ACCURATE SYMMETRIC EIGENREDUCTION BY A JACOBI METHOD

DISSERTATION
zur
Erlangung des Grades Dr. rer. nat.
des Fachbereichs Mathematik
der Fernuniversität – Gesamthochschule – Hagen

vorgelegt von
IVAN SLAPNIČAR
aus Split, Kroatien

Hagen 1992

Erstgutachter und Mentor: Prof. Dr. K. Veselić, Hagen
Zweitgutachter: Prof. Dr. J. Demmel, Berkeley

Acknowledgement

I would like to thank my mentor Prof. Dr. Krešimir Veselić for introducing me to the exciting field of relative error analysis, for devoting to me a lot of his time, and for sharing with me so many of his ideas.

I also thank my colleagues Eberhard Pietzsch, Zlatko Drmač und Xiaofeng Wang for the possibility to check my ideas in numerous discussions, and Prof. Dr. Jesse Barlow, Pennsylvania State University, for his comments.

Finally, I thank Prof. Dr. James Demmel, University of California, Berkeley, for his valuable and detailed remarks.

Mojoj obitelji

Contents

Chapter 1

Introduction

The paper considers the eigenvalue problem

$$Hx = \lambda x \ , \qquad x \neq 0 \ ,$$

where H is a real symmetric matrix of order n. Our aim is the following:

> *if* the matrix is "well-behaved", that is, if small relative changes of the matrix elements cause small relative changes in the eigenvalues, *then* perform the eigenreduction accurately in this sense.

Small relative changes in matrix elements typically occur when the matrix is being stored in the computer.

Our work generalizes the works by Barlow and Demmel [2] who considered scaled diagonally dominant matrices (which are described later), and by Demmel and Veselić [13] who considered positive definite matrices. Our results are, however, less definite than in the positive definite case. This is due to the fact that the structure of the set of all well–behaved indefinite matrices is more complicated than the structure of the set of all well–behaved positive definite matrices, and is not simply characterized as yet. Demmel and Veselić's [13] algorithm of choice was the Jacobi method. One of the versions of the algorithm that they used consists of two steps. First step is to calculate the Cholesky decomposition of a starting positive definite matrix. Second step is to apply the implicit (one–sided) version of the Jacobi method to the Cholesky factor as described by Veselić and Hari [31]. We use the algorithm which is an immediate generalization of this two–step algorithm, and was proposed by Veselić [28, 29].

The algorithm consists of two steps.

1. Decompose H as

$$H = GJG^T \ , \qquad J = I_{npos} \oplus (-I_{r-npos}) \ , \tag{1.1}$$

where G is a $n \times r$ matrix of a full column rank, $\text{rank}\,(H) = r$, and $npos$ is the number of the positive eigenvalues of H.

This decomposition is an extension of the known symmetric indefinite decomposition of Bunch and Parlett [6]. The eigensolutions of the matrix H and the pair G^TG, J are simply related. There always exists a matrix F which diagonalizes the pair G^TG, J such that

$$F^TG^TGF = \Delta \ , \qquad F^TJF = J \ .$$

where Δ is diagonal and positive definite. The matrices for which $F^TJF = J$ are called $J-$orthogonal. The non–zero eigenvalues of H are the diagonal elements of ΔJ, and the corresponding eigenvectors are the columns of $GF\Delta^{-1/2}$.

2. Apply the implicit (one–sided) $J-$orthogonal Jacobi method to the pair G, J to find the non–zero eigenvalues and the corresponding eigenvectors of H.

The implicit $J-$orthogonal Jacobi method consists of an iterative application of the transformation

$$G_{m+1} = G_m J_m \ ,$$

where $G \equiv G_0$ and J_m is a $J-$orthogonal Jacobi plane rotation. The $J-$orthogonality of J_m means that J_m performs a hyperbolic rotation if $1 \leq i \leq npos < j \leq r$, and a trigonometric rotation otherwise. Since the implicit Jacobi works only on the columns of G, it is suitable for parallel computing. The symmetric indefinite decomposition (1.1) is, however, not suitable for parallelization. The transition from the matrix H to the pair G^TG, J is, in fact, one step of the LR algorithm and usually has a diagonalizing effect. This reduces the number of iterative steps in our algorithm, and makes it faster than the standard Jacobi.

The algorithm has very favourable accuracy properties. For most well–behaved matrices we were able to prove relative error bounds for the eigenvalues and the norm error bounds for the eigenvectors similar to those in [13]. These errors are uniformly better than those for QR or the standard Jacobi algorithm applied directly to H.

Now we present our error bounds. They depend on new perturbation theory for eigenvalues and eigenvectors, error analysis of the symmetric indefinite decomposition, and error analysis of the $J-$orthogonal Jacobi methods. The statement that our algorithm is more accurate than QR or the standard Jacobi algorithm depends also on some empirical observations for which we have overwhelming numerical evidence, but somewhat weaker theoretical understanding. Our perturbation theory is an extension of those of Barlow and Demmel [2] and Demmel and Veselić [13].

We first consider known results. Let H be a real non–singular symmetric matrix. Let δH be a small symmetric perturbation of H such that

$$|\delta H_{ij}| \leq \varepsilon |H_{ij}| \ . \tag{1.2}$$

Let λ_i and λ_i' be the $i-$th eigenvalues of H and $H + \delta H$, respectively, numbered so that $\lambda_1 \leq \cdots \leq \lambda_n$. The standard perturbation theory [33] says that (1.2) implies

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq \frac{\|\delta H\|_2}{\lambda_i} \leq \varepsilon\sqrt{n}\|H\|_2 \cdot \|H^{-1}\|_2 = \varepsilon\sqrt{n}\kappa(H) \ , \tag{1.3}$$

where $\kappa(H) \equiv \|H\|_2 \cdot \|H^{-1}\|_2$ is the condition number of H. For the positive definite H, Demmel and Veselić [13] proved the following stronger result: write $H = DAD$ where $D = (\text{diag }(H))^{1/2}$ is a scaling so that $A_{ii} = 1$. Then (1.2) implies

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq \frac{\varepsilon n}{\lambda_{min}(A)} \leq \varepsilon n \kappa(A) \ . \tag{1.4}$$

By a theorem of Van der Sluis [27]

$$\kappa(A) \leq n \min_D \kappa(DHD) \ , \tag{1.5}$$

i.e. $\kappa(A)$ nearly minimizes the condition number of positive definite H over all possible diagonal scalings. Clearly, it is possible that $\kappa(A) \ll \kappa(H)$ and it is always true that $\kappa(A) \leq n\kappa(H)$, so the bound (1.4) is always at least about as good and can be much better than the bound (1.3). Demmel and Veselić [13] showed that (1.4) also holds under a more general perturbation of the type

$$|\delta H_{ij}| \leq \varepsilon (H_{ii} H_{jj})^{1/2} \ , \tag{1.6}$$

and that the standard Jacobi method computes the eigenvalues with nearly this accuracy. Barlow and Demmel [2] considered scaled diagonally dominant matrices, i.e. matrices of the form

$$H = DAD \ , \qquad A = E + M \ ,$$

where D is diagonal and non–singular, E is diagonal with elements ± 1, diag $(M) = 0$, and $\|M\|_2 = \zeta < 1$. They showed that for such matrices (1.2) implies

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq \frac{\varepsilon n^2}{1 - \zeta} \ , \tag{1.7}$$

and that a version of bisection without previous tridiagonalization computes the eigenvalues with nearly this accuracy.

Our perturbation bound for the non–singular but possibly indefinite matrix H is the following: set

$$|\!|H|\!| = D\widehat{A}D \ , \tag{1.8}$$

where $|\!|\cdot|\!|$ is the spectral absolute value ($|\!|H|\!|$ is symmetric square root of H^2), and $D = (\text{diag }(|\!|H|\!|))^{1/2}$. Then

$$\frac{|\lambda_i - \lambda_i'|}{|\lambda_i|} \leq \frac{\varepsilon n}{\lambda_{min}(\widehat{A})} \tag{1.9}$$

holds under the perturbations of types (1.2) and under

$$|\delta H_{ij}| \leq \varepsilon D_{ii} D_{jj} \ . \tag{1.10}$$

This bound is actually derived in the more general setting of positive definite Hermitian matrix pairs. By (1.5) it is always true that $\kappa(\widehat{A}) \leq n\kappa(|\!|H|\!|) = n\kappa(H)$, and it is possible that $\kappa(\widehat{A}) \ll \kappa(H)$. Therefore, our bound (1.9) is always at least about as good and can be much better than the bound (1.3). If H is positive definite, our bound reduces to the bound (1.4). If H is scaled diagonally dominant, our bound is similar to the bound (1.7) (see Chap. 2).

Since the implicit $J-$orthogonal Jacobi method works on the pair G, J, we also need the perturbation theory in the case when H is perturbed by its factors. Let λ_i' be the $i-$th eigenvalue of a perturbed matrix $(G+\delta G)J(G+\delta G)^T$. Set $G = BD$ where D is diagonal positive definite, and columns of B have unit norms. Set $\delta G = \delta BD$. If $\|\delta B\| \leq \varepsilon$ and $\varepsilon/\sigma_{min}(B) < 1$, where $\sigma_{min}(B)$ is the smallest singular value of B, then

$$(1-\varepsilon/\sigma_{min}(B))^2 \leq \frac{\lambda_i'}{\lambda_i} \leq (1+\varepsilon/\sigma_{min}(B))^2 \ . \tag{1.11}$$

Here H needs not to be non–singular, but G must have full column rank.

Error bounds for the eigenvalues computed by our algorithm follow from (1.9), (1.11), and the error analysis of our algorithm. Let H be non-singular. Suppose that both steps of our algorithm are performed in a floating–point arithmetic with precision ε. Let G, J be the output of the symmetric indefinite decomposition. Write $G = D_G B_G$, where D_G is diagonal positive definite, and rows of B_G have unit norms. For the matrices G_m obtained by the implicit $J-$orthogonal Jacobi method write $G_m = B_m D_m$, where D_m is diagonal and positive definite, and columns of B_m have unit norms. Let G_M, J be the last pair obtained by the implicit Jacobi, and let G_M satisfy the stopping criterion,

$$|(B_M^T B_M)_{ij}| \leq tol \ , \qquad \text{for all } i \neq j \ .$$

tol is a small constant, usually n times machine precision. This relative stopping criterion is a natural consequence of (1.11) and it has been used before [13, 29, 31]. Let λ_i' be the $i-$th calculated eigenvalue. Then

$$\frac{|\lambda_i - \lambda_i'|}{|\lambda_i|} \leq \frac{272 n^2 \varepsilon}{\lambda_{min}(D_G^{-1}|GJG^T|D_G^{-1})} + 2\varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol + n^2 \varepsilon \tag{1.12}$$

holds with the relative error of $O(\varepsilon)$. Here GJG^T denotes the exact product of the calculated factors of H, and C_m are moderate constants. Throughout the thesis the formulation "with the relative error of $O(\varepsilon)$" means that ε is replaced by $\varepsilon(1+K\varepsilon)$, where $0 < K \ll 1/\varepsilon$. The first quotient on the right hand side of (1.12) comes from (1.9) and the error analysis of the symmetric indefinite decomposition, and the rest comes from (1.11) and the error analysis of the implicit Jacobi. The bound (1.12) has the same order of magnitude as predicted by the perturbation theory of (1.9) and (1.11) if $\lambda_{min}(D_G^{-1}|GJG^T|D_G^{-1})$ is not much smaller than $\lambda_{min}(\widehat{A})$ of (1.9), and if the quantity $1/\sigma_{min}(B_m)$ does not grow much during the implicit Jacobi (note that in exact arithmetic $\lim_{m\to\infty} \sigma_{min}(B_m) = 1$). We have strong numerical evidence for both these facts, but our theoretical understanding is weaker. Moreover, we have observed that $1/\sigma_{min}(B_0)$ is usually very small. This means that:

- the error induced by symmetric indefinite decomposition is usually larger than the error induced by implicit Jacobi,
- our method becomes even more accurate if the (almost) exact factor G is readily supplied,
- our algorithm is usually faster than the standard Jacobi.

Similar observations were made by Demmel and Veselić [13] for the positive definite H. Moreover, since the theoretical results about the behaviour of $1/\sigma_{min}(B_m)$ are independent of the type of rotations used, we conclude that there is no reason to avoid hyperbolic rotations. Deichmöller [8] considered the solving of the generalized singular value problem with Jacobi–type methods, and obtained similar results about the growth of the condition of scaled matrices and a good error analysis for non–orthogonal rotations used there.

Our approach to the eigenvector perturbation theory is that of [20] which deals with the norm–estimates of the eigenprojections and thus allows the treatment of multiple and clustered eigenvalues. Our error bound holds, however, only for the eigenvectors corresponding to single eigenvalues. Let, as above, H and G both be non–singular. Let v_i and v_i' be the eigenvectors of λ_i and λ_i', respectively. Let $\lambda_{G,i}$ be the i–th eigenvalue of GJG^T. Then, less formally stated,

$$\|v_i' - v_i\|_2 \leq \frac{\sqrt{2}\eta}{rg(\lambda_i)} + \frac{4\sqrt{2}\bar{\eta}}{rg_G(\lambda_{G,i})} + O(n^2\varepsilon) \ . \tag{1.13}$$

Here η is the first quotient of the right hand side of (1.12), and $\bar{\eta}$ is approximately 1.5 times the rest of the right hand side of (1.12). $rg(\lambda)$ and $rg_G(\lambda)$ are two kinds of relative gaps between the eigenvalues, e.g. for $\lambda > 0$ we set

$$rg_G(\lambda) = \min\left\{1, \frac{\lambda_R - \lambda}{\lambda_R + \lambda}, \frac{\lambda - \lambda_L}{\lambda + \lambda_L}\right\} \ .$$

Here λ_L and λ_R are the left and right neighbours of λ in the spectrum, and the quotients containing them are defined only if λ_L, λ_R exist and are positive, respectively. This result applied to positive definite or scaled diagonally dominant H is similar to the corresponding results of [13, 2], although with a different definition of relative gap. The bound (1.13) compares favourably to the standard eigenvector result [22] which, for the perturbation of the type (1.2), says

$$\|v_i' - v_i\| \leq \frac{n\varepsilon\|H\|_2}{\min_{i\neq j}|\lambda_i - \lambda_j|} + O(\varepsilon^2).$$

In fact, if H has two or more tiny eigenvalues, then the above minimum is necessarily small for some i's, but the relative gaps may be large.

To illustrate our theory consider the matrix

$$H = \begin{bmatrix} 1600 & -300 & 14 & 300000 \\ -300 & 43.5 & -4.75 & -423212 \\ 14 & -4.75 & 0.1875 & 19800 \\ 300000 & -423212 & 19800 & 3207938 \cdot 10^3 \end{bmatrix}$$

whose all elements are sums of powers of 2, and are exactly stored in IEEE single precision, $\varepsilon \approx 10^{-8}$. We have

$$\frac{1}{\lambda_{min}(D_G^{-1}|GJG^T|D_G^{-1})} \approx 18 \ , \qquad \frac{1}{\sigma_{min}(B)} \approx 1.1 \ ,$$

so we expect that the single precision version of our algorithm ($\varepsilon \approx 10^{-8}$) returns six or seven correct decimal digits. The eigenvalues of H are

$$\begin{aligned} \lambda_1 &= -54.043364 \\ \lambda_2 &= -0.0283096849 \\ \lambda_3 &= 1613.74866 \\ \lambda_4 &= 3207938084.0105 \end{aligned}$$

Here the digits which are common to our algorithm and the LAPACK routine DSYEV which implements tridiagonalization followed by QR iteration (all performed in IEEE double precision, $\varepsilon \approx 10^{-16}$) are displayed. Our algorithm, QR algorithm from the LAPACK routine SSYEV, and the standard Jacobi, all in single precision, computed the following eigenvalues:

	OUR ALG.	*SSYEV*	*JACOBI*
λ_1	-54.043369	-55.990593	-54.043369
λ_2	-0.02830968	-0.0326757	-0.02830995
λ_3	1613.7487	1651.6652	1613.7486
λ_4	3207938000	3207938000	3207938000

Therefore, our algorithm computed the eigenvalues with the predicted relative accuracy, QR has totally missed the absolutely smallest eigenvalue (and two more are very inaccurate), and the standard Jacobi computed the absolutely smallest eigenvalue somewhat less accurately than our algorithm. Note that H is far from being scaled diagonally dominant which shows that our results are a non–trivial generalization of those of [2]. The algorithms behaved similarly on all such matrices for which the bound (1.12) is small and $\kappa(H)$ is large.

To explain the loss of accuracy in QR and the standard Jacobi algorithm note that both algorithms do all of their work on an indefinite matrix. Let H_m be the sequence of matrices generated in floating–point arithmetic by either of those algorithms. Further, let $\widehat{A}_m$ be obtained from H_m according to (1.8). In both algorithms it frequently happens that $\max_m \kappa(\widehat{A}_m) \gg \kappa(\widehat{A})$, which can, in turn, result in the loss of accuracy. In QR algorithm accuracy can be lost during the tridiagonalization, as well as during the iterative part. To illustrate the loss of accuracy during the tridiagonalization consider the matrix

$$H = \begin{bmatrix} 10^{20} & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 10^{20} \end{bmatrix},$$

for which $\kappa(\widehat{A}) \approx 1$ and $\kappa(H) \approx 10^{20}$. The tridiagonalization, which consists of one Givens rotation, yields the matrix

$$H_1 = \begin{bmatrix} 10^{20} & \sqrt{2} & 0 \\ \sqrt{2} & 10^{20} + \frac{3}{2} & 10^{20} - \frac{1}{2} \\ 0 & 10^{20} - \frac{1}{2} & 10^{20} - \frac{1}{2} \end{bmatrix},$$

for which $\kappa(\widehat{A}_1) \approx \kappa(H)$. In floating–point arithmetic with precision $\varepsilon = 10^{-16}$ the computed matrix H_1 is exactly singular indicating total loss of accuracy. Demmel [10] gives an example of a well–behaved tridiagonal matrix where $\kappa(\widehat{A}_m)$ almost reaches $\kappa(H)$ during QR iterations, which, in turn, results in the total loss of accuracy.

The main difference between indefinite non–singular and positive definite matrices is the following: for positive definite H the perturbations of the types (1.2) and (1.6) are equivalent in the sense that if H is insensitive to the one type, it is insensitive to the other type, and vice versa [13]. For indefinite H this is not the case. Indeed, let

$$H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & \epsilon \end{bmatrix},$$

where ϵ is small (this matrix is considered in Sections 2.3, 4.3). H is obviously very sensitive to perturbations of the type (1.10) so the bound (1.9) must necessarily be large. On the other side, H is insensitive to small relative componentwise perturbations (1.2). This shows that we are still unable to completely characterize all well–behaved symmetric matrices. Due to large errors in the symmetric indefinite decomposition, our algorithm computes the eigenvalues with large relative errors. We can, however, easily obtain an almost exact factorization of H (one way is to change the choice of pivots in the symmetric indefinite decomposition), and then the implicit Jacobi computes the eigensolution to nearly full working accuracy. This shows that we have not completely reached our ideal: if the matrix is well behaved, our algorithm should compute the eigenvalues with nearly this accuracy.

The thesis is organized as follows: Chapter 2 presents the new perturbation theory. This chapter, except Subsection 2.3.1, is due to Veselić and Slapničar [32]. The results of Veselić and Slapničar are included mainly for the sake of completeness. In Chapter 3 we first describe the J–orthogonal Jacobi method for the pair H, J, where H is positive definite, and give its error analysis. Although this explicit method is rarely used, its error analysis is the basis for the later analysis of the implicit method. The error analysis consists of two steps. We first show that one step of $J-$orthogonal Jacobi method satisfies the perturbation bounds of Chapter 2. Then we combine one–step error analysis with the perturbation bounds to obtain overall error bounds for the eigensolution computed by $J-$orthogonal Jacobi method. In Subsection 3.2.2 we give known and new results concerning the upper bound for $1/\sigma_{min}(B_m)$. Then we describe and analyse the implicit $J-$orthogonal Jacobi method, and do the same for the implicit method with fast and fast self–scaling rotations. The latter are used to suppress possible underflow/overflow when accumulating the diagonal of the fast rotations. In Chapter 4 we define the symmetric indefinite decomposition (1.1) and give its error analysis. In Section 4.3 we combine the error analysis of the symmetric indefinite decomposition, error analysis of the implicit $J-$orthogonal Jacobi method, and the perturbation bounds of Chap. 2, to obtain the final error bounds for the computed eigensolution of the real symmetric eigenvalue problem. There we also shortly refer to the singular case, and state some open problems. In Section 4.4 we give an interesting theoretical result saying that the condition of the scaled matrix G^TG, $\kappa(B^TB)$, is bounded by a function of n irrespective of the condition of the starting matrix H. In Chapter 5 we present results of our numerical experiments. Main tests were performed by comparing QR algorithms from LAPACK, standard Jacobi, and our algorithms in single and double precision. We also tested the behaviour of $\lambda_{min}(D_G^{-1}|GJG^T|D_G^{-1})$ and $\sigma_{min}(B_m)$, and compared computation times.

Chapter 2

Floating–point perturbations of Hermitian matrices[1]

2.1 Introduction and preliminaries

The standard perturbation result for the eigenvalue problem of a Hermitian matrix H of order n, $Hx = \lambda x$, reads [16]

$$|\delta\lambda_i| \leq \|\delta H\|_2 \ , \tag{2.1.1}$$

where

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \ ,$$
$$\lambda_1' = \lambda_1 + \delta\lambda_1 \leq \ldots \leq \lambda_n' = \lambda_n + \delta\lambda_n \ ,$$

are the eigenvalues of H and $H + \delta H$, respectively. The perturbation matrix δH is again Hermitian, and $\|\cdot\|_2$ is the spectral norm. The backward error analysis of various eigenvalue algorithms initiated by Wilkinson [33] follows the same pattern, i.e. the round–off error estimates are given in terms of norms. A more realistic perturbation theory starts from the fact that both the input entries of the matrix H and the output eigenvalues are given in the floating–point form. Thus, a desirable estimate would read

$$\max_i \left| \frac{\delta\lambda_i}{\lambda_i} \right| \leq C \max_{i,j} \left| \frac{\delta H_{ij}}{H_{ij}} \right| \ , \tag{2.1.2}$$

where we define $0/0 = 0$. Colloquially, "floating–point" perturbations are those with $|\delta H_{ij}| \leq \varepsilon |H_{ij}|$, ε small. Similarly, we call a matrix "well–behaved" if (2.1.2) holds with a "reasonable" C, i.e. if the small relative changes in the matrix elements cause small relative changes in the eigenvalues. For the floating–point perturbations (2.1.1)

[1]Sections 2.1, 2.2 and 2.3 of this chapter are due to Veselić and Slapničar [32]. Subsection 2.3.1 is new.

implies (2.1.2) with $C = \sqrt{n} \cdot \kappa(H) \equiv \sqrt{n} \cdot \|H\|_2 \|H^{-1}\|_2$, and this bound is almost attainable. This is illustrated by the positive definite matrix

$$H = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon \end{bmatrix}, \qquad 0 < \varepsilon \ll 1 .$$

The small eigenvalue of H is very sensitive to small relative changes in the matrix elements.

Our results generalize the results obtained in [12, 2, 13]. Demmel and Veselić [13] showed that for a positive definite matrix H (2.1.2) holds with

$$C = \frac{n}{\lambda_{min}(A)} ,$$

where

$$A = (\text{diag } (H))^{-1/2} H (\text{diag } (H))^{-1/2} \tag{2.1.3}$$

is *the standard scaled matrix.* The condition of A can be much smaller and is never much larger than that of H. Indeed, since $A_{ii} = 1$ it follows

$$\frac{1}{\lambda_{min}(A)} \le \kappa(A) \le \frac{n}{\lambda_{min}(A)} ,$$

whereas (1.5) implies

$$\kappa(A) \le n \cdot \kappa(H) . \tag{2.1.4}$$

Similar results hold for the singular value problem [13].

The aim of this paper is to extend the above result to general non–singular Hermitian matrices. The nature of the estimate (2.1.2) shows that the non-singularity is a natural condition to require. We show (Th. 2.2.3) that (2.1.2) holds for a non–singular Hermitian matrix H with

$$C = \||A|\|_2 \|\widehat{A}^{-1}\|_2 ,$$

where

$$H = DAD , \qquad \widehat{A} = D^{-1} |\!|H|\!| D^{-1} .$$

Here D is *any scaling matrix*, i.e. a positive definite diagonal matrix, and $|\cdot|$, $|\!|\cdot|\!|$ denote the two kinds of absolute value functions, "pointwise" and "spectral":

$$|A|_{ij} = |A_{ij}| , \qquad |\!|H|\!| = \sqrt{H^2} ,$$

respectively. Note that $\|A\|_2 \le \||A|\|_2 \le \sqrt{n}\|A\|_2$ holds for any matrix A. The scaling D is typically, but not necessarily of the standard form $D = (\text{diag } |\!|H|\!|)^{1/2}$. This result is stated and proved in a more general setting, namely that of a matrix pair H, K with K positive definite, thus properly generalizing corresponding results of [2, 13]. Our eigenvector result, stated in Subsect. 2.2.1, concerns the case of a single non–singular Hermitian matrix and it essentially generalizes the norm–estimates from [2, 13]. An unpleasant point of our theory is that the matrix $|H|$, which has to be scaled, is not easy to compute. Moreover, the set of well-behaved indefinite Hermitian matrices *is not scaling-invariant.*

Barlow and Demmel [2] showed that for matrices of the type

$$H = D(E+N)D \;, \tag{2.1.5}$$

where D, E are diagonal, $E^2 = I$, diag $(N) = 0$ and $\|N\|_2 < 1$, (2.1.2) holds with

$$C = \frac{n}{1 - \|N\|_2} \;. \tag{2.1.6}$$

The matrices (2.1.5) are called *scaled diagonally dominant* (s.d.d.). We show that for a s.d.d. matrix

$$\||A|\|_2 \|\widehat{A}^{-1}\|_2 \leq n \frac{1 + \||N|\|_2}{1 - \|N\|_2} \;.$$

Although this does not reproduce the constant C in (2.1.6) (there is an extra factor $1 + \||N|\|_2 \leq 1 + \sqrt{n}$), we see that s.d.d. matrices are included in our theory.

In the positive definite case the only well–behaved matrices are those which can be well scaled, i.e. for which the scaled matrix A from (2.1.3) is "reasonably" conditioned. More precisely, if (2.1.2) holds for sufficiently small δH, then $\lambda_{min}(A) \geq 2/(1+C)$ for A from (2.1.3). This, rather sharp result is proved in [32]. It improves a related result of [13] and also yields a slight improvement of the van der Sluis estimate (2.1.4).

In contrast to this, the choice of well–behaved indefinite matrices is, in a sense, richer. Writing

$$H = GJG^*$$

with G^*G positive definite (G need not be square) and J non-singular, the eigenvalue problem $Hx = \lambda x$ converts into the problem

$$\widehat{H} y = \lambda J^{-1} y \;, \qquad \widehat{H} = G^*G \;. \tag{2.1.7}$$

In Sect. 2.3 we prove the estimate of the type (2.1.2) for the problem (2.1.7) under the perturbations of the factor $|\delta G_{ij}| \leq \varepsilon |G_{ij}|$. The latter is a generalization of the singular value problem known as *hyperbolic singular value problem* [21]. The estimates again depend on the condition number of the matrix obtained by scaling G^*G. As an interesting application, we obtain floating-point perturbation estimates for matrices of the type

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^* & 0 \end{bmatrix} \;, \tag{2.1.8}$$

where $H_{12}H_{12}^*$ is positive definite. Note that this H may be singular. As could be expected, the only well-behaved singular matrices are those where the rank defect can be read-off from the zero pattern.

Similarly as in [2], [13] we note the remarkable fact that our eigenvalue estimates are independent of the condition number of the corresponding eigenvector matrices

- in generalized Hermitian eigenvalue problems they are not unitary and there is no upper bound for their condition. This phenomenon seems to be typical for the "floating-point" perturbation theory.

2.2 Well–conditioned scalings

In this section we present perturbation results which are natural extensions of those from [2] and [13]. We first give a general perturbation result for the eigenvalues of the pair H, K with K positive definite. (An eigenvalue of the pair H, K is a scalar λ for which $\det(H - \lambda K) = 0$.) For this purpose we introduce a new *absolute value of H relative to K* denoted by $|H|_K$. We then apply our general perturbation result to the floating–point perturbations of the matrices H and K. Theorems 2.2.3 and 2.2.4 give two simplifications of the perturbation bounds and Th. 2.2.5 gives bounds for another, more general, type of perturbation where perturbing the zero elements is also allowed. Our theory applied to a single positive definite matrix slightly improves the corresponding results of [13]. It also improves the van der Sluis estimate (2.1.4) in some cases [32]. Then we apply our theory to a single non–singular indefinite matrix. We prove that our theory includes scaled diagonally dominant matrices [2]. We also characterize the class of matrices with the best perturbation bounds. At the end we give some examples, and also consider some singular matrices. In Subsect. 2.2.1 we consider the perturbation of the eigenvectors of a single non–singular matrix H.

Theorem 2.2.1 *Let H, K be Hermitian and K positive definite. Set $K = ZZ^*$ and*

$$|H|_K = Z|Z^{-1}HZ^{-*}|Z^* \ . \qquad (2.2.1)$$

$|H|_K$ is independent of the freedom of choice in Z.[2] *Let δH, δK be Hermitian perturbations such that for all $x \in \mathbf{C}^n$*

$$|x^*\delta H x| \leq \eta_H x^* |H|_K x \ , \qquad |x^* \delta K x| \leq \eta_K x^* K x \ , \qquad \eta_H, \eta_K < 1 \qquad (2.2.2)$$

holds. Let λ_i and λ_i' be the increasingly ordered eigenvalues of the matrix pairs H, K and $H' \equiv H + \delta H, K' \equiv K + \delta K$, respectively. Then $\lambda_i' = 0$ if and only if $\lambda_i = 0$, and for non–vanishing λ_i's we have

$$\frac{1 - \eta_H}{1 + \eta_K} \leq \frac{\lambda_i'}{\lambda_i} \leq \frac{1 + \eta_H}{1 - \eta_K} \ . \qquad (2.2.3)$$

Proof. Let $K = ZZ^* = FF^*$. Then $Z = FU$, where U is a unitary matrix, and

$$Z|Z^{-1}HZ^{-*}|Z^* = FU|U^*F^{-1}HF^{-*}U|U^*F^* = F|F^{-1}HF^{-*}|F^* \ .$$

[2] For H positive definite we obviously have $|H|_K = H$.

Thus, $\mathbf{|}H\mathbf{|}_K$ is independent of the freedom of choice in Z. From (2.2.2) it follows

$$
\begin{aligned}
x^*(H - \eta_H \mathbf{|}H\mathbf{|}_K)x &\leq x^*(H + \delta H)x \leq x^*(H + \eta_H \mathbf{|}H\mathbf{|}_K)x \ , && (2.2.4)\\
(1 - \eta_K)x^*Kx &\leq x^*(K + \delta K)x \leq (1 + \eta_K)x^*Kx \ . && (2.2.5)
\end{aligned}
$$

Now note that the pair $H \pm \eta_H \mathbf{|}H\mathbf{|}_K, K$ has the same eigenvectors as the pair H, K with the (again increasingly ordered) eigenvalues $\lambda_i \pm \eta_H |\lambda_i|$. Let $\widehat{\lambda}_i$ be the increasingly ordered eigenvalues of the pair H', K. The monotonicity property of the eigenvalues together with (2.2.4) yields immediately

$$
1 - \eta_H \leq \frac{\widehat{\lambda}_i}{\lambda_i} \leq 1 + \eta_H \ . \tag{2.2.6}
$$

It is also clear that H and H' have the same inertia.[3] The transition form H', K to H', K' is similar. Note that both pairs have again the same inertia. If e.g. $\widehat{\lambda}_i \leq 0$, then $\lambda'_i \leq 0$ and (2.2.5) implies

$$
\min_{S_i} \max_{x \in S_i} \frac{x^* H' x}{(1 - \eta_K)x^*Kx} \leq \min_{S_i} \max_{x \in S_i} \frac{x^* H' x}{x^* K' x} \leq \min_{S_i} \max_{x \in S_i} \frac{x^* H' x}{(1 + \eta_K)x^*Kx},
$$

where S_i is any $i-$dimensional subspace of $\mathbf{C}^n$. In other words,

$$
\frac{\widehat{\lambda}_i}{1 - \eta_K} \leq \lambda'_i \leq \frac{\widehat{\lambda}_i}{1 + \eta_K} \ . \tag{2.2.7}
$$

Similarly, if $\widehat{\lambda}_i \geq 0$, then $\lambda'_i \geq 0$, and we obtain

$$
\frac{\widehat{\lambda}_i}{1 + \eta_K} \leq \lambda'_i \leq \frac{\widehat{\lambda}_i}{1 - \eta_K} \ . \tag{2.2.8}
$$

Now (2.2.7) and (2.2.8) combined with (2.2.6) give (2.2.3). Q.E.D.

We now apply this result to the floating–point perturbations of matrix entries. Set

$$
\widetilde{C}(H, K) = \sup_{x \neq 0} \frac{|x|^T |H| |x|}{x^* \mathbf{|}H\mathbf{|}_K x}
$$

and

$$
\widetilde{C}(H) = \widetilde{C}(H, I) \ .
$$

Obviously, $\widetilde{C}(H, K)$ is finite if and only if H is non–singular. For every H, K with K positive definite, we have

$$
\widetilde{C}(H, K) \geq 1 \ . \tag{2.2.9}
$$

Indeed, if $\widetilde{C}(H, K)$ were less than one, then the matrices H, K, $\delta H = -H$ and $\delta K = 0$ would satisfy the assumptions of Th. 2.2.1 and this would, in turn, imply that $H + \delta H$ is non–singular — a contradiction.

[3] In fact, H and H' have the same null–spaces.

Theorem 2.2.2 *Let H, K be Hermitian matrices with H non–singular and K positive definite. Let Hermitian perturbations δH and δK satisfy*

$$|\delta H_{ij}| \leq \varepsilon |H_{ij}| \ , \qquad |\delta K_{ij}| \leq \varepsilon |K_{ij}| \ , \tag{2.2.10}$$

such that

$$\eta_H = \varepsilon \widetilde{C}(H, K) < 1 \ , \qquad \eta_K = \varepsilon \widetilde{C}(K) < 1 \ .$$

Then the assumption (2.2.2) of Th. 2.2.1 is fulfilled, hence its assertion holds.

Proof. We have

$$|x^* \delta H x| \leq |x|^T |\delta H| |x| \leq \varepsilon |x|^T |H| |x| \leq \varepsilon \widetilde{C}(H, K) x^* \boldsymbol{|} H \boldsymbol{|}_K x \ ,$$

and similarly

$$|x^* \delta K x| \leq \varepsilon \widetilde{C}(K) x^* K x \ .$$

Q.E.D.

Th. 2.2.1 is a significant improvement over Lemma 1 and Th. 4 from [2] which require a more restrictive condition

$$|x^* \delta H x| \leq \eta_H |x^* H x|$$

which has non–trivial applications only for positive definite H.

The values $\widetilde{C}(H, K)$ and $\widetilde{C}(K)$ are not readily computable and we now exhibit a chain of simpler upper bounds for them.

Theorem 2.2.3 *Let H, K be as in Th. 2.2.2, and let A, $\widehat{A}$ and B be defined by*

$$H = DAD \ , \qquad \boldsymbol{|} H \boldsymbol{|}_K = D \widehat{A} D \ , \qquad K = D_1 B D_1 \ , \tag{2.2.11}$$

where D and D_1 are scaling matrices. Then

$$\begin{aligned} \widetilde{C}(H, K) &\leq \| |A| \|_2 \| \widehat{A}^{-1} \|_2 \equiv C(A, \widehat{A}) \ , \\ \widetilde{C}(K) &\leq \| |B| \|_2 \| B^{-1} \|_2 \equiv C(B) \ , \end{aligned} \tag{2.2.12}$$

and $\eta_H = \varepsilon C(A, \widehat{A}) < 1$, $\eta_K = \varepsilon C(B) < 1$ implies the assertion of Th. 2.2.1.

Proof. We have

$$\begin{aligned} |x|^T |H| |x| &= |x|^T D |A| D |x| \leq \| |A| \|_2 x^* D^2 x \\ &\leq C(A, \widehat{A}) x^* D \widehat{A} D x = C(A, \widehat{A}) x^* \boldsymbol{|} H \boldsymbol{|}_K x \ , \end{aligned}$$

and similarly

$$|x|^T |K| |x| \leq C(B) x^* D_1 B D_1 x = C(B) x^* K x \ .$$

Q.E.D

The constant $C(A,\widehat{A})$ cannot be uniformly improved. Indeed, take H as diagonal with $H^2 = I$ and let $H' = H + \delta H$ be obtained by setting to zero any of the diagonal elements of H. Then the assertion of the above theorem, applied to the pair $H, K = I$ with $\delta K = 0$, is obviously not true and we have $\eta_H = 1$, $\eta_K = 0$.

Of course, all this does not mean that Th. 2.2.3 covers all well behaved matrices. Next sections will show the contrary.

The constants $C(A,\widehat{A})$, $C(B)$ are further estimated as follows:

Theorem 2.2.4 *Let H, K be as in Th. 2.2.2, and let A, $\widehat{A}$ and B be defined by (2.2.11), where D, D_1 are scalings. Then*

$$C(A,\widehat{A}) \le \mathrm{Tr}\ \widehat{A}\|\widehat{A}^{-1}\|_2\ , \qquad C(B) \le \mathrm{Tr}\ B\|B^{-1}\|_2\ ,$$

and $\eta_H = \varepsilon \mathrm{Tr}\ \widehat{A}\|\widehat{A}^{-1}\|_2 < 1$, $\eta_K = \varepsilon \mathrm{Tr}\ B\|B^{-1}\|_2 < 1$ implies the assertion of Th. 2.2.1.

Proof. Let

$$Z^{-1}HZ^{-*} = U\Lambda U^*$$

be an eigenvalue decomposition of $Z^{-1}HZ^{-*}$ with U unitary and Λ diagonal. Then $\boldsymbol{|}Z^{-1}HZ^{-*}\boldsymbol{|} = U|\Lambda|U^*$ and from (2.2.1) it follows

$$\boldsymbol{|}H\boldsymbol{|}_K = ZU|\Lambda|U^*Z^* = GG^*\ ,$$

where $G = ZU\sqrt{|\Lambda|}$. Furthermore,

$$H = Z(Z^{-1}HZ^{-*})Z^* = ZU\Lambda U^*Z^* = GJG^*\ ,$$

where J is diagonal with ± 1's on the diagonal. Setting $F = D^{-1}G$ for some positive definite diagonal D and using the obvious estimate

$$|(FJF^*)_{ij}| \le \sqrt{(FF^*)_{ii}(FF^*)_{jj}}\ ,$$

we obtain $|A_{ij}|^2 \le \widehat{A}_{ii}\widehat{A}_{jj}$, and hence $\||A|\|_2 \le \mathrm{Tr}\ \widehat{A}$. Similarly, $\||B|\|_2 \le \mathrm{Tr}\ B$, and the theorem now follows from the definitions of $C(A,\widehat{A})$ and $C(B)$. Q.E.D.

For the standard scalings $D = (\mathrm{diag}\ \boldsymbol{|}H\boldsymbol{|}_K)^{1/2}$, $D_1 = (\mathrm{diag}\ K)^{1/2}$, Th. 2.2.4 yields

$$C(A,\widehat{A}) \le n\|\widehat{A}^{-1}\|_2\ , \qquad C(B) \le n\|B^{-1}\|_2\ .$$

In addition, the above upper bounds can accomodate another class of perturbations where perturbing the zero elements is also allowed.

Theorem 2.2.5 *Let H, K be Hermitian matrices with H non–singular and K positive definite. Let Hermitian perturbations δH and δK satisfy*

$$|\delta H_{ij}| \le \varepsilon D_{ii} D_{jj} \ , \qquad |\delta K_{ij}| \le \varepsilon D_{1,ii} D_{1,jj} \ , \tag{2.2.13}$$

such that

$$\eta_H = \varepsilon n \|\widehat{A}^{-1}\|_2 < 1 \ , \qquad \eta_K = \varepsilon n \|B^{-1}\|_2 < 1 \ .$$

Then the assumption (2.2.2) of Th. 2.2.1 is fulfilled, hence its assertion holds.

Proof. Let us define the matrix E with $E_{ij} = 1$. We have

$$|x^* \delta H x| \le |x|^T |\delta H| |x| \le \varepsilon |x|^T D E D |x| \le \varepsilon \|E\|_2 x^* D^2 x \le \varepsilon n \|\widehat{A}^{-1}\|_2 x^* \mathbf{|} H \mathbf{|}_K x \ ,$$

and similarly

$$|x^* \delta K x| \le \varepsilon n \|B^{-1}\|_2 x^* K x \ .$$

Q.E.D.

Remark 2.2.6 Note that for the standard scaling the bounds of Theorems 2.2.3 and 2.2.5 differ by at most a factor n. Therefore, the relative error bounds which use $C(A, \widehat{A})$ and $C(B)$ actually allow both kinds of perturbations, (2.2.10) and (2.2.13), which makes them inappropriate in some cases (see Rem. 2.2.11 below).

When we apply our general theory to a single positive definite matrix H ($K = I$), Th. 2.2.4 reproduces the main floating-point perturbation result of Th. 2.3 from [13], while Th. 2.2.2 is even sharper. The perturbations allowed by Th. 2.2.5 are of the form

$$|\delta H_{ij}| \le \varepsilon \sqrt{H_{ii} H_{jj}} \ . \tag{2.2.14}$$

We now turn to the case of the single non-singular indefinite matrix H. We first prove that the class of matrices H with well-behaved $C(A, \widehat{A})$ includes the already known class of scaled diagonally dominant matrices. We have

Theorem 2.2.7 *Let*

$$H = DAD \ , \qquad A = E + N \ ,$$

with $E = E^* = E^{-1}$, $ED = DE$, and $\|N\|_2 < 1$. If $\widehat{A}$ is defined by $\mathbf{|} H \mathbf{|} = D\widehat{A}D$, then

$$C(A, \widehat{A}) \le n \frac{1 + \||N|\|_2}{1 - \|N\|_2} \ . \tag{2.2.15}$$

PROOF. Since D commutes with E, there exists a unitary matrix U which simultaneously diagonalizes D and E, i.e.

$$U^*DU = \Delta \ , \qquad U^*EU = \mathrm{diag}\ (\pm 1) \ .$$

Since Δ is only a permuted version of the matrix D, there exists a permutation matrix P such that $\Delta = PDP^T$. Setting $V = UP$, we have

$$V^*DV = D \ , \qquad V^*EV = E_1 \ ,$$

where E_1 is diagonal with ± 1's on the diagonal. Now perform the unitary transformation

$$H_1 = V^*HV = D(V^*EV + V^*NV)D = D(E_1 + N_1)D \ .$$

Here we used the fact that D and V commute. Also, $\|N_1\|_2 = \|N\|_2$.

