SIAM J. Matrix Anal. Appl.
Vol. 39, No. 4, pp. 1591–1615

THE QR STEPS WITH PERFECT SHIFTS*

NICOLA MASTRONARDI[†] AND PAUL VAN DOOREN[‡]

Abstract. In this paper we revisit the problem of performing a QR step on an unreduced Hessenberg matrix H when we know an "exact" eigenvalue λ_0 of H. In exact arithmetic, this eigenvalue will appear on the diagonal of the transformed Hessenberg matrix $\tilde{H}$ and will be decoupled from the remaining part of the Hessenberg matrix, thus resulting in a *deflation*. But it is well known that in finite-precision arithmetic the so-called perfect shift can get *blurred* and that the eigenvalue λ_0 can then not be deflated and/or is perturbed significantly. In this paper, we develop a new strategy for computing such a QR step so that the deflation is almost always successful. We also show how to extend this technique to double QR steps with complex conjugate shifts.

Key words. Hessenberg form, QR step, eigenvalue problem, perfect shift

AMS subject classifications. 65F15, 65F25

DOI. 10.1137/17M1139321

1. Introduction.

Computing the eigenvalues of a matrix A is a widely studied problem in numerical linear algebra. Eigenvalues play an important role in the solution of explicit differential equations, which can be modeled as

$$\lambda \mathbf{x}(t) = A\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \quad A \in \mathbb{R}^{n\times n}, \tag{1.1}$$

where λ stands for the differential operator. The solutions of (1.1) depend heavily on the Jordan structure of A at each of its eigenvalues (see [11], [4] for more details).

There are two basic steps in the standard computation of the eigenvalues of a general matrix A. The first step is to reduce A to a Hessenberg matrix $H = UAU^T$ using an orthogonal similarity transformation U. This is a well-understood process, and its complexity is $\mathcal{O}(n^3)$ floating point operations. The second step is to perform a series of QR steps with so-called shifts that are computed during this process and "converge" to very good approximations of the eigenvalues of A. If such a shift is an exact eigenvalue λ_0 of A, then, in exact arithmetic, this eigenvalue will appear on the diagonal of the transformed Hessenberg matrix $\tilde{H}$ and will be decoupled from the remaining part of the Hessenberg matrix, thus resulting in a *deflation*. But it is well known that in finite-precision arithmetic the so-called perfect shift can get *blurred* and that the eigenvalue λ_0 can then not be deflated. In this paper, we develop a new strategy for computing such a QR step so that the deflation is almost always successful. The method is based on the preliminary computation of the corresponding eigenvector $\mathbf{x}$ such that the residual $(H - \lambda_0 I)\mathbf{x}$ is sufficiently small. The eigenvector

*Received by the editors July 18, 2017; accepted for publication (in revised form) by Z. Drmac August 21, 2018; published electronically October 23, 2018.

http://www.siam.org/journals/simax/39-4/M113932.html

Funding: The work of the first author was partially supported by GNCS–INdAM Project "Tecniche innovative per problemi di algebra lineare." The work of the second author was partially supported by the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office, and by CNR under the Short Term Mobility Program. The scientific responsibility rests with its authors.

[†] Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Bari I-70126, Italy (n.mastronardi@ba.iac.cnr.it).

[‡] Department of Mathematical Engineering, Université Catholique de Louvain, Louvain–la–Neuve B-1348, Belgium (paul.vandooren@uclouvain.be).

is then transformed to a unit vector $\mathbf{e}_1$ by a sequence of Givens transformations, which are also performed on the Hessenberg matrix. Notice that what we just described is a "backward" QR step in which we transform the eigenvalue λ_0 to the (1,1) position in $\tilde{H}$, whereas the standard (forward) QR step moves it to the (n, n) position. These two are in a sense dual to each other, but we chose to describe the backward variant here because it is more closely related to the calculation of the staircase form [19], [12], which was the inspiration for this new method.

The calculation of the Jordan form of a given eigenvalue λ_0, described in [6], [12], is perhaps the most direct application of the new perfect shift strategy developed in this paper. In order to find the complete Jordan structure of a given eigenvalue, one needs to perform a sequence of deflations with the exact eigenvalue λ_0. That eigenvalue is supposed to be known or computed to sufficiently high accuracy. This is needed in the calculation of the Drazin inverse of a matrix A [1], but it can also be applied to the (regular) generalized eigenvalue problem $\lambda B - A$ to deflate the so-called infinite eigenvalues. The length of the Jordan chains at the infinite eigenvalue then defines the so-called index of the corresponding system of differential-algebraic equations (DAEs), and the deflated system gives the differential subsystem of the DAEs [13], [10].

Another application is the use of perfect shifts in the implicitly restarted Arnoldi method [16]. As pointed out in [17, Chapter 5, section 2.1], the polynomial filter applied in this procedure will purge a shifted eigenvalue μ only if it can be deflated after the shifted QR step has been applied. The forward instability of the QR step with a perfect shift μ may therefore fail to deflate the eigenvalue μ, resulting in an undesired Ritz value μ being locked in the spectrum of the approximated eigenspace.

The problem of forward instability of perfect shifts also occurs in the symmetric case, where the Hessenberg matrix is now tridiagonal. This was pointed out in [3], where the authors also present an alternative deflation technique which requires a two-phase procedure with both a forward and a backward decomposition of the tridiagonal matrix. This procedure is shown experimentally to almost always work and is also implemented in the MRRR method of LAPACK [15], but it is more elaborate than the new method presented in this paper.

For the sake of simplicity, we consider only real matrices, since the extension to complex matrices is straightforward. On the other hand, we will consider the case where the real matrix A has complex conjugate eigenvalues λ_0 and $\bar{\lambda}_0$, and we then apply a real double shift QR step for these two eigenvalues. We will use the following notation. Matrices and submatrices are denoted by capital letters, i.e., A, B, H. The entry (i, j) of the matrix A is denoted by the lowercase letter $a_{i,j}$. Vectors are denoted by bold letters, i.e., $\mathbf{a}, \mathbf{b}, \dots$. The identity matrix of order n is denoted by I_n and its ith column by $\mathbf{e}_i^{(n)}$, or, if there is no ambiguity, simply by I and $\mathbf{e}_i$, respectively. Generic entries different from zero in matrices or vectors are denoted by "$\times$". The unit roundoff of a computer is denoted by u and the machine epsilon by ϵ_M. For a machine using the IEEE floating point standard in double precision, we have $u \approx$ 1.11e-16 and $\epsilon_M \approx$ 2.22e-16. We denote by

$$G_i = \begin{bmatrix} I_{i-1} & & & \\ & c & -s & \\ & s & c & \\ & & & I_{n-i-1} \end{bmatrix}, \quad \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix}^T = I_2,$$

the Givens rotation acting on the consecutive rows/columns i and $i+1$ of a compatible matrix.

The paper is organized as follows. We introduce the deflation problem of a real eigenvalue in the next section and do its error analysis in section 3. In section 4 we give sufficient conditions for a perfect shift QR step and give a procedure to compute it in section 5. We discuss the extension to a double QR step for complex conjugate eigenvalues in section 6 and give some numerical examples in section 7. We then end the paper with a number of concluding remarks in section 8.

2. Deflating a real eigenvalue.

We will suppose here that we are already given the Hessenberg form $H := UAU^T$ of the matrix A, and that H is unreduced. If not, the operations described below can be applied to each unreduced submatrix of a general Hessenberg matrix H.

In exact arithmetic, if λ_0 is an eigenvalue of the unreduced matrix H and we perform one backward QR step with shift λ_0, the matrix $\tilde{H} = QHQ^T$ is still in Hessenberg form, with its first column equal to $\lambda_0\mathbf{e}_1$, and Q is an unreduced Hessenberg matrix formed by the product of $n-1$ Givens rotations $G_{n-i},\ i = 1,\dots,n-1$. Unfortunately this may not be the case anymore in finite precision because of the phenomenon known as "blurring" [20] or because of the ill-conditioning of the eigenvalue λ_0. The first column of the computed $\tilde{H}$ might be far from $\lambda_0\mathbf{e}_1$, depending on the condition number of the eigenvalue λ_0.

Therefore, we need to consider alternative constructions of the QR step, for which we recall the following theorem. Since we want to relate the rotations used in these different constructions, we will make them unique by choosing the sign of s always positive when $s \neq 0$, and choosing $c = 1$ when $s = 0$. The results of this theorem are well known (see, e.g., [14]) but since we rephrase them for the backward QR step, we repeat it here.

THEOREM 2.1. *Let $H \in \mathbb{R}^{n\times n}$ be an unreduced Hessenberg matrix with eigenvalue λ_0. Then the following hold:*

1. *H has a normalized eigenvector $\mathbf{x}$ corresponding to λ_0,*

$$H\mathbf{x} = \lambda_0\mathbf{x}, \quad \|\mathbf{x}\|_2 = 1,$$

which is unique up to a scale factor ± 1, and has its last component x_n nonzero.

2. *There is an "essentially unique" sequence of Givens rotations $G_{n-1},\dots,G_1$ whose product*

$$Q := G_1G_2\cdots G_{n-1}$$

transforms the pair $(H,\mathbf{x})$ to a similar one,

$$(\tilde{H},\tilde{\mathbf{x}}) := (QHQ^T, Q\mathbf{x}),$$

where

$$\tilde{\mathbf{x}} = \pm\mathbf{e}_1, \quad \tilde{H}\mathbf{e}_1 = \lambda_0\mathbf{e}_1, \quad \tilde{H} \text{ is in Hessenberg form.}$$

3. *The Hessenberg matrix $(H - \lambda_0 I)$ is reduced to upper triangular form R with $r_{1,1} = 0$ by the orthogonal transformation $Q^T = G_{n-1}^T\cdots G_2^T G_1^T$, yielding the factorization*

$$H - \lambda_0 I = RQ.$$

Proof. The fact that the normalized eigenvector $\mathbf{x}$ is unique (up to a scaling factor ± 1) follows from the equation

$$(H - \lambda_0 I)\mathbf{x} = 0, \quad \|\mathbf{x}\|_2 = 1,$$

where $(H - \lambda_0 I)$ has rank $n-1$ since it is unreduced and Hessenberg. For the same reason, its last component x_n is nonzero, since otherwise the whole vector $\mathbf{x}$ would be zero. The reduction of $\mathbf{x}$ to $\tilde{\mathbf{x}} = Q\mathbf{x} = \pm\mathbf{e}_1$ requires a sequence of Givens rotations,

$$G_{i-1} \in \mathbb{R}^{n\times n}, \quad i = n, n-1, \ldots, 2, \tag{2.1}$$

in order to eliminate the entries $x_i,\ i = n, n-1, \ldots, 2$, of the vector $\mathbf{x}$. By choosing the sign of s in these Givens rotations positive, we made them unique. It follows from $(H - \lambda_0 I_n)\mathbf{x} = 0$ that $\begin{bmatrix} h_{n,n-1}, & h_{n,n} - \lambda_0 \end{bmatrix} \left[\begin{smallmatrix} x_{n-1} \\ x_n \end{smallmatrix}\right] = 0$. The orthogonality of these two vectors implies then that the Givens rotation G_{n-1} eliminating x_n in the product $G_{n-1}\mathbf{x}$ is the transpose of the rotation that eliminates $h_{n,n-1}$ in the product $(H - \lambda_0 I_n)G_{n-1}^T$. We then obtain the expression

$$\left((H - \lambda_0 I_n)G_{n-1}^T\right)(G_{n-1}\mathbf{x}) = \left[\begin{array}{cccc|c} \times & \times & \ldots & \times & \times \\ \times & \times & \ldots & \times & \times \\ & \ddots & \ddots & \vdots & \vdots \\ & & \times & \times & \times \\ \hline & & & 0 & \hat{\times} \end{array}\right] \left[\begin{array}{c} \times \\ \vdots \\ \times \\ \hat{\times} \\ \hline 0 \end{array}\right] = 0,$$

where the elements $\hat{\times}$ are nonzero. Deflating the last row and column in this expression yields again an unreduced Hessenberg matrix and a corresponding eigenvector. We can thus follow the same reasoning by induction to show that the rotations transforming the vector $Q\mathbf{x}$ to $\pm\mathbf{e}_1$ are the same rotations transforming $(H - \lambda_0 I_n)Q^T$ to triangular form:

$$(H - \lambda_0 I)Q^T = \begin{bmatrix} \diagdown \end{bmatrix} = R.$$

Since Q is an upper Hessenberg matrix, it then follows that the product

$$Q(H - \lambda_0 I)Q^T = \begin{bmatrix} \diagdown \end{bmatrix} = \tilde{H} - \lambda_0 I$$

is in Hessenberg form again. This therefore shows that the upper Hessenberg transformation Q transforming the eigenvector $\mathbf{x}$ to $Q\mathbf{x} = \pm\mathbf{e}_1$ is essentially the same as the one implementing an explicit QR step. For the equivalence between the explicit QR step and the implicit QR step, we refer the reader to [14].

