(\epsilon).\|X\|_F/\sigma_{\min}(X). \tag{3.20}$$

Large $\|U\|_F$ makes the analysis for $\widetilde{V}_m P$ more challenging. Consider the partitioning

$$\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2), \quad \Sigma_1 \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_k), \quad \sigma_1 \ge \cdots \ge \sigma_m,$$
$$P = [P_1, P_2], \quad Z = [Z_1, Z_2]; \quad P_1,\ Z_1 \in \mathbb{R}^{m\times k}.$$

Substituting $\widetilde{J} = P\Sigma Z^T$ into (3.13) gives

$$X(\widetilde{W}Z) = (\widetilde{V}_m P)\Sigma + F_1 Z, \quad X(\widetilde{W}Z_1) = (\widetilde{V}_m P_1)\Sigma_1 + F_1 Z_1, \tag{3.21}$$

while (3.17) with $\sigma_k > 0$ leads to

$$SP\Sigma = (I - S)F_3 Z, \qquad SP_1\Sigma_1 = (I - S)F_3 Z_1, \qquad SP_1 = (I - S)F_3 Z_1 \Sigma_1^{-1}.$$

If k in the partitioning is chosen so that $\|X\|_F/\sigma_k$ is not too large compared to 1, then from (3.17) we can say in this same sense that $\|SP_1\|_F \le O(\epsilon)$. But from (3.18) the first k columns of Q_1 are orthonormal, and therefore so are those of

$$Q_1 P_1 = \begin{bmatrix} SP_1 \\ \widetilde{V}_m(I - S)P_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \widetilde{V}_m P_1 \end{bmatrix} + \begin{bmatrix} I \\ -\widetilde{V}_m \end{bmatrix} SP_1 = \begin{bmatrix} 0 \\ \widetilde{V}_m P_1 \end{bmatrix} + O(\epsilon),$$

so $\widetilde{V}_m P_1$ has almost orthonormal columns. This along with (3.19) and (3.12) gives

$$\|X(\widetilde{W}Z_1) - (\widetilde{V}_m P_1)\Sigma_1\|_F \le O(\epsilon)\|X\|_F, \qquad \|(\widetilde{V}_m P_1)^T X - \Sigma_1(\widetilde{W}Z_1)^T\|_F \le O(\epsilon)\|X\|_F,$$
$$\|(\widetilde{W}Z_1)^T(\widetilde{W}Z_1) - I_k\|_F \le O(\epsilon), \qquad \|(\widetilde{V}_m P_1)^T(\widetilde{V}_m P_1) - I_k\|_F \le O(\epsilon),$$

proving that the columns of $\widetilde{V}_m P_1$ and $\widetilde{W}Z_1$ are excellent near-orthonormal left and right singular vectors of X. This is an alternative proof of the result obtained in [2, p. 35]: "The application of any backward stable singular value decomposition procedure to B" ($\widetilde{J}$ here) "recovers the left singular vectors associated with the leading (largest) singular values of X to near orthogonality."

Of course there is far more of interest and use in [2] than is mentioned in this quick sketch designed to exhibit the usefulness of Theorem 2.1, and those interested in this bidiagonalization algorithm must refer to the far more precise [2]. A second working of a result can often be shorter and clearer than the original it is based on, but the extent of the simplicity induced by the present full use of Theorem 2.1 ([2] did make use of Sheffield's original observation) still seems impressive, so we suspect it will bring clarity and simplicity to our understanding of other important algorithms, including those in [5, 6, 18].

Also, since the form of the orthonormal Q_1 in (3.18) is known in terms of $\widetilde{V}_m$ (see Theorem 2.1), it might be possible to use this knowledge to produce or apply *all* the left singular vectors of X in a backward stable manner from (3.18), perhaps paralleling what was done in [5, Algorithm 6.1] for solutions of least squares problems.

4. Transforming augmented into standard results. The use of Theorem 2.1 leads to expressions involving $Q_1^{(k)} \in \mathcal{U}^{(n+k)\times k}$ (see, for example, (3.18)), where the finite precision computation produced a matrix within $O(\epsilon)$ of $V_k \notin \mathcal{U}^{n\times k}$. There are several instances where we need to prove the existence of a closely related $\widehat{V}_k \in \mathcal{U}^{n\times k}$; see, for example, [5]. The following general theorem is useful in this regard. It is more flexible than [5, Lemma 3.1] and provides a new bound.

THEOREM 4.1. *For any matrices $n\times p$ X and E'', $m\times p$ E', $m\times s$ Q_{11}, $n\times s$ Q_{21}, and $s\times p$ R satisfying*

$$\begin{bmatrix}0\\X\end{bmatrix} + E \equiv \begin{matrix}m\{\\n\{\end{matrix}\begin{bmatrix}E'\\X+E''\end{bmatrix} = \begin{bmatrix}Q_{11}\\Q_{21}\end{bmatrix} R, \qquad Q_{11}^H Q_{11} + Q_{21}^H Q_{21} = I, \tag{4.1}$$

and any dimensions m, n, s, p with $s \le n$ (if originally $s > n$, the matrices X, E'', and Q_{21} can each be padded with $s-n$ rows of zeros to give a new $n' = s$), there exist $n\times s$ $\widehat{V}_s$ and F, and $n\times p$ $\widehat{E}$, such that

$$\begin{aligned} &X + \widehat{E} = \widehat{V}_s R, \quad \widehat{V}_s^H \widehat{V}_s = I_s, \quad \widehat{E} = F Q_{11}^H E' + E'', \quad \|\widehat{E}\|_2 \le \sqrt{2}\|E\|_2, \\ &\widehat{V}_s - Q_{21} = F Q_{11}^H Q_{11}, \quad (\widehat{V}_s - Q_{21})R = F Q_{11}^H E', \quad 0.5 \le \|F\|_2 \le 1. \end{aligned} \tag{4.2}$$

