Electronic Transactions on Numerical Analysis.
Volume 20, pp. 1-26, 2005.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

FAST GIVENS TRANSFORMATION FOR QUATERNION VALUED MATRICES APPLIED TO HESSENBERG REDUCTIONS*

DRAHOSLAVA JANOVSKÁ† AND GERHARD OPFER‡

Abstract. In a previous paper we investigated Givens transformations applied to quaternion valued matrices. Since arithmetic operations with quaternions are very costly it is desirable to reduce the number of arithmetic operations with quaternions. We show that the Fast Givens transformation, known for the real case, can also be defined for quaternion valued matrices, and we apply this technique to the reduction of an arbitrary quaternion valued matrix to upper Hessenberg form and also include a numerical example. We offer two algorithms. One is based on the classical real case using dynamically two transformation matrices, while the other is based on four transformation matrices where in each step that matrix is selected that has the smallest condition number. For the first algorithm we show that the essential information (namely the two numbers s and c which define the Givens transformation) can be stored in only one variable. This is apparently even new for the real case. We include, necessarily, some investigations on the determination of the relevant condition numbers. We show that in general the application of the Fast Givens transformation in the quaternion case is not as favorable as in the real case with respect to (relative) savings in arithmetic operations. We begin with some introduction into the field of quaternions. In the end in an appendix we present some results concerning the computation of roots of quaternions which in some cases are needed.

Key words. Fast Givens rotation, quaternions, quaternion valued matrices, Hessenberg form for quaternion valued matrices, roots of quaternions.

AMS subject classifications. 11R52, 12E15, 15A66, 65F30, 70Exx

1. Basic properties and definitions for quaternions. We start with some information on the algebra of quaternions. There are more details in our previous paper, Janovská and Opfer [10]. General information are contained in a book by Kuipers [13], results concerning matrices with quaternion elements are surveyed by Zhang [21]. Applications to quantum mechanics are treated by Dongarra et al., [4],[5], and application in chemistry are given by Rösch[15].

Let $m, n \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers starting with one. We denote by $\mathbb{H} = \mathbb{R}^4$ the skew field of quaternions. Let $a = (a_1, a_2, a_3, a_4),\ b = (b_1, b_2, b_3, b_4) \in \mathbb{H}$. Then, addition is defined elementwise and multiplication is governed by the following rule:

$$ab := (a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4, a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3, \\ a_1b_3 - a_2b_4 + a_3b_1 + a_4b_2, a_1b_4 + a_2b_3 - a_3b_2 + a_4b_1). \tag{1.1}$$

We see, that 16 real multiplications and 12 real additions are needed to compute the product ab, altogether 28 floating point operations (*flops*). The first component a_1 of $a = (a_1, a_2, a_3, a_4) \in \mathbb{H}$ is called the *real part* of a and denoted by $\Re a$. The second component a_2 is called the *imaginary part* of a and denoted by $\Im a$. A quaternion $a = (a_1, 0, 0, 0)$ will be identified with $a_1 \in \mathbb{R}$ and $a = (a_1, a_2, 0, 0)$ will be identified with $a_1 + \mathrm{i}a_2 \in \mathbb{C}$. The zero element $(0, 0, 0, 0) \in \mathbb{H}$ and the unit element $(1, 0, 0, 0) \in \mathbb{H}$ will be abbreviated by $0, 1$, respectively. Let $a = (a_1, a_2, a_3, a_4) \in \mathbb{H}$. The *conjugate* of a, denoted by $\overline{a}$, will be defined by

$$\overline{a} := (a_1, -a_2, -a_3, -a_4).$$

The *absolute value* of a, denoted by $|a|$, will be defined by

*Received January 19, 2005. Accepted for publication April 4, 2005. Recommended by L. Reichel.

†Institute of Chemical Technology, Prague, Department of Mathematics, Technická 5 166 28 Prague 6, Czech Republic (janovskd@vscht.cz).

‡University of Hamburg, MIN Faculty, Bundesstraße 55, 20146 Hamburg, Germany (opfer@math.uni-hamburg.de)

$$|a| := \sqrt{a_1^2 + a_2^2 + a_3^2 + a_4^2}.$$

There are the following important rules:

$$\begin{aligned}
\Re(ab) &= \Re(ba),\\
|ab| &= |ba| = |a||b|,\\
|a|^2 &= a\,\overline{a} = \overline{a}\,a,\\
\overline{a\,b} &= \overline{b}\,\overline{a},\\
a^{-1} &= \frac{\overline{a}}{|a|^2},\ a \neq 0,\\
(a\,b)^{-1} &= b^{-1}\,a^{-1},\ a, b \neq 0.
\end{aligned}$$

We denote by $\mathbb{H}^n$ the normed vector space of n-vectors formed by quaternions, where the norm of $\mathbf{x} := (x_1, x_2, \ldots, x_n) \in \mathbb{H}^n$ will be defined by

$$||\mathbf{x}|| := \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2},$$

and by $\mathbb{H}^{m\times n}$ the set of all $(m \times n)$-matrices with elements from $\mathbb{H}$. We note here, that these matrices act as *linear mappings* $\ell : \mathbb{H}^n \to \mathbb{H}^m$ only in the following sense:

$$\begin{aligned}
\ell(\mathbf{x} + \mathbf{y}) &= \ell(\mathbf{x}) + \ell(\mathbf{y}), \quad \mathbf{x},\ \mathbf{y} \in \mathbb{H}^n,\\
\ell(\mathbf{x}\alpha) &= \ell(\mathbf{x})\alpha, \quad \mathbf{x} \in \mathbb{H}^n,\ \alpha \in \mathbb{H}.
\end{aligned}$$

The converse is also true: A linear mapping ℓ defined by the above two properties is always represented by a matrix. This follows from standard arguments.

Let $\mathbf{A} \in \mathbb{H}^{m\times n}$. By $\mathbf{A}^{\mathrm{T}} \in \mathbb{H}^{n\times m}$ we understand the *transposed matrix* of $\mathbf{A}$ where the rows and columns are exchanged. By $\overline{\mathbf{A}} \in \mathbb{H}^{m\times n}$ we understand the matrix which is formed by conjugation of all its elements. Finally,

$$\mathbf{A}^* := (\overline{\mathbf{A}})^{\mathrm{T}} = \overline{\mathbf{A}^{\mathrm{T}}}.$$

In case $\mathbf{A}^* = \mathbf{A}$, we call $\mathbf{A}$ *Hermitean*. The zero element of $\mathbb{H}^n$ and of $\mathbb{H}^{m\times n}$ will be denoted by $\mathbf{0}$. From the context it will become clear which zero element is meant. Elements in $\mathbb{H}^n$ will be denoted by boldface Latin lower case letters, where for matrices in $\mathbb{H}^{m\times n}$ normally boldface Latin capital letters are used. A matrix $\mathbf{A} \in \mathbb{H}^{n\times n}$ will be called *unitary* if $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Unitary matrices $\mathbf{A}$ are characterized by $||\mathbf{A}\mathbf{x}|| = ||\mathbf{x}||$ for all $\mathbf{x} \in \mathbb{H}^n$. Eigenvalue problems for $\mathbf{A} \in \mathbb{H}^{n\times n}$ have to be posed in the form

$$\mathbf{A}\mathbf{x} = \mathbf{x}\lambda \tag{1.2}$$

and similar matrices have the same set of eigenvalues. The set of eigenvalues is in general not finite. If λ is an eigenvalue, the whole *equivalence class*

$$[\lambda] := \left\{\sigma \in \mathbb{H} : \sigma = h\lambda h^{-1} \text{ for all } h \in \mathbb{H}\backslash\{0\}\right\}$$

consists of eigenvalues. The number of different equivalence classes is, however, at most n.

LEMMA 1.1. *Two quaternions λ_1 and λ_2 are members of the same equivalence class if and only if $|\lambda_1| = |\lambda_2|$ and $\Re\lambda_1 = \Re\lambda_2$. As a consequence, two different complex numbers are equivalent if and only if they are conjugate two each other. Two real numbers are equivalent if and only if they coincide.*

Proof. Cf. Janovská and Opfer[10]. ☐

ETNA
Kent State University
etna@mcs.kent.edu

This lemma implies that in any equivalence class $[q]$ of quaternions there is exactly one complex quaternion $\tilde{q}$ with $\Re\tilde{q} \geq 0$. This will be called the *complex representative* of $[q]$. If $q = (q_1, q_2, q_3, q_4) \in [q]$, then $\tilde{q} = (q_1, \sqrt{q_2^2 + q_3^2 + q_4^2}, 0, 0)$ is the complex representative of $[q]$.

We should note here, that Hermitean matrices have only real eigenvalues and that all eigenvalues λ of unitary matrices obey $|\lambda| = 1$. In the later sections 4.1, 4.2 we will make some investigations on *condition numbers* of certain quaternion valued matrices. Since condition numbers for quaternion valued matrices $\mathbf{M}$ are apparently not defined in the literature we use the following definition.

DEFINITION 1.2. Let $\mathbf{M} \in \mathbb{H}^{n\times n}$ be nonsingular. We define the *condition number* of $\mathbf{M}$, denoted by $\mathrm{cond}(\mathbf{M})$ by

$$\mathrm{cond}(\mathbf{M}) := \sqrt{\lambda_{\max}(\mathbf{M}\mathbf{M}^*)/\lambda_{\min}(\mathbf{M}\mathbf{M}^*)}$$

where $\lambda_{\min}, \lambda_{\max}$ denote the smallest, the largest eigenvalue of $\mathbf{M}\mathbf{M}^*$, respectively.

This definition makes sense since all matrices of the type $\mathbf{M}\mathbf{M}^*$ have (only) n nonnegative eigenvalues λ and in case $\mathbf{M}$ is nonsingular these eigenvalues are positive. The definition implies $\mathrm{cond}(\mathbf{M}) \geq 1$ and $\mathrm{cond}(\mathbf{M}) = 1$ if and only if all eigenvalues of $\mathbf{M}\mathbf{M}^*$ are positive and identical.

In the above definition it is required to compute the eigenvalues of a quaternion valued, Hermitean matrix. For the (2×2) case we shall use the following lemma.

LEMMA 1.3. *Let $\mathbf{C} \in \mathbb{H}^{2\times 2}$ be a Hermitean matrix. Then it has necessarily the form*

$$\mathbf{C} = \begin{pmatrix} a & h \\ \overline{h} & b \end{pmatrix}, \quad a, b \in \mathbb{R},\ h \in \mathbb{H}$$

and the eigenvalues λ are real and obey the following quadratic equation with real coefficients:

$$p(\lambda) := \lambda^2 - (a+b)\lambda + ab - |h|^2 = 0.$$

The two solutions of this equation are

$$\lambda_{1,2} := \frac{a+b}{2} \pm \sqrt{\left(\frac{a-b}{2}\right)^2 + |h|^2}.$$

Proof. The given form of $\mathbf{C}$ and the fact that the eigenvalues are real are obvious. The solutions of the eigenvalue equation $\mathbf{C}\mathbf{x} = \mathbf{x}\lambda$ imply the given quadratic equation and its solutions. □

2. Flop counts for quaternion arithmetic. In connection with quaternion arithmetic not only quaternion×quaternion will be considered but many special cases for which the flop counts are more favorable. We repeat that *flop* stands for *floating point operation* which means an algebraic operation of the type addition, subtraction, multiplication, division of real numbers in floating point representation. Thus, a multiplication of two complex numbers in the classical form $(x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) = x_1x_2 - y_1y_2 + \mathrm{i}(y_1x_2 + x_1y_2)$ needs apparently 4 multiplications and 2 additions, together 6 flops. Actually, it is possible to use only three multiplications, but then, the number of additions/subtractions is increased. As

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 2.1
q-flop counts for various operations with quaternions

Operation	additions	multiplications	flops	q-flops
$p+q$	4	0	4	$1/7$
pq	12	16	28	1
rq	0	4	4	$1/7$
$\|q\|^2$	3	4	7	$1/4$
q^{-1}	3	8	11	$11/28$
pq^{-1}	15	24	39	$1+11/28$
cq	4	8	12	$3/7$
$\mathbf{a}^{\mathrm{T}}\mathbf{b}$	$16n-4$	$16n$	$32n-4$	$n+\frac{n-1}{7}$
$\mathbf{A}\mathbf{a}$	$m(16n-4)$	$16mn$	$m(32n-4)$	$m(n+\frac{n-1}{7})$
$\mathbf{A}+\mathbf{B}$	$4mn$	0	$4mn$	$\frac{mn}{7}$
$\mathbf{A}^{\mathrm{T}}\mathbf{B}$	$n^2(16m-4)$	$16n^2m$	$n^2(32m-4)$	$n^2(m+\frac{m-1}{7})$

already mentioned, see formula (1.1), the product of two quaternions needs 12 additions and 16 multiplications, altogether 28 flops.

DEFINITION 2.1. We will call the amount of 28 flops one *q-flop*. One q-flop is the algebraic work needed to multiply two quaternions.

We will summarize various flop counts in Table 2.1.

In the table we used the following notation:
- r for real numbers, • c for complex numbers, • p, q for quaternions,
- $\mathbf{a}, \mathbf{b}$ for quaternion valued column vectors with n components,
- $\mathbf{A}, \mathbf{B}$ for quaternion valued matrices with m rows and n columns.

3. Givens transformations for quaternions. We turn now to 2×2 Givens transformations. Let $\mathbf{G}$ be a matrix of the form

$$\mathbf{G} := \begin{pmatrix} \overline{c} & s \\ -\overline{s} & c \end{pmatrix}, \quad c, s \in \mathbb{H}. \tag{3.1}$$

Such a matrix $\mathbf{G}$ is unitary if and only if

$$\text{(a)} \quad cs = sc, \quad \text{(b)} \quad |c|^2 + |s|^2 = 1. \tag{3.2}$$

Let $\mathbf{x} = (x_1, x_2)^{\mathrm{T}} \in \mathbb{H}^2\backslash\{\mathbf{0}\}$. A matrix $\mathbf{G}$ of the form (3.1) will be called a *Givens transformation with respect to* $\mathbf{x}$ if it is unitary and if it has the property

$$\mathbf{G}^*\mathbf{x} =: \mathbf{u} = w\mathbf{e}_1, \quad \text{where} \quad \mathbf{e}_1 = (1,0)^{\mathrm{T}} \in \mathbb{H}^2, \quad w \in \mathbb{H}\backslash\{0\}. \tag{3.3}$$

In a recent paper we gave a characterization of Givens transformations.