By Lemma 3 of [2] for any eigenpair λ, y of H_1 we have

$$(1 - \|N_1\|_2)\|Dy\|_2^2 \le |\lambda|\|y\|_2^2 \le (1 + \|N_1\|_2)\|Dy\|_2^2 \ . \tag{2.2.16}$$

Note that formally [2] needs that N_1 have a zero diagonal. It is easily seen that this condition is not necessary. For any eigenpair λ, y of H, (2.2.16) implies

$$(1 - \|N\|_2)\|Dy\|_2^2 \le |\lambda|\|y\|_2^2 \le (1 + \|N\|_2)\|Dy\|_2^2 \ . \tag{2.2.17}$$

Now let $H = Y\Lambda Y^*$, $Y^*Y = I$, $\Lambda = \mathrm{diag}\ (\lambda_1, \cdots, \lambda_n)$, be an eigenvalue decomposition of H. Then $\lvert H \rvert = Y|\Lambda|Y^*$ and

$$\widehat{A}^{-1} = D\lvert H\rvert^{-1}D = DY|\Lambda|^{-1/2}|\Lambda|^{-1/2}Y^*D \ .$$

Therefore,

$$\|\widehat{A}^{-1}\|_2 = \|DY|\Lambda|^{-1/2}\|_2^2 \le n \max_i \|Dy_i\|_2^2 \frac{1}{|\lambda_i|} \le \frac{n}{1 - \|N\|_2} \ .$$

Here we have set $Y = [y_1, \cdots, y_n]$ and used (2.2.17) for every pair λ_i, y_i. The theorem now follows from[4]

$$\||A|\|_2 \le \|I + |N|\|_2 \le 1 + \||N|\|_2 \ .$$

Q.E.D.

The s.d.d. matrices are a special case of the matrices considered in Th. 2.2.7, that is, we do not require the diagonality of E. Note that the argument of [2] leading to the estimate (2.1.6) can be easily modified to hold under the conditions of Th. 2.2.7 as well.

[4]The case of the pair H, K of s.d.d. matrices is not covered by this result (cf. a similar claim in [2]), although it seems highly probable that such a generalization holds.

Even though we could only bound our measure $C(A,\widehat{A})$ by (2.2.15) which is somewhat weaker than (2.1.6), we expect that $C(A,\widehat{A})$ is actually much better. The following example illustrates the power of our theory. Set

$$\widehat{A} = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & & \\ & d & \\ & & d^2 \end{bmatrix}, \qquad d \geq 1 \; .$$

Then $\|\widehat{A}^{-1}\|_2 = 10$. For $d = 10^2$ the spectrum of $\lvert H \rvert = D\widehat{A}D$ is, properly rounded, $1.47 \cdot 10^{-1}$, $1.90 \cdot 10^3$, $1.00 \cdot 10^8$. Now H is obtained from $\lvert H \rvert$ by just turning the smallest eigenvalue into its negative. We obtain

$$H = \begin{bmatrix} 0.705 & 9.00 \cdot 10^1 & 9.00 \cdot 10^3 \\ 9.00 \cdot 10^1 & 1.00 \cdot 10^4 & 9.00 \cdot 10^5 \\ 9.00 \cdot 10^3 & 9.00 \cdot 10^5 & 1.00 \cdot 10^8 \end{bmatrix}$$

with

$$A = \begin{bmatrix} 0.705 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}, \qquad \|A\| \leq 3 \; .$$

Thus, $C(A,\widehat{A}) \leq 30$ and H is far from being s.d.d.

A natural question is to ask which matrix pairs or single non-singular matrices have the smallest η_H, η_K in Th. 2.2.3. Obviously, $C(B) \geq 1$ and the equality is attained, if and only if K is diagonal. In this case we can take $K = I$ and the whole problem reduces to the case of the single matrix H.

We first derive some useful inequalities. Set $x = K^{-1/2}y = D^{-1}z$. Then

$$|x^* H x| = |y^* K^{-1/2} H K^{-1/2} y| \leq y^* \lvert K^{-1/2} H K^{-1/2} \rvert y = x^* \lvert H \rvert_K x \; , \tag{2.2.18}$$

and thus

$$|z^* A z| \leq z^* \widehat{A} z \; . \tag{2.2.19}$$

Similarly, $|x^* H^{-1} x| \leq x^* \lvert H \rvert_K^{-1} x$, and

$$|z^* A^{-1} z| \leq z^* \widehat{A}^{-1} z \; . \tag{2.2.20}$$

Now we have $\|A^{-1}\|_2 \leq \|\widehat{A}^{-1}\|_2$, and

$$C(A,\widehat{A}) \geq \|A\|_2 \|\widehat{A}^{-1}\|_2 \geq \|A\|_2 \|A^{-1}\|_2 \geq 1 \; . \tag{2.2.21}$$

Theorem 2.2.8 *Let $H = DAD$ be Hermitian and non-singular and let $\lvert H \rvert = D\widehat{A}D$. Then*

$$C(A,\widehat{A}) = \||A|\|_2 \|\widehat{A}^{-1}\|_2 = 1 \tag{2.2.22}$$

if and only if A is proportional to $P\,\mathrm{diag}\,(A_1, \cdots, A_p)P^T$, where each of the blocks A_i has one of the forms

$$1\ , \qquad -1\ , \qquad \begin{bmatrix} 0 & e^{i\varphi} \\ e^{-i\varphi} & 0 \end{bmatrix}\ ,$$

A and D commute, and P is a permutation matrix.

PROOF. If H has the form described above, then $|H| = D^2|A| = D^2$, i.e. $\widehat{A} = I$ and (2.2.22) holds.

Conversely, if (2.2.22) holds, then all inequalities in (2.2.21) go into equalities. Without loss of generality we can assume that

$$\widehat{A}_{11} = 1\ . \tag{2.2.23}$$

Now the equality $\|A\|_2\|A^{-1}\|_2 = 1$ means that

$$A = cV\ , \qquad c > 0\ , \qquad V = V^{-1} = V^*\ . \tag{2.2.24}$$

From $|H|^2 = H^2$ it follows that

$$c^2 V D^2 V = \widehat{A} D^2 \widehat{A}\ . \tag{2.2.25}$$

This is equivalent to the unitarity of the matrix

$$W = cD^{-1}\widehat{A}^{-1}VD\ .$$

This, in turn, implies that W is similar to $c\widehat{A}^{-1/2}V\widehat{A}^{-1/2}$. Since the latter matrix is also Hermitian, it must be unitary, i.e.

$$c^2\widehat{A}^{-1/2}V\widehat{A}^{-1}V\widehat{A}^{-1/2} = I\ .$$

This is equivalent to

$$V\left(\frac{\widehat{A}}{c}\right)^{-1}V = \frac{\widehat{A}}{c}\ . \tag{2.2.26}$$

We now use $\|A\|_2\|\widehat{A}^{-1}\|_2 = \|(\widehat{A}/c)^{-1}\|_2 = 1$ which, together with (2.2.26) and (2.2.23), implies $\widehat{A} = I$, $c = 1$. Now we can write (2.2.25) as $D^2A = AD^2$, i.e. A and D commute. Finally, we use $\||A|\|_2\|\widehat{A}^{-1}\|_2 = \||A|\|_2 = 1$. By $c = 1$, the relation (2.2.24) gives

$$A = A^{-1} = A^*\ .$$

Here we need the following

Lemma 2.2.9 *Let $U^*U = I$ and $\||U|\|_2 = 1$. Then $|U|^T|U| = I$, i.e. each row of U contains at most one non–vanishing element. If, in addition, U is square, then U is a (one sided) permutation of a diagonal matrix. Conversely, $|U|^T|U| = I$ implies $U^*U = I$ and $\||U|\|_2 = 1$.*

PROOF. From $U^*U = I$ it follows $(|U|^T|U|)_{ii} \equiv 1$. If $a_{ij} = (|U|^T|U|)_{ij} \neq 0$ for some pair $i \neq j$, then the submatrix

$$\begin{bmatrix} 1 & a_{ij} \\ a_{ij} & 1 \end{bmatrix}$$

of $|U|^T|U|$ has an eigenvalue greater than one – a contradiction to the assumption $\||U|\|_2 = 1$. The rest of the assertion is trivial. Q.E.D.

To finish the proof of the theorem just use the lemma above and the hermiticity of A. Thus, up to a simultaneous permutation of rows and columns, A is a direct sum of

$$A_i \in \left\{1, -1, \begin{bmatrix} 0 & e^{i\varphi} \\ e^{-i\varphi} & 0 \end{bmatrix}\right\}, \qquad i = 1, \cdots, p .$$

Q.E.D.

The simple upper bounds in Th. 2.2.4 take their minimum n on a much larger class of matrices, namely those with A unitary and commuting with D. Indeed, from the proof of Th. 2.2.8 we immediately obtain

Corollary 2.2.10 *Let H, D, A, and $\widehat{A}$ be as in Th. 2.2.8 such that $\widehat{A}_{11} = 1$. Then the following assertions are equivalent:*

(i) Tr $\widehat{A}\|\widehat{A}^{-1}\|_2 = n$,

(ii) $\widehat{A} = I$,

(iii) A is unitary and commutes with D.

An example of such matrix is given by

$$A = \begin{bmatrix} c & s & 0 \\ s & -c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} d_1 & & \\ & d_1 & \\ & & d_3 \end{bmatrix},$$

where $s^2 + c^2 = 1$ and $d_1, d_3 > 0$. Note that Th. 2.2.7 concerns a certain sort of small perturbations of such matrices. Also note that the only positive definite matrices satisfying Cor. 2.2.10 are again diagonal ones.

The next natural question is: how good are the matrices $H = DAD$ with A unitary, but not necessarily commuting with D? As an example take the matrix $H = DAD$ with

$$A = \frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} d & & & \\ & 1 & & \\ & & 1 & \\ & & & d \end{bmatrix}, \tag{2.2.27}$$

where $d > 0$. Here A is unitary, but it does not commute with D. The eigenvalues of H are $\lambda_1 = d^2$, $\lambda_2 = d$, $\lambda_3 = -d$, $\lambda_4 = 1$, and the corresponding eigenvectors are

$$U = \begin{bmatrix} 1/\sqrt{2} & 1/2 & 1/2 & 0 \\ 0 & -1/2 & 1/2 & 1/\sqrt{2} \\ 0 & -1/2 & 1/2 & -1/\sqrt{2} \\ -1/\sqrt{2} & 1/2 & 1/2 & 0 \end{bmatrix} .$$

If we choose a relative perturbation of the form

$$\delta H = \varepsilon d^2 w w^T , \qquad w = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^T ,$$

and set $H' = H + \delta H$, we have $|\delta H_{ij}| \leq 2\varepsilon |H_{ij}|$ and

$$U^T H' U = \text{diag}\,(d^2, d, -d, 1) + \varepsilon d^2 U^T w w^T U = \begin{bmatrix} d^2 & 0 & 0 & 0 \\ 0 & d + \varepsilon d^2 & \varepsilon d^2 & 0 \\ 0 & \varepsilon d^2 & -d + \varepsilon d^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

Therefore, $\lambda_2' = d(\varepsilon d + \sqrt{1 + \varepsilon^2 d^2})$ and $|\delta\lambda_2|/|\lambda_2| > \varepsilon d$, so H is not well–behaved for large d. Since the matrix

$$HA = \frac{1}{2} \begin{bmatrix} d^2 + d & 0 & 0 & -d^2 + d \\ 0 & d + 1 & d - 1 & 0 \\ 0 & d - 1 & d + 1 & 0 \\ -d^2 + d & 0 & 0 & d^2 + d \end{bmatrix}$$

is symmetric and positive definite, we conclude that $\boldsymbol{|}H\boldsymbol{|} = HA$. For $x = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}^T$ we have

$$\frac{|x|^T |H| |x|}{x^* \boldsymbol{|}H\boldsymbol{|} x} = d ,$$

and thus $\widetilde{C}(H) \to \infty$ as $d \to \infty$. This example shows that the properties of the matrix A *alone* are in general not enough for the good behaviour of the indefinite matrix $H = DAD$. In other words, *contrary to the positive definite case*, an additional scaling $H_1 = D_1 H D_1$ of a well–behaved H need not produce a well–behaved H_1.

Remark 2.2.11 Contrary to the positive definite case, for the indefinite matrices we do not have the result telling us that the matrix behaves well under the perturbations of the type (2.2.10) if and only if $\widetilde{C}(H)$ is small. Moreover, estimating $\widetilde{C}(H)$ with $C(A, \widehat{A})$ is in some cases not appropriate. For example, matrices of the type (2.1.8) behave well under the perturbations of the type (2.2.10) (see the following sections), but are very sensitive to the perturbations of the type (2.2.13) for the standard scaling. Therefore, η_H from Th. 2.2.5 and then, in turn, η_H from Th. 2.2.4 must neccessarily be large and some other kind of analysis is required.

Remark 2.2.12 (*Some singular matrices*). Although Th. 2.2.1 does not require the non–singularity of the unperturbed matrix H, the subsequent theory, as it stands, cannot handle singular matrices. However, for a single matrix of the type

$$H = \begin{bmatrix} \widetilde{H} & 0 \\ 0 & 0 \end{bmatrix}, \qquad \widetilde{H} \text{ non–singular}, \tag{2.2.28}$$

the condition $|\delta H_{ij}| \le \varepsilon |H_{ij}|$ obviously preserves the zero structure and the problem trivially reduces to the perturbation of $\widetilde{H}$ to which our theory can be applied. For a pair H, K with H as above and K positive definite of the form

$$K = \begin{bmatrix} K_{11} & K_{12}^* \\ K_{12} & K_{22} \end{bmatrix}$$

we proceed as follows: from the proof of Th. 2.2.2 we see that the perturbation on K does not need the non–singularity of H. Furthermore, the non–zero eigenvalues of the pair H, K coincide with the eigenvalues of the pair $\widetilde{H}, \widetilde{K}$, where $\widetilde{K} = K_{11} - K_{12}K_{22}^{-1}K_{12}^*$. Thus, in perturbing H the zero eigenvalues do not change and we can apply Th. 2.2.2 to the pair $\widetilde{H}, \widetilde{K}$. We obtain the full assertion of Th. 2.2.2 with $\widetilde{C}(\widetilde{H}, \widetilde{K})$ instead of $\widetilde{C}(H, K)$.

Similarly, Th. 2.2.3 holds where A, $\widehat{A}$ and B are obtained by scaling $\widetilde{H}$, $[\![\widetilde{H}]\!]_{\widetilde{K}}$ and K, respectively. If, in addition, H is positive semidefinite, then $[\![\widetilde{H}]\!]_{\widetilde{K}} = \widetilde{H}$, and Th. 2.2.3 and the subsequent theory hold with $A = \widehat{A}$ and B obtained by scaling $\widetilde{H}$ and K, respectively.

It is readily seen that (2.2.28) is the only form (up to a permutation) of a positive semidefinite matrix whose eigenvalues behave well under the floating–point perturbations. As we shall see later, the indefinite case is more complicated in this aspect.

2.2.1 Perturbation of the eigenvectors

In this subsection we consider the behaviour of the eigenvectors under the perturbations as in Th. 2.2.1. We consider the case of a single non–singular Hermitian matrix H (i.e. $K = I$, $\delta K = 0$). Like in [2, 13], this behaviour is influenced by a relative gap between the neighbouring eigenvalues. Our definition of relative gap is similar but not identical with the ones from [2, 13] which makes an exact comparison of (actually similar) results difficult. Our approach – in contrast to the one from [2, 13] – is that of [20] which deals with the norm–estimates of the spectral projections and thus allows the treatment of multiple and clustered eigenvalues. We also expect our bounds to be better than those of [2, 13], since they do not depend on n.

We now define *the relative gap*, $rg(\lambda)$, for the possibly multiple eigenvalue λ of H. To simplify the notation, as well as the statement and the proof of the following theorem, we shall assume that λ is positive. Negative eigenvalues of H are considered as the positive eigenvalues of the matrix $-H$. By λ_L and λ_R we denote the left and the right neighbour of λ in the spectrum $\sigma(H)$ of H, respectively. We set

$$rg(\lambda) = \begin{cases} \min\left\{\dfrac{\sqrt{\lambda}-\sqrt{\lambda_L}}{\sqrt{\lambda}}, \dfrac{\sqrt{\lambda_R}-\sqrt{\lambda}}{\sqrt{\lambda_R}}\right\} & \text{if } \lambda_L > 0 \ , \\ \min\left\{2(\sqrt{2}-1), \dfrac{\lambda_R-\lambda}{\lambda_R+\lambda}\right\} & \text{otherwise} \ . \end{cases} \tag{2.2.29}$$

Theorem 2.2.13 *Let λ be a positive (possibly multiple) eigenvalue of a non–singular Hermitian matrix H, and let*

$$P = \frac{1}{2\pi i}\int_\Gamma R_\mu d\mu \ , \qquad R_\mu = (\mu I - H)^{-1} \ , \tag{2.2.30}$$

be the corresponding eigenprojection. Here Γ is a curve around λ which separates λ from the rest of the spectrum. Let $P+\delta P$ be the corresponding spectral projection of the matrix $H+\delta H$ with $|x^*\delta H x| \leq \eta x^* |H| x$. Then

$$\|\delta P\|_2 \leq \begin{cases} \dfrac{\eta}{rg(\lambda)} \cdot \dfrac{1}{1-\left(1+\dfrac{1}{rg(\lambda)}\right)\eta} & \text{for } \lambda_L > 0, \ 2\sqrt{\lambda}-\sqrt{\lambda_L} < \sqrt{\lambda_R} \ , \\ \dfrac{\eta}{rg(\lambda)} \cdot \dfrac{1}{1-\dfrac{\eta}{rg(\lambda)}} & \text{otherwise} \ , \end{cases} \tag{2.2.31}$$

provided that the right hand side is positive.

Proof. By setting

$$\Delta = |H|^{-1/2}\delta H |H|^{-1/2}, \qquad z_\mu = R_\mu |H|^{1/2}, \qquad w_\mu = |H|^{1/2} R_\mu |H|^{1/2},$$

we obtain $\|\Delta\|_2 \leq \eta$ and

$$\delta P = \frac{1}{2\pi i}\int_\Gamma z_\mu \Delta \sum_{k=0}^{\infty} (w_\mu \Delta)^k z_\mu d\mu \ .$$

Choosing Γ as a circle around λ with the radius r, we obtain

$$\|\delta P\|_2 \leq r z^2 \eta \frac{1}{1-w\eta}$$

with

$$\begin{aligned} z^2 &= \max_{\mu\in\Gamma} \|z_\mu\|_2^2 = \max_{\mu\in\Gamma} \max_{\nu\in\sigma(H)} \frac{|\nu|}{|\mu-\nu|^2} \\ w &= \max_{\mu\in\Gamma} \|w_\mu\|_2 = \max_{\mu\in\Gamma} \max_{\nu\in\sigma(H)} \frac{|\nu|}{|\mu-\nu|} \ , \end{aligned}$$

provided that $\eta < 1/w$. We obviously have

$$\begin{aligned} z^2 &= \max\left\{\frac{|\lambda_L|}{(\lambda - r - \lambda_L)^2}, \frac{\lambda}{r^2}, \frac{\lambda_R}{(\lambda_R - \lambda - r)^2}\right\} \\ w &= \max\left\{\frac{|\lambda_L|}{\lambda - r - \lambda_L}, \frac{\lambda}{r}, \frac{\lambda_R}{\lambda_R - \lambda - r}\right\} . \end{aligned} \tag{2.2.32}$$

We first consider the case $\lambda_L > 0$. If $2\sqrt{\lambda} - \sqrt{\lambda_L} < \sqrt{\lambda_R}$, then by setting

$$r = \sqrt{\lambda}(\sqrt{\lambda} - \sqrt{\lambda_L}) \tag{2.2.33}$$

we obtain

$$z^2 = \frac{1}{(\sqrt{\lambda} - \sqrt{\lambda_L})^2}, \qquad w \le \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \sqrt{\lambda_L}} + 1 .$$

Here we used our assumption and the fact that both rightmost terms in (2.2.32) are decreasing functions of λ_R. Therefore,

$$\|\delta P\|_2 \le \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \sqrt{\lambda_L}}\eta \frac{1}{1 - \left(1 + \dfrac{\sqrt{\lambda}}{\sqrt{\lambda} - \sqrt{\lambda_L}}\right)\eta} ,$$

and (2.2.31) holds. Positivity of the right hand side of (2.2.31) justifies, in turn, our choice of the same Γ in the definitions of P and $P + \delta P$ as follows: perturbation theorem for the eigenvalues implies that λ_L can increase to at most $\lambda_L(1 + \eta)$, λ_R can decrease to at least $\lambda_R(1 - \eta)$, and the eigenvalues of $H + \delta H$ which correspond to λ remain in the interval $[\lambda(1 - \eta), \lambda(1 + \eta)]$. Positivity of the right hand side of (2.2.31) always implies $rg(\lambda) > \eta$. This, together with our choice of r, implies that Γ contains no points of the spectrum of $H + \delta H$ and that the interior of Γ contains exactly those eigenvalues of $H + \delta H$ which correspond to λ. This remark holds for the subsequent cases, as well.

If $2\sqrt{\lambda} - \sqrt{\lambda_L} \ge \sqrt{\lambda_R}$, then by setting

$$r = \sqrt{\lambda}(\sqrt{\lambda_R} - \sqrt{\lambda})$$

we obtain

$$z^2 = \frac{1}{(\sqrt{\lambda_R} - \sqrt{\lambda})^2}, \qquad w = \frac{\sqrt{\lambda_R}}{\sqrt{\lambda_R} - \sqrt{\lambda}} .$$

Here we used our assumption and the fact that both leftmost terms in the right hand side of (2.2.32) are increasing functions of $\lambda_L > 0$. Therefore,

$$\|\delta P\|_2 \le \frac{\sqrt{\lambda}}{\sqrt{\lambda_R} - \sqrt{\lambda}}\eta \frac{1}{1 - \dfrac{\sqrt{\lambda_R}}{\sqrt{\lambda_R} - \sqrt{\lambda}}\eta} ,$$

and (2.2.31) holds. If λ is the largest positive eigenvalue (i.e. λ_R does not exist), then by setting r as in (2.2.33) we obtain

$$z^2 = \frac{1}{(\sqrt{\lambda} - \sqrt{\lambda_L})^2}, \qquad w = \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \sqrt{\lambda_L}} ,$$

and (2.2.31) holds again.

If $\lambda_L < 0$ or if λ_L does not exist, we proceed as follows: if $rg(\lambda) = 2(\sqrt{2} - 1)$ (if λ_R exists, this implies $\lambda(4\sqrt{2} + 5) \leq \lambda_R$), then by setting

$$r = 2(\sqrt{2} - 1)\lambda$$

we obtain

$$z^2 = \frac{1}{4(\sqrt{2} - 1)^2\lambda} , \qquad w = \frac{1}{2(\sqrt{2} - 1)} ,$$

so (2.2.31) holds. Finally, if $rg(\lambda) = (\lambda_R - \lambda)/(\lambda_R + \lambda)$, then by setting

$$r = \lambda \frac{\lambda_R - \lambda}{\lambda_R + \lambda}$$

we obtain

$$z^2 = \frac{1}{\lambda} \left(\frac{\lambda_R + \lambda}{\lambda_R - \lambda} \right)^2 , \qquad w = \frac{\lambda_R + \lambda}{\lambda_R - \lambda} ,$$

and (2.2.31) holds again. Q.E.D.

2.3 Perturbations by factors

In this section we consider perturbations of the eigenvalues of a single Hermitian matrix H given in a factorized form

$$H = GJG^* , \tag{2.3.1}$$

where G need not to be square but must have full column rank, whereas J is Hermitian and non-singular. A typical J is

$$J_1 = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} . \tag{2.3.2}$$

Here the unit blocks need not have the same dimension and one of them may be void. Such factorization is obtained e.g. by the symmetric indefinite decomposition of Chap. 4. We consider the changes of the eigenvalues and eigenvectors of H under perturbation of G while J remains unchanged. Here it is natural to use the one–sided scaling $G = BD$.

For $J = I$ the problem reduces to considering singular values of G. We reproduce the result of [13] with somewhat better constants. The same technique allows an interesting floating–point estimate for the eigenvalues of G (see [32]).

The section is organized as follows. Th. 2.3.1 gives a general perturbation theory, while Th. 2.3.2 applies this theory to the floating–point perturbations. In the following discussion we simplify the perturbation bounds analogously to the previous section. As an application we derive floating–point perturbation estimates for some classes of matrices not covered by Sect. 2.2.

Theorem 2.3.1 *Let $H = GJG^*$ be as above and let $H' = G'JG'^*$ with*

$$G' = G + \delta G \ , \qquad \|\delta G x\|_2 \le \eta \|Gx\|_2 \ , \tag{2.3.3}$$

for all $x \in \mathbf{C}^n$ and some $\eta < 1$. Then H and H' have the same inertia and their non–vanishing eigenvalues λ_k, λ'_k, respectively, satisfy the inequalities

$$(1-\eta)^2 \le \frac{\lambda'_k}{\lambda_k} \le (1+\eta)^2 \ . \tag{2.3.4}$$

PROOF. We first show that the non–vanishing eigenvalues of H coincide with the eigenvalues of the pair G^*G, J^{-1}. Indeed, since G^*G is positive definite, there exists a non–singular F such that

$$F^*G^*GF = \Delta \tag{2.3.5}$$

and

$$F^*J^{-1}F = J_1 \tag{2.3.6}$$

are diagonal matrices, and J_1 is from (2.3.2). Then the eigenvalues of the pair G^*G, J^{-1} are found on the diagonal of $\Delta J_1 = J_1\Delta$. Set $U = GF\Delta^{-1/2}$. By (2.3.5) we have $U^*U = I$ (but not necessarily $UU^* = I$). Using (2.3.5) and (2.3.6) we obtain

$$\begin{aligned} HU &= GJG^*GF\Delta^{-1/2} = GJF^{-*}F^*G^*GF\Delta^{-1/2} \\ &= GJF^{-*}\Delta^{1/2} = GFF^{-1}JF^{-*}\Delta^{1/2} \\ &= GF(F^*J^{-1}F)^{-1}\Delta^{1/2} = UJ_1\Delta \ . \end{aligned}$$

Thus, the columns of U are eigenvectors of H and the eigenvalues of H coincide with those of G^*G, J^{-1}. Furthermore, $U^*x = 0$ implies $Hx = 0$, so the eigenvalues of G^*G, J^{-1} are exactly all non–vanishing eigenvalues of H. By (2.3.3) we have

$$(1-\eta)\|Gx\|_2 \le \|G'x\|_2 \le (1+\eta)\|Gx\|_2 \ , \tag{2.3.7}$$

so that everything said for H holds for H' as well. In particular, H and H' have the same inertia. Now square (2.3.7), use the monotonicity property from the proof of Th. 2.2.1 for the pairs J^{-1}, G^*G and J^{-1}, G'^*G', and take reciprocals in (2.2.7) and (2.2.8). Q.E.D.

We now consider floating–point perturbations and scalings.

Theorem 2.3.2 *Let $H = GJG^*$ be as in (2.3.1) and (2.3.2). Let $H' = G'JG'^*$ where $G' = G + \delta G$, and for all i, j and some $\varepsilon > 0$ holds*

$$|\delta G_{ij}| \le \varepsilon |G_{ij}| \ . \tag{2.3.8}$$

Set

$$\eta \equiv \frac{\varepsilon \||B|\|_2}{\sigma_{min}(B)} \ ,$$

where $B = GD^{-1}$, D is diagonal and positive definite, and $\sigma_{min}(B)$ is the smallest singular value of B. If $\eta < 1$ then the assumptions of Th. 2.3.1 are fulfilled, hence its assertion holds.

PROOF. For $x \in \mathbf{C}^n$ we have

$$\begin{aligned} \|\delta G x\|_2 &\le \varepsilon \||B|D|x|\|_2 \le \varepsilon \||B|\|_2 \|Dx\|_2 \\ &\le \frac{\varepsilon \||B|\|_2 \|BDx\|_2}{\sigma_{min}(B)} = \frac{\varepsilon \||B|\|_2 \|Gx\|_2}{\sigma_{min}(B)} \ . \end{aligned}$$

Q.E.D.

By $\||B|\|_2 \ge \|B\|_2$ we have

$$\frac{\||B|\|_2}{\sigma_{min}(B)} \ge \frac{\sigma_{max}(B)}{\sigma_{min}(B)} \ge 1 \ .$$

Here both inequalities go over into equalities, if and only if B has the property

$$B^*B = \gamma^2 I \ , \qquad \gamma > 0 \ , \qquad \||B|\|_2 = \gamma \ ,$$

or, equivalently (Lemma 2.2.9), if and only if $|B|^T|B| = \gamma^2 I$. Similarly as in Sect. 2.2 we can make a simplifying estimate

$$\frac{\||B|\|_2}{\sigma_{min}(B)} \le \frac{(\mathrm{Tr}\ (B^*B))^{1/2}}{\sigma_{min}(B)} \ ,$$

so that

$$\eta = \frac{\varepsilon (\mathrm{Tr}\ (B^*B))^{1/2}}{\sigma_{min}(B)} < 1 \tag{2.3.9}$$

again implies (2.3.3) and therefore (2.3.4). This yields a new "condition number"

$$\frac{(\mathrm{Tr}\ (B^*B))^{1/2}}{\sigma_{min}(B)} \ge \sqrt{n} \ ,$$

where the equality is attained if and only if $B^*B = \gamma^2 I$. For the standard scaling where $(B^*B)_{ii} = 1$ the relation (2.3.4) is implied by

$$\eta = \frac{\varepsilon \sqrt{n}}{\sigma_{min}(B)} < 1 \ . \tag{2.3.10}$$

This is a slight improvement over [13] for the case $J = I$ (our constant is $\sqrt{n}$ times better).

For $J = I$ (or $J = -I$) we can handle the matrix $H = GG^*$ in two ways. If G has full column rank, then we apply our theory as described in Theorems 2.3.1 and 2.3.2. If G^* has full column rank, then we apply our theory to the matrix $\widehat{H} = G^*G$, whose non–vanishing eigenvalues are the eigenvalues of H. In the indefinite case ($J \neq \pm I$) the situation is different. The following simple example illustrates this important asymmetry. Take

$$G = [a, b] \;, \qquad \delta G = [\delta a, \delta b] \;.$$

Our theory cannot be applied to

$$H = GG^* = |a|^2 + |b|^2 \;,$$

but it works on

$$H = G^*G \;,$$

where $G^* = \widetilde{B}\widetilde{D}$, $\widetilde{B} = \left[\begin{array}{cc} 1/\sqrt{2} & 1/\sqrt{2} \end{array}\right]^T$, $\widetilde{D} = (|a|^2 + |b|^2)^{1/2}$, thus giving $\eta = \varepsilon$ independently of a and b. On the contrary, no theory can "save" the matrix

$$H = G \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} G^* = |a|^2 - |b|^2$$

since

$$\frac{|a + \delta a|^2 - |b + \delta b|^2}{|a|^2 - |b|^2}$$

cannot be made small uniformly in a, b if $|\delta a/a|$ and $|\delta b/b|$ are sufficiently small.[5]

Similarly as in Th. 2.2.5 we can show that a perturbation result holds under perturbations δG defined by

$$|\delta G_{ij}| \leq \varepsilon D_j \qquad \text{for all } i, j,$$

where D is a scaling. The above type of perturbation is less restrictive than (2.3.8), e.g. it allows us to change zero elements. We have

$$\begin{aligned} \|\delta G x\|_2^2 &= \sum_{i,j,k} \bar{x}_i \delta \bar{G}_{ji} \delta G_{jk} x_k \leq n \left(\varepsilon \sum_j |D_j x_j| \right)^2 \\ &\leq n^2 \varepsilon^2 \|Dx\|_2^2 \leq \frac{n^2 \varepsilon^2 \|Gx\|_2^2}{\lambda_{min}(B^*B)} \;, \end{aligned}$$

hence (2.3.4) is implied by

$$\eta = \frac{n\varepsilon}{\sigma_{min}(B)} < 1 \;. \qquad (2.3.11)$$

[5] In the indefinite case the values $\mu_k = \sqrt{|\lambda_k|}\text{sign } \lambda_k$ are called the *hyperbolic singular values* [21].

Similarly one shows that the estimate (2.3.4) is obtained under the perturbation

$$\delta G = \delta B D \ , \qquad \eta = \frac{\|\delta B\|_2}{\sigma_{min}(B)} < 1 \ . \tag{2.3.12}$$

The following two examples show how Th. 2.3.2 can accomodate floating–point perturbations of some matrices which, in spite of Rem. 2.1, cannot be handled by the theory from Sect. 2.2. For the first example set

$$H = \begin{bmatrix} A & F^* \\ F & 0 \end{bmatrix} , \tag{2.3.13}$$

where A is of order m and $m \le n - m$. Then $H = GJG^*$ with

$$G = \begin{bmatrix} \frac{1}{2}A & I \\ F & 0 \end{bmatrix} , \qquad J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} ,$$

where the unit blocks have the order m. Now the perturbation δH of H with $|H_{ij}| \le \varepsilon |H_{ij}|$ gives rise to a perturbation δG of G with $|\delta G_{ij}| \le \varepsilon |G_{ij}|$, and Th. 2.3.2 holds e.g. with

$$B = \begin{bmatrix} \frac{1}{2}A & I \\ F & 0 \end{bmatrix} \begin{bmatrix} D^{-1} & 0 \\ 0 & I \end{bmatrix} ,$$

where D is the standard scaling

$$D_{ii}^2 = \left(\frac{1}{4}A^2 + F^*F\right)_{ii} .$$

The requirement that G have full column rank is equivalent to the same requirement on F. Note that this allows singular matrices H.

An even simpler case is the one with $A = 0$. Then we can apply the theory to

$$H = \begin{bmatrix} 0 & F^* \\ F & 0 \end{bmatrix} = \begin{bmatrix} 0 & I \\ F & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & F^* \\ I & 0 \end{bmatrix} , \tag{2.3.14}$$

as well as to

$$H = \begin{bmatrix} 0 & F \\ F^* & 0 \end{bmatrix} = \begin{bmatrix} 0 & F \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ F^* & 0 \end{bmatrix} .$$

In any case, the non–vanishing eigenvalues of H coincide with the singular values of F taken with both signs. Now $|\delta G_{ij}| \le \varepsilon |G_{ij}|$ means $|\delta F_{ij}| \le \varepsilon |F_{ij}|$ and we can apply our theory in two ways:

(i) take e.g. (2.3.14) and use Th. 2.3.2 to obtain (2.3.4) with

$$\eta = \frac{\||B|\|_2}{\sigma_{min}(B)} ,$$

where $B = FD^{-1}$, $(B^*B)_{ii} = 1$, or

(ii) apply Th. 2.3.2 to the factorized matrix FF^* (with the same B) which yields a slightly better estimate

$$(1-\eta)^2 \le \frac{\lambda_k'^2}{\lambda_k^2} \le (1+\eta)^2 \ .$$

In both cases the theory from Sect. 2 would require both BB^* and B^*B to scale well, which is certainly a further unnecessary restriction.

As a second example set

$$H = \begin{bmatrix} a & b & c \\ b & 0 & 0 \\ c & 0 & \alpha^2 \end{bmatrix} .$$

We can e.g. decompose H as

$$H = \begin{bmatrix} a/2 & 1 & 0 \\ b & 0 & 0 \\ c & 0 & \alpha \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a/2 & b & c \\ 1 & 0 & 0 \\ 0 & 0 & \alpha \end{bmatrix} . \tag{2.3.15}$$

Now $|\delta H_{ij}| \le \varepsilon |H_{ij}|$ again implies $|\delta G_{ij}| \le \varepsilon |G_{ij}|$ and we can apply our theory as in the previous example. For e.g. $a = b = c = 1$ we obtain $\| |B| \|_2 \|B^{-1}\|_2 = 2 + \sqrt{3}$, independently of α. Especially, if α is small then even the absolutely smallest eigenvalue $\alpha^2/2 + O(\alpha^4)$ is well defined by the matrix elements of H. On the other side, the theory from Sect. 2 applied to H, I gives nothing useful here. Indeed, as $\alpha \to 0$ we have

$$|H| = \frac{1}{3} \begin{bmatrix} 5 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} + O(\alpha^2) \ , \tag{2.3.16}$$

so that $C(A, \widehat{A}) = O(1/\alpha^2)$. Moreover, numerical experiments show that $\widetilde{C}(H) > 1/|\alpha|$. Another very interesting approach to matrices of the above type is given by Demmel and Gragg [11].

2.3.1 Perturbation of the eigenvectors

In this subsection we give the perturbation bounds for the eigenvectors of the non–singular Hermitian matrix

$$H = GJG^* \ ,$$

under the perturbations as in Th. 2.3.1, i.e.

$$\|\delta G x\|_2 \le \eta \|Gx\|_2 \ ,$$

for every x.

As in [2, 13] and Subsect. 2.2.1, the behaviour of the eigenvectors is influenced by a relative gap between the neighbouring eigenvalues. Our definition of relative gap is similar but not identical with the one from [2, 13] and Subsect. 2.2.1, and our approach is again that of [20].

We now define *the relative gap*, $rg_G(\lambda)$, and the eigenprojection P for the possibly multiple eigenvalue λ of H. To simplify the notation, as well as the statement and the proof of the following theorem, we shall assume that λ is positive. Negative eigenvalues of H are considered as the positive eigenvalues of the matrix $-H$. By λ_L and λ_R we denote the left and the right neighbour of λ in the spectrum $\sigma(H)$ of H, respectively. We set

$$\begin{aligned} rg_G(\lambda) &= \min\left\{1, \frac{\lambda_R-\lambda}{\lambda_R+\lambda}, \frac{\lambda-\lambda_L}{\lambda+\lambda_L}\right\} , \\ P &= \frac{1}{2\pi i}\int_\Gamma R_\mu d\mu , \qquad R_\mu = (\mu I - H)^{-1} , \end{aligned} \tag{2.3.1}$$

where Γ is a curve around λ which separates λ from the rest of the spectrum of H. Here, as well as throughout the section, the terms containing λ_L, λ_R are defined if λ_L, λ_R exist and are positive, respectively.

Theorem 2.3.3 *Let λ be a positive (possibly multiple) eigenvalue of a non–singular Hermitian matrix $H = GJG^*$, and let P be the corresponding eigenprojection. Let P' be the corresponding spectral projection of the matrix $H' = G'J(G')^*$, where $G' = G + \delta G$ and $\|\delta G x\|_2 \leq \eta \|Gx\|_2$ for every x.*

Then

$$\|P'-P\|_2 \leq \frac{4\bar\eta}{rg_G(\lambda)} \cdot \frac{1}{1-\dfrac{3\bar\eta}{rg_G(\lambda)}} , \tag{2.3.2}$$

where

$$\bar\eta = \eta(2+\eta) ,$$

provided that the right hand side in (2.3.2) is positive.

Proof. Since H and H^{-1} have the same eigenvectors, we can define P as

$$P = \frac{1}{2\pi i}\int_\Gamma S_\mu d\mu , \qquad S_\mu = (\mu I - H^{-1})^{-1} ,$$

where Γ is now a curve around $1/\lambda$ which separates $1/\lambda$ from the rest of the spectrum of H^{-1}. Therefore,

$$P'-P = \frac{1}{2\pi i}\int_\Gamma (S'_\mu - S_\mu) d\mu , \tag{2.3.3}$$

where

$$S'_\mu = (\mu I - H'^{-1})^{-1} .$$

We can write

$$S_\mu = (\mu I - G^{-*}JG^{-1})^{-1} = G(\mu G^* G - J)^{-1} G^* \equiv G T_\mu G^* \ , \tag{2.3.4}$$

and analogously

$$S'_\mu = G' T'_\mu (G')^* \ , \qquad T'_\mu = (\mu (G')^* G' - J)^{-1} \ .$$

Now

$$S'_\mu - S_\mu = G(T'_\mu - T_\mu) G^* + \Phi \ , \tag{2.3.5}$$

where

$$\Phi = \delta G T'_\mu G^* + G T'_\mu \delta G^* + \delta G T'_\mu \delta G^* \ . \tag{2.3.6}$$

Further,

$$G(T'_\mu - T_\mu)G^* = G T_\mu (T_\mu^{-1} - (T'_\mu)^{-1}) T'_\mu G^* = G T_\mu \mu \gamma T'_\mu G^* \ , \tag{2.3.7}$$

where

$$\gamma = -\delta G^* G - G^* \delta G - \delta G^* \delta G \ .$$

Inserting

$$\gamma = G^* \Delta G \tag{2.3.8}$$

and (2.3.4) into (2.3.7), we obtain

$$G(T'_\mu - T_\mu)G^* = S_\mu \mu \Delta G T'_\mu G^* \ . \tag{2.3.9}$$

Using (2.3.4) and (2.3.8), we obtain

$$\begin{aligned} G T'_\mu G^* &= G(T_\mu^{-1} - \mu G^* \Delta G)^{-1} G^* \\ &= G(I - \mu T_\mu G^* \Delta G)^{-1} T_\mu G^* \\ &= G((T_\mu G^*)^{-1} - \mu \Delta G)^{-1} \\ &= S_\mu (I - \mu \Delta S_\mu)^{-1} \ . \end{aligned} \tag{2.3.10}$$

Inserting (2.3.10), (2.3.9), (2.3.6) and (2.3.5) into (2.3.3), we obtain

$$\begin{aligned} P' - P = \ & \frac{1}{2\pi i} \int_\Gamma [\mu S_\mu \Delta S_\mu (I - \mu \Delta S_\mu)^{-1} \\ & + \delta G G^{-1} S_\mu (I - \mu \Delta S_\mu)^{-1} + S_\mu (I - \mu \Delta S_\mu)^{-1} G^{-*} \delta G^* \\ & + \delta G G^{-1} S_\mu (I - \mu \Delta S_\mu)^{-1} G^{-*} \delta G^*] d\mu \ . \end{aligned} \tag{2.3.11}$$

Our assumption on δG and the definition of Δ in (2.3.8) imply

$$\begin{aligned} \|\delta G G^{-1}\|_2 &\leq \eta \ , \\ \|\Delta\|_2 &\leq 2\|\delta G G^{-1}\|_2 + \|\delta G G^{-1}\|_2^2 \leq \bar{\eta} \ . \end{aligned}$$

Choosing Γ as a circle around $1/\lambda$ with radius r, taking norms in (2.3.11), and using the above relations, we obtain

$$\|P' - P\|_2 \leq rz(w+1)\bar{\eta}\frac{1}{1-\bar{\eta}w} \ , \tag{2.3.12}$$

where

$$\begin{aligned} w &= \max_{\mu\in\Gamma} \|\mu S_\mu\|_2 = \max_{\mu\in\Gamma} \max_{\nu\in\sigma(H^{-1})} \frac{|\mu|}{|\mu-\nu|} \ , \\ z &= \max_{\mu\in\Gamma} \|S_\mu\|_2 = \max_{\mu\in\Gamma} \max_{\nu\in\sigma(H^{-1})} \frac{1}{|\mu-\nu|} \ . \end{aligned}$$

Since Γ is a circle, the maxima in the above relations are attained for μ's which lie on the real axis.

If λ_R exists, then we choose r as

$$r = \frac{1}{2}\min\left\{\frac{1}{\lambda} - \frac{1}{\lambda_R}, \frac{1}{\lambda_L} - \frac{1}{\lambda}\right\} \ ,$$

and if λ_R does not exist, then we choose r as

$$r = \frac{1}{2}\min\left\{\frac{1}{\lambda}, \frac{1}{\lambda_L} - \frac{1}{\lambda}\right\} \ .$$

It is easy to see that we always have

$$z = \frac{1}{r} \ .$$

Since $\mu = 1/\lambda \pm r$, we have

$$w = \max\left\{\frac{1/\lambda - r}{1/\lambda - r - 1/\lambda_R}, \frac{1/\lambda + r}{r}, \frac{1/\lambda + r}{1/\lambda_L - 1/\lambda - r}\right\} \ .$$

Now if $r = (1/\lambda - 1/\lambda_R)/2$, then

$$w = 1 + \frac{2}{\dfrac{\lambda_R - \lambda}{\lambda_R}} \leq 1 + \frac{2}{rg_G(\lambda)} \leq \frac{3}{rg_G(\lambda)} \ ,$$

and (2.3.2) follows by inserting this and $z = 1/r$ into (2.3.12).

