Highly Accurate Symmetric Eigenvalue Decomposition and Hyperbolic SVD*

Ivan Slapničar†

August 10, 2002

Abstract

Let G be a $m \times n$ real matrix with full column rank and let J be a $n \times n$ diagonal matrix of signs, $J_{ii} \in \{-1, 1\}$. The hyperbolic singular value decomposition (HSVD) of the pair (G, J) is defined as $G = U\Sigma V^{-1}$, where U is orthogonal, Σ is positive definite diagonal, and V is J-orthogonal matrix, $V^T JV = J$. We analyze when it is possible to compute the HSVD with high relative accuracy. This essentially means that each computed hyperbolic singular value is guaranteed to have some correct digits, even if they have widely varying magnitudes. We show that one-sided J-orthogonal Jacobi method method computes the HSVD with high relative accuracy. More precisely, let $B = GD^{-1}$, where D is diagonal such that the columns of B have unit norms. Essentially, we show that the computed hyperbolic singular values of the pair (G, J) will have $\log_{10}(\varepsilon/\sigma_{\min}(B))$ correct decimal digits, where ε is machine precision. We give the necessary relative perturbation bounds and error analysis of the algorithm. Our numerical tests confirmed all theoretical results.

For the symmetric non-singular eigenvalue problem $Hx = \lambda x$, we analyze the two-step algorithm which consists of factorization $H = GJG^T$ followed by the computation of the HSVD of the pair (G, J). Here G is square and non-singular. Let $\widehat{B} = \widehat{D}G$, where $\widehat{D}$ is diagonal such that the rows of $\widehat{B}$ have unit norms, and let B be defined as above. Essentially, we show that the computed eigenvalues of H will have $\log_{10}(\varepsilon/\sigma^2_{\min}(\widehat{B}) + \varepsilon/\sigma_{\min}(B))$ correct decimal digits. This accuracy can be much higher then the one obtained by the classical QR and Jacobi methods applied to H, where the accuracy depends on the spectral condition number of H, particularly if the matrices B and $\widehat{B}$ are well conditioned, and we are interested in the accurate computation of tiny eigenvalues. Again, we give the perturbation and error bounds, and our theoretical predictions are confirmed by a series of numerical experiments.

We also give the corresponding results for eigenvectors and hyperbolic singular vectors.

Keywords: Hyperbolic singular value decomposition; Symmetric eigenvalue problem; Symmetric indefinite decomposition; Jacobi method; Relative perturbation theory; High relative accuracy

AMS classification: 65F15, 65G05, 65F35, 15A18

*This work is partially contained in the author's Ph.D Thesis [23] which was done at the Fernuniversität Hagen, Germany, under the supervision of Professor Krešimir Veselić. The author acknowledges the grant 037012 of the Croatian Ministry of Science and Technology.

†University of Split, Faculty of Electrical Engineering, Mechanical Engineering, and Naval Architecture, R. Boškovića b.b., 21000 Split, Croatia (Ivan.Slapnicar@fesb.hr). The penultimate revision of this work was written while the author was visiting the Department of Mathematics and Statistics, Utah State University, Logan, UT.

1 Introduction

The problem of computing eigenvalue and singular value decompositions of real matrices with high relative accuracy has been considered by many authors, for example by Barlow and Demmel [3], Demmel and Kahan [8], Demmel and Gragg [7], Demmel et al. [6], Drmač [10, 11], Mathias [18], Slapničar [23] and Veselić [31]. The term "high relative accuracy" means that the algorithm is capable of computing eigenvalues or singular values with higher relative accuracy than can be obtained by classical QR algorithm [14, §8.3], [20, §8] or divide and conquer algorithm [14, §8.5], [16]. More precisely, the latter two algorithms are backward stable and compute the eigenvalues of a real symmetric matrix H with absolute error $|\lambda_i - \lambda_i'| \leq f(n)\varepsilon\|H\|_2$. Here the original eigenvalues λ_i and the computed eigenvalues λ_i' are in the same order, $f(n)$ is a moderately growing function of the matrix dimension n, ε is the machine precision, and $\|H\|_2$ is the spectral norm of the matrix. For the relative error this implies