THEOREM 3.1. *Let* $\mathbf{x} = (x_1, x_2)^{\mathrm{T}} \in \mathbb{H}^2$ *be given with*
(a) $x_1x_2 \neq 0$. *Then,* $\mathbf{G}$ *as defined in (3.1) is a Givens transformation with respect to* $\mathbf{x}$ *if and only if*

$$s = -\sigma\frac{\overline{x_2}}{||\mathbf{x}||}, \quad c = \sigma\frac{\overline{x_1}}{||\mathbf{x}||}, \quad \textit{where} \tag{3.4}$$

ETNA
Kent State University
etna@mcs.kent.edu

$$\sigma \in \Sigma := \left\{ \sigma \in \mathbb{H} : \sigma = \frac{\alpha x_1 + \beta x_2}{|\alpha x_1 + \beta x_2|},\ \alpha, \beta \in \mathbb{R},\ |\alpha| + |\beta| > 0 \right\}, \tag{3.5}$$

provided that the components x_1, x_2 are linearly independent over $\mathbb{R}$. In case there is a real constant $\alpha \neq 0$ such that $x_1 = \alpha x_2$, then, $\sigma \in \mathbb{H}$ is arbitrary with $|\sigma| = 1$. In addition, we have $w = \sigma||\mathbf{x}||$.
(b) *Let $x_1 x_2 = 0$, but $\mathbf{x} \neq \mathbf{0}$. Then, the general solution of (3.3) is:*

$$\begin{cases} s = -\sigma \dfrac{\overline{x_2}}{|x_2|},\ c = 0 & \text{if } x_1 = 0, \\ c = \quad \sigma \dfrac{\overline{x_1}}{|x_1|},\ s = 0 & \text{if } x_2 = 0, \end{cases} \qquad w = \sigma||\mathbf{x}||, \ \text{where } \sigma \in \mathbb{H} \text{ with } |\sigma| = 1. \tag{3.6}$$

Proof. Cf. Janovská and Opfer[10]. ∎

This theorem roughly says, that different from the real and complex case we cannot freely choose the parameter σ with $|\sigma| = 1$, but we have to follow the rule given in (3.5). In the above theorem we have not considered the case $\mathbf{x} = \mathbf{0}$. In this case, let $\mathbf{G} = \mathbf{I}$. There is another advantage of this conventional method. In the real case it is known, Stewart[17], that both constants c, s could be combined into one constant from which c, s then can be recovered. This is also valid in the quaternion case if we assume that c or s is real. But this is not a restriction. A look at formulas (3.4), (3.5) shows that it is always possible to choose either c or s not only real but nonnegative. So we can define

$$\mu_c := \frac{s}{1+c} \text{ for } c \geq 0 \Rightarrow c = \frac{1 - |\mu_c|^2}{1 + |\mu_c|^2},\ s = (1+c)\mu_c = \frac{2\mu_c}{1+|\mu_c|^2}, \tag{3.7}$$

$$\mu_s := \frac{c}{1+s} \text{ for } s \geq 0 \Rightarrow s = \frac{1 - |\mu_s|^2}{1 + |\mu_s|^2},\ c = (1+s)\mu_s = \frac{2\mu_s}{1+|\mu_s|^2}. \tag{3.8}$$

In terms of $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ with $x_1 x_2 \neq 0$, we have

$$\mu_c = -\frac{x_1 \overline{x_2}}{|x_1|(|x_1| + ||\mathbf{x}||)} \ (\sigma := \frac{x_1}{|x_1|}), \quad \mu_s = -\frac{x_2 \overline{x_1}}{|x_2|(|x_2| + ||\mathbf{x}||)} \ (\sigma := -\frac{x_2}{|x_2|}). \tag{3.9}$$

For the special cases $x_1 = 0$ or $x_2 = 0$ but $\mathbf{x} \neq \mathbf{0}$ we obtain

$$\mu_c = 0 \text{ for } x_2 = 0 \Rightarrow c = 1, s = 0; \quad \mu_s = 0 \text{ for } x_1 = 0 \Rightarrow c = 0, s = 1.$$

It should be noted that cancellation in the numerator of the formulas for c and s cannot take place if we choose formula (3.7) only in case $c^2 \geq 0.5 (\Leftrightarrow |s|^2 \leq 0.5 \Leftrightarrow |x_1| \geq |x_2|)$ which is equivalent to $|\mu_c|^2 \leq 3 - 2\sqrt{2} \approx 0.1716$. If $c^2 < 0.5$ we choose the other formula. Since μ_c, μ_s are undistinguishable by the sizes of their absolute values, the idea of Stewart[17] was to store a new constant, namely,

$$\mu := \begin{cases} \mu_c & \text{if } |x_1| \geq |x_2|, \\ \mu_s^{-1} & \text{if } 0 < |x_1| < |x_2|, \\ 1 & \text{if } x_1 = 0. \end{cases} \tag{3.10}$$

Now, if $\mu < 1$, then we can use the recovery formulas given in (3.7), with $\mu_c := \mu$. If $\mu > 1$ we use (3.8), however, μ_s has to be replaced by $\mu_s := \mu^{-1}$. If $\mu = 1$, we have $c = 0, s = 1$. Actually, Stewart's[17] definition was a little different. The advantage of the given formulas is, that no square roots are needed in the recovery process.

Givens transformation usually requires many evaluations of $\mathbf{y} := \mathbf{G}^*\mathbf{z}$ for arbitrary $\mathbf{z}$ and fixed $\mathbf{G}$. Let $\mathbf{y} := (y_1, y_2)^{\mathrm{T}}, \mathbf{z} := (z_1, z_2)^{\mathrm{T}}$. Then,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := \mathbf{G}^*\mathbf{z} = \begin{pmatrix} cz_1 - sz_2 \\ \overline{s}z_1 + \overline{c}z_2 \end{pmatrix}. \tag{3.11}$$

By rearranging and using (3.7), (3.8) we can write this as follows:

$$\begin{pmatrix} y_2 \\ y_1 \end{pmatrix} = \begin{pmatrix} \overline{s}z_1 + cz_2 \\ -\mu_c(z_2 + y_2) + z_1 \end{pmatrix} \text{ for } c \geq 0, \tag{3.12}$$

$$\begin{pmatrix} y_2 \\ y_1 \end{pmatrix} = \begin{pmatrix} sz_1 + \overline{c}z_2 \\ \mu_s(z_1 + y_2) - z_2 \end{pmatrix} \quad \text{for } s \geq 0. \tag{3.13}$$

Let c or s be real. Then, formulas (3.11), (3.12), (3.13) all need (2+4/7) q-flops. However, one multiplication real$\times$quaternion in (3.11) is replaced with one addition of two quaternions in (3.12), (3.13).

4. Fast Givens transformations for quaternions. Let $n \in \mathbb{N}$ be given with $n \geq 2$ and let $\mathbf{G} \in \mathbb{H}^{2\times 2}$ be of the form already defined in (3.1) and let $\mathbf{G}$ be unitary, i. e. (3.2) is valid. With given $2 \leq j_0 < k_0 \leq n, \mathbf{G}$, we define

$$\mathbf{\Gamma} := \mathbf{\Gamma}(j_0, k_0) \quad =: \quad (\gamma_{jk}) \in \mathbb{H}^{n\times n}, \tag{4.1}$$

where

$$\begin{pmatrix} \gamma_{j_0 j_0} & \gamma_{j_0 k_0} \\ \gamma_{k_0 j_0} & \gamma_{k_0 k_0} \end{pmatrix} := \mathbf{G} \quad \text{and} \quad \gamma_{jk} \quad := \quad \delta_{jk} \text{ otherwise.}$$

We note that with $\mathbf{G}$ also $\mathbf{\Gamma}$ is unitary. Let $\mathbf{A} = (a_{jk}) \in \mathbb{H}^{n\times n}$ be an arbitrary matrix. We call $\mathbf{\Gamma} = \mathbf{\Gamma}(j_0, k_0)$ a *Givens transformation* if $\mathbf{G}$ has this property with respect to the vector $\mathbf{x} := (a_{j_0,j_0-1}, a_{k_0,j_0-1})^{\mathrm{T}}$. That means, $\mathbf{\Gamma}(j_0, k_0)$ annihilates the element in position $(k_0, j_0 - 1)$ and the formulas of Theorem 3.1 are valid. Then, $\mathbf{A}' = (a'_{jk})$ defined by $\mathbf{A}' = \mathbf{\Gamma}^*\mathbf{A}$ is

$$a'_{jk} = \begin{cases} a_{jk} & \text{if } j \neq j_0, j \neq k_0, \\ ca_{j_0 k} - sa_{k_0 k} & \text{if } j = j_0,\ k = 1, 2, \dots, n, \\ \overline{s}a_{j_0 k} + \overline{c}a_{k_0 k} & \text{if } j = k_0,\ k = j_0, j_0 + 1, \dots, n, \\ \overline{s}a_{j_0,j_0-1} + \overline{c}a_{k_0,j_0-1} = 0 & \text{if } j = k_0,\ k = j_0 - 1. \end{cases} \tag{4.2}$$

In order to compute $\mathbf{A}'$ we need $(2 + 1/7)(2n - (j_0 - 1))$ q-flops. Since algebraic operations with quaternions are expensive an idea of Gentleman[6] and Hammarling[9] can be used to reduce the number of operations. Compare also Rath[14]. An application to complex matrices was given by Xu [20].

As we see from (4.2), only the rows j_0, k_0 are affected. Therefore, in this section, we may assume that $\mathbf{A} \in \mathbb{H}^{2\times n}$. We see from (4.2) that the two possible cases $s = 0$ (implying $|c| = 1$) and $c = 0$ (implying $|s| = 1$) already reduce the computational work of carrying out (4.2) to $(2n - (j_0 - 1))$ q-flops. In our further considerations in this section, we may therefore assume that

$$s, c \in \mathbb{H}\backslash\{0\}. \tag{4.3}$$

For computing $\mathbf{A}' = \mathbf{G}^*\mathbf{A}$, the idea is to use decompositions of $\mathbf{A}$ and of $\mathbf{A}'$ of the form

$$\mathbf{A} = \mathbf{D}\mathbf{B},\ \mathbf{A}' = \mathbf{D}'\mathbf{B}' \text{ with } \mathbf{D} := \operatorname{diag}(d_1, d_2),\ \mathbf{D}' := \operatorname{diag}(d'_1, d'_2),$$

where the diagonal matrices $\mathbf{D}, \mathbf{D}'$ are supposed to be quaternion-valued, nonsingular, but otherwise arbitrary matrices. From the decomposition $\mathbf{A}' = \mathbf{D}'\mathbf{B}' = \mathbf{G}^*\mathbf{A} = \mathbf{G}^*\mathbf{D}\mathbf{B}$, it

ETNA
Kent State University
etna@mcs.kent.edu

follows that

$$\mathbf{B}' = \underbrace{(\mathbf{D}')^{-1}\mathbf{G}^*\mathbf{D}}_{\mathbf{H}^*}\mathbf{B}, \tag{4.4}$$

where $\mathbf{H}^*$, defined as indicated, should be as simple as possible. By this, the computational work for finding $\mathbf{A}'$ is split into three parts:

$$(1.)\ \mathbf{B} = \mathbf{D}^{-1}\mathbf{A}, \quad (2.)\ \mathbf{B}' = \mathbf{H}^*\mathbf{B}, \quad (3.)\ \mathbf{A}' = \mathbf{D}'\mathbf{B}', \tag{4.5}$$

and we will see later whether this splitting can lead to less computational work. Occasionally, we will use the notation

$$\mathbf{A} =: (a_{jk}),\ \mathbf{A}' =: (a'_{jk}),\ \mathbf{B} =: (b_{jk}),\ \mathbf{B}' =: (b'_{jk}),\ j = 1,2;\ k = 1,2,\ldots,n.$$

The matrix $\mathbf{H}^*$ is determined by the two diagonal matrices $\mathbf{D}, \mathbf{D}'$. It is reasonable to discuss two cases for $\mathbf{H}^*$, namely the *standard matrix forms*

$$1.)\ \widetilde{\mathbf{H}}^* := \begin{pmatrix} 1 & -\tilde{u} \\ \tilde{v} & 1 \end{pmatrix}, \quad 2.)\ \widehat{\mathbf{H}}^* := \begin{pmatrix} \hat{u} & -1 \\ 1 & \hat{v} \end{pmatrix}, \tag{4.6}$$

and the *alternative matrix forms*

$$3.)\ \widetilde{\mathbf{H}}^*_{\mathrm{alt}} := \begin{pmatrix} \tilde{u}_{\mathrm{alt}} & -1 \\ \tilde{v}_{\mathrm{alt}} & 1 \end{pmatrix}, \quad 4.)\ \widehat{\mathbf{H}}^*_{\mathrm{alt}} := \begin{pmatrix} 1 & -\hat{u}_{\mathrm{alt}} \\ 1 & \hat{v}_{\mathrm{alt}} \end{pmatrix}. \tag{4.7}$$

where the quantities $\tilde{u}, \tilde{v}, \hat{u}, \hat{v}, \tilde{u}_{\mathrm{alt}}, \tilde{v}_{\mathrm{alt}}, \hat{u}_{\mathrm{alt}}, \hat{v}_{\mathrm{alt}}$ still have to be determined. We shall treat both cases in the two subsequent subsections.

