行政院國家科學委員會專題研究計畫　成果報告

應用在正交矩陣上之二重移位 QR 算則收斂之探討

計畫類別：個別型計畫
計畫編號：NSC93-2115-M-004-002-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立政治大學應用數學學系

計畫主持人：王太林

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 12 月 23 日

ON CONVERGENCE OF THE DOUBLE-SHIFT QR ALGORITHM FOR REAL ORTHOGONAL HESSENBERG MATRICES*

TAI-LIN WANG†

Abstract. In applying the double-shift QR algorithm to compute the eigenvalues of real *orthogonal* Hessenberg matrices, a *unimodular* shift strategy with a distinct monotonicity property and a *cubic* asymptotic rate of convergence is devised and tested, which is more efficient than the conventional Francis shift in producing swift convergence and, in extreme circumstances, significantly accelerates the deflation process, as numerical experiments demonstrate.

Key words. QR algorithm, shift strategy, convergence, orthogonal Hessenberg matrices

AMS subject classification. 65F15, 15A18

1. Introduction. In applying the double-shift QR algorithm [7][10] to compute the eigenvalues of real *orthogonal* Hessenberg matrices, we seek to customize the conventional Francis shift strategy to make the QR algorithm perform more efficiently. Specifically, a *unimodular* shift is proposed and analyzed, with which the QR iteration has more rapid numerical convergence and still enjoys a unique monotonicity property and a cubic asymptotic rate; the distinctive advantage of using this customized shift is that, for an iterating matrix with elements under extreme conditions, the number of iterations required for deflation can be significantly reduced. Furthermore, since the *explicit* invariant form under the double-shift QR can be fully determined in this case, preventive measures are devised in advance to avoid the unusual situations of slow- and no-convergence through the inclusion of an auxiliary shift. For orthogonal matrices, improvements in practical convergence of QR are demonstrated through numerical testing.

2. Notation and Background. In this paper, A will represent a *real* (upper) Hessenberg matrix of order n, with *positive* elements $\beta_k := \mathbf{e}_{k+1}^* A \mathbf{e}_k$, $1 \le k < n$, on the subdiagonal and zero entries below it, where $\mathbf{e}_k$ is the kth column of the identity matrix, and the superscript $*$ denotes the transposition of a given vector or matrix. The Euclidean norm $||\bullet||$ is used to measure the length of a row or column vector. In case A is *orthogonal*, we use the letter U to respresent the matrix. The same structure and similar expressions are also extended to $\widehat{A}$ and $A^{(k)}$, which are defined later. A Hessenberg matrix is said to be *unreduced* if none of its subdiagonal elements is zero, and with no loss of generality in convergence analysis we may assume that the Hessenberg matrices considered in this paper have only *positive* subdiagonal elements, and hence are unreduced.

2.1. Double-shift QR iteration. Given a real Hessenberg matrix A, consider the orthogonal-triangular factorization of

$$(A - \mu_1 I)(A - \mu_2 I) \equiv \chi_2(A) =: QR, \tag{2.1}$$

where Q is orthogonal, R is upper triangular with nonnegative diagonal elements, and μ_1, μ_2 are the eigenvalues of the *trailing* principal 2-by-2 submatrix of A. Observe

*Research supported by the National Science Council of the Republic of China under grant 93-2115-M-004-002

†Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan. E-mail: wang@math.nccu.edu.tw

that χ_2 is a quadratic polynomial with *real* coefficients, though μ_1, μ_2 may appear as a complex conjugate pair. With Q we define $\widehat{A}$, the double QR transform of A, by

$$\widehat{A} := Q^* A Q.$$

It can be shown [8][13] that $\widehat{A}$ is a unique, real Hessenberg matrix with *positive* subdiagonal elements if μ_1, μ_2 are not eigenvalues of A, and it is easy to check that

$$(\widehat{A} - \mu_1 I)(\widehat{A} - \mu_2 I) \equiv \chi_2(\widehat{A}) = RQ. \tag{2.2}$$

The real Hessenberg QR algorithm (HQR) [10] iterates the transformation $A \to \widehat{A}$ (usually called a double QR step), and a sequence $\{A^{(k)}\}$ of *unreduced*, orthogonally similar Hessenberg matrices is produced if the $\chi_2^{(k)}(A^{(k)})$ are nonsingular. It was shown [4][14] that, if A is *normal*, then

$$\widehat{\beta}_{n-2}\widehat{\beta}_{n-1} \le \beta_{n-2}\beta_{n-1} \tag{2.3}$$

for each double step; hence in the QR iteration, $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)}$ form a bounded decreasing sequence for which the limit is essentially 0, if $A^{(k)}$ is not invariant under double QR; the asymptotic rate of $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)} \to 0$ is very fast (see [17] or §3.4 for an analysis); consequently, as $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)}$ becomes negligibly small, one or two (approximate) eigenvalues of A are obtained after deflating $A^{(k)}$. The procedure continues on the deflated submatrix until all the eigenvalues are found. See [10] for more details.

In this paper the HQR is employed on an *orthogonal* matrix U, and the specific structure of U is exploited to make the iterative process converge more efficiently by customizing the conventional shift polynomial χ_2 while maintaining the merits, namely, with a monotonicity property similar to (2.3) and a cubic rate of convergence. We begin with a parametric representation of U.