Finally, since $\mathbf{x} = \pm Q^T\mathbf{e}_1$, we also have

$$R\mathbf{e}_1 = r_{1,1}\mathbf{e}_1 = 0, \quad (\tilde{H} - \lambda_0 I)\mathbf{e}_1 = 0,$$

from which it follows that $\tilde{H}\mathbf{e}_1 = \lambda_0\mathbf{e}_1$. □

Remark 2.1. The implicit Q theorem is closely related to Theorem 2.1. It implies that the transformation Q can also be determined from the first rotation G_{n-1} that computes

$$\begin{bmatrix} h_{n,n-1}, & h_{n,n} - \lambda_0 \end{bmatrix} G_{n-1}^T = \begin{bmatrix} 0 & \times \end{bmatrix} \tag{2.2}$$

and from the fact that QHQ^T is still Hessenberg. This is known as "chasing the bulge" [21].

Theorem 2.1 also says that there are three alternative ways to determine the sequence of Givens rotations (2.1):

1. Determine Q from $Q\mathbf{x} = \pm\mathbf{e}_1$.
2. Determine Q from $(H - \lambda_0 I) = RQ$.
3. Determine G_{n-1} from (2.2) and the rest of Q from the Hessenberg form of $\tilde{H} := QHQ^T$.

Although these three different approaches are equivalent in exact arithmetic, their numerical implementations are different. In the following two little examples the eigenvector approach is clearly the most reliable method.

Example 2.1. Let H be the following 3×3 unreduced Hessenberg matrix with eigenvalue $\lambda_0 = 0$, given by its factorized form:

$$H = RQ, \quad R := \begin{bmatrix} 0 & 1 & 0 \\ 0 & \sqrt{\epsilon_M} & 1 \\ 0 & 0 & \sqrt{\epsilon_M} \end{bmatrix}, \quad Q := \begin{bmatrix} \sqrt{2} & -1 & 1 \\ \sqrt{2} & 1 & -1 \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} /2,$$

where R is upper triangular, Q is orthogonal and Hessenberg, and ϵ_M is the machine epsilon in double precision. The product is given by

$$H = \begin{bmatrix} 0.707106781186548 & 0.500000000000000 & -0.500000000000000 \\ 0.000000010536712 & 0.707106788637128 & 0.707106773735967 \\ 0 & 0.000000010536712 & 0.000000010536712 \end{bmatrix},$$

and the exact QR step would thus yield the transformed Hessenberg matrix

$$\tilde{H}_{exact} = QR = \begin{bmatrix} 0 & 0.707106773735967 & -0.499999992549419 \\ 0 & 0.707106788637128 & 0.499999992549419 \\ 0 & 0.000000010536712 & 0.707106791723260 \end{bmatrix},$$

which clearly yields a deflated problem with the eigenvalue $\lambda_0 = 0$ in the (1,1) position. But when recalculating the RQ factorization of H by using two Givens rotations constructed from the elements of H, we obtain

$$R_1 = \begin{bmatrix} 0.000000001471273 & 1.000000000000000 & 0 \\ 0 & 0.000000014901161 & 1.000000000000000 \\ 0 & 0 & 0.000000014901161 \end{bmatrix},$$

and for the resulting product $Q_1 R_1$,

$$\tilde{H}_1 = \begin{bmatrix} 0.000000001040347 & 0.707106773735967 & -0.499999992549419 \\ 0.000000001040347 & 0.707106788637128 & 0.499999992549419 \\ 0 & 0.000000010536712 & 0.707106791723260 \end{bmatrix},$$

which shows the blurring of the shift. The same is observed when using the implicit shift strategy, which yields essentially the same updated Hessenberg matrix (they only differ in the 16th digit)

$$\tilde{H}_2 = \begin{bmatrix} 0.000000001040347 & 0.707106773735967 & -0.499999992549419 \\ 0.000000001040347 & 0.707106788637128 & 0.499999992549419 \\ 0 & 0.000000010536712 & 0.707106791723260 \end{bmatrix}.$$

On the other hand, if we compute the right eigenvector $\mathbf{x}$ of H, it appears to be (up to ϵ_M accuracy) the vector

$$\mathbf{x} = \begin{bmatrix} 0.707106781186548 & -0.500000000000000 & 0.500000000000000 \end{bmatrix}^T,$$

and the Givens rotations used to transform this vector to $\mathbf{e}_1$ appear to implement correctly the perfect shift QR step, since now we obtain the desired result up to ϵ_M accuracy:

$$\tilde{H}_3 = \begin{bmatrix} 0 & 0.707106773735967 & -0.499999992549419 \\ 0 & 0.707106788637128 & 0.499999992549419 \\ 0 & 0.000000010536712 & 0.707106791723260 \end{bmatrix}.$$

We point out that we have put equal to zero all elements below the machine accuracy.

The next example is a symmetric one, which was considered in [3] in order to analyze the failure of perfect shifts for symmetric matrices.

Example 2.2. Let T be the symmetric tridiagonal matrix

$$T = \begin{bmatrix} 2 & 1 & & & \\ 1 & 1+\rho & \rho & & \\ & \rho & 2\rho & \rho & \\ & & \rho & 1+\rho & 1 \\ & & & 1 & 2 \end{bmatrix}.$$

For $\rho \ll 1$, the smallest eigenvalue λ_1 of T lies in the interval $(0, 2\rho)$. Let λ_1^M be the smallest eigenvalue of T computed by `eig` of MATLAB (version 2014b), and let us try to perform a perfect shift with this approximation λ_1^M. For different values of ρ, we computed one step of the implicit QR method (IQR) obtaining the matrix T^I; one step of the explicit QR method (EQR) obtaining the matrix T^E; one step of the *"eigenvector"* method (Eigv) obtaining the matrix T^V, where the eigenvector $\mathbf{x}_1$ was computed applying one step of inverse iteration to $(T - \lambda_1^M I_5)$; and one step of the *"eigenvector"* method with balancing (Eigv_B), as explained later in section 5, obtaining the matrix T^B.

In Table 2.1 the difference between the entry $(1,1)$ and λ_1^M and the absolute value of the entry in position $(2,1)$ of the computed matrices are displayed. Moreover, in the last row we give the 2-norm of the part of the matrix obtained applying the eigenvector method, below the first subdiagonal. If the implementation of the QR step is successful, the corresponding values should be of the order of $\epsilon_M \|T\|_2 \approx$ 5.8132e-16.

For the other eigenvalues of T the deflation is performed in an accurate way with all the considered methods.

TABLE 2.1
Numerical errors in the perfect shift strategies IQR, EQR, Eigv, and Eig_B.

ρ		10^{-8}	10^{-10}	10^{-12}	10^{-14}
IQRs	$\lvert t_{1,1}^I - \lambda_1^M \rvert$	2.2204e-16	8.0483e-12	3.3766e-07	2.5821e-07
IQRs	$\lvert t_{2,1}^I \rvert$	7.5101e-09	2.8369e-06	5.8108e-04	1.6067e-02
EQRs	$\lvert t_{1,1}^E - \lambda_1^M \rvert$	4.4000e-16	8.1391e-12	2.8734e-07	2.5467e-06
EQRs	$\lvert t_{2,1}^E \rvert$	1.6564e-08	2.8529e-06	5.3604e-04	1.5958e-03
Eigv	$\lvert t_{1,1}^V - \lambda_1^M \rvert$	1.3235e-23	2.5849e-26	8.0779e-28	3.1554e-30
Eigv	$\lvert t_{2,1}^V \rvert$	6.0072e-15	2.9330e-17	3.6704e-16	1.2927e-17
Eigv	$\mathrm{tril}(T^V, -2)$	3.2725e-15	2.2572e-16	1.3975e-16	4.9607e-17
Eigv_B	$\lvert t_{1,1}^B - \lambda_1^M \rvert$	1.3235e-23	2.5849e-26	4.0390e-28	3.1554e-30
Eigv_B	$\lvert t_{2,1}^B \rvert$	2.1766e-24	5.1699e-26	8.0779e-28	3.1554e-30
Eigv_B	$\mathrm{tril}(T^B, -2)$	4.8057e-24	8.7043e-26	1.6339e-28	3.5734e-30

In [3] an alternative method for implementing the perfect shifts was presented for this example, but it is more involved than the eigenvector approach presented in this paper. We will also analyze when the eigenvector method can be proved to be successful.

3. Error analysis of a QR step.

In this section we briefly recall the error analysis of a QR step. For this, we use the model described in [8] for inexact arithmetic on a machine with unit roundoff u and $\gamma_n := \frac{nu}{1-nu}$, and we will ignore the effects of gradual underflow. For the proof, we refer the reader to the appendix.

THEOREM 3.1. *Let H be an unreduced Hessenberg matrix, and let λ_0 be an arbitrary shift. Let the sequence of Givens transformations G_i be constructed from the explicit QR step, and let Q be the accumulated product of the corresponding exactly orthogonal transformations $\tilde{G}_i$. Then, in inexact arithmetic, the computed Hessenberg matrix $\tilde{H}$ satisfies*

$$Q(H + \Delta_H)Q^T = \tilde{H} + \Delta_{\tilde{H}} \tag{3.1}$$

with

$$\|\Delta_H\|_F \leq \gamma_{cn}\|H - \lambda_0 I_n\|_F, \quad \|\Delta_{\tilde{H}}\|_F \leq \gamma_{cn}\|\tilde{H} - \lambda_0 I_n\|_F,$$

where the perturbations Δ_H and $\Delta_{\tilde{H}}$ are Hessenberg as well and where c is a moderate constant of the order of 1*, provided the rotation parameters are computed via the standard construction.*

Remark 3.1. We point out here that Theorem 3.1 does not apply to the implicit QR step. For this, one can prove the weaker result that $Q(H + \Delta_H)Q^T = \tilde{H}$, where the backward error Δ_H satisfies $\|\Delta_H\|_F \leq 2\gamma_{cn}\|H\|_F$, but without the constraint that it is Hessenberg. We refer the reader again to the appendix for its proof.

Remark 3.2. Notice that the bounds for $\|\Delta_H\|_F$ in Theorem 3.1 and Remark 3.1 are comparable, since $\|H - \lambda_0 I_n\|_F$ and $\|H\|_F$ are almost equal when the shift λ_0 is selected according to the standard QR procedure.