4.1. The standard matrix forms. We first treat the two matrices defined in (4.6). If we do not make special assumptions for the two diagonal matrices $\mathbf{D}, \mathbf{D}'$ steps (1.) and (3.) of (4.5) require together $4n$ q-flops and step (2.) requires $(2 + 2/7)n$ q-flops, more than the original formula (4.2). We first discuss case 1.) of (4.6). From (3.1), (4.6), case 1.), (4.4) we have

$$\begin{pmatrix} cd_1 & -sd_2 \\ \overline{s}d_1 & \overline{c}d_2 \end{pmatrix} = \mathbf{G}^*\mathbf{D} = \mathbf{D}'\widetilde{\mathbf{H}}^* = \begin{pmatrix} d'_1 & -d'_1\tilde{u} \\ d'_2\tilde{v} & d'_2 \end{pmatrix}.$$

This leads to the following equations:

$$d'_1 = cd_1, \quad d'_2 = \overline{c}d_2, \tag{4.8}$$

$$\tilde{u} = (d_1)^{-1}c^{-1}sd_2, \quad \tilde{v} = (d_2)^{-1}(\overline{c})^{-1}\overline{s}d_1. \tag{4.9}$$

From (3.2)(a) we know that $cs = sc$. If we use, in addition, the quaternion formula $q^{-1} = \frac{\overline{q}}{|q|^2}$ for $q \in \mathbb{H}\backslash\{0\}$ we find

$$\tilde{u}\tilde{v} = (d_1)^{-1}c^{-1}s(\overline{c})^{-1}\overline{s}d_1 = \frac{|s|^2}{|c|^2} \text{ and } 1/(1 + \tilde{u}\tilde{v}) = |c|^2 \tag{4.10}$$

which is independent of $\mathbf{D}$. Because of the denominator $|c|^2$ in (4.10), we also discuss case 2.) of (4.6). The corresponding equation reads

$$\begin{pmatrix} cd_1 & -sd_2 \\ \overline{s}d_1 & \overline{c}d_2 \end{pmatrix} = \mathbf{G}^*\mathbf{D} = \mathbf{D}'\widehat{\mathbf{H}}^* = \begin{pmatrix} d'_1\hat{u} & -d'_1 \\ d'_2 & d'_2\hat{v} \end{pmatrix}$$

and leads to:

$$d_1' = sd_2, \quad d_2' = \overline{s}d_1, \tag{4.11}$$

$$\hat{u} = (d_2)^{-1}s^{-1}cd_1, \quad \hat{v} = (d_1)^{-1}(\overline{s})^{-1}\overline{c}d_2. \tag{4.12}$$

We will call the equations (4.8), (4.11) *transition equations*. Now we see, that (4.11), (4.12) are obtained from (4.8), (4.9) by interchanging d_1, d_2 and s, c. Thus, it is sufficient to study case 1.) of (4.6) in more detail. Then, case 2.) follows immediately, e. g.

$$\hat{u}\hat{v} = \frac{|c|^2}{|s|^2} \text{ and } 1/(1+\hat{u}\hat{v}) = |s|^2. \tag{4.13}$$

It is also clear, that we shall use formulas (4.6), case 1.), (4.8), (4.9) in case $|c|^2 \geq 0.5$ and formulas (4.6), case 2.), (4.11), (4.12) when $|s|^2 > 0.5$. Since $|c|^2 + |s|^2 = 1$, exactly one of these cases will occur. This distinction allows us to include the cases $|c| = 1, s = 0$ and $c = 0, |s| = 1$. See, Theorem 3.1, part (b).

The above formulas describe the transition from $\mathbf{D}$ to $\mathbf{D}'$ and the computation of the constants $\tilde{u}, \tilde{v}, \hat{u}, \hat{v}$ in terms of general c, s. To use the annihilation property (last equation of (4.2)) we set

$$\mathbf{a} := (a_1, a_2)^{\mathrm{T}} := (d_1 b_{11}, d_2 b_{21})^{\mathrm{T}}, \tag{4.14}$$

where $\mathbf{a}$ stands for one of the columns of $\mathbf{A}$ and use the terminology of Theorem 3.1. In formulas (4.9), (4.12) there are four products $c^{-1}s, (\overline{c})^{-1}\overline{s}, (s)^{-1}c, (\overline{s})^{-1}\overline{c}$. If we insert the results from Theorem 3.1 we obtain

$$c^{-1}s = -\frac{a_1\overline{a_2}}{|a_1|^2}, \quad (\overline{c})^{-1}\overline{s} = -\sigma\frac{\overline{a_1}a_2}{|a_1|^2}\overline{\sigma}, \tag{4.15}$$

$$s^{-1}c = -\frac{a_2\overline{a_1}}{|a_2|^2}, \quad (\overline{s})^{-1}\overline{c} = -\sigma\frac{\overline{a_2}a_1}{|a_2|^2}\overline{\sigma}. \tag{4.16}$$

There is some freedom in choosing σ, where the details about σ are explained in Theorem 3.1. If in (4.15) we choose $\sigma = a_1/|a_1|$ (implying $c \geq 0$), then, the second product reads

$$(\overline{c})^{-1}\overline{s} = -\frac{a_2\overline{a_1}}{|a_1|^2}. \tag{4.17}$$

If in (4.16) we choose $\sigma = -a_2/|a_2|$ (implying $s \geq 0$), then, the second product simplifies to

$$(\overline{s})^{-1}\overline{c} = -\frac{a_1\overline{a_2}}{|a_2|^2}. \tag{4.18}$$

We compute the coefficients $\tilde{u}, \tilde{v}, \hat{u}, \hat{v}$, defined in (4.9), (4.12) with the help of (4.15) to (4.18) and obtain

$$\begin{cases} \tilde{v} = -b_{21}b_{11}^{-1}, \quad \tilde{u} = \dfrac{|d_2|^2}{|d_1|^2}\,\overline{\tilde{v}} \quad \text{for } \tilde{u}\tilde{v} \leq 1, \\[2ex] \hat{v} = -b_{11}b_{21}^{-1}, \quad \hat{u} = \dfrac{|d_1|^2}{|d_2|^2}\,\overline{\hat{v}} \quad \text{for } \hat{u}\hat{v} < 1. \end{cases} \tag{4.19}$$

Since $1/(1+\tilde{u}\tilde{v}) = |c|^2$, the condition $|c|^2 \geq 0.5$ is equivalent to $\tilde{u}\tilde{v} \leq 1$ and $|c|^2 \leq 1$ implies $\tilde{u}\tilde{v} \geq 0$. The second condition $\hat{u}\hat{v} < 1$ follows similarly. It should be observed that $(\tilde{u}\hat{u}) = (\tilde{v}\hat{v}) = 1$ and that the products $\tilde{u}\tilde{v}, \hat{u}\hat{v}$ are real and nonnegative with $(\tilde{u}\tilde{v})(\hat{u}\hat{v}) = 1$. As a byproduct we see, that no square roots are needed. It is important to note that in case 1

($\tilde{u}\tilde{v} \le 1$) the quantity c is real and nonnegative whereas in case 2 the quantity s is real and nonnegative. Therefore, the transition equations (4.8), (4.11) imply that the diagonal elements always stay real and nonnegative if the starting diagonal matrix has this property. We can square the transition equations and replace c^2, s^2 with $|c|^2, |s|^2$, respectively. Using (4.10),(4.13) the transition equations can be put into the form

$$\begin{aligned} {d'_1}^2 &= \frac{1}{1+\tilde{u}\tilde{v}}\, d_1^2, \quad {d'_2}^2 = \frac{1}{1+\tilde{u}\tilde{v}}\, d_2^2 \quad \text{for } \tilde{u}\tilde{v} \le 1, \\ {d'_1}^2 &= \frac{1}{1+\hat{u}\hat{v}}\, d_2^2, \quad {d'_2}^2 = \frac{1}{1+\hat{u}\hat{v}}\, d_1^2 \quad \text{for } \hat{u}\hat{v} < 1. \end{aligned} \tag{4.20}$$

Our strategy has the consequence, that the factors applied to d_1^2, d_2^2 are always in the range $[0.5, 1]$. We observe that here and in (4.19) only the squares of the diagonal elements are needed. The new "Fast" formula $\mathbf{B}' = \mathbf{H}^*\mathbf{B}$ with $\mathbf{B}' = (b'_{jk})$ reads explicitly

$$b'_{jk} = \begin{cases} \begin{cases} b_{1k} - \tilde{u} b_{2k} & \text{if } j = 1,\ k = 1, 2, \dots, n, \\ \tilde{v} b_{1k} + b_{2k} & \text{if } j = 2,\ k = 2, 3, \dots, n, \end{cases} & \text{for } \tilde{u}\tilde{v} \le 1, \\[2ex] \begin{cases} \hat{u} b_{1k} - b_{2k} & \text{if } j = 1,\ k = 1, 2, \dots, n, \\ b_{1k} + \hat{v} b_{2k} & \text{if } j = 2,\ k = 2, 3, \dots, n, \end{cases} & \text{for } \hat{u}\hat{v} < 1. \end{cases} \tag{4.21}$$

We summarize the algorithm for a given $\mathbf{A} \in \mathbb{H}^{2\times n}$ and a given $\mathbf{D} = \mathrm{diag}(d_1, d_2) \in \mathbb{H}^{2\times 2}$ with $d_1 \ne 0, d_2 \ne 0$.

ALGORITHM 4.1. Fast Givens algorithm

1. Determine $(b_{jk}) := \mathbf{B} := \mathbf{D}^{-1}\mathbf{A}$,
2. compute $(c_1, c_2) := (|d_1|^2|b_{11}|^2, |d_2|^2|b_{21}|^2)$,
3. if $c_1 < c_2$ compute $\hat{u}, \hat{v}, \hat{u}\hat{v} = c_1/c_2$, else compute $\tilde{u}, \tilde{v}, \tilde{u}\tilde{v} = c_2/c_1$ according to (4.19),
4. compute $(b'_{jk}) := \mathbf{B}'$ according to (4.21), use 3.,
5. compute d'_1, d'_2 according to (4.20) and define $\mathbf{D}' := \mathrm{diag}(d'_1, d'_2)$, use 3.,
6. compute $\mathbf{A}' = \mathbf{D}'\mathbf{B}'$.

Since only the squares of the d's appear one could add the following step after step 1.

1a. Replace $\mathbf{D}$ with $\mathbf{D}^2$ (elementwise) and replace all squares of d's in (4.19), (4.20) and in step 2. by d alone.

In this case, step 6. must be replaced by

6$'$. compute $\mathbf{A}' = \sqrt{\mathbf{D}'}\,\mathbf{B}'$, where $\sqrt{}$ has to be computed elementwise. See appendix for roots of quaternions.

Both algorithms (1. to 6. and $1., 1a., 2., \dots, 5., 6'.$) work with general quaternion entries in step 1. However, the choice $\mathbf{D} := \mathrm{diag}(1, 1)$ simplifies 1. to $\mathbf{B} := \mathbf{A}$ and further simplifies step 2. and also step 1a.

We will mention an alternative algorithm. Wolfgang Rath says in his paper [14, p. 51] that "the two parameters ... can unfortunately not be put together and stored in one element". The same comment appears in Schwarz and Köckler[16, p. 239]. This is actually not true. By using μ_c, μ_s from (3.7), (3.8) and also using (4.20) we can write

$$\tilde{u} = \frac{2}{1-|\mu_c|^2} d_1^{-1}\mu_c d_2, \quad \tilde{v} = \frac{2}{1-|\mu_c|^2} d_2^{-1}\overline{\mu_c} d_1 = \frac{|d_1|^2}{|d_2|^2}\,\overline{\tilde{u}}, \tag{4.22}$$

$$
\begin{aligned}
& d_1' = \left(\frac{1-|\mu_c|^2}{1+|\mu_c|^2}\right) d_1, \quad d_2' = \left(\frac{1-|\mu_c|^2}{1+|\mu_c|^2}\right) d_2 \text{ for } |\mu_c|^2 \le 3-2\sqrt{2},\\
& \hat{u} = \frac{2}{1-|\mu_s|^2} d_2^{-1}\mu_s d_1, \quad \hat{v} = \frac{2}{1-|\mu_s|^2} d_1^{-1}\overline{\mu_s} d_2 = \frac{|d_2|^2}{|d_1|^2}\,\overline{\hat{u}},\\
& d_1' = \left(\frac{1-|\mu_s|^2}{1+|\mu_s|^2}\right) d_2, \quad d_2' = \left(\frac{1-|\mu_s|^2}{1+|\mu_s|^2}\right) d_1 \text{ otherwise.}
\end{aligned} \tag{4.23}
$$

Apart from d_1, d_2, all information is stored in one element, namely in μ_c or in μ_s and the above formulas show that we can recover all necessary quantities from $\mu_c(\mu_s)$ alone. Neither squares of the d's nor square roots appear in the above formula. However, the computation of $\mu_c(\mu_s)$ requires one square root of a (nonnegative) real number. This one parameter algorithm seems to be new, even in the real case. Let $\mathbf{A} \in \mathbb{H}^{2\times n}$ and $\mathbf{D} = \mathrm{diag}(d_1, d_2) \in \mathbb{H}^{2\times 2}$ with $d_1 \neq 0, d_2 \neq 0$ be given.

ALGORITHM 4.2. Fast Givens algorithm with only one parameter μ
1. Determine $(b_{jk}) := \mathbf{B} := \mathbf{D}^{-1}\mathbf{A}$,
2. compute $(a_1, a_2) := (d_1 b_{11}, d_2 b_{21})$,
3. if $|a_1| \ge |a_2|$ compute μ_c, else μ_s using (3.9),
4. compute $\tilde{u}, \tilde{v}, d_1', d_2'$ according to (4.22) or $\hat{u}, \hat{v}, d_1', d_2'$ according to (4.23),
5. compute $\mathbf{B}'$ according to (4.21),
6. compute $\mathbf{A}' = \mathbf{D}'\mathbf{B}'$.

It would have been possible to work with μ defined in (3.10) alone. In this case μ_s has to be replaced with μ^{-1} in (4.23).