If $r = (1/\lambda_L - 1/\lambda)/2$, then

$$w = \frac{\lambda - \lambda_L}{\lambda + \lambda_L} \leq \frac{1}{rg_G(\lambda)} \ ,$$

and (2.3.2) follows by inserting this and $z = 1/r$ into (2.3.12).

Finally, if $r = 1/(2\lambda)$ (λ_R does not exist), then $w = 3$ and (2.3.2) follows by inserting this and $z = 1/r$ into (2.3.12).

Positivity of the right hand side of (2.3.2) justifies, in turn, our choice of the same Γ in the definitions of P and P' in (2.3.3) as follows: perturbation theorem for the eigenvalues implies that $1/\lambda_R$ can increase to at most $1/(\lambda_R(1-\eta)^2)$, $1/\lambda_L$ can decrease to at least $1/(\lambda_L(1+\eta)^2)$ and the eigenvalues of H'^{-1} which correspond to $1/\lambda$ remain in the interval $[1/(\lambda(1+\eta)^2), 1/(\lambda(1-\eta)^2)]$. Positivity of the right hand side of (2.3.2) always implies $rg_G(\lambda) > 6\eta$. This, together with our choice of r, implies that Γ contains no points of the spectrum of H'^{-1} and that the interior of Γ contains exactly those eigenvalues of H'^{-1} which correspond to $1/\lambda$. Q.E.D.

Remark 2.3.4 It is possible to prove theorem similar to Th. 2.3.3 for a cluster of eigenvalues, as well. All eigenvalues of the cluster must be either positive or negative. The relative gap for the cluster is then defined using λ_L (λ_R) and the leftmost (rightmost) member of the cluster, respectively. The $r \cdot z$ term of (2.3.12) is then larger than 1, and smaller than the inverse of the relative gap of the cluster.

Note that we can in some cases actually prove better bounds than (2.3.2), but the differences are small, so we have decided to state and to prove the simpler version. Th. 2.3.3 is a generalization of the corresponding results from [13] since it allows $J \neq I$ and multiple eigenvalues.

Now suppose that λ and λ' are both simple. Let v and $v' = v + \delta v$ be the corresponding unit eigenvectors, and let ϕ be the angle between them. Then $P = vv^*$, $P' = v'(v')^*$, and $P' - P$ is a matrix of rank 2 with the non–trivial eigenvalues, say, γ_1 and γ_2. Since $\mathrm{Tr}\,(P' - P) = 0$, we have $|\gamma_1| = |\gamma_2| \equiv \gamma$. Now

$$2\gamma^2 = \mathrm{Tr}\,[(P' - P)(P' - P)] = 2\sin^2\phi \ ,$$

so that

$$\|P' - P\|_2 = |\sin\phi| \ .$$

This finally implies

$$\|\delta v\|_2 = 2|\sin(\phi/2)| \leq \sqrt{2}\|P' - P\|_2 \ . \qquad (2.3.13)$$

Combining the above relation with Th. 2.3.3 we obtain the bound on $\|\delta v\|_2$. We expect this bound to compare favourably to the corresponding bounds from [2, 13] since it does not contain the factors $(n-1)$ or $(n-1)^{1/2}$, respectively.

Chapter 3

Error analysis of the J–orthogonal Jacobi methods

3.1 J–orthogonal Jacobi method

The J–orthogonal Jacobi method solves the problem

$$Hx = \lambda Jx, \qquad x \neq 0, \tag{3.1.1}$$

where $H = (H_{ij})$ is a positive definite matrix,

$$J = I_{npos} \oplus (-I_{n-npos}) \ ,$$

$npos$ is the number of the positive, and $n - npos$ is the number of the negative eigenvalues of the pair H, J. The algorithm, including the convergence theory, was proposed by Veselić [29]. For the sake of completeness we give the algorithm of the method and state the known convergence results.

In Chap. 2, we showed that there exists a nonsingular matrix V which simultaneously diagonalizes H and J in the manner that

$$V^T HV = D, \qquad V^T JV = J, \tag{3.1.2}$$

where $D = (D_i)$ is a positive definite diagonal matrix. The eigenvalues of the pair H, J are the values $D_i \cdot J_i$ and the eigenvectors are the corresponding columns of V. The matrices for which $V^T JV = J$ are called *J–orthogonal* and they form a multiplicative group. (For a fixed J, of course.)

The J–orthogonal Jacobi method consists of an iterative application of the congruence transformation

$$H' = C^T HC \ ,$$

where C is the J–orthogonal plane rotation. From now on let $\widehat{A}$ denote the 2×2 pivot submatrix of the square matrix A. The matrix C is defined as

$$\widehat{C} = \begin{bmatrix} c_{ii} & c_{ij} \\ c_{ji} & c_{jj} \end{bmatrix},$$

and the non–displayed elements are those of the identity matrix. The pair (i,j) is the pivot pair. The J–orthogonality of the matrix C implies that

$$\begin{bmatrix} c_{ii} & c_{ij} \\ c_{ji} & c_{jj} \end{bmatrix} = \begin{cases} \begin{bmatrix} ch & sh \\ sh & ch \end{bmatrix}, & \text{for } 1 \leq i \leq npos < j \leq n, \\ \\ \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix}, & \text{otherwise .} \end{cases}$$

Here $ch = \cosh y$, $sh = \sinh y$, $cs = \cos x$ and $sn = \sin x$ for some y and x, respectively. These two types of rotations are called the *hyperbolic* and the *trigonometric rotation*, respectively. The parametar x or y is chosen so that the i,j–element of the transformed matrix is annihilated. Let

$$\widehat{H} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}.$$

Then

$$\tan 2x = \frac{2c}{b-a}, \qquad -\frac{\pi}{4} \leq x \leq \frac{\pi}{4},$$

or

$$\tanh\ 2y = -\frac{2c}{a+b}.$$

We obtain the following algorithm (in the notation of [13]):

Algorithm 3.1.1 *Two-sided $J-$orthogonal Jacobi method for the problem (3.1.1). tol is a user defined stopping criterion. The matrix V whose columns return the computed eigenvectors initially contains the identity.*

repeat
    for all pairs $i < j$
    / compute the parameter hyp: $hyp = 1$ for the hyperbolic and*
        *$hyp = -1$ for the trigonometric rotation, respectively */*
        if $1 \leq i \leq npos < j \leq n$ then
            $hyp = 1$
        else
            $hyp = -1$
        endif
        / compute the $J-$orthogonal Jacobi rotation which diagonalizes*
$$\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad */$$
            $\zeta = -hyp * (b + hyp * a)/(2c)$
            $t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2 - hyp})$
            $h = \sqrt{1 - hyp * t^2}$
            $cs = 1/h$
            $sn = t/h$
            $sn1 = hyp * sn$
            / update the 2 by 2 pivot submatrix */*
            $H_{ii} = a + hyp * c * t$
            $H_{jj} = b + c * t$
            $H_{ij} = H_{ji} = 0$
            / update the rest of rows and columns i and j */*
            for $k = 1$ to n except i and j
                $tmp = H_{ik}$
                $H_{ik} = cs * tmp + sn1 * H_{jk}$
                $H_{jk} = sn * tmp + cs * H_{jk}$
                $H_{ki} = H_{ik}$
                $H_{kj} = H_{jk}$
            endfor
            / update the eigenvector matrix V */*
            for $k = 1$ to n
                $tmp = V_{ki}$
                $V_{ki} = cs * tmp + sn1 * V_{kj}$
                $V_{kj} = sn * tmp + cs * V_{kj}$
            endfor
    endfor
until convergence (all $|H_{ij}|/(H_{ii}H_{jj})^{1/2} \leq tol$)
/ the computed eigenvalues of the pair H, J are $\lambda_j = H_{jj}J_{jj}$ */*

/ the computed eigenvectors of the pair H, J are the columns of the final matrix V */*

Our algorithm is essentially the standard one introduced by Rutishauser [22]. The formulae for the hyperbolic case are derived in the same manner as for the trigonometric one [29]. In the following section we analyse this (simple) version of the algorithm. We omitt enhancements like delayed updates of the diagonals and fast rotations, to make the analysis clearer. Analysis of the fast rotations is given for the implicit method in Sect. 3.4. One of the differences between our algorithm and the standard one is the stopping criterion. This criterion is also used in [13, 29, 31]. Our justification of this criterion is the same as in [13]: according to Th. 2.2.1, the accuracy of the eigenvalues depends on $1/\lambda_{min}(A)$ (or $\kappa(A)$) and not on $\kappa(H)$, so that we set H_{ij} to zero only if $|H_{ij}|/(H_{ii}H_{jj})^{1/2}$ is small, not just if $|H_{ij}|/\max_{kl}|H_{kl}|$ is small.

One difference between trigonometric and hyperbolic rotations is that $\mathrm{Tr}\,(H') = \mathrm{Tr}\,(H)$ after trigonometric, and $\mathrm{Tr}\,(H') < \mathrm{Tr}\,(H)$ after hyperbolic rotation. Using this trace reduction argument Veselić [29] proved that the hyperbolic parameter t tends to zero. The second difference is that the condition of the transformation matrix is in the trigonometric case one, while in the hyperbolic case it can be large. Note, however, that

$$|\tanh y| \leq \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\ ,$$

where A is the scaled matrix, i.e. $H = DAD$, $\mathrm{diag}\,(A) = I$. Moreover, if G, J is the output of the symmetric indefinite decomposition, then the scaled condition of the matrix G^TG is generally small (see Sect. 4.4, Chap. 5), and it does not grow much during the Jacobi process (see Sect. 3.2.2, Chap. 5), so the hyperbolic parameters are generally moderate. In Subsect. 3.2.1 we show how to modify hyperbolic rotations in order to bound the condition of the transformation matrix. This modification improves the theoretical bounds, but it does not seem to be of importance in practice.

Veselić [29] proved that the $J-$orthogonal Jacobi method is globally convergent for the optimal strategy, threshold strategies, row–cyclic strategy, and all other strategies which are equivalent to the row–cyclic one (for example, the modulus parallel strategy [18]). He also proved a very interesting fact that all $J-$orthogonal matrices V which satisfy (3.1.2) have the same condition number. Moreover, if V_1 and V_2 are two such matrices, then

$$V_2 = V_1U, \qquad U = \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix},$$

where U_1, U_2 are othogonal matrices of order m, $n-m$, respectively.

Drmač and Hari [15] proved that the $J-$orthogonal Jacobi method is quadratically convergent.

3.2 Error bounds for the eigenvalues

In this section we prove that the two-sided J−orthogonal Jacobi method in floating–point arithmetic applied to the problem (3.1.1) computes eigenvalues with the error bounds of Chap. 2. Since the computed eigenvector matrix is not orthogonal and is not needed when applying our algorithm to a single indefinite matrix, we do not investigate the accuracy of the computed eigenvectors.

Let $H_0 = D_0 A_0 D_0$ be the initial matrix, and $H_m = D_m A_m D_m$ where H_m is obtained from H_{m-1} by applying a single J−orthogonal Jacobi rotation. Here D_m is diagonal and A_m has unit diagonal as before. All the error bounds in this section contain the quantities $1/\lambda_{min}(A)$ (or $\kappa(A_m)$), whereas the perturbation bounds of Chap. 2 are proportional to $\kappa(A_0)$. Therefore, our claim that J−orthogonal Jacobi method solves the eigenproblem as accurately as predicted in Chap. 2 depends, as in [13], on the ratios $\max_m \lambda_{min}(A_0)/\lambda_{min}(A_m)$ (or $\max_m \kappa(A_m)/\kappa(A_0)$) being modest in size. Note that the convergence of H_m to diagonal form is equivalent to the convergence of A_m to the identity, or $\kappa(A_m)$ to 1. Thus we expect $\kappa(A_m)$ to be less than $\kappa(A_0)$ eventually. Demmel and Veselić [13] have overwhelming numerical evidence that in the positive definite case ($J = I$) the above ratios are modest in size. Our experiments of Chap. 5 reveal the same for $J \neq I$. Our theoretical understanding of why these ratios are so small is somewhat weaker; we present our theoretical bounds in Subsect. 3.2.2.

The section is organized as follows: we first show that one step of the method satisfies the perturbation bounds of Chap. 2, and that we can extend this result to an overall error bound (modulo the assumption that the quotients $\max_m \lambda_{min}(A_0)/\lambda_{min}(A_m)$ are modest). In Subsect. 3.2.1 we show how to modify the method in order to bound potentially large hyperbolic angles, which, in turn, results in better error bounds.

We now present our model of the finite precision floating–point arithmetic. The floating–point result $fl(\cdot)$ of the operation $(\cdot)$ is given by [33, 13]

$$\begin{aligned} fl(a \pm b) &= a(1+\varepsilon_1) \pm b(1+\varepsilon_2) \\ fl(a \times b) &= (a \times b)(1+\varepsilon_3) \\ fl(a/b) &= (a/b)(1+\varepsilon_4) \\ fl(\sqrt{a}) &= \sqrt{a}(1+\varepsilon_5) \end{aligned} \tag{3.2.1}$$

where $|\varepsilon_i| \leq \varepsilon$, and $\varepsilon \ll 1$ is the machine precision. This is somewhat more general than the usual model which uses $fl(a \pm b) = (a \pm b)(1+\varepsilon_1)$ and includes machines like the Cray which do not have a guard digit. This does not greatly complicate the error analysis, but it is possible that the computed rotation angle may be less accurate. This may adversely affect convergence, but as we will see it does not affect the one-step error analysis.

Numerically subscripted ε's will denote independent quantities bounded in magnitude by ε. As usual (e.g. [13]), we will make approximations like $(1+i\varepsilon_1)(1+j\varepsilon_2) = 1 + (i+j)\varepsilon_3$ and $(1+i\varepsilon_1)/(1+j\varepsilon_2) = 1 + (i+j)\varepsilon_3$.

The next theorem and its corollary justify our accuracy claims for eigenvalues computed by two-sided $J-$orthogonal Jacobi method .

Theorem 3.2.1 *Let H_m be the sequence of matrices generated by Algorithm 3.1.1 in floating–point arithmetic with precision ε; that is, H_{m+1} is obtained from H_m by applying a single J–orthogonal Jacobi rotation. Then the following diagram commutes.*

$$\begin{tikzcd} H_m \arrow[r, "\text{floating}"] \arrow[d] & H_{m+1} \\ H_m + \delta H_m \arrow[ur, "\text{exact}"'] & \end{tikzcd}$$

The top arrow indicates that H_{m+1} is obtained from H_m by applying one J–orthogonal Jacobi rotation in floating–point arithmetic. The diagonal arrow indicates that H_{m+1} is obtained from $H_m + \delta H_m$ by applying one J–orthogonal Jacobi rotation in exact arithmetic; thus H_{m+1} amd $H_m + \delta H_m$ are exactly similar. δH_m is bounded as follows. Let $\kappa = \kappa(A_m)$, and write $\delta H_m = D_m \delta A_m D_m$. Then, with the relative error of order ε,[1]

$$\|\delta A_m\|_2 \leq C_m \, \varepsilon \ , \tag{3.2.2}$$

where

$$C_m = \begin{cases} 60 + 58\sqrt{n-2} & \textit{in trigon. case} \ , \\ 35.5 + (\sqrt{\kappa}+3)(30.93 + 8.24\sqrt{n-2}) & \textit{in hyperb. case} \ , \quad |\zeta| \leq \dfrac{3}{2\sqrt{2}} \ , \\ 222.42 + 46.77\sqrt{n-2} & \textit{in hyperb. case} \ , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \ , \\ & \qquad\qquad\qquad\qquad b \geq \dfrac{1}{2}a \ , \\ 225.5 + 62.45\sqrt{n-2} & \textit{in hyperb. case} \ , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \ , \\ & \qquad\qquad\qquad\qquad b < \dfrac{1}{2}a \ . \end{cases}$$

In other words, one step of Jacobi satisfies the assumptions needed for the perturbation bounds of Sect. 2.2.

The bound (3.2.2) seems to be highly discontinuous at $|\zeta| = 3/(2\sqrt{2})$. This discontinuity can be removed as decribed in Rem. 3.2.4, or by using the modified method of Subsect. 3.2.1.

[1] This formulation is explained after the relation (1.12).

PROOF. The proof of the commuting diagram is a tedious computation. We shall prove the diagram separately for the trigonometric and for the hyperbolic case. We assume that multiplications with the parameter *hyp* in Alg. 3.1.1 have no errors. Write the 2 by 2 submatrix of the current matrix H_m as

$$\widehat{H}_m = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv \begin{bmatrix} d_i^2 & zd_id_j \\ zd_id_j & d_j^2 \end{bmatrix} \tag{3.2.3}$$

In both cases we can assume without loss of generality that $a \geq b$. By positive definiteness we have

$$0 < |z| \leq \bar{z} \equiv (\kappa - 1)/(\kappa + 1) < 1 \ . \tag{3.2.4}$$

Let a' and b' be the new values of H_{ii} and H_{jj} computed by the algorithm, respectively.

Trigonometric case. This case was analysed by Demmel and Veselić [13]. Our proof is essentially the same as theirs, and we repeat it for the sake of completeness. Small differences in the proof lead to a somewhat better bound for $\|\delta A_m\|_2$.

Systematic application of the formulae (3.2.1) shows that

$$\begin{aligned} \zeta &= fl((b-a)/(2*c)) \\ &= (1+\varepsilon_4)(((1+\varepsilon_1)b - (1+\varepsilon_2)a)/((1+\varepsilon_3)2c)) \\ &= \frac{(1+\varepsilon_4)(1+\varepsilon_2)}{1+\varepsilon_3}\left(\frac{\tilde{b}-a}{2c}\right) \end{aligned}$$

where

$$\tilde{b} \equiv \frac{1+\varepsilon_1}{1+\varepsilon_2} b \equiv (1+\varepsilon_b)b \ , \qquad |\varepsilon_b| \leq 2\varepsilon \ .$$

Thus

$$\zeta = (1+\varepsilon_\zeta)\frac{\tilde{b}-a}{2c} \ , \qquad |\varepsilon_\zeta| \leq 3\varepsilon \ .$$

Let $\tilde{t}$, $\widetilde{cs}$, $\widetilde{sn}$ and $-\widetilde{sn}$ denote the true values of t, cs, sn and $sn1 = -sn$ (i.e. without rounding error) as a function of a, $\tilde{b}$ and c. Using (3.2.1) again one can show that

$$t = (1+\varepsilon_t)\tilde{t}, \qquad cs = (1+\varepsilon_{cs})\widetilde{cs}, \qquad sn = (1+\varepsilon_{sn})\widetilde{sn},$$

where[2]

$$|\varepsilon_t| \leq 7\varepsilon, \qquad |\varepsilon_{cs}| \leq 10\varepsilon, \qquad |\varepsilon_{sn}| \leq 17\varepsilon.$$

$\widetilde{cs}$ and $\widetilde{sn}$ define the exact trigonometric Jacobi rotation

$$J_m \equiv \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix}$$

[2] Calculating sn as $sn = t/h$ instead of $sn = t \cdot cs$ [13] saves one ε in bounding ε_{sn}. This was noticed by Drmač [14].

which transforms $H_m + \delta H_m$ to H_{m+1} in the diagram in the statement of the theorem:

$$J_m^T (H_m + \delta H_m) J_m = H_{m+1} \; .$$

Now we begin constructing δH_m. δH_m will be nonzero only in the rows and columns i and j. We first compute its entries outside the 2 by 2 pivot submatrix. Let H'_{ik} and H'_{jk} denote the updated quantities computed by the algorithm. Then

$$\begin{array}{rcl} H'_{ik} & = & fl(cs * H_{ik} - sn * H_{jk}) \\ & = & (1+\varepsilon_4)(1+\varepsilon_5) cs H_{ik} - (1+\varepsilon_6)(1+\varepsilon_7) sn H_{jk} \\ & = & (1+\varepsilon_4)(1+\varepsilon_5)(1+\varepsilon_{cs}) \widetilde{cs} H_{ik} - (1+\varepsilon_6)(1+\varepsilon_7)(1+\varepsilon_{sn}) \widetilde{sn} H_{jk} \\ & \equiv & \widetilde{cs} H_{ik} - \widetilde{sn} H_{jk} + \epsilon(H'_{ik}), \end{array}$$

where

$$\epsilon(H'_{ik}) = \varepsilon'_1 \widetilde{cs} H_{ik} - \varepsilon'_2 \widetilde{sn} H_{jk}, \qquad |\varepsilon'_1| \le 12\varepsilon, \;\; |\varepsilon'_2| \le 19\varepsilon.$$

Similarly,

$$\begin{array}{rcl} H'_{jk} & = & fl(sn * H_{ik} + cs * H_{jk}) \\ & = & \widetilde{sn} H_{ik} + \widetilde{cs} H_{jk} + \epsilon(H'_{jk}), \end{array}$$

where

$$\epsilon(H'_{jk}) = \varepsilon'_3 \widetilde{cs} H_{jk} + \varepsilon'_4 \widetilde{sn} H_{ik}, \qquad |\varepsilon'_3| \le 12\varepsilon, \;\; |\varepsilon'_4| \le 19\varepsilon.$$

Thus

$$\begin{array}{rcl} \begin{bmatrix} H'_{ik} \\ H'_{jk} \end{bmatrix} & = & J_m^T \begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + \begin{bmatrix} \epsilon(H_{ik}) \\ \epsilon(H_{jk}) \end{bmatrix} \\ & = & J_m^T \left(\begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + J_m \begin{bmatrix} \epsilon(H_{ik}) \\ \epsilon(H_{jk}) \end{bmatrix} \right) \\ & \equiv & J_m^T \left(\begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + \begin{bmatrix} \delta H_{ik} \\ \delta H_{jk} \end{bmatrix} \right) \;, \end{array}$$

where

$$\begin{array}{rcl} \delta H_{ik} & = & \varepsilon'_1 \widetilde{cs}^2 H_{ik} - \varepsilon'_2 \widetilde{cs}\widetilde{sn} H_{jk} + \varepsilon'_3 \widetilde{cs}\widetilde{sn} H_{jk} + \varepsilon'_4 \widetilde{sn}^2 H_{ik} \\ \delta H_{jk} & = & -\varepsilon'_1 \widetilde{cs}\widetilde{sn} H_{ik} + \varepsilon'_2 \widetilde{sn}^2 H_{jk} + \varepsilon'_3 \widetilde{cs}^2 H_{jk} + \varepsilon'_4 \widetilde{cs}\widetilde{sn} H_{ik} \; . \end{array}$$

Using

$$|H_{ij}| \le d_i d_j, \qquad \widetilde{cs} = \frac{1}{\sqrt{1+\tilde{t}^2}}, \qquad \widetilde{sn} = \frac{\tilde{t}}{\sqrt{1+\tilde{t}^2}},$$

we have

$$|\delta A_{ik}| \le \frac{1}{1+\tilde{t}^2} \left(12 + 31|\tilde{t}| \frac{d_j}{d_i} + 19\tilde{t}^2 \right) \varepsilon \; , \tag{3.2.5}$$

which is an increasing function for $|\tilde{t}| \in [0,1]$.

Set $x \equiv d_j/d_i$. Note that $x \leq 1$. In estimating $|\delta A_{jk}|$ we consider two cases: $x < \bar{x} \equiv .48$, and $x \geq \bar{x}$. If $x < \bar{x}$, then, with the relative error of $O(\varepsilon)$, we have

$$|\tilde{t}| = \frac{1}{\dfrac{1-x^2}{2|z|x} + \left(1 + \left(\dfrac{1-x^2}{2|z|x}\right)^2\right)^{1/2}} \leq \frac{x}{1-\bar{x}^2} \ .$$

Since we want to bound $|\delta A_{jk}|$ with a bound of order ε, we neglect the relative error of $O(\varepsilon)$ in the above inequality. Therefore,

$$|\delta A_{jk}| \leq \frac{1}{1+\tilde{t}^2} \left(12 + 31\frac{1}{1-\bar{x}^2} + 19\tilde{t}^2\right) \varepsilon \ , \tag{3.2.6}$$

which is a decreasing function of $\tilde{t}^2$. Substituting 1 for $\tilde{t}$ and $\bar{x}$ for d_j/d_i in (3.2.5), and 0 for $\tilde{t}$ in (3.2.6), we obtain

$$\sqrt{\delta A_{ik}^2 + \delta A_{jk}^2} \leq 57.3\varepsilon \ . \tag{3.2.7}$$

If $x \geq \bar{x}$, then

$$|\delta A_{jk}| \leq \frac{1}{1+\tilde{t}^2} \left(12 + 31|\tilde{t}|\frac{1}{\bar{x}} + 19\tilde{t}^2\right) \varepsilon \ , \tag{3.2.8}$$

which is an increasing function of $|\tilde{t}| \in [0,1]$. Substituting 1 for $\tilde{t}$ and d_j/d_i in (3.2.5) and (3.2.8), we obtain

$$\sqrt{\delta A_{ik}^2 + \delta A_{jk}^2} \leq 57.4\varepsilon \ . \tag{3.2.9}$$

Note that our choice of $\bar{x}$ makes bounds in relations (3.2.7) and (3.2.9) almost equal.

Now we construct the 2 by 2 submatrix $\delta\widehat{H}_m$ of δH_m at the intersection of the rows and columns i and j. We will construct it of three components

$$\delta\widehat{H}_m = \Delta_1 + \Delta_2 + \Delta_3 \ .$$

Applying the relations (3.2.1), we obtain

$$\begin{aligned} b' &= fl(b + ct) = \frac{1+\varepsilon_2}{1+\varepsilon_1}(1+\varepsilon_8)\tilde{b} + (1+\varepsilon_9)(1+\varepsilon_{10})(1+\varepsilon_t)\, c\tilde{t} \\ &= (1+\varepsilon_9)(1+\varepsilon_{10})(1+\varepsilon_t)\left(\frac{(1+\varepsilon_2)(1+\varepsilon_8)}{(1+\varepsilon_1)(1+\varepsilon_9)(1+\varepsilon_{10})(1+\varepsilon_t)}\tilde{b} + c\,\tilde{t}\right) \\ &\equiv (1+\varepsilon_{b'})(\tilde{b} + c\,\tilde{t} + \varepsilon_b'\tilde{b}) \ , \end{aligned}$$

where $|\varepsilon_{b'}| \leq 9\varepsilon$ and $|\varepsilon_b'| \leq 12\varepsilon$. Similarly,

$$\begin{aligned} a' &= fl(a - ct) = (1+\varepsilon_{11})a - (1+\varepsilon_{12})(1+\varepsilon_{13})(1+\varepsilon_t)\, c\tilde{t} \\ &= (1+\varepsilon_{a'})(a - c\,\tilde{t}) \ , \end{aligned}$$

where $|\varepsilon_{a'}| \leq 9\varepsilon$. Here we used the fact that $c\tilde{t} < 0$.

Now let

$$\Delta_1 = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon_b b \end{bmatrix} + J_m \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon_b' \tilde{b} \end{bmatrix} J_m^T .$$

From earlier discussion we see that

$$J_m^T \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 \right) J_m = \begin{bmatrix} a - c\tilde{t} & 0 \\ 0 & b + c\tilde{t} + \varepsilon_b' \tilde{b} \end{bmatrix} .$$

Next let

$$\Delta_2 = \varepsilon_{a'} \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 \right) .$$

Thus

$$\begin{aligned} J_m^T \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 + \Delta_2 \right) J_m &= (1 + \varepsilon_{a'}) \begin{bmatrix} a - c\tilde{t} & 0 \\ 0 & b + c\tilde{t} + \varepsilon_b' \tilde{b} \end{bmatrix} \\ &= \begin{bmatrix} a' & 0 \\ 0 & b' \dfrac{1 + \varepsilon_{a'}}{1 + \varepsilon_{b'}} \end{bmatrix} . \end{aligned}$$

Now let

$$\Delta_3 = J_m \begin{bmatrix} 0 & 0 \\ 0 & b' \left(1 - \dfrac{1 + \varepsilon_{a'}}{1 + \varepsilon_{b'}} \right) \end{bmatrix} J_m^T \equiv \begin{bmatrix} \widetilde{sn}^2 \varepsilon_{b''} b' & \widetilde{cs}\widetilde{sn} \varepsilon_{b''} b' \\ \widetilde{cs}\widetilde{sn} \varepsilon_{b''} b' & \widetilde{cs}^2 \varepsilon_{b''} b' \end{bmatrix} ,$$

where $|\varepsilon_{b''}| \leq |\varepsilon_{a'}| + |\varepsilon_{b'}| \leq 18\varepsilon$. Then

$$J_m^T \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 + \Delta_2 + \Delta_3 \right) J_m = \begin{bmatrix} a' & 0 \\ 0 & b' \end{bmatrix}$$

as desired. This completes the construction of $\delta \widehat{H}_m$. Since $\tilde{b} = b(1 + \varepsilon_b)$ and $b' < b$,

$$\|\delta \widehat{A}_m\|_2 \leq |\varepsilon_b| + |\varepsilon_b'| + 2 \cdot |\varepsilon_{a'}| + |\varepsilon_{b''}| \leq 60\varepsilon$$

holds with the relative error of $O(\varepsilon)$. From (3.2.7), (3.2.9), and the above relation, it finally follows

$$\|\delta A_m\|_2 \leq (60 + 58\sqrt{n-2})\,\varepsilon . \tag{3.2.10}$$

This bound improves the bound $\|\delta A_m\|_2 \leq (257\sqrt{n-2} + 104)\varepsilon$ from [13].

Hyperbolic case. To avoid the confusion with the trigonometric case, we denote the quantities cs, sn and $sn1 = sn$ computed by Alg. 3.1.1 with ch and sh, respectively. We first compute $\tilde{t}$, $\tilde{h}$, $\widetilde{ch}$ and $\widetilde{sh}$ as the exact values of the parameters computed without rounding errors from a, b and some $\tilde{c} \equiv (1+\varepsilon_c)c$. Generally, the following bounds hold:

$$\begin{aligned} |\zeta| &\geq \frac{\kappa+1}{\kappa-1} > 1 \ , \\ |t| &\leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} < 1 \ , \\ |sh| < ch &\leq \frac{1}{2}\left(\sqrt[4]{\kappa} + \frac{1}{\sqrt[4]{\kappa}}\right) \ . \end{aligned} \tag{3.2.11}$$

Let $\zeta_0 \equiv (a+b)/(-2c)$ be the exact value of $(a+b)/(-2c)$. Systematic application of (3.2.1) gives

$$\zeta_1 = fl\left(\frac{a+b}{-2c}\right) = (1+\varepsilon_{\zeta_1})\zeta_0 \ ,$$

where $|\varepsilon_{\zeta_1}| \leq 3\varepsilon$, and

$$\begin{aligned} t_1 &= fl\left(\frac{\text{sign }(\zeta_1)}{|\zeta_1| + \sqrt{\zeta_1^2 - 1}}\right) \\ &= \frac{(1+\varepsilon_1)\text{sign }(\zeta_1)}{(1+\varepsilon_2)(|\zeta_1| + (1+\varepsilon_3)\sqrt{\zeta_1^2(1+\varepsilon_4)(1+\varepsilon_5) - (1+\varepsilon_6)}} \\ &= \frac{1+\varepsilon_1}{1+\varepsilon_2} \cdot \frac{\text{sign }(\zeta_1)}{|\zeta_1| + (1+\varepsilon_3)\sqrt{1+\varepsilon_6}\sqrt{\zeta_1^2 \frac{(1+\varepsilon_4)(1+\varepsilon_5)}{1+\varepsilon_6} - 1}} \ . \end{aligned}$$

Now let

$$\zeta_2^2 \equiv \zeta_1^2 \frac{(1+\varepsilon_4)(1+\varepsilon_5)}{1+\varepsilon_6} \ .$$

Then $\zeta_2 \equiv (1+\varepsilon_{\zeta_2})\zeta_1$, where $|\varepsilon_{\zeta_2}| \leq 1.5\varepsilon$. This implies that

$$\begin{aligned} t_1 &= \frac{1+\varepsilon_1}{1+\varepsilon_2} \cdot \frac{\text{sign }(\zeta_1)}{|\zeta_2 \frac{1}{1+\varepsilon_{\zeta_2}}| + (1+\varepsilon_3)\sqrt{1+\varepsilon_6}\sqrt{\zeta_2^2 - 1}} \\ &\equiv (1+\varepsilon_{t_1}) \frac{\text{sign }(\zeta_2)}{|\zeta_2| + \sqrt{\zeta_2^2 - 1}} \ , \end{aligned}$$

where $|\varepsilon_{t_1}| \leq 3.5\varepsilon$. Furthermore,

$$\begin{aligned} h &= fl(\sqrt{1-t_1^2}) = (1+\varepsilon_7)\sqrt{(1+\varepsilon_8) - (1+\varepsilon_9)(1+\varepsilon_{10})t_1^2} \\ &= (1+\varepsilon_7)\sqrt{1+\varepsilon_8}\sqrt{1 - \frac{(1+\varepsilon_9)(1+\varepsilon_{10})}{1+\varepsilon_8}t_1^2} \\ &\equiv (1+\varepsilon_h)\sqrt{1-t_2^2} \ , \end{aligned} \tag{3.2.12}$$

where

$$|\varepsilon_h| \le 1.5\varepsilon \ , \qquad t_2^2 \equiv \frac{(1+\varepsilon_9)(1+\varepsilon_{10})}{1+\varepsilon_8} t_1^2 \ ,$$

i.e.

$$t_2 = (1+\varepsilon_{t_2})t_1 \ , \qquad |\varepsilon_{t_2}| \le 1.5\varepsilon \ . \tag{3.2.13}$$

Therefore,

$$\frac{\text{sign }(\zeta_2)}{|\zeta_2| + \sqrt{\zeta_2^2 - 1}} = \frac{1}{1+\varepsilon_{t_1}} t_1 = \frac{1}{(1+\varepsilon_{t_1})(1+\varepsilon_{t_2})} t_2 \equiv (1+\varepsilon'_{t_2})t_2 \ ,$$

where $|\varepsilon'_{t_2}| \le |\varepsilon_{t_1}| + |\varepsilon_{t_2}| \le 5\varepsilon$. In exact arithmetic

$$t = \frac{\text{sign }(\zeta)}{|\zeta| + \sqrt{\zeta^2 - 1}}$$

implies

$$\zeta = \frac{1}{2}\left(t + \frac{1}{t}\right) \ .$$

Therefore, in exact arithmetic for ζ_2 we have

$$\zeta_2 = \frac{1}{2}\left((1+\varepsilon'_{t_2})t_2 + \frac{1}{(1+\varepsilon'_{t_2})t_2}\right) \ ,$$

which, in turn, implies

$$\frac{1}{2}\left(t_2 + \frac{1}{t_2}\right) = (1+\varepsilon'_{\zeta_2})\zeta_2 = (1+\varepsilon'_{\zeta_2})(1+\varepsilon_{\zeta_2})\zeta_0 \equiv (1+\varepsilon''_{\zeta_2})\zeta_0 \ , \tag{3.2.14}$$

where

$$|\varepsilon'_{\zeta_2}| \le |\varepsilon'_{t_2}| \le 5\varepsilon \ , \qquad |\varepsilon''_{\zeta_2}| \le |\varepsilon'_{\zeta_2}| + |\varepsilon_{\zeta_2}| \le 9.5\varepsilon \ .$$

Therefore, we can choose ε_c,

$$|\varepsilon_c| \le |\varepsilon''_{\zeta_2}| \le 9.5\varepsilon \ ,$$

such that for $\tilde{c} = (1+\varepsilon_c)c$ in exact arithmetic[3]

$$\frac{a+b}{-2\tilde{c}} = (1+\varepsilon''_{\zeta_2})\zeta_0 \ . \tag{3.2.15}$$

From (3.2.14) and (3.2.15) it follows that $\tilde{t} = t_2$ is the exact value of the parameter t computed without rounding errors from a, b and $\tilde{c}$. Set

$$\tilde{h} = \sqrt{1-\tilde{t}^2}, \qquad \widetilde{ch} = \frac{1}{\tilde{h}}, \qquad \widetilde{sh} = \frac{\tilde{t}}{\tilde{h}}.$$

[3] In the trigonometric case we had to perturb b. Here we can perturb either a or c, and we perturb c since it is absolutely smaller.

For the computed quantities we have

$$\begin{aligned}
t &= t_1 = (1+\varepsilon_t)\widetilde{t}\ , \qquad |\varepsilon_t| \le |\varepsilon_{t_2}| \le 1.5\varepsilon\ , \\
h &= (1+\varepsilon_h)\widetilde{h}\ , \qquad |\varepsilon_h| \le 1.5\varepsilon\ , \\
ch &= fl(1/h) = (1+\varepsilon_{ch})\widetilde{ch}\ , \qquad |\varepsilon_{ch}| \le \varepsilon + |\varepsilon_h| \le 2.5\varepsilon\ , \\
sh &= fl(t/h) = (1+\varepsilon_{sh})\widetilde{sh}\ , \qquad |\varepsilon_{sh}| \le \varepsilon + |\varepsilon_t| + |\varepsilon_h| \le 4\varepsilon\ .
\end{aligned}$$

Here the first line follows from $\widetilde{t} = t_2$ and (3.2.12), the second line follows from (3.2.13), and the last two lines follow from the first two lines and the formulae (3.2.1). $\widetilde{ch}$ and $\widetilde{sh}$ define the exact hyperbolic rotation

$$J_m \equiv \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix}$$

which transforms $H_m + \delta H_m$ to H_{m+1} in the diagram in the statement of the theorem:

$$J_m(H_m + \delta H_m)J_m = H_{m+1}\ .$$

Now we begin constructing δH_m. δH_m will be nonzero only in the rows and columns i and j. First we compute its entries outside the 2 by 2 (i,j) submatrix. Let H'_{ik} and H'_{jk} denote the updated quantities computed by the algorithm. Then, similarly to the trigonometric case, we have

$$H'_{ik} = fl(ch * H_{ik} + sh * H_{jk}) = \widetilde{ch}H_{ik} + \widetilde{sh}H_{jk} + \epsilon(H'_{ik}),$$

where

$$\epsilon(H'_{ik}) = \varepsilon'_1\widetilde{ch}H_{ik} + \varepsilon'_2\widetilde{sh}H_{jk}, \qquad |\varepsilon'_1| \le 4.5\varepsilon,\ |\varepsilon'_2| \le 6\varepsilon\ ,$$

and

$$H'_{jk} = fl(sh * H_{ik} + ch * H_{jk}) = \widetilde{sh}H_{ik} + \widetilde{ch}H_{jk} + \epsilon(H'_{jk}),$$

where

$$\epsilon(H'_{jk}) = \varepsilon'_3\widetilde{ch}H_{jk} + \varepsilon'_4\widetilde{sh}H_{ik}, \qquad |\varepsilon'_3| \le 4.5\varepsilon,\ |\varepsilon'_4| \le 6\varepsilon.$$

Thus

$$\begin{aligned}
\begin{bmatrix} H'_{ik} \\ H'_{jk} \end{bmatrix} &= J_m^T \begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + \begin{bmatrix} \epsilon(H_{ik}) \\ \epsilon(H_{jk}) \end{bmatrix} \\
&= J_m^T \left(\begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + J_m^{-1} \begin{bmatrix} \epsilon(H_{ik}) \\ \epsilon(H_{jk}) \end{bmatrix} \right) \\
&\equiv J_m \left(\begin{bmatrix} H_{ik} \\ H_{jk} \end{bmatrix} + \begin{bmatrix} \delta H_{ik} \\ \delta H_{jk} \end{bmatrix} \right)\ ,
\end{aligned}$$

where

$$\begin{aligned}
\delta H_{ik} &= \varepsilon'_1\widetilde{ch}^2 H_{ik} + \varepsilon'_2\widetilde{ch}\widetilde{sh}H_{jk} - \varepsilon'_3\widetilde{ch}\widetilde{sh}H_{jk} - \varepsilon'_4\widetilde{sh}^2 H_{ik}\ , \\
\delta H_{jk} &= -\varepsilon'_1\widetilde{ch}\widetilde{sh}H_{ik} - \varepsilon'_2\widetilde{sh}^2 H_{jk} + \varepsilon'_3\widetilde{ch}^2 H_{jk} + \varepsilon'_4\widetilde{ch}\widetilde{sh}H_{ik}\ .
\end{aligned} \qquad (3.2.16)$$

Contrary to the trigonometric case, we analyse two cases. The first case is when $|\zeta|$ is near the bound (3.2.11), and the second case is when $|\zeta|$ is bounded away from (3.2.11). Set, as in the trigonometric case, $x \equiv d_j/d_i$. Set[4]

$$\alpha = \frac{3}{2\sqrt{2}} , \qquad \beta \equiv \frac{1}{\alpha + \sqrt{\alpha^2 - 1}} = \frac{1}{\sqrt{2}} . \tag{3.2.17}$$

Case I. $|\zeta| \le \alpha$.

From our assumption, the definition of ζ, and $|c| \le \sqrt{ab}$ it follows

$$a + b \le \alpha \cdot 2\sqrt{ab},$$

i.e.

$$\frac{1}{x} + x \le 2\alpha .$$

This implies

$$\begin{aligned} x &\ge \frac{1}{\alpha + \sqrt{\alpha^2 - 1}} \equiv \beta = \frac{1}{\sqrt{2}} , \\ d_j \le d_i \le \frac{1}{\beta} d_j , & \qquad b \le a \le \frac{1}{\beta^2} b , \end{aligned} \tag{3.2.18}$$

i.e. when $|\zeta|$ is near its lower bound, then a and b do not differ much. We now show that

$$\widetilde{ch} \le \frac{1}{2}\left(\sqrt[4]{\kappa} + \frac{1}{\sqrt[4]{\kappa}}\right) \tag{3.2.19}$$

holds with the relative error of $O(\varepsilon)$. Indeed, $c = z d_i d_j$ implies $\tilde{c} = \tilde{z} d_i d_j = z(1 + \varepsilon_c) d_i d_j$. Set

$$\kappa_1 \equiv (1 + z(1 + \varepsilon_c))/(1 - z(1 + \varepsilon_c)) .$$

Then

$$\widetilde{ch} \le (\sqrt[4]{\kappa_1} + \frac{1}{\sqrt[4]{\kappa_1}})/2 .$$

A simple calculation shows that

$$1 - z(1 + \varepsilon_c) = (1 + \varepsilon')(1 - z) , \qquad |\varepsilon'| \le |\varepsilon_c|(\kappa + 1)/2 .$$

Therefore,

$$\kappa_1 \equiv \frac{1 + z(1 + \varepsilon_c)}{1 - z(1 + \varepsilon_c)} = \frac{1 + z}{1 - z}(1 + \varepsilon_c)(1 + \varepsilon') \le \kappa(1 + |\varepsilon_c| + |\varepsilon'|) ,$$

and (3.2.19) holds. We neglect the relative error of $O(\varepsilon)$ since it adds only the relative error of $O(\varepsilon)$ in the final estimate. Therefore,

$$\widetilde{sh}^2 \le |\widetilde{sh}|\widetilde{ch} \le \widetilde{ch}^2 \le \frac{1}{4}(\sqrt{\kappa} + 3) . \tag{3.2.20}$$

[4]Note that we can choose some other α, as well. This choice is explained in Rem. 3.2.4 below.