We will then say that the QR step with shift λ_0 is "perfect," provided $\tilde{H}$ is Hessenberg and if, moreover, the $(1,1)$ and $(2,1)$ entries of $\tilde{H} + \Delta_{\tilde{H}}$ satisfy

$$\tilde{h}_{1,1} + \tilde{\delta}_{1,1} = \lambda_0, \quad \tilde{h}_{2,1} + \tilde{\delta}_{2,1} = 0,$$

or equivalently, up to machine accuracy, one would have

$$\tilde{h}_{1,1} \approx \lambda_0, \quad \tilde{h}_{2,1} \approx 0.$$

This would imply that $\mathbf{e}_1$ is the eigenvector of $\tilde{H} + \Delta_{\tilde{H}}$ corresponding to the eigenvalue λ_0 and that the vector $\mathbf{x} := Q^T\mathbf{e}_1$ is the *exact* eigenvector of a perturbed Hessenberg matrix $H + \Delta_H$ corresponding to its *exact* eigenvalue λ_0. Notice that the use of the forward error $\Delta_{\tilde{H}}$ is needed for this interpretation. Usually, a tolerance τ is specified for the errors Δ_H and $\Delta_{\tilde{H}}$ in (3.1) that is of the order $\epsilon_M\|H\|_F$ and compatible with the bound of Theorem 3.1 or Remark 3.1, i.e., $\tau \geq \gamma_{cn}\max(\|H - \lambda_0 I_n\|_F, 2\|H\|_F)$. In what follows, we will insist that the backward error Δ_H is Hessenberg, because we will be able to construct such a perturbation.

DEFINITION 3.2. *A (backward) QR step with shift λ_0 is "perfect" if it corresponds to a perturbed Hessenberg matrix $H + \Delta_H$ with $\|\Delta_H\|_F \leq \tau$ for which the (backward) QR step satisfies* (3.1) exactly *and for which* $(\lambda_0, \mathbf{x})$ *is an* exact *eigenvalue/eigenvector pair. Moreover, the property that λ_0 is an exact eigenvalue of the transformed* *matrix $\tilde{H}$ is made possible by a perturbation $\Delta_{\tilde{H}}$ of norm $\|\Delta_{\tilde{H}}\|_F \le \tau$ by setting the elements $\tilde{h}_{1,1}$ and $\tilde{h}_{2,1}$ to the nearby values λ_0 and 0, respectively.*

Notice that this error analysis does not say if, for a given matrix H, a shift λ_0 will be "perfect" and show up in the $(1,1)$ position of the computed matrix $\tilde{H}$ since we do not know what backward errors correspond to it, and these can affect the forward errors a lot. One would think that it suffices to have the property

$$\|(H - \lambda_0 I_n)\mathbf{x}\|_2 \approx \epsilon_M \|H - \lambda_0 I_n\|_2, \tag{3.2}$$

where $\mathbf{x}$ is the presumed eigenvector since it yields a small residual, and where ϵ_M is the machine epsilon of the computer used. This would imply that the first column of $(\tilde{H} - \lambda_0 I)$ is of the order of $\epsilon_M \|H - \lambda_0 I_n\|_2$, or in other words

$$\tilde{h}_{1,1} - \lambda_0 \approx \epsilon_M \|H - \lambda_0 I_n\|_2, \quad \tilde{h}_{2,1} \approx \epsilon_M \|H - \lambda_0 I_n\|_2,$$

and this would mean that the $(1,1)$-element is very close to λ_0 and can be deflated. But clearly this is not what happens in Example 3.1 below, since that condition is not sufficient there.

Example 3.1. Let H be the following 3×3 unreduced Hessenberg matrix,

$$H := \begin{bmatrix} 0 & 1 & \sqrt[3]{\epsilon_M} \\ \sqrt[3]{\epsilon_M} & 0 & 1 \\ 0 & \sqrt[3]{\epsilon_M} & 0 \end{bmatrix}, \quad \|H\|_2 \approx 1,$$

and let us suppose that we have an approximate eigenvalue/eigenvector pair $(\lambda_0, \mathbf{x})$, where $\lambda_0 = 0$ and $\mathbf{x} = \begin{bmatrix} 1 & \sqrt[3]{\epsilon_M^2} & -\sqrt[3]{\epsilon_M} \end{bmatrix}^T$, with residual

$$(H - \lambda_0 I)\mathbf{x} = \epsilon_M \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The ϵ_M-small residual indicates that there exists an ϵ_M-perturbation Δ that makes $H + \Delta$ singular, but it is not Hessenberg. For example,

$$(H + \Delta - \lambda_0 I)\mathbf{x} = (H + \Delta) \begin{bmatrix} 1 \\ \sqrt[3]{\epsilon_M^2} \\ -\sqrt[3]{\epsilon_M} \end{bmatrix} = 0 \quad \text{for} \quad \Delta := \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\epsilon_M & 0 & 0 \end{bmatrix}.$$

If we insist that the perturbation is Hessenberg and with the same eigenvector $\mathbf{x}$, then we have as minimum norm solution

$$\Delta_H = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -\epsilon_M & \sqrt[3]{\epsilon_M^2} \end{bmatrix} \Big/ \left(1 + \sqrt[3]{\epsilon_M^2}\right), \quad \text{with} \quad \|\Delta_H\|_2 \approx \sqrt[3]{\epsilon_M^2}.$$

So in order to be able to perform a perfect QR step, we must know the eigenvalue/eigenvector pair $(\lambda_0, \mathbf{x})$ to higher accuracy. But what precision is actually needed?

4. Sufficient conditions for a perfect shift $\boldsymbol{QR}$ step.

In general, we do not have an exact eigenvalue/eigenvector pair of a given matrix H, but rather just an estimate of the eigenvector $\mathbf{x}$ and corresponding eigenvalue λ_0. We have shown in Example 3.1 that this can be quite problematic in general. But as we will see below, it suffices to know some of the components of the eigenvector to a higher accuracy.

Let us assume that $H-\lambda_0 I$ is nearly singular in the sense that its smallest singular value $\underline{\sigma} := \sigma_{\min}(H-\lambda_0 I)$ is equal to $\epsilon\|H-\lambda_0 I\|_2$, with ϵ of the order of the machine accuracy. This suggests that λ_0 might be a good choice for a perfect shift. Let us then choose as approximate eigenvector the vector $\mathbf{v}$ minimizing the residual

$$\min_{\mathbf{v}} \|(H-\lambda_0 I)\mathbf{v}\|_2, \quad \|\mathbf{v}\|_2 = 1. \tag{4.1}$$

An optimal solution $\mathbf{v}$ to this problem is given by the right singular vector of $(H-\lambda_0 I)$:

$$(H-\lambda_0 I)\mathbf{v} = \mathbf{u}, \quad \|\mathbf{u}\|_2 = \underline{\sigma} := \sigma_{\min}(H-\lambda_0 I). \tag{4.2}$$

From this, one also finds the minimum norm perturbation $\Delta = -\mathbf{u}\mathbf{v}^T$ of 2-norm $\underline{\sigma}$ ensuring that $\mathbf{v}$ is a true eigenvector of $H+\Delta$,

$$(H+\Delta-\lambda_0 I)\mathbf{v} = \mathbf{0},$$

but this solution is not Hessenberg in general. In the next lemma we look for the minimum norm perturbation while imposing this Hessenberg structure, starting from an arbitrary pair of vectors $(\mathbf{u},\mathbf{v})$ satisfying

$$\|\mathbf{v}\|_2 = 1, \quad \mathbf{u} = (H-\lambda_0 I)\mathbf{v}. \tag{4.3}$$

LEMMA 4.1. *The minimum Frobenius norm solution Δ_H of Hessenberg form*

$$\Delta_H = \begin{bmatrix} \delta h_{1,1} & \delta h_{2,1} & \cdots & \delta h_{1,n} \\ \delta h_{2,1} & \delta h_{2,2} & \cdots & \delta h_{2,n} \\ & \ddots & \cdots & \vdots \\ & & \delta h_{n,n-1} & \delta h_{n,n} \end{bmatrix}$$

to the system

$$(H+\Delta_H-\lambda_0 I)\mathbf{v} = \mathbf{0}, \quad (H-\lambda_0 I)\mathbf{v} = \mathbf{u}, \tag{4.4}$$

where $\|\mathbf{v}\|_2 = 1$ and $v_n \neq 0$, has Frobenius norm equal to

$$\|\Delta_H\|_F = \|u_1/\nu_1, u_2/\nu_2, \ldots, u_n/\nu_n\|_2, \tag{4.5}$$

where

$$\nu_1 = 1, \quad \nu_i = \|[v_{i-1}, v_i, \ldots, v_{n-1}, v_n]\|_2, \ i = 2, \ldots, n.$$

Proof. It follows from (4.4) that

$$\Delta_H \mathbf{v} = -\mathbf{u},$$

which is a linear set of equations. It has a minimum Frobenius norm solution which we solve row by row. Let row i of this equation be

$$\left[0, \ldots, 0, \underbrace{\boldsymbol{\delta h}_i^T}_{\min(n,n-i+2)} \right] \mathbf{v} = -u_i, \tag{4.6}$$

with

$$\boldsymbol{\delta h}_i^T = [\delta h_{i,i-1}, \delta h_{i,i}, \dots, \delta h_{i,n}],$$

and let

$$\mathbf{v}_i^T = [v_{i-1}, v_i, \dots, v_{n-1}, v_n], \quad \nu_i := \|\mathbf{v}_i\|_2,$$

be the subvector of $\mathbf{v}$ involved in (4.6). Then clearly the minimum norm solution of this equation is

$$\boldsymbol{\delta h}_i^T = -\frac{u_i \mathbf{v}_i^T}{\|\mathbf{v}_i\|_2^2}, \quad \|\boldsymbol{\delta h}_i\|_2 = \frac{|u_i|}{\nu_i} \quad \text{for} \quad i = 2, \dots, n$$

and

$$\boldsymbol{\delta h}_1^T = -u_1 \mathbf{v}^T, \quad \|\boldsymbol{\delta h}_1\|_2 = |u_1|.$$

Since these equations all involve independent rows of Δ_H, the result follows. □

Remark 4.1. It is easy to see that the subsequent vector norms ν_i satisfy the inequalities

$$\nu_n \le \cdots \le \nu_3 \le \nu_2 = \nu_1 = 1,$$

where $\nu_n = \| \begin{bmatrix} v_{n-1} & v_n \end{bmatrix} \|_2$. Therefore the Frobenius norm for Δ_H is bounded by

$$\|\Delta_H\|_F \le \frac{\|\mathbf{u}\|_2}{\nu_n},$$

and hence the Hessenberg perturbation Δ_H is then of the same order as the unstructured perturbation Δ if $\nu_n \approx 1$.

Another way to guarantee a bound for Δ_H that is of the same order as the unstructured error Δ is to compute the approximate eigenvector $\mathbf{x}$ in such a way that the residual vector (which we now denote by $\mathbf{r} = (H - \lambda_0 I)\mathbf{x}$) satisfies stricter conditions. This is shown in the next theorem.