$$\frac{|\lambda_i - \lambda_i'|}{|\lambda_i|} \leq \frac{f(n)\varepsilon\|H\|_2}{|\lambda_i|} \leq f(n)\varepsilon\kappa(H), \tag{1}$$

provided H is non-singular. Here $\kappa(H) = \|H\|_2\|H^\dagger\|_2$ denotes the spectral condition number, where $H^\dagger$ is the pseudo-inverse of H. Similarly, the QR algorithm [14, §8.6] or divide and conquer algorithm [15] compute the singular values of a full column rank matrix G with the relative accuracy

$$\frac{|\sigma_i - \sigma_i'|}{\sigma_i} \leq f_1(n)\varepsilon\kappa(G), \tag{2}$$

where $f_1(n)$ is a moderately growing function of n.

There are many classes of matrices for which such accuracy results are inadequate, in particular for tiny eigenvalues or singular values, like bidiagonal matrices [8], acyclic matrices [7], scaled diagonally dominant matrices [3] and well-scaled positive definite matrices [9] which apply in finite elements applications [21]. In all cases algorithms were given which compute the solutions with higher accuracy then given in (1) or (2). The scheme of the analysis is always the following:

relative perturbation theory + relative error analysis = relative error bounds.

In [9] Demmel and Veselić proved that the Jacobi method [14, §8.4], [20, §9] computes the eigenvalues of the positive definite symmetric matrix H with optimal relative accuracy. More precisely: if we write $H = DAD$ where $D = \mathrm{diag}([H_{ii}]^{1/2})$ and $A_{ii} = 1$, then

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq f_2(n)\varepsilon\kappa(A), \tag{3}$$

where $f_2(n)$ is a moderately growing function of n. This bound will hold even if the initial matrix entries have ε-relative uncertainties, that is, if one computes the eigenvalues of the matrix $H + \delta H$ where $|\delta H_{ij}| \leq \varepsilon|H_{ij}|$. Such uncertainties typically occur when the matrix is stored in the computer. Notice that $\log_{10}$ of the left hand side of (3) is the number of the accurate decimal digits. It is important to notice that the matrix A is nearly optimally scaled in the sense that (see [30])

$$\kappa(A) \leq n \min_{\Delta=\mathrm{diag}} \kappa(\Delta A \Delta).$$

This inequality trivially implies that

$$\kappa(A) \leq n\kappa(H),$$

which, in turn, implies an important fact that the bound (3) can never be much worse than the classical bound (1). Clearly, if the matrix H is strongly scaled in the sense that A is well-conditioned and H is not, then the bound (3) will be much better than (1). Therefore, in such cases the Jacobi method is the method of choice if one wants to compute eigenvalues with small relative error. It is important to stress a caveat which is present in [9]: the Jacobi method forms sequence of orthogonally similar matrices H_k which converges to a diagonal matrix whose diagonal elements are the desired eigenvalues. To this sequence there corresponds the sequence of scaled matrices A_k, defined by

$$H_k = D_k A_k D_k, \qquad D_k = \mathrm{diag}([H_k]_{ii}^{1/2}), \tag{4}$$

such that $[A_k]_{ii} = 1$. The convergence of the series H_k to a diagonal matrix is equivalent to convergence of the sequence A_k to the identity matrix. However, for (3) to hold, $\kappa(A_k)$ should not grow much over $\kappa(A)$ during the algorithm. There is no theoretical proof that this is true, instead a strong numerical evidence was given in [9].

Demmel and Veselić also proved that essentially the same accuracy as in (3) is attained by the following two step method: in the first step H is decomposed by the Cholesky factorization as $H = LL^T$; in the second step one-sided Jacobi method is applied from the right to L in order to compute the singular value decomposition $L = U\Sigma V^T$. Then $\lambda_i = \Sigma_{ii}^2$, and the columns of U are the corresponding eigenvectors.