2.2. Parametrization of the orthogonal matrix. It is known that each *orthogonal* Hessenberg matrix $A =: U$ of order n with *positive* subdiagonal elements $\{\beta_j\}_{j=1}^{n-1}$ can be *uniquely* expressed in a parametric form [1][9]

$$U = \begin{bmatrix} -\alpha_0\alpha_1 & -\alpha_0\beta_1\alpha_2 & \cdots & -\alpha_0\beta_1\beta_2\cdots\beta_{n-2}\alpha_{n-1} & -\alpha_0\beta_1\beta_2\cdots\beta_{n-1}\alpha_n \\ \beta_1 & -\alpha_1\alpha_2 & \cdots & -\alpha_1\beta_2\cdots\beta_{n-2}\alpha_{n-1} & -\alpha_1\beta_2\cdots\beta_{n-1}\alpha_n \\ & \beta_2 & \ddots & \vdots & \vdots \\ & & \ddots & -\alpha_{n-2}\alpha_{n-1} & -\alpha_{n-2}\beta_{n-1}\alpha_n \\ & & & \beta_{n-1} & -\alpha_{n-1}\alpha_n \end{bmatrix}, \tag{2.4}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are usually called the *Schur parameters* of U, $\alpha_0 := 1$, $|\alpha_n| = 1$, and

$$|\alpha_j|^2 + \beta_j^2 = 1, \quad \alpha_j \in \mathbb{R}, \ \beta_j > 0, \quad 1 \le j < n. \tag{2.5}$$

Observe from (2.4) that these parameters can be determined successively from the top row and the subdiagonal of U [9][15]:

$$\begin{aligned} \alpha_1 &= -\mathbf{e}_1^* U \mathbf{e}_1, \\ \alpha_j &= -\mathbf{e}_1^* U \mathbf{e}_j / \beta_1\beta_2\cdots\beta_{j-1}, \qquad 2 \le j \le n. \end{aligned}$$

For unitary eigenvalue problems, this parametrization plays a fundamental role in simplifying intricate relations and deriving useful properties, because of the convenience of using n parameters to fully represent a structured matrix with more than $n^2/2$ elements; see, for example, a recent paper by Bohnhorst, Bunse-Gerstner, and Faßbender [3], and the numerous references cited therein.

To find the eigenvalues of U, there is no loss of generality in assuming that

$$n \text{ is even and } \alpha_n = 1; \tag{2.6}$$

hence the eigenvalues are all in complex conjugate pairs, which further implies that, in the double QR iteration,

$$\beta_{2j-1}^{(k)},\ j = 1, 2, 3, \ldots, n/2, \text{ are bounded away from 0.} \tag{2.7}$$

This is because existence of the real eigenvalues ($+1$ and/or -1) of U can be detected beforehand by inspecting the parity of n and the sign of α_n (for more details, see [1] or [14]), and removal of these eigenvalues from U can be carried out through QR steps with shifts ± 1.

3. Convergence of the QR Iteration. In applying the double-shift QR to an orthogonal matrix U with the form (2.4), there are circumstances in which the use of the Francis shifts μ_1, μ_2 in (2.1) may not be appropriate; the shifts could be fairly small or even zero in extreme cases, since they are the eigenvalues of a 2-by-2 submatrix of U and hence with magnitude always less than unity (because $\beta_{n-2} > 0$) [16]. If $\mu_1 = \mu_2 = 0$ (corresponding to $\beta_{n-2} = \beta_{n-1} = 1$), then $\widehat{U} = U$ and the sequence $U^{(k)}$ is invariant under double QR because $\chi_2(U)$ is orthogonal. If μ_1, μ_2 are very small, then $\chi_2(U)$ is still close to being orthogonal and $\widehat{U}$ would not be quite different from U, resulting in very slow decreases of $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)}$ in the early stages of the QR iteration; hence it may take numerous steps before $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)}$ ultimately gets into the asymptotic regime, where the convergence is cubic and swift. (See the data presented in Experiment 2 of Section 4.) Besides, μ_1, μ_2 could be real numbers with different magnitude, which apparently are not close to any conjugate pair of eigenvalues of U.

3.1. Construction of a unimodular-shift strategy. Based upon the fact that the eigenvalues of U are *all* situated on the unit circle and by (2.6) *all* appear in complex conjugate pairs, we should choose instead a *unimodular* conjugate pair $\{\mu, \overline{\mu}\}$ as shifts, and it is only natural (by anticipating that $\beta_{n-2} \approx 0$ and $\alpha_{n-2} \approx 1$ in (2.4)) to take the eigenvalues of the 2-by-2 *orthogonal* matrix

$$\begin{bmatrix} -\alpha_{n-1} & -\beta_{n-1} \\ \beta_{n-1} & -\alpha_{n-1} \end{bmatrix},$$

which are $-\alpha_{n-1} \pm i\beta_{n-1}$, as μ and $\overline{\mu}$. This matrix has the characteristic polynomial

$$\phi_2(\lambda) = \lambda^2 + 2\alpha_{n-1}\lambda + 1 \tag{3.1}$$

which depends only on α_{n-1} and is used in place of χ_2 in (2.1) when $A =: U$ is *orthogonal*. For brevity, we shall call ϕ_2 the unimodular shift (polynomial).