Using the fact that $|H_{ij}| \leq d_i d_j$, and inserting (3.2.18) and (3.2.20) into (3.2.16), we obtain

$$\begin{aligned} |\delta H_{ik}| &\leq \frac{1}{4}(\sqrt{\kappa}+3)(|\varepsilon_1'|+|\varepsilon_2'|+|\varepsilon_3'|+|\varepsilon_4'|)d_i d_k \\ &\leq 5.25(\sqrt{\kappa}+3)d_i d_k \varepsilon \\ |\delta H_{jk}| &\leq \frac{1}{4}(\sqrt{\kappa}+3)(|\varepsilon_2'|+|\varepsilon_3'|+\frac{1}{\beta}(|\varepsilon_1'|+|\varepsilon_4'|))d_j d_k \\ &\leq 6.34(\sqrt{\kappa}+3)d_j d_k \varepsilon \ . \end{aligned} \tag{3.2.21}$$

Now we construct the 2 by 2 submatrix $\delta\widehat{H}_m$ of δH_m at the intersection of the rows and columns i and j. We will construct it of three components, $\delta\widehat{H}_m = \Delta_1 + \Delta_2 + \Delta_3$. The analysis is somewhat different from the analysis in the trigonometric case because $a' < a$, $b' < b$, so that, due to subtraction, a' and b' can both have large relative errors. We have

$$\begin{aligned} a' &= fl(a+ct) = (1+\varepsilon_{12})a + (1+\varepsilon_{13})(1+\varepsilon_{14})(1+\varepsilon_c)(1+\varepsilon_t)\,\tilde{c}\tilde{t} \\ &= (1+\varepsilon_{13})(1+\varepsilon_{14})(1+\varepsilon_c)(1+\varepsilon_t)\left(\frac{1+\varepsilon_{12}}{(1+\varepsilon_{13})(1+\varepsilon_{14})(1+\varepsilon_c)(1+\varepsilon_t)}\,a + \tilde{c}\tilde{t}\right) \\ &= (1+\varepsilon_{a'})(a+\tilde{c}\tilde{t}+\varepsilon_a a) \ , \end{aligned}$$

where

$$\begin{aligned} |\varepsilon_{a'}| &\leq 2\varepsilon + |\varepsilon_c| + |\varepsilon_t| \leq 13\varepsilon \ , \\ |\varepsilon_a| &\leq 3\varepsilon + |\varepsilon_c| + |\varepsilon_t| \leq 14\varepsilon \ . \end{aligned}$$

Similarly,

$$b' = fl(b+ct) = (1+\varepsilon_{b'})(b+\tilde{c}\tilde{t}+\varepsilon_b b) \ ,$$

where

$$|\varepsilon_{b'}| \leq 13\varepsilon \ , \qquad |\varepsilon_b| \leq 14\varepsilon \ .$$

Let

$$\begin{aligned} \Delta_1 &= \begin{bmatrix} 0 & \varepsilon_c c \\ \varepsilon_c c & 0 \end{bmatrix} + J_m^{-1}\begin{bmatrix} \varepsilon_a a & 0 \\ 0 & \varepsilon_b b \end{bmatrix} J_m^{-1} \\ &= \begin{bmatrix} 0 & \varepsilon_c c \\ \varepsilon_c c & 0 \end{bmatrix} + \begin{bmatrix} \widetilde{ch}^2\varepsilon_a a + \widetilde{sh}^2\varepsilon_b b & -\widetilde{ch}\,\widetilde{sh}(\varepsilon_a a+\varepsilon_b b) \\ -\widetilde{ch}\,\widetilde{sh}(\varepsilon_a a+\varepsilon_b b) & \widetilde{sh}^2\varepsilon_a a + \widetilde{ch}^2\varepsilon_b b \end{bmatrix} \ , \end{aligned}$$

and

$$\Delta_2 = \varepsilon_{a'}\left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1\right) \ .$$

From earlier discussion we see that

$$J_m\left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 + \Delta_2\right)J_m = \begin{bmatrix} a' & 0 \\ 0 & b'\dfrac{1+\varepsilon_{a'}}{1+\varepsilon_{b'}} \end{bmatrix} \ .$$

Now let

$$\Delta_3 = J_m^{-1} \left[\begin{array}{cc} 0 & 0 \\ 0 & b'\left(1 - \dfrac{1+\varepsilon_{a'}}{1+\varepsilon_{b'}}\right) \end{array} \right] J_m^{-1} = \left[\begin{array}{cc} \widetilde{sh}^2 \varepsilon_{b''} b' & -\widetilde{ch}\,\widetilde{sh}\varepsilon_{b''} b' \\ -\widetilde{ch}\,\widetilde{sh}\varepsilon_{b''} b' & \widetilde{ch}^2 \varepsilon_{b''} b' \end{array} \right] ,$$

where $|\varepsilon_{b''}| \le |\varepsilon_{a'}| + |\varepsilon_{b'}| \le 26\varepsilon$. Then

$$J_m \left(\left[\begin{array}{cc} a & c \\ c & b \end{array} \right] + \Delta_1 + \Delta_2 + \Delta_3 \right) J_m = \left[\begin{array}{cc} a' & 0 \\ 0 & b' \end{array} \right]$$

to the first order of ε, as desired. This completes the construction of $\delta\widehat{H}_m$ and we have

$$\begin{aligned} \|\delta\widehat{A}_m\|_2 &\le |\varepsilon_c| + \frac{1}{4}(\sqrt{\kappa}+3)\max\{|\varepsilon_a|,|\varepsilon_b|\}(2\alpha + 1 + \frac{1}{\beta^2}) + 2|\varepsilon_{a'}| + \frac{1}{2}(\sqrt{\kappa}+3)|\varepsilon_{b''}| \\ &\le (35.5 + (\sqrt{\kappa}+3)\,30.93)\,\varepsilon\ . \end{aligned}$$

Here we used (3.2.18), (3.2.20), and $b' < b$. Combining (3.2.21) with the above relation, we finally obtain

$$\|\delta A_m\|_2 \le (35.5 + (\sqrt{\kappa}+3)(30.93 + 8.24\sqrt{n-2}))\,\varepsilon\ . \tag{3.2.22}$$

Case II. $|\zeta| > \alpha$.

Our assumption implies

$$\begin{aligned} |t| &\le \beta = \frac{1}{\sqrt{2}}\ , \\ ch &\le \frac{1}{\sqrt{1-\beta^2}} = \sqrt{2}\ , \\ |sh| &\le \frac{\beta}{\sqrt{1-\beta^2}} = 1\ . \end{aligned} \tag{3.2.23}$$

These bounds hold with the relative error of $O(\varepsilon)$ for $\tilde{t}$, $\widetilde{sh}$ and $\widetilde{ch}$, as well. We split this case into two subcases.

Subcase IIa. $x \equiv d_j/d_i \ge \beta$.

The analysis is identical to the analysis in the first case; only the upper bounds for $\widetilde{sh}^2$, $|\widetilde{sh}|\widetilde{ch}$ and $\widetilde{ch}^2$ are now obtained from (3.2.23) and not from (3.2.20). Therefore,

$$\begin{aligned} |\delta H_{ik}| &\le (2|\varepsilon_1'| + \sqrt{2}|\varepsilon_2'| + \sqrt{2}|\varepsilon_3'| + |\varepsilon_4'|)d_i d_k \le 29.85 d_i d_k\,\varepsilon\ , \\ |\delta H_{jk}| &\le (\sqrt{2}\sqrt{2}|\varepsilon_1'| + |\varepsilon_2'| + 2|\varepsilon_3'| + \sqrt{2}\sqrt{2}|\varepsilon_4'|)d_j d_k \le 36 d_j d_k\,\varepsilon\ , \\ \|\delta\widehat{A}_m\|_2 &\le |\varepsilon_c| + \max\{|\varepsilon_a|,|\varepsilon_b|\} \left\| \left[\begin{array}{cc} 3 & 3\sqrt{2} \\ 3\sqrt{2} & 4 \end{array} \right] \right\|_2 + 2|\varepsilon_{a'}| + |\varepsilon_{b''}| \left\| \left[\begin{array}{cc} 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{array} \right] \right\|_2 \\ &\le 222.42\,\varepsilon\ , \end{aligned}$$

and, altogether,

$$\|\delta A_m\|_2 \leq (222.42 + 46.77\sqrt{n-2})\,\varepsilon \ . \tag{3.2.24}$$

Subcase IIb. $x \equiv d_j/d_i < \beta$.

The above assumption implies $|c| < \beta a$. From

$$\widetilde{\zeta} = -(1+x^2)/(2\widetilde{z}x) \ , \qquad \widetilde{z} = z(1+\varepsilon_c) \ ,$$

it follows

$$|\widetilde{t}| \leq 2|\widetilde{z}|x/(1+x^2) \leq 2x \ .$$

Here we ignored the relative error of $O(\varepsilon)$ in z and and used the fact that $|\widetilde{z}| < 1$. Therefore,

$$\widetilde{ch}|\widetilde{sh}| = \widetilde{ch}^2|\widetilde{t}| \leq 2\widetilde{ch}^2\frac{d_j}{d_i} \ . \tag{3.2.25}$$

From (3.2.16), (3.2.23), (3.2.25), and our assumption, it follows

$$\begin{aligned} |\delta H_{ik}| &\leq (2|\varepsilon_1'| + \sqrt{2}|\varepsilon_2'|\frac{1}{\sqrt{2}} + \sqrt{2}|\varepsilon_3'|\frac{1}{\sqrt{2}} + |\varepsilon_4'|)d_i d_k \leq 25.5 d_i d_k\,\varepsilon \ , \\ |\delta H_{jk}| &\leq (\sqrt{2}\sqrt{2}|\varepsilon_1'| + |\varepsilon_2'| + 2|\varepsilon_3'| + \sqrt{2}\sqrt{2}|\varepsilon_4'|)d_j d_k \leq 57 d_j d_k\,\varepsilon \ . \end{aligned} \tag{3.2.26}$$

Now we construct the 2 by 2 submatrix $\delta\widehat{H}_m$. The analysis is similar to the analysis in the trigonometric case because a' can be computed with small relative error. We have

$$a' = fl(a + ct) = (1+\varepsilon_{a'})((1+\varepsilon_a)a + \widetilde{c}\widetilde{t}) \ ,$$

where

$$|\varepsilon_{a'}| \leq 13\varepsilon \ , \qquad |\varepsilon_a| \leq 14\varepsilon \ .$$

Since $|t| \leq \beta$ and $|c| < \beta a$, we can write (with the relative error of $O(\varepsilon)$, of course)

$$(1+\varepsilon_a)a + \widetilde{c}\widetilde{t} = (1+\varepsilon_a')(a + \widetilde{c}\widetilde{t}) \ ,$$

where

$$|\varepsilon_a'| = \left|\varepsilon_a\frac{\widetilde{c}\widetilde{t}}{a+\widetilde{c}\widetilde{t}}\right| \leq |\varepsilon_a|\frac{a\beta^2}{a(1-\beta^2)} = |\varepsilon_a| \ .$$

Therefore,

$$a' = (1+\varepsilon_{a'}')(a + \widetilde{c}\widetilde{t}) \ , \qquad |\varepsilon_{a'}'| \leq |\varepsilon_{a'}| + |\varepsilon_a'| \leq 27\cdot\varepsilon \ .$$

Also

$$b' = fl(b + ct) = (1+\varepsilon_{b'})(b + \widetilde{c}\widetilde{t} + \varepsilon_b b) \ , \qquad |\varepsilon_{b'}| \leq 13\varepsilon, \ |\varepsilon_b| \leq 14\varepsilon \ .$$

Let

$$\begin{aligned}
\Delta_1 &= \begin{bmatrix} 0 & \varepsilon_c c \\ \varepsilon_c c & 0 \end{bmatrix} + J_m^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon_b b \end{bmatrix} J_m^{-1} , \\
\Delta_2 &= \varepsilon'_{a'} \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 \right) , \\
\Delta_3 &= J_m^{-1} \begin{bmatrix} 0 & 0 \\ 0 & b'\left(1 - \dfrac{1+\varepsilon'_{a'}}{1+\varepsilon_{b'}}\right) \end{bmatrix} J_m^{-1} = \begin{bmatrix} \widetilde{sh}^2 \varepsilon'_{b''} b' & -\widetilde{ch}\,\widetilde{sh} \varepsilon'_{b''} b' \\ -\widetilde{ch}\,\widetilde{sh} \varepsilon'_{b''} b' & \widetilde{ch}^2 \varepsilon'_{b''} b' \end{bmatrix} ,
\end{aligned}$$

where $|\varepsilon'_{b''}| \leq |\varepsilon'_{a'}| + |\varepsilon_{b'}| \leq 40\varepsilon$. Then

$$J_m \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 + \Delta_2 + \Delta_3 \right) J_m = \begin{bmatrix} a' & 0 \\ 0 & b' \end{bmatrix} ,$$

and

$$\|\delta \widehat{A}_m\|_2 \leq |\varepsilon_c| + 3|\varepsilon_b| + 2|\varepsilon'_{a'}| + 3|\varepsilon'_{b''}| \leq 225.5\,\varepsilon \ .$$

Combining (3.2.26) with the above relation, we finally obtain

$$\|\delta A_m\|_2 \leq (225.5 + 62.45\sqrt{n-2})\,\varepsilon \ . \tag{3.2.27}$$

The theorem now follows from the relations (3.2.10), (3.2.22), (3.2.24) and (3.2.27). Q.E.D.

Corollary 3.2.2 *Assume Algorithm 3.1.1 converges, and that H_M, J is the final pair. Write $H_m = D_m A_m D_m$ with D_m diagonal and A_m with ones on the diagonal for $0 \leq m \leq M$. Let λ_j be the j-th eigenvalue of the pair $H, J \equiv H_0, J$ and $\lambda'_j = (H_M)_{jj} J_{jj}$. Then, with the relative error of $O(\varepsilon)$, the following error bound holds:*

$$\frac{|\lambda_j - \lambda'_j|}{|\lambda_j|} \leq \varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\lambda_{min}(A_m)} + n \cdot tol \ . \tag{3.2.28}$$

Proof. For every vector x and positive definite H we have

$$\begin{aligned}
|x^* \delta H x| &\leq |x^* D \delta A D x| \leq \|\delta A\|_2 |x^* D D x| \leq \|\delta A\|_2 \frac{|x^* D A D x|}{\lambda_{min}(A)} \\
&= \frac{\|\delta A\|_2}{\lambda_{min}(A)} x^* H x \ .
\end{aligned}$$

Let $\lambda_{m,j}$ denote the $j-$th eigenvalue of the pair H_m, J. Applying Th. 2.2.1 with $\delta J = 0$ and $\eta_J = 0$, and Th. 3.2.1 to the pairs J, H_m for $0 \leq m \leq M-1$, we obtain

$$1 - \eta_m \leq \frac{\lambda_{m+1,j}}{\lambda_{m,j}} \leq 1 + \eta_m \ , \tag{3.2.29}$$

where

$$\eta_m = \frac{C_m}{\lambda_{min}(A_m)}\varepsilon \ .$$

Applying Th. 2.2.1 and the stopping criterion to the pair J, H_M, and ignoring the $O(tol^2)$ term, we obtain

$$1 - n \cdot tol \le \frac{\lambda_j'}{\lambda_{M,j}} \le 1 + n \cdot tol \ . \tag{3.2.30}$$

Here we also used the fact that $\lambda_{min}(A_M) \ge 1 - n \cdot tol$. Since

$$\frac{\lambda_j'}{\lambda_J} = \frac{\lambda_{1,j}}{\lambda_j} \cdot \frac{\lambda_{2,j}}{\lambda_{1,j}} \cdots \frac{\lambda_{M,j}}{\lambda_{M-1,j}} \cdot \frac{\lambda_j'}{\lambda_{M,j}} \ , \tag{3.2.31}$$

the corollary follows by inserting (3.2.29) and (3.2.30) in the above relation, and ignoring the relative error of $O(\varepsilon)$. Q.E.D.

Here are some remarks about Th. 3.2.1 and Cor. 3.2.2. The remarks hold for all subsequent theorems and corollaries of the above type.

Remark 3.2.3 In the hyperbolic case for $\zeta \le \alpha = 3/2\sqrt{2}$ (Case I), the constant C_m depends additionally on $\sqrt{\kappa(A_m)}$. Deichmöller [8] also obtained a similar bound for some non–orthogonal transformations.

Remark 3.2.4 In practical computation Case I of Th. 3.2.1 occurs rarely, and almost never if we transform the pair $H, I \equiv GJG^T, I$ to the pair G^TG, J (due to diagonalizing effect of this transformation). Thus, our choice of α (and its function β) in (3.2.17) implies that the discontinuity of the bound (3.2.2) at $|\zeta| = 3/(2\sqrt{2})$ has little practical importance. This discontinuity can be removed by considering Case I, $|\zeta| \le 3/(2\sqrt{2})$, as Case II, $|\zeta| > \alpha'$, for some $\alpha' < 3/(2\sqrt{2})$. Also note that β cannot have an optimizing function as $\bar{x}$ in the trigonometric case, where the choice of $\bar{x}$ makes the bounds in the relations (3.2.7) and (3.2.9) almost equal. We can choose another approach when analysing the hyperbolic case in Th. 3.2.1, namely to analyse only the cases $d_j/d_i \ge \beta$ and $d_j/d_i < \beta$. Then the bounds (3.2.22) and (3.2.27) hold in the first and the second case, respectively. The approach of Th. 3.2.1 is, however, more enlightening and it simplifies the analysis of the modified method in the following subsection.

Remark 3.2.5 The $\sqrt{n-2}$ part of C_m may be multiplied by $\max_{m,i\ne j} |A_{m,ij}| < 1$. Thus if the matrices A_m are strongly diagonally dominant, the part of the error term which depends on n is suppressed.

Remark 3.2.6 Numerical experiments indicate that (3.2.28) grows only slowly with the increase of n or M.

3.2.1 The modified method

In order to avoid potentially large C_m in Th. 3.2.1 in the hyperbolic case for $|\zeta| \leq \alpha = 3/(2\sqrt{2})$, we modify the $J-$orthogonal Jacobi method by bounding the hyperbolic angle as suggested in [29]. Since the original method converges, large hyperbolic angles can occur only finitely many times. We first show that the modification does not affect convergence properties. We then prove that one step of the modified method satisfies the assumptions needed for the error bounds of Chap. 2, i.e. that Th. 3.2.1 and Cor. 3.2.2 hold with small modifications. The algorithm of the modified $J-$orthogonal Jacobi method is similar to Alg. 3.1.1. The only changes are the computation of the hyperbolic rotation parameters and the update of the pivot submatrix $\widehat{H}$.

Algorithm 3.2.7 *Modified two-sided $J-$orthogonal Jacobi method for the problem (3.1.1).*

/* *compute the parameter hyp: hyp* $= 1$ *for the hyperbolic and*
  $hyp = -1$ *for the trigonometric rotation, respectively* */
  if $1 \leq i \leq npos < j \leq n$ *then*
    $hyp = 1$
  else
    $hyp = -1$
  endif
/* *compute the hyperbolic Jacobi rotation which diagonalizes*
  $\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix}$, *and update the 2 by 2 pivot submatrix* */
  $\zeta = -hyp * (b + hyp * a)/(2c)$
  if $hyp = 1$ *and* $|\zeta| \leq \alpha \equiv 3/(2\sqrt{2})$ *then*
    $cs = \sqrt{2}$
    $sn = sign(\zeta)$
    $sn1 = sn$
    $H_{ii} = 2 * a + b - 2 * \sqrt{2} * |c|$
    $H_{jj} = a + 2 * b - 2 * \sqrt{2} * |c|$
    $H_{ij} = H_{ji} = sn * \sqrt{2} * (a + b) + 3 * c$
  else
    $t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2 - hyp})$
    $h = \sqrt{1 - hyp * t^2}$
    $cs = 1/h$
    $sn = t/h$
    $sn1 = hyp * sn$
    $H_{ii} = a + hyp * c * t$
    $H_{jj} = b + c * t$
    $H_{ij} = H_{ji} = 0$
  endif

proceed as in Algorithm 3.1.1

The convergence proof for the modified method [29] rests on the trace reduction which takes place in our case, too. Let $|\zeta| \leq \alpha$ and let H' and H'' denote the matrices after an unmodified and modified step, respectively. Then

$$a'' = 2a + b - 2\sqrt{2}|c| \ , \qquad b'' = a + 2b - 2\sqrt{2}|c| \ ,$$

and

$$\begin{aligned} \delta Tr' &\equiv Tr(H) - Tr(H') = 2|c||t| \ , \\ \delta Tr'' &\equiv Tr(H) - Tr(H'') = -2a - 2b + 4\sqrt{2}|c| \ . \end{aligned}$$

The quotient $\delta Tr''/\delta Tr' \leq 1$ is bounded below with β. For α and β from (3.2.17), we have

$$\frac{\delta Tr''}{\delta Tr'} = \frac{1}{|t|}(2\sqrt{2} - 2|\zeta|) \geq 2\sqrt{2} - 2\alpha = \frac{1}{\sqrt{2}} \ .$$

This trace reduction is quite acceptable. Also, modified steps do not affect the quadratic convergence. Indeed, Drmač and Hari [15] showed that the hyperbolic tangent is bounded by $|t| \leq \sqrt{2}/6$ after the quadratic convergence starts. This, in turn, implies $|\zeta| \geq 3/\sqrt{2}$, so the modified steps do not occur after the quadratic convergence starts.

The next theorem is an analog of Th. 3.2.1 and Cor. 3.2.2 for the modified $J-$orthogonal Jacobi method.

Theorem 3.2.8 *Let H_m be the sequence of matrices generated by Algorithm 3.2.7 in floating–point arithmetic with precision ε. Then Theorem 3.2.1 holds except that in the hyperbolic case for $|\zeta| \leq 3/(2\sqrt{2})$ the value of C_m is reduced to*

$$C_m = 82 + 19.63\sqrt{n-2} \ . \tag{3.2.32}$$

Corollary 3.2.2 holds with this exception, too.

Proof. The technique of the proof is the same as in Th. 3.2.1. We assume without loss of generality that $sh = +1$. Then sign $(c) = -1$. Using (3.2.1) we obtain

$$\begin{aligned} H''_{ik} &= fl(\sqrt{2}H_{ik} + H_{jk}) = \sqrt{2}H_{ik} + H_{jk} + \varepsilon'_1 H_{ik} + \varepsilon_1 H_{jk} \\ H''_{jk} &= fl(\sqrt{2}H_{jk} + H_{ik}) = \sqrt{2}H_{jk} + H_{ik} + \varepsilon'_2 H_{jk} + \varepsilon_2 H_{ik} \ , \end{aligned}$$

where $|\varepsilon'_1|, \ |\varepsilon'_2| \leq 3\sqrt{2}\varepsilon$. Since $d_j/d_i \geq \beta$, we have

$$|\delta H_{ik}| \leq (7 + 4\sqrt{2})d_i d_k \, \varepsilon \ , \qquad |\delta H_{jk}| \leq 15 d_j d_k \, \varepsilon \ .$$

Further,

$$\begin{aligned}
a'' &= fl(2a+b-2\sqrt{2}|c|) = 2a+b-2\sqrt{2}|c| + \varepsilon_3' a + \varepsilon_3 b + \varepsilon_4'|c| \\
b'' &= fl(a+2b-2\sqrt{2}|c|) = a+2b-2\sqrt{2}|c| + \varepsilon_4 a + \varepsilon_5' b + \varepsilon_6'|c| \\
c'' &= fl(\sqrt{2}(a+b)+3c) = \sqrt{2}(a+b)+3c + \varepsilon_7'(a+b) + \varepsilon_8' c
\end{aligned}$$

where

$$|\varepsilon_3'|,\ |\varepsilon_5'| \leq 4\varepsilon\ , \quad |\varepsilon_4'|,\ |\varepsilon_6'| \leq 8\sqrt{2}\varepsilon\ , \quad |\varepsilon_7'| \leq 4\sqrt{2}\varepsilon\ , \quad |\varepsilon_8'| \leq 6\varepsilon\ .$$

Setting

$$\Delta = \begin{bmatrix} \varepsilon_3' a + \varepsilon_3 b + \varepsilon_4'|c| & \varepsilon_7'(a+b) + \varepsilon_8' c \\ \varepsilon_7'(a+b) + \varepsilon_8' c & \varepsilon_4 a + \varepsilon_5' b + \varepsilon_6'|c| \end{bmatrix}$$

we have

$$J_m \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + J_m^{-1} \Delta J_m^{-1} \right) J_m = \begin{bmatrix} a'' & c'' \\ c'' & b'' \end{bmatrix} .$$

Using $d_j/d_i \geq \beta$, we obtain

$$\|\delta \widehat{A}_m\|_2 \leq 82\,\varepsilon\ ,$$

and finally (3.2.32). Q.E.D.

We have thus eliminated $\kappa(A_m)$ from C_m in the hyperbolic case for $\zeta \leq 3/(2\sqrt{2})$. This makes the one–step error bounds for the modified method of the same type as the corresponding bounds from [13], that is, the bounds depend only on $\sqrt{n-2}$. For 2×2 matrices, the use of modified rotations makes obviously no improvement. For $n \geq 3$, however, numerical experiments show that the use of modified rotations generally does not affect the convergence. Thus, the use of modified rotations generally decreases relative error estimates.

3.2.2 Growth of the condition of the scaled matrix

As we have seen in Cor. 3.2.2, the behaviour of the quotient $\lambda_{min}(A_0)/\lambda_{min}(A_m)$ (or $\kappa(A_m)/\kappa(A_0)$) is essential for the overall error bound of the $J-$orthogonal Jacobi method. In this subsection we first state known results. We then show that $\kappa(A_m)/\kappa(A_0) \leq n$ if $\kappa(A) \geq \kappa(H)$. After that we give a simple pattern for the behaviour of the upper bound for $\lambda_{min}(A_0)/\lambda_{min}(A_m)$. As a corollary we show that, with the appropriate choice of pivots, we can perform $n' \leq n-1$ successive steps such that $\lambda_{min}(A_0)/\lambda_{min}(A_m) \leq n$ for every $1 \leq m \leq n'$. In the conclusion, we define an algorithm for calculating the upper bound for $\lambda_{min}(A_0)/\lambda_{min}(A_m)$ in Jacobi process. Results of numerical experiments are given in Chap. 5. The results of this subsection are partially contained in [26].

We now state the known bounds for $1/\lambda_{min}(A_m)$, which were originally proved for the case $npos = n$ by Demmel and Veselić [13]. Later Veselić [28] noticed that the results also hold if the hyperbolic rotations are used, since the proofs do not require the orthogonality of rotation matrices. Let the pair H_m, J be obtained from the pair

H_0, J by applying m Jacobi rotations in pairwise nonoverlapping rows and columns (this means $m \leq n/2$), and let (i_k, j_k) be the pivot pair in the $k-$th step. We use the standard scaling, i.e.

$$H_m = D_m A_m D_m \ , \tag{3.2.33}$$

where D_m is positive definite diagonal matrix, and A_m has ones on the diagonal. The spectrum of A_m coincides with the spectrum of the pencil $A_0 - \lambda A_0'$, where A_0' coincides with A_0 on every rotated element and is the identity otherwise. This implies

$$\frac{1}{\lambda_{min}(A_m)} = \max_{x\neq 0} \frac{x^T A_0' x}{x^T A_0 x} \leq \frac{\max_{x\neq 0, \|x\|_2=1} x^T A_0' x}{\min_{x\neq 0, \|x\|_2=1} x^T A_0 x} = \frac{1 + \max_{0\leq k\leq m-1} |A_{0,i_k j_k}|}{\lambda_{min}(A_0)} \ . \tag{3.2.34}$$

After m arbitrary steps we have

$$\frac{1}{\lambda_{min}(A_m)} \leq \frac{\prod_{k=0}^{m-1}(1 + |A_{k,i_k j_k}|)}{\lambda_{min}(A_0)} \ .$$

The above upper bound for $1/\lambda_{min}(A_m)$ is usually a large overestimate.

The second bound is based on the Hadamard measure of a symmetric positive definite matrix H,

$$\mathcal{H}(H) \equiv \frac{\det(H)}{\prod_i H_{ii}}$$

It is easy to see that $\mathcal{H}(H) \leq 1$ and $\mathcal{H}(H) = 1$ if and only if H is diagonal. $\mathcal{H}(H)$ is independant of the scaling so that

$$\mathcal{H}(H) = \mathcal{H}(A) = \det A \ .$$

Furthermore,

$$\frac{1}{\lambda_{min}(A_m)} \leq \frac{e}{\mathcal{H}(H_m)} \ , \tag{3.2.35}$$

where $e = \exp(1)$, and

$$\frac{1}{\mathcal{H}(H_{m+1})} = \frac{1 - A_{m,ij}^2}{\mathcal{H}(H_m)} \leq \frac{1}{\mathcal{H}(H_m)} \ , \tag{3.2.36}$$

where (i, j) is the pivot pair in the $m-$th step. The above two relations can be used to monitor the convergence of $1/\lambda_{min}(A_m)$ to 1, but they can be a large overestimate in the beginning of the diagonalization process. Finally, (3.2.35) and (3.2.36) give the guaranteed upper bound

$$\max_m \frac{1}{\lambda_{min}(A_m)} \leq \frac{e}{\det(A_0)} = \frac{e}{\mathcal{H}(H_0)} \ .$$

The following simple result seems not to have attracted attention:

Proposition 3.2.9 *Let $npos = n$ and $\kappa(A) \geq \kappa(H)$. Let H_m, J be the sequence of pairs obtained by the $J-$orthogonal Jacobi method from the starting pair H, J. Then*

$$\kappa(A_m)/\kappa(A_0) \leq n \ ,$$

where matrices A_m are defined by (3.2.33).

Proof. The assumption $npos = n$ implies that all rotation matrices are orthogonal. The assumption $\kappa(A) \geq \kappa(H)$ and (1.5) imply

$$\kappa(A_m) \leq n \min_D \kappa(DA_mD) \leq n\kappa(H_m) = n\kappa(H) \leq n\kappa(A).$$

Q.E.D.

Now we come to the central result of this subsection:

Theorem 3.2.10 *Let $H_m = D_m A_m D_m$ be the sequence of matrices obtained by Algorithm 3.1.1 from the starting matrix $H \equiv H_0$, i.e.*

$$H_m = J_{m-1}^T H_{m-1} J_{m-1} \ .$$

Let us define the sequence of matrices T_m by

$$\begin{aligned} T_0 &= I \\ T_m &= T_{m-1} U_m \\ U_m &= D_{m-1}^{-1} J_{m-1}^{-T} D_m \ . \end{aligned}$$

Then for $m \geq 1$

$$\begin{aligned} A_m &= T_m^{-1} A_0 T_m^{-T} \ , \\ \frac{1}{\lambda_{min}(A_m)} &\equiv \|A_m^{-1}\|_2 \leq \|A_0^{-1}\|_2 \|T_m\|_2^2 \leq \|A_0^{-1}\|_2 \|T_m\|_E^2 = \frac{\|T_m\|_E^2}{\lambda_{min}(A_0)} \end{aligned} \quad (3.2.37)$$

and

$$\|T_m\|_E^2 = \|T_{m-1}\|_E^2 + 2A_{m-1,ij} T_{m-1,\cdot i}^T T_{m-1,\cdot j} \ . \quad (3.2.38)$$

Here (i,j) is the pivot pair in the $m-$th step, and $T_{m-1,\cdot i}$ denotes the $i-$th column of T_{m-1}, etc.

Proof. The first two statements of the theorem are obvious. Moreover, since $A_m \to I$ as $m \to \infty$, the relation

$$T_m A_m T_m^T = A_0$$

implies

$$\lim_{m\to\infty} T_m T_m^T = A_0 \ ,$$

and

$$\lim_{k\to\infty} U_{m+1}\cdot\dots\cdot U_k U_k^T \cdot\dots\cdot U_{m+1}^T = A_m \ .$$

It remains to prove the relation (3.2.38). From the definition of U_m we see that only its pivot submatrix $\widehat{U}_m$ differs from the identity matrix, and that

$$\widehat{U}_m \widehat{U}_m^T = \widehat{A}_{m-1} \ .$$

Also, $U_m^{-1} A_{m-1} U_m^{-T} = A_m$. Now we show that

$$U_m = \bar{J}_m^{-T} \bar{D}_m R_m \ , \tag{3.2.39}$$

where $\bar{J}_m$ is a $J-$orthogonal Jacobi rotation on A_{m-1}, $\bar{D}_m^{-1}$ scales $\bar{J}_m^T A_{m-1} \bar{J}_m$, and R_m is orthogonal. Indeed, set

$$\widehat{R}_m = \widehat{\bar{D}}_m \widehat{\bar{J}}_m^{-1} \widehat{D}_{m-1} \widehat{J}_{m-1} \widehat{D}_m^{-1} \ .$$

Then (3.2.39) is satisfied and R_m is orthogonal since

$$\begin{aligned}
\widehat{R}_m^T \widehat{R}_m &= \widehat{D}_m^{-1} \widehat{J}_{m-1}^T \widehat{D}_{m-1} \widehat{\bar{J}}_m^{-T} \widehat{\bar{D}}_m^{2} \widehat{\bar{J}}_m^{-1} \widehat{D}_{m-1} \widehat{J}_{m-1} \widehat{D}_m^{-1} \\
&= \widehat{D}_m^{-1} \widehat{J}_{m-1}^T \widehat{D}_{m-1} \widehat{A}_{m-1} \widehat{D}_{m-1} \widehat{J}_{m-1} \widehat{D}_m^{-1} \\
&= \widehat{D}_m^{-1} \widehat{J}_{m-1}^T \widehat{H}_{m-1} \widehat{J}_{m-1} \widehat{D}_m^{-1} \\
&= \widehat{D}_m^{-1} \widehat{H}_m \widehat{D}_m^{-1} = \widehat{I}
\end{aligned}$$

Note that the above relation holds for trigonometric as well as for hyperbolic rotations.

In the multiplication

$$T_m = T_{m-1} U_m$$

only the $i-$th and $j-$th column of T_{m-1} change, i.e.

$$\begin{bmatrix} T_{m,\cdot i} & T_{m,\cdot j} \end{bmatrix} = \begin{bmatrix} T_{m-1,\cdot i} & T_{m-1,\cdot j} \end{bmatrix} \widehat{U}_m \ . \tag{3.2.40}$$

Let $a \equiv A_{m-1,i_m j_m}$ and

$$\widehat{R}_m = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \ .$$

In the trigonometric case we have

$$\widehat{\bar{J}}_m = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \ , \qquad \widehat{\bar{D}}_m = \begin{bmatrix} \sqrt{1-a} & 0 \\ 0 & \sqrt{1+a} \end{bmatrix} \ .$$

which, together with (3.2.39) and (3.2.40), implies

$$\begin{aligned}
\begin{bmatrix} T_{m,ki} & T_{m,kj} \end{bmatrix} &= \begin{bmatrix} T_{m-1,ki} & T_{m-1,kj} \end{bmatrix} \cdot \\
&\cdot \frac{1}{\sqrt{2}} \begin{bmatrix} c\sqrt{1-a} - s\sqrt{1+a} & s\sqrt{1-a} + c\sqrt{1+a} \\ -c\sqrt{1-a} - s\sqrt{1+a} & -s\sqrt{1-a} + c\sqrt{1+a} \end{bmatrix} \ .
\end{aligned}$$

After a simple but rather long calculation, we obtain

$$T_{m,ki}^2 + T_{m,kj}^2 = T_{m-1,ki}^2 + T_{m-1,kj}^2 + 2aT_{m-1,ki}T_{m-1,kj} \ . \qquad (3.2.41)$$

The relation (3.2.38) now follows by summing up (3.2.41) for $k = 1, \ldots, n$.

In the hyperbolic case we have

$$\hat{\hat{J}}_m = \begin{bmatrix} ch & sh \\ sh & ch \end{bmatrix} , \qquad \hat{\hat{D}}_m = \begin{bmatrix} \sqrt{1+ta} & 0 \\ 0 & \sqrt{1+ta} \end{bmatrix} ,$$

where (see Alg. 3.1.1)

$$\zeta = -\frac{1}{a} , \qquad t = -\frac{a}{1+\sqrt{1-a^2}} ,$$
$$ch^2 + sh^2 = \frac{1}{\sqrt{1-a^2}} , \qquad sh \cdot ch = -\frac{a}{2\sqrt{1-a^2}} .$$

This, together with (3.2.39) and (3.2.40), implies

$$\begin{bmatrix} T_{m,ki} & T_{m,kj} \end{bmatrix} = \begin{bmatrix} T_{m-1,ki} & T_{m-1,kj} \end{bmatrix} \sqrt{1+ta} \begin{bmatrix} c \cdot ch + s \cdot sh & s \cdot ch - c \cdot sh \\ -c \cdot sh - s \cdot ch & -s \cdot sh + c \cdot ch \end{bmatrix} .$$

After a simple but rather long calculation, we obtain again (3.2.41) and the theorem is proved. Q.E.D.

Corollary 3.2.11 *Let the matrices $H_m = D_m A_m D_m$, T_m, and U_m be defined as in Theorem 3.2.10. Let us perform $n' \leq n-1$ successive steps of the $J-$orthogonal Jacobi method such that for every $m \in \{1, \ldots, n'\}$ and every $k \in \{1, \ldots, n\}$ either $T_{m-1,ki_m} = 0$ or $T_{m-1,kj_m} = 0$. Here (i_m, j_m) denotes the pivot pair in the $m-$th step. (These assumptions are fulfilled e.g. if we choose pivot pairs along the first row, or along the last column, or along the first off–diagonal.) Then*

$$1/\lambda_{min}(A_m) \leq n/\lambda_{min}(A_0)$$

for every $m \in \{1, \ldots, n'\}$.

Proof. By definition is $\|T_0\|_E^2 = n$. The corollary follows from the assumptions and the relations (3.2.38) and (3.2.37). Q.E.D.

Now we derive an efficient algorithm for calculating the upper bound for $1/\lambda_{min}(A_m)$ in Jacobi process. The inequality (3.2.37) implies that

$$1/\lambda_{min}(A_m) \leq \|T_m\|_E^2/\lambda_{min}(A_0) \ .$$

We can calculate $\|T_m\|_E^2$ using the recursive equation (3.2.38) in the following manner: instead of keeping the eigenvector matrix V according to Alg. 3.1.1,

$$\begin{aligned} V_0 &= I \\ V_m &= J_0 J_1 \cdots J_{m-1} = V_{m-1} J_{m-1} \ , \end{aligned}$$

we keep the matrix S defined by

$$\begin{aligned} S_0 &= D_0^{-1} \\ S_m &= D_0^{-1} J_0^{-T} J_1^{-T} \cdots J_{m-1}^{-T} = S_{m-1} J_{m-1}^{-T} \ . \end{aligned}$$

In the trigonometric case we have $J_{m-1}^{-T} = J_m$, and in the hyperbolic case we have

$$\hat{J}_{m-1}^{-T} = \hat{J}_{m-1}^{-1} = \begin{bmatrix} ch & -sh \\ -sh & ch \end{bmatrix} \ .$$

Also

$$V_m^{-T} = D_0 S_m \ , \qquad T_m = S_m D_m \ . \tag{3.2.42}$$

In order to apply (3.2.38), we need to calculate the scalar product of the $i-$th and $j-$th column of T_m. From (3.2.42), we see that

$$T_{m,\cdot i}^T T_{m,\cdot j} = S_{m,\cdot i}^T S_{m,\cdot j} D_{m,ii} D_{m,jj} \ .$$

Therefore, the sequence $\|T_m\|_E^2$ is given by the recursion

$$\begin{aligned} \|T_0\|_E^2 &= n \\ \|T_m\|_E^2 &= \|T_{m-1}\|_E^2 + 2H_{m-1,ij} S_{m-1,\cdot i}^T S_{m-1,\cdot j} \ , \end{aligned} \tag{3.2.43}$$

at a cost of $n+2$ multiplications and n additions in each step.

Suppose that the algorithm converges, and that H_M, J is the final pair. Then (3.2.42) implies that

$$V_M^{-T} = D_0 S_M \ ,$$

but we want to obtain the eigenvector matrix V_M. Since V_M is $J-$orthogonal, i.e. $V_M^T J V_M = J$, we have

$$V_M = J V_M^{-T} J \ .$$

Multiplication with D_0 from left has relative error ε and multiplications with J have no error at all.

In numerical experiments sequence $\|T_m\|_E^2$ behaved extremely well in the sense that it was approximately n for all m. However, the recursion (3.2.43) does not reveal the fact that $1/\lambda_{min}(A_m)$ tends to one. This convergence can be monitored using the monotonically decreasing upper bound (3.2.35). This bound is usually large in the beginning of the diagonalization process, and it meets the bound given by (3.2.43) after one or two cycles. After that point (3.2.43) is not needed any more. Updating $\mathcal{H}(H_m)$ according to (3.2.36) is very simple. The only additional effort is to calculate $\mathcal{H}(H_0)$ (for example by using the Cholesky decomposition of H_0).

Remark 3.2.12 The theoretical results of this section, as well as numerical observations, do not depend upon whether only trigonometric ($J = I$), or trigonometric *and* hyperbolic rotations are used. This once more justifies the use of the hyperbolic rotations.

3.3 Implicit J–orthogonal Jacobi method

In this section we present and analyse the implicit (one–sided) J–orthogonal Jacobi method for solving the eigenvalue problem

$$Hx = \lambda x \ , \qquad x \neq 0 \ , \tag{3.3.1}$$

where H is a $n \times n$ real symmetric matrix of rank $\operatorname{rank}(H) = r \leq n$. Let H be decomposed as

$$H = GJG^T \ , \tag{3.3.2}$$

where G is a $n \times r$ matrix (i.e. G has full column rank), $J = I_{npos} \oplus (-I_{r-npos})$, and $npos$ is number of the positive eigenvalues of H. The symmetric indefinite decomposition (3.3.2) is described in Chap. 4. Since $J^{-1} = J$, Th. 2.3.1 implies that the eigenvalues of the pair G^TG, J are the nonzero eigenvalues of H, and that there exists a J–orthogonal matrix F ($F^TJF = J$) such that the matrix

$$F^TG^TGF \equiv \Delta$$

is diagonal and positive definite. Therefore, nonzero eigenvalues of the problem (3.3.1) are the diagonal elements of the diagonal matrix ΔJ, and the corresponding eigenvectors are the columns of the matrix

$$U = GF\Delta^{-1/2} \ .$$

Instead of forming explicitly the matrix G^TG and applying Alg. 3.1.1 to the pair G^TG, J, we apply the implicit J–orthogonal Jacobi method to the pair G, J. The method, originally proposed by Veselić [29], consists of an iterative application of the one–sided transformation

$$G_{m+1} = G_m J_m \ ,$$

where $G \equiv G_0$ and J_m is a J–orthogonal Jacobi plane rotation.

If G is square and non–singular, the method also solves the hyperbolic singular value problem [21] for the pair G, J.

Note that in the positive definite case [13], the implicit method can be applied either to G *or* G^T (since $J = I$, the matrices G^TG *and* GG^T have the same eigenvalues and simply related eigenvectors). Here, even if H is non–singular (G is non–singular and square), only one application makes sense, i.e. from the right on G or from the left on G^T (see also Sect. 2.3).