THEOREM 4.2. *Let H be an unreduced Hessenberg matrix, and let us have an estimate of an eigenvalue/eigenvector pair $(\lambda_0, \mathbf{x})$ satisfying*

$$\mathbf{r} := (H - \lambda_0 I)\mathbf{x}, \quad \|\mathbf{x}\|_2 = 1,$$

where

$$\begin{aligned} &\mathbf{x}_i^T := [x_{i-1}, x_i, \dots, x_{n-1}, x_n], \quad \nu_i := \|\mathbf{x}_i\|_2, \; i = 2, \dots, n, \\ &\nu_1 = 1, \quad \hat\epsilon_i := r_i/\nu_i, \quad \text{and} \quad \|[\hat\epsilon_1, \hat\epsilon_2, \dots, \hat\epsilon_n]\|_2 \le \epsilon_M \|H\|_F, \end{aligned} \tag{4.7}$$

and let the Givens rotations G_i be computed to annihilate element x_{i+1} for $i = n-1, \dots, 1$ of the approximate eigenvector $\mathbf{x}$ and transforming it to $\mathbf{e}_1$. Then the product

$$Q := \tilde G_1 \tilde G_2 \cdots \tilde G_{n-1},$$

where each $\tilde G_i$ is the exactly orthogonal Givens rotation corresponding to G_i, yields a perfect shift implementation of the QR step applied to H, in the sense that there exist backward perturbations Δ_H and $\Delta_{\mathbf{x}}$ such that

$$Q(H + \Delta_H)Q^T = \tilde H, \quad Q(\mathbf{x} + \Delta_{\mathbf{x}}) = \mathbf{e}_1, \quad \|\Delta_H\|_F \le c\epsilon_M \|H\|_F, \quad \|\Delta_{\mathbf{x}}\|_2 \le c\epsilon_M,$$

where c is a constant of the order of 1.

Proof. We first prove that the Givens rotations constructed to annihilate the successive entries of x_{i+1} for $i = n-1, \dots, 1$ of the approximate eigenvector $\mathbf{x}$ also yield the RQ factorization of a nearby Hessenberg matrix, as indicated in Theorem 3.1 and Lemma 4.1. We can solve for the backward perturbation Δ_H recursively by only considering problems in $\mathbb{R}^2$. For the first step we define

$$\mathbf{h} := \begin{bmatrix} h_{n-1,n} \\ h_{n,n} - \lambda_0 \end{bmatrix}, \quad \mathbf{v} := \begin{bmatrix} x_{n-1} \\ x_n \end{bmatrix}.$$

Then our conditions imply that $\mathbf{h}^T\mathbf{v}/\|\mathbf{v}\|_2 = r_n/\nu_n = \hat{\epsilon}_n \le \epsilon_M \|H\|_F$. Obviously we are satisfying the condition of Lemma A.2, and the transformation constructed to eliminate x_n in $\mathbf{v}$ can thus be applied safely so that there exists an $\hat{\epsilon}_n$-small perturbation in the bottom row of $(H - \lambda_0 I)$ that yields a "0" in the $(n, n-1)$ position of the matrix R (the relevant elements are indicated with a $\hat{\ }$). After applying this first Givens transformation G_{n-1}, the effect of the first Δ_H thus yields

$$(H + \Delta_H - \lambda_0 I) = \begin{bmatrix} \times & \times & \dots & \times & \times \\ \times & \times & \dots & \times & \times \\ & \ddots & & \vdots & \vdots \\ & & \times & \times & \times \\ & & & \hat{\times} & \hat{\times} \end{bmatrix} = \left[\begin{array}{ccccc|c} \times & \times & \dots & \dots & & \times \\ \times & \times & \dots & \dots & & \times \\ & \ddots & & & & \vdots \\ & & & \times & \times & \times \\ \hline & & & & \hat{0} & \hat{r} \end{array}\right] \tilde{G}_{n-1},$$

where $\tilde{G}_{n-1}$ is the exactly orthogonal Givens rotation eliminating x_n, and the guaranteed bound for Δ_H is

$$\|\Delta_H\|_F \le r_n/\nu_n = \hat{\epsilon}_n \le \epsilon_M \|H\|_F.$$

Since now we also have

$$\tilde{G}_{n-1}\mathbf{x} = \left[\begin{array}{c} \times \\ \vdots \\ \times \\ \hline 0 \end{array}\right],$$

we need only consider the "deflated" problem of order $n-1$, which is again in Hessenberg form, and the deflated eigenvector $\mathbf{x}$ corresponding to it. Taking Lemma A.2 into account, the next set of Givens rotations $G_{n-i},\ i = 2, \dots, n-1$, is computed in a similar way, and for each of them the conditions (4.7) are exactly those of Lemma A.2. Moreover, the squared Frobenius norm of Δ_H grows with $\hat{\epsilon}^2_{n-i} = (r_{n-i}/\nu_{n-i})^2$ at each step. Therefore we obtain by induction that the RQ factorization step of the QR step yields the required backward bound $\|\Delta_H\|_F \le c.\epsilon_M\|H\|_F$. The bound $\|\Delta_\mathbf{x}\|_2 \le \epsilon_M \|\mathbf{x}\|_2 = c.\epsilon_M$ follows from the standard analysis of Givens rotations applied to a vector. One then still needs to perform the Givens transformations on the left of the matrix R to complete the similarity transformation, but it follows from the proof of Theorem 3.1 that the increase of Δ_H in this second step is of the same order as in the first step, and that it stays Hessenberg. □

We summarize this so-called *eigenvector* method below by giving a pseudocode.

```
1) function  $[H, \lambda, \mathbf{x}]$ = eigenvector_method($H, \mathbf{x}, n$);
2)   for $i = n-1 : -1 : 1$,
3)       $G$ = givens($x_i, x_{i+1}$);
4)       $\mathbf{x}_{i:i+1} = G\mathbf{x}_{i:i+1}$;
```

```
        $H_{i:i+1,:} = GH_{i:i+1,:}$;
        $H_{:,i:i+1} = H_{:,i:i+1}G^T$;
    end;
    $\lambda = H_{1,1}$;
```

The key point in this theorem is of course that we need an approximate eigenvector $\mathbf{x}$ with a sufficiently small residual, especially in the components where each trailing subvector $\mathbf{x}_i$ has small norm ν_i. We explain in the next section how to compute such an approximation.

5. Computing an eigenvector by scaled inverse iteration. In this section we show how to compute an approximate eigenvector $\mathbf{x}$ of an unreduced Hessenberg matrix such that its residual $\mathbf{r}$ satisfies the conditions (4.7) requested by Theorem 4.2.

The basic idea here is to apply a diagonal scaling (with $d \geq 1$)

$$D := \mathrm{diag}(1, d, d^2, \ldots, d^{n-1}) \tag{5.1}$$

that "balances" the entries of $\mathbf{x}$ without affecting too much the norm of H.

THEOREM 5.1. *Let H be an unreduced Hessenberg matrix, and let $(\lambda_0, \mathbf{x})$ be an approximate eigenvalue/eigenvector pair. Then there always exists a scaling D of the form (5.1) such that the transformed pair $(H_D, \mathbf{x}_D) := (DHD^{-1}, D\mathbf{x})$ satisfies the constraints*

$$\|H_D\|_F \leq d\|H\|_F, \quad d := \max\left(\min[\max_{i\leq n-2}\{|x_i/x_{n-1}|^{1/n-i-1}\}, \max_{i\leq n-2}\{|x_i/x_n|^{1/n-i}\}], 1\right)$$

and such that the largest component of $\mathbf{x}_D$ is one of its last two components.

Proof. The elements of the scaled vector $\mathbf{x}_D$ are all nonzero and are given by

$$\mathbf{x}_D = \begin{bmatrix} x_1 & dx_2 & \ldots & d^{n-2}x_{n-1} & d^{n-1}x_n \end{bmatrix}.$$

The element $d^{n-2}|x_{n-1}|$ will be larger than all $d^{i-1}|x_i|$ for $i \leq n-2$ if and only if

$$d^{n-2}|x_{n-1}| \geq d^{i-1}|x_i| \quad \Rightarrow \quad d^{n-i-1} \geq |x_i/x_{n-1}| \quad \Rightarrow \quad d \geq |x_i/x_{n-1}|^{1/n-i-1}.$$

If $x_{n-1} = 0$, this quantity is clearly infinite and must be dismissed. We also compare the element $d^{n-1}|x_n|$ to all $d^{i-1}|x_i|$ for $i \leq n-2$, implying

$$d^{n-1}|x_n| \geq d^{i-1}|x_i| \quad \Rightarrow \quad d^{n-i} \geq |x_i/x_n| \quad \Rightarrow \quad d \geq |x_i/x_n|^{1/n-i}.$$

Since $x_n \neq 0$ this value for d is finite. If one of these two values for d turns out smaller than 1, then the largest element of $\mathbf{x}$ was already in position $n-1$ or n, and we should then choose $d = 1$ (i.e., no scaling). If both values are larger than 1, we choose the best conditioned transformation, i.e., the smallest of both values for d. This explains the value for d. For the bound on $\|H_D\|_F$ it suffices to point out that all nonzero elements of H are scaled by a number smaller than or equal to d. $\square$

Remark 5.1. The bound $\|H_D\|_2 \leq d/(1-d^{-2})\|H\|_2$ for the 2-norm follows from [9, section 5.5.18]. For $d \geq 2$ this can also be bounded by $\|H_D\|_2 \leq \frac{4d}{3}\|H\|_2$. Moreover, in practice we typically have $\|H_D\|_2 \approx \|H\|_2$ and $\|H_D\|_F \approx \|H\|_F$ since only $n-1$ elements of H are scaled with d, while most of the other elements decrease. The same obviously holds for the shifted matrices $H_D - \lambda_0 I$ and $H - \lambda_0 I$. Finally, for matrices of dimension $n \geq 10$ the values for d are often smaller than 10 since the scaling on the last two elements of $\mathbf{x}$ are of the order of d^{n-2}.

Remark 5.2. In order to make sure that the scaling process with the diagonal transformation D does not introduce too much roundoff noise, we can approximate d by the nearest power of 2. In that way, the transformation to the scaled system $(H_D, \mathbf{x}_D) := (DHD^{-1}, D\mathbf{x})$, as well as its inverse transformation, can be implemented without any roundoff.

LEMMA 5.2. *Let $\mathbf{x}_D$ be an approximate normalized eigenvector of H_D corresponding to eigenvalue λ_0, and let the residual $\mathbf{r}_D := (H_D - \lambda_0 I)\mathbf{x}_D$ be computed with accuracy $\|\mathbf{r}_D\|_2 \le \epsilon \|H_D - \lambda_0 I\|_2$. Then*

$$r_i^{(D)} \le \|\mathbf{r}_D\|_2 \le \epsilon\sqrt{n}\|H_D - \lambda_0 I\|_2 \nu_i^{(D)}, \tag{5.2}$$

and the rescaled residual $\mathbf{r} := D^{-1}\mathbf{r}_D$ satisfies the bound

$$r_i \le \epsilon\sqrt{n}\|H_D - \lambda_0 I\|_2 \nu_i^{(D)}/d^{i-1}. \tag{5.3}$$

If we assume $d \ge 2$ and use the bound

$$\nu_i^{(D)}/d^{i-2} \le c_\nu \nu_i, \tag{5.4}$$

we obtain the simplified inequality

$$r_i \le (4c_\nu\sqrt{n}/3)\epsilon\|H - \lambda_0 I\|_2 \nu_i \tag{5.5}$$

in terms of the rescaled vector $\mathbf{x} := D^{-1}\mathbf{x}_D$ and its subnorms ν_i.