For the singular value decomposition, Demmel and Veselić proved that the one-sided Jacobi method applied from the right to a $m \times n$ full-column rank matrix G computes the singular values with the relative accuracy bounded by

$$\frac{|\sigma_i - \sigma_i'|}{\sigma_i} \leq f_3(n)\varepsilon\kappa(B), \tag{5}$$

where

$$G = BD, \qquad D = \mathrm{diag}(\|G_{\cdot i}\|_2), \tag{6}$$

that is, the columns of B have unit norms, and $f_3(n)$ is a moderately growing function of n. Here $B_{\cdot i}$ denotes the i-th column of the matrix B. In analogy to the symmetric positive definite case described above, the bound (5) will be better than the classical bound (2) if the matrix B is strongly scaled from the right in the sense that B is well-conditioned and G is not. There is also a caveat analogous to the one in the symmetric positive definite case: the one-sided Jacobi method from [9] forms a sequence of matrices G_k which converges to a matrix with orthogonal columns; the column-norms of the final matrix being the desired singular values. To this sequence there corresponds the sequence of scaled matrices B_k defined by $B_k = G_k D_k^{-1}$, where $D_k = \mathrm{diag}(\|[G_k]_{\cdot i}\|_2)$ such that $\|[B_k]_{\cdot i}\|_2 = 1$. The convergence of the series G_k to a matrix with orthogonal columns is equivalent to convergence of the sequence B_k to a matrix with orthonormal columns implying that $\kappa(B_k) \to 1$ as k increases. However, for (5) to hold, $\kappa(B_k)$ should not grow much over $\kappa(B)$ during the algorithm. Again, there is no theoretical proof that this is true, instead a strong numerical evidence was given in [9].

When considering the classical SVD this caveat can be removed by applying the one-sided Jacobi method from the left, and not from the right (e.g. for square non-singular G, see [10, 12] for details). Then the error analysis does not depend on growth of $\kappa(B_k)$, since this quantity does not change when performing rotations from the left. The disadvantage of this approach is that it is in general slower than when one-sided Jacobi is applied from the right, this is, from the side from which the matrix is well scaled. When considering the hyperbolic SVD, we cannot apply this approach, since in the hyperbolic case the rotations must be performed from the

right, as we shall see later. The problem of computing the singular value decomposition with high relative accuracy was further analyzed in [6].

To summarize, bounds (3) and (5) essentially show that the accuracy of the computed values is determined by the condition of the scaled matrix, rather than the condition of the original matrix. In particular, the singular values can be computed to high relative accuracy only if the right hand side of (5) is less than one, and remains less than one during the algorithm. In this paper we prove that the same is the case for the hyperbolic singular value decomposition algorithm.

In this paper we consider two problems:

- the hyperbolic singular value decomposition (HSVD) for the pair (G, J), and
- the classical eigenvalue problem for the non-singular indefinite symmetric matrix H.

The reason for considering such two different problems here, is that the HSVD is a part of our highly accurate algorithm for the real symmetric eigenvalue problem.

The HSVD of the pair (G, J), where G is a $m \times n$ full column rank matrix and J is a $n \times n$ diagonal matrix of signs, $J = \mathrm{diag}(\pm 1)$, is defined as [19, 33]

$$G = U\Sigma V^{-1}, \tag{7}$$

were U is a $m \times m$ orthogonal matrix, $\Sigma = \mathrm{diag}(\sigma_i)$ is a $m \times n$ diagonal matrix with $\sigma_i > 0$, and V is a $n \times n$ J-orthogonal matrix, that is, $V^T J V = J$. The diagonal entries σ_i are the hyperbolic singular values of the pair (G, J), the columns of U are the left singular vectors, and the columns of V are the right singular vectors. We prove that the one-sided J-orthogonal Jacobi method applied to the matrix G from the right computes the hyperbolic singular values σ_i with the accuracy given by (5).