Proof. Suppose Q_{21} has the SVD $Q_{21} = W_1 \Sigma Z^H$, where $W \equiv [W_1, W_2]$ and Z are square unitary matrices, and Σ is $s \times s$ diagonal, $0 \le \Sigma \le I$. If Σ is singular, or even zero, W_1 and Z are somewhat arbitrary. Then (4.1) gives

$$Z^H Q_{11}^H Q_{11} Z = I - \Sigma^2 = (I+\Sigma)(I-\Sigma), \quad \text{with } 0.5I \le (I+\Sigma)^{-1} \le I.$$

Define $\widehat{V}_s \equiv W_1 Z^H$ so that $\widehat{V}_s^H \widehat{V}_s = I$. Here $\widehat{V}_s$ is the closest matrix with orthonormal columns to $Q_{21} = W_1 \Sigma Z^H$ in any unitarily invariant norm. Then

$$\begin{aligned} \widehat{V}_s - Q_{21} &= W_1(I-\Sigma)Z^H = W_1(I+\Sigma)^{-1} Z^H Q_{11}^H Q_{11}, \\ (\widehat{V}_s - Q_{21})R &= W_1(I+\Sigma)^{-1} Z^H Q_{11}^H E'. \end{aligned}$$

Setting $F \equiv W_1(I+\Sigma)^{-1}Z^H$ gives the second line of (4.2). With $N \equiv [F Q_{11}^H, I]$,

$$\widehat{E} = \widehat{V}_s R - X = (\widehat{V}_s - Q_{21})R + E'' = F Q_{11}^H E' + E'' = NE.$$

But $NN^H = I + F Q_{11}^H Q_{11} F^H = I + W_1(I+\Sigma)^{-1}(I-\Sigma^2)(I+\Sigma)^{-1} W_1^H$, showing that $\|NN^H\|_2 \le 2$ and completing the proof. ∎

We can see by taking $s = p = m$ and $R = \widetilde{J}\widehat{W}^T$ that (4.1) includes (3.18), and Theorem 4.1 could have been used there.

5. Some properties of V_k, S_k, and $Q^{(k)}$ in Theorem 2.1. Since Theorem 2.1 appears to be a generally useful result, for later reference we will give some corollaries to Theorem 2.1 that describe properties of V_k, S_k, and $Q^{(k)}$.

COROLLARY 5.1. *In Theorem* 2.1, S_k *and* U_k *are SUT, so* $S_k^k = U_k^k = 0$. *Also* $I+U_k=(I-S_k)^{-1}$, $S_k=U_k-S_kU_k$, $\|S_k\|_2 \le 1$, *so* $\|S_k\|_2 \le 2\|U_k\|_2$ *and*

$$(I-S_k)^{-1} = I + S_k + \cdots + S_k^{k-1}, \qquad \|(I-S_k)^{-1}\|_2 \le k;$$
$$U_k = S_k(I-S_k)^{-1} = (I-S_k)^{-1}S_k = S_k + \cdots + S_k^{k-1}, \quad \|U_k\|_2 \le (k-1)\|S_k\|_2 \le k-1;$$
$$S_k = (I+U_k)^{-1}U_k = U_k(I+U_k)^{-1} = U_k - U_k^2 + \cdots (-1)^k U_k^{k-1};$$
$$\|(I-S_k)^{-1}\|_2 \le \begin{cases} 1/(1-\|S_k\|_2), & \textit{which is best when } \|S_k\|_2 \le 1-k^{-1}; \\ k, & \textit{which is best when } \|S_k\|_2 > 1-k^{-1}; \end{cases}$$
$$\|S_k\|_2/2 \le \|S_k\|_2/(1+\|S_k\|_2) \le \|U_k\|_2 \le \begin{cases} \|S_k\|_2/(1-\|S_k\|_2) \textit{ always}; \\ (k-1)\|S_k\|_2, \textit{best when } \|S_k\|_2 > \frac{(k-2)}{(k-1)}; \end{cases}$$
$$\|U_k\|_2/k \le \|U_k\|_2/(1+\|U_k\|_2) \le \|S_k\|_2 \le \begin{cases} \|U_k\|_2/(1-\|U_k\|_2) \textit{ if } \|U_k\|_2 \le 1; \\ 2\|U_k\|_2, \textit{ best when } \|U_k\|_2 > 0.5. \end{cases}$$

Proof. These results follow from simple matrix and norm manipulation; remember that $(1+\|X\|_2)^{-1} \le \|(I \pm X)^{-1}\|_2$, while $\|X\|_2 < 1 \Leftrightarrow \|(I \pm X)^{-1}\|_2 \le (1-\|X\|_2)^{-1}$. Here since $(I-X)^{-1}$ always exists, we allow $\|X\|_2 \le 1$. ∎