Proof. We have the result $\|\mathbf{r}_D\|_2 = \|(H_D - \lambda_0 I)\mathbf{x}_D\|_2$, for which we have reached a bound $\epsilon\|H_D - \lambda_0 I\|_2$. Because of the scaling technique and the construction of d we know that there are two “largest” elements of $\mathbf{x}_D$ of equal magnitude. One of them is in the last two components, and the other one is in the first $n-2$ components. As a consequence of this, we know that the norms $\nu_i^{(D)}$ of the subvectors of $\mathbf{x}_D$ satisfy

$$1/\sqrt{n} \le \nu_i^{(D)} \le 1,$$

and hence are approximately of the same size. This then yields the desired bound (5.2). Since $r_i = r_i^{(D)}/d^{i-1}$ we also have the bound (5.3). Finally, using the bound (5.4) and Remark 5.1 yields the inequality (5.5). ◻

Remark 5.3. In the above lemma, we would have hoped for the stronger bound

$$r_i \le \epsilon\|H - \lambda_0 I\|_2 \nu_i.$$

The following remarks encourage us to believe that such a result is often true. One can expect that $\mathbf{x}_D$ and $\mathbf{r}_D$ are both “balanced” in some sense, because they result from a minimization problem with a well-distributed eigenvector and residual:

$$r_i^{(D)} \approx \|\mathbf{r}_D\|_2/\sqrt{n} \le \epsilon\|H_D - \lambda_0 I\|_2 \nu_i^{(D)}.$$

This type of inequality is preserved when transforming back $r_i = r_i^{(D)}/d^{i-1}$ and $\nu_i \approx \nu_i^{(D)}/d^{i-2}$, provided the elements of $\mathbf{x}_D$ are randomly distributed. Under these assumptions, one can expect that $c_\nu \approx 1$ and that the factor $\sqrt{n}$ in (5.2)–(5.5) can be dismissed.

The described algorithm is summarized in the following pseudocode, which uses the function `eigenvector_method` described earlier.

```
    Proposed algorithm
Input:   $H$, Hessenberg matrix;
         $(\lambda_0, \mathbf{x})$, approximation of an eigenpair of $H$;
         $\epsilon$, the tolerance
Output:  $\hat{H}, (\hat{\lambda}, \hat{\mathbf{x}})$, the deflated Hessenberg matrix and the computed eigenpair

compute $\mathbf{r}(H, \lambda_0, \mathbf{x}) = (H - \lambda_0 I)\mathbf{x}$;
compute $\nu_i := \|\mathbf{x}_{i-1:n}\|_2, i = 2, \ldots, n,\ \nu_1 := 1$; and $\hat{\epsilon}_i := r_i/\nu_i,\ i = 1, \ldots, n,$
if $\|[\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n\|_2 \leq \epsilon\|H\|_F$,
    $[\hat{H}, \hat{\lambda}, \hat{\mathbf{x}}] =$ eigenvector_method$(H, \mathbf{x})$;
else
    compute $d$ as in § 4
    $H_D = DHD^{-1}$;
    $\mathbf{x}_D = D\mathbf{x}$; $\mathbf{x}_D = \mathbf{x}_D/\|\mathbf{x}_D\|_2$;
    % apply one step of inverse iteration
    $\mathbf{x}_D = (H_D - \lambda_0 I)\mathbf{x}_D$; $\mathbf{x}_D = \mathbf{x}_D/\|\mathbf{x}_D\|_2$;
    $\hat{\lambda} = \mathbf{x}_D^T H_D \mathbf{x}_D$;
    % back to the initial coordinate system
    $\mathbf{x}_b = D^{-1}\mathbf{x}_D$; $\mathbf{x}_b = \mathbf{x}_b/\|\mathbf{x}_b\|_2$;
    $[\hat{H}, \hat{\lambda}, \hat{\mathbf{x}}] =$ eigenvector_method$(H, \mathbf{x}_b)$;
end
```

6. Complex conjugate roots and the perfect double shift.

If a matrix H has a pair of complex conjugate eigenvalues $\lambda_0 = \alpha + \jmath\beta$ and $\bar{\lambda}_0 = \alpha + \jmath\beta$, then it still has a real invariant subspace $\mathcal{V}$ of dimension 2, corresponding to the eigenvectors

$$H(\mathbf{v} + \jmath\mathbf{w}) = (\alpha + \jmath\beta)(\mathbf{v} + \jmath\mathbf{w}), \quad H(\mathbf{v} - \jmath\mathbf{w}) = (\alpha - \jmath\beta)(\mathbf{v} - \jmath\mathbf{w}). \tag{6.1}$$

This can indeed be rewritten as a real matrix equation,

$$HV = V\Lambda_0, \quad \text{where} \quad V := \left[\begin{array}{cc} \mathbf{v} & \mathbf{w} \end{array}\right],\ \Lambda_0 := \left[\begin{array}{cc} \alpha & \beta \\ -\beta & \alpha \end{array}\right], \tag{6.2}$$

which defines the invariant subspace $\mathcal{V}$ of H spanned by the columns of V, and with constrained spectrum $\mathrm{sp}(\Lambda_0) = \{\lambda_0, \bar{\lambda}_0\}$. When choosing a different basis for the same space, such as an orthonormal basis $\hat{V}$ obtained from the QR decomposition of V, one obtains a similar equation with the same constrained spectrum,

$$H\hat{V} = \hat{V}\hat{\Lambda}_0, \quad \text{where} \quad \hat{V} := VR^{-1},\ \hat{\Lambda}_0 := R\Lambda_0 R^{-1},\ \mathrm{sp}(\hat{\Lambda}_0) = \mathrm{sp}(\Lambda_0). \tag{6.3}$$

Finally, $\mathcal{V}$ is also in the kernel of the real matrix

$$H_{\Lambda_0} := H^2 - (\lambda_0 + \bar{\lambda}_0)H + (\lambda_0\bar{\lambda}_0)I_n = H^2 - 2\alpha H + (\alpha^2 + \beta^2)I_n$$

since both eigenvectors $\mathbf{v} \pm \jmath\mathbf{w}$ are in its kernel and span the real subspace $\mathcal{V}$. Moreover, if H is an unreduced Hessenberg matrix, H_{Λ_0} has a nonzero second subdiagonal, and its nullity is at most 2. Therefore its kernel is exactly $\mathcal{V}$. For the same reason, the bottom 2×2 submatrix of $\hat{V}$ is invertible, since otherwise the basis $\hat{V}$ would be rank deficient.

We now give different alternative constructions of the double real QR step for pairs of complex conjugate eigenvalues of a real Hessenberg matrix (see also [5, section 7.4.5] for further details). For this purpose, we make the orthonormal basis of $\mathcal{V}$ essentially unique by applying an additional orthogonal column transformation U to the orthonormal basis $\hat{V}$ to annihilate the $(n,1)$-element in the product $\hat{V}U$. In other words, U triangularizes the bottom 2×2 submatrix of $\hat{V}$. Let us denote this product as X and its columns as $\mathbf{x}$ and $\mathbf{y}$:

$$X := \hat{V}U = \left[\begin{array}{cc} \mathbf{x} & \mathbf{y} \end{array}\right], \quad \text{where} \quad x_n = 0,\ x_{n-1} \neq 0,\ y_n \neq 0.$$

It then follows that all such orthonormal bases are unique up to a diagonal sign scaling. In the subsequent lemma, we again make the Givens rotations unique by choosing the sign of s always positive when $s \neq 0$, and $c = 1$ when $s = 0$.

Lemma 6.1. *Let $H \in \mathbb{R}^{n\times n}$ be an unreduced Hessenberg matrix with complex conjugate eigenvalues λ_0 and $\bar{\lambda}_0$. Then the following hold:*

1. *H has an invariant subspace corresponding to the spectrum $\lambda_0, \bar{\lambda}_0$ with orthonormal basis X,*

$$HX = X\Lambda_0, \quad \mathrm{sp}(\Lambda_0) = \{\lambda_0, \bar{\lambda}_0\}, \quad X^TX = I_2,$$

which is unique up to a sign scaling $S := \left[\begin{smallmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{smallmatrix}\right]$ and has $x_n = 0$ and nonzero x_{n-1} and y_n,

2. *There are "essentially unique" sequences of Givens rotations $G^{(x)}_{n-2}, \ldots, G^{(x)}_1$ and $G^{(y)}_{n-1}, \ldots, G^{(y)}_2$ whose product*

$$Q := (G^{(y)}_2 G^{(y)}_3 \cdots G^{(y)}_{n-1})(G^{(x)}_1 G^{(x)}_2 \cdots G^{(x)}_{n-2}) \tag{6.4}$$

transforms the pair (H, X) into a similar one,

$$(\tilde{H}, \tilde{X}) := (QHQ^T, QX),$$

where

$$\tilde{X} = E_1.S := \left[\begin{array}{cc} \mathbf{e}_1 & \mathbf{e}_2 \end{array}\right].S, \quad \tilde{H}\tilde{X} = \tilde{X}\Lambda_0, \quad \tilde{H} \text{ is in Hessenberg form.}$$

3. *The matrix $(H - \lambda_0 I)(H - \bar{\lambda}_0 I)$ is reduced to upper-triangular form R, with leading 2×2 block $R_{1,1} = 0$, by the "essentially unique" column orthogonal transformation Q^T, yielding the factorization*

$$H_{\Lambda_0} = H^2 - (\lambda_0 + \bar{\lambda}_0)H + (\lambda_0\bar{\lambda}_0)I_n = RQ.$$

Proof. The fact that the normalized basis X is unique (up to a diagonal sign scaling factor S) follows from the equation

$$H_{\Lambda_0}X = 0, \quad X^TX = I_n,$$

where H_{Λ_0} has rank $n-2$ since it has a nonzero second subdiagonal. The only degree of freedom is a 2×2 orthogonal transformation acting on the columns of X, but since the bottom 2×2 block of X is $\left[\begin{smallmatrix} x_{n-1} & y_{n-1} \\ 0 & y_n \end{smallmatrix}\right]$ with nonzero diagonal, that degree of freedom reduces to a diagonal sign scaling S.

The reduction of X to $\tilde{X} = QX = E_1S$ requires a first sequence of Givens rotations,

$$G^{(x)}_{i-1} \in \mathbb{R}^{n\times n}, \quad i = n-1, \ldots, 2, \tag{6.5}$$

in order to eliminate the entries $x_i,\ i = n-1, \dots, 2$, of the vector $\mathbf{x}$. Then it requires a second sequence of Givens rotations,

$$G_{i-1}^{(y)} \in \mathbb{R}^{n\times n}, \quad i = n, \dots, 3, \tag{6.6}$$

in order to eliminate the entries $y_i,\ i = n, \dots, 3$, of the vector $\mathbf{y}$. The only degrees of freedom for these Givens rotations lie in a left and right diagonal scaling with elements of modulus 1. After these rotations have been applied, $\tilde{X} := \left[\begin{array}{cc} \tilde{\mathbf{x}} & \tilde{\mathbf{y}} \end{array}\right]$ still has orthonormal columns, which implies that $\tilde{\mathbf{x}} = \pm\mathbf{e}_1$ and $\tilde{\mathbf{y}} = \pm\mathbf{e}_2$, and hence proves point 1.

These Givens rotations are the same ones that reduce

$$H_{\Lambda_0} Q^T = \left[\diagdown\right] = R$$

to upper-triangular form. To see this, we note that most of the Givens rotations in Q commute with each other, which implies that we can also write

$$Q := (G_2^{(y)} G_1^{(x)})(G_3^{(y)} G_2^{(x)}) \cdots (G_{n-1}^{(y)} G_{n-2}^{(x)}). \tag{6.7}$$

After applying the first pair of rotations $Q_1 := (G_{n-1}^{(y)} G_{n-2}^{(x)})$ to X and their transpose to H_{Λ_0}, we clearly have the pattern

$$(H_{\Lambda_0} Q_1^T)(Q_1 X) = \left[\begin{array}{ccccc|c} \times & \times & \times & \dots & \times & \times \\ \times & \times & \times & \dots & \times & \times \\ \times & \times & \times & \dots & \times & \times \\ & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & \times & \times & \times & \times \\ \hline & & & \hat{0} & \hat{0} & \hat{\times} \end{array}\right] \left[\begin{array}{cc} \times & \times \\ \vdots & \vdots \\ \times & \times \\ \hat{\times} & \hat{\times} \\ \hat{0} & \hat{\times} \\ \hline \hat{0} & \hat{0} \end{array}\right] = 0$$

since the bottom row of $(H_{\Lambda_0} Q_1^T)$ must be orthogonal to $(Q_1 X)$ (the relevant elements are indicated with a $\hat{\ }$). Repeating this recursively to $Q_{i-1} \cdots Q_1 X$ with two rotations $Q_i := (G_{n-i}^{(y)} G_{n-i-1}^{(x)})$ for $i = 2, \dots, n-2$ shows indeed that $H_{\Lambda_0} Q^T = R$ must be upper triangular, except for the first two columns of R. These must be zero because of the orthogonality constraint with $\tilde{X}$, which completes the proof of point 3.