For the symmetric indefinite eigenvalue problem $Hx = \lambda x$, we analyze the following two-step algorithm originally proposed by Veselić [31]:

- in the first step H is decomposed by the symmetric indefinite factorization [24] as $H = GJG^T$ where J is a diagonal matrix with $J_{ii} \in \{-1, 1\}$;
- in the second step one-sided J-orthogonal Jacobi method is applied from the right to G in order to compute the hyperbolic singular value decomposition (7).

Note that (7), $H = GJG^T$ and $V^T J V = J$ imply $H = U\Sigma^2 J U^T$. Hence, $\lambda_i = \sigma_i^2 J_{ii}$ are the eigenvalues, and the columns of U are the corresponding eigenvectors of H. For this algorithm we prove that it computes the eigenvalues λ_i with the accuracy essentially given by (3), where A is obtained from $|H| = DAD$, where $D = \mathrm{diag}(|H|_{ii}^{1/2})$ such that $A_{ii} = 1$, and $|H| = \sqrt{H^2}$ is the positive definite polar factor of H.

Since we consider problems which involve the sign matrix J, our results generalize the corresponding results from [9, 18, 10, 6], where $J = I$, to larger classes of problems.

For the computed hyperbolic singular vectors we prove relative norm-wise error bounds. Roughly speaking, these bounds are proportional to the condition of the scaled matrix B and inversely proportional to relative gaps between singular values. Similarly, for the computed eigenvectors we prove relative norm-wise error bounds which are proportional to the condition of the scaled matrix A and inversely proportional to relative gap between eigenvalues. These bounds are also proper generalization of the corresponding results from [9, 18, 10, 6].

The results of this paper are partially contained in [23]. Let us briefly outline the major differences. In [23], the one-sided J-orthogonal Jacobi method was analyzed only to the extent

necessary for its use in the eigenvalue computations. Here we also discuss additional details when this method is used as the HSVD solver. In particular, in Theorem 3 we give error bounds for the computed right hyperbolic singular vectors (matrix V from (7)). These bounds were not derived in [23], due to lack of the adequate perturbation bounds. Further, the proof of the error bound for the computed eigenvectors in [23] was based on the perturbation bound from [32, Th. 2.48]. The proof of the eigenvector bound from our Theorem 5 is, on the other hand, based on the perturbation bound from [29, Th. 6]. This is a better approach, since here the eigenvalues that correspond to the observed invariant subspace need not be adjacent. Also, for the symmetric indefinite factorization (see §3), instead of the error bound from [23, §4], we use the sharper error bound from [24].

The paper is organized as follows. In §2 we describe the hyperbolic singular value decomposition. In §2.1 we state the existing relative perturbation results. In §2.2 we describe the one-sided J-orthogonal Jacobi method, and in §2.3 we analyze one step of the method. In §2.4 we plug the error bounds from §2.3 into the perturbation bounds of §2.1 to obtain the overall error bounds for the method. In §2.5 we give results of numerical experiments.

Section §3 deals with the symmetric eigenvalue problem. We first describe the above two-step algorithm in more details, and state the existing error analysis. In §3.1 we state the existing relative perturbation results for the symmetric eigenvalue problem. In §3.2 we give overall error bounds for the two-step algorithm. Finally, in §3.3 we give results of numerical experiments.