The section is organized as follows: we first present the algorithm. Then we prove that in floating–point arithmetic the method computes the non–zero eigenvalues of H with the error bounds of Chap. 2. We analyse the simple version of the algorithm, omitting enhancements like keeping the diagonal in a separate vector and fast rotations, to make the error analysis clearer. In Subsect. 3.3.1 we analyse the version of the algorithm where diagonal of G^TG is kept in a separate vector. In Subsect. 3.3.2 we give the norm error bounds for the computed eigenvectors if H is non–singular (non–singularity is neccessary since we use the eigenvector perturbation bounds from Chap. 2). In Sect. 3.4 we analyse the fast version of the algorithm. In Subsect. 3.4.1 we analyse the fast method which uses self–scaling rotations. These rotations, introduced and analysed by Anda and Park [1] for the trigonometric case, are used to suppress possible underflow/overflow when accumulating the diagonal of the fast rotations.

We now present our algorithm:

Algorithm 3.3.1 *Implicit $J-$orthogonal Jacobi method for the pair G, J. tol is a user defined stopping criterion.*

```
repeat
    for all pairs i < j
        /* compute [a c; c b] ≡ the (i,j) submatrix of G^T G */
            a = Σ_{k=1}^n G_{ki}^2
            b = Σ_{k=1}^n G_{kj}^2
            c = Σ_{k=1}^n G_{ki} * G_{kj}
        /* compute the parameter hyp: hyp = 1 for the hyperbolic and
            hyp = −1 for the trigonometric rotation, respectively */
            if 1 ≤ i ≤ npos < j ≤ r then
                hyp = 1
            else
                hyp = −1
            endif
        /* compute the J−orthogonal Jacobi rotation which diagonalizes
                [H_ii H_ij; H_ji H_jj] ≡ [a c; c b] */
            ζ = −hyp * (b + hyp * a)/(2c)
            t = sign(ζ)/(|ζ| + sqrt(ζ^2 − hyp))
            h = sqrt(1 − hyp * t^2)
            cs = 1/h
            sn = t/h
            sn1 = hyp * sn
        /* update columns i and j of G */
            for k = 1 to n
                tmp = G_ki
                G_ki = cs * tmp + sn1 * G_kj
                G_kj = sn * tmp + cs * G_kj
            endfor
    endfor
until convergence (all |c|/sqrt(ab) ≤ tol)
/* the computed non–zero eigenvalues of H = GJG^T (and of the pair G^T G, J) are
        λ_j = (Σ_{k=1}^n G_{kj}^2) J_jj */
/* the computed eigenvectors of H are the normalized columns of the final G */
```

Remark 3.3.2 If G is square and non–singular, then the computed hyperbolic singular values [21] of the pair G, J are

$$\sigma_j = \sqrt{\sum_{k=1}^{n} G_{kj}^2 \, J_{jj}} \ ,$$

and the computed hyperbolic singular vectors are the normalized columns of the final G. This remark holds for all subsequent implicit methods in this chapter.

The perturbation theory for the problem (3.3.1), as well as for the hyperbolic singular value problem [21], is given by Theorems 2.3.1 and 2.3.2. Let G_m be the sequence of matrices obtained by Alg. 3.3.1 from the starting matrix $G \equiv G_0$. For every $m \geq 0$ write $G_m = B_m D_m$, where D_m is diagonal positive definite, and the columns of B_m have unit norms. All error bounds in this section contain the quantities $1/\sigma_{min}(B_m)$, whereas the perturbation bounds in Chap. 2 are proportional to $1/\sigma_{min}(B_0)$ (or $\kappa(B_0)$). Therefore, as in Sect. 3.2, our claim that the implicit $J-$orthogonal Jacobi method is as accurate as predicted in Sect. 2.3 depends on the ratio $\max_m \sigma_{min}(B_0)/\sigma_{min}(B_m)$ (or $\max_m \kappa(B_m)/\kappa(B_0)$) being modest. In exact arithmetic, one-sided Jacobi on $G = BD$ is identical to two-sided Jacobi on $H = G^T G = DB^T BD = DAD$. Thus, all convergence properties of the explicit method carry naturally over to the implicit one, and the question of the growth of $\kappa(B_m) = \kappa(A_m)^{1/2}$ is essentially identical to the question of the growth of $\kappa(A_m)$ in the case of two-sided Jacobi. Therefore, the results of Subsect. 3.2.2 apply here, as well.

The following theorem and its corollary justify our accuracy claims for the non–zero eigenvalues of the matrix $H = GJG^T$ computed by the implicit $J-$orthogonal Jacobi method.

Theorem 3.3.3 *Let G_m be the sequence of matrices generated by the implicit $J-$orthogonal Jacobi algorithm in floating–point arithmetic with precision ε; that is G_{m+1} is obtained from G_m by applying a single $J-$orthogonal Jacobi rotation. Then the following diagram commutes:*

$$\begin{tikzcd} G_m \arrow[r, "\text{floating}", "\text{Jacobi}"'] \arrow[dd] & G_{m+1} \\ & \\ G_m + \delta G_m \arrow[uur, "\text{exact rotation}"'] & \end{tikzcd}$$

The top arrow indicates that G_{m+1} is obtained from G_m by applying one $J-$orthogonal Jacobi rotation in floating–point arithmetic. The diagonal arrow indicates that G_{m+1} is obtained from $G_m + \delta G_m$ by applying one $J-$orthogonal plane rotation in exact arithmetic; thus $G_{m+1} J G_{m+1}^T$ and $(G_m + \delta G_m) J (G_m + \delta G_m)^T$ have identical non–zero eigenvalues and the corresponding eigenvectors. δG_m is bounded as follows: let $\kappa = \kappa^2(B_m)$, and write $\delta G_m = \delta B_m D_m$, where D_m is diagonal such that B_m in $G_m = B_m D_m$ has unit columns. Let a_T and b_T be the true values of $\sum_k G_{ki}^2$ and $\sum_k G_{kj}^2$, respectively. Then, with the relative error of order ε,

$$\|\delta B_m\|_2 \leq C_m \, \varepsilon \ , \tag{3.3.3}$$

where

$$C_m = \begin{cases} 26 & \text{in trigonometric case} \; , \\ \kappa + 13\sqrt{\kappa} + 29 & \text{in hyperbolic case} \; , \quad |\zeta| \le \dfrac{3}{2\sqrt{2}} \; , \\ 77 & \text{in hyperbolic case} \; , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \; , \\ & \qquad b_T \ge \dfrac{1}{2} a_T \; , \\ 96 & \text{in hyperbolic case} \; , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \; , \\ & \qquad b_T < \dfrac{1}{2} a_T \; . \end{cases}$$

In other words, one step of the implicit $J-$orthogonal Jacobi method satisfies the assumptions needed for the perturbation bounds of Sect. 2.3.

PROOF. The proof of the commuting diagram is a tedious computation. We shall prove the diagram separately for the trigonometric and for the hyperbolic case. Let c_T be the true value of $\sum_k G_{ki} G_{kj}$. As in (3.2.3), we set

$$a_T = d_i^2 \; , \qquad b_T = d_j^2 \; , \qquad c_T = z d_i d_j \; .$$

We may assume without loss of generality that $a_T \ge b_T$ and $c_T > 0$. As in (3.2.4), we have

$$0 < z \le \bar{z} \equiv (\kappa^2(B_m) - 1)/(\kappa^2(B_m) + 1) < 1 \; . \tag{3.3.4}$$

Set $x \equiv d_j/d_i$. Note that $x \le 1$. Systematic application of formulae (3.2.1) shows that

$$\begin{aligned} a &= a_T(1 + \varepsilon_a) \quad \text{where} \quad |\varepsilon_a| \le n\varepsilon \\ b &= b_T(1 + \varepsilon_b) \quad \text{where} \quad |\varepsilon_b| \le n\varepsilon \\ c &= c_T + \varepsilon_c \sqrt{a_T b_T} \quad \text{where} \quad |\varepsilon_c| \le n\varepsilon \; . \end{aligned}$$

Trigonometric case. This case was analysed by Demmel and Veselić [13] and we present it for the sake of completeness. Small differences in the proof give here, again, a somewhat better bound for $\|\delta B_m\|_2$.

Let

$$\widetilde{cs} \equiv 1/\sqrt{1+t^2} \; , \qquad \widetilde{sn} \equiv t/\sqrt{1+t^2} \; .$$

From (3.2.1) we get

$$sn = (1 + \varepsilon_{sn})\widetilde{sn} \; , \qquad cs = (1 + \varepsilon_{cs})\widetilde{cs} \; , \qquad |\varepsilon_{sn}|, \; |\varepsilon_{cs}| \le 3\varepsilon \; .$$

$\widetilde{cs}$ and $\widetilde{sn}$ define the exact rotation

$$J_m = \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix}$$

which takes $G_m + \delta G_m$ to G_{m+1}:

$$(G_m + \delta G_m) J_m = G_{m+1} \ .$$

Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. Then

$$\begin{aligned} G'_{ki} &= fl(cs * G_{ki} - sn * G_{kj}) \\ &= (1+\varepsilon_1)(1+\varepsilon_2) cs G_{ki} - (1+\varepsilon_3)(1+\varepsilon_4) sn G_{kj} \\ &= (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_{cs})\widetilde{cs} G_{ki} - (1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_{sn})\widetilde{sn} G_{kj} \\ &\equiv \widetilde{cs} G_{ki} - \widetilde{sn} G_{kj} + E_{ki} \ , \end{aligned} \tag{3.3.5}$$

and, similarly,

$$G'_{kj} = fl(sn * G_{ki} + cs * G_{kj}) = \widetilde{sn} G_{ki} + \widetilde{cs} G_{kj} + E_{kj} \ , \tag{3.3.6}$$

where

$$\begin{aligned} \|E_{\cdot i}\|_2 &\le 5(\widetilde{cs}\|G_{\cdot i}\|_2 + |\widetilde{sn}|\|G_{\cdot j}\|_2)\varepsilon \\ \|E_{\cdot j}\|_2 &\le 5(|\widetilde{sn}|\|G_{\cdot i}\|_2 + \widetilde{cs}\|G_{\cdot j}\|_2)\varepsilon \ . \end{aligned}$$

Here $G_{\cdot i}$ refers to the i-th column of G, etc. Thus

$$\begin{aligned} \begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} &= \begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \\ &= \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{cs} & -\widetilde{sn} \\ \widetilde{sn} & \widetilde{cs} \end{bmatrix} \right) \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix} \\ &\equiv \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} F_{\cdot i} & F_{\cdot j} \end{bmatrix} \right) \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix} \ , \end{aligned} \tag{3.3.7}$$

where

$$\begin{aligned} \|F_{\cdot i}\|_2 &\le \widetilde{cs}\|E_{\cdot i}\|_2 + |\widetilde{sn}|\|E_{\cdot j}\|_2 \\ &\le (5\|G_{\cdot i}\|_2 + 10\widetilde{cs}|\widetilde{sn}|\|G_{\cdot j}\|_2)\,\varepsilon \\ &\le 5(1+x) d_i \varepsilon \ , \end{aligned} \tag{3.3.8}$$

and

$$\begin{aligned} \|F_{\cdot j}\|_2 &\le |\widetilde{sn}| \cdot \|E_{\cdot i}\|_2 + \widetilde{cs}\|E_{\cdot j}\|_2 \\ &\le (5\|G_{\cdot j}\|_2 + 10\widetilde{cs}|\widetilde{sn}|\|G_{\cdot i}\|_2)\,\varepsilon \\ &\le 5(1 + 2\widetilde{cs}|\widetilde{sn}|/x) d_j \varepsilon \ . \end{aligned} \tag{3.3.9}$$

We consider two cases, $x < \bar{x} \equiv 0.48$, and $x \ge \bar{x}$. First consider $x < \bar{x}$. By inserting $\bar{x}$ for x in (3.3.8) we obtain

$$\|F_{\cdot i}\|_2 \le 7.4 d_i \varepsilon \ .$$

Our assumption further implies that the subtraction $1 - x^2$ has a low relative error, and that $z + n\varepsilon < 1$ with a relative error of $O(\varepsilon)$. Therefore

$$\begin{aligned} |t| &\leq \frac{|c|}{|b-a|} = \frac{|c_T + \varepsilon_c\sqrt{a_T b_T}|}{|b_T + \varepsilon_b b_T - a_T - \varepsilon_a a_T|} \\ &= \frac{|zx + \varepsilon_c x|}{|x^2 - 1 + \varepsilon_b x^2 - \varepsilon_a|} \leq \frac{x(z+n\varepsilon)}{1-\bar{x}^2}(1+O(\varepsilon)) \ . \end{aligned} \tag{3.3.10}$$

We can ignore the $(z + n\varepsilon)(1 + O(\varepsilon))$ term, so that $|t| \leq x/(1 - \bar{x}^2)$. Inserting this inequality into (3.3.9) we obtain

$$\|F_{\cdot j}\|_2 \leq 5(1 + \frac{2}{1-\bar{x}^2})d_j\varepsilon \leq 18 d_j \varepsilon \ .$$

Here we also used $\widetilde{cs}|\widetilde{sn}| \leq \widetilde{cs}^2|t| \leq |t|$. Therefore,

$$\|\delta B_m\|_2 \leq \frac{\|F_{\cdot i}\|_2}{d_i} + \frac{\|F_{\cdot j}\|_2}{d_j} \leq 26\varepsilon \tag{3.3.11}$$

Now consider the case $x \geq \bar{x}$. Inserting 1 for x in (3.3.8) we obtain

$$\|F_{\cdot i}\|_2 \leq 10 d_i \varepsilon \ .$$

Inserting $\widetilde{cs}|\widetilde{sn}| \leq 1/2$ and $1/\bar{x}$ for $1/x$ in (3.3.9), we obtain

$$\|F_{\cdot j}\|_2 \leq 15.5 d_j \varepsilon \ ,$$

so that (3.3.11) holds again, thus improving the bound $\|\delta B_m\|_2 \leq 72\varepsilon$ from [13].

Hyperbolic case. For the sake of the clarity, we denote the quantities cs, sn and $sn1 = sn$ computed by Alg. 3.1.1 with ch and sh, respectively. Let

$$\widetilde{ch} \equiv 1/\sqrt{1-t^2} \ , \qquad \widetilde{sh} \equiv t/\sqrt{1-t^2} \ .$$

Using (3.2.1) we can show that the bounds (3.2.11) hold for t, $\widetilde{ch}$ and $\widetilde{sh}$ with a relative error of $O(\varepsilon)$. Suppose that we can write

$$sh = (1 + \varepsilon_{sh})\widetilde{sh} \ , \qquad ch = (1 + \varepsilon_{ch})\widetilde{ch} \ .$$

$\widetilde{ch}$ and $\widetilde{sh}$ define the exact rotation

$$J_m = \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix}$$

which takes $G_m + \delta G_m$ to G_{m+1}:

$$(G_m + \delta G_m)J_m = G_{m+1} \ .$$

Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. Then

$$\begin{aligned} G'_{ki} &= fl(ch * G_{ki} + sh * G_{kj}) \\ &= (1+\varepsilon_1)(1+\varepsilon_2) ch G_{ki} + (1+\varepsilon_3)(1+\varepsilon_4) sh G_{kj} \\ &= (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_{ch})\widetilde{ch} G_{ki} + (1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_{sh})\widetilde{sh} G_{kj} \\ &\equiv \widetilde{ch} G_{ki} + \widetilde{sh} G_{kj} + E_{ki} \ , \end{aligned} \tag{3.3.12}$$

and, similarly,

$$G'_{kj} = fl(sh * G_{ki} + ch * G_{kj}) = \widetilde{sh} G_{ki} + \widetilde{ch} G_{kj} + E_{kj} \ , \tag{3.3.13}$$

where

$$\begin{aligned} \|E_{\cdot i}\|_2 &\le |\varepsilon'_1| \widetilde{ch} \|G_{\cdot i}\|_2 + |\varepsilon'_2| |\widetilde{sh}| \|G_{\cdot j}\|_2 \\ \|E_{\cdot j}\|_2 &\le |\varepsilon'_3| |\widetilde{sh}| \|G_{\cdot i}\|_2 + |\varepsilon'_4| \widetilde{ch} \|G_{\cdot j}\|_2 \ . \end{aligned}$$

Here

$$|\varepsilon'_1|, |\varepsilon'_4| = |\varepsilon_{ch}| + 2\varepsilon \ , \qquad |\varepsilon'_2|, |\varepsilon'_3| = |\varepsilon_{sh}| + 2\varepsilon \ . \tag{3.3.14}$$

Thus

$$\begin{aligned} \begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} &= \begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \\ &= \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{ch} & -\widetilde{sh} \\ -\widetilde{sh} & \widetilde{ch} \end{bmatrix} \right) \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix} \\ &\equiv \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} F_{\cdot i} & F_{\cdot j} \end{bmatrix} \right) \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix} \ , \end{aligned} \tag{3.3.15}$$

where

$$\begin{aligned} \|F_{\cdot i}\|_2 &\le \widetilde{ch} \|E_{\cdot i}\|_2 + |\widetilde{sh}| \|E_{\cdot j}\|_2 \\ &\le (|\varepsilon'_1| \widetilde{ch}^2 + |\varepsilon'_3| \widetilde{sh}^2) \|G_{\cdot i}\|_2 + (|\varepsilon'_2| + |\varepsilon'_4|) \widetilde{ch} |\widetilde{sh}| \|G_{\cdot j}\|_2 \\ &\le (|\varepsilon'_1| \widetilde{ch}^2 + |\varepsilon'_3| \widetilde{sh}^2 + (|\varepsilon'_2| + |\varepsilon'_4|) \widetilde{ch} |\widetilde{sh}| x) d_i \ , \end{aligned} \tag{3.3.16}$$

and

$$\begin{aligned} \|F_{\cdot j}\|_2 &\le |\widetilde{sh}| \cdot \|E_{\cdot i}\|_2 + \widetilde{ch} \|E_{\cdot j}\|_2 \\ &\le (|\varepsilon'_4| \widetilde{ch}^2 + |\varepsilon'_2| \widetilde{sh}^2) \|G_{\cdot j}\|_2 + (|\varepsilon'_1| + |\varepsilon'_3|) \widetilde{ch} |\widetilde{sh}| \|G_{\cdot i}\|_2 \\ &\le \left(|\varepsilon'_4| \widetilde{ch}^2 + |\varepsilon'_2| \widetilde{sh}^2 + (|\varepsilon'_1| + |\varepsilon'_3|) \widetilde{ch} |\widetilde{sh}| \frac{1}{x} \right) d_j \ . \end{aligned} \tag{3.3.17}$$

Now we have to calculate the upper bounds for $|\varepsilon'_i|$'s, $\widetilde{ch}^2$, $\widetilde{sh}^2$ and $\widetilde{ch}|\widetilde{sh}|$, and to insert them into relations (3.3.16) and (3.3.17). We consider two cases, $|\zeta| \le \alpha$ and $|\zeta| > \alpha$, where α is defined by (3.2.17).

First consider $|\zeta| \le \alpha$. As in the proof of Th. 3.2.1 we can show that the relations (3.2.18) and (3.2.20) hold. From (3.2.11) it follows that

$$fl(\sqrt{1-t^2}) = (1+\varepsilon_h)\sqrt{1-t^2} \ ,$$

where

$$|\varepsilon_h| \le \left(\frac{3}{8}\sqrt{\kappa} + \frac{4}{3}\right)\varepsilon \ .$$

Therefore,

$$|\varepsilon_{sh}|, |\varepsilon_{ch}| \le \left(\frac{3}{8}\sqrt{\kappa} + \frac{7}{3}\right)\varepsilon \ ,$$

so that

$$|\varepsilon_i'| \le \left(\frac{3}{8}\sqrt{\kappa} + \frac{13}{3}\right)\varepsilon \ , \qquad i = 1, \ldots, 4 \ .$$

Inserting $1/x \le \sqrt{2}$, (3.2.20), and the above relation in (3.3.16) and (3.3.17), we obtain

$$\begin{aligned} \|F_{\cdot i}\|_2 &\le (0.375\kappa + 5.46\sqrt{\kappa} + 13)d_i\varepsilon \\ \|F_{\cdot j}\|_2 &\le (0.46\kappa + 6.6\sqrt{\kappa} + 15.67)d_j\varepsilon \ . \end{aligned}$$

This, in turn, implies $\|\delta B_m\|_2 \le C_m$ as desired.

Now consider the case $|\zeta| > \alpha$. As in the proof of Th. 3.2.1, we can show that the relations (3.2.23) hold for t, $\widetilde{ch}$ and $\widetilde{sh}$ with a relative error of $O(\varepsilon)$. Now

$$fl(\sqrt{1-t^2}) = (1+\varepsilon_h)\sqrt{1-t^2} \ , \qquad |\varepsilon_h| \le 3\varepsilon \ ,$$

so that

$$|\varepsilon_{sh}|, |\varepsilon_{ch}| \le 4\varepsilon \ , \qquad |\varepsilon_i'| \le 6\varepsilon \ , \qquad i = 1, \ldots, 4 \ . \tag{3.3.18}$$

We have two subcases, $x \ge \beta$ and $x < \beta$, where β is defined by (3.2.17). If $x \ge \beta$, then inserting $1/x \le \sqrt{2}$, (3.3.18) and (3.2.23) into (3.3.16) and (3.3.17) yields

$$\|F_{\cdot i}\|_2 \le 35 d_i\varepsilon \ , \qquad \|F_{\cdot j}\|_2 \le 42 d_j\varepsilon \ , \qquad \|\delta B_m\|_2 \le 77\varepsilon \ ,$$

as desired.

If $x < \beta$, then

$$\begin{aligned} |t| &\le \frac{1}{|\zeta|} = \frac{2|c|}{|a+b|} = \frac{2|c_T + \varepsilon_c\sqrt{a_T b_T}|}{|a_T + \varepsilon_a a_T + b_T + \varepsilon_b b_T|} \\ &= \frac{2|zx + \varepsilon_c x|}{|1 + x^2 + \varepsilon_a + \varepsilon_b x^2|} \le 2x(z + n\varepsilon)(1 + O(\varepsilon)) \ . \end{aligned} \tag{3.3.19}$$

We can ignore the $(z+n\varepsilon)(1+O(\varepsilon))$ term, so that $|t| \le 2x$. Therefore,

$$\widetilde{ch}|\widetilde{sh}| = \widetilde{ch}^2|t| \le 2\widetilde{ch}^2 x \ .$$

Inserting this, (3.3.18) and (3.2.23) into (3.3.16) and (3.3.17), we obtain

$$\|F_{\cdot i}\|_2 \leq 30 d_i \varepsilon \ , \qquad \|F_{\cdot j}\|_2 \leq 66 d_j \varepsilon \ , \qquad \|\delta B_m\|_2 \leq 96\varepsilon \ ,$$

and the theorem is proved. Q.E.D.

Corollary 3.3.4 *Assume Algorithm 3.3.1 converges, and that G_M, J is the final pair which satisfies the stopping criterion. For $0 \leq m \leq M$ write $G_m = B_m D_m$ with D_m diagonal and B_m with unit columns. Let λ_j be the j-th non-zero eigenvalue of $G_0 J G_0^T$, and let λ'_j be the j-th computed non–zero eigenvalue. Then, with the relative error of $O(\varepsilon)$,*

$$(1-\gamma)^2 \leq \frac{\lambda'_j}{\lambda_j} \leq (1+\gamma)^2 \ , \tag{3.3.20}$$

where

$$\gamma = \varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol/2 + r \cdot n \cdot \varepsilon/2 \ .$$

Proof. Let $\lambda_{m,j}$ denote the j–th non–zero eigenvalue of the matrix $G_m J G_m^T$. By substituting (3.3.3) into (2.3.12) and then applying Th. 2.3.1 for every $0 \leq m \leq M-1$, we obtain

$$(1-\eta_m)^2 \leq \frac{\lambda_{m+1,j}}{\lambda_{m,j}} \leq (1+\eta_m)^2 \ , \tag{3.3.21}$$

where

$$\eta_m = \varepsilon C_m / \sigma_{min}(B_m) \ .$$

Also,

$$\operatorname{diag}(\lambda'_j) = \operatorname{diag}(fl(G_M^T G_M)) = fl(G_M^T G_M) + F \ ,$$

where

$$|F_{ij}| \leq (n\varepsilon + tol) \|G_{M\cdot i}\|_2 \|G_{M,\cdot j}\|_2 \ .$$

Here $G_{M,\cdot i}$ denotes the i–th column of G_M. The tol term comes from the stopping criterion. The $n\varepsilon$ term comes for the off–diagonal elements of F from the fact that $c/\sqrt{ab}$ in the stopping criterion may be underestimated by as much as $n\varepsilon$, and for the diagonal elements of F from computing the norms of the columns of G_M. Therefore,

$$1 - r \cdot n \cdot \varepsilon - n \cdot tol \leq \frac{\lambda'_j}{\lambda_{M,j}} \leq 1 + r \cdot n \cdot \varepsilon + n \cdot tol \ ,$$

and (3.3.20) follows by inserting (3.3.21) and the above relation into (3.2.31), and ignoring the relative error of $O(\varepsilon)$. Q.E.D.

An alternative way to prove this corollary is given in the proof of Th. 3.3.9.

Remark 3.3.5 If G is square and non–singular, then Cor. 3.3.4 can be applied to the hyperbolic singular value problem. Let σ_j be the j-th hyperbolic singular value of G_0, J and σ'_j the j-th computed hyperbolic singular value. Then, by taking square roots in (3.3.20) and ignoring relative errors of $O(\varepsilon)$, we obtain

$$1-\gamma-\varepsilon \leq \frac{\sigma'_j}{\sigma_j} \leq 1+\gamma+\varepsilon \ , \tag{3.3.22}$$

where

$$\gamma = \varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol/2 + r \cdot n\varepsilon/2 \ .$$

Extra ε in (3.3.22) comes from the fact that $\sigma'_j = fl(\sqrt{\lambda'_j})$.

This remark holds for all subsequent implicit methods in this chapter.

As we did in Subsect. 3.2.1, we can modify the implicit $J-$orthogonal Jacobi method in order to avoid potentially large C_m in Th. 3.4.2 in the hyperbolic case for $|\zeta| \leq \alpha$. The algorithm of the modified method is obtained by combining Algorithms 3.3.1 and 3.2.7 in the obvious manner. The comments from Subsect. 3.2.1 hold here, as well. We have the following:

Theorem 3.3.6 *Let G_m be the sequence of matrices generated by the modified implicit $J-$orthogonal Jacobi method in finite precision arithmetic with precision ε. Then Theorem 3.3.3 holds except that in the hyperbolic case for $|\zeta| \leq 3/(2\sqrt{2})$ the value C_m is changed to $C_m = 28$. Corollary 3.3.4 holds with this exception, too.*

Proof. The technique of proof is the same as in Th. 3.3.3. We assume without loss of generality that $\widetilde{sh} = sh = +1$. Also, $\widetilde{ch} = \sqrt{2}$, $ch = fl(\sqrt{2})$, so that

$$|\varepsilon_{ch}| \leq \varepsilon \ , \qquad \varepsilon_{sh} = 0 \ .$$

Therefore,

$$|\varepsilon'_1|, |\varepsilon'_4| \leq 3\varepsilon \ , \qquad |\varepsilon'_2|, |\varepsilon'_3| \leq 2\varepsilon \ ,$$

and the theorem follows by inserting these values and $1/x \leq \sqrt{2}$ into (3.3.16) and (3.3.17). Q.E.D.

3.3.1 Keeping the diagonal in a separate vector

The approximate operation count for the implicit $J-$orthogonal Jacobi method of Alg. 3.3.1 is the following: we need $3n$ multiplications and $3(n-1)$ additions to calculate a, b, and c, and $4n$ multiplications and $2n$ additions to update vectors $G_{.i}$, $G_{.j}$ per rotation. This gives the total of approximately $3.5n^3$ multiplications and $2.5n^3$ additions per cycle ($n(n-1)/2$ rotations). Keeping the diagonal elements of the matrix $G^T G$ in a separate vector makes the calculation of the parameters a and b via scalar product in each step unnecessary, which leaves the total of $2.5n^3$ multiplications and $1.5n^3$ additions per cycle.

The main idea (in the notation of Alg. 3.3.1) is the following: at the beginning of each cycle we calculate

$$\Delta_i = \sum_{k=1}^{n} G_{ki}^2 \ .$$

At the beginning of each step we set

$$a = \Delta_i \ , \qquad b = \Delta_j \ , \qquad c = \sum_{k=1}^{n} G_{ki} G_{kj} \ .$$

We update Δ_i and Δ_j by the formulae

$$\Delta_i = \Delta_i - c * t \ , \qquad \Delta_j = \Delta_j + c * t \ ,$$

in the trigonometric, and

$$\Delta_i = \Delta_i + c * t \ , \qquad \Delta_j = \Delta_j + c * t \ ,$$

in the hyperbolic case, respectively.

Due to subtractions in updating Δ_i's, they can become inaccurate, i.e. the relative error of Δ_i to $\|G_{\cdot i}\|^2$ can be larger then $O(\varepsilon)$. Suppose that $\Delta_i = \|G_{\cdot i}\|^2$. After one subtraction we have

$$\Delta_i' = \|G_{\cdot i}'\|^2 (1 + \epsilon) \ , \qquad |\epsilon| \leq \frac{\kappa^2(B_m) + 1}{2} \varepsilon \ ,$$

where the maximum is attained when z tends to its upper bound (3.3.4) and $a = b$. Therefore, the relative error of Δ_i can grow considerably, which can affect the convergence by making the rotation angles inaccurate. This is why the vector Δ should be updated at the beginning of each cycle from the columns of the current matrix G. We did not use the well known Rutishauser's delayed updates of the diagonal, since they do not guarantee high relative accuracy of the diagonal at the beginning of each cycle.

When the pair $G^T G, J$ is obtained from the pair GJG^T, I, then the probability that the convergence is actually spoiled is very low. This is due to a non-trivial diagonalizing effect of the above transition.

We now turn to the one–step error analysis of the method. In the notation of Th. 3.3.3 we have

$$\Delta_i = d_i^2 \ , \qquad \Delta_j = d_j^2 \ , \qquad x = \sqrt{\Delta_i / \Delta_j} \ .$$

If $x \geq \bar{x}$ in the trigonometric, and $x \geq \beta$ in the hyperbolic case, then Th. 3.3.3 holds irrespectively of the accuracy of Δ_i and Δ_j.

If $x < \bar{x}$ in the trigonometric, and $x < \beta$ in the hyperbolic case, then Th. 3.3.3 holds if the relations (3.3.10) and (3.3.19) are satisfied, respectively. This is always the case if

$$c^2 < \Delta_i \Delta_j \ .$$

If the above inequality does not hold, then we have to refresh Δ_i and Δ_j. Note that hyperbolic rotations cause no additional problems over trigonometric ones.

The following algorithm is only a slight modification of Alg. 3.3.1, so only the parts where the two algorithms differ are stated.

Algorithm 3.3.7 *Implicit $J-$orthogonal method for the pair G, J. The vector Δ contains diagonal elements of the matrix $G^T G$.*

repeat
  /* *at the beginning of each cycle refresh the vector*
    Δ *which contains diagonal of* $G^T G$ */
    for j=1 to r
      $\Delta_j = \sum_{k=1}^n G_{kj}^2$
    endfor
  for all pairs $i < j$
    /* *compute* $\begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv$ *the* (i, j) *submatrix of* $G^T G$ */
      $c = \sum_{k=1}^n G_{ki} * G_{kj}$
      if $c^2 < \Delta_i \Delta_j$ *then*
        $a = \Delta_i$
        $b = \Delta_j$
      else
        $a = \sum_{k=1}^n G_{ki}^2$
        $b = \sum_{k=1}^n G_{kj}^2$
      endif
    /* *compute the parameter hyp: hyp* $= 1$ *for the hyperbolic and*
      hyp $= -1$ *for the trigonometric rotation, respectively* */
    /* *compute the $J-$orthogonal Jacobi rotation which diagonalizes*
$$\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix} \quad */$$
    /* *update columns* i *and* j *of* G */
    /* *update* Δ_i *and* Δ_j */
      $\Delta_i = a + hyp * c * t$
      $\Delta_j = b + c * t$
  endfor
until convergence (all $|c|/\sqrt{ab} \leq tol$*)*
/* *the computed non–zero eigenvalues of* $H = GJG^T$ *(and of the pair* $G^T G, J$*) are*
    $\lambda_j = (\sum_{k=1}^n G_{kj}^2) J_{jj}$ */
/* *the computed eigenvectors of* H *are the normalized columns of the final* G */

Numerical experiments of Chap. 5 showed no difference in the accuracy between Alg. 3.3.7 and other implicit algorithms.

3.3.2 Error bounds for the eigenvectors

Theorems which give one–step error analysis of the implicit $J-$orthogonal Jacobi methods in Sections 3.3 and 3.4 imply that one step of any of those methods satisfies the eigenprojection perturbation bounds of Th. 2.3.3. As a consequence, the eigen (spectral) projections computed by any of those methods also satisfy those bounds. We prove the following theorem for the method defined by Alg. 3.3.1. The proof for other implicit methods is similar. In the proof of the theorem we use the following lemma due to Veselić [30]:

Lemma 3.3.8 *Let*

$$F^*F = I + E \ , \qquad \|E\|_2 = \epsilon < 1 \ ,$$

where F is any matrix with full column rank. Then there exists a matrix Q such that $Q^*Q = I$ and $\|F - Q\|_2 \leq \epsilon$.

PROOF. We make the polar decomposition $F = QP$ where $Q^*Q = I$ and P is Hermitian positive definite matrix. Since $QQ^*F = F$, we have $P^2 = I + E$, or

$$(P + I)(P - I) = E \ .$$

Thus

$$\|P - I\|_2 \leq \epsilon/(1 + \sqrt{1-\epsilon}) \leq \epsilon \ ,$$

so that

$$\|F - Q\|_2 = \|QP - Q\|_2 = \|P - I\|_2 \ ,$$

and the lemma is proved. Q.E.D.

Theorem 3.3.9 *Let G, J, where G is non–singular, be the starting pair for Alg. 3.3.1. Assume algorithm converges, and that G_M, J is the final pair which satisfies the stopping criterion. For $0 \leq m \leq M$ write $G_m = B_m D_m$, where D_m diagonal and B_m has unit columns. Let λ be an eigenvalue of the matrix GJG^T and let P be its eigenprojection. Let P' be the approximation of the corresponding spectral projection, i.e. P' is obtained from the final eigenvectors which are obtained by dividing the columns of G_M by their norms. Then, with the relative error of $O(\varepsilon)$,*

$$\|P' - P\|_2 \leq \frac{4\bar{\eta}}{rg_G(\lambda)} \frac{1}{1 - \dfrac{3\bar{\eta}}{rg_G(\lambda)}} + 2n \cdot tol + n(3n+4)\varepsilon \ , \qquad (3.3.23)$$

where $\bar\eta = \eta(\eta+2)$, *and*

$$\eta = \varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol + n^2\varepsilon \ ,$$

provided $3\bar\eta / rg_G(\lambda) < 1$. *Here* $rg_G(\lambda)$ *is defined by (2.3.1) and the quantities* C_m *are defined by Th. 3.3.3.*

PROOF. We first show that for every $1 \le m \le M$, the matrix G_m is obtained by the sequence of exact transformations on some perturbed matrix $G + \delta G^{(m-1)}$ in the sense of Th. 3.3.3, i.e.

$$G_m = (G + \delta G^{(m-1)}) R_0 \cdot \dots \cdot R_{m-1} \ , \tag{3.3.24}$$

where

$$\|\delta G^{(m-1)} x\|_2 \le \varepsilon \sum_{k=0}^{m-1} \frac{C_k}{\sigma_{min}(B_k)} \|Gx\|_2 \tag{3.3.25}$$

holds with the relative error of ε. The proof is by induction on m. For $m = 1$ the statement follows from Th. 3.3.3. Now suppose that (3.3.24) holds for some $m \ge 1$. By Th. 3.3.3 and the induction assumption we have

$$\begin{aligned} G_{m+1} &= (G_m + \delta G_m) R_m \\ &= [(G + \delta G^{(m-1)}) R_0 \cdot \dots \cdot R_{m-1} + \delta G_m] R_m \\ &= (G + \delta G^{(m)}) R_0 \cdot \dots \cdot R_m \ , \end{aligned}$$

where

$$\delta G^{(m)} = \delta G^{(m-1)} + \delta G_m (R_0 \cdot \dots \cdot R_{m-1})^{-1} \ .$$

Set $\delta G_m = \delta B_m D_m$. Then

$$\|\delta G^{(m)} x\|_2 \le \|\delta G^{(m-1)} x\|_2 + \frac{\|\delta B_m\|_2}{\sigma_{min}(B_m)} \|B_m D_m G_m^{-1} (G + \delta G^{(m-1)}) x\|_2 \ ,$$

and (3.3.24) follows from (3.3.25) and Th. 3.3.3, ignoring the relative errors of $O(\varepsilon)$.

Since the final pair satisfies the stopping criterion, we have

$$B_M^T B_M = I + E \ , \qquad \|E\|_2 \le n \cdot tol + n^2\varepsilon \ .$$

The $n^2\varepsilon$ term comes from the fact that $c/\sqrt{ab}$ in the stopping criterion may be underestimated by as much as $n\varepsilon$. Lemma 3.3.8 implies that there exists an orthogonal matrix

$$B'_M = B_M + \delta B_M \ ,$$

where

$$\|\delta B_M\|_2 \le n \cdot tol + n^2\varepsilon \ .$$

Set $G'_M = B'_M D_M$. As in the first part of the proof, we can show that

$$G'_M = (G + \delta G^{(M)}) R_0 \cdot \ldots \cdot R_{M-1} \ ,$$

where $\|\delta G^{(M)} x\|_2 \leq \eta \|Gx\|_2$. Since $\sigma_{min}(B_M) \geq 1 - (n \cdot tol + n^2 \varepsilon)$, we ignore the factor $1/\sigma_{min}(B_M)$ when applying Th. 3.3.3. Let P'_M denote the spectral projection of the matrix $G'_M J G'^T_M$ which corresponds to the eigenprojection P. Th. 2.3.3 now implies

$$\|P - P'_M\|_2 \leq \frac{4\bar{\eta}}{rg_G(\lambda)} \cdot \frac{1}{1 - 3\bar{\eta}/rg_G(\lambda)} \ . \tag{3.3.26}$$

The spectral projection P'_M is obtained from columns of the matrix B'_M, while the approximation P' is obtained from columns of the matrix

$$fl(G_M \cdot |\mathrm{diag}\,(\lambda'_j)|^{-1/2}) = B_M + F \ ,$$

where

$$|F_{ij}| \leq |B_{M,ij}|(n+4)\varepsilon/2 \ .$$

Here we used $|\lambda'_j|/D_{M,j} \leq 1 + n\varepsilon$ and ignored the relative error of $O(\varepsilon)$. Using $\|B_M\|_2 \leq 1 + n \cdot tol + n^2\varepsilon$, and ignoring again the relative error of $O(\varepsilon)$, we finally have

$$\begin{aligned} \|P'_M - P'\|_2 &\leq \|(B_M + \delta B_M)(B_M + \delta B_M)^T - (B_M + F)(B_M + F)^T\|_2 \\ &\leq 2\|\delta B_M\|_2 + 2\|F\|_2 \\ &\leq 2n \cdot tol + n(3n+4)\varepsilon \ , \end{aligned}$$

which, together with (3.3.26), implies (3.3.23). Q.E.D.

3.4 Fast implicit method

In this section we define and analyse the fast implicit $J-$orthogonal Jacobi method for the pair G, J. The remarks from Sect. 3.3 hold here as well. The section is also organized as Sect. 3.3. We first present the algorithm. We then give one–step error analysis and overall error bound for the eigenvalues. In Th. 3.4.4 we give one–step error analysis of the modified method. After that we shortly discuss the version of the algorithm where the diagonal of $G^T G$ is kept in a separate vector. In Subsect. 3.4.1 we consider fast self–scaling rotations used in order to avoid possible underflow/overflow when updating the scaling matrix.

The idea of fast rotations is to use transformation matrices of the form

$$J_m = \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} \ , \tag{3.4.1}$$

instead of matrices of the form

$$\begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} , \qquad \begin{bmatrix} ch & sh \\ sh & ch \end{bmatrix} .$$

This saves $2n$ multiplications in each step, or approximately n^3 multiplications in each cycle. The use of matrices of the type (3.4.1) is possible if the matrices G_m are stored in factorized form

$$G_m = \bar{G}_m \bar{D}_m \ ,$$

where $\bar{D}_m$ is diagonal positive definite.

In the $m-$th step of the implicit method only the columns i and j of the matrix G_m are changed. Let $G_m \equiv G$ and $G_{m+1} \equiv G'$. If we use the ordinary rotation, then we have

$$\begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} = \begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} ,$$

in the trigonometric, or

$$\begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} = \begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} \begin{bmatrix} ch & sh \\ sh & ch \end{bmatrix} ,$$

in the hyperbolic case. Now suppose that $G = \bar{G}\bar{D}$, i.e.

$$\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} = \begin{bmatrix} \bar{G}_{\cdot i} & \bar{G}_{\cdot j} \end{bmatrix} \begin{bmatrix} \bar{D}_i & \\ & \bar{D}_j \end{bmatrix} . \tag{3.4.2}$$

Simple calculation shows that

$$\begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} = \begin{bmatrix} \bar{G}'_{\cdot i} & \bar{G}'_{\cdot j} \end{bmatrix} \begin{bmatrix} \bar{D}'_i & \\ & \bar{D}'_j \end{bmatrix} ,$$

where

$$\begin{bmatrix} \bar{G}'_{\cdot i} & \bar{G}'_{\cdot j} \end{bmatrix} = \begin{bmatrix} \bar{G}_{\cdot i} & \bar{G}_{\cdot j} \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} .$$

Here

$$\begin{aligned} &\alpha = \frac{\bar{D}_i}{\bar{D}_j} t \ , \qquad && \beta = -\frac{\bar{D}_j}{\bar{D}_i} t \ , \qquad && t = sn/cs \ , \\ &\bar{D}'_i = \bar{D}_i cs \ , && \bar{D}'_j = \bar{D}_j cs \ , && \end{aligned} \tag{3.4.3}$$

in the trigonometric, and

$$\begin{aligned} &\alpha = \frac{\bar{D}_i}{\bar{D}_j} t \ , \qquad && \beta = \frac{\bar{D}_j}{\bar{D}_i} t \ , \qquad && t = sh/ch \ , \\ &\bar{D}'_i = \bar{D}_i ch \ , && \bar{D}'_j = \bar{D}_j ch \ , && \end{aligned} \tag{3.4.4}$$

in the hyperbolic case.