3.2. Monotonicity property of ϕ_2. Under the assumption that A is *normal*, a basic property of the QR transformation $A \to \widehat{A} : p(A) = QR$, $\widehat{A} = Q^*AQ$ with *any* shift polynomial p is

$$||\mathbf{e}_n^* p(\widehat{A})|| \leq ||\mathbf{e}_n^* p(A)||, \tag{3.2}$$

which states that the last row of $p(A)$ is decreasing after each QR step. To see this, observe that

$$\begin{aligned} ||\mathbf{e}_n^* p(\widehat{A})|| &= ||\mathbf{e}_n^* RQ|| = ||\mathbf{e}_n^* R|| \\ &\leq ||R\mathbf{e}_n|| = ||QR\mathbf{e}_n|| = ||p(A)\mathbf{e}_n|| \\ &= ||\mathbf{e}_n^* p(A)||, \end{aligned} \tag{3.3}$$

in which properties of the unitary invariance of $||\bullet||$, triangular structure of R, and normality of $p(A)$ have been used.

Applying the parametric form (2.4) of U and relation (2.5) to ϕ_2 defined by (3.1), we obtain

$$\begin{aligned} \mathbf{e}_n^* \phi_2(U) = \mathbf{e}_{n-2}^* \beta_{n-2}\beta_{n-1} + \mathbf{e}_{n-1}^* (1-\alpha_{n-2})\alpha_{n-1}\beta_{n-1} \\ + \mathbf{e}_n^* (1-\alpha_{n-2})\beta_{n-1}^2 \end{aligned} \tag{3.4}$$

and

$$||\mathbf{e}_n^* \phi_2(U)|| = \sqrt{2(1-\alpha_{n-2})}\beta_{n-1}. \tag{3.5}$$

For each double QR step $U \to \widehat{U}$, it is easy to check that, if $\widehat{\alpha}_{n-1} \neq \alpha_{n-1}$, then

$$||\mathbf{e}_n^* \widehat{\phi}_2(\widehat{U})|| < ||\mathbf{e}_n^* \phi_2(\widehat{U})||, \tag{3.6}$$

where $\widehat{\phi}_2(\lambda) = \lambda^2 + 2\widehat{\alpha}_{n-1}\lambda + 1$. Summing up these calculations, we have

$$\begin{array}{lll} & \sqrt{2(1-\widehat{\alpha}_{n-2})}\,\widehat{\beta}_{n-1} & \\ = & ||\mathbf{e}_n^* \widehat{\phi}_2(\widehat{U})||, & \text{from (3.5)}, \\ \leq & ||\mathbf{e}_n^* \phi_2(\widehat{U})||, & \text{by (3.6)}, \\ \leq & ||\mathbf{e}_n^* \phi_2(U)||, & \text{by (3.2) with } p = \phi_2 \text{ and } A = U, \\ = & \sqrt{2(1-\alpha_{n-2})}\beta_{n-1}, & \text{from (3.5)}, \end{array} \tag{3.7}$$

for each double step, which implies that $\sqrt{2(1-\alpha_{n-2}^{(k)})}\beta_{n-1}^{(k)}$ is monotonically decreasing in the QR iteration. If the limit of this sequence is zero, then clearly

$$\beta_{n-2}^{(k)} = \sqrt{1-\alpha_{n-2}^{(k)2}} \to 0$$

is implied, because $\beta_{n-1}^{(k)}$is bounded below from 0 by (2.7), and a pair of complex conjugate eigenvalues is obtained after deflation. This is the ordinary case and it happens almost all the time. In the following subsection we consider the (unstable) situation in which the limit of the declining sequence is not zero.

3.3. Asymptotic analysis in the extreme case. Now suppose

$$\sqrt{2(1-\alpha_{n-2}^{(k)})}\beta_{n-1}^{(k)} \to \beta > 0 \tag{3.8}$$

as $k \to \infty$; then from the above analysis outlined in (3.7) and (3.3) we know

$$\beta \le ||\mathbf{e}_n^* R^{(k)}|| \le ||R^{(k)}\mathbf{e}_n|| \to \beta$$

and hence

$$R^{(k)}\mathbf{e}_n \to \beta \mathbf{e}_n,$$

which implies, through the QR factorization of

$$\phi_2^{(k)}(U^{(k)}) =: Q^{(k)}R^{(k)},$$

that *asymptotically*, the last column of $\phi_2^{(k)}(U^{(k)})$ is orthogonal to *all* its preceding $n-1$ columns and has a length of β, or equivalently,

$$\phi_2^{(k)}(U^{(k)})^*\phi_2^{(k)}(U^{(k)})\mathbf{e}_n \to \beta^2\mathbf{e}_n. \tag{3.9}$$

This indicates that β and $\mathbf{e}_n$ are approached by a singular value and vector of $\phi_2^{(k)}(U^{(k)})$. We seek to extract from (3.9) all the basic asymptotic relations among the entries of $U^{(k)}$ in terms of the parameters $\alpha_j^{(k)}$. The inner product of $\mathbf{e}_j$, $j < n$, with (3.9) gives $\mathbf{e}_j^*\phi_2^{(k)}(U^{(k)})^*\phi_2^{(k)}(U^{(k)})\mathbf{e}_n \to 0$ which, after switching the order of $\phi_2^{(k)}(U^{(k)})^*$ and $\phi_2^{(k)}(U^{(k)})$, is equivalent to

$$\mathbf{e}_j^*\phi_2^{(k)}(U^{(k)})\phi_2^{(k)}(U^{(k)})^*\mathbf{e}_n \to 0 \quad \text{for } j < n. \tag{3.10}$$

