



# Roundoff error analysis of orthogonal factorizations of upper Hessenberg rectangular matrices

Mario Arioli

February 26, 2008

© Science and Technology Facilities Council

Enquires about copyright, reproduction and requests for additional copies of this report should be addressed to:

Library and Information Services  
SFTC Rutherford Appleton Laboratory  
Harwell Science and Innovation Campus  
Didcot  
OX11 0QX  
UK  
Tel: +44 (0)1235 445384  
Fax: +44(0)1235 446403  
Email: [library@rl.ac.uk](mailto:library@rl.ac.uk)

The STFC ePublication archive (epubs), recording the scientific output of the Chilbolton, Daresbury, and Rutherford Appleton Laboratories is available online at:  
<http://epubs.cclrc.ac.uk/>

**ISSN 1358-6254**

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigation

# Roundoff error analysis of orthogonal factorizations of upper Hessember rectangular matrices

M. Arioli<sup>1</sup>

## ABSTRACT

Krylov space methods minimizing the 2-norm of the residual (GMRES and MINRES are classical examples) requires the solution of relative small linear least squares problems. The matrix modelling this least square problem is of upper Hessenberg type and the right-hand side is a multiple of the first column of the identity. We specialize some classical roundoff results for Givens (Householder) method to this case pointing out some peculiarities that are useful in the error analysis of Krylov methods such as GMRES, MINRES, and Flexible GMRES.

**Keywords:** Upper Hessember matrix, Least-square problems, mixed precision arithmetic, round-off error.

**AMS(MOS) subject classifications:** 65F05, 65F50, 65F10, 65G50

---

Current reports available by anonymous ftp to <ftp.numerical.rl.ac.uk> in directory pub/reports.

<sup>1</sup> [m.arioli@rl.ac.uk](mailto:m.arioli@rl.ac.uk), [i.s.duff@rl.ac.uk](mailto:i.s.duff@rl.ac.uk) Rutherford Appleton Laboratory,

The work was supported by EPSRC grant GR/S42170/01.

Computational Science and Engineering Department  
Atlas Centre  
Rutherford Appleton Laboratory  
Oxon OX11 0QX

March 26, 2008

# Contents

1	Introduction	1
2	Roundoff error analysis	2
3	Convergence problems for GMRES	4
4	Conclusions	5

# 1 Introduction

We consider linear least-squares problems

$$\min_y \|b - Hy\|_2 \quad (1.1)$$

where  $H \in \mathbb{R}^{(k+1) \times k}$  is an upper Hessenberg matrix. In particular, we are interested in the case when the vector  $b$  is a simple multiple of  $e_1$  the first column of the identity matrix of order  $k + 1$ . This special case arises from the use of the Arnoldi process in GMRES and Flexible GMRES methods (see Arioli and Fassino (1996), Drkosova, Geenbaum, Rozložník and Strakoš (1995), Paige, Rozložník and Strakoš (2006), and Saad (2003)). The computation of  $y$  is performed in two stages

The Givens (or the Householder) algorithm computes elementary rotations (or reflections in the case of Householder)  $G^{(i)}$  in order to reduce the matrix  $H$  to the upper triangular form  $U$ :

$$G^{(i)} = \begin{bmatrix} I_{i-1} & & & \\ & c_i & s_i & \\ & -s_i & c_i & \\ & & & I_{k-i-1} \end{bmatrix} \quad i = 1, \dots, k$$

$$c_i = \frac{(h)_{i,i}}{\sqrt{(h)_{i,i}^2 + (h)_{i+1,i}^2}} \quad \text{and} \quad s_i = \frac{(h)_{i+1,i}}{\sqrt{(h)_{i,i}^2 + (h)_{i+1,i}^2}}$$

Then, we compute

$$q = \prod_{i=1}^k G^{(i)} b = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix},$$

and solve the system

$$Uy = q_1. \quad (1.2)$$

We must immediately note that

$$\|b - Hy\| = |q_2|. \quad (1.3)$$

Moreover, if  $b = \beta e_1$  we have that

$$q_i = \beta \prod_{j=1}^i (-1)^j s_j c_j \quad i = 1, \dots, k \quad (1.4)$$

$$q_k = \beta \prod_{j=1}^k (-1)^j s_j \quad (1.5)$$

Therefore, if  $s_i < 1$  for all  $i$  then the vector  $q$  entries are decreasing in absolute values when  $i$  increases. Finally, we point out that in this case and when  $H \in \mathbb{R}^{n \times n}$ , (1.1) is consistent and the residual is zero.