We now state the algorithm:

Algorithm 3.4.1 *Fast implicit $J-$orthogonal Jacobi method for the pair G, J. tol is a user defined stopping criterion.*

$$
\begin{array}{l}
\textit{for } k = 1 \textit{ to } r \\
\quad D_k = 1 \\
\textit{endfor} \\
\textit{repeat} \\
\quad \textit{for all pairs } i < j \\
\qquad /^* \textit{ compute } \begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv \textit{the } (i,j) \textit{ submatrix of } G^T G \ ^*/ \\
\qquad\quad a = D_i^2 \sum_{k=1}^n G_{ki}^2 \\
\qquad\quad b = D_j^2 \sum_{k=1}^n G_{kj}^2 \\
\qquad\quad c = D_i D_j \sum_{k=1}^n G_{ki} * G_{kj} \\
\qquad /^* \textit{ compute the parameter hyp: } hyp = 1 \textit{ for the hyperbolic and} \\
\qquad\quad hyp = -1 \textit{ for the trigonometric rotation, respectively } ^*/ \\
\qquad\quad \textit{if } 1 \le i \le npos < j \le r \textit{ then} \\
\qquad\qquad hyp = 1 \\
\qquad\quad \textit{else} \\
\qquad\qquad hyp = -1 \\
\qquad\quad \textit{endif} \\
\qquad /^* \textit{ compute the } J-\textit{orthogonal Jacobi rotation which diagonalizes} \\
\qquad\qquad \begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix} \ ^*/ \\
\qquad\quad \zeta = -hyp * (b + hyp * a)/(2c) \\
\qquad\quad t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2 - hyp}) \\
\qquad\quad cs = 1/\sqrt{1 - hyp * t^2} \\
\qquad\quad \alpha = t * D_i / D_j \\
\qquad\quad \beta = hyp * t * D_j / D_i \\
\qquad /^* \textit{ update columns } i \textit{ and } j \textit{ of } G \ ^*/ \\
\qquad\quad \textit{for } k = 1 \textit{ to } n \\
\qquad\qquad tmp = G_{ki} \\
\qquad\qquad G_{ki} = tmp + \alpha * G_{kj} \\
\qquad\qquad G_{kj} = \beta * tmp + G_{kj} \\
\qquad\quad \textit{endfor} \\
\qquad /^* \textit{ update } D_i \textit{ and } D_j \ ^*/ \\
\qquad\quad D_i = D_i * cs \\
\qquad\quad D_j = D_j * cs \\
\quad \textit{endfor} \\
\textit{until convergence (all } |c|/\sqrt{ab} \le tol) \\
/^* \textit{ the computed non–zero eigenvalues of } H = GJG^T \textit{ (and of the pair } G^T G, J) \textit{ are} \\
\qquad \lambda_j = (\sum_{k=1}^n G_{kj}^2) D_j^2 J_{jj} \ ^*/ \\
/^* \textit{ the computed eigenvectors of } H \textit{ are the normalized columns of the final } G \ ^*/
\end{array}
$$

The following theorem and its corollary justify our accuracy claims for the eigenvalues of the matrix $H = GJG^T$ computed by the fast implicit $J-$orthogonal Jacobi method .

Theorem 3.4.2 *Let $\bar{G}_m$, $\bar{D}_m$ be the sequences of matrices generated by the fast implicit $J-$orthogonal Jacobi algorithm in floating–point arithmetic with precision ε; that is $\bar{G}_{m+1}$ is obtained from $\bar{G}_m$ by applying a single fast rotation, and D_{m+1} is obtained from D_m according to (3.4.3) or (3.4.4). Let $G_m \equiv \bar{G}_m \cdot \bar{D}_m$. Since G_m is needed only for theoretical consideration, we suppose that this matrix multiplication is exact. Then the following diagram commutes.*

$$\begin{tikzcd}
\bar{G}_m \cdot D_m \equiv G_m \arrow[rr, "\text{fast floating}", "\text{rotation}"'] \arrow[d] & & \bar{G}_{m+1} \cdot \bar{D}_{m+1} \equiv G_{m+1} \\
G_m + \delta G_m \arrow[urr, "\text{exact}", "\text{rotation}"'] & &
\end{tikzcd}$$

The top arrow indicates that G_{m+1} is obtained from G_m by applying one fast rotation in floating–point arithmetic. The diagonal arrow indicates that G_{m+1} is obtained from $G_m + \delta G_m$ by applying one $J-$orthogonal plane rotation in exact arithmetic; thus $G_{m+1}JG_{m+1}$ and $(G_m + \delta G_m)J(G_m + \delta G_m)^T$ have identical eigenvalues. δG_m is bounded as follows. Let $\kappa = \kappa^2(B_m)$ and write $\delta G_m = \delta B_m D_m$, where D_m is diagonal such that B_m in $G_m = B_m D_m$ has unit columns. Let a_T and b_T be the true values of $\sum_k G_{ki}^2$ and $\sum_k G_{kj}^2$, respectively. Then, with the relative error of order ε,

$$\|\delta B_m\|_2 \leq C_m \cdot \varepsilon \ , \tag{3.4.5}$$

where

$$C_m = \begin{cases} 33 & \textit{in trigonometric case} \ , \\ \kappa + 16\sqrt{\kappa} + 39 & \textit{in hyperbolic case} \ , \quad |\zeta| \leq \dfrac{3}{2\sqrt{2}} \ , \\ 102 & \textit{in hyperbolic case} \ , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \ , \\ & \qquad\qquad\qquad\qquad\quad b_T \geq \dfrac{1}{2}a_T \ , \\ 125 & \textit{in hyperbolic case} \ , \quad |\zeta| > \dfrac{3}{2\sqrt{2}} \ , \\ & \qquad\qquad\qquad\qquad\quad b_T < \dfrac{1}{2}a_T \ . \end{cases}$$

In other words, one step of the fast implicit $J-$orthogonal Jacobi method satisfies the assumptions needed for the perturbation bounds of Sect. 2.3.

Proof. The proof of the commuting diagram is a tedious computation. We shall prove the diagram separately for the trigonometric and for the hyperbolic case. Let

a_T, b_T and c_T be the true values of $\bar{D}_i^2 \sum_k \bar{G}_{ki}^2$, $\bar{D}_j^2 \sum_k \bar{G}_{kj}^2$ and $\bar{D}_i \bar{D}_j \sum_k \bar{G}_{ki}\bar{G}_{kj}$. We may assume without loss of generality that $a_T \geq b_T$ and $c_T > 0$. As in (3.2.3), we have

$$a_T = d_i^2 \ , \qquad b_T = d_j^2 \ , \qquad c_T = z d_i d_j \ .$$

As in (3.2.4), we can show that (3.3.4) holds. Also let $x \equiv d_j/d_i \leq 1$. Systematic application of formulae (3.2.1) shows that

$$\begin{aligned} a = a_T(1+\varepsilon_a) &\quad \text{where} \quad |\varepsilon_a| \leq (n+2)\varepsilon \\ b = b_T(1+\varepsilon_b) &\quad \text{where} \quad |\varepsilon_b| \leq (n+2)\varepsilon \\ c = c_T + \varepsilon_c\sqrt{a_T b_T} &\quad \text{where} \quad |\varepsilon_c| \leq (n+2)\varepsilon \end{aligned}$$

Trigonometric case. This case was analysed by Anda and Park [1] for the Givens rotation in the QR–algorithm. Our proof is similar to theirs.

Let

$$\begin{aligned} \widetilde{cs} &\equiv 1/\sqrt{1+t^2} \ , \qquad \widetilde{sn} \equiv t/\sqrt{1+t^2} \\ \tilde{\alpha} &\equiv t\bar{D}_i/\bar{D}_j \ , \qquad \tilde{\beta} \equiv -t\bar{D}_j/\bar{D}_i \ . \end{aligned} \tag{3.4.6}$$

For the calculated transformation parameters we have

$$\begin{aligned} cs &= (1+\varepsilon_{cs})\widetilde{cs} \ , \qquad |\varepsilon_{cs}| \leq 3\varepsilon \ , \\ \alpha &= (1+\varepsilon_\alpha)\tilde{\alpha} \ , \qquad \beta = (1+\varepsilon_\beta)\tilde{\beta} \ , \qquad |\varepsilon_\alpha|, |\varepsilon_\beta| \leq 2\varepsilon \ . \end{aligned}$$

$\widetilde{cs}$ and $\widetilde{sn}$ define the exact rotation

$$J_m = \begin{bmatrix} \widetilde{cs} & \widetilde{sn} \\ -\widetilde{sn} & \widetilde{cs} \end{bmatrix}$$

which takes $G_m + \delta G_m$ to G_{m+1}:

$$(G_m + \delta G_m) J_m = G_{m+1} \ .$$

Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. We have

$$\begin{aligned} \bar{G}'_{ki} &= fl(\bar{G}_{ki} + \beta\bar{G}_{kj}) = (1+\varepsilon_1)\bar{G}_{ki} + (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_\beta)\tilde{\beta}\bar{G}_{kj} \\ &= \bar{G}_{ki} + \tilde{\beta}\bar{G}_{kj} + \varepsilon_1\bar{G}_{ki} + (\varepsilon_2+\varepsilon_3+\varepsilon_\beta)\tilde{\beta}\bar{G}_{kj} \\ \bar{D}'_i &= fl(\bar{D}_i\,\widetilde{cs}) = \bar{D}_i\widetilde{cs} + (\varepsilon_4+\varepsilon_{cs})\bar{D}_i\widetilde{cs} \ . \end{aligned}$$

Using

$$G_{\cdot i} = \bar{G}_{\cdot i}\,\bar{D}_i \ , \qquad G'_{\cdot i} = \bar{G}'_{\cdot i}\,\bar{D}'_i \ ,$$

and (3.4.6), and ignoring the relative error of $O(\varepsilon)$, we obtain

$$G'_{\cdot i} = \widetilde{cs}G_{\cdot i} - \widetilde{sn}G_{\cdot j} + E_{\cdot i}$$

where

$$\|E_{\cdot i}\|_2 \le (5\widetilde{cs}\|G_{\cdot i}\|_2 + 8|\widetilde{sn}|\|G_{\cdot j}\|_2)\varepsilon \ .$$

Here $G_{\cdot i}$ refers to the i-th column of G, etc. Similarly,

$$G'_{\cdot j} = \widetilde{sn}G_{\cdot i} + \widetilde{cs}G_{\cdot j} + E_{\cdot j}$$

where

$$\|E_{\cdot j}\|_2 \le (8|\widetilde{sn}|\|G_{\cdot i}\|_2 + 5\widetilde{cs}\|G_{\cdot j}\|_2)\varepsilon \ .$$

Now (3.3.7) holds with

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\le \widetilde{cs}\|E_{\cdot i}\|_2 + |\widetilde{sn}|\|E_{\cdot j}\|_2 \\
&\le \frac{1}{1+t^2}(5 + 8t^2 + 13|t|x)d_i\varepsilon \qquad (3.4.7) \\
\|F_{\cdot j}\|_2 &\le |\widetilde{sn}|\,\|E_{\cdot i}\|_2 + \widetilde{cs}\|E_{\cdot j}\|_2 \\
&\le \frac{1}{1+t^2}(5 + 8t^2 + 13|t|/x)d_j\varepsilon \ . \qquad (3.4.8)
\end{aligned}$$

We consider two cases, $x < \bar{x} \equiv 0.51$, and $x \ge \bar{x}$. First consider $x < \bar{x}$. Inserting $\bar{x}$ for x in (3.4.7) we obtain

$$\|F_{\cdot i}\|_2 \le 9.82 d_i\varepsilon \ .$$

Inserting (3.3.10) into (3.4.8) we obtain

$$\|F_{\cdot j}\|_2 \le 22.57 d_j\varepsilon \ ,$$

and

$$\|B_m\|_2 \le 33\varepsilon \ . \qquad (3.4.9)$$

Now consider the case $x \ge \bar{x}$. Inserting 1 for x in (3.4.7) we obtain

$$\|F_{\cdot i}\|_2 \le 13 d_i\varepsilon \ .$$

Inserting $1/\bar{x}$ for $1/x$ in (3.4.8), we obtain

$$\|F_{\cdot j}\|_2 \le 19.25 d_j\varepsilon \ ,$$

so that (3.4.9) holds again.

Hyperbolic case. The proof is a combination of the above proof for the trigonometric case and the proof for the hyperbolic case of Th. 3.3.3. We denote the quantities cs, sn and $sn1 = sn$ computed by Alg. 3.4.1 with ch and sh, respectively. Let

$$\begin{aligned}
\widetilde{ch} &\equiv 1/\sqrt{1-t^2}\ , \qquad \widetilde{sh} \equiv t/\sqrt{1-t^2} \\
\tilde{\alpha} &\equiv t\bar{D}_i/\bar{D}_j\ , \qquad \tilde{\beta} \equiv t\bar{D}_j/D_i\ . \qquad (3.4.10)
\end{aligned}$$

For the calculated transformation parameters we have

$$\begin{aligned} ch &= (1+\varepsilon_{ch})\widetilde{ch} \ , \\ \alpha &= (1+\varepsilon_\alpha)\tilde{\alpha} \ , \qquad \beta = (1+\varepsilon_\beta)\tilde{\beta} \ , \qquad |\varepsilon_\alpha|, |\varepsilon_\beta| \le 2\varepsilon \ . \end{aligned}$$

$\widetilde{ch}$ and $\widetilde{sh}$ define the exact rotation

$$J_m = \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix}$$

which takes $G_m + \delta G_m$ to G_{m+1}:

$$(G_m + \delta G_m) J_m = G_{m+1} \ .$$

Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. As in the proof for the trigonometric case, we obtain

$$G'_{\cdot i} = \widetilde{ch} G_{\cdot i} + \widetilde{sh} G_{\cdot j} + E_{\cdot i}$$

where

$$\|E_{\cdot i}\|_2 \le (2\varepsilon + |\varepsilon_{ch}|)\widetilde{ch}\|G_{\cdot i}\|_2 + (5\varepsilon + |\varepsilon_{ch}|)|\widetilde{sh}|\|G_{\cdot j}\|_2 \ ,$$

and

$$G'_{\cdot j} = \widetilde{sh} G_{\cdot i} + \widetilde{ch} G_{\cdot j} + E_{\cdot j}$$

where

$$\|E_{\cdot j}\|_2 \le (5\varepsilon + |\varepsilon_{ch}|)|\widetilde{sh}|\|G_{\cdot i}\|_2 + (2\varepsilon + |\varepsilon_{ch}|)\widetilde{ch}\|G_{\cdot j}\|_2 \ .$$

Now (3.3.15) holds with

$$\begin{aligned} \|F_{\cdot i}\|_2 &\le ((2\varepsilon + |\varepsilon_{ch}|)\widetilde{ch}^2 + (5\varepsilon + |\varepsilon_{ch}|)\widetilde{sh}^2 + (7\varepsilon + 2|\varepsilon_{ch}|)\widetilde{ch}|\widetilde{sh}|x)d_i \\ \|F_{\cdot j}\|_2 &\le ((2\varepsilon + |\varepsilon_{ch}|)\widetilde{ch}^2 + (5\varepsilon + |\varepsilon_{ch}|)\widetilde{sh}^2 + (7\varepsilon + 2|\varepsilon_{ch}|)\widetilde{ch}|\widetilde{sh}|/x)d_j \end{aligned} \quad (3.4.11)$$

As in Th. 3.3.3 we consider two cases, $|\zeta| \le 3/(2\sqrt{2})$ and $|\zeta| > 3/(2\sqrt{2})$. First consider $|\zeta| \le 3/(2\sqrt{2})$. Then (3.2.18) and (3.2.20) hold, and

$$|\varepsilon_{ch}| \le \left(\frac{3}{8}\sqrt{\kappa} + \frac{7}{3}\right)\varepsilon \ .$$

The assertion of the theorem now follows by inserting $1/x \le \sqrt{2}$, (3.2.20), and the above relation into (3.4.11).

Now consider $|\zeta| > 3/(2\sqrt{2})$. Then the relations (3.2.23) hold for t, $\widetilde{ch}$ and $\widetilde{sh}$ with a relative error of $O(\varepsilon)$, and

$$|\varepsilon_{ch}| \le 4\varepsilon \ . \quad (3.4.12)$$

We have two subcases, $x \ge 1/\sqrt{2}$ and $x < 1/\sqrt{2}$. If $x \ge 1/\sqrt{2}$, then the assertion of the theorem follows by inserting $1/x \le \sqrt{2}$, (3.4.12), and (3.2.23) into (3.4.11).

If $x < 1/\sqrt{2}$, then (3.3.19) holds, and the assertion of the theorem follows by inserting (3.3.19), (3.4.12), and (3.2.23) into (3.4.11). Q.E.D.

Corollary 3.4.3 *Assume Algorithm 3.4.1 converges, and that $G_M, J \equiv \bar{D}_M \bar{G}_M, J$ is the final pair which satisfies the stopping criterion. For $0 \leq m \leq M$ write $G_m = B_m D_m$ with D_m diagonal and B_m with unit columns.*

Let λ_j be the j-th non-zero eigenvalue of $G_0 J G_0^T$, and let λ_j' be the j-th computed eigenvalue. Then, with the relative error of $O(\varepsilon)$,

$$(1-\gamma)^2 \leq \frac{\lambda_j'}{\lambda_j} \leq (1+\gamma)^2 , \tag{3.4.13}$$

where

$$\gamma = \varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol/2 + r(n+2)\varepsilon/2 .$$

PROOF. See the proof of Cor. 3.3.4. The $r(n+2)\varepsilon/2$ term comes from the facts that $c/\sqrt{ab}$ in the stopping criterion may now be underestimated by as much as $(n+2)\varepsilon$, and that the squares of the norms of the columns of G_M are computed with a relative error not greater than $(n+2)\varepsilon$. Q.E.D.

As in Subsect. 3.2.1, we can modify the fast implicit $J-$orthogonal Jacobi method in order to avoid potentially large C_m in Th. 3.4.2 in the hyperbolic case for $|\zeta| \leq 3/(2\sqrt{2})$. The algorithm of the modified method is obtained by combining Algorithms 3.4.1 and 3.2.7 in the obvious manner. We have the following:

Theorem 3.4.4 *Let G_m be the sequence of matrices generated by the modified fast implicit $J-$orthogonal Jacobi method in floating–point arithmetic with precision ε. Then Theorem 3.4.2 holds except that in the hyperbolic case for $|\zeta| \leq 3/(2\sqrt{2})$ the value C_m is changed to $C_m = 55$. Corollary 3.4.3 holds with this exception, too.*

PROOF. See the proof of Th. 3.3.6. Q.E.D.

As in Subsect. 3.3.1, we can keep the diagonal of the matrix $G^T G$ in a separate vector, thus saving $2(n+1)$ multiplications and $2(n-1)$ additions in every step. This is done as in Alg. 3.3.7, except that Δ_i's are now refreshed using $\bar{G}_m$ and $\bar{D}_m$. All remarks about Alg. 3.3.7 from Subsect. 3.3.1 hold here, as well.

3.4.1 Self–scaling rotations

Analysing the fast rotation formulae (3.4.3) and (3.4.4), we see that these rotations make both values $\bar{D}_i$ and $\bar{D}_j$ smaller or larger, respectively. This can lead to underflow/overflow in some $\bar{D}_i$ during floating–point computation. As already mentioned, the probability that this happens is in the case of transition from the matrix $H = GJG^T$ to the pair G, J very low. The probability of underflow/overflow can further be reduced by using self–scaling rotations suggested by Anda and Park [1].

The main idea is to "push" the diagonal element of $\bar{D}$ which is further away from 1 towards 1. We use the "two way branch algorithm" of [1] and generalize it to the hyperbolic case. This adds four new fast rotations to the already existing ones (3.4.3) and (3.4.4). In this subsection we define these rotations, give the algorithm of the method, and present the error analysis.

The trigonometric self–scaling rotations from [1] are the following: suppose that (3.4.2) holds. Simple calculation shows that either

$$\begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} = \begin{bmatrix} \bar{G}_{\cdot i} & \bar{G}_{\cdot j} \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \bar{D}'_i & \\ & \bar{D}'_j \end{bmatrix} , \tag{3.4.14}$$

where

$$\begin{aligned} &\alpha = \frac{\bar{D}_i}{\bar{D}_j} t , \qquad \beta = -\frac{\bar{D}_j}{\bar{D}_i} cs \cdot sn , \qquad t = sn/cs , \\ &\bar{D}'_i = \bar{D}_i/cs , \qquad \bar{D}'_j = \bar{D}_j cs , \end{aligned} \tag{3.4.15}$$

or

$$\begin{bmatrix} G'_{\cdot i} & G'_{\cdot j} \end{bmatrix} = \begin{bmatrix} \bar{G}_{\cdot i} & \bar{G}_{\cdot j} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \bar{D}'_i & \\ & \bar{D}'_j \end{bmatrix} , \tag{3.4.16}$$

where

$$\begin{aligned} &\beta = -\frac{\bar{D}_j}{\bar{D}_i} t , \qquad \alpha = \frac{\bar{D}_i}{\bar{D}_j} cs \cdot sn , \qquad t = sn/cs , \\ &\bar{D}'_i = \bar{D}_i cs , \qquad \bar{D}'_j = \bar{D}_j/cs . \end{aligned} \tag{3.4.17}$$

The hyperbolic versions of the above rotations are either (3.4.14) with

$$\begin{aligned} &\alpha = \frac{\bar{D}_i}{\bar{D}_j} t , \qquad \beta = \frac{\bar{D}_j}{\bar{D}_i} ch \cdot sh , \qquad t = sh/ch , \\ &\bar{D}'_i = \bar{D}_i/ch , \qquad \bar{D}'_j = \bar{D}_j ch , \end{aligned} \tag{3.4.18}$$

or (3.4.16) with

$$\begin{aligned} &\beta = \frac{\bar{D}_j}{\bar{D}_i} t , \qquad \alpha = \frac{\bar{D}_i}{\bar{D}_j} ch \cdot sh , \qquad t = sh/ch , \\ &\bar{D}'_i = \bar{D}_i ch , \qquad \bar{D}'_j = \bar{D}_j/ch . \end{aligned} \tag{3.4.19}$$

The rotation (3.4.3) makes both $\bar{D}_i$ and $\bar{D}_j$ smaller. We use it in the trigonometric case when $\bar{D}_i, \bar{D}_j \geq 1$. The rotation (3.4.4) makes both $\bar{D}_i$ and $\bar{D}_j$ larger. We use it in the hyperbolic case when $\bar{D}_i, \bar{D}_j < 1$.

The rotations (3.4.14), (3.4.15) and (3.4.16), (3.4.19) make $\bar{D}_i$ larger and $\bar{D}_j$ smaller so they are always used when $\bar{D}_i < 1 \leq \bar{D}_j$. The first is also used in the trigonometric case when $\bar{D}_i \le \bar{D}_j < 1$ and the second is used in the hyperbolic case when $1 \le \bar{D}_i \le \bar{D}_j$.

The rotations (3.4.16), (3.4.17) and (3.4.14), (3.4.18) make $\bar{D}_i$ smaller and $\bar{D}_j$ larger so they are always used when $\bar{D}_i \ge 1 > \bar{D}_j$. The first is also used in the trigonometric case when $1 > \bar{D}_i > \bar{D}_j$, and the second is used in the hyperbolic case when $\bar{D}_i > \bar{D}_j \ge 1$.

Thus, we have the following

Algorithm 3.4.5 *Fast implicit $J-$orthogonal Jacobi method with self–scaling rotations for the pair G, J. tol is a user defined stopping criterion.*

for $k = 1$ *to* r
    $D_k = 1$
endfor
repeat
    for all pairs $i < j$
        /* *compute* $\begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv$ *the* (i,j) *submatrix of* G^TG */
            $a = D_i^2 \sum_{k=1}^n G_{ki}^2$
            $b = D_j^2 \sum_{k=1}^n G_{kj}^2$
            $c = D_i D_j \sum_{k=1}^n G_{ki} * G_{kj}$
        /* *compute the parameter hyp:* $hyp = 1$ *for the hyperbolic and*
            $hyp = -1$ *for the trigonometric rotation, respectively* */
            if $1 \le i \le npos < j \le r$ *then*
                $hyp = 1$
            else
                $hyp = -1$
            endif
        /* *compute the $J-$orthogonal Jacobi rotation which diagonalizes*
                $\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix}$ */
            $\zeta = -hyp * (b + hyp * a)/(2c)$
            $t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2 - hyp})$
            $h = \sqrt{1 - hyp * t^2}$
            $cs = 1/h$
            $sn = t/h$
        /* *update columns* i *and* j *of* G *and* D_i *and* D_j */
            if ($hyp = 1$ *and* $D_i, D_j < 1$) *or* ($hyp = -1$ *and* $D_i, D_j \ge 1$) *then*
                $\alpha = t * D_i/D_j$
                $\beta = hyp * t * D_j/D_i$
                for $k = 1$ *to* n
                    $tmp = G_{ki}$
                    $G_{ki} = tmp + \alpha * G_{kj}$

```
            G_kj = β * tmp + G_kj
        endfor
        D_i = D_i * cs
        D_j = D_j * cs
    elseif (hyp = 1 and (D_j < 1 ≤ D_i or D_i > D_j ≥ 1)) or
        (hyp = −1 and (D_i < 1 ≤ D_j or D_i ≤ D_j ≤ 1)) then
        α = t * D_i/D_j
        β = hyp * cs * sn * D_j/D_i
        for k = 1 to n
            G_kj = α * G_ki + G_kj
            G_ki = G_ki + β * G_kj
        endfor
        D_i = D_i/cs
        D_j = D_j * cs
    else
        β = hyp * t * D_j/D_i
        α = cs * sn * D_i/D_j
        for k = 1 to n
            G_ki = G_ki + β * G_kj
            G_kj = α * G_ki + G_kj
        endfor
        D_i = D_i * cs
        D_j = D_j/cs
    endif
  endfor
until convergence (all |c|/√(ab) ≤ tol)
/* the computed non–zero eigenvalues of H = GJG^T (and of the pair G^T G, J) are
      λ_j = (Σ_{k=1}^n G_kj^2) D_j^2 J_jj */
/* the computed eigenvectors of H are the normalized columns of the final G */
```

The version of the algorithm where the diagonal of G^TG is kept in a separate vector is obtained by combining Algorithms 3.4.5 and 3.3.7. The only exception from Alg. 3.3.7 is that Δ_i's are refreshed using $\bar{G}_m$ and $\bar{D}_m$. Further, the modified method is obtained by combining Algorithms 3.4.5 and 3.2.7. Error analysis of the self–scaling rotations is similar to the analysis of the fast rotations from previous section. The following theorem gives error analysis of the modified method:

Theorem 3.4.6 *Let $\bar{G}_m$, $\bar{D}_m$ be the sequences of matrices generated by the modified fast implicit $J-$orthogonal Jacobi algorithm with self–scaling rotations in floating–point arithmetic with precision ε; that is $\bar{G}_{m+1}$ is obtained from $\bar{G}_m$ by applying one of the fast rotations, and D_{m+1} is obtained from D_m by one of the formulae (3.4.3), (3.4.4), (3.4.15), (3.4.17 – 3.4.19). Then Th. 3.4.2 holds with*

$$C_m = 191 \tag{3.4.20}$$

in all cases. Corollary 3.4.3 holds as well.

PROOF. For the standard fast rotations (3.4.3) and (3.4.4), the theorem follows from Theorems 3.4.2 and 3.4.4.

Suppose that we apply the hyperbolic self–scaling rotation defined with (3.4.16) and (3.4.19). Let the quantities cs and sn computed by Alg. 3.4.5 be denoted by ch and sh, respectively. Let

$$\begin{aligned} \widetilde{ch} &\equiv 1/\sqrt{1-t^2}\ , & \widetilde{sh} \equiv t/\sqrt{1-t^2} \\ \widetilde{\beta} &\equiv t\bar{D}_j/\bar{D}_i\ , & \widetilde{\alpha} \equiv \bar{D}_i/\bar{D}_j\widetilde{ch}\cdot\widetilde{sh}\ . \end{aligned} \tag{3.4.21}$$

Since we are using the modified method, the relations (3.2.23) always hold, and for the calculated transformation parameters we have

$$\begin{aligned} ch &= (1+\varepsilon_{ch})\widetilde{ch}\ , \quad sh = (1+\varepsilon_{sh})\widetilde{sh}\ , \qquad |\varepsilon_{ch}|, |\varepsilon_{sh}| \le 4\varepsilon \\ \beta &= (1+\varepsilon_\beta)\widetilde{\beta}\ , \qquad |\varepsilon_\beta| \le 2\varepsilon\ . \\ \alpha &= (1+\varepsilon_\alpha)\widetilde{\alpha}\ , \qquad |\varepsilon_\alpha| \le 11\varepsilon\ , \end{aligned}$$

$\widetilde{ch}$ and $\widetilde{sh}$ define the exact rotation

$$J_m = \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix}$$

which takes $G_m + \delta G_m$ to G_{m+1}, i.e. $(G_m+\delta G_m)J_m = G_{m+1}$. Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. From Alg. 3.4.5 we have

$$\begin{aligned} \bar{G}'_{ki} &= fl(\bar{G}_{ki}+\beta\bar{G}_{kj}) = (1+\varepsilon_1)\bar{G}_{ki} + (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_\beta)\widetilde{\beta}\bar{G}_{kj} \\ &= \bar{G}_{ki} + \widetilde{\beta}\bar{G}_{kj} + \varepsilon_1\bar{G}_{ki} + (\varepsilon_2+\varepsilon_3+\varepsilon_\beta)\widetilde{\beta}\bar{G}_{kj} \\ \bar{D}'_i &= fl(\bar{D}_j\, ch) = \bar{D}_i\widetilde{ch} + (\varepsilon_4+\varepsilon_{ch})\bar{D}_i\widetilde{ch}\ . \end{aligned}$$

Using

$$G_{\cdot j} = \bar{G}_{\cdot j}\,\bar{D}_j\ , \qquad G'_{\cdot j} = \bar{G}'_{\cdot j}\,\bar{D}'_j\ ,$$

and (3.4.21), we obtain

$$G'_{\cdot i} = \widetilde{ch}G_{\cdot i} + \widetilde{sh}G_{\cdot j} + E_{\cdot i}$$

where

$$\|E_{\cdot i}\|_2 \le (6\widetilde{ch}\|G_{\cdot i}\|_2 + 9|\widetilde{sh}|\|G_{\cdot j}\|_2)\varepsilon\ .$$

Further,

$$\begin{aligned} \bar{G}'_{kj} &= fl(\alpha\bar{G}'_{ki}+\bar{G}_{kj}) = (1+\varepsilon_5)(1+\varepsilon_6)(1+\varepsilon_\alpha)\widetilde{\alpha}\bar{G}'_{ki} + (1+\varepsilon_7)\bar{G}_{kj} \\ \bar{D}'_j &= fl(\bar{D}_j/ch) = \bar{D}_j/ch + (\varepsilon_8+\varepsilon'_{ch})\bar{D}_j/ch\ , \end{aligned}$$

where $|\varepsilon'_{ch}| \leq 4\varepsilon$, so that

$$G'_{\cdot j} = \bar{G}'_{\cdot j}\,\bar{D}'_j = \widetilde{sh}G_{\cdot i} + \widetilde{ch}G_{\cdot j} + E_{\cdot j}$$

where

$$\|E_{\cdot j}\|_2 \leq \left(19|\widetilde{sh}|\|G_{\cdot i}\|_2 + \left(5\widetilde{ch} + \frac{1}{\widetilde{ch}} + 17\frac{\widetilde{sh}^2}{\widetilde{ch}}\right)\right)\varepsilon\ .$$

Now (3.3.15) holds with

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\leq \widetilde{ch}\|E_{\cdot i}\|_2 + |\widetilde{sh}|\|E_{\cdot j}\|_2 \\
&\leq (6\widetilde{ch}^2 + 19\widetilde{sh}^2 + (14\widetilde{ch}|\widetilde{sh}| + 17|t|\widetilde{sh}^2 + |t|)x)\varepsilon d_i\ , \\
\|F_{\cdot j}\|_2 &\leq |\widetilde{sh}|\cdot\|E_{\cdot i}\|_2 + \widetilde{ch}\|E_{\cdot j}\|_2 \\
&\leq (25\widetilde{ch}|\widetilde{sh}|/x + 26\widetilde{sh}^2 + 5\widetilde{ch}^2 + 1)\varepsilon d_j\ . \qquad (3.4.22)
\end{aligned}$$

Here

$$\|G_{\cdot i}\|_2 \equiv d_i\ , \qquad \|G_{\cdot j}\|_2 \equiv d_j\ , \qquad , \qquad x \equiv d_j/d_i\ .$$

We consider two cases, $x \geq 1/\sqrt{2}$ and $x < 1/\sqrt{2}$. If $x \geq 1/\sqrt{2}$, then by inserting $1/x \leq \sqrt{2}$ and (3.2.23) into (3.4.22) we have

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\leq 64 d_i\varepsilon\ , \qquad \|F_{\cdot j}\|_2 \leq 86 d_j\varepsilon\ , \\
\|\delta B_m\|_2 &\leq \|F_{\cdot i}\|_2/d_i + \|F_{\cdot j}\|_2/d_j \leq 150\varepsilon\ .
\end{aligned}$$

If $x < 1/\sqrt{2}$, then (3.3.19) holds, and by inserting (3.3.19), and (3.2.23) into (3.4.22) we have

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\leq 54 d_i\varepsilon\ , \qquad \|F_{\cdot j}\|_2 \leq 137 d_j\varepsilon\ , \\
\|\delta B_m\|_2 &\leq 191\varepsilon\ .
\end{aligned}$$

The analysis of the three remaining types of the self–scaling rotations is similar. Q.E.D.

Chapter 4

Symmetric indefinite decomposition

4.1 Introduction and algorithm

In order to solve the eigenvalue problem

$$Hx = \lambda x \ , \qquad x \neq 0 \ , \tag{4.1.1}$$

where H is a $n \times n$ real symmetric matrix with $\mathrm{rank}\,(H) = r \leq n$, by any of the implicit (one–sided) Jacobi methods of Chap. 3 for which we have good error bounds, we first decompose H as

$$PHP^T = GJG^T \ , \qquad J = I_{npos} \oplus -I_{r-npos} \ . \tag{4.1.2}$$

Here G is a $n \times r$ matrix (i.e. G has full column rank), P is a permutation matrix, and $npos$ is the number of positive eigenvalues of H. The decomposition (1.1) is then obtained by multiplying (4.1.2) by P^T from the left and P from the right, that is, the implicit Jacobi is applied to the pair $P^T G, J$.

The chapter is organized as follows: in this section we give the algorithm of the symmetric indefinite decomposition (4.1.2). In Sect. 4.2 we give the error analysis of the method. In Sect. 4.3 we give the final error bounds for the computed eigensolution of the symmetric eigenvalue problem. Finally, in Sect. 4.4 we show an interesting fact that the scaled condition of the matrix $G^T G$ is bounded by a function of n irrespectively of the condition of the starting matrix H.

We now give the algorithm of the symmetric indefinite decomposition (4.1.2). Our method is essentially the method of Bunch and Parlett [6] with some modifications. The method of Bunch and Parlett decomposes H as

$$PHP^T = LTL^T, \tag{4.1.3}$$

where L is lower triangular matrix with unit diagonal, and T is block diagonal matrix with (1×1) and (2×2) blocks. We shortly describe one step of the algorithm. Let $\bar{P}$ be a permutation matrix such that

$$\bar{P}H\bar{P}^T = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix} , \tag{4.1.4}$$

where X is nonsingular $k \times k$ matrix, $k \in \{1, 2\}$, C is a $(n-k) \times k$ matrix, and Y is a $(n-k) \times (n-k)$ matrix. Such $\bar{P}$ always exists because H is nonsingular. We can decompose $\bar{P}H\bar{P}^T$ as

$$\begin{aligned} \bar{P}H\bar{P}^T &= \bar{L} \begin{bmatrix} X & 0 \\ 0 & H_1 \end{bmatrix} \bar{L}^T, \\ \bar{L} &= \begin{bmatrix} I_k & 0 \\ CX^{-1} & I_{n-k} \end{bmatrix} , \\ H_1 &= Y - CX^{-1}C^T . \end{aligned} \tag{4.1.5}$$

Recursive application of (4.1.5) yields (4.1.3) in the obvious manner. We choose 1×1 or 2×2 pivot according to the unequilibrated diagonal pivoting from [6] [1]: set

$$\alpha = (1 + \sqrt{17})/8 ,$$

and calculate

$$\nu_0 = \max_{i \neq j} |H_{ij}| , \qquad \nu_1 = \max_i |H_{ii}| . \tag{4.1.6}$$

We choose a 1×1 pivot if and only if $\nu_1 \geq \alpha\nu_0$, and a 2×2 pivot otherwise. For a 1×1 pivot, we choose $\bar{P}$ in (4.1.4) to interchange row and column 1 with s, where s is the least integer such that $\nu_1 = |H_{ss}|$. Therefore, $|X| = \nu_1$. For a 2×2 pivot, we choose $\bar{P}$ to interchange rows and columns 1 with q and 2 with p, where q is the least column integer and p is the least row integer in the $q-$th column such that $\nu_0 = |H_{pq}|$ (note that $p > q$). Therefore,

$$\begin{aligned} |(\bar{P}H\bar{P}^T)_{21}| &= \nu_0 , \\ -\det(X) &= |\det X| \geq \nu_0^2 - \nu_1^2 . \end{aligned} \tag{4.1.7}$$

Bunch and Parlett [6] showed that the above choice of α minimizes the element growth which can take place in transition from H to H_1, and that for any pivoting strategy which satisfies (4.1.7)

$$|L_{ij}| \leq \begin{cases} 1.562 & \text{if } L_{ij} \text{ is obtained after a } 1 \times 1 \text{ pivot} \\ 2.781 & \text{otherwise .} \end{cases} \tag{4.1.8}$$

---

[1] See Rem. 4.2.3.

To obtain decomposition (4.1.2), we futher decompose PHP^T as

$$PHP^T = LQQ^T TQQ^T L^T \ ,$$

where Q is orthogonal block diagonal matrix with the same structure as T. The 1×1 blocks of Q are 1, and the 2×2 blocks of Q are elementary orthogonal plane rotation matrices of the form

$$\begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} , \qquad cs^2 + sn^2 = 1 \ ,$$

chosen to diagonalize corresponding 2×2 blocks of T. Denoting $L_1 = LQ$ and $D_1 = Q^T TQ$ we can write $PHP^T = L_1 D_1 L_1^T$, where L_1 is lower block triangular and D_1 is diagonal matrix. Due to (4.1.7) the 2×2 diagonal blocks of D_1 which correspond to 2×2 diagonal blocks of T always have one positive and one negative element. Further we have

$$PHP^T = L_1 \sqrt{|D_1|} J_1 \sqrt{|D_1|} L_1^T,$$

where J_1 is diagonal with $J_{1,ii} \in \{-1, 1\}$. Finally,

$$PHP^T = L_1 \sqrt{|D_1|} P_1 P_1^T J_1 P_1 P_1^T \sqrt{|D_1|} L_1^T, \tag{4.1.9}$$

where P_1 is a permutation matrix chosen to sort elements of J_1 according to the relation (4.1.2). Setting $G = L_1 \sqrt{|D_1|} P_1$ and $J = P_1^T J_1 P_1$ we obtain the decomposition (4.1.2).

If H is positive definite, the above algorithm reduces to the Cholesky decomposition with complete pivoting (see e.g. [13]), that is

$$PHP^T = LL^T \ . \tag{4.1.10}$$

Combining (4.1.5) and (4.1.9), and using

$$Q^T XQ = D \ ,$$

where D is a 1×1 or 2×2 diagonal matrix, we obtain (in the notation of (4.1.5))

$$\begin{aligned} \bar{P} H \bar{P}^T &= \bar{G} \begin{bmatrix} J & 0 \\ 0 & H_1 \end{bmatrix} \bar{G}^T, \\ \bar{G} &= \begin{bmatrix} B & 0 \\ Z & I \end{bmatrix} , \\ B &= Q|D|^{1/2} \ , \\ Z &= CQ|D|^{-1/2} J \ , \\ H_1 &= Y - ZJZ^T \ . \end{aligned} \tag{4.1.11}$$

Thus, we have the following:

Algorithm 4.1.1 *Symmetric indefinite decomposition (4.1.2) of a real symmetric matrix H. Vector P is initially defined by $P_i = i,\ i = 1, \ldots, n$. The symbol $\leftrightarrow$ denotes interchanging of two elements.*

```
α = (1 + √17)/8
i = 1
npos = 0
r = 0
repeat
/* finding ν0, ν1, p, q and s */
    ν0 = 0
    ν1 = |H_ii|
    for k = i + 1 to n
        if |H_kk| > ν1 then
            ν1 = |H_kk|
            s = k
        endif
        for l = i to k − 1
            if |H_kl| > ν0 then
                ν0 = |H_kl|
                p = k
                q = l
            endif
        endfor
    endfor
    if ν1 ≥ α · ν0 then
    /* 1 × 1 pivot; we first check for the non–singularity */
        if ν1 = 0 then
            r = i − 1
            i = n + 1
        else
        /* permuting */
            for k = i to n
                H_ki ↔ H_ks
            endfor
            for k = 1 to n
                H_ik ↔ H_sk
            endfor
            P_i ↔ P_s
        /* updating H */
            J_i = sign(H_ii)
            if J_i = −1 then
                npos = npos + 1
```

$endif$

$temp = \sqrt{|H_{ii}|}$

$H_{ii} = temp$

$for\ k = i+1\ \ to\ n$

$\quad H_{ki} = H_{ki} * J_i/temp$

$\quad H_{ik} = 0$

$\quad for\ l = i+1\ \ to\ k$

$\quad\quad H_{kl} = H_{kl} - H_{ki} * H_{li} * J_i$

$\quad\quad H_{lk} = H_{kl}$

$\quad endfor$

$endfor$

$i = i+1$

$endif$

$else$

$/*\ 2 \times 2\ pivot;\ we\ first\ permute\ */$

$for\ k = i\ \ to\ n$

$\quad H_{ki} \leftrightarrow H_{kq}$

$\quad H_{k,i+1} \leftrightarrow H_{kp}$

$endfor$

$for\ k = 1\ \ to\ n$

$\quad H_{ik} \leftrightarrow H_{qk}$

$\quad H_{i+1,k} \leftrightarrow H_{pk}$

$endfor$

$P_i \leftrightarrow P_q$

$P_{i+1} \leftrightarrow P_p$

$/*\ calculating\ the\ orthogonal\ matrix\ which\ diagonalizes \begin{bmatrix} H_{ii} & H_{i,i+1} \\ H_{i+1,i} & H_{i+1,i+1} \end{bmatrix}\ */$

$\zeta = (H_{i+1,i+1} - H_{ii})/(2 * H_{i+1,i})$

$t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2+1})$

$h = \sqrt{1+t^2}$

$cs = 1/h$

$sn = t/h$

$/*\ updating\ H\ \ */$

$a = H_{ii} - H_{i+1,i} * t$

$b = H_{i+1,i+1} + H_{i+1,i} * t$

$J_i = sign(a)$

$J_{i+1} = sign(b)$

$npos = npos + 1$

$a = |a|$

$b = |b|$

$H_{ii} = cs * a$

$H_{i,i+1} = sn * b$

```
        H_{i+1,i} = -sn * a
        H_{i+1,i+1} = cs * b
        for k = i + 2 to n
            temp = H_{ki}
            H_{ki} = (temp * cs - H_{k,i+1} * sn) * J_i / a
            H_{k,i+1} = (temp * sn + H_{k,i+1} * cs) * J_{i+1} / b
            H_{ik} = 0
            H_{i+1,k} = 0
            for l = i + 2 to k
                H_{kl} = H_{kl} - H_{ki} * H_{li} * J_i - H_{k,i+1} * H_{l,i+1} * J_{i+1}
                H_{lk} = H_{kl}
            endfor
        endfor
        i = i + 2
    endif
until i > n
/* if non–singularity did not occur, then rank equals dimension */
if r = 0 then
    r = n
endif
/* permuting the columns of H to sort J */
k = npos + 1
for l = 1 to npos
    if J_l = -1 then
        while J_k = -1
            k = k + 1
        endwhile
        for m = 1 to n
            H_{ml} ↔ H_{mk}
        endfor
        k = k + 1
    endif
endfor
/* r is equal to rank(H) and to the number of columns of G */
/* Matrix G is stored in the first r columns of the array H */
/* Matrix J is given implicitly by npos and r */
/* Vector P describes the pivoting which took place in the sense that H(P, P) = GJG^T */
```

4.2 Error analysis

In this section we give error analysis of the symmetric indefinite decomposition defined by Alg. 4.1.1. In our proof we use the approach from Th. 3.3.1 of [16]. We compare our result with the existing analysis of the algorithm of Bunch and Parlett [6] by Bunch [3].