2 Hyperbolic singular value decomposition

In this section we consider the HSVD (7) of the matrix pair (G, J). From now on, we assume that G is a real matrix with full-column rank. Since $V^T J V = J$ implies $V^{-1} = J V^T J$, the HSVD may also be written as $G = U \Sigma J V^T J$. Similarly to the classical singular value decomposition (when $J = I$), the HSVD is closely related to two eigenvalue problems. Matrix U is the eigenvector matrix of the symmetric indefinite non-singular eigenvalue problem

$$H = GJG^T = U \Sigma J \Sigma^T U^T, \tag{8}$$

the eigenvalues of H being $\lambda_i = \sigma_i^2 J_{ii}$, $i = 1, \ldots, n$, and $\lambda_i = 0$, $i = n+1, \ldots, m$. Furthermore, matrix V^{-1} is the eigenvector matrix of the hyperbolic eigenvalue problem $G^T G x = \lambda J x$ [25],

$$G^T G = V^{-T} \Sigma^T \Sigma V^{-1}, \quad V^{-T} J V^{-1} = J,$$

the hyperbolic eigenvalues being $\lambda_i = \sigma_i^2$, $i = 1, \ldots, n$. Also, U and V are related by

$$U_{\cdot,1:n} = GV \operatorname{diag}(\sigma_i^{-1}), \tag{9}$$

where $U_{\cdot,1:n}$ denotes the matrix of the first n columns of U.

The HSVD is a natural way to find the eigenvalues of a difference of two outer products

$$H = G_1 G_1^T - G_2 G_2^T.$$

This is done by writing H in the product form

$$H = GJG^T, \quad G = [\, G_1 \quad G_2 \,], \quad J = \operatorname{diag}(I, -I),$$

and then solving the problem (8) via the HSVD (7) (see [19], [33]).

As already mentioned, one of the major applications of the HSVD is its use in the highly accurate algorithm for solving the classical symmetric eigenvalue problem (see [31, 23]), as we describe in §3.

2.1 Relative perturbation bounds

The relative perturbation bounds for the HSVD have been proved in [32, 25]. As already mentioned, we consider $G = BD$ scaled from the right as in (6). Let $(G + \delta G, J)$ be the perturbed pair, where

$$\delta G = \delta B D.$$

We set

$$\beta = \|\delta B B^{\dagger}\|_2, \qquad \beta_F = \|\delta B B^{\dagger}\|_F. \tag{10}$$

Obviously, if $\|\delta B\|_2$ or $\|\delta B\|_F$ are known, which will be the case in our subsequent error analysis, then

$$\beta \leq \frac{\|\delta B\|_2}{\sigma_{\min}(B)}, \qquad \beta_F \leq \frac{\|\delta B\|_F}{\sigma_{\min}(B)}.$$

In particular, for the element-wise perturbation of G of the form

$$|\delta G| \leq \varepsilon |G|,$$

which typically appears when the matrix is being stored in computer memory, we have

$$\beta \leq \varepsilon \frac{\| \, |B| \, \|_2}{\sigma_{\min}(B)} \leq \varepsilon \frac{\sqrt{n}}{\sigma_{\min}(B)}, \qquad \beta_F \leq \varepsilon \frac{\sqrt{n}}{\sigma_{\min}(B)}.$$

According to [32, Th. 3.3], if δG is such that $\|\delta G x\|_2 \leq \beta \|G x\|_2$ for all vectors x and some $\beta < 1$, then the singular values of the pairs (G, J) and $(G + \delta G, J)$, σ_i and $\widetilde{\sigma}_i$, respectively, satisfy the inequalities

$$1 - \beta \leq \frac{\widetilde{\sigma}_i}{\sigma_i} \leq 1 + \beta. \tag{11}$$

Here we assume that σ_i and $\widetilde{\sigma}_i$ are in the increasing order. Since

$$\|\delta G x\|_2 = \|\delta B D x\|_2 = \|\delta B B^{\dagger} B D x\|_2 \leq \|\delta B B^{\dagger}\|_2 \|G x\|_2,$$

(11) holds with β defined by (10), as well.