An other important related problem is the updating of  $H$  by the addition of a column and a row that leave the new  $H$  still upper Hessenberg. Let  $H^{(i+1)}$  the updated matrix obtained from  $H^{(i)} = H$ , i.e.:

$$H^{(i+1)} = \begin{bmatrix} H^{(i)} & h_1 \\ 0 & h_2 \end{bmatrix} \quad (1.6)$$

Moreover, if  $b = \beta e_1$  and  $s_i < 1$  for all  $i$  the residual of the updated problem decreases.

## 2 Roundoff error analysis

In the following, we will denote by  $c_p(n, j)$  functions that depend only on the dimension  $n$  and the integer  $j$ . We will avoid a precise formulation of these dependencies, but we assume that each  $c_p(n, j)$  grows moderately with  $n$  and  $j$ . Finally, if  $B \in \mathbf{R}^{n \times m}$ ,  $n \geq m$  is a full rank matrix, we denote by  $\kappa(B) = \|B\| \|B^+\|$  its spectral condition number where  $B^+ = (B^T B)^{-1} B$ . For all matrices and vectors we denote by  $|B|$  the matrix or vector of the absolute values.

**Lemma 2.1.** *Applying the QR factorization with Givens rotations to solve*

$$\min_y \|\beta e_1 - Hy\| \quad (2.7)$$

using finite-precision arithmetic conforming to IEEE standard with relative precision  $\varepsilon$  and under the condition

$$0.1 > c_0(n)\varepsilon \kappa(\bar{H}_k) + \mathcal{O}(\varepsilon^2) \quad \forall k, \quad (2.8)$$

there exist an orthonormal matrix  $\hat{G}^{[k]}$ , a vector  $g^{[k]}$ , and an upper Hessenberg matrix  $\Delta H$  such that the computed value  $\bar{y}_k$  satisfies the following relations

$$\begin{cases} \bar{y}_k = \arg \min_y \|\hat{G}^{[k]}(\bar{\beta} e_1 + g^{[k]} - (\bar{H}_k + \Delta H_k)y)\|, \\ \|\Delta H_k\| \leq c_1(k, 1)\varepsilon \|\bar{H}_k\| + \mathcal{O}(\varepsilon^2) \text{ and } \|g^{[k]}\| \leq c_2(k, 1)\varepsilon \bar{\beta} + \mathcal{O}(\varepsilon^2). \end{cases} \quad (2.9)$$

Moreover, the residuals

$$\alpha_k = \|\hat{G}^{[k]}(\bar{\beta} e_1 + g^{[k]} - (\bar{H}_k + \Delta H_k)\bar{y}_k)\|,$$

satisfy the equations

$$\begin{cases} \alpha_k = \bar{\beta} \left( \prod_{j=0}^k |\bar{s}_j| \right) \left( \prod_{j=0}^k (1 + \zeta_j) \right) \\ |\zeta_j| \leq \varepsilon \quad \forall j. \end{cases} \quad (2.10)$$

If

$$|\bar{s}_j| < 1 - \varepsilon, \quad (2.11)$$

we have that  $\alpha_k$  is strictly decreasing to zero and  $\alpha_{\hat{k}} = 0$  for some value of  $\hat{k} \leq n$ .

*Proof.* The floating-point computation of the matrices  $G^{(i)}$  gives

$$fl(G^{(i)}) = \tilde{G}^{(i)} = \begin{bmatrix} I_{i-1} & & & \\ & \bar{c}_i & -\bar{s}_i & \\ & \bar{s}_i & \bar{c}_i & \\ & & & I_{n-i-1} \end{bmatrix} \quad i = 1, \dots, k$$

$$\bar{c}_i = fl\left(\frac{(\bar{H}_k)_{i,i}}{\sqrt{(\bar{H}_k)_{i,i}^2 + (\bar{H}_k)_{i+1,i}^2}}\right) \quad \text{and} \quad \bar{s}_i = fl\left(\frac{(\bar{H}_k)_{i+1,i}}{\sqrt{(\bar{H}_k)_{i,i}^2 + (\bar{H}_k)_{i+1,i}^2}}\right)$$