The trick used here is that, due to the (upper) Hessenberg structure of $U^{(k)}$, the inner product of $\mathbf{e}_j^*\phi_2^{(k)}(U^{(k)})$ and $\mathbf{e}_n^*\phi_2^{(k)}(U^{(k)})$ in (3.10) is much simpler to compute, when j is close to n. After somewhat lengthy but elementary calculations using (2.4), (3.1) to compute $\mathbf{e}_j^*\phi_2^{(k)}(U^{(k)})$ and applying (3.4) to $\mathbf{e}_n^*\phi_2^{(k)}(U^{(k)})$, we reach the following conclusions from (3.10):

- for $j = n-1$, after rearrangements and factorizations,

$$\alpha_{n-3}^{(k)}\left(\frac{1+\alpha_{n-2}^{(k)}}{3-\alpha_{n-2}^{(k)}}\right) - \alpha_{n-1}^{(k)} \to 0 \tag{3.11}$$

  because $(1-\alpha_{n-2}^{(k)})\beta_{n-1}^{(k)}$ is bounded away from 0, by (3.8);
- for $j = n-2$, after collecting terms and simplifying with (3.11),

$$\alpha_{n-4}^{(k)} \to 1 \ \text{ and hence } \ \beta_{n-4}^{(k)} \to 0, \tag{3.12}$$

  because $\beta_{n-3}^{(k)2}\beta_{n-2}^{(k)}\beta_{n-1}^{(k)}$ is bounded away from 0, by (2.7) and (3.8).

From the asymptotic condition (3.12), it follows that a 4-by-4 submatrix is ultimately decoupled from $U^{(k)}$ in the lower right corner with

$$E_4^* U^{(k)} E_4 \to U_4 \quad \text{as } k \to \infty, \tag{3.13}$$

where

$$E_4 := [\mathbf{e}_{n-3}, \mathbf{e}_{n-2}, \mathbf{e}_{n-1}, \mathbf{e}_n],$$

and U_4 is a 4-by-4 orthogonal matrix with Schur parameters $\{a_j\}_{j=1}^4$ satisfying the conditions

$$a_1\left(\frac{1+a_2}{3-a_2}\right) = a_3, \tag{3.14}$$

with $0 \le |a_j| < 1$ for $1 \le j \le 3$, and $a_4 = 1$, because of (3.11) and (2.6). And there is no need to pursue further information from (3.10) for $j < n-2$ because of the decoupling.

Some comments are given on the asymptotic behavior of (3.13) and the limit form U_4 defined by (3.14):

1. Through the derivation of (3.11) and (3.12) from (3.8), it is obvious that the *invariance* of $\sqrt{2(1-\alpha_{n-2})}\beta_{n-1}$ under double QR implies *that* of

$$\alpha_{n-3}\left(\frac{1+\alpha_{n-2}}{3-\alpha_{n-2}}\right) = \alpha_{n-1} \quad \text{and} \quad \alpha_{n-4} = 1 \ (\beta_{n-4} = 0);$$

hence *no* unreduced U with $n > 4$ can be invariant because $\beta_{n-4} > 0$. If $n = 4$, then clearly U_4 is invariant, since $\phi_2(U_4)$ gives a multiple β of an orthogonal matrix with $0 < \beta = \sqrt{2(1-a_2)(1-a_3^2)} < 2$, and this is the *only* invariant form with ϕ_2. Also note that given the real numbers α $(0 \le |\alpha| < 1)$, β $(0 < \beta < 2)$, and the quadratic polynomial

$$\phi(z) = z^2 + 2\alpha z + 1,$$

there exist exactly 4 distinct numbers $\{z_j\}_{j=1}^4$, on the unit circle and in two conjugate pairs, such that $|\phi(z_j)| = \beta$ for $1 \le j \le 4$; more precisely, if x_1 and x_2 are the real parts of these two conjugate pairs, then

$$\begin{aligned} -\tfrac{1}{2}(x_1 + x_2) &= \alpha \\ 2|x_j + \alpha| &= \beta, \quad j = 1, 2. \end{aligned}$$

2. The highly unusual situation of (3.13) might exist only if *four* of the eigenvalues λ_j of U (in two conjugate pairs) satisfy the asymptotic condition

$$|\phi_2^{(k)}(\lambda_j)| \to \beta,$$

a direct consequence of (3.9), which is equivalent to

$$||\mathbf{e}_n^* \phi_2^{(k)}(U^{(k)})|| \to \beta. \tag{3.15}$$

This extreme case is *unstable* [4][2], in the sense that any small perturbations (e.g., rounding errors in numerical computation) in the limit space of (3.15) will translate into perturbations of $\alpha_{n-1}^{(k)}$ in the shift $\phi_2^{(k)}$ which, due to the strict property of (3.6), may well cause $\sqrt{2(1-\alpha_{n-2}^{(k)})}\beta_{n-1}^{(k)}$ to drive past the value β and eventually down to 0. So we would not observe this kind of decoupling of $U^{(k)}$ to occur in practical problems.

3. However, for an iterate $U^{(k)}$ with

$$E_4^* U^{(k)} E_4 \approx U_4, \tag{3.16}$$

it usually takes numerous steps in the QR iteration for

$$\sqrt{2(1-\alpha_{n-2}^{(k)})}\beta_{n-1}^{(k)} \searrow 0$$

before the asymptotic regime is established, though convergence is ultimately *cubic* (which will be shown in the next subsection). Condition (3.16) therefore provides the very specific circumstances under which ϕ_2 does not function effectively; fortunately, the situation of (3.16) is predictable and measurable through (3.11) and (3.12) before each QR step $U^{(k)} \to U^{(k+1)}$ is carried out. To hasten convergence, an *auxiliary* shift can be used to break up the closeness in (3.16) and give the iteration a new starting U. See the formula devised later in (3.32). This special treatment is implemented and tested in Section 4.