COROLLARY 5.2 (see [18, Lemma 5.1]). *For* V_k, U_k, *and* S_k *in Theorem* 2.1,

$$1 - 2\|U_k\|_2 \le \frac{1-\|S_k\|_2}{1+\|S_k\|_2} \le \sigma_i^2(V_k) \le 1 + 2\|U_k\|_2 \le \frac{1+\|S_k\|_2}{1-\|S_k\|_2}, \tag{5.1}$$

$$\sigma_{min}(V_k) \le 1 \le \sigma_{max}(V_k), \qquad \kappa_2(V_k) \le \frac{1+\|S_k\|_2}{1-\|S_k\|_2}. \tag{5.2}$$

Proof. The first and fourth inequalities in (5.1) follow from Corollary 5.1. The third and second are proven as follows. From (2.3) $(I-S_k)^H V_k^H V_k (I-S_k) = I - S_k^H S_k$, so for any $y \in \mathbb{C}^k$ such that $\|y\|_2 = 1$, define $z \equiv (I-S_k)y$ so $\|z\|_2 \le 1 + \|S_k\|_2$ to give

$$1 + 2\|U_k\|_2 \ge 1 + \frac{z^H(U_k+U_k^H)z}{z^Hz} = \frac{z^HV_k^HV_kz}{z^Hz} = \frac{1-y^HS_k^HS_ky}{z^Hz} \ge \frac{1-\|S_k\|_2^2}{(1+\|S_k\|_2)^2},$$

and then cancel $1+\|S_k\|_2$. For (5.2) taking $y = e_1$ in $\sigma_{min}^2(V_k) \le \|V_ky\|_2^2 \le \sigma_{max}^2(V_k)$, $\|y\|_2 = 1$, proves the first part, while the bound on $\kappa_2(V_k)$ follows from (5.1). ∎

Corollary 5.2 can be very useful; see, for example, [18, (5.7) and section 6]. Giraud and Langou [8] proved under mild conditions that for MGS applied to $n \times k$ B, the matrix V_k of computed supposedly orthonormal vectors is well conditioned. Theorem 2.1 generalizes their work, leading to the more general result (5.2), where for *any* algorithm, bounding $\|S_k\|_2 < 1$ bounds $\kappa_2(V_k)$; see, for example, the bound for MGS in (2.12). We see that V_k can be very well conditioned even when significant orthogonality is lost. For example if $\|S_k\|_2 = .9$, corresponding to a severe loss of orthogonality in V_k, (5.2) shows that $\kappa_2(V_k) \le 19$, which is surprisingly and pleasingly small.

COROLLARY 5.3. *For* V_k *and* $k \times k$ *SUT* S_k *in Theorem* 2.1,

(5.3) *for the Frobenius norm,* $\|(I_k-S_k)^{-1}\|_F^2 \le k(k+1)/2$, *which is tight,*

(5.4) $\|S_k\|_F^2 + \|V_k(I_k-S_k)\|_F^2 = k = \|S_k\|_F^2 + \|V_k(I_k-S_k)^H\|_F^2$,

(5.5) $\|S_k\|_F^2 \le k-1$, so $\|V_k(I_k-S_k)\|_F^2 = \|V_k(I_k-S_k)^H\|_F^2 = k - \|S_k\|_F^2 \ge 1$,

(5.6) $\|I_n - V_k(I_k-S_k)V_k^H\|_F^2 = n - \|V_k(I_k-S_k)\|_F^2 = n - k + \|S_k\|_F^2 \le n-1$.

Proof. For (5.3), write $G_k \equiv (I-S_k)^{-1}$. Then for $k=1$, $\|G_k\|_F^2 = 1 = k(k+1)/2$. Suppose we have proven (5.3) for some $k \geq 1$; then from (2.7)

$$G_{k+1} = \begin{bmatrix} (I-S_k)^{-1} & (I-S_k)^{-1}s_{k+1} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} G_k & V_k^H v_{k+1} \\ 0 & 1 \end{bmatrix},$$
$$\|G_{k+1}\|_F^2 = \|G_k\|_F^2 + \|V_k^H v_{k+1}\|_2^2 + 1 \leq \|G_k\|_F^2 + k + 1 \leq (k+1)(k+2)/2,$$

proving the bound in (5.3). Now consider $V_k = ve^H$, $v^H v = 1$, $e \equiv (1,1,\ldots,1)^H$. Here $V_j^H v_{j+1} = e$, so $\|G_k\|_F^2 = 1+2+\cdots+k = k(k+1)/2$, proving it is tight. The others hold because $Q^{(k)}$ is a unitary matrix in (2.6). ∎

If we add v_{k+1} to V_k, then from (2.7)

$$\begin{bmatrix} [S_k, s_{k+1}] \\ V_{k+1}(I-S_{k+1}) \end{bmatrix} = \begin{bmatrix} S_k & s_{k+1} \\ V_k(I-S_k) & v_{k+1} - V_k s_{k+1} \end{bmatrix} \equiv [Q_1^{(k)}, q^{(k)}], \quad \text{say,} \tag{5.7}$$

has orthonormal columns (it is just $Q_1^{(k+1)}$ with row $(k+1)$, which is a zero row, removed). We now relate $Q_2^{(k)}$ in (2.6) to $q^{(k)}$ in (5.7). From (2.7)

$$q^{(k)} = \begin{bmatrix} s_{k+1} \\ v_{k+1} - V_k s_{k+1} \end{bmatrix} = \begin{bmatrix} (I-S_k)V_k^H v_{k+1} \\ v_{k+1} - V_k(I-S_k)V_k^H v_{k+1} \end{bmatrix} = Q_2^{(k)} v_{k+1}, \tag{5.8}$$

again showing that $Q_1^{(k)T} q^{(k)} = 0$. Since this develops $Q_1^{(k+1)}$ from $Q_2^{(k)}$ and v_{k+1}, this is a more thorough development than that of S_k in (2.7). Note that if $V_k^H v_{k+1} = 0$, then $s_{k+1} = 0$ in (5.7), and the last column of $Q_1^{(k+1)}$ is $\left[\begin{smallmatrix} 0 \\ v_{k+1} \end{smallmatrix}\right]$.