In the end we will make some investigations on the condition numbers of the two matrices defined in (4.6).

THEOREM 4.3. *Fix $d_1 \neq 0, d_2 \neq 0$ and put $d := \dfrac{|d_1|^2}{|d_2|^2}$. Denote both matrices $\widetilde{\mathbf{H}}, \widehat{\mathbf{H}}$ by $\mathbf{H}$. (a) The condition numbers of the two matrices $\mathbf{H}$ defined in (4.6) are both bounded by*

$$
\mathrm{cond}(\mathbf{H})_{\max}(d) := \max(d, \frac{1}{d}) = \mathrm{cond}(\mathbf{H})_{\max}(\frac{1}{d}) \tag{4.24}
$$

and this bound is sharp. (b) At the midpoint $|s|^2 = 0.5$ the condition number for both matrices is

$$
\begin{aligned}
\mathrm{cond}(\mathbf{H})_{\mathrm{mid}}(d) : &= \sqrt{\frac{(1+d)^2 + \sqrt{(1+d)^4 - 16d^2}}{(1+d)^2 - \sqrt{(1+d)^4 - 16d^2}}}\\
&= \sqrt{\frac{1+\sqrt{1-\frac{16d^2}{(1+d)^4}}}{1-\sqrt{1-\frac{16d^2}{(1+d)^4}}}} = \mathrm{cond}(\mathbf{H})_{\mathrm{mid}}(\frac{1}{d}).
\end{aligned} \tag{4.25}
$$

Proof. (a) We write u for $\tilde{u}, \hat{u}$, and v for $\tilde{v}, \hat{v}$. The eigenvalues of the matrices $\mathbf{H}^*\mathbf{H}$ (and of $\mathbf{H}\mathbf{H}^*$ as well) obey all the same eigenvalue equation (cf. Lemma 1.3). If by (4.19) we use $u = \gamma\,\overline{v}$, where

$$
\gamma = \begin{cases} 1/d & \text{for } \mathbf{H} = \widetilde{\mathbf{H}},\\ d & \text{for } \mathbf{H} = \widehat{\mathbf{H}}, \end{cases} \quad |v|^2 = \begin{cases} |\tilde{v}|^2 & \text{for } \mathbf{H} = \widetilde{\mathbf{H}},\\ |\hat{v}|^2 = \frac{1}{|\tilde{v}|^2} & \text{for } \mathbf{H} = \widehat{\mathbf{H}}, \end{cases}
$$

ETNA
Kent State University
etna@mcs.kent.edu

this eigenvalue equation has by Lemma 1.3 the two solutions

$$\lambda_{\max} = p_1 + p_2, \quad \lambda_{\min} = p_1 - p_2, \text{ where}$$
$$p_1 := \frac{(1+\gamma^2)|v|^2 + 2}{2}, \quad p_2 := \frac{1}{2}\sqrt{(1-\gamma^2)^2|v|^4 + 4(1-\gamma)^2|v|^2}.$$

Thus,

$$\operatorname{cond}(\mathbf{H}) = \sqrt{Q(|v|^2)} \quad \text{where} \quad Q := \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{p_1+p_2}{p_1-p_2} = \frac{1+q}{1-q}, \tag{4.26}$$

$$q(|v|^2) := \frac{p_2(|v|^2)}{p_1(|v|^2)} = \frac{\sqrt{(1-\gamma^2)^2|v|^4 + 4(1-\gamma)^2|v|^2}}{(1+\gamma^2)|v|^2+2}. \tag{4.27}$$

Observe, that $0 \le q < 1$ and that q is strictly increasing with $|v|^2$ for $\gamma \ne 1$. Therefore, Q is also strictly increasing with $|v|^2$. Now, we divide the numerator and denominator in the expression for q by $|v|^2 > 0$ and find

$$\lim_{|v|^2\to\infty} q(|v|^2) = \frac{|\gamma^2-1|}{\gamma^2+1}.$$

From here we deduce

$$\lim_{|v|^2\to\infty} Q(|v|^2) = \begin{cases} \gamma^2 & \text{for } \gamma \ge 1, \\ 1/\gamma^2 & \text{otherwise.} \end{cases}$$

Observing that Q is the square of the condition number and taking into account the meaning of the quantity d yields (4.24). Now, by (4.10), (4.13) it follows that $|v|^2 \to \infty$ implies $|s|^2 \to 1$, $|s|^2 \to 0$ for the two matrices, respectively. Therefore, the bound is sharp. (b) Let $|s|^2 = 0.5$. Then $|c|^2 = 0.5$ and by (4.10) and (4.13) it follows that $uv = 1$ and by (4.19) it follows that $|v|^2 = 1/\gamma$ with $\gamma > 0$. Thus, we have to evaluate $Q(\frac{1}{\gamma})$. We find

$$Q(\frac{1}{\gamma}) = \frac{(1+\gamma)^2 + \sqrt{(1+\gamma)^4 - 16\gamma^2}}{(1+\gamma)^2 - \sqrt{(1+\gamma)^4 - 16\gamma^2}} = Q(\gamma)$$

which proves (4.25). $\square$

Let us mention some test values for $\operatorname{cond}(\mathbf{H})_{\text{mid}}(d)$ for $d = 1, 2, 3, 4, 5, 10, 100$:

$$1,\ 1.64038\,82032\,02208,\ 2.21525\,04370\,21530,\ 2.76308\,57945\,18659,$$
$$3.29666\,29547\,09577,\ 5.87992\,99466\,01142,\ 50.98538\,65368\,50569.$$

From the representation (4.25) it follows that asymptotically we have $\operatorname{cond}(\mathbf{H})_{\text{mid}}(d) \sim \frac{1}{2}\max(d, 1/d)$ for $d \to \infty$ and $d \to 0$.

Though s, c are not explicitly used, we used $|s|^2$ as abscissa for several plots. The formula is (see (4.10), (4.13), (4.19))

$$|s|^2 = \begin{cases} \frac{|\tilde{v}|^2}{d+|\tilde{v}|^2} & \text{for matrix form 1.)} \\ \frac{1}{1+d|\hat{v}|^2} & \text{for matrix form 2.)} \end{cases}, \quad d = \frac{|d_1|^2}{|d_2|^2}. \tag{4.28}$$

From (4.19) it follows that $|\hat{v}|^2 = |\tilde{v}|^{-2}$ implying that both parts in (4.28) are identical.

ETNA
Kent State University
etna@mcs.kent.edu

4.2. The alternative matrix forms. We shall investigate the two matrices defined in (4.7), because we have seen in examples, that the corresponding condition numbers of these matrices may be smaller than the condition number of the matrices investigated in the previous subsection. This case is normally neglected (see Schwarz and Köckler[16, p. 236]). Following the pattern of the last subsection we have to solve the two matrix equations

$$\text{(a)}\quad \begin{pmatrix} cd_1 & -sd_2 \\ \overline{s}d_1 & \overline{c}d_2 \end{pmatrix} = \mathbf{D}'\widetilde{\mathbf{H}}^*_{\text{alt}} = \begin{pmatrix} d_1'\tilde{u}_{\text{alt}} & -d_1' \\ d_2'\tilde{v}_{\text{alt}} & d_2' \end{pmatrix};$$

$$\text{(b)}\quad \begin{pmatrix} cd_1 & -sd_2 \\ \overline{s}d_1 & \overline{c}d_2 \end{pmatrix} = \mathbf{D}'\widehat{\mathbf{H}}^*_{\text{alt}} = \begin{pmatrix} d_1' & -d_1'\hat{u}_{\text{alt}} \\ d_2' & d_2'\hat{v}_{\text{alt}} \end{pmatrix}.$$

The solutions for $s \neq 0$ and $c \neq 0$ are

$$\text{(4.29)}\qquad \begin{aligned} &\text{(a)}\quad d_1' = sd_2,\ d_2' = \overline{c}d_2,\ \tilde{u}_{\text{alt}} = d_2^{-1}s^{-1}cd_1,\ \tilde{v}_{\text{alt}} = d_2^{-1}(\overline{c})^{-1}\overline{s}d_1;\\ &\text{(b)}\quad d_1' = cd_1,\ d_2' = \overline{s}d_1,\ \hat{u}_{\text{alt}} = d_1^{-1}c^{-1}sd_2,\ \hat{v}_{\text{alt}} = d_1^{-1}(\overline{s})^{-1}\overline{c}d_2. \end{aligned}$$

Since $sc = cs$ we obtain from (a) and (b)

$$\text{(4.30)}\qquad \begin{aligned} &\text{(a)}\quad \tilde{u}_{\text{alt}}(\tilde{v}_{\text{alt}})^{-1} = \frac{|c|^2}{|s|^2} \Rightarrow |s|^2 = \frac{1}{1+\tilde{u}_{\text{alt}}(\tilde{v}_{\text{alt}})^{-1}};\\ &\text{(b)}\quad \hat{u}_{\text{alt}}(\hat{v}_{\text{alt}})^{-1} = \frac{|s|^2}{|c|^2} \Rightarrow |c|^2 = \frac{1}{1+\hat{u}_{\text{alt}}(\hat{v}_{\text{alt}})^{-1}}. \end{aligned}$$

In both formulas (4.29) the quantities s^{-1} and c^{-1} appear simultaneously. Thus, we should avoid both formulas in cases where either $|s|$ or $|c|$ is small. By the same techniques (using (4.15)–(4.18)) as in the previous subsection, we obtain

$$\text{(4.31)}\qquad \begin{cases} \tilde{v}_{\text{alt}} = -b_{21}b_{11}^{-1}, & \tilde{u}_{\text{alt}} = \dfrac{|d_1|^2}{|d_2|^2}\dfrac{\tilde{v}_{\text{alt}}}{|\tilde{v}_{\text{alt}}|^2},\\ \hat{v}_{\text{alt}} = -b_{11}b_{21}^{-1}, & \hat{u}_{\text{alt}} = \dfrac{|d_2|^2}{|d_1|^2}\dfrac{\hat{v}_{\text{alt}}}{|\hat{v}_{\text{alt}}|^2}. \end{cases}$$

The transition equations (4.29) should be given the following form:

$$\text{(4.32)}\qquad \begin{aligned} &\text{(a)}\quad d_1' = |d_2|^2\sqrt{e|s|^2}\,(\overline{d_1})^{-1}\overline{\tilde{v}_{\text{alt}}},\quad d_2' = \sqrt{e|s|^2}\,d_2,\\ &\text{where } e|s|^2 := \frac{e}{1+e},\quad d := \frac{|d_1|^2}{|d_2|^2},\quad b := \frac{|b_{11}|^2}{|b_{21}|^2},\quad e := db,\\ &\text{(b)}\quad d_1' = |d_1|^2\sqrt{|s|^2}\,(\overline{d_2})^{-1}\overline{\hat{v}_{\text{alt}}},\quad d_2' = \sqrt{|s|^2}\,d_1,\quad |s|^2 = \frac{1}{1+e}. \end{aligned}$$

The new "Fast" formula $\mathbf{B}' = \mathbf{H}^*\mathbf{B}$ with $\mathbf{B}' = (b'_{jk})$ reads in this case explicitly

$$\text{(4.33)}\qquad b'_{jk} = \begin{cases} \begin{cases} \tilde{u}_{\text{alt}}b_{1k} - b_{2k} & \text{if } j=1,\ k=1,2,\ldots,n,\\ \tilde{v}_{\text{alt}}b_{1k} + b_{2k} & \text{if } j=2,\ k=2,3,\ldots,n, \end{cases}\\ \\ \begin{cases} b_{1k} - \hat{u}_{\text{alt}}b_{2k} & \text{if } j=1,\ k=1,2,\ldots,n,\\ b_{1k} + \hat{v}_{\text{alt}}b_{2k} & \text{if } j=2,\ k=2,3,\ldots,n, \end{cases} \end{cases}$$

ETNA
Kent State University
etna@mcs.kent.edu

where the use of the first or second form will be specified later. Because of the singularities of $\tilde{u}_{\rm alt}, \tilde{v}_{\rm alt}, \hat{u}_{\rm alt}, \hat{v}_{\rm alt}$ at $s = 0$ and $c = 0$ (see 4.29) an algorithm cannot be based on the alternative forms alone. Another disadvantage of this form is that the transition equations for the d's (the first two equations in (a) and (b) of (4.29)) have different factors with the consequence that in general the d's will not remain real. Since the condition numbers of both matrices $\widetilde{\mathbf{H}}_{\rm alt}$, $\widehat{\mathbf{H}}_{\rm alt}$ has poles at both ends $|s| = 0$ and $|s| = 1$ more information from plots of the reciprocal of the condition numbers (rather than from the condition numbers direct) can be gained. See Figure 4.1.

Since the condition number of these two matrices is small "in the middle" of the interval $[0, 1]$, a strategy to use them has been developed which is summarized in Table 4.1.