3.4. Rate of convergence. In this subsection we first present, for *normal* matrices, a simple and illustrative analysis on the asymptotic rate of $\beta_{n-2}^{(k)}\beta_{n-1}^{(k)} \to 0$ in the double-step QR iteration with shift χ_2, using an elementary technique applied to the single-step QR by Wilkinson [18]. We then demonstrate that, with the *unimodular* shift ϕ_2 described by (3.1), $\beta_{n-2}^{(k)} \to 0$ in the QR iteration $U^{(k)}$ still enjoys a *cubic* asymptotic rate, the same as that with the Francis shift χ_2. Note that a general treatment for a generalized problem (including the convergence of multishift QR and LR) has been given by Watkins and Elsner [17] using subspace iteration techniques. Let

$$A =: \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{3.17}$$

where $A_{22} \in \mathbb{R}^{2\times 2}$, and let

$$(A-\mu_1 I)(A-\mu_2 I) =: \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ O & R_{22} \end{bmatrix} := QR, \tag{3.18}$$

where the partitionings are conformal with that of A. Premultiplying this equation by Q^*, we have

$$\begin{bmatrix} Q_{11}^* & Q_{21}^* \\ Q_{12}^* & Q_{22}^* \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ O & R_{22} \end{bmatrix}$$

and, by equating respectively the $(2,1)$ and $(2,2)$ block entries on each side,

$$Q_{12}^* B_{11} + Q_{22}^* B_{21} = O \tag{3.19}$$

$$Q_{12}^* B_{12} + Q_{22}^* B_{22} = R_{22} \tag{3.20}$$

where, from (3.17) and (3.18),

$$B_{11} = (A_{11} - \mu_1 I)(A_{11} - \mu_2 I) + A_{12}A_{21} \in \mathbb{R}^{(n-2)\times(n-2)} \tag{3.21}$$

$$B_{22} = (A_{22} - \mu_1 I)(A_{22} - \mu_2 I) + A_{21}A_{12} \in \mathbb{R}^{2\times 2}. \tag{3.22}$$

In the following it is illustrated that, if A is *normal* Hessenberg, μ_1, μ_2 are the eigenvalues of A_{22}, and $\beta_{n-2} = \varepsilon$ for some arbitrarily small number $\varepsilon > 0$, then $\widehat{\beta}_{n-2}\widehat{\beta}_{n-1} = O(\varepsilon^3)$ after one double-QR step. We use the Frobenius norm $||C||_{\mathsf{F}} := \sqrt{\sum_{i,j} |c_{ij}|^2}$ to measure the magnitude of a matrix $C = [c_{ij}]$. Since A is normal and (unreduced) Hessenberg, we know

$$||A_{12}||_{\mathsf{F}} = ||A_{21}||_{\mathsf{F}} = \beta_{n-2} = \varepsilon \tag{3.23}$$

and, in (3.22) with $\chi_2(A_{22}) = (A_{22} - \mu_1 I)(A_{22} - \mu_2 I) = O$ (by the Cayley-Hamilton theorem),

$$||B_{22}||_{\mathsf{F}} \leq \varepsilon^2. \tag{3.24}$$

From the partitions defined by (3.17) and (3.18) we have

$$B_{21} = A_{21}(A_{11} - \mu_2 I) + (A_{22} - \mu_1 I)A_{21},$$

and hence we know

$$||B_{12}||_{\mathsf{F}} = ||B_{21}||_{\mathsf{F}} = O(\varepsilon) \tag{3.25}$$

because of (3.23) and the fact that $||A||_{\mathsf{F}}$ is always bounded. Furthermore, since all the eigenvalues of A are simple, we see from (3.21) that, if ε is sufficiently small and (3.23), (3.24) hold, then B_{11} is "bounded away" from being singular (see Lemma 6.4 and Theorem 6.5 of [17]) and

$$||B_{11}^{-1}||_{\mathsf{F}} \leq \kappa \tag{3.26}$$

for some finite number κ. Next, let us examine the partitioned block of Q. From (3.19) we have

$$Q_{12}^* = -Q_{22}^* B_{21} B_{11}^{-1}$$

and therefore

$$||Q_{12}||_{\mathsf{F}} = ||Q_{21}||_{\mathsf{F}} = O(\varepsilon) \tag{3.27}$$

because of (3.25), (3.26) and the fact that Q is orthogonal and

$$||Q_{22}||_{\mathsf{F}} \leq ||Q||_{\mathsf{F}} = \sqrt{n}. \tag{3.28}$$

Hence from (3.20) we infer that

$$||R_{22}||_{\mathsf{F}} = O(\varepsilon^2) \tag{3.29}$$

by (3.24), (3.25), (3.27), and (3.28). Now let the partition of

$$\widehat{A} =: \begin{bmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{bmatrix}$$

be conformal with that of A and equate the $(2,1)$ block entry on each side of (2.2):

$$\widehat{A}_{21}(\widehat{A}_{11} - \mu_2 I) + (\widehat{A}_{22} - \mu_1 I)\widehat{A}_{21} = R_{22}Q_{21} \in \mathbb{R}^{2\times(n-2)}.$$