5.1. Some bounds involving $\|V_k\|_2$. Here are some general results for $V_k \equiv [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$ with $\|v_j\|_2 = 1$, $j = 1, \ldots, k$. First we will show that

$$\|[V_k, I_n]\|_F^2 = n + k; \qquad \|[V_k, I_n]\|_2^2 \leq k+1, \quad \text{which is tight,} \tag{5.9}$$
$$\|[-I_k, V_k^H]\|_F^2 = 2k; \qquad \|[-I_k, V_k^H]\|_2^2 \leq k+1, \quad \text{which is tight.} \tag{5.10}$$

In (5.9) the F-norm result is obvious, while the 2-norm result follows from

$$\|[V_k, I_n]\|_2^2 = \|I_n + V_k V_k^H\|_2 = 1 + \|V_k V_k^H\|_2 = 1 + \|V_k\|_2^2 \leq 1 + \|V_k\|_F^2 = 1 + k,$$

with the example $V_k = ve^T$, $\|v\|_2 = 1$. A similar argument proves (5.10).

LEMMA 5.4. *Let $V_k \equiv [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$, $\|v_j\|_2 = 1$, $j = 1, \ldots, n+1$. Then*

$$\|V_k^H v_{k+1}\|_2^2 \geq \frac{k}{n(n+1)} \quad \textit{for at least one } k \in \{1, \ldots, n\}.$$

Proof. Let $V_k^H V_k \equiv I_k + U_k + U_k^H$, U_k SUT. Since V_{n+1} has linearly dependent columns, there exists y, $\|y\|_2 = 1$, such that $V_{n+1}y = 0$, and then $(U_{n+1} + U_{n+1}^H)y = -y$, and $U_{n+1} + U_{n+1}^H$ has a singular value of one. It follows that

$$1 \leq \|U_{n+1} + U_{n+1}^H\|_F^2 = 2\|U_{n+1}\|_F^2 = 2\sum_{k=1}^{n} \|V_k^H v_{k+1}\|_2^2.$$

Suppose the lemma is false; then

$$\sum_{k=1}^{n} \|V_k^H v_{k+1}\|_2^2 < \frac{1}{n(n+1)} \sum_{k=1}^{n} k = \frac{1}{2},$$

which is a contradiction, so the lemma is true. ∎

Among other things, Corollary 5.2 gave a lower bound on the minimum singular value of V_k in terms of $\|S_k\|_2$ or $\|U_k\|_2$. If we give particular bounds on the columns of U_k, we can obtain more precise results, as we now show.

COROLLARY 5.5. *With the notation in Lemma* 5.4,

$$\|V_i^H v_{i+1}\|_2^2 < \frac{i}{k(k+1)} \quad \textit{for} \quad i = 1, \ldots, k-1 \quad \Rightarrow \quad \sigma_{\min}(V_k)^2 > \frac{1}{k+1}.$$

Proof. For any $y \in \mathbb{C}^k$ with $\|y\|_2 = 1$, the left-hand side of the implication gives

$$\begin{aligned}
&\|V_k y\|_2^2 = y^H V_k^H V_k y = 1 + y^H (U_k + U_k^H) y, \\
&(1 - \|V_k y\|_2^2)^2 \le \|U_k + U_k^H\|_2^2 \le \|U_k + U_k^H\|_F^2 = 2\|U_k\|_F^2 \\
&\qquad = 2 \sum_{i=1}^{k-1} \|V_i^H v_{i+1}\|_2^2 < \frac{(k-1)k}{k(k+1)} = \frac{k^2-1}{(k+1)^2} < \frac{k^2}{(k+1)^2},
\end{aligned}$$

so $(1 - \|V_k y\|_2^2)(k+1) < k$. The result follows since

$$1 \ge \sigma_{\min}(V_k) = \min_{\|y\|_2 = 1} \|V_k y\|_2. \qquad \square$$

5.2. Some properties of the eigensystem of $Q^{(k)}$. Finally, there are some interesting results on the eigendecomposition of $Q^{(k)}$.

COROLLARY 5.6. *With the notation in Theorem* 2.1 *we have left and right eigensubspaces of* $Q^{(k)}$

$$[V_k, I_n] Q^{(k)} = [V_k, I_n], \qquad Q^{(k)} \begin{bmatrix} V_k^H \\ I_n \end{bmatrix} = \begin{bmatrix} V_k^H \\ I_n \end{bmatrix}, \tag{5.11}$$

so this gives n *left eigenvectors and the* same n *right eigenvectors of* $(n+k)\times(n+k)$ $Q^{(k)}$*, all with eigenvalue* 1. *For the remaining* k*-dimensional left and right eigenspaces*

$$[-I_k, V_k^H] Q^{(k)} = -(I_k - S_k)^{-H} (I_k - S_k)[-I_k, V_k^H], \tag{5.12}$$

$$Q^{(k)} \begin{bmatrix} -I_k \\ V_k \end{bmatrix} = -\begin{bmatrix} -I_k \\ V_k \end{bmatrix} (I_k - S_k)(I_k - S_k)^{-H}. \tag{5.13}$$

$$\textit{Also} \quad \begin{bmatrix} -I_k & V_k^H \\ V_k & I_n \end{bmatrix} \begin{bmatrix} -I_k & V_k^H \\ V_k & I_n \end{bmatrix} = \begin{bmatrix} I_k + V_k^H V_k & 0 \\ 0 & I_n + V_k V_k^H \end{bmatrix}, \tag{5.14}$$

so the remaining k *eigenvectors will be orthogonal to the* n *eigenvectors from* (5.11). *Since* $Q^{(k)}$ *is a normal matrix, it necessarily has a complete set of orthonormal eigenvectors. The above gives simple representations of two relevant eigenspaces.*