Finally, let us apply Q_1 as a similarity to the Hessenberg matrix H. Then we have the following pattern of nonzero elements:

$$Q_1 H Q_1^T = \left[\begin{array}{cccc|cccc} \times & \times & \dots & \times & \times & \times & \times & \times \\ \times & \times & \dots & \times & \times & \times & \times & \times \\ & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \times & \times & \times & \times & \times & \times \\ \hline & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times \\ & & & & \hat{h} & \hat{h} & \hat{h} & \times \end{array}\right], \quad Q_1 X = \left[\begin{array}{cc} \times & \times \\ \times & \times \\ \vdots & \vdots \\ \times & \times \\ \hline \hat{\times} & \hat{\times} \\ \hat{\times} & \hat{\times} \\ \hat{0} & \hat{\times} \\ 0 & 0 \end{array}\right] = 0,$$

which shows a 4×4 bulge at the bottom of the (transformed) Hessenberg matrix $Q_1 H Q_1^T$. But because of the equality

$$(Q_1 H Q_1^T)(Q_1 X) = (Q_1 X)\Lambda_0,$$

it follows that the submatrices marked with a $\hat{}$ are orthogonal:

$$\begin{bmatrix} \hat{h} & \hat{h} & \hat{h} \end{bmatrix} \begin{bmatrix} \hat{\times} & \hat{\times} \\ \hat{\times} & \hat{\times} \\ \hat{0} & \hat{\times} \end{bmatrix} = 0.$$

Therefore, the next matrix $Q_2 = G^{(y)}_{n-2}G^{(x)}_{n-3}$, which acts on the right factor, automatically annihilates the first two elements of the left factor as well:

$$\left(\begin{bmatrix} \hat{h} & \hat{h} & \hat{h} \end{bmatrix} Q_2^T \right) \left(Q_2 \begin{bmatrix} \hat{\times} & \hat{\times} \\ \hat{\times} & \hat{\times} \\ \hat{0} & \hat{\times} \end{bmatrix} \right) = \begin{bmatrix} 0 & 0 & \hat{h} \end{bmatrix} \begin{bmatrix} \hat{\times} & \hat{\times} \\ 0 & \hat{\times} \\ 0 & 0 \end{bmatrix} = 0. \tag{6.8}$$

It then follows that in $Q_2Q_1HQ_1^TQ_2^T$ the bulge has moved up one row and column. Repeating this argument for all transformations $Q_i = G^{(y)}_{n-i}G^{(x)}_{n-i-1}$ then completes the proof. □

The equivalence of points 1 and 2 is well known in the literature, but since we will now analyze backward error bounds in these transformations, it was important to give these relations in more detail. The transformation described in (6.8) will especially play an important role in this analysis.

6.1. Backward errors in the double *QR* step.

In this subsection, we give the double shift analogue of the discussion given in section 3.

THEOREM 6.2. *Let H be an unreduced Hessenberg matrix, and let λ_0 and $\bar{\lambda}_0$ be a pair of complex conjugate shifts. Let the sequence of Givens transformations $G^{(x)}_i$ and $G^{(y)}_i$ be constructed from the implicit double QR step, and let Q be the accumulated product of the corresponding exactly orthogonal transformations $\tilde{G}^{(x)}_i$ and $\tilde{G}^{(y)}_i$. Then in inexact arithmetic, the computed Hessenberg matrix $\tilde{H}$ satisfies*

$$Q(H + \Delta_H)Q^T = \tilde{H},$$

with

$$\|\Delta_H\|_F \le 4\gamma_{cn}\|H\|_F,$$

where c is a moderate constant of the order of 1*, provided the rotation parameters are computed via the standard construction.*

We will then say that the double QR step with shifts $\mathrm{sp}(\Lambda_0) = \{\lambda_0, \bar{\lambda}_0\}$ is "perfect," provided $\tilde{H}$ is Hessenberg, and moreover the leading two columns of $\tilde{H}$ satisfy, up to machine accuracy,

$$\begin{bmatrix} \tilde{h}_{1,1} & \tilde{h}_{1,2} \\ \tilde{h}_{2,1} & \tilde{h}_{2,2} \\ 0 & \tilde{h}_{3,2} \end{bmatrix} \approx \begin{bmatrix} S\Lambda_0 S \\ 0 \quad 0 \end{bmatrix},$$

which we can "absorb" in a forward error $\Delta_{\tilde{H}}$. This would imply that E_1 spans an invariant subspace of

$$\tilde{H} + \Delta_{\tilde{H}} = Q(H + \Delta_H)Q^T \tag{6.9}$$

corresponding to the spectrum of Λ_0 and that the span of $X := Q^TE_1$ is the *exact* invariant subspace of a perturbed Hessenberg matrix $H + \Delta_H$ corresponding to the spectrum of Λ_0. Notice that the use of the forward error $\Delta_{\tilde{H}}$ is needed for this interpretation. Usually, a tolerance τ is specified for the errors Δ_H and $\Delta_{\tilde{H}}$ that is of the order of $\epsilon_M \|H\|_F$ and compatible with the bound of Theorem 6.2.

This is again a backward error analysis, and it does not say anything about the forward errors, but it implies the following, where $E_1 := \left[\begin{array}{cc} \mathbf{e}_1 & \mathbf{e}_2 \end{array}\right]$.

Definition 6.3. *A double (backward) QR step with shifts λ_0 and $\bar{\lambda}_0$ is "perfect" if it corresponds to a perturbed Hessenberg matrix $H+\Delta_H$ with $\|\Delta_H\|_F \leq \tau$ for which the (backward) QR step satisfies* (6.9) exactly *and for which (Λ_0, QE_1) is an* exact *"spectral pair." Moreover, the property that λ_0 and $\bar{\lambda}_0$ are an exact eigenvalue pair of the transformed matrix $\tilde{H}$ is enforced if we set the leading 2×2 block and $\tilde{h}_{3,2}$ to the nearby values $S\Lambda_0 S$ and 0, respectively.*

Notice that here again we constrain ourselves to a Hessenberg backward error Δ_H, because we show how to construct such a perturbation.

6.2. Structured backward errors in the double *QR* step. We now derive bounds for the structured backward errors. For this, we assume that we have computed an approximate basis X for an invariant subspace of H with presumed spectrum $(\lambda_0, \bar{\lambda}_0)$:

$$HX - X\Lambda_0 = U, \quad X^T X = I_2, \quad \|U\|_F \ll 1.$$

Lemma 6.4. *The minimum Frobenius norm solution Δ_H of Hessenberg form*

$$\Delta_H = \left[\begin{array}{cccc} \delta h_{1,1} & \delta h_{2,1} & \cdots & \delta h_{1,n} \\ \delta h_{2,1} & \delta h_{2,2} & \cdots & \delta h_{2,n} \\ & \ddots & \cdots & \vdots \\ & & \delta h_{n,n-1} & \delta h_{n,n} \end{array}\right]$$

to the system

$$(H+\Delta_H)X - X\Lambda_0 = 0, \quad HX - X\Lambda_0 = U, \tag{6.10}$$

where $X^T X = I_2$, has Frobenius norm bounded by

$$\|\operatorname{diag}(\nu_1, \nu_2, \ldots, \nu_n)^{-1} U\|_F,$$

where

$$\nu_1 = 1, \quad \nu_i = \sigma_{min} X(i-1:n,:),\ i = 2, \ldots, n.$$

Proof. It follows from (6.10) that

$$\Delta_H X = -U,$$

which is a linear set of equations. It has a minimum Frobenius norm solution which we can solve row by row. Let row i of this equation be

$$\left[0, \ldots, 0, \underbrace{\boldsymbol{\delta h}_i^T}_{\min(n, n-i+2)} \right] X = -\mathbf{u}_i^T, \tag{6.11}$$

where $\mathbf{u}_i^T$ denotes the ith row of U, and

$$\boldsymbol{\delta h}_i^T = [\delta h_{i,i-1}, \delta h_{i,i}, \ldots, \delta h_{i,n}].$$

Since ν_i is defined as the smallest singular value of the submatrix $X(i-1:n,:)$ involved in (6.11), clearly the minimum norm solution for this row satisfies

$$\boldsymbol{\delta h}_i^T = -\mathbf{u}_i^T X(i-1:n,:)^+, \quad \|\boldsymbol{\delta h}_i\|_2 \leq \|\mathbf{u}_i\|_2/\sigma_{min}X(i-1:n,:),$$

where $X(i-1:n,:)^+$ denotes the pseudoinverse of that matrix and has 2-norm bounded by the inverse of its smallest singular value. Notice that for rows 1 and 2 the matrix X is complete and its smallest singular value is 1. Going over to the Frobenius norm of Δ_H then completes the proof. ◻

As was also shown in the single real shift case, we now give sufficient conditions for attaining a satisfactory result for the backward structured error.

THEOREM 6.5. *Let H be an unreduced Hessenberg matrix, and let us have an estimate of an invariant subspace X with constrained spectrum $(\lambda_0, \bar{\lambda}_0)$ satisfying*

$$U := (H-\lambda_0 I)X - X\Lambda_0, \quad X^T X = I_2,$$

where

$$\|\operatorname{diag}(\nu_1,\nu_2,\ldots,\nu_n)^{-1}U\|_F \leq \epsilon\|H\|_F, \tag{6.12}$$
$$\nu_1 = 1,\ \ \nu_i = \sigma_{min}X(i-1:n,:),\ i=2,\ldots,n,$$

and let the Givens rotations $G_i^{(x)}$ and $G_i^{(y)}$ be computed to annihilate the successive entries of x_{i+1} for $i=n-2,\ldots,1$ and y_{i+1} for $i=n-1,\ldots,2$ of the approximate invariant subspace basis vectors $\mathbf{x}$ and $\mathbf{y}$, and transforming them into $\mathbf{e}_1$ and $\mathbf{e}_2$. Then the product

$$Q := (\tilde{G}_1^{(y)}\cdots\tilde{G}_{n-1}^{(y)})(\tilde{G}_1^{(x)}\cdots\tilde{G}_{n-2}^{(x)})$$

of the corresponding exactly orthogonal Givens rotations yields a perfect shift implementation of the double QR step applied to H, in the sense that there exist backward perturbations Δ_H and Δ_X such that

$$Q(H+\Delta_H)Q^T = \tilde{H},\ Q(X+\Delta_X) = \left[\,\pm\mathbf{e}_1\ \pm\mathbf{e}_2\,\right],\ \|\Delta_H\|_F \leq c\epsilon_M\|H\|_F,\ \|\Delta_X\|_F \leq c\epsilon_M,$$

where c is a constant of the order of 1.

Proof. The proof is completely analogous to that for the single real shift case but relies this time on Lemma A.3. ◻

The key point in this theorem is again that we need an approximate basis X for an invariant subspace, with a sufficiently small residual, especially in the components where each trailing submatrix of X has small values ν_i. Such a basis can again be obtained using a diagonal scaling technique very similar to the one explained in section 5, except that now we apply a diagonal scaling,

$$D := \operatorname{diag}(1, d, d^2, \ldots, d^{n-2}, d^{n-2}), \tag{6.13}$$

where $d \geq 1$, that "balances" the row vectors of X without affecting too much the norm of H.