Further equating the $(2, n-2)$ entry on each side of this matrix equation (observing that $\widehat{A}_{21} = \widehat{\beta}_{n-2}\mathbf{e}_2\mathbf{e}_{n-2}^*$ and that $\mathbf{e}_2^*\widehat{A}_{22}\mathbf{e}_1 = \widehat{\beta}_{n-1}$) gives

$$\widehat{\beta}_{n-2}\widehat{\beta}_{n-1} = \mathbf{e}_2^* R_{22}Q_{21}\mathbf{e}_{n-2} \le ||R_{22}Q_{21}||_{\mathsf{F}} = O(\varepsilon^3),$$

because of (3.27) and (3.29). If $\beta_{n-1}^{(k)}$ are bounded away from zero (e.g., the eigenvalues sought are in complex conjugate pairs, as they are in the orthogonal case), then

$$||\widehat{A}_{21}||_{\mathsf{F}} = \widehat{\beta}_{n-2} = O(\varepsilon^3)$$

and this demonstrates that convergence of the double-shift QR is of cubic order. In reaching this conclusion, note that condition (3.29) is crucial, which in turn (cf. (3.20)) hinges on condition (3.24) that $||B_{22}||_{\mathsf{F}}$ is of order ε^2, which is derived from (3.22), where we take advantage of the double shift χ_2 (i.e., $\chi_2(A_{22}) = O$) and normality of A (cf. (3.23)). Note also that in the *orthogonal* case, cubic convergence of the QR process is still preserved if instead the *unimodular* shift ϕ_2 is used. This is because in (3.22) (with the A's replaced by the U's)

$$\begin{aligned} B_{22} &= \phi_2(U_{22}) + U_{21}U_{12} \\ &= \phi_2(U_{22}) - \chi_2(U_{22}) + U_{21}U_{12} \\ &= (1-\alpha_{n-2})(\alpha_{n-1}U_{22} + I_2) + U_{21}U_{12}, \end{aligned}$$

and we have

$$||B_{22}||_{\mathsf{F}} \le 2\sqrt{2}(1-\alpha_{n-2}) + ||U_{21}||_{\mathsf{F}}^2 \approx (\sqrt{2}+1)\varepsilon^2,$$

since $|\alpha_{n-1}| \le 1$, $||U_{22}||_{\mathsf{F}} \le ||I_2||_{\mathsf{F}} = \sqrt{2}$, $||U_{21}||_{\mathsf{F}} = \beta_{n-2} = \varepsilon$, and

$$1 - \alpha_{n-2} = 1 - \sqrt{1-\varepsilon^2} \approx \frac{1}{2}\varepsilon^2$$

for sufficiently small ε[1]. Therefore, $||B_{22}||_{\mathsf{F}}$ is still of order ε^2 as before, and hence cubic convergence of the iteration.

3.5. Sufficient conditions for convergence. We summarize our analyses in the previous subsections and conclude with a theorem on the convergence of double-step QR for orthogonal matrices; a more efficient shift strategy based on the assumptions of this theorem for such specifically structured matrices is proposed.

THEOREM 3.1. *Let the double-step QR algorithm with the unimodular shift polynomial ϕ_2 be applied to an unreduced orthogonal Hessenberg matrix U of order $n \ge 4$ expressed in the Schur parametric form* (2.4) *under the assumption* (2.6). *If there is a number $\delta > 0$ such that either*

$$\left| \alpha_{n-3}^{(j)} \left(\frac{1+\alpha_{n-2}^{(j)}}{3-\alpha_{n-2}^{(j)}} \right) - \alpha_{n-1}^{(j)} \right| \ge \delta \tag{3.30}$$

or

$$\beta_{n-4}^{(j)} \ge \delta \text{ for } n > 4 \tag{3.31}$$

[1]Note that eventually $\alpha_{n-2} = |\alpha_{n-2}| = \sqrt{1-\beta_{n-2}^2} \to 1$ if $\beta_{n-2} = \varepsilon \to 0$, since α_{n-2} is becoming the *last* Schur parameter of the deflated U of order $n-2$, which is $+1$ by (2.6).

holds for some subsequence with index j in the QR iteration $U^{(k)}$, then

$$\beta_{n-2}^{(k)} \to 0 \quad as \quad k \to \infty$$

and two eigenvalues of U in a complex conjugate pair are being approximated; the rate of convergence is cubic.

Proof. If $\beta_{n-2}^{(k)} \nrightarrow 0$ as $k \to \infty$, then from the previous analysis, (3.8) must be true, which implies both (3.11) and (3.12), a contradiction to either (3.30) or (3.31), whichever is supposed to hold in the QR iteration. That the rate of convergence is cubic has been proved in the last subsection.□

Note that in the forecast of slow- or no-convergence (cf. (3.16) and (3.13)), condition (3.11) is of primary concern to us and more informative than (3.12), as *all* the subdiagonal elements of $U^{(k)}$ are inherently becoming smaller in the iterative process. In practice we may choose a "suitable" value for δ and examine, *before* each QR step, if condition (3.30) (with $j := k$) in the above theorem is satisfied; and the *auxiliary* shift

$$\widetilde{\phi}_2(\lambda) = \lambda^2 + 2\lambda + 1 \tag{3.32}$$

is employed if it is not, to give the iteration a different starting U. (Another choice is $\widetilde{\phi}_2(\lambda) = \lambda^2 - 2\lambda + 1$.) The aim of using such an ad hoc shift is mainly to accelerate numerical convergence, when condition (3.16) is close enough and ϕ_2 may not operate efficiently.