Proof. Using (2.6) and expanding each of the left-hand sides of (5.11) shows that (5.11) is true. For (5.12), from (2.6) with (2.3)

$$\begin{aligned}
&(I_k - S_k)^H [-I_k, V_k^H] Q^{(k)} = (I_k - S_k)^H [-I_k, V_k^H] \left[\begin{array}{c|c} S_k & (I_k - S_k) V_k^H \\ V_k (I_k - S_k) & I_n - V_k (I_k - S_k) V_k^H \end{array}\right] \\
&= [S_k^H S_k - S_k + I_k - S_k^H S_k | - (I_k - S_k)^H (I_k - S_k) V_k^H + (I_k - S_k)^H V_k^H - (I_k - S_k^H S_k) V_k^H] \\
&= -(I_k - S_k) \left[\begin{array}{c|c} -I_k & V_k^H \end{array}\right],
\end{aligned}$$

proving (5.12). Similarly, (5.13) follows from (2.6) with (2.4):

$$\begin{aligned} Q^{(k)}\begin{bmatrix}-I_k\\ V_k\end{bmatrix}(I_k-S_k)^H &= -\begin{bmatrix}S_k-(I_k-S_k)V_k^HV_k\\ V_k(I_k-S_k)-V_k+V_k(I_k-S_k)V_k^HV_k\end{bmatrix}(I_k-S_k)^H\\ &= \begin{bmatrix}-I_k\\ V_k\end{bmatrix}[S_k-(I_k-S_k)V_k^HV_k](I_k-S_k)^H\\ &= \begin{bmatrix}-I_k\\ V_k\end{bmatrix}[S_k-S_kS_k^H-(I_k-S_k)V_k^HV_k(I_k-S_k)^H] = -\begin{bmatrix}-I_k\\ V_k\end{bmatrix}(I_k-S_k). \end{aligned}$$

Equation (5.14) is obvious. ∎

COROLLARY 5.7. *For the matrices in Theorem* 2.1*, let* W_k *be such that*

$$W_k^H(S_k+S_k^H)W_k = \operatorname{diag}(\delta_i), \quad \delta_i \text{ real}, \quad W_k^H = W_k^{-1}; \quad \text{then if} \tag{5.15}$$

$$Y_1^{(k)} \equiv \begin{bmatrix}-I_k\\ V_k\end{bmatrix}(I_k-S_k)^HW_k, \qquad Y_2^{(k)} \equiv \begin{bmatrix}-I_k\\ V_k\end{bmatrix}(I_k-S_k)W_k,$$

each of $Y_1^{(k)}$ *and* $Y_2^{(k)}$ *has orthogonal columns, where*

$$\|Y_1^{(k)}e_i\|_2 = \|Y_2^{(k)}e_i\|_2 = 2-\delta_i, \quad i=1,\dots,k, \quad \text{and} \quad -Q^{(k)}Y_1^{(k)} = Y_2^{(k)} \tag{5.16}$$

so that $Q^{(k)}$ *rotates the columns of* $Y_1^{(k)}$ *into the columns of* $Y_2^{(k)}$ *(both sets of* k *orthogonal vectors lie in the same subspace), and*

$$\delta_i < 2, \quad i=1,\dots,k. \tag{5.17}$$

Proof. $-Q^{(k)}Y_1^{(k)} = Y_2^{(k)}$ follows from (5.13). Then with (2.3) and (5.15)

$$\begin{aligned} &Y_1^{(k)H}Y_1^{(k)} = Y_2^{(k)H}Y_2^{(k)} = W_k^H(I_k-S_k)^H(I_k+V_k^HV_k)(I_k-S_k)W_k\\ &= W_k^H[(I_k-S_k)^H(I_k-S_k)+I_k-S_k^HS_k]W_k = W_k^H[2I_k-S_k-S_k^H]W_k = 2I_k-\operatorname{diag}(\delta_i), \end{aligned}$$

which is diagonal, proving that $Y_i^{(k)}$ has orthogonal columns for $i=1,2$ and completing (5.16). $Y_1^{(k)H}Y_1^{(k)}$ is also Hermitian positive definite, proving (5.17). ∎

COROLLARY 5.8. $Q^{(k)}$ *in Theorem* 2.1 *has* n *eigenvalues of* 1*, and* k *eigenvalues on the unit circle which are not* 1.