THEOREM 4.4. *Let $d_1 \neq 0, d_2 \neq 0$ and put $d := \frac{|d_1|^2}{|d_2|^2}$. Denote both matrices defined in (4.7) by $\mathbf{H}_{\rm alt}$ and use the notation $v_{\rm alt}$ for both $\tilde{v}_{\rm alt}$ and $\hat{v}_{\rm alt}$. Put*

$$\gamma_{\rm alt} := \begin{cases} d & \textit{for } \mathbf{H}_{\rm alt} = \widetilde{\mathbf{H}}_{\rm alt}, \\ 1/d & \textit{for } \mathbf{H}_{\rm alt} = \widehat{\mathbf{H}}_{\rm alt}, \end{cases} \qquad |v_{\rm alt}|^2 := \begin{cases} |\tilde{v}_{\rm alt}|^2 = |\tilde{v}|^2 & \textit{for } \mathbf{H}_{\rm alt} = \widetilde{\mathbf{H}}_{\rm alt}, \\ |\hat{v}_{\rm alt}|^2 = \frac{1}{|\tilde{v}_{\rm alt}|^2} & \textit{for } \mathbf{H}_{\rm alt} = \widehat{\mathbf{H}}_{\rm alt}. \end{cases}$$

(a) The condition number of $\mathbf{H}_{\rm alt}$ is

$$\text{cond}(\mathbf{H}_{\rm alt}) = \sqrt{Q_{\rm alt}(|v_{\rm alt}|^2)} \quad \textit{where} \quad Q_{\rm alt} := \frac{1 + q_{\rm alt}}{1 - q_{\rm alt}} \quad \textit{and} \tag{4.34}$$

$$q_{\rm alt}(|v_{\rm alt}|^2) := \frac{\sqrt{\left(|v_{\rm alt}|^2 + \frac{\gamma_{\rm alt}^2}{|v_{\rm alt}|^2}\right)^2 + 4 - 8\gamma_{\rm alt}}}{|v_{\rm alt}|^2 + \frac{\gamma_{\rm alt}^2}{|v_{\rm alt}|^2} + 2}. \tag{4.35}$$

(b) At the midpoint $|s|^2 := 0.5$ which corresponds to $|v_{\rm alt}|^2 := \gamma_{\rm alt}$ the condition number is

$$\text{cond}(\mathbf{H}_{\rm alt})_{\rm mid}(d) = \sqrt{\max(d, 1/d)} = \text{cond}(\mathbf{H}_{\rm alt})_{\rm mid}(\frac{1}{d}). \tag{4.36}$$

Proof. (a) Define $\mathbf{M} := \mathbf{H}_{\rm alt}^* \mathbf{H}_{\rm alt}$. Then,

$$\mathbf{M} = \begin{pmatrix} \frac{\gamma_{\rm alt}^2}{|v_{\rm alt}|^2} + 1 & \pm(\gamma_{\rm alt} - 1) \\ \pm(\gamma_{\rm alt} - 1) & |v_{\rm alt}|^2 + 1 \end{pmatrix},$$

where the $+$-sign refers to the first matrix $\widetilde{\mathbf{H}}_{\rm alt}$. Applying Definition 1.2 to $\mathbf{H}_{\rm alt}$ and employing Lemma 1.3 for $\mathbf{M}$, we find the expressions given in (4.34) and in (4.35).
(b) Follows directly from (4.34), (4.35) by inserting $|v_{\rm alt}|^2 = \gamma_{\rm alt}$. ∎

It is interesting to see the condition numbers of all four matrices for a fixed value of d in only one plot. Since the condition number of $\mathbf{H}_{\rm alt}$ is singular at both endpoints of $|s|^2 \in [0, 1]$ we plot the reciprocal of the condition numbers of all four matrices. Then, instead of looking at the minima, we have to look at the maxima. In Figure 4.1, we have graphed the reciprocal of the condition numbers of all four matrix forms (see (4.6),(4.7)) for a fixed quotient $d := |d_1|^2/|d_2|^2$ over the quantity $|s|^2 \in [0, 1]$ and, in addition, the maximum of all four curves (dashed). In the given figure we have selected a $d > 1$. If we would replace d with $1/d < 1$ only the graphs corresponding to the matrix forms 3.) and 4.)

ETNA
Kent State University
etna@mcs.kent.edu

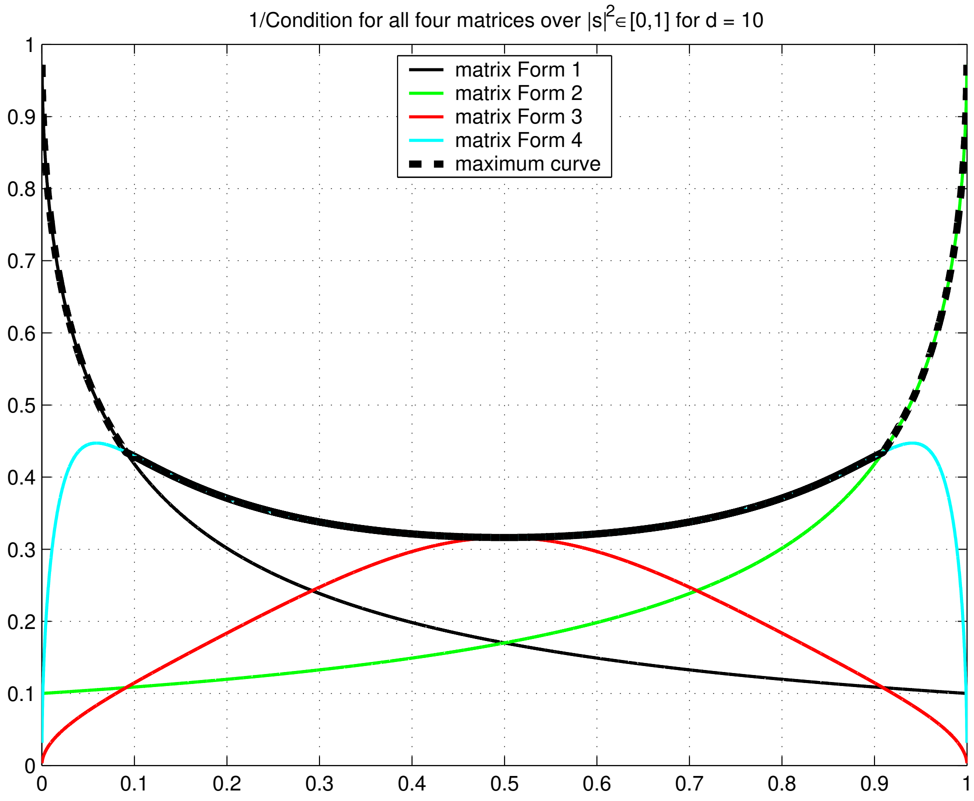

FIG. 4.1. *Reciprocal of the condition numbers for all four matrices for* $d = 10$.

would interchange. The curves corresponding to matrix forms 1.) and 2.) show monotonous behavior with respect to the abscissa $|s|^2$ and do not change if we would replace d by $1/d$. Thus, the graph of the maximum for all four curves is invariant under $d \to 1/d$.

The relation between $|s|^2$ and $|v|^2$ (for matrix forms 1.) and 2.)) is given in (4.28). The new formula for $|s|^2$ is (see (4.30), (4.31))

$$|s|^2 = \begin{cases} \frac{1}{1+d|\tilde{v}_{\mathrm{alt}}|^2} & \text{for matrix form 3.)} \\ \frac{|\hat{v}_{\mathrm{alt}}|^2}{d+|\hat{v}_{\mathrm{alt}}|^2} & \text{for matrix form 4.)} \end{cases}, \quad d = \frac{|d_1|^2}{|d_2|^2}. \tag{4.37}$$

From (4.31) it follows that $|\hat{v}_{\mathrm{alt}}|^2 = |\tilde{v}_{\mathrm{alt}}|^{-2}$ implying that both parts in (4.37) are identical. By comparing the expressions for $\tilde{v}$ and $\tilde{v}_{\mathrm{alt}}$ at (4.19) and (4.31) we see that $|\tilde{v}|^2 = |\tilde{v}_{\mathrm{alt}}|^{-2}$ (though $\tilde{v}$ and $\tilde{v}_{\mathrm{alt}}^{-1}$ differ). Figure 4.1 shows that at the midpoint $|s|^2 = 0.5$ the alternative matrix forms have smaller condition numbers than the standard forms.

COROLLARY 4.5. *Let $d > 0$ be fixed and $d \neq 1$. At the midpoint $|s|^2 = 0.5$, the condition number of $\mathbf{H}_{\mathrm{alt}}$ denoted by $\mathrm{cond}(\mathbf{H}_{\mathrm{alt}})_{\mathrm{mid}}(d)$ is smaller than the condition number $\mathrm{cond}(\mathbf{H})_{\mathrm{mid}}(d)$ of $\mathbf{H}$, or*

$$\mathrm{cond}(\mathbf{H}_{\mathrm{alt}})_{\mathrm{mid}}(d) < \mathrm{cond}(\mathbf{H})_{\mathrm{mid}}(d). \tag{4.38}$$

TABLE 4.1
Logic of Fast algorithm using all four matrix forms

	$\lvert s\rvert^2 \in [0, a]$	$\lvert s\rvert^2 \in]a, 1-a]$	$\lvert s\rvert^2 \in]1-a, 1]$
$d \geq 1$	$\begin{pmatrix} 1 & -\tilde{u} \\ \tilde{v} & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -\hat{u}_{\rm alt} \\ 1 & \hat{v}_{\rm alt} \end{pmatrix}$	$\begin{pmatrix} \hat{u} & -1 \\ 1 & \hat{v} \end{pmatrix}$
$d < 1$	$\begin{pmatrix} 1 & -\tilde{u} \\ \tilde{v} & 1 \end{pmatrix}$	$\begin{pmatrix} \tilde{u}_{\rm alt} & -1 \\ \tilde{v}_{\rm alt} & 1 \end{pmatrix}$	$\begin{pmatrix} \hat{u} & -1 \\ 1 & \hat{v} \end{pmatrix}$

Proof. The corresponding formulas (in the above order) are given in (4.36) and in (4.25). By using the notation $q := \sqrt{(1+d)^4 - 16d^2}$ the inequality (4.38) is equivalent to the inequality

$$\max(d, \frac{1}{d}) < \frac{(1+d)^2 + q}{(1+d)^2 - q}. \tag{4.39}$$

Let $d > 1$. Then, this inequality is equivalent to

$$q > d^2 - 1.$$

By squaring and putting in the meaning of q we obtain

$$(1+d)^4 - 16d^2 > (d^2 - 1)^2.$$

From here we have $(1+d)^4 - 16d^2 - (d^2-1)^2 = 4d(d-1)^2 > 0$. Very similar arguments work for $0 < d < 1$. □

Inequality (4.38) implies that (in case $d \neq 1$) it is also true in a certain (symmetric) neighborhood of $|s|^2 = 0.5$. And a look at Figure 4.1 shows this neighborhood, explicitly. Thus, we can base an algorithm on the partition of $|s|^2 \in [0, 1]$ into three subintervals according to Table 4.1:

It remains to determine $0 < a < 0.5$. Let us call the curve of the condition number corresponding to matrix form j.) *the curve* j, $j = 1, 2, 3, 4$. The quantity a is the smallest positive abscissa $|s|^2$ of the intersection of curves 1 and 4 if $d > 1$, and $1 - a$ is the largest positive abscissa of the intersection of curves 2 and 3 if $d < 1$.

LEMMA 4.6. *The above quantity a is given by*

$$a = |s_0|^2 := \frac{1}{1 + \max(d, 1/d)} \text{ for } d > 0. \tag{4.40}$$

Proof. Let $d > 1$. Then, a is the smallest, positive solution (in terms of $|s|^2$) of $\Delta := q^2(|v|^2) - q_{\rm alt}^2(|v_{\rm alt}|^2) = 0$ where the corresponding expressions are defined in (4.27), (4.35). In these expressions we put $w := |v|^2 = |\tilde{v}|^2 = 1/|v_{\rm alt}|^2, \gamma = 1/d$ such that (via first line of (4.28)).

$$|s|^2 = w/(d + w).$$

ETNA
Kent State University
etna@mcs.kent.edu

Then, Δ is a rational function in the variable w of numerator and denominator degree six with positive denominator. The numerator can be factored with the help of `maple`* in the form

$$-4d^2(w+d)^2(w-1)\{w^3+(-d^4+2d^2)w^2-d^4w-d^4\}.$$

The zero $w := -d < 0$ is irrelevant and the zero $w = 1$ corresponds to $|s_0|^2 := 1/(1+d)$. The polynomial $p(w;d) := \{w^3+(-d^4+2d^2)w^2-d^4w-d^4\}$ has exactly one positive root $w_1 > d^2$ which implies that $|s_1|^2 := w_1/(d+w_1) > 1/(1+d) = |s_0|^2$. Similar arguments work for the case $d < 1$. $\square$

This lemma tells us that even for d of moderate size the interval $[a, 1-a]$ in which the alternative matrix forms have smaller condition numbers is already considerably large. Already for $d > 3$ the interval $[a, 1-a]$ is longer than the two subintervals $[0, a], [1-a, 1]$ together.

4.3. Fast Givens algorithm with four matrix forms. The principal algorithm is laid out in Table 4.1. The decision whether $|s|^2$ is in one of the intervals $[0, a],]a, 1-a],]1-a, 1]$ turns out to be very easy. We assume that $d_1 \neq 0$, $d_2 \neq 0$, b_{11}, b_{21} are given quaternions. We set

$$\begin{aligned} &d := |d_1|^2/|d_2|^2, \quad b := |b_{11}|^2/|b_{21}|^2 \text{ for } b_{21} \neq 0,\\ &a := 1/(1+\max(d, 1/d)), \quad e := bd. \end{aligned}$$

Then $|s|^2$ can easily be expressed in the form

$$|s|^2 = \begin{cases} 0 & \text{if } b_{21} = 0,\\ \frac{1}{1+e} & \text{otherwise.} \end{cases}$$

Then, with the help of Lemma 4.6 and Table 4.1 we find

$$|s|^2 \in \begin{cases} [0, a] & \text{if } b \geq \max(1, 1/d^2) \text{ or } b_{21} = 0 \text{ [matrix form 1.) is used]},\\ ]a, 1-a] & \text{if } \min(1, 1/d^2) \leq b < \max(1, 1/d^2) \text{ [matrix form 3.)}\\ & \text{for } d < 1, \text{ matrix form 4.) for } d \geq 1 \text{ is used]},\\ ]1-a, 1] & \text{if } b < \min(1, 1/d^2) \text{ [matrix form 2.) is used]}. \end{cases}$$

Now, the corresponding formulas are:

1. Matrix form 1.): $\tilde{u}, \tilde{v}$ in (4.19), d_1', d_2' in (4.20), but no squares, use $\tilde{u}\tilde{v} = 1/e$.
2. Matrix form 2.): $\hat{u}, \hat{v}$ in (4.19), d_1', d_2' in (4.20), but no squares, use $\hat{u}\hat{v} = e$.
3. Matrix form 3.): to be used only if $d < 1$, $\tilde{u}_{\text{alt}}, \tilde{v}_{\text{alt}}$ in (4.31), d_1', d_2' in (4.32 (a)). use $\tilde{u}_{\text{alt}}\tilde{v}_{\text{alt}}^{-1} = e$,
4. Matrix form 4.): to be used only if $d \geq 1$, $\hat{u}_{\text{alt}}, \hat{v}_{\text{alt}}$ in (4.31), d_1', d_2' in (4.32 (b)), use $\hat{u}_{\text{alt}}\hat{v}_{\text{alt}}^{-1} = 1/e$.