The  $\bar{G}^{(i)}$  matrices are also applied to the vector  $\bar{\beta}e_1$  and, from the error analysis presented by Wilkinson (1965) (see also Gentleman (1973), Golub (1965), and Higham (2002)), the floating-point arithmetic will produce an exact orthogonal matrix  $G^{[k]}$  such that

$$\begin{aligned} fl(\bar{G}^{(k)} \dots \bar{G}^{(1)} \bar{\beta}e_1) &= G^{[k]}(\bar{\beta}e_1 + g^{[k]}) \\ fl(\bar{G}^{(k)} \dots \bar{G}^{(1)} \bar{H}_k) &= G^{[k]}(\bar{H}_k + \Delta\bar{H}_k^{(1)}), \\ \|\Delta\bar{H}_k^{(1)}\| &\leq c_3(k, 1)\varepsilon \|\bar{H}_k\| + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Second, the  $\bar{y}_k$  vector is computed by solving the upper triangular system. The standard backward substitution algorithm will introduce an additional perturbation  $\Delta\bar{H}_k^{(2)}$  of  $\bar{H}_k$  but will leave the perturbation  $g^{[k]}$  untouched. The perturbation  $\Delta\bar{H}_k^{(2)}$  will also have the same upper Hessenberg structure of  $\bar{H}_k$  and

$$\|\Delta\bar{H}_k^{(2)}\| \leq c_4(k, 1)\varepsilon \|\bar{H}_k\| + \mathcal{O}(\varepsilon^2).$$

This follows from the upper triangular structure of the perturbation to  $U_k$  induced by the backward substitution algorithm, and from the structure and orthogonality of  $G^{[k]}$ . Finally, we point out that in the relations (2.9) we have see (2.8)

$$\begin{cases} \Delta\bar{H}_k = \Delta\bar{H}_k^{(1)} + \Delta\bar{H}_k^{(2)}, & \text{and} \\ c_1(k, 1) = c_3(k, 1) + c_4(k, 1). \end{cases}$$

where  $c_1(k, 1)$  is the constant of (2.8) Moreover, because of the special structure of  $\bar{\beta}e_1$  and the orthogonality of  $G^{[k]}$  we have

$$g_j^{[k]} = 0 \quad j = k+1, \dots, n,$$

and, denoting by  $\bar{h}^{[k]} = fl(\bar{G}^{(k)} \dots \bar{G}^{(1)} \bar{\beta}e_1)$ , we have

$$\begin{aligned} \alpha_k &= \|G^{[k]}(\bar{\beta}e_1 + g^{[k]} - (\bar{H}_k + \Delta\bar{H}_k)\bar{y}_k)\| \\ &= \|(G^{[k]}(\bar{\beta}e_1 + g^{[k]}))_{k+1}\| = |\bar{h}_{k+1}^{[k]}|. \end{aligned} \tag{2.12}$$

A direct analysis of  $\bar{h}^{[k]}$  shows that

$$\begin{cases} \bar{h}_1^{[1]} &= \bar{\beta}\bar{c}_1(1 + \mu_1) & |\mu_1| \leq \varepsilon \\ \bar{h}_2^{[1]} &= \bar{\beta}\bar{s}_1(1 + \zeta_1) & |\zeta_1| \leq \varepsilon \\ \bar{h}_j^{[k]} &= \bar{h}_j^{[k-1]} & j = 1, \dots, k-1, \quad k \geq 2, \\ \bar{h}_k^{[k]} &= \bar{h}_k^{[k-1]}\bar{c}_k(1 + \mu_k) & |\mu_k| \leq \varepsilon \\ \bar{h}_{k+1}^{[k]} &= \bar{h}_k^{[k-1]}\bar{s}_k(1 + \zeta_k) & |\zeta_k| \leq \varepsilon. \end{cases}$$

Formula (2.10) follows immediately by recurrence. □

### 3 Convergence problems for GMRES

The analysis presented in the previous sections is related to the Arnoldi process and in particular to the GMRES algorithm and its variants such as Flexible GMRES, see Saad (2003), Arioli and Fassino (1996), Drkosova et al. (1995), and Paige et al. (2006).

The Arnoldi algorithm applied to the matrix  $A$  computes, starting with an arbitrary vector  $w_1$  and in exact arithmetic, an orthonormal matrix  $W$  and an upper Hessenberg  $H$  with entries  $h_{i+1,i} \geq 0$   $i = 1, \dots, n$ . In particular, the first column of  $W$  is  $w_1$ .

The decomposition

$$AW = WH \tag{3.13}$$

is one of many possible decompositions that can be computed changing the initial vector  $w_1$ .

Let  $E_k \in \mathbb{R}^{n \times k}$  be the matrix of the first  $k$  column of the  $n \times n$  identity. The GMRES method can be seen as a truncation of (3.13)

$$AWE_k = WHE_k = W_{k+1}H_k. \tag{3.14}$$

From the previous analysis it is straight forward to see that GMRES residual can stagnate if and only if the first row of  $H$  has its first  $n - 1$  entries equal to zero. In this case all the residuals will be equal to the norm of  $w_1$  until step  $n$  when the final residual collapses to zero if  $A$  is non singular.