Proof. The first part follows from (5.11). We prove the second part by contradiction. Suppose in (5.12) that $-(I_k-S_k)^{-H}(I_k-S_k)$ had an eigenvalue of 1; then there exists y with $y^Hy=1$ such that

$$(I_k-S_k)y = -(I_k-S_k)^Hy, \quad 2y = (S_k+S_k^H)y, \quad y^H(S_k+S_k^H)y = 2,$$

which is impossible from (5.15) and (5.17). ∎

To summarize, $\mathcal{R}(\left[\begin{smallmatrix}V_k^H\\ I_n\end{smallmatrix}\right])$ is the eigenspace of unitary $Q^{(k)}$ corresponding to all its eigenvalues of 1, while $\mathcal{R}(\left[\begin{smallmatrix}-I_k\\ V_k\end{smallmatrix}\right])$ is the remaining eigenspace. If $V_k^HV_k = I_k$ so $S_k = 0$, then from (5.13) the remaining k eigenvalues are -1.

6. Thoughts on optimality of $Q^{(k)}$ and the use of these ideas. There are other ways of obtaining unitary matrices like $Q^{(k)}$ from a sequence of unit 2-norm vectors $v_1,\dots,v_k$, and it might seem natural to ask if the form (2.6) in Theorem 2.1 is optimal in some sense, such as $Q_1^{(k)}$ being as close as possible in some sense to $\left[\begin{smallmatrix}0\\ V_k\end{smallmatrix}\right]$. For example, is SUT S_k in (2.1) as small as possible in some sense, subject to $Q^{(k)}$

being unitary in (2.6)? This seems unlikely, since S_{k+1} is obtained by appending s_{k+1} and a zero row to S_k (see (2.7)), and such sequential additions do not usually lead to overall optima.

We saw from (2.1) that $V_k^H V_k = I \;\Leftrightarrow\; S_k = 0$, so at least $Q_1^{(k)}$ exhibits this characteristic of optimality. Also we saw in Theorem 2.1 that S_k in (2.1) is the *unique* SUT $k \times k$ matrix satisfying (2.6) for a given sequence $v_1, \ldots, v_k$ of unit 2-norm vectors. Thus any further optimality analysis would have to consider some wider range of possibilities. What that range could be is not clear, nor does it seem important for the uses that Theorem 2.1 was designed for, as we will now argue.

The steps in using Theorem 2.1 for a rounding error analysis of the type considered here have so far taken the following form. The $v_1, \ldots, v_k$ are the correctly normalized versions of the supposedly orthonormal vectors computed by the algorithm. Then $U_k \equiv \mathrm{sut}(V_k^H V_k)$ and $S_k \equiv (I_k + U_k)^{-1} U_k$ are fully determined, and either can be used to represent the loss of orthogonality in $v_1, \ldots, v_k$. Initially U_k might seem more natural, but the fact that $\|S_k\|_2 \le 1$, with the rest of (2.5), is very useful.

The crucial step for each analysis is to find an expression involving either U_k or S_k from which it can be rigorously bounded. If U_k is bounded, then S_k can be bounded using (2.1) or (2.2). In the analysis in section 3.1, the 2-norms of U_k and S_k were bounded in (3.16) and (3.17). Theorem 2.1 can then be used to produce an expression involving the computed objects related by a matrix $Q_1^{(k)}$ with *exactly* orthonormal columns. This is done in (3.18). From this expression conclusions can be made regarding the numerical stability of the algorithm, and possible improvements might be suggested. See, for example, the text from (3.18) to the end of section 3.1.

Since S_k is the unique SUT matrix giving the unitary matrix in (2.6), it does not seem to matter whether $Q^{(k)}$ is optimal or not—it appears to be the most useful matrix for the analysis.

7. Augmented biorthogonality. Theorem 2.1 was designed for the analysis of a class of orthogonalization algorithms. At the Householder Symposium 2008 in Zeuthen, Germany, Ron Morgan [16] suggested that this approach might be applicable to certain biorthogonalization algorithms as well. Following his suggestion, we now state a generalization of Theorem 2.1 which might be useful for the rounding error analyses of some biorthogonalization algorithms. In particular, it might be useful for analyzing the numerical behavior of Lanczos's [13, pp. 266 et seq.] process for tridiagonalizing an *unsymmetric* square matrix; see also, for example, Wilkinson [25, pp. 388–394].

THEOREM 7.1. *For any integers $n \ge 1$ and $k \ge 1$, with $V \equiv [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$, $W \equiv [w_1, \ldots, w_k] \in \mathbb{C}^{n\times k}$, where $w_j^H v_j = 1, j = 1, \ldots, k$, define the SUT matrices U, S, and R, and the lower triangular matrix L, as well as the augmented matrices Q and P, as follows:*

$$
\begin{aligned}
&U \equiv \mathrm{sut}(W^H V), \qquad S \equiv (I+U)^{-1}U = I - (I+U)^{-1},\\
&L \equiv \mathrm{slt}(W^H V), \qquad R \equiv (I+L^H)^{-1}L^H = I - (I+L^H)^{-1},\\
&Q \equiv \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \equiv \left[\begin{array}{c|c} S & (I-S)W^H \\ V(I-S) & I - V(I-S)W^H \end{array}\right],\\
&P \equiv \left[\begin{array}{c|c} P_1 & P_2 \end{array}\right] \equiv \left[\begin{array}{c|c} R & (I-R)V^H \\ W(I-R) & I - W(I-R)V^H \end{array}\right].
\end{aligned} \tag{7.1}
$$