THEOREM 6.6. *Let H be an unreduced Hessenberg matrix, and let X be a basis of an approximate invariant subspace with constrained spectrum Λ_0. Then there always exists a scaling D of the form* (6.13) *such that the transformed pair $(H_D, X_D) := (DHD^{-1}, DX)$ satisfies the constraints*

$$\|H_D\|_F \leq d\|H\|_F, \quad d := \max(\max_{i\leq n-2}\{(n_i/\sigma)^{1/n-i-1}\}, 1),$$

where $n_i := \|[x_i, y_i]\|_2$ *and* $\sigma := \sigma_{min} \left[\begin{smallmatrix} x_{n-1} & y_{n-1} \\ 0 & y_n \end{smallmatrix}\right]$ *and such that the transformed matrix* X_D *has a bottom* 2×2 *block with* σ_{min} *larger than or equal to the row norms from* 1 *to* $n-2$*.*

Proof. The proof is very similar to that in section 5 by reasoning on the vector

$$\begin{bmatrix} n_1 & dn_2 & \dots & d^{n-3}n_{n-2} & d^{n-2}\sigma \end{bmatrix}$$

and using $\sigma \neq 0$. The fact that the last two rows are lumped together in a 2×2 block explains why the diagonal scaling D has its last two powers of d equal. □

It then follows from this scaling that the smallest singular value $\sigma_{min}^{(i)}$ of the submatrices $DX[i:n,:]$ are all bounded as follows:

$$\sigma_{min}^{(n-1)} \leq \sigma_{min}^{(i)} \leq \sqrt{n}\sigma_{min}^{(n-1)}, \quad \text{where} \quad \sigma_{min}^{(n-1)} = d^{n-2}\sigma.$$

The proof of this relies on the fact that the scaled vector norms $d^i n_i$ are all smaller than or equal to $\sigma_{min}^{(n-1)}$ and that when appending rows $DX[i:n-2,:]$ to $DX[n-1:n,:]$ to form $DX[i:n,:]$, the squared singular values all grow, but not more than the squared Frobenius norm of the added rows:

$$\sigma_{min}^2 DX[n-1:n,:] \leq \sigma_{min}^2 DX[i:n,:] \leq \|DX[i:n-2,:]\|_F^2 + \sigma_{min}^2 DX[n-1:n,:].$$

As a result we can hope again (as in Lemma 5.2) that computing the invariant subspace for this scaled matrix will go well and that transforming it back by inverting the scaling again will yield an invariant subspace for which the residual satisfies the desired bound

$$\|\operatorname{diag}(\nu_1, \nu_2, \dots, \nu_n)^{-1}U\|_F \leq \epsilon.$$

Guaranteeing such a result is now much harder, even under mild conditions. Therefore we will illustrate this scaling technique in the numerical section.

7. Numerical examples. In this section we illustrate our new perfect shift QR step implementation using the preliminary computation of an eigenvector or eigenspace of the eigenvalue(s) that need to be deflated. All computations were done in MATLAB (version 2014b) in double precision with a machine epsilon of $\epsilon_M \approx$ 2.22e-16. In all the examples, we used the normalized eigenvector $\mathbf{x}$ or the normalized eigenspace basis X to construct the Givens rotations. As a consequence, the deflation of the eigenvalues is ensured. What is *not* automatically guaranteed is that the transformed matrix H is again Hessenberg, and this is verified by looking at the relative norm $\frac{\|\mathtt{tril}(H,-2)\|_F}{\|H\|_F}$. To this end we define the functions $\mathbf{r}(H, \mu, \mathbf{x}) = (H - \mu I)\mathbf{x}$ and $b(H) = \|\mathtt{tril}(H, -2)\|_F$, i.e., the residual and the Frobenius norm of the part of the matrix H below the first subdiagonal.

We first show an example from a test case described in [6, Appendix B] arising in a problem of surface subdivision. The aim in this problem is to compute the Jordan chains at the eigenvalue $\lambda_0 = 0$ of a given matrix A.

Example 7.1. Let $A \in \mathbb{R}^{18 \times 18}$ be the second of the two matrices considered in [6, Appendix B]. The 2-norm of A is about 2, and its two smallest singular values, computed by MATLAB, are $\sigma_{17} = 9.16 \times 10^{-11}$ and $\sigma_{18} = 2.57 \times 10^{-16}$, while the smallest eigenvalue λ_{18} computed by `eig` of MATLAB has modulus equal to 3.3473×10^{-5}. This is to be expected since the condition number of that eigenvalue is $\kappa(\lambda_{18}) = 1.8088 \times 10^{11}$. Let H be the similar Hessenberg matrix computed via

TABLE 7.1
Results obtained by applying to H the eigenvector method and the balanced eigenvector method.

$\|\mathbf{r}(H,\tilde{\lambda},\tilde{\mathbf{x}})\|_2$	$b(\tilde{H})$	$\|\tilde{h}_{21}\|$	$\|\mathbf{r}(H,\hat{\lambda},\hat{\mathbf{x}})\|_2$	$b(\hat{H})$	$\|\hat{h}_{21}\|$
3.5335e-15	1.3459e-04	5.6362e-16	4.7075e-16	1.2739e-16	4.6847e-16

orthogonal transformation [5, pp. 378–379]. Let $\tilde{\mathbf{x}} \in \mathbb{R}^{18}$ be a normalized vector obtained by applying one step of inverse iteration with zero shift to H with a random initial guess, $\tilde{\lambda} = \tilde{\mathbf{x}}^T H \tilde{\mathbf{x}}$, and let $\hat{H}, \hat{\lambda}, \hat{\mathbf{x}}$ be, respectively, the matrix and the estimated eigenvalue and eigenvector obtained by applying the balancing eigenvector method.

The results are displayed in Table 7.1.

For this example, the value of d, the coefficient computed to construct D as in (5.1), is 16. Notice that the matrix computed by the eigenvector method is far from a Hessenberg matrix, while the part below the first subdiagonal of the matrix computed by the balanced eigenvector method is negligible and the entries $(1,1)$ and $(2,1)$ are of the order of the machine precision.

For the second example, we refer back to Example 2.2, and more precisely to the use of balancing in this example. It can be seen from this example that the eigenvector method (Eigv) did perform quite well, but that when applying balancing (Eigv_B) we managed to reduce the errors by an additional 9 to 10 digits!

We now give a set of examples with both real and complex conjugate eigenvalues.

Example 7.2. In this example, unsymmetric matrices from the University of Florida Sparse Matrix Collection [2] of order between 50 and 150 are considered. In particular, each considered matrix is first transformed into Hessenberg form H. Then its real Schur form is computed by the proposed procedure, $H = U_E R_E U_E^T$, starting from the eigenvalue decomposition computed by `eig` of MATLAB. The results are reported in Table 7.2. The name of the matrix, its dimension, and its Frobenius norm are displayed in columns 1, 2, and 3, respectively, while columns 4 and 5 show the norm of the part of the matrix below the first subdiagonal and the relative error of the Schur form computed by the proposed method.

We point out that only for the matrix `gent113` have we used the balancing eigenvector method, and d was chosen equal to 1.4, in order to avoid cancellation errors.

In the third example, we consider test matrices H from [7] and included in MATLAB, whose eigenvalues are explicitly known.

Example 7.3. In this example, we consider two matrices whose eigenvalues are explicitly known. In particular, the `clement` matrix of order n has eigenvalues $\lambda_i = -n + i,\ i = 1, 3, 5, \ldots, 2n-3, 2n-1$, while the `chow(1,0)` matrix has $p = \lfloor n/2 \rfloor$ eigenvalues equal to zero, and the rest of the eigenvalues are equal to $4\cos(\frac{k\pi}{n+2})^2, k = 1, 2, \ldots, n-p-1, n-p$.

The eigenvector method, and the balancing eigenvector method when needed, was applied to each matrix, for each eigenvalue $\lambda_i,\ i = 1, \ldots, n$, obtaining the matrices $\tilde{H}_i$.

The results are reported in Table 7.3. The size and the average of the relative norm of the part below the first subdiagonal of $\tilde{H}_i$ are displayed in columns 2 and 3, respectively. The average of the entries in position $(2,1)$ and differences of the entries in position $(1,1)$ of the computed matrices and the eigenvalues, are reported in columns 4 and 5, respectively. Finally, the average of the differences between the eigenvalues and the eigenvalues computed by using the function `eig` of MATLAB are

TABLE 7.2
Real Schur form computation with `Schur` *of MATLAB and with the proposed procedure.*

Matrix	n	$\|H\|_F$	$b(R_E)$	$\frac{\|HU_E-U_ER_E\|_F}{\|H\|_F}$
`ww_36_pmec_36`	66	4.0892e+03	2.1775e-16	1.0458e-15
`west0067`	67	1.3122e+01	5.1330e-16	1.4205e-15
`lesmis`	77	1.0923e+02	1.4534e-16	1.7730e-15
`steam3`	80	1.0633e+00	2.0434e-16	1.3713e-15
`cat_ears_2_1`	85	1.5937e+01	5.1767e-16	1.6393e-15
`d_dyn`	87	1.2456e+02	4.6675e-16	1.3426e-15
`dwt_87`	87	2.3259e+01	1.4021e-16	1.1961e-15
`cage6`	93	6.2516e+00	3.6287e-16	1.7929e-15
`tub100`	100	1.1856e+04	6.2818e-16	2.1834e-15
`olm100`	100	4.1167e+03	1.1197e-16	1.5951e-15
`rotor1`	100	9.1659e+04	3.1043e-15	1.9861e-15
`pivtol`	102	1.4867e+00	1.1929e-16	2.7175e-15
`ck104`	104	1.1068e+01	3.1347e-16	1.9016e-15
`gent113`	113	2.5593e+01	3.6680e-15	1.2587e-15
`gre_115`	115	7.2468e+00	4.9680e-16	1.9885e-15
`rajat11`	135	2.3711e+01	1.4639e-16	1.9667e-15
`rw136`	136	7.2356e+00	5.2296e-16	2.2214e-15
`impcol_c`	137	1.4521e+02	3.3655e-16	1.9118e-15
`lop163`	163	8.7365e+00	5.3851e-16	2.5749e-15
`rajat14`	180	1.0789e+00	9.7592e-17	2.4943e-16

TABLE 7.3
Results obtained by applying one step of deflation to matrices from the "MATLAB gallery."

Matrix	n	$\frac{\sum_{i=1}^n b(\tilde{H}_i)}{n\|H\|_2}$	$\frac{\sum_{i=1}^n \|\bar{h}_{21}\|}{n\|H\|_2}$	$\frac{\sum_{i=1}^n \|\bar{h}_{11}-\lambda_i\|}{n\|H\|_2}$	$\frac{\sum_{i=1}^n \|\lambda_i^M-\lambda_i\|}{n\|H\|_2}$
`clement`	100	2.7363e-16	1.5060e-18	3.3710e-16	4.2813e-06
`chow(1,0)`	100	7.0223e-18	1.7738e-17	6.8588e-17	4.5622e-03

displayed in column 6. We observe that `eig` computes the small eigenvalues in an inaccurate way. In particular, the eigenvalues corresponding to the zero eigenvalues of `chow(1,0)` are all complex.

Remark 7.1. In all of the examples, the final results obtained by our new method were computed with a satisfactory error. Moreover, the balancing techniques were almost never needed. One drawback is of course that we need to have a good starting value for the "perfect shift." We can also expect difficulties when the approximate eigenvector has small elements in the last two entries, x_{n-1} and x_n, while the previous entry, x_{n-2}, is very large. The scaling technique will then produce a very large value d, which may cause underflows in the computations.