Reversely, if the residual at step  $k$  does not decrease then the value of  $s_k$  in the Givens matrix  $G^{(k)}$  will be equal to 1 and  $G^{(k)}$  is the permutation matrix swapping rows  $k$  and  $k + 1$ . The value of  $s_k = 1$  if and only if  $h_{kk} = 0$  and  $h_{k+1,k} > 0$ , thus, if all residuals are equals the first  $n - 1$  entries of the first row of  $H$  must be zero.

We point out that the any upper Hessenberg matrix having this property is the permutation of an upper triangular matrix with positive entries on the main diagonal

$$H = PU \quad (U)_{ii} > 0 \quad i = 1, \dots, n \tag{3.15}$$

where  $P$  is the circulant shifting permutation matrix such that

$$\begin{aligned} Pe_i &= e_{i+1} & i = 1, \dots, n-1 \\ Pe_n &= e_1. \end{aligned}$$

Therefore, the manifold

$$\mathfrak{M} = \{A : A = WPUW^T, W^T W = I, U_{i,j} = 0 \quad i < j, U_{ii} > 0 \quad i = 1, \dots, n\}$$

is the manifold of all the matrices for which exists a vector  $w_1$  such that if we apply GMRES to the system

$$Ax = w_1$$

we will have convergence only after  $n$  steps. This generalizes the result presented by Nachtigal, Reddy, and Trefethen (1994) (see example C page 788).

Finally, owing the presence in (3.13) of both  $W$  and its transpose we can assume that  $W \in \mathcal{SO}(n)$  (Hall 2004, Rossmann 2002) the special Lie group of orthogonal matrices. Moreover,  $U \in \mathcal{U}$  where  $\mathcal{U}$  is the Lie Group of the upper triangular matrices with positive diagonal entries. Following Hall (2004) and Rossmann (2002), we can also write each matrix  $A$  in  $\mathfrak{M}$  as

$$A = e^S P e^V e^{-S} \tag{3.16}$$

with  $S \in \mathfrak{so}$  i.e.  $S^T = -S$  ( $S$  skew-symmetric), and  $V$  an upper triangular matrix with any assumption on the non singularity of  $V$ .

From Lemma 2.1, it is straight forward to see that in presence of roundoff the matrix  $H$  will be perturbed by small quantities in each entries. The resulting perturbed matrix  $\bar{H}$  will have small entries in the first row. However, the convergence will be still quite slow.

## 4 Conclusions

We have proved that the Givens algorithm applied to problem (2.7) is backward stable. The main result is useful in proving the backward stability of the GMRES (Flexible GMRES) algorithms: in particular equation (2.10) has a critical role in the proof of the backward stability of both GMRES and Flexible GMRES.

The backward stability will generate perturbation that will marginally influence the theoretical convergence of the residual  $\alpha_k$  to zero. Finally, we introduce a novel characterization of the class of nonsingular matrices for which GMRES will converge only after  $n$  steps.

## References

- Arioli, M. and Fassino, C. (1996), ‘Roundoff Error Analysis of Algorithms Based on Krylov Subspace Methods’, *BIT* **36**, 189–206.
- Drkosova, J., Geenbaum, A., Rozložník, M., and Strakoš, Z. (1995), ‘Numerical stability of GMRES method’, *BIT* **35**, 308–330.
- Gentleman, W. M. (1973), ‘Least squares computations by Givens transformations without square roots problems’, *J. Inst. Maths Applics* **12**, 329–336.
- Golub, G. H. (1965), ‘Numerical methods for solving linear least squares problems’, *Numer. Math.* **7**, 206–216.
- Hall, B. C. (2004), *Lie Groups, Lie Algebras, and Representations. An Elementary Introduction, Second Edition*, GTM 22, Springer, New York, USA.
- Higham, N. J. (2002), *Accuracy and Stability of Numerical Algorithms, Second Edition*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- Nachtigal, N. M., Reddy, S. C., and Trefethen, L. N. Paige, C., Rozložník, M. and Strakoš, Z. (2006), ‘How Fast are Nonsymmetric Matrix Iterations?’, *SIAM Journal on Matrix Analysis and Applications* **13**(1), 778–795.
- Paige, C., Rozložník, M. and Strakoš, Z. (2006), ‘Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES’, *SIAM Journal on Matrix Analysis and Applications* **28**(1), 264–284.
- Rossmann, W. (2002), *Lie Groups. An Introduction Through Linear Groups*, Oxford Graduated Texts in Mathematics,5, Oxford University Press, Oxford, UK.
- Saad, Y. (2003), *Iterative Methods for Sparse Linear Systems Second Edition*, Society for Industrial and Applied Mathematics.
- Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press.