The use of all four matrix forms implies that the transition equations must be used in the square free form. The advantage is, that in the end no square roots need to be computed.

4.4. Avoiding underflows in the diagonal elements. If we have a look at the transition equations ((4.8), (4.11), (4.29)) which describe the transition from $\mathbf{D}$ to $\mathbf{D}'$, we see that the matrix elements of $\mathbf{D}$ are always multiplied by quantities which are smaller than one in modulus. The effect is that they are always decreasing in modulus. In Hessenberg reduction each diagonal element (apart from the first) of $\mathbf{D}$ is multiplied $n-2$ times. Thus, if the

---

* Actually, one can directly check, that $q^2(1) - q_{\text{alt}}^2(1) = 0$.

multiplicator is in the average 0.75 and $n = 100$, then the diminishing factor is already $(3/4)^{98} \approx 6 \cdot 10^{-13}$. Therefore, we will discuss shortly a method introduced by Anda and Park[1] for the real case in which they try to avoid this phenomenon by introducing another matrix decomposition. In our context, using matrix form $\widetilde{\mathbf{H}}^*$ (see (4.6)) the main idea is contained in the following identity:

$$\mathbf{D}' = \begin{pmatrix} cd_1 & 0 \\ 0 & \overline{c}d_2 \end{pmatrix} = \begin{pmatrix} cd_1 & 0 \\ 0 & c^{-1}d_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix} =: \widetilde{\mathbf{D}} \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix}.$$

The final clue is the following equation and decomposition:

$$\mathbf{D}'\widetilde{\mathbf{H}}^* = \widetilde{\mathbf{D}} \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix} \widetilde{\mathbf{H}}^* = \widetilde{\mathbf{D}} \begin{pmatrix} 1 & -u \\ |c|^2 v & |c|^2 \end{pmatrix} = \widetilde{\mathbf{D}} \begin{pmatrix} 1 & 0 \\ |c|^2 v & 1 \end{pmatrix} \begin{pmatrix} 1 & -u \\ 0 & 1 \end{pmatrix}. \tag{4.41}$$

The last decomposition shows, that a multiplication with a vector can still be carried out by using only two (quaternion) multiplications. We also remark here that this matrix form is only used in case c is real and $|c|^2 \geq 0.5$. In case $|c|^2 < 0.5$ we use matrix form 2.) with a similar expansion.

Matrix form 3.) is $\widetilde{\mathbf{H}}^*_{\text{alt}} = \begin{pmatrix} u & -1 \\ v & 1 \end{pmatrix}$ and for the transition equations we can make a corresponding development. See (4.29):

$$\mathbf{D}' = \begin{pmatrix} s & 0 \\ 0 & \overline{c} \end{pmatrix} d_2 = \begin{pmatrix} s & 0 \\ 0 & c^{-1} \end{pmatrix} d_2 \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix} =: \widetilde{\mathbf{D}} \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix}.$$

Therefore,

$$\mathbf{D}'\widetilde{\mathbf{H}}^*_{\text{alt}} = \widetilde{\mathbf{D}} \begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix} \begin{pmatrix} u & -1 \\ v & 1 \end{pmatrix}.$$

The last product can also be factored in the form

$$\begin{pmatrix} 1 & 0 \\ 0 & |c|^2 \end{pmatrix} \begin{pmatrix} u & -1 \\ v & 1 \end{pmatrix} = \begin{pmatrix} u & 0 \\ |c|^2 v & 1 \end{pmatrix} \begin{pmatrix} 1 & -u^{-1} \\ 0 & 1 \end{pmatrix} \tag{4.42}$$

where we may assume that c is real and positive. However, as before, we still have two quaternion multiplications (56 flops) but we need one additional multiplication real $\times$ quaternion (4 flops). If we compare the two different decompositions of the last equation, we see, that the product on the right has no advantages over the product on the left and therefore we do not use the right decomposition. When using matrix form 3.) we assume, that both $|s|^2, |c|^2$ are near the middle of $[0, 1]$ such that there is no preference of s over c or vice versa. Therefore, we may also use the form

$$\mathbf{D}'\widetilde{\mathbf{H}}^*_{\text{alt}} = \widetilde{\mathbf{D}} \begin{pmatrix} |s|^2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u & -1 \\ v & 1 \end{pmatrix} \text{ with } \widetilde{\mathbf{D}} := \begin{pmatrix} (\overline{s})^{-1} d_2 & 0 \\ 0 & \overline{c}d_2 \end{pmatrix}. \tag{4.43}$$

Anda and Park[1] reduce or avoid the possibility of an underflow in the diagonal elements of $\mathbf{D}$. However, instead, they introduce a multiplication by $|c|^2$ or $|s|^2$ of the matrix elements (see (4.41) to (4.43)) and thus, increase the danger of underflows in the matrix elements. Therefore, we have not implemented this technique. The use of matrix form $\mathbf{H}^*_{\text{alt}}$ was not considered by Anda and Park[1].

ETNA
Kent State University
etna@mcs.kent.edu

5. Hessenberg reduction of quaternion valued matrices by Givens transformations. Assume that $\mathbf{A} \in \mathbb{H}^{n\times n}$ is an arbitrary matrix. By similarity transformations with Givens transformations we want to reduce $\mathbf{A} = (a_{jk})$ to upper Hessenberg form, see Janovská and Opfer [11]. Such a Hessenberg form is usually the starting point for the application of a QR algorithm to find the eigenvalues of $\mathbf{A}$. A QR algorithm for quaternion valued matrices was developed by Bunse-Gerstner, Byers, and Mehrmann [2]. The reduction steps with $\mathbf{\Gamma} := \mathbf{\Gamma}(j_0, k_0)$, defined in (4.1) have the form

$$\mathbf{A}'' := \mathbf{\Gamma}^*\mathbf{A}\mathbf{\Gamma}. \tag{5.1}$$

The matrix $\mathbf{A}''$ will differ from $\mathbf{A}$ only in rows and columns with numbers j_0, k_0. Let us introduce $\mathbf{A}' := (a'_{jk})$ where $\mathbf{A}' = \mathbf{\Gamma}^*\mathbf{A}$. Then, $\mathbf{A}'' = (a''_{jk}) := \mathbf{A}'\mathbf{\Gamma}$ with

$$a''_{jk} = \begin{cases} a'_{jk} & \text{for } k \neq j_0, k \neq k_0, \\ a'_{jj_0}\overline{c} - a'_{jk_0}\overline{s} & \text{for } k = j_0,\ j = 1, 2, \ldots, n, \\ a'_{jj_0}s + a'_{jk_0}c & \text{for } k = k_0,\ j = 1, 2, \ldots, n. \end{cases} \tag{5.2}$$

A formula for computing (a'_{jk}) is given in (4.2), and we already found that $(2 + 1/7)(2n - (j_0 - 1))$ q-flops were needed. In order to compute all elements of $\mathbf{A}''$ from $\mathbf{A}'$, we need $(4+2/7)n$ q-flops. Since the total number of q-flops is a cubic polynomial p in n we can use $p(0) = p(1) = p(2) = 0, p(3) = (23 + 4/7)$ to obtain

$$p(n) = (4 - 2/28)(n^3 - 3n^2 + 2n) \quad \text{q-flops} \tag{5.3}$$

as the total number of computing $\mathbf{A}'$ and $\mathbf{A}''$. This number is valid under the assumption that all entries are quaternion entries. If c or s is real, then the factor $(4 - 2/28) = 55/14$ at $p(n)$ has to be replaced with $(2 + 10/28) = 33/14$.

6. Hessenberg reduction with Fast Givens transformations for quaternion valued matrices. We start with the same setting as in Section 4 and 5. There is an arbitrary matrix $\mathbf{A} \in \mathbb{H}^{n\times n}$ and a unitary matrix $\mathbf{\Gamma}$ defined in (4.1) depending on a fixed matrix position (j_0, k_0) where $1 < j_0 < k_0 \leq n$. This matrix has the effect that in $\mathbf{A}' := \mathbf{\Gamma}^*\mathbf{A}$ it annihilates the element in position $(k_0, j_0 - 1)$. In $\mathbf{A}'' := \mathbf{A}'\mathbf{\Gamma}$ only the columns j_0, k_0 differ from those in $\mathbf{A}'$. Therefore, also in $\mathbf{A}''$ the element at $(k_0, j_0 - 1)$ is zero. If $\mathbf{A}$ happens to be Hermitean, then, so will be $\mathbf{A}''$.

The “Fast Idea” is the same as in Section 4, namely to introduce a matrix decomposition into (5.1) which has the following form:

$$\mathbf{A} := \mathbf{D}\mathbf{B}\mathbf{D}^*, \quad \mathbf{A}'' := \mathbf{D}'\mathbf{B}''\mathbf{D}'^*, \tag{6.1}$$

where $\mathbf{D} := \operatorname{diag}(d_1, d_2, \ldots, d_n)$, $\mathbf{D}' := \operatorname{diag}(d'_1, d'_2, \ldots, d'_n)$ are suitably chosen, nonsingular but otherwise arbitrary diagonal matrices in $\mathbb{H}^{n\times n}$. Then, (6.1), (5.1) imply:

$$\mathbf{B}'' = \underbrace{(\mathbf{D}')^{-1}\mathbf{\Gamma}^*\mathbf{D}}_{\mathbf{H}^*}\,\mathbf{B}\,\underbrace{\mathbf{D}^*\mathbf{\Gamma}(\mathbf{D}'^*)^{-1}}_{\mathbf{H}}. \tag{6.2}$$

We see that (apart from the dimension) the above defined matrix $\mathbf{H}^*$ coincides with the former matrix $\mathbf{H}^*$ introduced in (4.4). If we choose the diagonal elements of $\mathbf{D}, \mathbf{D}'$ according to

$$d_j = d'_j \text{ for all } j \neq j_0, j \neq k_0, \tag{6.3}$$

$\mathbf{H}^*$ coincides with the identity matrix apart from the rows and columns indexed j_0, k_0, respectively. The complete analysis from Section 4 is therefore, valid here. In the formulas (4.19), (4.20) we only have to replace d_1, d_1' by d_{j_0}, d_{j_0}'; d_2, d_2' by d_{k_0}, d_{k_0}'; a_{11}, b_{11} by $a_{j_0,j_0-1}, b_{j_0,j_0-1}$, and a_{21}, b_{21} by $a_{k_0,j_0-1}, b_{k_0,j_0-1}$, respectively. Let us rewrite the iteration steps derived from (6.2) in the form

$$\mathbf{B}_{\ell+1} := \mathbf{H}_\ell^* \mathbf{B}_\ell \mathbf{H}_\ell = (\mathbf{D}_{\ell+1})^{-1} \mathbf{\Gamma}_\ell^* \mathbf{D}_\ell \, \mathbf{B}_\ell \, \mathbf{D}_\ell^* \mathbf{\Gamma}_\ell (\mathbf{D}_{\ell+1}^*)^{-1}, \tag{6.4}$$
$$\ell := 0, 1, \ldots, m-1, \quad m := (n-2)(n-1)/2, \text{ then,}$$
$$\mathbf{B}_m := (\mathbf{D}_m)^{-1} \mathbf{\Gamma}_{m-1}^* \cdots \mathbf{\Gamma}_0^* \, \mathbf{D}_0 \mathbf{B}_0 \mathbf{D}_0^* \, \mathbf{\Gamma}_0 \cdots \mathbf{\Gamma}_{m-1} (\mathbf{D}_m^*)^{-1}. \tag{6.5}$$

Then, the wanted Hessenberg form is

$$\mathbf{A}_m := \mathbf{D}_m \mathbf{B}_m \mathbf{D}_m^*, \text{ where } \mathbf{B}_0 := \mathbf{D}_0^{-1} \mathbf{A} (\mathbf{D}_0^*)^{-1}. \tag{6.6}$$

For algebraic simplicity it is reasonable to choose $\mathbf{D}_0 := \mathbf{I}$, then $\mathbf{B}_0 := \mathbf{A}$. But see the comment at the end of the numerical example. If all $\mathbf{D}_\ell$ already contain the squares of the diagonal entries, then, the last step must read

$$\mathbf{A}_m := \sqrt{\mathbf{D}_m} \, \mathbf{B}_m \sqrt{\mathbf{D}_m^*}, \tag{6.7}$$

where $\sqrt{}$ should be applied elementwise. For square roots of quaternions, see the appendix.