Then

$$US = SU, \quad U = (I-S)^{-1}S \equiv S(I-S)^{-1}, \quad (I-S)^{-1} = I+U, \tag{7.2}$$

$$L^H R = RL^H, \quad L^H = (I-R)^{-1}R \equiv R(I-R)^{-1}, \quad (I-R)^{-1} = I+L^H, \tag{7.3}$$

$$(I-R)^H W^H V(I-S) = I - R^H S, \tag{7.4}$$

$$(I-S)W^H V(I-R)^H = I - SR^H, \tag{7.5}$$

$$W^H V = I \Leftrightarrow R = S = 0; \tag{7.6}$$

$$W^H V \text{ singular} \Leftrightarrow R^H S \text{ has an eigenvalue } 1 \quad (\text{and so} \quad \|R^H S\|_2 \geq 1). \tag{7.7}$$

Also R and S are the unique *SUT $k \times k$ matrices such that*

$$P^H Q = I \quad (\text{and so} \quad QP^H = I). \tag{7.8}$$

If we add a column to each of V and W so the expanded matrices are $\widehat{V} \equiv [V, v]$, $\widehat{W} \equiv [W, w]$, $\widehat{U} \equiv \left[\begin{smallmatrix} U & u \\ 0 & 0 \end{smallmatrix}\right]$, $\widehat{R} \equiv \left[\begin{smallmatrix} R' & r \\ 0 & 0 \end{smallmatrix}\right]$, $\widehat{S} \equiv \left[\begin{smallmatrix} S' & s \\ 0 & 0 \end{smallmatrix}\right]$ (giving $\widehat{P}^H \widehat{Q} = I_{n+k+1}$), we also have

$$R' = R, \quad S' = S, \quad \text{and} \quad s = (I-S)W^H v, \quad r = (I-R)V^H w, \tag{7.9}$$

$$\begin{bmatrix} \widehat{S} \\ \widehat{V}(I-\widehat{S}) \end{bmatrix} = \left[\begin{array}{c|c} S & s \\ 0 & 0 \\ V(I-S) & v - Vs \end{array}\right], \qquad \begin{bmatrix} \widehat{R} \\ \widehat{W}(I-\widehat{R}) \end{bmatrix} = \left[\begin{array}{c|c} R & r \\ 0 & 0 \\ W(I-R) & w - Wr \end{array}\right].$$

Proof. We start with the leading part of (7.8). For any $k \times k$ SUT matrices $\widetilde{S}$ and $\widetilde{R}$, define $N \equiv \widetilde{P}_1^H \widetilde{Q}_1 - I$ for $\widetilde{P}_1 \equiv \left[\begin{smallmatrix} \widetilde{R} \\ W(I-\widetilde{R}) \end{smallmatrix}\right]$, $\widetilde{Q}_1 \equiv \left[\begin{smallmatrix} \widetilde{S} \\ V(I-\widetilde{S}) \end{smallmatrix}\right]$; see (7.1). Since by definition $W^H V = L + I + U$, we have

$$\begin{aligned} I + N = \widetilde{P}_1^H \widetilde{Q}_1 &= \widetilde{R}^H \widetilde{S} + (I-\widetilde{R})^H (I-\widetilde{S}) + (I-\widetilde{R})^H (L+U)(I-\widetilde{S}), \\ N &= -(I-\widetilde{R})^H \widetilde{S} - \widetilde{R}^H (I-\widetilde{S}) + (I-\widetilde{R})^H (L+U)(I-\widetilde{S}), \\ (I-\widetilde{R})^{-H} N (I-\widetilde{S})^{-1} &= -\widetilde{S}(I-\widetilde{S})^{-1} - (I-\widetilde{R})^{-H} \widetilde{R}^H + L + U. \end{aligned}$$

But $U - \widetilde{S}(I-\widetilde{S})^{-1}$ is SUT, while $L - (I-\widetilde{R})^{-H}\widetilde{R}^H$ is SLT, so $N = 0$ if and only if $U = \widetilde{S}(I-\widetilde{S})^{-1}$ and $L = (I-\widetilde{R})^{-H}\widetilde{R}^H$. But then $\widetilde{S} = U - U\widetilde{S}$ and so $(I+U)\widetilde{S} = U$, while $\widetilde{R}^H = L - \widetilde{R}^H L$ so $\widetilde{R}^H (I+L) = L$, proving that R and S in (7.1) are the unique $k \times k$ SUT matrices giving $P_1^H Q_1 = I$.

From (7.1) $U = S + US$, so $U(I-S) = S$, $U = S(I-S)^{-1} = (I-S)^{-1}S$. Also $U(I-S) + (I-S) = I$, so $U + I = (I-S)^{-1}$, and then $US = SU$, proving (7.2). A similar argument for L^H and R proves (7.3). Note that (7.4) follows from $P_1^H Q_1 = I$, and using it gives for the rest of (7.8)

$$\begin{aligned} P_1^H Q_2 &= R^H (I-S) W^H + (I-R)^H W^H - (I-R)^H W^H V (I-S) W^H = 0, \\ P_2^H Q_1 &= V(I-R)^H S + V(I-S) - V(I-R)^H W^H V (I-S) = 0, \\ P_2^H Q_2 &= V(I-R)^H (I-S) W^H + I - V(I-R)^H W^H - V(I-S) W^H \\ &\quad + V(I-R)^H W^H V (I-S) W^H = I, \end{aligned}$$

so that (7.8) holds. Then (7.5) follows from the leading principal $k \times k$ submatrix of $QP^H = I$. The definitions in (7.1) give (7.6) directly, while (7.7) follows from (7.4).