8. Conclusions.

In this paper we revisited the problem of performing a QR step with a so-called perfect shift, which is the eigenvalue λ_0 that we want to deflate. We gave a new procedure that is based on the preliminary computation of the eigenvector $\mathbf{x}$ corresponding to that shift, but with the requirement that it be computed to a certain relative precision. This condition involves the residual $r = (H - \lambda_0 I)\mathbf{x}$ and the norms of the trailing subvectors of $\mathbf{x}$. We also gave a scaling technique that yields, under some mild assumptions, an eigenvalue/eigenvector pair $(\lambda_0, \mathbf{x})$ with a relative precision that meets this perfect shift condition. We also showed how to extend these ideas to a real double shift corresponding to a pair of complex conjugate eigenvalues. For this case it is less obvious how to perform a scaling that will allow us to get an eigenspace X with an appropriate residual that achieves a relatively small backward error. The numerical experiments indicate that the perfect shift technique works very well.

We did not cover in this paper the corresponding problem of computing a perfect QZ step of a (regular) generalized eigenvalue problem $A - \lambda B$ in an unreduced Hessenberg–Schur form, but we conjecture that results along the same lines should be possible, provided one computes the eigenvector corresponding to that shift, with a suitable precision. But this is left for future research.

Appendix A.

Proof of Theorem 3.1. We first recall the analysis of the QR decomposition given in [8, Theorem 18.9]. We modified it here to apply it to an RQ decomposition of an $n \times n$ Hessenberg matrix, rather than a QR decomposition of a dense $m \times n$ matrix.

LEMMA A.1. *Consider the sequence of transformations $H_{k+1} = H_k G_{n-k}^T$ for $k = 1 \ldots, n-1$, where $H_1 = H \in \mathbb{R}^{n\times n}$, and where each Givens rotation G_i is computed to annihilate the $(i+1, i)$ element of an unreduced Hessenberg matrix H. Then the computed (triangular) matrix $\hat{H}_n$ satisfies*

$$\hat{H}_n = (H + \Delta_H)Q^T, \quad Q = \tilde{G}_1 \tilde{G}_2 \cdots \tilde{G}_{n-1}, \tag{A.1}$$

where each $\tilde{G}_i$ is the exactly orthogonal Givens rotation that annihilates exactly the position $(i+1, i)$ in the current matrix of the computed sequence $\tilde{H}_1, \ldots, \tilde{H}_n$, where Δ_H satisfies the normwise bound

$$\|\Delta_H\|_F \le \gamma_{cn} \|H\|_F, \quad \gamma_{cn} := \frac{cnu}{1 - cnu}, \tag{A.2}$$

and where c is a moderate number of order 1.

If we apply this to $H_1 = H - \lambda_0 I_n$ and multiply (A.1) with Q, we obtain

$$H - \lambda_0 I_n + \Delta_H = \hat{R}Q, \quad \|\Delta_H\|_F \le \gamma_{cn} \|H - \lambda_0 I_n\|_F,$$

where $\hat{R}$ is the computed upper triangular matrix. Since $(H - \lambda_0 I_n)$ and $(\hat{R}Q)$ are Hessenberg, it then follows that Δ_H is Hessenberg as well.

Let us now consider the second part of the (backward) explicit QR step, which consists of the multiplication $Q\hat{R}$. There also we can apply the results of Lemma A.1 since we just multiply a matrix with a sequence of $(n-1)$ Givens rotations. We do not use these Givens transformations to assign a zero, but the bounds still apply. The starting matrix is now $\hat{R}$ and one computes $Q\hat{R}$. Thus, in inexact arithmetic this yields $Q(\hat{R} + \Delta R)$, which is equal to $\tilde{H} - \lambda_0 I_n$ since by shifting this product back by $\lambda_0 I_n$, one gets $\tilde{H}$. We thus have

$$\tilde{H} - \lambda_0 I_n = Q(\hat{R} + \Delta R).$$

If we now define $\Delta_{\tilde{H}} := -Q\Delta R$, we finally have

$$\tilde{H} - \lambda_0 I_n + \Delta_{\tilde{H}} = Q\hat{R},$$

which implies that $\Delta_{\tilde{H}}$ must also be Hessenberg. □

Proof of Remark 3.1. For the implicit QR step we note that the Givens rotations are always performed either on the rows or on the columns of a matrix $H_{\ell r} := Q_\ell H Q_r^T$ that is related to the original Hessenberg matrix H via orthogonal left and right transformations Q_ℓ and Q_r. Each Givens rotation results in an additive error that is bounded by $\gamma_c \|H_{\ell r}\|_F = \gamma_c \|H\|_F$ that can be mapped to the original matrix H without changing its norm since Q_ℓ and Q_r are orthogonal. The results then easily follow. □

Proof of Theorem 6.2. For the double implicit QR step the proof is completely analogous to that of the single implicit QR step. The only difference is that the number of Givens rotations is essentially doubled, and so is the bound on the backward error.

LEMMA A.2. *Let $\mathbf{h}, \mathbf{v} \in \mathbb{R}^2$ be two vectors that are nearly orthogonal, i.e.,*

$$\mathbf{h}^T \mathbf{v} = \eta, \quad |\eta| \ll \max(\|\mathbf{h}\|_2, \|\mathbf{v}\|_2).$$

Then the smallest perturbation $\delta\mathbf{h}$ measured in the 2-norm to make them orthogonal is given by

$$\delta\mathbf{h} = -\eta \mathbf{v} / \|\mathbf{v}\|_2^2$$

and has 2-norm $|\eta|/\|\mathbf{v}\|_2$. Moreover, if $\eta \leq \hat{\epsilon}\|\mathbf{v}\|_2$ and we construct a Givens transformation G such that

$$G\mathbf{v} = \begin{bmatrix} \|\mathbf{v}\|_2 \\ 0 \end{bmatrix} = \|\mathbf{v}\|_2 \mathbf{e}_1, \tag{A.3}$$

it also follows that the first element of the transformed vector $G\mathbf{h}$ is $\hat{\epsilon}$-small and can be neglected if this is below the tolerance level.

Proof. Let $\mathbf{v}^T = [v_1, v_2]$. Then the construction of the Givens rotation in (A.3) just uses $d = \sqrt{v_1^2 + v_2^2}, c = v_1/d, s = -v_2/d$. Moreover, the first element of $G\mathbf{h}$ is $\mathbf{e}_1^T G\mathbf{h} = \begin{bmatrix} c & -s \end{bmatrix} \mathbf{h} = \eta/\|\mathbf{v}\|_2$ and that is precisely supposed to be $\hat{\epsilon}$-small. The stability of the Givens rotation implementation then guarantees that in inexact arithmetic, this is also negligible with respect to $\|\mathbf{h}\|_2$. □

LEMMA A.3. *Let*

$$\mathbf{h} := \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} \in \mathbb{R}^3 \quad \text{and} \quad V := \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ 0 & y_3 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

be nearly orthogonal, i.e.,

$$\mathbf{h}^T V = \eta := [\eta_1, \eta_2], \quad \|\eta\|_2 \ll \max(\|\mathbf{h}\|_2, \|V\|_2).$$

Then the smallest perturbation $\delta\mathbf{h}$ measured in the 2-norm to make them orthogonal is given by

$$\delta\mathbf{h} = -\eta V^+, \quad V^+ := (V^T V)^{-1} V^T$$

and has 2-norm smaller than $\|\eta\|_2 \|V^+\|_2$, where V^+ is the pseudoinverse of V. Moreover, if $\|\eta\|_2 \leq \hat{\epsilon}/\|V^+\|_2$ and we construct two Givens transformations $G_1^{(x)}$ and $G_2^{(y)}$ such that

$$G_2^{(y)} G_1^{(x)} V = \begin{bmatrix} \hat{x}_1 & \hat{y}_1 \\ 0 & \hat{y}_2 \\ 0 & 0 \end{bmatrix}, \tag{A.4}$$

it also follows that the first two elements of the transformed vector $\hat{\mathbf{h}} := G_2^{(y)} G_1^{(x)} \mathbf{h}$ *are* $\hat{\epsilon}$ *small and can be neglected if this is below the tolerance level.*

Proof. The subvectors $\left[\begin{smallmatrix} h_1 \\ h_2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$ are obviously nearly orthogonal, and we can apply the previous lemma to show that the first element of the transformed vector $\hat{\mathbf{h}}$ is $\hat{\epsilon}$ small. After that, we repeat the same argument on the (transformed) subvectors $\left[\begin{smallmatrix} \hat{h}_2 \\ h_3 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} \hat{y}_2 \\ y_3 \end{smallmatrix}\right]$. Again we rely on the stability of Givens rotation implementations to guarantee that in inexact arithmetic, these elements also are negligible with respect to $\|\mathbf{h}\|_2$. □

Acknowledgment. The authors wish to thank the anonymous reviewers and the handling editor, Zlatko Drmac, for their constructive remarks that helped improving the presentation of the results. The first author would like to acknowledge the hospitality of the Université Catholique de Louvain in Louvain-la-Neuve, Belgium, where part of this work took place.

REFERENCES

[1] S. L. Campbell and C. D. Meyer, *Continuity properties of the Drazin pseudoinverse*, Linear Algebra Appl., 10 (1975), pp. 77–83.

[2] T. A. Davis and Y. Hu, *The University of Florida Sparse Matrix Collection*, ACM Trans. Math. Softw., 38 (2011), pp. 1:1–1:25.

[3] I. S. Dhillon and A. N. Malyshev, *Inner deflation for symmetric tridiagonal matrices*, Linear Algebra Appl., 358 (2003), pp. 139–144.

[4] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.

[6] N. Guglielmi, M. Overton, and G. W. Stewart, *An efficient algorithm for computing the generalized null space decomposition*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 38–54, https://doi.org/10.1137/140956737.

[7] N. J. Higham, *The Matrix Computation Toolbox*, http://www.ma.man.ac.uk/~higham/mctoolbox.

[8] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 1st ed., SIAM, Philadelphia, 1996.

[9] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[10] P. Kunkel and V. Mehrmann, *Differential-algebraic Equations: Analysis and Numerical Solution*, European Math. Soc., 2006.

[11] P. Lancaster and M. Tismenetsky, *Theory of Matrices*, Academic Press, Orlando, FL, 1985.

[12] N. Mastronardi and P. Van Dooren, *Computing the Jordan structure of an eigenvalue*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 949–966, https://doi.org/10.1137/16M1083098.

[13] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Springer, Berlin, 1991.

[14] G. S. Miminis and C. C. Paige, *Implicit shifting in the QR and related algorithms*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 385–400, https://doi.org/10.1137/0612028.

[15] B. N. Parlett and I. S. Dhillon, *Relatively robust representations of symmetric tridiagonals*, Linear Algebra Appl., 309 (2000), pp. 121–151.

[16] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385, https://doi.org/10.1137/0613025.

[17] G. W. Stewart, *Matrix Algorithms. Volume* II: *Eigensystems*, SIAM, Philadelphia, 2001, https://doi.org/10.1137/1.9780898718058.

[18] F. Tisseur, *Backward Stability of the QR Algorithm*, Tech. Report UMR 5585, Université de Lyon Saint-Etienne, 1996.

[19] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.

[20] D. S. Watkins, *The transmission of shifts and shift blurring in the QR algorithm*, Linear Algebra Appl., 241–243 (1996), pp. 877–896.

[21] D. S. Watkins, *The Matrix Eigenvalue Problem*, SIAM, Philadelphia, 2007, https://doi.org/10.1137/1.9780898717808.