In order to compute $\mathbf{B}'' := \mathbf{B}_{\ell+1}$ from $\mathbf{B}' := \mathbf{B}_\ell$ according to formula (6.4) we first compute $(b_{jk}') := \mathbf{B}'$ according to a slightly modified (4.21), namely

$$b_{jk}' = \begin{cases} \begin{cases} b_{j_0,k} - \tilde{u} b_{k_0,k} & \text{if } j = j_0, \ k = 1, 2, \ldots, n, \\ 0 & \text{if } j = k_0, \ k = 1, 2, \ldots, j_0 - 1, \\ \tilde{v} b_{j_0,k} + b_{k_0,k} & \text{if } j = k_0, \ k = j_0, j_0 + 1, \ldots, n, \end{cases} & \text{for } \tilde{u}\tilde{v} \le 1, \\ \\ \begin{cases} \hat{u} b_{j_0,k} - b_{k_0,k} & \text{if } j = j_0, \ k = 1, 2, \ldots, n, \\ 0 & \text{if } j = k_0, \ k = 1, 2, \ldots, j_0 - 1, \\ b_{j_0,k} + \hat{v} b_{k_0,k} & \text{if } j = k_0, \ k = j_0, j_0 + 1, \ldots, n, \end{cases} & \text{for } \hat{u}\hat{v} < 1, \end{cases} \tag{6.8}$$

and then $(b_{jk}'') := \mathbf{B}''$ according to

$$b_{jk}'' = \begin{cases} \begin{cases} b_{j,j_0}' - b_{j,k_0}' \overline{\tilde{u}} & \text{if } k = j_0, \ j = 1, 2, \ldots, n, \\ b_{j,j_0}' \overline{\tilde{v}} + b_{j,k_0}' & \text{if } k = k_0, \ j = 1, 2, \ldots, n, \end{cases} & \text{for } \tilde{u}\tilde{v} \le 1, \\ \\ \begin{cases} b_{j,j_0}' \overline{\hat{u}} - b_{j,k_0}' & \text{if } k = j_0, \ j = 1, 2, \ldots, n, \\ b_{j,j_0}' + b_{j,k_0}' \overline{\hat{v}} & \text{if } k = k_0, \ j = 1, 2, \ldots, n, \end{cases} & \text{for } \hat{u}\hat{v} < 1. \end{cases} \tag{6.9}$$

If we count the q-flops together for evaluating (6.8) and for (6.9) we arrive at

$$p(n) = (44/21)(n^3 - 3n^2 + 2n) \quad \text{q-flops.} \tag{6.10}$$

A comparison with (5.3) using the realistic factor $33/14$ instead of $4 - 2/28$ shows that we have a moderate gain factor of 1.125 which will later appear also in Table 6.1.

The general relation between the iteration number ℓ and the corresponding matrix position (j, k) can be expressed by

$$\mathbf{\Gamma}_{n(j-2)-j(j+1)/2+k+1} = \mathbf{\Gamma}(j, k); \quad 1 < j < k \le n, \tag{6.11}$$

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 6.1
Flop comparisons: ordinary and Fast Givens transformation

		$y_1 = cx_1 - sx_2$ $y_2 = \overline{s}x_1 + \overline{c}x_2$			$y_1 = x_1 - ux_2$ $y_2 = vx_1 + x_2$			gain inF	gain inM	gainfactor in F
		A	M	F	A	M	F			
1.	all$\in \mathbb{R}$	2	4	6	2	2	4	2	2	1.5
2.	all$\in \mathbb{C}$	12	16	28	8	8	16	12	8	
3.	$c \in \mathbb{R}$	8	12	20	8	8	16	4	4	1.25
4.	all $\in \mathbb{H}$	56	64	120	32	32	64	56	32	
5.	$c \in \mathbb{R}$	32	40	72	32	32	64	8	8	1.125

where $\mathbf{\Gamma}(j,k)$ was already defined in (4.1). If we want to recover j,k from ℓ at $\mathbf{\Gamma}_\ell, 1 \le \ell \le m$, we introduce

$$K(i) := \frac{(n-2)(n-1)}{2} - \frac{(n-i)(n-i+1)}{2} + 1, \quad 1 < i < n,$$

and determine j,k by

$$j = \arg \max_{1<i<n} K(i) \le \ell, \quad k = j + 1 + (\ell - K(j)).$$

Let us make a little example with $n = 6$. Then, $(2,3),(2,4),(2,5),(2,6)$; $(3,4),(3,5),(3,6);(4,5),(4,5);(5,6)$ correspond to $k = 1,2,3,4,5,6,7,8,9,10$. Let us try to find (j,k) [$=$ $(3,5)$] from $\mathbf{\Gamma}_6$, i. e. we have $\ell = 6$. We see that $K(3) = 5 \le \ell$, $K(4) = 8 > \ell$. Hence, $j = 3$, $k = j + 1 + (\ell - K(j)) = 5$, as required.

We make a comparative flop count for various cases of ordinary and Fast Givens transformation in Table 6.1.

In Table 6.1, we used A for additions, M for multiplications, F for flops. The Fast formulas in all cases are based on the fact that either c or s is real. Thus, we have to compare the counts in lines 3. and 5. in Table 6.1. The savings deduced from lines 2. and 4. are unrealistic. Thus, the (relative) saving is optimal in the real case, but moderate in the complex and quaternion case. The question is whether there is a strategy for choosing the diagonal elements d_1, d_2 such that the quantities u, v become real. We could choose $(d_1, d_2) = (a_{11}, a_{21})$ in case both quantities are not zero. Then, the first column of $\mathbf{B}$ would consist of two ones and both u, v would be real with $v = 1$. This would reduce (4.21) to $3n/7$ q-flops, however, by the cost of a premultiplication with $2n$ q-flops, together $(2 + 3/7)n$ q-flops. The direct evaluation needs $(2 + 2/7)n$ q-flops.

7. Stability considerations. The three papers by Gentleman, Hammarling, Rath, [6],[9],[14], all contain investigations on stability of the Fast Givens method. If we consider the computation of $b'_{jk}, j = 1,2,\ k = 1,2,\ldots,n$ in formula (4.21), we combine the two vectors b'_{1k}, b'_{2k} to one vector $b' := (b'_k), k = 1,2,\ldots,2n$, and compare that with the computed vector $\bar{b}' := (\bar{b}'_k)$. We also combine the two given vectors b_{1k}, b_{2k} to one vector b. Then, the estimates given are all of the form

$$||b' - \bar{b}'||_2 \le c\,\mathbf{m}||\,b\,||_2 \tag{7.1}$$

where the constant c is moderate and varies depending on specific assumptions and where $\mathbf{m}$ is the machine precision. In all cases the authors refer to Wilkinson, [19, p. 134]. However, it seems a little easier to follow the concept of the *condition of a function* which one can find e. g. in Demmel's book, [3, p. 5].

The stability is treated as the problem of evaluating a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at a known, computed neighbor $\bar{x} := x + h_x$ of an unknown value x, where h_x is assumed to be a small perturbation of x. For a function $f : \mathbb{R}^m \to \mathbb{R}$ let $x = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ be the true argument and $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)^{\mathrm{T}}$ a disturbed argument and define $e_{x_j} := (x_j - \bar{x}_j)/x_j$, the relative error of component number $j = 1, 2, \ldots, m$. We define

$$f_j(x, \bar{x}) := f(\bar{x}_1, \bar{x}_2, \ldots \bar{x}_j, x_{j+1}, \ldots, x_m), \quad j = 0, 1, \ldots, m.$$

In particular, $f_0(x, \bar{x}) = f(x)$ and $f_m(x, \bar{x}) = f(\bar{x})$. Then,

$$G(f)(x) := \frac{f(x) - f(\bar{x})}{f(x)} = \sum_{j=1}^{m} \frac{f_{j-1}(x, \bar{x}) - f_j(x, \bar{x})}{x_j - \bar{x}_j} \, \frac{x_j}{f(x)} \, \frac{x_j - \bar{x}_j}{x_j}. \tag{7.2}$$

And we define the condition of f at x by the vector

$$\begin{aligned} \kappa(f)(x) &:= \left(f_{x_1}(x) \frac{x_1}{f(x)}, f_{x_2}(x) \frac{x_2}{f(x)}, \ldots, f_{x_m}(x) \frac{x_m}{f(x)} \right)^{\mathrm{T}} \\ &=: (\kappa_1, \kappa_2, \ldots, \kappa_m)^{\mathrm{T}}, \end{aligned} \tag{7.3}$$

where f_{x_j} denotes the partial derivative of f in direction $x_j, j = 1, 2, \ldots, m$. According to (7.2), component j of $\kappa(f)(x)$ is the amplification factor of the relative error e_{x_j} or

$$G(f)(x) \approx \sum_{j=1}^{m} \kappa_j e_{x_j}. \tag{7.4}$$

Finally, if we have a function $f : \mathbb{R}^m \to \mathbb{R}^n$, we define a condition vector for each component of f. The result will be an $(m \times n)$ matrix of condition numbers. From (7.4) we can deduce

$$|G(f)(x)| \approx |\sum_{j=1}^{m} \kappa_j e_{x_j}| \le m \max_j |\kappa_j| \max_j |e_{x_j}|. \tag{7.5}$$

Let us apply this to the multiplication of two quaternions. The formula for multiplication is given in (1.1). Each component of the product is a function of eight variables $a := (a_1, a_2, a_3, a_4)$, $b := (a_5, a_6, a_7, a_8)$. For the first component of the product (1.1) defined by $p(a, b) := a_1 a_5 - a_2 a_6 - a_3 a_7 - a_4 a_8$ we obtain from (7.5)

$$|\sum_{j=1}^{8} \kappa_j e_{a_j}| \le \frac{8}{p} \max_{1 \le j \le 4} |a_j a_{j+4}| \max_{1 \le j \le 8} (|e_{a_j}|) \le \frac{8\mathbf{m}}{p} \max_{1 \le j \le 4} |a_j a_{j+4}|,$$

where $\mathbf{m}$ stands for the machine precision. Since p is an ordinary scalar product, it may have the well known fallacies of a scalar product, or in other words, the four components of the product of two quaternions must be computed carefully.

8. A numerical example. Given a matrix $\mathbf{A} \in \mathbb{H}^{5 \times 5} =: (a_{jk}), j, k = 1, 2, \ldots, 5$ defined by

$$\mathbf{A} := \left(\begin{array}{rrrrrrrrrr} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline 5 & 4 & 3 & -3 & -2 & 0 & 0 & 1 & 4 & -5 \\ 0 & 2 & 3 & 4 & 0 & 0 & 0 & -4 & -1 & 0 \\ -4 & 5 & 4 & 2 & -4 & -3 & 3 & -4 & 2 & 4 \\ -4 & 3 & -4 & -1 & 0 & 4 & 2 & 1 & -4 & 3 \end{array} \right.$$

ETNA
Kent State University
etna@mcs.kent.edu

a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}
3	3	−4	−2	2	4	1	−1	3	3	−5	−2	5	−2	−5
−3	−3	1	0	−4	−5	4	−1	4	−1	−1	−2	−2	−4	0
2	−3	−3	3	3	−4	0	−1	3	−1	3	0	3	−2	1
−3	0	4	−2	0	2	0	4	4	−3	−2	−3	4	2	2

The reduction of $\mathbf{A}$ to upper Hessenberg form $\mathbf{H} =: (h_{jk}), j, k = 1, 2, \ldots, 5$ by the prescribed Fast Givens transformation with four matrices and using the identity matrix as starting diagonal matrix $\mathbf{D}$ yields

$\mathbf{H} :=$

h_{11}	h_{12}	h_{13}	h_{14}	h_{15}
5	−0.74775650110597	−3.95327600209997	2.96961127198873	0.29808603448429
0	−6.51410951925005	−2.20380757846644	3.40622786697461	−3.00578422682992
−4	2.25764943603147	−3.83464360352797	2.76338195772684	2.39657240898337
−4	2.67466748472519	3.35725239384719	1.83028985634685	2.45382201023755

h_{21}	h_{22}	h_{23}	h_{24}	h_{25}
−6.72980850995369	0.08333333333333	−5.81010025559997	2.97160998711501	−0.41862959053221
6.72980850995369	−3.30624483043838	2.55855956052311	1.40121322988457	2.99852081752252
−4.48653900663580	−2.23945409429280	−2.95394126306036	−1.48234496331670	0.94736829485381
6.72980850995369	0.66583953680728	−2.64698611503065	−4.65096304953245	3.03127807183587

h_{31}	h_{32}	h_{33}	h_{34}	h_{35}
0	−5.60809863079113	−4.42335169964168	2.81328889050477	−1.76867759787944
0	2.57030463609747	−2.62149689810323	0.99739994003402	−0.66532778056974
0	−2.54116573298713	−4.13525389269467	−1.02170998081773	3.31842681724252
0	6.49645510300699	1.87303821660327	−0.94909428323479	3.12994058167427

h_{41}	h_{42}	h_{43}	h_{44}	h_{45}
0	0	−0.14684700502573	−2.45117954729734	−2.28477951278127
0	0	5.39325768134285	−1.64396407983639	−1.47533658343006
0	0	2.85553112722675	−3.48282028844388	−2.68719252790753
0	0	4.55960852186855	2.30234114446700	1.42019804718412

h_{51}	h_{52}	h_{53}	h_{54}	h_{55}
0	0	0	1.92401313127372	0.79119791360569
0	0	0	−2.68560326801966	−4.27095881717863
0	0	0	−6.01180247912326	−1.34098991661015
0	0	0	−4.21206594139542	1.54951849179966

It should be pointed out, that a change of the strategy or a change of the diagonal matrix $\mathbf{D}$ at the start would change the result. The matrix $\mathbf{D}$ could be used to scale $\mathbf{A}$. If we choose $\mathbf{D} := \sqrt{\mathrm{diag}(\mathbf{A})}$ where $\sqrt{\ }$ is applied elementwise, with the exception of zero diagonal elements of $\mathbf{A}$ where one defines the corresponding diagonal element of $\mathbf{D}$ to be one. Then, the effect is, that all non zero diagonal elements of $\mathbf{B}_0$ (see (6.6)) will have absolute value one. This could be an important argument, not to use the identity matrix $\mathbf{I}$ for $\mathbf{D}$ in the beginning.

9. Conclusion.

We have developed two techniques for applying the Fast Givens transform to matrices with quaternion valued matrices. The first technique follows from what is known for the real case, namely, to use two different matrices according to the size of the quantity $|s|^2 \in [0, 1]$. In this case we also show, that in contrast to the standard opinion it is also possible to store the essential information in only one parameter. As in the real case, two options are available, namely either to use the squares of the diagonal elements or to use the diagonal elements directly. In the first case square roots are only needed in the very end. Since the diagonal elements may be quaternions there may be the necessity to compute square roots of quaternions. Therefore, we have added an appendix in which we show how to compute roots from quaternions.