To prove (7.9) note that $u \equiv W^H v$ so that $\widehat{U} e_{k+1} = \left[\begin{smallmatrix} u \\ 0 \end{smallmatrix}\right]$. Now from (7.2) $\widehat{S} = (I_{k+1} - \widehat{S})\widehat{U}$. One proof follows by using this and the fact that $\widehat{U}$ and $\widehat{S}$ are SUT:

$$\left[\frac{S'}{0}\right] = \widehat{S}\left[\frac{I_k}{0}\right] = (I_{k+1} - \widehat{S})\left[\frac{U}{0}\right] = \left[\frac{(I_k - S')U}{0}\right] = \left[\frac{S}{0}\right],$$
$$\left[\frac{s}{0}\right] = \widehat{S} e_{k+1} = (I_{k+1} - \widehat{S})\left[\frac{u}{0}\right] = \left[\frac{(I_k - S)u}{0}\right];$$

see the analogous argument in Theorem 2.1. The proofs for R' and r are similar. ∎

Thus if V and W are supposedly biorthogonal matrices from some computation and satisfy $w_j^H v_j = 1$, $j = 1, \ldots, k$, then the augmented matrices Q_1 and P_1 are *truly* biorthogonal. Note in this case that for any real nonsingular $k \times k$ diagonal matrix D that $D^{-1} P_1^H Q_1 D = I$, and by normalizing the columns of both P_1 and Q_1 we can ensure that $Q_1 D$, for example, has unit length columns if we wish.

The augmented matrices P_1 and Q_1 are of primary interest, but analogously to Theorem 2.1, we have shown that by using V, W, S, and R, and no other matrices, P_1 and Q_1 can be extended to square matrices P and Q with biorthogonal columns, that is, $P^H = Q^{-1}$. Also, if we take $W = V$, then $L = U^H$, $R = S$, and $P = Q$ is the matrix $Q^{(k)}$ in Theorem 2.1, so that is a special case of this theorem. But whereas in Theorem 2.1 we always have $\|S_k\|_2 \leq 1$, here we can have *any* SUT S and R, for from (7.2) and (7.3) a chosen SUT S and R give a U and L, and then, for example, taking $W = I$ and $V = L + I + U$ would result in this S and R. Thus while $0 \leq \|S_k\|_2 \leq 1$ was a beautiful measure of the loss of orthogonality in V_k, it is not clear that there is any better measure of loss of biorthogonality here than $\|L + U\|_{2,F}$.

8. Future research. This paper could be used to simplify the understanding of the analyses in [5, 6, 18], as well as that in [2]. More importantly, it has led to the realization that Lanczos's tridiagonalization of a symmetric matrix [13] is backward stable for a remarkably strange augmented system; see [17]. This might facilitate the analyses of this and many other algorithms for solving such large sparse problems as the symmetric and unsymmetric matrix eigenproblems [13, 1], symmetric positive definite linear systems [11], symmetric indefinite linear systems [19], and unsymmetric linear systems, least squares, total least squares, and scaled total least squares [20, 21, 4, 22], as well as some of the other practical methods which produce supposedly orthonormal vectors, or two sets of vectors which are supposedly biorthogonal, in the manner discussed here.

Acknowledgments. I thank Ivo Panayotov for his careful reading of this paper and his help which led, in particular, to clearer formulations of both Theorems 2.1 and 7.1, and Julien Langou for his suggestions which improved both the text of this paper and the motivation for and result of Corollary 5.5. I also thank Ron Morgan for his perceptive and very valuable remark that led to section 7 here, and two referees for their very useful comments that improved both the scholarship and readability of the paper.

REFERENCES

[1] W. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] J.L. Barlow, N. Bosner, and Z. Drmač, *A new stable bidiagonal reduction algorithm*, Linear Algebra Appl., 397 (2005), pp. 35–84.

[3] Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.
[4] Å, Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[5] Å. Björck and C.C. Paige, *Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190.
[6] Å. Björck and C.C. Paige, *Solution of augmented linear systems using orthogonal factorizations,* BIT, 34 (1994), pp. 1–24.
[7] L. Giraud, S. Gratton, and J. Langou, *A rank-k update procedure for reorthogonalizing the orthogonal factor from modified Gram–Schmidt*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1163–1177.
[8] L. Giraud and J. Langou, *When modified Gram-Schmidt generates a well-conditioned set of vectors*, IMA J. Numer. Anal., 22 (2002), pp. 521–528.
[9] G.H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224.
[10] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[11] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[12] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[13] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282.
[14] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.
[15] J. Langou, *Résolution de systèmes linéaires de grande taille avec plusieurs seconds membres (Iterative methods for solving linear systems with multiple right-hand sides)*, Ph.D. thesis, Institut Nationales des Sciences Appliquées de Toulouse, Toulouse, France, 2003.
[16] R.B. Morgan, *Personal communication*, 2008.
[17] C.C. Paige, *Augmented Backward Stability of Lanczos's Symmetric Matrix Tridiagonalization Process*, in preparation.
[18] C.C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram–Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284.
[19] C.C. Paige and M.A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[20] C.C. Paige and M.A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[21] C.C. Paige and M.A. Saunders, *ALGORITHM* 583, *LSQR: Sparse linear equations and sparse least squares problems*, ACM Trans. Math. Software, 8 (1982), pp. 195–209.
[22] C.C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., 91 (2002), pp. 117–146.
[23] Y. Saad and M.H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] C. Sheffield (June 25, 1935 – November 2, 2002), *comment to Gene Golub.*
[25] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.