It is clear that this ad hoc shift $\widetilde{\phi}_2$ can be viewed as the limit case of ϕ_2 with $\alpha_{n-1} \to 1$. Note also that $\widetilde{\phi}_2$ cannot be ineffective under condition (3.16) without the parameters $\alpha_{n-1}^{(k)}$, $\alpha_{n-2}^{(k)}$, and $\alpha_{n-3}^{(k)}$ *all* being very close to unity (cf. (3.11)), in other words, the subdiagonal elements $\beta_{n-1}^{(k)}$, $\beta_{n-2}^{(k)}$, and $\beta_{n-3}^{(k)}$ of $U^{(k)}$ all becoming vanishingly small in addition to $\beta_{n-4}^{(k)} \approx 0$ (cf. (3.12)), an extreme situation in which two conjugate pairs of eigenvalues are clustered near the point -1, and in this case, $\widetilde{\phi}_2$ (with double zeros at -1) should indeed be considered as a most desired shift polynomial to deal with these clustered eigenvalues.

4. Numerical Experiments. We demonstrate in this section that, for *orthogonal* matrices, shift ϕ_2 provides more efficient numerical convergence than shift χ_2 does, especially before the establishment of the asymptotic regime, though both were shown in §3.5 to have cubic asymptotic rates; the latter is incorporated with double QR in software packages. The testing was focused on determining the numbers of iterations required for $\beta_{j-2}^{(k)}\beta_{j-1}^{(k)}$, $j = n, n-1, \ldots, 3$, to become negligible [10]. The computations were done in double-precision Fortran on an IBM compatible PC-80486 with unit roundoff $\varepsilon \approx 10^{-19}$.

The following contractions are used to describe how the shift scheme in each case

is devised or modified in carrying out the double QR iteration:

HQR	the subroutine in EISPACK [12] with shift χ_2 and an exceptional shift [10][2]introduced at steps $k = 11$ and $k = 21$ to combat the extreme situations;
HQR0	the HQR with shift χ_2 but *skipping* the exceptional shift;
HQRu	the HQR with shift ϕ_2 instead and a check on condition (3.30) in Theorem 3.1 with $j := k$ and $\delta = 10^{-12}$ before using the auxiliary shift $\widetilde{\phi}_2$.

We also use the data from HQR0 to accentuate, for orthogonal matrices in particular, the significance of the incorporated exceptional shift.

Given an orthogonal Hessenberg matrix U of size n (with $\alpha_n = 1$ and n even), let

$$\textbf{itmax} = \left\{ \begin{array}{l} \text{the } \textit{maximum} \text{ number of double-step QR iterations} \\ \text{required to get one conjugate pair of eigenvalues} \end{array} \right\}.$$

To test the effectiveness of the various shift strategies, we measure and list in the following tables the numerical averages of **itmax** over $10,000$ sample matrices of size n, each of which is constructed by a set of "randomly selected" real Schur parameters $\{\alpha_j\}_{j=1}^{n}$ with certain prescribed constraints as stated in each experiment.

Experiment 1. Randomly selected $\{\alpha_j\}_{j=1}^{n-1}$ with $\alpha_n = 1$.

itmax	HQR0	HQR	HQRu
$n = 4$	5.06	5.02	4.11
$n = 10$	5.88	5.77	5.16
$n = 20$	6.48	6.30	5.81
$n = 30$	6.70	6.61	6.18

Observe the slight differences in numbers between the columns HQR0 and HQR, which indicate that, even for randomly selected orthogonal matrices, the exceptional shift does have a role to play; in other words, there is still a small percentage (around 1-3%) of matrices for which the itmax exceeds at least 10 iterations. On average, HQRu is only about 7-18% more efficient (in terms of itmax) than HQR for matrices of size ≤ 30, as the data listed in Experiment 1 show. However, under "specific" circumstances, the differences among the three are very substantial, and HQRu works significantly better than HQR, as Experiments 2 and 3 demonstrate.

Experiment 2. Randomly selected $\{\alpha_j\}_{j=1}^{n-1}$ with $|\alpha_{n-2}| \leq 10^{-7}$, $|\alpha_{n-1}| \leq 10^{-7}$, and $\alpha_n = 1$.

itmax	HQR0	HQR	HQRu
$n = 4$	23.6	15.4	5.44
$n = 10$	22.6	16.1	5.67
$n = 20$	22.8	16.3	6.10
$n = 30$	22.9	16.4	6.34

[2]Later versions of both LAPACK and Matlab have made further modifications over the exceptional shift in the double-step QR [6][5][11].

Sluggish convergence shown in the columns HQR0 and HQR of this table was predicted and explained in the opening paragraph of Section 3.

Moreover, the conventional shift χ_2 also becomes ineffective for those U's with parameters under the conditions

$$|\alpha_{n-3}\alpha_{n-2} - \alpha_{n-1}| \approx \varepsilon \quad \text{and} \quad |\alpha_{n-4} - 1| \approx \varepsilon,$$

where ε is small [14]; orthogonal matrices of this special type that cause HQR to fail are not hard to come by, even with the exceptional shift, as the next table indicates. Those "extreme matrices" for which convergence (to a pair of eigenvalues) does not occur within the iteration limit $30n$ set in HQR are not counted in the itmax averages. Instead, the *number* of such matrices (out of $10,000$ for each n) is placed in parentheses after the (itmax) average.