The second technique involves four matrices and in each step we select that matrix which has the smallest condition number. We have shown, that this technique yields smaller condition numbers in comparison with the standard technique with only two matrices. This technique would even be new for the real case. There is also some disadvantage. The diagonal elements do not stay real (in case we do not work with real matrix entries) if we would start with real diagonal elements, and in computing the new matrix forms more algebraic work is needed. However, when computing the Hessenberg form which is our main application the algebraic work for transforming the original matrix is $O(n^3)$, whereas the additional work is only $O(n^2)$. This additional work may be applied well, when using the start matrix $\mathbf{D}$ for scaling the given matrix $\mathbf{A}$. More details are given at the end of Section 8.

There is one general drawback when using Fast Givens transformation for complex or quaternion valued matrices. This can be seen from looking at Table 6.1. There is only a gain factor of 1.125 with respect to the flop count over the classical Givens transformation when applied to quaternions. We also see that the Fast Givens transformation in the quaternion case needs 16 times as many flops as the Fast Givens transformation applied to real matrices.

Appendix. Roots of quaternions. In Section 6, there is the possibility of using quaternion entries in the diagonal matrix $\mathbf{D}$. In step (6.7), there is the task of computing square roots of quaternions. Since this is not obvious, we will give some hints for computing roots in general. For a given $a \in \mathbb{H}\backslash\{0\}$ and fixed $n \geq 2$ let us consider the polynomial $p : \mathbb{H} \to \mathbb{H}$ defined by

$$p(z) := z^n - a, \quad n \in \mathbb{N}, \quad n \geq 2, \quad a \in \mathbb{H}\backslash\{0\}. \tag{A.1}$$

Any $z \in \mathbb{H}$ with $p(z) = 0$ will be called a *root* (sometimes also n-*root*) of a or a *zero* of p. We will use the already introduced notion of *equivalence* of quaternions. In particular we refer to Lemma 1.1. We will see that p defined in (A.1) has always n zeros. But it may have even infinitely many zeros. In order to see this, let z be a root of a. Then, for any $h \in \mathbb{H}\backslash\{0\}$ we have $h^{-1}p(z)h = (h^{-1}zh)^n - h^{-1}ah = 0$. In particular, if a is real, we obtain $p(h^{-1}zh) = 0$. Thus, with z also $h^{-1}zh$ is a root of $a \in \mathbb{R}$. Therefore, if z is not real (take $a = -1$ as an example) there will be infinitely many roots of a.

LEMMA A.1. *Let p of (A.1) be given and $\tilde{a}$ be the corresponding complex representative of $[a]$ with representation $a = h\tilde{a}h^{-1}$. Let $\hat{z} := \tilde{a}^{1/n}$. Then $z := h\hat{z}h^{-1}$ is a root of a and $\Im\hat{z} \geq 0$.*

Proof. Apparently we have $\hat{z}^n - \tilde{a} = \hat{z}^n - h^{-1}ah = 0$. Multiplying from the left by h, from the right by h^{-1} we obtain $h\hat{z}^n h^{-1} - a = (h\hat{z}h^{-1})^n - a = z^n - a = 0$, hence, z is a root of a. Since $\Im\{\tilde{a}\} \geq 0$ the quantity $\tilde{a}$ has a polar representation of the form $\tilde{a} = |\tilde{a}|(\cos\varphi + \mathrm{i}\sin\varphi)$ with $0 \leq \varphi \leq \pi$. Therefore, $\Im\hat{z} = |\tilde{a}|^{1/n}\sin\frac{\varphi}{n} \geq 0$. ∎

Let us keep the notation of the lemma. We assume, that we are able to compute all n-roots of a complex number. Then, a root of a can be found by first computing $\hat{z} := \tilde{a}^{1/n}$ and then, by finding $h \in \mathbb{H}\backslash\{0\}$ with $a = h\tilde{a}h^{-1}$. This is equivalent to

$$h\tilde{a} - ah = 0. \tag{A.2}$$

Let us put $a := (a_1, a_2, a_3, a_4)$, $\tilde{a} := (a_1, \tilde{a_2}, 0, 0)$, $h := (h_1, h_2, h_3, h_4)$ and recall that $\tilde{a_2} = \sqrt{a_2^2 + a_3^2 + a_4^2}$. Then, (A.2) can be written as a real, homogeneous linear (4×4) system

$$\text{(A.3)}\qquad \mathbf{A}h = \mathbf{0} \text{ with } \mathbf{A} = \begin{pmatrix} 0 & -a_2+\widetilde{a_2} & -a_3 & -a_4 \\ a_2-\widetilde{a_2} & 0 & -a_4 & a_3 \\ a_3 & a_4 & 0 & -a_2-\widetilde{a_2} \\ a_4 & -a_3 & a_2+\widetilde{a_2} & 0 \end{pmatrix} \in \mathbb{R}^{4\times 4},$$

where $\mathbf{A}$ is of rank 2 if $a \notin \mathbb{R}$. For the details see Janovská and Opfer [10]. If a solution $h \in \mathbb{H}\backslash\{0\}$ has been found, the wanted root is (according to Lemma A.1) $z = h\hat{z}h^{-1}$. In case a is real or complex there is no problem finding $\sqrt[n]{a}$. We may, therefore, assume that $|a_3| + |a_4| > 0$. This is equivalent to $\widetilde{a_2} > |a_2|$.

One solution h of (A.3) is:

$$\text{(A.4)}\qquad h := \begin{pmatrix} h_1 \\ h_2 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} := \begin{pmatrix} \dfrac{a_3-a_4}{\widetilde{a_2}-a_2} \\ \dfrac{a_3+a_4}{\widetilde{a_2}-a_2} \end{pmatrix}.$$

Another, independent solution η is

$$\text{(A.5)}\qquad \begin{aligned} \eta &:= \begin{pmatrix} 1 \\ 1 \\ h_3 \\ h_4 \end{pmatrix}, \begin{pmatrix} h_3 \\ h_4 \end{pmatrix} := \begin{pmatrix} \dfrac{a_3-a_4}{\widetilde{a_2}+a_2} \\ \dfrac{a_3+a_4}{\widetilde{a_2}+a_2} \end{pmatrix} \text{ for } a_4 \neq 0; \\ \eta &:= \begin{pmatrix} 1/h_3 \\ 2 \\ 2h_3 \\ 1 \end{pmatrix} \text{ in case } a_4 = 0. \end{aligned}$$

By this technique all n-roots of the complex $\tilde{a}$ produce a quaternion root of a, therefore, there are n such roots. Let us treat a little example with $n = 2$, $a := (1, 0, 3, 4)$. Then, $\tilde{a} = (1, 5, 0, 0)$ and $\hat{z} = \sqrt{1 + 5\mathrm{i}} = 1.7463 + 1.4316\mathrm{i}$. The solution (A.4) of (A.3) is $h = (-0.2, 1.4, 1, 1)$ and the final solution is $z = h\hat{z}h^{-1} = (1.7463, 0, 0.8590, 1.1453)$. We mention a just released article by Kuba[12] where in particular polar representations of roots of quaternions are given. However, there are no explicit formulas to find the roots in the presented cartesian form.

Appendix A. Why we do need quaternion arithmetic. There is the well known isomorphism, see van der Waerden[18, p. 55]

$$\imath : \mathbb{H} \to \mathbb{C}^{2\times 2},$$

which is defined by

$$\imath(x_1, x_2, x_3, x_3) = \begin{pmatrix} x_1 + x_2\mathrm{i} & x_3 + x_4\mathrm{i} \\ -x_3 + x_4\mathrm{i} & x_1 - x_2\mathrm{i} \end{pmatrix}.$$

Thus, in pure mathematical terms, we could transfer all quaternion problems to complex matrix problems. However, we will see by the smallest possible example, that the results are disastrous from a numerical point of view. Let $a = (a_1, a_2, a_3, a_4)$, $b = (b_1, b_2, b_3, b_4) \in \mathbb{H}$ and $\alpha_1 = a_1 + a_2\mathrm{i}$, $\alpha_2 = a_3 + a_4\mathrm{i}, \beta_1 = b_1 + b_2\mathrm{i}, \beta_2 = b_3 + b_4\mathrm{i} \in \mathbb{C}$ be given and define

$$\mathbf{A} := \begin{pmatrix} a & 0 \\ b & a \end{pmatrix} \in \mathbb{H}^{2\times 2}.$$

Then, the corresponding complex matrix is

$$\tilde{\mathbf{A}} := \imath(\mathbf{A}) := \begin{pmatrix} \imath(a) & \imath(0) \\ \imath(b) & \imath(a) \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & 0 & 0 \\ -\overline{\alpha_2} & \overline{\alpha_1} & 0 & 0 \\ \beta_1 & \beta_2 & \alpha_1 & \alpha_2 \\ -\overline{\beta_2} & \overline{\beta_1} & -\overline{\alpha_2} & \overline{\alpha_1} \end{pmatrix} \in \mathbb{C}^{4\times 4}.$$

All eigenvalues of $\mathbf{A}$ are $[a]$ and $\tilde{a} := a_1 + \sqrt{a_2^2 + a_3^2 + a_4^2}\,\mathrm{i}$ is its complex representative. The four eigenvalues of $\tilde{\mathbf{A}}$ are $\tilde{a}, \overline{\tilde{a}}$, each double. If we choose $a = (1, 0, 3, 4)$, then $\tilde{a} = 1+5\mathrm{i}$ and `eig(Ã)` of MATLAB (version 5.3) produces

$$\begin{aligned} &1.00000\,00263\,4178 - 4.99999\,99947\,3164\,\mathrm{i}, \\ &0.99999\,99736\,5822 - 5.00000\,00052\,6835\,\mathrm{i}, \\ &0.99999\,99635\,5056 + 4.99999\,99733\,2945\,\mathrm{i}, \\ &1.00000\,00364\,4944 + 5.00000\,00266\,7055\,\mathrm{i}. \end{aligned}$$

There is a loss of 6-7 decimal digits. The same is true in newer versions of MATLAB. Even the pairs of conjugate eigenvalues are not recognized. For matrices only a little larger, the same technique produces results with even fewer correct places. Therefore, the idea to go to complex matrices if quaternions should be treated numerically, has in general to be avoided because of the danger of significant error propagation.

There is also a (4×4) real matrix analogue for quaternions. See Gürlebeck and Sprößig[8, p. 12]. An application to the same example displays errors of the same nature.

Acknowledgment. The authors acknowledge with pleasure the support of the Grant Agency of the Czech Republic (grant No. 201/02/0844, grant No. 201/02/0595) and support of the project MSM 223400007 of the Czech Ministry of Education.

REFERENCES

[1] A. A. Anda and H. Park, *Fast plane rotations with dynamic scaling*, SIAM J. Matrix Anal. Appl. 15 (1994), pp. 162–174.

[2] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *A quaternion QR algorithm*, Numer. Math. 55 (1989), pp. 83–95.

[3] J. W. Demmel, *Applied numerical linear algebra*, SIAM, Philadelphia, 1997.

[4] J. J. Dongarra, J. R. Gabriel, D. D. Koelling, and J. H. Wilkinson, *Solving the secular equation including spin orbit coupling for systems with inversion and time reversal symmetry*, J. Comput. Phys., 54 (1984), pp. 278–288.

[5] ———, *The eigenvalue problem for hermitian matrices with time reversal symmetry*, Linear Algebra Appl., 60 (1984), pp. 27–42.

[6] M. Gentleman, *Least squares computations by Givens transformations without square roots*, J. Inst. Math. Appl. 12 (1973), pp. 329–336.

[7] J. W. Givens, *Numerical computation of the characteristic values of a real symmetric matrix*, Oak Ridge Natl. Lab. Report: ORNL 1574, 1954.

[8] K. Gürlebeck and W. Sprössig, *Quaternionic analysis and elliptic boundary value problems*, Birkhäuser, Basel, 1990, ISNM 89.

[9] S. Hammarling, *A note on modifications to the Givens plane rotations*, J. Inst. Math. Appl. 13 (1974), pp. 215–218.

[10] D. Janovská and G. Opfer, *Givens' transformation applied to quaternion valued vectors*, BIT Numerical Mathematics, 43 (2003), pp. 991–1002. [typographical error in title caused by publisher]

[11] D. Janovská and G. Opfer, *Givens' reduction of a quaternion-valued matrix to upper Hessenberg form*, to appear in ENUMAT 2005.

[12] G. Kuba, *Wurzelziehen aus Quaternionen*, Mitt. Math. Ges. Hamburg 23/1 (2004), pp. 81–94. (in German: Finding the roots of quaternions)

ETNA
Kent State University
etna@mcs.kent.edu

[13] J. B. KUIPERS, *Quaternions and rotation sequences, a primer with applications to orbits, aerospace, and virtual reality*, Princeton University Press, Princeton, NJ, 1999.
[14] W. RATH, *Givens rotations for orthogonal similarity transformations*, Numer. Math., 40 (1982), pp. 47–56.
[15] N. RÖSCH, *Time-reversal symmetry, Kramers' degeneracy and the algebraic eigenvalue problem*, Chemical Physics, 80 (1983), pp. 1–5.
[16] H. R. SCHWARZ AND N. KÖCKLER, *Numerische Mathematik*, 5. Aufl. Teubner, Stuttgart, 2004.
[17] G. W. STEWART, *The economical storage of plane rotations*, Numer. Math. 25 (1976), pp. 137–138.
[18] B. L. VAN DER WAERDEN, *Algebra I*, 5. Aufl., Springer, Berlin, Göttingen, Heidelberg, 1960.
[19] J. H. WILKINSON, *The algebraic eigenvalue problem*, Oxford University Press, Oxford, 1965.
[20] L. XU, *A fast Givens transformation for complex matrix*, J. East China Norm. Univ. Sci. Ed. 3 (1988), pp. 15–21. (Chinese, English Summary, private translation into German by L. Xia)
[21] F. ZHANG, *Quaternions and matrices of quaternions*, Linear Algebra Appl. 251 (1997), pp. 21–57.