Theorem 4.2.1 *Let G and J be the factors of a real symmetric matrix H computed by Alg. 4.1.1 in floating–point arithmetic with precision ε. Then, with the relative error of $O(\varepsilon)$, G and J satisfy*

$$\begin{aligned} GJG^T &= PHP^T + E \ , \\ |E| &\leq 136n(P|H|P^T + |G||G|^T)\varepsilon \ . \end{aligned} \tag{4.2.1}$$

PROOF. The proof is by induction on n. The theorem obviously holds for all matrices of order 1. To begin the induction, we must also analyse matrices of order 2 for a 2×2 pivot. Let

$$\begin{aligned} \tilde{\zeta} &= (H_{22} - H_{11})/(2H_{21}) \ , \\ \tilde{t} &= \mathrm{sign}\,(\tilde{\zeta})/(|\tilde{\zeta}| + \sqrt{1+\tilde{\zeta}^2}) \ , \\ \widetilde{cs} &= 1/\sqrt{1+\tilde{t}^2} \ , \\ \widetilde{sn} &= \tilde{t}/\sqrt{1+\tilde{t}^2} \ , \\ \tilde{a} &= H_{11} - H_{21}\tilde{t} \ , \\ \tilde{b} &= H_{22} + H_{21}\tilde{t} \ , \end{aligned} \tag{4.2.2}$$

and $\tilde{G}_{ij}$ denote the exact quantities computed by Alg. 4.1.1, i.e. without rounding errors. Since

$$|H_{21}| = \nu_0 \ , \qquad \max\{|H_{11}|, |H_{22}|\} \leq \nu_1 \ , \tag{4.2.3}$$

the fact that we perform a 2×2 step implies

$$|\tilde{\zeta}| \leq \begin{cases} \alpha & \text{if sign}\,(H_{11}) = -\text{sign}\,(H_{22}) \ , \\ \alpha/2 & \text{otherwise} \ . \end{cases} \tag{4.2.4}$$

Now we show that the computed quantities t, cs, sn, a and b have small relative errors with respect to the exact quantities from (4.2.2). Single subscribed ε's denote quantities of absolute value less than or equal to ε. Most of the subsequent inequalities hold with a relative error of $O(\varepsilon)$. Using (4.2.3) and the maximum in (4.2.4), we have

$$\begin{aligned} \zeta &= fl(\frac{H_{22} - H_{11}}{2H_{21}}) = \frac{H_{22}(1+\varepsilon_1) - H_{11}(1+\varepsilon_2)}{2H_{21}(1+\varepsilon_3)}(1+\varepsilon_4) \\ &= \tilde{\zeta} + \varepsilon_\zeta \ , \end{aligned}$$

where $|\varepsilon_\zeta| \le 3\alpha\varepsilon$. This implies that the equality

$$fl(1+\zeta^2) = (1+\varepsilon_5)(1+(\tilde{\zeta}+\varepsilon_\zeta)^2(1+\varepsilon_6)) = (1+\tilde{\zeta}^2)(1+\varepsilon') \ ,$$

holds for some

$$|\varepsilon'| \le 2|\varepsilon_\zeta \tilde{\zeta}| + (|\varepsilon_5| + |\varepsilon_6|)\tilde{\zeta}^2 + |\varepsilon_5| \le 4.3\varepsilon \ .$$

Further, the equality

$$\begin{aligned} fl(|\zeta| + \sqrt{1+\zeta^2}) &= (1+\varepsilon_7)(|\tilde{\zeta}+\varepsilon_\zeta| + (1+\varepsilon_8)(1+\varepsilon'/2)\sqrt{1+\tilde{\zeta}^2}) \\ &= (1+\varepsilon'')(|\tilde{\zeta}| + \sqrt{1+\tilde{\zeta}^2}) \end{aligned}$$

holds for some

$$|\varepsilon''| \le |\varepsilon_\zeta| + |\varepsilon_7|(1+|\tilde{\zeta}|) + |\varepsilon_8| + |\varepsilon'|/2 \le 7\varepsilon \ ,$$

so that finally

$$\begin{aligned} t &= \tilde{t}(1+\varepsilon_t) \ , & |\varepsilon_t| &\le 8\varepsilon \ , \\ cs &= fl(1/\sqrt{1+t^2}) = \widetilde{cs}(1+\varepsilon_{cs}) \ , & |\varepsilon_{cs}| &\le 11\varepsilon \ , \\ sn &= fl(t/\sqrt{1+t^2}) = \widetilde{sn}(1+\varepsilon_{sn}) \ , & |\varepsilon_{sn}| &\le 19\varepsilon \ . \end{aligned} \tag{4.2.5}$$

Let

$$\begin{aligned} a &= fl(H_{11} - H_{21}t) \\ b &= fl(H_{22} + H_{21}t) \ . \end{aligned} \tag{4.2.6}$$

If sign $(H_{11}) = -$sign (H_{22}), then a and b are both calculated by addition and have small relative errors, i.e.

$$a = \tilde{a}(1+\varepsilon_a) \ , \qquad b = \tilde{b}(1+\varepsilon_b) \ , \qquad |\varepsilon_a|, |\varepsilon_b| \le 10\varepsilon \ . \tag{4.2.7}$$

Let sign $(H_{11}) =$ sign (H_{22}). Assume further that $H_{11} \ge H_{22} \ge 0$. Then a is again calculated by addition and (4.2.7) holds for it. Using (4.2.4), $|H_{21}| = \nu_0$, and $|H_{22}| \le \nu_1$, we have

$$\begin{aligned} |\tilde{b}| &= |H_{22} + H_{21}\tilde{t}| \ge |\tilde{t}||H_{21}| - |H_{22}| \\ &\ge \nu_0 \left(\frac{1}{\alpha/2 + \sqrt{1+\alpha^2/4}} - \alpha \right) \ge 0.088\nu_0 \ . \end{aligned}$$

Therefore,

$$\begin{aligned} b = H_{22}(1+\varepsilon_9) + (1+\varepsilon_{10})(1+\varepsilon_{11})(1+\varepsilon_t)H_{21}\tilde{t} = \tilde{b}(1+\varepsilon_b') \ , \\ |\varepsilon_b'| \le (|H_{22}\varepsilon_9| + (|\varepsilon_{10}| + |\varepsilon_{11}| + |\varepsilon_t|)|H_{21}||\tilde{t}|)/\tilde{b} \le 121\varepsilon \ . \end{aligned}$$

We conclude that in any case

$$a = \tilde{a}(1+\varepsilon_a)\ , \qquad b = \tilde{b}(1+\varepsilon_b)\ , \qquad |\varepsilon_a|, |\varepsilon_b| \le 121\varepsilon\ . \tag{4.2.8}$$

This implies that, e.g.

$$\begin{aligned} G_{21} &= fl(-sn\sqrt{|a|}) = \tilde{G}_{21}(1+\varepsilon_G)\ , \\ |\varepsilon_G| &\le |\varepsilon_{sn}| + |\varepsilon_a|/2 + 2\varepsilon \le 81.5\varepsilon\ , \end{aligned}$$

so that

$$B = G = \tilde{G} + \delta G\ , \qquad |\delta G| \le 81.5|\tilde{G}|\varepsilon\ . \tag{4.2.9}$$

Thus,

$$\begin{aligned} BJB^T &= GJG^T = (\tilde{G}+\delta G)J(\tilde{G}+\delta G)^T = H + E\ , \\ |E| &\le 2\cdot 81.5|\tilde{G}||\tilde{G}|^T\varepsilon \\ &= 163|G||G|^T\varepsilon\ , \end{aligned} \tag{4.2.10}$$

and (4.2.1) holds.

The induction step must also be done separately for a 1×1 and a 2×2 pivot. We can assume without loss of generality that $\bar{P}$ from (4.1.4) is the identity. Moreover, permuting the columns of G in order to sort the elements of J (see (4.1.9)) does not influence the statement of the theorem. From (4.1.4) and (4.1.11) we conclude that

$$H = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix} = \begin{bmatrix} B & 0 \\ Z & I \end{bmatrix} \begin{bmatrix} J & 0 \\ 0 & H_1 \end{bmatrix} \begin{bmatrix} B^T & Z^T \\ 0 & I \end{bmatrix}\ . \tag{4.2.11}$$

Suppose that we do a 1×1 step, and that (4.2.1) holds for all matrices of order $n-1$. Then (4.2.11) holds with

$$\begin{aligned} B &= fl(|H_{11}|^{1/2}) = |H_{11}|^{1/2} + \delta B\ , \\ |\delta B| &\le |H_{11}|^{1/2}\varepsilon\ , \\ Z &= fl(CJ/B) = CJ|H_{11}|^{-1/2} + \delta Z\ , \\ |\delta Z| &\le 2\varepsilon|C||H_{11}|^{-1/2}\ , \\ H_1 &= fl(Y - ZJZ^T) = Y - ZJZ^T + F_1\ , \\ |F_1| &\le 2\varepsilon(|Y| + |Z||Z|^T)\ . \end{aligned} \tag{4.2.12}$$

The induction assumption implies that

$$G_1 J_1 G_1^T = P_1 H_1 P_1^T + E_1\ , \tag{4.2.13}$$

where

$$|E_1| \le 136(n-1)\varepsilon(P_1|H_1|P_1^T + |G_1||G_1|^T)\ .$$

Now Alg. 4.1.1 yields

$$G = \begin{bmatrix} B & 0 \\ P_1 Z & G_1 \end{bmatrix} ,$$

so that

$$G \begin{bmatrix} J & \\ & J_1 \end{bmatrix} G^T = \begin{bmatrix} BJB^T & BJZ^T P_1^T \\ P_1 ZJB^T & P_1 ZJZ^T P_1^T + G_1 J_1 G_1^T \end{bmatrix} . \tag{4.2.14}$$

Setting $P = I \oplus P_1$ and using (4.2.12), we obtain

$$\begin{aligned} G \begin{bmatrix} J & \\ & J_1 \end{bmatrix} G^T &= P \begin{bmatrix} H_{11} & C^T \\ C & Y \end{bmatrix} P^T \\ &\quad + \begin{bmatrix} 2\delta BJ|H_{11}|^{1/2} & \delta C^T \\ \delta C & E_1 + P_1 F_1 P_1^T \end{bmatrix} (1 + O(\varepsilon)) \\ &\equiv H + E , \end{aligned}$$

where

$$\delta C = P_1(\delta ZJ|H_{11}|^{1/2} + C|H_{11}|^{-1/2}\delta B) .$$

Using this and (4.2.12), we obtain

$$|E| \leq \begin{bmatrix} 2|H_{11}| & 3|C|^T P_1^T \\ 3P_1|C| & |E_1 + P_1 F_1 P_1^T| \end{bmatrix} \varepsilon . \tag{4.2.15}$$

From (4.2.12) it follows that

$$|H_1| \leq (1 + 2\varepsilon)(|Y| + |Z||Z|^T) .$$

By using (4.2.13), we have

$$|E_1 + P_1 F_1 P_1^T| \leq (136(n-1) + 2)(P_1(|Y| + |Z||Z|^T)P_1^T + |G_1||G_1|^T)\varepsilon . \tag{4.2.16}$$

Inserting the above relation into (4.2.15) we obtain

$$|E| \leq (136(n-1) + 3)(P|H|P^T + |G||G|^T)\varepsilon ,$$

which, in turn, implies (4.2.1).

Now suppose that we do a 2×2 pivot, and that (4.2.1) holds for all matrices of order $n - 2$. From the analysis of the 1×1 step, we see that we can without loss of generality assume that P_1 equals identity. Let H be partitioned as in (4.1.4), and let $\widetilde{Q}^T X \widetilde{Q} = \widetilde{D}$ be the exact spectral decomposition of X. Let Q and D be the computed matrices $\widetilde{Q}$ and $\widetilde{D}$, respectively. The analysis of the 2×2 case for $n = 2$ holds for the floating–point spectral decomposition of X, as well. Thus, (4.2.5) and (4.2.8) imply that

$$\begin{aligned} Q &= \widetilde{Q} + \delta Q , & |\delta Q| &\leq 19|\widetilde{Q}|\varepsilon , \\ D &= \widetilde{D} + \delta D , & |\delta D| &\leq 121|\widetilde{D}|\varepsilon . \end{aligned} \tag{4.2.17}$$

Now (4.2.11) holds with H_1 defined by (4.2.12), and B and Z as follows: from (4.2.9) it follows directly that

$$\begin{aligned} B &= fl(Q|D|^{1/2}) = \widetilde{Q}|\widetilde{D}|^{1/2} + \delta B \ , \\ |\delta B| &\leq 81.5|\widetilde{Q}||\widetilde{D}|^{1/2}\varepsilon \ , \end{aligned} \tag{4.2.18}$$

and from (4.1.11) and (4.2.17) it follows that

$$\begin{aligned} Z &= fl(CQ|D|^{-1/2}J) = C\widetilde{Q}|\widetilde{D}|^{-1/2}J + \delta Z \ , \\ |\delta Z| &\leq 83.5|C||\widetilde{Q}||\widetilde{D}|^{-1/2}\varepsilon \ . \end{aligned} \tag{4.2.19}$$

As in the 1×1 case, the induction assumption (4.2.13), where now

$$|E_1| \leq 136(n-2)\varepsilon(|H_1| + |G_1||G_1|^T) \ ,$$

implies (4.2.14). This, (4.2.18), (4.2.19), (4.2.12), and (4.2.13), imply that

$$G\begin{bmatrix} J & \\ & J_1 \end{bmatrix}G^T = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix} + \begin{bmatrix} \delta X & \delta C^T \\ \delta C & E_1 + F_1 \end{bmatrix} \equiv H + E \ , \tag{4.2.20}$$

where

$$\delta C = \delta Z J |\widetilde{D}|^{1/2}\widetilde{Q}^T + C\widetilde{Q}|\widetilde{D}|^{-1/2}J\delta B^T \ .$$

From (4.2.10) it follows directly that

$$|\delta X| \leq 163|B||B|^T\varepsilon \ . \tag{4.2.21}$$

As in the proof of (4.2.16), we have

$$|E_1 + F_1| \leq (136(n-2)+2)(|Y| + |Z||Z|^T + |G_1||G_1|^T)\varepsilon \ , \tag{4.2.22}$$

and it remains to bound $|\delta C|$ from above. From (4.2.18) and (4.2.19) it follows

$$|\delta C| \leq 165|C||\widetilde{Q}||\widetilde{Q}|^T\varepsilon \ .$$

It is easy to see that

$$|C||\widetilde{Q}||\widetilde{Q}|^T \leq |C| + \begin{bmatrix} |C_{\cdot 2}| & |C_{\cdot 1}| \end{bmatrix} \ ,$$

where $C_{\cdot j}$ denotes the $j-$th column of C. Further,

$$\begin{aligned} |Z||B|^T &= |C\widetilde{Q}|\widetilde{D}|^{-1/2} + \delta Z| \cdot ||\widetilde{D}|^{1/2}\widetilde{Q}^T + \delta B^T| \\ &\geq |C\widetilde{Q}||\widetilde{Q}|^T - 165|C||\widetilde{Q}||\widetilde{Q}|^T\varepsilon \ . \end{aligned} \tag{4.2.23}$$

Now

$$(|C\widetilde{Q}||\widetilde{Q}|^T)_{\cdot i} = |C_{\cdot 1}\widetilde{cs} - C_{\cdot 2}\widetilde{sn}|\widetilde{cs} + |C_{\cdot 1}\widetilde{sn} + C_{\cdot 2}\widetilde{cs}||\widetilde{sn}| \ .$$

Simple checking of all possible combinations for the signs of C_{ij} and $\widetilde{sn}$ shows that either

$$|C_{i1}\widetilde{cs} - C_{i2}\widetilde{sn}| = |C_{i1}|\widetilde{cs} + |C_{i2}||\widetilde{sn}| \ , \tag{4.2.24}$$

or

$$|C_{i1}\widetilde{sn} + C_{i2}\widetilde{cs}| = |C_{i1}||\widetilde{sn}| + |C_{i2}|\widetilde{cs} \ . \tag{4.2.25}$$

If (4.2.24) holds for some i, then

$$(|C\widetilde{Q}||\widetilde{Q}|^T)_{i1} \geq |C_{i1}|(\widetilde{cs}^2 - \widetilde{sn}^2) + 2|C_{i2}||\widetilde{sn}|\widetilde{cs} \ .$$

From (4.2.4) it follows that

$$|\widetilde{t}| \geq \frac{1}{\alpha + \sqrt{1+\alpha^2}} \ . \tag{4.2.26}$$

Therefore,

$$2\widetilde{cs}|\widetilde{sn}| \geq (1+\alpha^2)^{-1/2} \geq 0.842 \ ,$$

and

$$(|C\widetilde{Q}||\widetilde{Q}|^T)_{i1} \geq 0.842|C_{i2}| \ .$$

If (4.2.25) holds for some i, then

$$(|C\widetilde{Q}||\widetilde{Q}|^T)_{i1} \geq 2|C_{i2}||\widetilde{sn}|\widetilde{cs} - |C_{i1}|(\widetilde{cs}^2 - \widetilde{sn}^2) \ .$$

From (4.2.26) it follows that

$$\widetilde{cs}^2 - \widetilde{sn}^2 \leq \alpha(1+\alpha^2)^{-1/2} \leq 0.54 \ ,$$

so that

$$(|C\widetilde{Q}||\widetilde{Q}|^T)_{i1} \geq 0.842|C_{i2}| - 0.54|C_{i1}| \ .$$

The similar analysis holds for the second column of $|C\widetilde{Q}||\widetilde{Q}|^T$, too, and we conclude that

$$\begin{aligned} |C| + \left[\ |C_{\cdot 2}| \quad |C_{\cdot 1}| \ \right] \ &\leq \ \frac{1}{0.842}|C\widetilde{Q}||\widetilde{Q}|^T + \left(1 + \frac{0.54}{0.842}\right)|C| \\ &\leq \ 1.642\,(|C\widetilde{Q}||\widetilde{Q}|^T + |C|) \ . \end{aligned}$$

Using this and (4.2.23), and ignoring the relative error of $O(\varepsilon)$, we obtain

$$\begin{aligned} |\delta C| \ &\leq \ 165 \cdot 1.642\,(|Z||B|^T + |C| + 165\,|C||\widetilde{Q}||\widetilde{Q}|^T\varepsilon)\varepsilon \\ &\leq \ 271\,(|Z||B|^T + |C|)\varepsilon \ . \end{aligned}$$

Finally, (4.2.1) follows by inserting this, (4.2.21) and (4.2.22) into (4.2.20), and the theorem is proved. Q.E.D.

Bunch [3] showed that the decomposition (4.1.3) with the unequilibrated diagonal pivoting (which is also used in Alg. 4.1.1) is stable in the following sense: let L and T be the factors of H computed in floating–point arithmetic with precision ε. Then

$$LTL^T = PHP^T + F \ ,$$

where

$$\|F\|_1 \leq \max_k \nu_0^{(k)}(21.6n + 7.9n^2)\varepsilon \ ,$$

and $\nu_0^{(k)}$ is the value of ν_0 in the $k-$th reduction step. The quantity $\max_k \nu_0^{(k)}$ is further bounded by

$$\max_k \nu_0^{(k)} \leq \max_{i,j} |H_{ij}| 3.07 f(n)\sqrt{n}(n-1)^{0.446} \ ,$$

where

$$f(n) = \left(\prod_{k=2}^{n} k^{1/(k-1)}\right)^{1/2} \leq 2n^{(1/4)\log n} \ .$$

The bound of Th. 4.2.1 compares favourably to the above bounds, since it does not contain the $n^2\varepsilon$ term. The quantity $\max_k \nu_0^{(k)}$ is implicitly included in the $|G||G|^T$ term of (4.2.1). Note that Th. 4.2.1 holds for a singular H, as well.

From the proof of Th. 4.2.1 we see that 2×2 steps contribute much more to the error bound than 1×1 steps. If only 1×1 steps are performed (which is always the case when we decompose a positive definite matrix, and is often the case when we decompose scaled diagonally dominant matrices of [2]), then the bound (4.2.1) reduces to

$$|E| \leq 3n(P|H|P^T + |G||G|^T)\varepsilon \ .$$

In the positive definite case Alg. 4.1.1 reduces to the Cholesky decomposition with complete pivoting, and only 1×1 steps are performed. The above inequality then implies

$$|E_{ij}| \leq 6n((PHP^T)_{ii}(PHP^T)_{jj})^{1/2}\varepsilon \ ,$$

which is similar to the result of Demmel [9]. There the constant $6n$ is replaced by $(n+1)/(1-(n+1)\varepsilon)$. Note, however, that the above bound holds for the outer product version of the Cholesky decomposition (Alg. 4.2.2 of [16]) with the addition of the complete pivoting, while Demmel [9] analysed the Gaxpy version (Alg. 4.2.1 of [16]).

Remark 4.2.2 Numerical experiments of Chap. 5 indicate that the bound (4.2.1) increases only slowly with n.

Remark 4.2.3 *Other pivot strategies.* Note that Th. 4.2.1 and then, in turn, Th. 4.3.1, hold for any pivot strategy for which (4.2.4) holds when we apply a 2×2 step. In particular, these theorems hold for the partial pivot strategy of [5] and for the pivot strategy given by Algorithm C of [4], which both require $O(n^2)$ search. We have chosen the unequilibrated diagonal pivoting since it has better bounds for the element growth, as well as the uniform upper bound for the scaled condition of the matrix $G^T G$ (see Sect. 4.4). Moreover, since the symmetric indefinite decomposition takes about 10% of the computing time, an $O(n^3)$ search, which is needed by the unequilibrated diagonal pivoting, does not considerably slow down the algorithm. However, theoretical and practical investigation of Algorithm C of [4] (for positive definite matrices this algorithm also reduces to Cholesky decomposition with complete pivoting), is certainly of interest.

4.3 Overall error bounds

The results of the previous parts of the thesis suggest the following procedure to solve the real symmetric eigenvalue problem (4.1.1):

1. decompose H as $H = GJG^T$ by first using Alg. 4.1.1 to obtain the decomposition (4.1.2), and then setting $G = P^T G$ as follows (in the notation of Alg. 4.1.1):

   ```
   /* Back–permuting the rows of H to obtain the final factor */
       for k = 1 to n
           for l = k + 1 to n
               if P_l = k then
                   P_l ↔ P_k
                   for m = 1 to r
                       H_km ↔ H_lm
                   endfor
               endif
           endfor
       endfor
   ```

2. solve the problem (4.1.1) by applying any of the implicit J–orthogonal Jacobi methods of Chap. 3 on the pair G, J.

In this section we combine the error analysis of the symmetric indefinite decomposition, error analysis of the implicit J–orthogonal Jacobi methods, and the perturbation bounds of Chap. 2, to obtain error bounds for the computed eigensolution of the real symmetric eigenvalue problem. Bounds hold only in the non–singular case, since we cannot otherwise apply the perturbation theory of Sect. 2.2 to Th. 4.2.1. We give error bounds for the case when the implicit method of Alg. 3.3.1 is used. Error bounds for other implicit methods of Chap. 3 are obtained by simply substituting error bounds for those methods in the final estimate. We then show that an approximation for the error bounds can be obtained using only computed quantities. We also discuss what happens in the singular case. We give an interesting example how a change of the pivoting in the symmetric indefinite decomposition can considerably improve the accuracy of the obtained eigensolution. In the conclusion, we summarize some open problems.

Theorem 4.3.1 *Let H be a real symmetric non–singular matrix and let λ be the $i-th$ eigenvalue of H. Let G, J be the decomposition of H obtained by Alg. 4.1.1 in floating–point arithmetic with precision ε, and let $G = D_G B_G$, where D_G is diagonal and the rows of B_G have unit norms. Let λ_G be the $i-th$ eigenvalue of GJG^T. Let G_m, J be the sequence of pairs obtained from the pair G, J by Alg. 3.3.1 in floating–point arithmetic with precision ε, and let G_M, J be the final pair which satisfies the stopping criterion. For $m \geq 0$ write $G_m = B_m D_m$, where D_m is diagonal and columns of B_m have unit norms. Let λ' be the $i-th$ calculated eigenvalue. Then, with the relative error of $O(\varepsilon)$, we have*

$$1 - \eta - \eta_1 \leq \frac{\lambda'}{\lambda} \leq 1 + \eta + \eta_1, \tag{4.3.1}$$

where

$$\begin{aligned} \eta &= \frac{272\, n^2 \varepsilon}{\lambda_{min}(D_G^{-1} \| GJG^T \| D_G^{-1})} , \\ \eta_1 &= 2\varepsilon \sum_{m=0}^{M-1} \frac{C_m}{\sigma_{min}(B_m)} + n \cdot tol + n^2 \varepsilon , \end{aligned} \tag{4.3.2}$$

$\| \cdot \|$ is the spectral absolute value defined in Sect. 2.1, and C_m are constants from Th. 3.3.3.

Now suppose λ is simple. Let v be the corresponding eigenvector. Let v' be the eigenvector corresponding to λ', i.e. the $i-th$ column of G_M divided by its norm. Then

$$\begin{aligned} \|v' - v\|_2 \leq\; & \frac{\sqrt{2}\eta}{rg(\lambda)} \cdot \frac{1}{1 - \left(1 + \frac{1}{rg(\lambda)}\right)\eta} \\ & + \frac{4\sqrt{2}\bar{\eta}}{rg_G(\lambda_G)} \cdot \frac{1}{1 - \frac{3\bar{\eta}}{rg_G(\lambda_G)}} + 2n \cdot tol + n(3n+4)\varepsilon , \end{aligned} \tag{4.3.3}$$

provided $1 < (1 + 1/rg(\lambda))\eta$ and $rg_G(\lambda_G) < 3\bar{\eta}$. Here $rg(\lambda)$ and $rg_G(\lambda)$ are defined by (2.2.29) and (2.3.1), respectively, and

$$\begin{aligned} \bar{\eta} &= \eta_2(2 + \eta_2) , \\ \eta_2 &= (\eta_1 + n \cdot tol + n^2 \varepsilon)/2 , \end{aligned}$$

where η is defined by (4.3.2).

Proof. From Th. 4.2.1, by multiplying (4.2.1) by P^T from the left and by P from the right, and then setting $G = P^T G$, it follows that

$$H = GJG^T + \delta H \ ,$$

where

$$|\delta H| \leq 136\, n(|H| + |G||G|^T)\varepsilon \ .$$

Also,

$$|H| \leq |GJG^T| + |\delta H| \leq |G||G|^T + |\delta H| \ ,$$

so that, by ignoring the relative error of $O(\varepsilon)$, we have

$$|\delta H| \leq 272\, n|G||G|^T \varepsilon \ .$$

Further,

$$\begin{aligned} |x^T \delta H x| &\leq |x|^T |\delta H||x| \leq 272\, n|x|^T |D_G E D_G||x|\varepsilon \\ &\leq 272\, n^2 x^T D_G^2 x\, \varepsilon \\ &\leq \frac{272\, n^2 \varepsilon}{\lambda_{min}(D_G^{-1} \boldsymbol{|} GJG^T \boldsymbol{|} D_G^{-1})} x^T \boldsymbol{|} GJG^T \boldsymbol{|} x \ . \end{aligned}$$

Applying Th. 2.2.1 to the pair GJG^T, I with

$$\eta \equiv \eta_H = \frac{272\, n^2\, \varepsilon}{\lambda_{min}(D_G^{-1} \boldsymbol{|} GJG^T \boldsymbol{|} D_G^{-1})} \ , \qquad \eta_I = 0 \ ,$$

we obtain

$$1 - \eta \leq \frac{\lambda_G}{\lambda} \leq 1 + \eta \ .$$

This and Cor. 3.3.4, by ignoring the relative error of $O(\varepsilon)$, imply (4.3.1).

Let v_G be the eigenvector of λ_G. Applying (2.3.13) and Th. 2.2.13 to the matrix GJG^T yields

$$\|v_G - v\|_2 \leq \frac{\eta}{rg(\lambda)} \cdot \frac{1}{1 - \left(1 + \dfrac{1}{rg(\lambda)}\right)\eta} \ .$$

The relation (4.3.3) now follows from the above relation, (2.3.13), Th. 3.3.9, and the triangle inequality. The assumptions on $rg(\lambda)$ and $rg_G(\lambda_G)$ together with the proofs of Theorems 2.2.13 and 3.3.9, implies that λ is throughout the algorithm well separated from the rest of the spectrum. Q.E.D.

Remark 4.3.2 Th. 4.3.1 also holds if we substitute $\boldsymbol{|} GJG^T \boldsymbol{|}$ by $\boldsymbol{|} H \boldsymbol{|}$ in (4.3.2). Indeed, if we consider GJG^T as $H - \delta H$, then

$$|x^T \delta H x| \leq \frac{272\, n^2 \varepsilon}{\lambda_{min}(D_G^{-1} \boldsymbol{|} H \boldsymbol{|} D_G^{-1})} x^T \boldsymbol{|} H \boldsymbol{|} x\, \varepsilon \ ,$$

and we can apply Theorems 2.2.1 and 2.2.13 directly to H. We are using $|GJG^T|$ since G is the *computed* factor and D_G its exact scaling.

Th. 4.3.1 implies that the error bounds depend on how D_G^{-1} scales $|GJG^T|$. In the positive definite case $|GG^T| = GG^T$ and the scaling with D_G is optimal in the sense of (2.1.4). Our numerical experiments show that in the indefinite case the scaling with D_G is also not far from the almost optimal one by $(\mathrm{diag}\, |GJG^T|)^{-1/2}$.

It is natural to want to approximate the bounds (4.3.1) and (4.3.3) by using only computed quantities. We can substitute $rg(\lambda)$ and $rg_G(\lambda_G)$ with $rg(\lambda')$ and $rg_G(\lambda')$, respectively. Although $\lambda' = \lambda(1+O(\varepsilon)) = \lambda_G(1+O(\varepsilon))$, the above substitutions can have large relative errors. However, in numerical tests they are shown to be realistic. Further, we can substitute $|GJG^T|$ with $G_M G_M^T$. This is justified as follows: if F is a $J-$orthogonal matrix which diagonalizes some $G^T G$ as in the proof of Th. 2.3.1, then $|GJG^T| = GFF^TG^T$. Now consider the matrix

$$G_M' \equiv (B_M + \delta B_M)D_M \equiv G_M + \delta B_M D_M \equiv (G + \delta G^{(M)})R_0 \cdot \ldots \cdot R_{M-1}$$

from the proof of Th. 3.3.9. This matrix has orthogonal columns so that

$$G_M' G_M'^T = |(G + \delta G^{(M)})J(G + \delta G^{(M)})^T| ,$$

and $G_M' G_M'^T$ is, in turn, "not far" from $G_M G_M^T$. We have no theoretical results about the quality of this approximation, but its use is also justified by numerical experiments. Moreover, since we observed that the actual errors increase only slowly as n and M increase, and that the condition of the scaled matrix grows only little during the Jacobi process, we expect that

$$\begin{aligned} \left|\frac{\lambda'-\lambda}{\lambda}\right| &\leq \left(\frac{1}{\lambda_{min}(D_G^{-1}G_M G_M^T D_G^{-1})} + \frac{2}{\sigma_{min}(B)}\right)\varepsilon , \qquad (4.3.4) \\ \|v'-v\|_2 &\leq \frac{\eta}{rg(\lambda')} \cdot \frac{1}{1-\left(1+\dfrac{1}{rg(\lambda')}\right)\eta} + \frac{4\bar{\eta}}{rg_G(\lambda')} \cdot \frac{1}{1-\dfrac{3\bar{\eta}}{rg_G(\lambda')}} , \end{aligned}$$

where

$$\begin{aligned} \eta &= \frac{\varepsilon}{\lambda_{min}(D_G^{-1}G_M G_M^T D_G^{-1})} , \\ \bar{\eta} &= \frac{3\varepsilon}{\sigma_{min}(B)} . \end{aligned}$$

We cannot apply Th. 4.3.1 to singular matrices, since $|GJG^T|^{-1}$ is not defined. However, if we obtain a componentwise accurate factor G, as in the example (2.3.13), relative errors of the computed eigenvalues are bounded by Cor. 3.3.4. In this concrete example, we first have to bring J to the form $I \oplus (-I)$. This is equivalent to performing

m trigonometric rotations for $\pi/4$ on G from the right. These rotations add m terms to γ of (3.3.20). Since Th. 3.3.9 (Th. 2.3.3) requires the non–singularity of G, we have no error bounds for the eigenvectors in this case.

The following example opens an interesting problem about the pivot choice in the symmetric indefinite decomposition. The example underlines once more the importance of exact factors, and shows what difficulties we have when trying to do deflation. Consider the matrix

$$H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & \alpha^2 \end{bmatrix}, \tag{4.3.5}$$

where $\alpha > 0$ is small. Alg. 4.1.1 decomposes H as $H = GJG^T$ with

$$G = \begin{bmatrix} 1 & & \\ 1 & 1 & \\ 1 & 1 & \sqrt{\alpha^2 - 1 + 1} \end{bmatrix}, \qquad J = \operatorname{diag}(1, -1, 1) .$$

Since $|H|$ is given by (2.3.16), the error bounds of Th. 4.3.1 are large. Since in calculating $fl(\sqrt{\alpha^2 - 1 + 1})$ we obtain only $\log\alpha^2 - \log\varepsilon$ accurate digits, these error bounds are almost attained. However, since $1/\sigma_{min}(B) \approx 2.5$, any implicit Jacobi method will compute the eigensolution of the pair G, J with high accuracy. This means that when using Alg. 4.1.1, we can do deflation only if the submatrix which is to be reduced at some stage is exactly zero.[2] One way to accurately decompose H is given by (2.3.15). Here we give another one: let us first choose 2×2 pivot in (4.1.5). Then we have

$$H = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & 1 & \alpha \end{bmatrix} \begin{bmatrix} 1 & 1 & \\ 1 & 0 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ & 1 & 1 \\ & & \alpha \end{bmatrix} .$$

It is easy to see that with this pivot choice Alg. 4.1.1 returns the factor G which has componentwise small relative errors. Therefore, the first terms of (4.3.1) and (4.3.3) are superfluous, and, since $1/\sigma_{min}(B) \approx 2$, the obtained eigensolution is accurate. This underlines the importance of accurate factors, and shows that the unequilibrated diagonal pivoting is not always the best choice.

Now we shortly summarize some open problems:

- finding a realistic upper bound for the growth of $1/\sigma_{min}(B_M)$,
- how well does D_G^{-1} scale $|GJG^T|$, and how well does $\lambda_{min}(D_G^{-1} G_M G_M^T D_G^{-1})$ approximate $\lambda_{min}(D_G^{-1} |GJG^T| D_G^{-1})$,
- proving Th. 2.3.3 for the non–square full column rank G,
- improving the pivot strategy in Alg. 4.1.1, to avoid the unnecessary errors as in the example (4.3.5).

[2] In [6, 5] deflation is also performed only in this case.

The last problem is very difficult, and it is similar to the problem of finding the best pivots in Gaussian elimination. It is easy to see (see also Rutishauser [23]) that Gauss' algorithm with complete pivoting is also inaccurate when applied to (4.3.5).

4.4 Bound for the scaled condition of G^TG

The symmetric indefinite decomposition of Alg. 4.1.1 enables us to transform the eigenvalue problem (4.1.1) to the eigenvalue problem for the pair G^TG, J on which we can use implicit methods of Chap. 3. In this section we show that the scaled condition of the matrix G^TG is bounded by a function of n irrespectively of the condition of the starting matrix H. This bound is nearly attainable. For related results see [19]. Numerical experiments of [13] and Chap. 5 show that the scaled condition of G^TG is generally much smaller than our bound. This has a positive effect on the speed of the implicit methods applied on the pair G^TG, J. The results of this section are partially contained in [25].

For any positive definite matrix H we define the scaled matrix $A \equiv \mathrm{Scal}\,(H)$ by $H = DAD$, where D is diagonal positive definite, and A has ones on the diagonal.

We analyse separately the positive definite and indefinite case. For H positive definite Alg. 4.1.1 reduces to the Cholesky decomposition with complete pivoting (4.1.10), $PHP^T = LL^T$. Complete pivoting is equivalent to the fact that

$$L_{ii}^2 \geq \sum_{k=i}^{j} L_{jk}^2\ , \qquad i = 1, \ldots, n-1, \quad j > i.$$

This implies

$$L_{ii} \geq L_{jj}\ , \quad L_{ii} > |L_{ji}|\ , \quad i = 1, \ldots, n-1,\ j > i\ . \tag{4.4.1}$$

Set

$$H_1 = L^T L\ . \tag{4.4.2}$$

If (λ, x) is an eigenpair of H, then $(\lambda, L^{-1}x)$ is an eigenpair of H_1. Let $A_1 = \mathrm{Scal}\,(H_1)$, i.e.

$$H_1 = D_1 A_1 D_1\ . \tag{4.4.3}$$

Demmel and Veselić [13] showed that $\kappa(A_1)$ is bounded by a constant depending only on the dimension n. For example, for $n = 2$ it is easy to see that $\kappa(A_1) < 3 + 2\sqrt{2}$. For general n their upper bound is

$$\kappa(A_1) < e \cdot n \cdot n!\ , \tag{4.4.4}$$

which is, as they stated, a large overestimate. Here $e = \exp\,(1)$.

Now we analyse matrix A_1 in more detail and give a better bound which can be almost attained. We first illustrate the idea of the analysis on a 3×3 example. Let

$PHP^T = LL^T$ be the Cholesky decomposition with complete pivoting of a 3×3 positive definite matrix H. By

$$D_1 = \mathrm{diag}\,(\sqrt{L_{11}^2 + L_{21}^2 + L_{31}^2}, \sqrt{L_{22}^2 + L_{32}^2}, L_{33})$$

we have

$$A_1 = \begin{bmatrix} 1 & \dfrac{L_{21}L_{22} + L_{31}L_{32}}{\sqrt{L_{11}^2 + L_{21}^2 + L_{31}^2}\sqrt{L_{22}^2 + L_{32}^2}} & \dfrac{L_{31}}{\sqrt{L_{11}^2 + L_{21}^2 + L_{31}^2}} \\ & 1 & \dfrac{L_{32}}{\sqrt{L_{22}^2 + L_{32}^2}} \\ sym. & & 1 \end{bmatrix}.$$

Now we need two monotonicity properties of the norm $\|\cdot\|_2$,

$$\|A\|_2 \leq \||A|\|_2 \leq \sqrt{n}\|A\|_2 \ , \tag{4.4.5}$$

where $|A| = |A_{ij}|$, and

$$|A_{ij}| \leq B_{ij} \Longrightarrow \|A\|_2 \leq \|B\|_2 \ . \tag{4.4.6}$$

From (4.4.5) and (4.4.1) we conclude that $\|A_1\|_2 \leq \|A'\|_2$ where $A' = D^{-1}|L|^T|L|D^{-1}$, i.e. the worst case is when all L_{ij}, $i \neq j$, are non–negative. Treating A'_{23} as a monotonically increasing function of the (positive) variable L_{32}, from (4.4.1) it follows

$$A'_{23} < \frac{L_{22}}{\sqrt{L_{22}^2 + L_{22}^2}} = \sqrt{\frac{1}{2}}.$$

Treating A'_{13} as an increasing function of L_{31} we have

$$A'_{13} < \frac{L_{11}}{\sqrt{L_{11}^2 + L_{11}^2 + L_{21}^2}} \leq \sqrt{\frac{1}{2}}.$$

The element A'_{12} is an increasing function in three (positive) variables L_{21}, L_{31} and L_{32}. Therefore,

$$A'_{12} < \frac{L_{11}L_{22} + L_{11}L_{22}}{\sqrt{L_{11}^2 + L_{11}^2 + L_{11}^2}\sqrt{L_{22}^2 + L_{22}^2}} = \sqrt{\frac{2}{3}}.$$

Finally, from (4.4.6) we conclude that

$$\|A_1\|_2 < \left\| \begin{bmatrix} 1 & \sqrt{2/3} & \sqrt{1/2} \\ \sqrt{2/3} & 1 & \sqrt{1/2} \\ \sqrt{1/2} & \sqrt{1/2} & 1 \end{bmatrix} \right\|_2 \leq 1 + \sqrt{\frac{2}{3}} + \sqrt{\frac{1}{2}}.$$

Further, we have $A_1^{-1} = D_1 L^{-1} L^{-T} D_1$, where

$$L^{-1} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & \dfrac{1}{L_{33}} \end{bmatrix} \begin{bmatrix} 1 & & \\ & \dfrac{1}{L_{22}} & \\ & -\dfrac{L_{32}}{L_{22}} & 1 \end{bmatrix} \begin{bmatrix} \dfrac{1}{L_{11}} & & \\ -\dfrac{L_{21}}{L_{11}} & 1 & \\ -\dfrac{L_{31}}{L_{11}} & & 1 \end{bmatrix} .$$

From (4.4.5), (4.4.6) and (4.4.1) we see that $\|A_1^{-1}\|_2 < \|D'L'(L')^T D'\|_2$, where

$$L' = \begin{bmatrix} 1 & & \\ & 1 & \\ & & \dfrac{1}{L_{33}} \end{bmatrix} \begin{bmatrix} 1 & & \\ & \dfrac{1}{L_{22}} & \\ & 1 & 1 \end{bmatrix} \begin{bmatrix} \dfrac{1}{L_{11}} & & \\ 1 & 1 & \\ 1 & & 1 \end{bmatrix},$$

and $D' = \mathrm{diag}\,(\sqrt{3}L_{11}, \sqrt{2}L_{22}, L_{33})$. Therefore

$$D'L'(L')^T D' = D' \begin{bmatrix} \dfrac{1}{L_{11}^2} & \dfrac{1}{L_{11}L_{22}} & \dfrac{2}{L_{11}L_{33}} \\ & \dfrac{2}{L_{22}^2} & \dfrac{3}{L_{22}L_{33}} \\ sym. & & \dfrac{6}{L_{33}^2} \end{bmatrix} D' = \begin{bmatrix} 3 & \sqrt{6} & 2\sqrt{3} \\ \sqrt{6} & 4 & 3\sqrt{2} \\ 2\sqrt{3} & 3\sqrt{2} & 6 \end{bmatrix} ,$$

and

$$\|A_1^{-1}\|_2 < \mathrm{Tr}\,(A) = 13 .$$

Alltogether we have

$$\kappa(A_1) < 13(1 + \sqrt{2/3} + \sqrt{1/2}) \approx 32.81.$$

(The bound (4.4.4) for $n = 3$ is $18e \approx 48.96$.)