Experiment 3. Randomly selected $\{\alpha_j\}_{j=1}^{n-5} \cup \{\alpha_{n-3}, \alpha_{n-2}\}$ with $\alpha_{n-4} = \sqrt{1-(10^{-7})^2}$ (for $n > 4$), $\alpha_{n-1} = \alpha_{n-3}\alpha_{n-2}$, and $\alpha_n = 1$.

itmax	HQR0	HQR	HQRu
$n = 4$	49.0 (1349)	16.0 (16)	6.18
$n = 10$	40.5 (358)	16.0 (8)	6.30
$n = 20$	32.6 (69)	15.6 (2)	6.66
$n = 30$	30.4 (35)	15.4 (1)	6.93

The unimodular shift ϕ_2 does have its Achilles' heel, as described by (3.16). But HQRu takes the precautionary measure by checking this extreme condition before each QR sweep; if (3.16) is close enough, for example,

$$\left| \alpha_{n-3}^{(k)} \left(\frac{1+\alpha_{n-2}^{(k)}}{3-\alpha_{n-2}^{(k)}} \right) - \alpha_{n-1}^{(k)} \right| < \delta = 10^{-12},$$

then the auxiliary shift $\widetilde{\phi}_2$ is applied to break up the unusual configuration. Guarded with this measure, HQRu is again doing better than HQR, as the data listed in the following table indicate.

Experiment 4. Randomly selected $\{\alpha_j\}_{j=1}^{n-5} \cup \{\alpha_{n-3}, \alpha_{n-2}\}$ with $\alpha_{n-4} = \sqrt{1-(10^{-7})^2}$ (for $n > 4$), $\alpha_{n-1} = \alpha_{n-3}(1+\alpha_{n-2})/(3-\alpha_{n-2})$, and $\alpha_n = 1$.

itmax	HQR0	HQR	HQRu
$n = 4$	7.76	7.76	4.72
$n = 10$	7.82	7.80	4.98
$n = 20$	7.89	7.93	5.62
$n = 30$	7.99	8.04	6.01

It is interesting to observe that, under the specific conditions set in this experiment, HQR0 in general works slightly better than HQR as n becomes larger.

5. Concluding Remarks. For orthogonal matrices, it appears just natural to use *unimodular* shifts in the QR iteration to approximate the eigenvalues. In theory we have derived the distinctive convergence properties that it possesses, and in practice

we have demonstrated its superiority in numerical convergence through both general and specific examples.

Though rarely encountered in practical problems, extreme examples of orthogonal matrices of order n up to 30, for which the subroutine HQR in EISPACK fails to reveal the eigenvalues (within the iteration limit $30n$), can easily be found, as in Experiment 3 of Section 4, if a sufficient number of sample matrices, say $10,000$, is taken.

Through a better understanding of the convergence properties with the orthogonal matrices, a more efficient shift strategy can be specifically devised in the QR iteration for such structured matrices; still, more investigations and experiments are needed for further improvements.

REFERENCES

[1] G. S. Ammar, W. B. Gragg, and L. Reichel, *On the eigenproblem for orthogonal matrices*, in Proceedings of the 25th IEEE conference on Decision and Control, Athens, Greece, 1986, pp. 1963–1966.

[2] *S. Batterson*, *Convergence of the Francis shifted QR algorithm on normal matrices*, Linear Algebra Appl., 207 (1994), pp. 181–195.

[3] B. Bohnhorst, A. Bunse-Gerstner, and H. Fassbender, *On the perturbation theory for unitary eigenvalue problems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 809–824.

[4] H. J. Buurema, *A geometric proof of convergence for the QR method*, doctoral dissertation, Rijksuniversiteit Te Groningen, Groningen, The Netherlands, 1970.

[5] D. Day, *How the QR algorithm fails to converge and how to fix it*, Technical Report 96-0913J, Sandia National Laboratory, Albuquerque, NM, April 1996.

[6] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[7] J. G. F. Francis, *The QR transformation: a unitary analogue to the LR transformation, parts I and II*, Comput. J., 4 (1961–1962), pp. 265–271, 332–345.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, third edition, Johns Hopkins Univ. Press, Baltimore, 1996.

[9] W. B. Gragg, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[10] R. S. Martin, G. Peters, and J. H. Wilkinson, *The QR algorithm for real Hessenberg matrices*, Numer. Math., 14 (1970), pp. 219–231.

[11] C. Moler, *The QR algorithm*, Cleve's Corner, MATLAB News & Notes, The MathWorks Inc., Natick, MA, Summer 1995.

[12] B. T. Smith, J. M. Boyle, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines–EISPACK Guide*, Springer-Verlag, Berlin, 1974.

[13] G. W. Stewart, *Matrix Algorithms Volume II: Eigensystems*, SIAM, Philadelphia, 2001.

[14] T.-L. Wang, *Invariant normal matrices under the double-shift QR iteration*, preprint submitted to Linear Algebra Appl. for publication, 2003.

[15] T.-L. Wang and W. B. Gragg, *Convergence of the shifted QR algorithm for unitary Hessenberg matrices*, Math. Comp., 71 (2002), pp. 1473–1496.

[16] T.-L. Wang and W. B. Gragg, *Convergence of the unitary QR algorithm with a unimodular Wilkinson shift*, Math. Comp., 72 (2003), pp. 375–385

[17] D.S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.