Theorem 4.4.1 *Let H be a real symmetric positive definite matrix of order n, and let $PHP^T = LL^T$ be its Cholesky decomposition with complete pivoting. Let $A_1 = D_1^{-1} L^T L D_1^{-1}$, where*

$$D_1 = \mathrm{diag}\,(D_{1,ii}, \ldots, D_{1,nn}) = \mathrm{diag}\,(L^T L) \ , \quad D_{1,ii} = \left(\sum_{k=i}^{n} L_{ki}^2 \right)^{1/2} .$$

Then

$$\kappa(A_1) < \left(1 + \sum_{i=1}^{n-1} \sqrt{\frac{i}{i+1}} \right) \sum_{i=1}^{n} \left(1 + \frac{2^{2(i-1)} - 1}{3} \right) (n + 1 - i) . \qquad (4.4.7)$$

PROOF. Reasoning as we did in the 3×3 example, we conclude that $\|A_1\|_2 < \|A'\|_2$ where

$$A' = \begin{bmatrix} 1 & \sqrt{\frac{n-1}{n}} & \sqrt{\frac{n-2}{n-1}} & \cdots & \sqrt{\frac{1}{2}} \\ & 1 & \sqrt{\frac{n-2}{n-1}} & \cdots & \sqrt{\frac{1}{2}} \\ & & 1 & \cdots & \vdots \\ & & & \ddots & 1 \end{bmatrix}.$$

Therefore

$$\|A_1\|_2 < 1 + \sum_{i=1}^{n-1} \sqrt{\frac{i}{i+1}}, \tag{4.4.8}$$

which proves the first part of (4.4.7). As in the 3×3 example, we also conclude that

$$\|A_1^{-1}\|_2 = \|D_1 L^{-1} L^{-T} D_1\|_2 \le \|D' L' (L')^T D'\|_2,$$

where

$$\begin{aligned} D' &= \operatorname{diag}(\sqrt{n}, \sqrt{n-1}, \ldots, \sqrt{2}, 1), \\ L' &= L^{(n)} L^{(n-1)} \cdot \ldots \cdot L^{(1)}, \\ L_{jk}^{(i)} &= \begin{cases} 1, & j = k \ , \\ 1, & k = i, \ j = i+1, \ldots, n \ , \\ 0, & \text{otherwise} \ . \end{cases} \end{aligned}$$

For the elements of the matrix L' we now have

$$L'_{ij} = \begin{cases} 1, & i = j \ , \\ 2^{i-1-j}, & i > j \ , \\ 0, & i < j \ . \end{cases}$$

Set $B = L'(L')^T$ and $C = D'L'(L')^T D' = D'BD'$. Then

$$\begin{aligned} B_{ii} &= 1 + \sum_{j=1}^{i-1} (L'_{ij})^2 = 1 + \sum_{j=1}^{i-1} 2^{2(i-1-j)} = 1 + \sum_{k=0}^{i-2} 2^{2k} \\ &= 1 + \frac{2^{2(i-1)} - 1}{3} \ , \qquad\qquad i = 1, \ldots, n \ , \end{aligned}$$

and

$$\begin{aligned} B_{ij} &= L'_{ij} + \sum_{k=1}^{j-1} L'_{ik} L'_{jk} = 2^{i-1-j} + \sum_{k=1}^{j-1} 2^{i-1-k} 2^{j-1-k} \\ &= 2^{i-1-j} + 2^{i+j-2} \left(\sum_{k=0}^{j-1} 2^{-2k} - 1 \right) \\ &= 2^{i-j} \left(\frac{1}{2} + \frac{2^{2(j-1)} - 1}{3} \right) \ , \quad i = 1, \ldots, n; \ i > j \ . \end{aligned}$$

Of course, $B_{ij} = B_{ji}$. Furthermore,

$$C_{ij} = B_{ij}\sqrt{n+1-i}\sqrt{n+1-j} \ .$$

Finally, $\|C\|_2 < \mathrm{Tr}\,(C)$ and the second part of (4.4.7) is proved. Q.E.D.

In Th. 4.4.1 we have essentially proved that for any positive definite matrix H, the value $\kappa(A_1)$ is smaller than the product $\|A'\|_2\|C\|_2$. We must, however, emphasize that the second (dominant) part of (4.4.7) is a very good approximation for $\|C\|_2$ in the sense that for all n

$$\left(\sum_{i=1}^{n}\left(1+\frac{2^{2(i-1)}-1}{3}\right)(n+1-i)\right)/\|C\|_2 < 1.0001 \ .$$

We can further symplify the inequality (4.4.7) by bounding $\|A'\|_2$ by n and $\|C\|_2$ by

$$\int_1^{n+1}\left(1+\frac{2^{2(x-1)}-1}{3}\right)(n+1-x)dx < \frac{1}{3}\left(n^2+\frac{1}{\ln^2 4}2^{2n}\right) \ ,$$

which yields

$$\kappa(A_1) < \frac{n}{3}\left(n^2+\frac{1}{\ln^2 4}2^{2n}\right) \ . \tag{4.4.9}$$

We have experimentally observed that

$$\frac{1}{3}\left(n^2+\frac{1}{\ln^2 4}2^{2n}\right)/\|C\|_2 < 1.1708 \ .$$

Now we show that the transition from the matrix H to the matrix L^TL cannot spoil the condition of the scaled matrix too much. We use the technique from [14]. Set $A = \mathrm{Scal}\,(PHP^T) = D^{-1}LL^TD^{-1}$, $B = D^{-1}L$, and $B_1 = LD_1^{-1}$. Then $A = BB^T$, $A_1 = B_1^TB_1$ and

$$B_1^{-1} = D_1L^{-1} = D_1B^{-1}D^{-1} \ .$$

From (4.4.1) for every $1 \le j \le i \le n$ it follows

$$\begin{aligned} |D_1B^{-1}D^{-1}|_{ij} &= \sqrt{\frac{L_{ii}^2+L_{i+1,i}^2+\cdots+L_{ni}^2}{L_{j1}^2+L_{j2}^2+\cdots+L_{jj}^2}}\,|B^{-1}|_{ij} \\ &\le \sqrt{n-i+1}\,\frac{L_{ii}}{L_{jj}}\,|B^{-1}|_{ij} \le \sqrt{n-i+1}\,|B^{-1}|_{ij} \ . \end{aligned}$$

Thus, (4.4.5) and (4.4.6) imply

$$\|B_1^{-1}\|_2 \le \||D_1B^{-1}D^{-1}|\|_2 \le \sqrt{n}\||B^{-1}|\|_2 \le n\|B^{-1}\|_2 \ ,$$

that is, $\|A_1^{-1}\|_2 \le n^2\|A^{-1}\|_2$.

The bound (4.4.7) is almost attained for the matrices of the form $H = LL^T$, where

$$\begin{aligned} L &= L_0 D_0 \ , \\ D_0 &= \operatorname{diag}(1, s, s^2, \ldots, s^{n-1}) \ , \\ (L_0)_{ij} &= \begin{cases} 1, & i = j \ , \\ -c, & i > j \ , \\ 0, & i < j \ . \end{cases} \\ s^2 + c^2 &= 1 \ . \end{aligned} \tag{4.4.10}$$

These matrices are due to Kahan and are described in [16]. When $c \to 1$, then H and $\operatorname{Scal}(H)$ both tend to singular matrices. Since $H_1 = L^T L = D_0 L_0^T L_0 D_0$ and L is itself the optimal Cholesky factor of H, we conclude that

$$A_1 = \operatorname{Scal}(H_1) = \operatorname{Scal}(L_0^T L_0) = D_1^{-1} L_0^T L_0 D_1^{-1} \ ,$$

where

$$D_1 = \operatorname{diag}\left(\sqrt{1+(n-1)c^2}, \sqrt{1+(n-2)c^2}, \ldots, 1\right) \ .$$

It is easy to verify that $\lim_{c\to 1} A_1^{-1} = C$. Therefore, the quotient between the bound (4.4.7) and $\kappa(A_1)$ is in this case equal to $\|A'\|_2/\|A_1\|_2$ which is smaller than the first part of (4.4.7) (smaller than n).

At the end we have to point out, even though the bound of Th. 4.4.1 may seem pessimistic, experiments from Demmel and Veselić [13] and Chap. 5 show that $\kappa(A_1)$ is in practice *considerably* better than $\kappa(A)$ and, thus, the examples like that of Kahan are very rare. Moreover, for the matrices defined by (4.4.10) it is possible to obtain much better $\kappa(A_1)$. Since $H_{ii} = 1$, the optimal Cholesky decomposition requires no pivoting. However, permuting the matrix H so that e.g. H_{nn} comes to the position (1,1) does not contradict the complete pivoting and results in $\kappa(A_1) < n^2$. Demmel and Veselić [13] showed that for positive definite matrix H

$$\lambda_{min}(A) \leq \frac{H_{ii}}{\lambda_i} \leq \lambda_{max}(A) \ ,$$

where $A = \operatorname{Scal}(H)$, λ_i denotes the $i-$th eigenvalue of H, and H_{ii}'s and λ_i's have the same ordering. This means that the diagonal entries of H can differ from the eigenvalues only by factors bounded by $\kappa(A)$. Applying this result to $H_1 = L^T L$, wee see that the Cholesky decomposition usually has rank–revealing property. The complete pivoting usually gives satisfactory results, but the choice of the optimal pivoting as in the above example in an open problem. For related results about the rank–revealing QR decomposition see [7].

The following theorem holds for a non–singular but possibly indefinite H:

Theorem 4.4.2 *Let H be a nonsingular symmetric matrix and let $PHP^T = GJG^T$ be its decomposition. Let $\mu = 2.781$ denote the maximal value of the quantities $|L_{ij}|$ from (4.1.8), and let $A_1 = \operatorname{Scal}(G^T G)$. Then*

$$\kappa(A_1) < n(1 + 15n)3.781^{2n} \ . \tag{4.4.11}$$

PROOF. From (4.1.9) it follows

$$\begin{aligned}
\|A_1\|_2 &= \|\mathrm{Scal}\,(G^T G)\|_2 = \|\mathrm{Scal}\,(P_1^T \sqrt{|D_1|} L_1^T L_1 \sqrt{|D_1|} P_1)\|_2 \\
&= \|\mathrm{Scal}\,(\sqrt{|D_1|} L_1^T L_1 \sqrt{|D_1|})\|_2 \\
&= \|\mathrm{Scal}\,(L_1^T L_1)\|_2 = \|D^{-1} L_1^T L_1 D^{-1}\|_2 \ ,
\end{aligned}$$

where D is diagonal with elements $D_{ii} = (L_1^T L_1)_{ii} = (Q^T L^T L Q)_{ii}$. Note that in estimating $\|A_1\|_2$ and $\|A_1^{-1}\|_2$ we can without loss of generality assume that $P_1 = I$. The matrix A_1 is positive definite and has unit diagonal, so that

$$\|A_1\|_2 < n \ . \tag{4.4.12}$$

Further,

$$\|A_1^{-1}\|_2 = \|D L_1^{-1} L_1^{-T} D\|_2 = \|D Q^T L^{-1} L^{-T} Q D\|_2 \ .$$

Now we shall maximize elements of the matrices D and $Q^T L^{-1}$ and use the monotonicity properties of the norm $\|\cdot\|_2$ as we did in Th. 4.4.1. The elements of L^{-1} are largest in modulus if all under–diagonal elements of L are equal to $-\mu$. Let us denote this "maximal" L^{-1} by $\bar{L}$. Then

$$\bar{L}_{ij} = \begin{cases} 1, & i = j \ , \\ \mu(1+\mu)^{i-1-j}, & i > j \ , \\ 0, & i < j \ . \end{cases}$$

Now

$$|Q^T L^{-1}| \le |Q^T| \bar{L} \le L' \ ,$$

where

$$L'_{ij} = \begin{cases} 1+\mu, & i = j \ , \\ \mu(2+\mu)(1+\mu)^{i-1-j}, & i > j \ , \\ 1, & i = j-1 \ , \\ 0, & i < j-1 \ . \end{cases}$$

Element D_{ii} is the norm of the i–th column of LQ. It is easy to verify that

$$D_{ii} = \sqrt{1 + L_{i+1,i}^2 + \dots L_{ni}^2} \ ,$$

when the index i corresponds to a 1×1 pivot, and

$$\begin{aligned}
D_{ii} &= \sqrt{cs^2 + sn^2 + (L_{i+2,i} cs - L_{i+2,i+1} sn)^2 + (L_{ni} cs - L_{n,i+1} sn)^2} \ , \\
D_{i+1,i+1} &= \sqrt{cs^2 + sn^2 + (L_{i+2,i+1} cs + L_{i+2,i} sn)^2 + (L_{n,i+1} cs + L_{ni} sn)^2} \ ,
\end{aligned}$$

when the indices $i, i+1$ correspond to a 2×2 pivot. Therefore, it is always $D_{ii} \le D'_{ii}$, where D' is diagonal matrix with elements

$$D'_{ii} = \sqrt{1 + 2(n-i)\mu^2} \ .$$

Now we have

$$
\begin{aligned}
\|A_1^{-1}\|_2 &= \|DQ^T L^{-1} L^{-T} Q D\|_2 \leq \|D' L' (L')^T D'\|_2 \leq \operatorname{Tr}(D' L' (L')^T D') \\
&= \sum_{i=1}^{n} \left[1 + (1+\mu)^2 + \sum_{j=1}^{i-1} \left(\mu (1+\mu)^{(i-1-j)} (2+\mu) \right)^2 \right] (1 + 2(n-i)\mu^2) \\
&= \sum_{i=1}^{n} \left[1 + \mu(2+\mu)((1+\mu)^{2(i-1)} - 1) \right] (1 + 2(n-i)\mu^2) \\
&\leq \mu(2+\mu)(1+2n\mu^2) \sum_{i=1}^{n} (1+\mu)^{2(i-1)} \\
&\leq (1+2n\mu^2)(1+\mu)^{2n} ,
\end{aligned}
$$

which completes the proof of the theorem. Q.E.D.

Due to the fact that some of the worst cases assumed in the above proof are impossible, the statement of Th. 4.4.2 is an overestimate. Numerical experiments of Chap. 5 show that $\kappa(\operatorname{Scal}(G^T G))$ is, as in the positive definite case, generally very small.

If H is singular, then Alg. 4.1.1 returns an $n \times r$ matrix G of the full column rank. The nature of the proof of Th. 4.4.2 implies that (4.4.11) holds in this case, too (and that even with better constants, since some summations have fewer terms).

Chapter 5

Numerical experiments

In this chapter, we present the results of our numerical experiments. Briefly, we tested the algorithm of Sect. 4.3 and verified that error bounds of that section held in all examples. The comparison of our algorithms with the QR and the standard Jacobi algorithm showed that our algorithms are uniformly more accurate. In fact, the performance is better than we were able to explain theoretically, both because we could observe little or no growth in actual errors for increasing dimension, and because of small values attained by $\max_m \kappa(B_m)/\kappa(B_0)$ during the Jacobi part. The relative errors in eigenvalues were given by (4.3.4) multiplied by small coefficients which increased only slowly with n. The norm errors in eigenvectors were smaller than those predicted by (4.3.4) by an order of magnitude .

Tests were performed using FORTRAN on an IBM RISC/6000. The arithmetic is IEEE arithmetic with machine precision $\varepsilon_S \approx 5.9604 \cdot 10^{-8}$ in single, and $\varepsilon_D \approx 1.1102 \cdot 10^{-16}$ in double precision. Overflow/underflow tresholds are approximately $10^{\pm 38}$ in single, and $10^{\pm 308}$ in double precision. The machine has a special multiply–and–add function, *maf*, which computes $a = b + c * d$ as a single instruction. In single precision, *maf* first computes $c*d$ in double precision, adds b, and then rounds a back to single precision. For IEEE arithmetic with *maf*, the constants 272 and C_m from (4.3.2) are somewhat, but not essentially, smaller.

In our tests we used five different algorithms:

JGJ – the symmetric indefinite decomposition of Alg. 4.1.1 followed by the standard implicit method of Alg. 3.3.1,

JGJF – the symmetric indefinite decomposition of Alg. 4.1.1 followed by the fast implicit method of Alg. 3.4.1,

JGJFS – the symmetric indefinite decomposition of Alg. 4.1.1 followed by the fast implicit method with self–scaling rotations of Alg. 3.4.5,

JAC – the standard Jacobi algorithm (We used Alg. 3.1.1 with $J = I$. Then no hyperbolic rotations are performed and H does not have to be positive definite.), SSYEV – LAPACK single precision routine which implements tridiagonalization followed by QR iteration.

In all three implicit Jacobis the diagonal was kept separately according to Alg. 3.3.7.

We tested the accuracy as follows: we considered real symmetric non–singular eigenproblems. We first solved every problem using JGJ and JAC in double precision. We assumed that the digits of the computed eigenvalues which overlap in those two algorithms are correct. We took the eigenvectors computed by JGJ as the ones of reference. Then we solved the same problem with the single precision versions of JGJ, JGJF and JGJFS, and compared the answers with the double precision solution to see if they were as accurate as predicted (which they were). We also compared the solutions obtained by SSYEV and the single precision version of JAC. Absolutely small eigenvalues computed by SSYEV were often of the wrong sign, indicating total loss of relative accuracy. All Jacobi algorithms used the stopping criterion $tol = n \cdot \varepsilon$ and the parallel cyclic pivot strategy of [24].

The rest of the chapter is organized as follows: we first discuss the test matrix generation. We then discuss accuracy of the computed eigensolutions. We make an interesting remark about the sensitivity of the QR and the standard Jacobi algorithms to the initial permutations of the input matrix. After that we discuss behaviour of $\lambda_{min}(D_G^{-1} G_M G_M^T D_G^{-1})$, growth of $1/\sigma_{min}(B_m)$ during the implicit Jacobi process, and behaviour of the diagonal in fast rotations. Finally, we discuss convergence rates.

Test matrix generation. We generated two types of random matrices. The first type is divided in several categories according to dimension n, $\kappa(\widehat{A})$ (where $\widehat{A}_{ii} \equiv 1$, so that $\kappa(\widehat{A})$ is at most factor n from $C(A, \widehat{A})$ from (2.2.12)), and $\kappa(H)$. We first describe the algorithm used to generate a random matrix from these parameters and then the sets of parameters used. All steps were preformed in double precision. Given $\kappa(\widehat{A})$, we generated a positive definite diagonal matrix D whose entries' logarithms are uniformly distributed between $[-0.5 \log \kappa(\widehat{A}), 0.5 \log \kappa(\widehat{A})]$. On D we applied five sweeps of random trigonometric plane rotations, thus obtaining matrix A_0. On A_0 we applied five sweeps of the "anti–Jacobi" method, thus obtaining matrix $\bar{A}$. This method, due to Veselić, consists of an iterative application of trigonometric plane rotations, $A_{m+1} = J_m^T A_m J_m$, where J_m is obtained in the following manner: let

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} , \qquad \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix} ,$$

be the pivot submatrices of A_m and J_m, respectively. Then $cs = 1/h$ and $sn = -t/h$, where

$$\zeta = \frac{2c}{b-a} , \quad t = \frac{\text{sign } \zeta}{|\zeta| + \sqrt{1+\zeta^2}} , \quad h = \sqrt{1+t^2} .$$

The sequence of matrices obtained by the anti–Jacobi method converges to a matrix A where $A_{ii} \equiv \text{Tr}\, D/n$, i.e. $\kappa(\text{Scal}\,(A)) = \kappa(A)$. The convergence is very slow. It often required 50 or more sweeps for $n = 30$. However, after five sweeps $\kappa(\bar{A})$ and $\kappa(A) = \kappa(D)$ differ by no more than 10 %. Given $\kappa(H)$, we generated a positive definite diagonal matrix D_1 whose entries' logarithms are uniformly distributed between $[-0.5 \log \kappa(H), 0.5 \log \kappa(H)]$, and formed a positive definite matrix $\bar{H} = D_1 \bar{A} D_1$. We then calculated the eigendecomposition $\bar{H} = U^T \bar{\Lambda} U$ by our algorithm, and changed some randomly selected eigenvalues into negative ones, thus obtaining matrix Λ. Our random test matrix was then $H = U^T \Lambda U$.

The values for $\kappa(\widehat{A})$ were 10, 10^2 and 10^3, the values for $\kappa(H)$ were 10^2, 10^5, 10^9, 10^{14} and 10^{20}, and the values for n were 10, 20, 50, 100 and 200. This makes a total of $3 \times 5 \times 5 = 75$ different classes of matrices. In each class of dimension $n = 10$ matrices we generated 500 random matrices, in each class of $n = 20$ we generated 300 random matrices, in each class of $n = 50$ we generated 200 random matrices, in each class of $n = 100$ we generated 100 random matrices, and in each class of $n = 200$ we generated 50 random matrices. This makes a total of 17250 different test matrices.

The second type of test matrices were block scaled diagonally dominant (b.s.d.d) matrices of Th. 2.2.7 generated according to two parameters, dimension n and $\kappa(H)$. We first randomly generated number of diagonal blocks $2 \leq n_b \leq n-1$, and the size of the blocks. We then generated a random symmetric orthogonal matrix A with this block structure (the elements outside blocks are 0), and formed matrix $\bar{A} = A + N$, were N is a random symmetric matrix with $\|N\|_2 \leq 0.5$. Given $\kappa(H)$, we generated a positive definite diagonal matrix D whose entries' logarithms are uniformly distributed between $[-0.5 \log \kappa(H), 0.5 \log \kappa(H)]$. D is constant on the blocks which correspond to the blocks of A, so that A and D commute. Finally, we formed our test matrix $H = D\bar{A}D$. As above, we have chosen $\kappa(H) \in \{10^2, 10^5, 10^9, 10^{14}, 10^{20}\}$ and $n \in \{10, 20, 50, 100, 200\}$. In each class of dimension $n = 10$ and $n = 20$ matrices we generated 100 random matrices, in each class of $n = 50$ we generated 50 random matrices, in each class of $n = 100$ we generated 30 random matrices, and in each class of $n = 200$ we generated 10 random matrices.

Accuracy of the computed eigensolution. For every matrix we first calculated expected relative error in eigenvalues and expected norm error in eigenvectors according to (4.3.4) with $\varepsilon = \varepsilon_S = 5.9604 \cdot 10^{-8}$. For every eigenvalue we calculated relative error

$$\frac{|\lambda_{D,i} - \lambda_{S,i}|}{|\lambda_{D,i}|},$$

where $\lambda_{D,i}$ denotes the $i-$th reference eigenvalue, and $\lambda_{S,i}$ denotes the $i-$th single precision eigenvalue. For every eigenvector we calculated the error $\|v_{D,i} - v_{S,i}\|_2$, where $v_{D,i}$ and $v_{S,i}$ are the eigenvectors corresponding to $\lambda_{D,i}$ and $\lambda_{S,i}$, respectively. Table 1 shows quotients of the maximum of the relative errors in single precision eigenvalues and the expected relative error of (4.3.4). For all quantities we give mean value, standard deviation, maximum and minimum attained on the respective

	Table 1: $\dfrac{\max_i\{|\lambda_{D,i}-\lambda_{S,i}|/|\lambda_{D,i}|\}}{\text{expected relative error}}$				
n		MEAN	STD	MAX	MIN
10	JGJ	1.551	1.342	6.710	.0676
	JGJF	1.562	1.372	7.658	.0554
	JGJFS	1.225	1.024	6.347	.0565
20	JGJ	2.267	2.137	10.53	.1105
	JGJF	2.330	2.199	10.32	.1231
	JGJFS	1.618	1.509	8.216	.0984
50	JGJ	4.282	4.165	17.01	.2256
	JGJF	4.355	4.282	18.34	.2332
	JGJFS	2.737	2.625	11.14	.1872
100	JGJ	6.653	6.528	26.56	.3609
	JGJF	6.803	6.721	27.45	.3595
	JGJFS	4.191	4.168	20.06	.2357
200	JGJ	12.13	11.53	38.97	.9087
	JGJF	12.26	11.60	39.11	.9693
	JGJFS	7.546	7.239	25.62	.5904

class of test matrices. We see that the expectations were fulfilled up to a slowly growing constant, thus the statements of Remarks 3.2.6 and 4.2.2 that the actual errors increase only slowly as n or M increases. Note that the quotients in Table 1 increase at most linearly in n, which is still far below the theoretical growth of $O(n^2)$ from (4.3.2). Comparing the data for JGJ and JGJF indicates that the use of *maf* makes no difference in practice (*maf* is theoretically fully exploited by fast rotations in JGJF, and only partially exploited in JGJ). Note that JGJFS is slightly more accurate than JGJ and JGJF.

Table 2 shows quotients of the maximum of the norm errors in single precision eigenvectors and the expected norm error. We see that the actual errors are considerably smaller than the expected ones, for which we have no explanation. Note, also, that the quotients are almost independent of n, and that JGJFS is now somewhat less accurate than JGJ and JGJF.

Table 3 shows quotients between maximal relative errors in eigenvalues of SSYEV (JAC) and JGJFS. We see that SSYEV and JAC often had no accurate digits, and are therefore unreliable. SSYEV and JAC performed as well or even slightly better than our algorithms on those matrices for which parameter $\kappa(H)$ was small, i.e. on the matrices where our perturbation theory and the standard one do not differ much.

Tables 1, 2 and 3 are obtained from the first type of test matrices. Data for b.s.d.d matrices are similar, except that JAC is for those matrices as accurate as our algorithms.

Table 2: $\dfrac{\max_i \|v_{D,i} - v_{S,i}\|_2}{\text{expected norm error}}$					
n		MEAN	STD	MAX	MIN
10	JGJ	.0144	.0106	.0895	.0002
	JGJF	.0147	.0118	.1258	.0003
	JGJFS	.0149	.0111	.0945	.0003
20	JGJ	.0138	.0120	.1095	.0008
	JGJF	.0145	.0133	.1099	.0009
	JGJFS	.0159	.0144	.1112	.0002
50	JGJ	.0168	.0152	.1056	.0004
	JGJF	.0181	.0169	.1018	.0007
	JGJFS	.0230	.0232	.1364	.0014
100	JGJ	.0177	.0175	.1397	.0008
	JGJF	.0195	.0197	.1356	.0010
	JGJFS	.0285	.0292	.1938	.0012
200	JGJ	.0198	.0191	.0808	.0001
	JGJF	.0231	.0223	.1045	.0003
	JGJFS	.0365	.0349	.1467	.0011

Table 3: Quotients of maximal relative errors in eigenvalues					
n		MEAN	STD	MAX	MIN
10	SSYEV/JGJFS	$6.6 \cdot 10^5$	$9.3 \cdot 10^5$	$4.5 \cdot 10^6$	.1687
	JAC/JGJFS	$1.0 \cdot 10^4$	$1.3 \cdot 10^5$	$3.1 \cdot 10^6$	.1055
20	SSYEV/JGJFS	$4.2 \cdot 10^5$	$4.9 \cdot 10^5$	$2.1 \cdot 10^6$	.1812
	JAC/JGJFS	$4.8 \cdot 10^4$	$1.6 \cdot 10^5$	$1.2 \cdot 10^6$	.1282
50	SSYEV/JGJFS	$2.2 \cdot 10^5$	$2.1 \cdot 10^5$	$8.3 \cdot 10^5$	.1136
	JAC/JGJFS	$1.2 \cdot 10^5$	$1.9 \cdot 10^5$	$7.6 \cdot 10^5$	.1595
100	SSYEV/JGJFS	$1.2 \cdot 10^5$	$1.1 \cdot 10^5$	$4.5 \cdot 10^5$	.0631
	JAC/JGJFS	$1.0 \cdot 10^5$	$1.1 \cdot 10^5$	$4.5 \cdot 10^5$	.1608
200	SSYEV/JGJFS	$4.1 \cdot 10^4$	$4.4 \cdot 10^4$	$1.4 \cdot 10^5$	.0553
	JAC/JGJFS	$3.7 \cdot 10^4$	$4.4 \cdot 10^4$	$1.4 \cdot 10^5$	.1877

Table 4: $\dfrac{\lambda_{min}(\widehat{A})}{\lambda_{min}(D_G^{-1}G_M G_M^T D_G^{-1})}$					
n		MEAN	STD	MAX	MIN
10	TYPE 1	1.216	.2970	3.076	0.9166
	TYPE 2	2.742	1.249	6.000	1.100
20	TYPE 1	1.412	.1665	4.411	.9696
	TYPE 2	3.816	1.505	8.300	1.100
50	TYPE 1	1.821	.6617	5.000	1.000
	TYPE 2	6.944	3.318	17.00	1.100
100	TYPE 1	2.347	.9997	5.588	1.200
	TYPE 2	12.12	7.186	25.00	1.500
200	TYPE 1	3.522	1.654	7.272	1.608
	TYPE 2	20.85	8.900	37.00	6.500

Remark. We have observed that the QR and the standard Jacobi algorithm often improved in accuracy when the starting matrix was permuted so that the symmetric indefinite decomposition needs no permutations. In many cases even the accuracy of our algorithms was achieved. This phenomenon in an interesting open problem, and can serve as an empirical advice to someone using QR or the standard Jacobi.

Behaviour of $1/\lambda_{min}(D_G^{-1}G_M G_M^T D_G^{-1})$. Table 4 displays values of

$$\frac{\lambda_{min}(\widehat{A})}{\lambda_{min}(D_G^{-1}G_M G_M^T D_G^{-1})} ,$$

where the denominator comes from (4.3.4), and $\widehat{A} = (\mathrm{diag}\,|H|)^{-1/2}|H|(\mathrm{diag}\,|H|)^{-1/2}$. We see that the quotients are small, thus implying that the errors induced by the symmetric indefinite decomposition satisfy the perturbation bounds of Sect. 2.2 almost optimally. The same values were obtained by all three of our algorithms. There are small differences between test matrices of the first and the second type.

Behaviour of $1/\sigma_{min}(B_m)$. Let $G_m = B_m D_m$ denote the sequence of matrices which was obtained by the implicit Jacobi from the starting pair G_0, J. As usual, the columns of B_m have unit norms. Also, let $A_m = D_m^{-1}G_m^T G_m D_m^{-1}$. We calculated upper bounds for $\max_m \sigma_{min}(B_0)/\sigma_{min}(B_m)$ in two ways. Table 5 gives four values: SIGMA, HAD/SIGMA, BOUND and ROT. Here

$$\mathrm{SIGMA} = 1/\sigma_{min}(B_0) , \qquad \mathrm{HAD} = (\exp(1)/\det(A_0))^{1/2} .$$

BOUND and ROT were computed as follows: we computed a decreasing sequence h_m as

$$h_0 \;=\; \mathrm{HAD}^2 ,$$

$$h_{m+1} = h_m(1 - A^2_{m,ij}) , \qquad m \geq 0 .$$

Each sweep of the parallel pivot strategy of [24] has n parallel steps each having $p = (n-1)/2$ rotations for n odd, and $n-1$ parallel steps each having $p = n/2$ rotations for n even. We computed a non–decreasing sequence s_m defined by

$$\begin{aligned}
s_0 &= \text{SIGMA}^2 , \\
s_m &= s_{m-1} , \qquad m \geq 1 , \quad m \bmod p \neq 0 , \\
s_m &= s_{m-p}(1 + \max_{0 \leq k \leq p-1} |A_{m-p+k,ij}|) , \qquad m \geq 1 , \quad m \bmod p = 0 .
\end{aligned}$$

Recursive application of (3.2.34) implies that $1/\sigma^2_{min}(B_m) \leq s_m$. Recursive application of (3.2.36), together with (3.2.35), implies that $1/\sigma^2_{min}(B_m) \leq h_m$. Therefore, $1/\sigma^2_{min}(B_m) \leq \min\{s_m, h_m\}$ for every $m \geq 0$. Also, $s_0 \leq h_0$. Let m' be the largest m such that $s_m \leq h_m$. Then

$$\text{BOUND} = (s_m/s_0)^{1/2} , \qquad \text{ROT} = m' .$$

In other words, BOUND is the guaranteed upper bound for $\max_m \sigma_{min}(B_0)/\sigma_{min}(B_m)$.

The values of $1/\sigma_{min}(B_0)$ in Table 5 are very small, thus showing the non–trivial diagonalizing effect of the transition from matrix H to pair G^TG, J. We also see that the guaranteed upper bound is reliable only for smaller dimensions, and that s_m and h_m usually meet in the first sweep. The data of Table 5 come from test matrices of the first type. Data for b.s.d.d matrices are similar.

A much better upper bound for $\max_m \sigma_{min}(B_0)/\sigma_{min}(B_m)$ was obtained by the algorithm of Sect. 3.2.2 (which, however, requires additional computational effort). This bound is by its nature always greater or equal $\sqrt{n}$, and the largest value attained *in all experiments* was $1.05\sqrt{n}$. In fact, accuracy of computed eigensolutions implies that this is also an overestimate, that is, $1/\sigma_{min}(B_m)$ can grow only little before converging to 1.

Behaviour of the diagonal in fast rotations. Table 6 shows four values: MINF is the smallest element of the diagonal of fast rotations obtained by JGJF, MINF/MINS is the quotient of this element and the smallest element of the diagonal of fast self–scaling rotations obtained by JGJFS, MAXF is the largest element of the diagonal of JGJF, and MAXF/MAXS is the quotient of this element and the largest element of the diagonal of JGJFS. We see that, even for large n, there is actually no danger of underflow/overflow.

Convergence rates. We compared computing times of JGJF and SSYEV, computing times of JGJ and JGJFS, and number of sweeps and rotations of JAC and JGJF. The speed ratio of JGJF and SSYEV is the following: for $n = 200$, mean value, standard deviation, maximum and minimum are for matrices of the first type $(4.9, 0.5, 5.8, 3.6)$, and for b.s.d.d matrices $(4.9, 0.8, 6.4, 3.3)$. These ratios are realistic

Table 5: Behaviour of $1/\sigma_{min}(B_m)$					
n		MEAN	STD	MAX	MIN
10	SIGMA	1.940	.7408	5.193	1.032
	HAD/SIGMA	2.014	1.103	10.86	1.217
	BOUND	1.331	.2513	2.877	1.649
	ROT	23	10	60	5
20	SIGMA	2.813	1.249	9.481	1.130
	HAD/SIGMA	29.92	87.97	100.2	1.277
	BOUND	2.606	1.771	14.06	1.649
	ROT	85	28	170	10
50	SIGMA	4.696	2.707	14.65	1.524
	HAD/SIGMA	$1.0 \cdot 10^{10}$	$1.4 \cdot 10^{11}$	$2.9 \cdot 10^{12}$	2.182
	BOUND	330.1	784.8	550.5	1.649
	ROT	653	223	1175	75
100	SIGMA	7.146	4.654	23.07	2.003
	HAD/SIGMA	$3.2 \cdot 10^{31}$	$6.9 \cdot 10^{32}$	$1.4 \cdot 10^{34}$	57.41
	BOUND	$4.1 \cdot 10^{9}$	$8.6 \cdot 10^{10}$	$1.8 \cdot 10^{12}$	9.220
	ROT	3247	1251	15500	105

Table 6: Behaviour of the diagonal in fast rotations					
n		MEAN	STD	MAX	MIN
100	MINF	.2839	.1869	.7100	.0051
	MINF/MINS	.4150	.2584	.9838	.0086
	MAXF	1.323	.1076	1.700	1.100
	MAXF/MAXS	.9633	.0767	1.230	.7857
200	MINF	.0876	.0855	.3300	.0005
	MINF/MINS	.1418	.1352	.5409	.0009
	MAXF	1.439	.1158	1.900	1.300
	MAXF/MAXS	1.028	.0827	1.357	.9285

although JGJF could be made slightly faster. Namely, SSYEV uses BLAS routines which are distributed together with RISC/6000 (and are therefore highly optimized), while our algorithm uses some extra BLAS type routines written by us (e.g. hyperbolic plane rotation).

Use of fast rotations, JGJFS, brought only about 5% speed up over JGJ.

We begin the comparison of sweeps and rotations needed for convergence of JAC and JGJF with a few details. JAC stopped when the last $n(n-1)/2$ stopping tests $|H_{ij}| \leq tol\sqrt{|H_{ii}||H_{jj}|}$ succeeded. Since our implicit algorithms keep the diagonal in a separate vector, JGJF stopped after an empty sweep. Since one scalar product is needed to determine $(G^T G)_{ij}$ even if no rotation is performed, an empty sweep in JGJF requires approximately 1/3 of the computation time of the full sweep, which is a slight dissadvantage. The symmetric indefinite decomposition used in JGJF amounts to no more than 2/9 of one sweep and is neglected. Table 7 shows number of sweeps and rotations for JAC and JGJF, and quotient of numbers of rotations for JAC and JGJF.

We see that JAC needed averagely twice as much rotations as JGJF. Another important phenomenon, not readily seen in this table, is that number of rotations in JGJF is somewhat stable, that is, it did not depend much on parameters $\kappa(\widehat{A})$ and $\kappa(H)$, while in JAC number of rotations grew as $\kappa(H)$ grew. Data in Table 7 come from matrices of the first type. For b.s.d.d matrices, JAC performs better, that is, it needs averagely 1.5 times more rotations than JGJF.

Table 7: Sweeps and rotations for JAC and JGJF					
n		MEAN	STD	MAX	MIN
10	SWEEP JAC	5.1	.96	9	3
	ROT JAC	166	30	257	105
	SWEEP JGJF	4.1	.65	6	3
	ROT JGJF	107	30	191	43
	JAC/JGJF	1.6	.4	4.2	.98
20	SWEEP JAC	7.2	1.7	12	4
	ROT JAC	935	191	1556	530
	SWEEP JGJF	4.8	.72	7	3
	ROT JGJF	545	152	917	254
	JAC/JGJF	1.8	.6	4.2	1.0
50	SWEEP JAC	10.7	2.7	17	4
	ROT JAC	8305	1740	12719	4089
	SWEEP JGJF	5.7	.92	8	4
	ROT JGJF	4317	1361	7427	2084
	JAC/JGJF	2.1	.9	4.9	.96
100	SWEEP JAC	13.2	2.7	19	6
	ROT JAC	40431	10814	213460	19908
	SWEEP JGJF	6.5	1.1	9	5
	ROT JGJF	20502	7816	92408	9059
	JAC/JGJF	2.2	1.0	5.6	.91
200	SWEEP JAC	14.3	2.7	19	9
	ROT JAC	173952	25326	231892	135121
	SWEEP JGJF	8.0	1.3	10	6
	ROT JGJF	108607	31874	161841	57715
	JAC/JGJF	1.7	.72	3.6	.85

Bibliography

[1] A. A. Anda, H. Park, *Fast Plane Rotations with Dynamic Scaling*, preprint, Computer Science Department, University of Minnesota , Minneapolis, 1990, to appear in SIAM J. Mat. Anal. Appl.

[2] J. Barlow, J. Demmel, *Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices*, SIAM J. Numer. Anal., Vol. 27, No. 3, (762–791) 1990.

[3] J. R. Bunch, *Analysis of the diagonal pivoting method*, SIAM J. Numer. Anal., Vol. 8, No. 4, (656–680) 1971.

[4] J. R. Bunch, L. Kaufmann, *Some Stable Methods for Calculating Inertia and Solving Symmetric Linear Systems*, Math. of Comp., Vol. 31, No. 137, (163–179) 1977.

[5] J. R. Bunch, L. Kaufmann, B. N. Parlett, *Decomposition of a Symmetric Matrix*, Numer. Math., No. 27, (95–109) 1976.

[6] J. R. Bunch, B. N. Parlett, *Direct Methods for Solving Symmetric Indefinite Systems of Linear Equations*, SIAM J. Numer. Anal., Vol. 8, No. 4, (639–655) 1971.

[7] T. F. Chan, *Rank Revealing QR Factorizations*, Linear Algebra and its Appl. 88/89, (67–82) 1987.

[8] A. Deichmöller, *Über die Berechnung verallgemeinerter singulärer Werte mittels Jacobi–ähnlicher Verfahren*, Dissertation, Fernuniversität, Hagen 1991.

[9] J. Demmel, *On Floating Point Errors in Cholesky*, LAPACK Working Note 14, Computer Science Dept. Report, University of Tennessee, Knoxville, October 1989.

[10] J. Demmel, *The inherent inaccuracy of implicit tridiagonal QR*, IMA Preprint Series 963, University of Minnesota, Minneapolis, April 1992.

[11] J. Demmel, W. Gragg, *Accurate Eigenvalues of Acyclic Matrices*, Computer Science Division and Department of Mathematics preprint, University of California, Berkeley, 1992.

[12] J. Demmel, W. Kahan, *Accurate Singular Values of Bidiagonal Matrices*, SIAM J. Sci. Stat. Comp., Vol. 11, No. 5, (873-912) 1990

[13] J. Demmel, K. Veselić, *Jacobi's method is more accurate than QR*, Computer Science Dept. Report 468, Courant Institute, New York, October 1989, to appear in SIAM J. Mat. Anal. Appl.

[14] Z. Drmač, *unpublished manuscript*, 1991.

[15] Z. Drmač, V. Hari, *On quadratic convergence bounds for the J-symmetric Jacobi method*, to appear in Numer. Math.

[16] G. H. Golub, C. F. Van Loan, *Matrix Computations*, University Press, Baltimore and London, 1989.

[17] K. P. Hadeler, *Variationsprinzipien bei nichtlinearen Eigenwertaufgaben*, Arch. Rat Mech. Anal., Vol. 30, (297–307) 1968.

[18] V. Hari, K. Veselić, *On Jacobi methods for singular value decompositions*, SIAM J. Sci. Stat. Comp., Vol. 8, (741–754) 1987.

[19] N. Higham, *Analysis of the Cholesky decomposition of a semi–definite matrix*, in Reliable Numerical Computation, Clarendon Press, eds. M. G. Cox and S. Hammarling, 1990.

[20] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1966.

[21] R. Onn, A. O. Steinhardt, A. Bojanczyk, *Hyperbolic Singular Value Decompositions and Applications*, IEEE Trans. on Acoustics, Speech, and Signal Processing, (1575–1588) July 1991.

[22] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Engelwood Cliffs, NJ, 1980.

[23] H. Rutishauser, *Vorlesungen über numerische Mathematik*, Vol. 1, Birkhäuser Verlag, Basel, Stuttgart, 1976.

[24] A. H. Sameh, *On Jacobi and Jacobi–Like Algorithms for a Parallel Computer*, Math. of Comp., Vol. 25, No. 115, (579–590) 1971.

[25] I. Slapničar, *Upper bound for the condition of the scaled matrix of the symmetric eigenvalue problem*, preprint, Seminarberichte aus dem Fachbereich Mathematik und Informatik Nr. 39, Fernuniversität, Hagen, 1990.

[26] I. Slapničar, *On the growth of the condition of the scaled matrix in Jacobi methods*, preprint, Seminarberichte aus dem Fachbereich Mathematik und Informatik Nr. 41, Fernuniversität, Hagen, 1992.

[27] A. van der Sluis, *Condition Numbers and Equilibration of Matrices*, Numer. Math., Vol. 14, (14–23) 1969.

[28] K. Veselić, *Jacobi's method is more accurate than QR II*, unpublished manuscript, 1989.

[29] K. Veselić, *An Eigenreduction Algorithm for Definite Matrix Pairs*, preprint, FB Mathematik, Fernuniversität, Hagen, 1989, 1992, to appear in Numer. Math.

[30] K. Veselić, *unpublished note*, 1987.

[31] K. Veselić, V. Hari, *A Note on a One–Sided Jacobi Algorithm*, Numer. Math. 56, (627–633) 1989.

[32] K. Veselić, I. Slapničar, *Floating–point perturbations of Hermitian matrices*, preprint, FB Mathematik, Fernuniversität, Hagen, 1991, submitted to Lin. Alg. Appl.

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

Lebenslauf

13.07.1961	Geburt in Split, Kroatien Vater: Prof. Dr. Petar Slapničar Mutter: Antonia Slapničar, geb. Jurić Beruf des Vaters: Universitätsprofessor Staatsangehörigkeit: kroatisch Konfession: römisch–katholisch
1967–1973	Volksschule in Split
1973–1974	Volksschule in Palo Alto, Kalifornien
1974–1978	Gymnasium in Split
1978–1984	Studium der Mathematik an der Universität Zagreb
1981–1984	Studium des Klaviers an der Universität Zagreb
Juli 1984	Hauptdiplom in Mathematik
1984–1988	Postdiplomstudium der Mathematik an der Universität Zagreb
1985–1989	Wissenschaftlicher Assistent an der Universität Split
Juli 1988	Magister der Naturwissenschaften aus dem Gebiet der Mathematik
seit 1.06.1990	Wissenschaftlicher Angestellter an der Fernuniversität Gesamthochschule Hagen

Hagen, den 29.05.1992