Perturbation bounds for left and right singular vectors are given in terms of relative variants of the well-known $\sin\Theta$ theorems [5]. Let $\mathcal{U}$ and $\widetilde{\mathcal{U}}$ be two subspaces of the same dimension. The sines of the canonical angles between the subspaces $\mathcal{U}$ and $\widetilde{\mathcal{U}}$ are the diagonal entries of the matrix $\sin\Theta(\mathcal{U},\widetilde{\mathcal{U}})$ which is defined as follows [28, Cor. I.5.4]: let $U_{\perp}$ and $\widetilde{U}$ form orthonormal basis for $\mathcal{U}_{\perp}$ and $\widetilde{\mathcal{U}}$, respectively, where $\mathcal{U}_{\perp}$ is the orthogonal complement of $\mathcal{U}$, and let QSW^* be a singular value decomposition of $U_{\perp}^* \widetilde{U}$. Then $\sin\Theta(\mathcal{U},\widetilde{\mathcal{U}}) = S$.

In order to state the bounds, we introduce the following notation: let the HSVD of the pair (G, J) be written as

$$G = \begin{bmatrix} U_1 & U_2 & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1}, \tag{12}$$

where U_1 is $m \times k$ matrix, U_2 is $m \times (n-k)$ matrix, and the rest of the matrices have the corresponding dimensions. Similarly, let

$$\widetilde{G} = G + \delta G = \begin{bmatrix} \widetilde{U}_1 & \widetilde{U}_2 & \widetilde{U}_0 \end{bmatrix} \begin{bmatrix} \widetilde{\Sigma}_1 & \\ & \widetilde{\Sigma}_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \widetilde{V}_1 & \widetilde{V}_2 \end{bmatrix}^{-1}.$$

Here we assume that σ_i and $\widetilde{\sigma}_i$ are in the same order (not necessarily increasing or decreasing). More precisely, σ_i denotes the k-th largest hyperbolic singular value of the pair (G, J), and $\widetilde{\sigma}_i$ denotes the k-th largest hyperbolic singular value of the perturbed pair $(\widetilde{G}, J)$. Similarly to the eigenvector and singular vector bounds which are used in [3, 9, 6], the bounds which we use also depend on a relative gap between the singular values from $\widetilde{\Sigma}_1$ and those from Σ_2. We use the relative gap which is defined by

$$\mathrm{rg}(\widetilde{\Sigma}_1, \Sigma_2) = \min_{\substack{1 \le p \le k \\ k+1 \le q \le n}} \frac{|\widetilde{\sigma}_p J_{pp} - \sigma_q J_{qq}|}{2 \max\{\widetilde{\sigma}_p, \sigma_q\}}, \tag{13}$$

Notice that the relative gap contains diagonal elements of the sign matrix J. This, for example, implies that the hyperbolic singular values which correspond to diagonal elements of J of different signs are always well separated (the relative gap is in that case greater than $1/2$).

We are now ready to state the perturbation bounds for singular subspaces. Let $\mathcal{U}_1$ and $\widetilde{\mathcal{U}}_1$ be the subspaces spanned by the columns of U_1 and $\widetilde{U}_1$, respectively. According to [26, Th. 3], if $\beta < 1$, then

$$\| \sin \Theta(\mathcal{U}_1, \widetilde{\mathcal{U}}_1) \|_F \le \frac{2\,\beta_F}{1-\beta} \cdot \frac{1}{\mathrm{rg}(\widetilde{\Sigma}_1, \Sigma_2)}. \tag{14}$$

Further, let $\mathcal{V}_1$ and $\widetilde{\mathcal{V}}_1$ be the subspaces spanned by the columns of V_1 and $\widetilde{V}_1$, respectively. According to [26, Th. 4] (see also [25, Th. 4]), if $\beta, \beta_F < 1/3$, then

$$\| \sin \Theta(\mathcal{V}_1, \widetilde{\mathcal{V}}_1) \|_F \le \|V\|_2^2 \left(\frac{1}{2}\psi + \sqrt{1 + \frac{1}{4}\psi^2} \right) \frac{\psi}{\mathrm{rg}(\widetilde{\Sigma}_1, \Sigma_2)}, \tag{15}$$

where

$$\psi = \frac{3\,\beta_F}{\sqrt{1 - 3\,\beta}}.$$

We can further simplify the above bound as follows: according to [27, Th. 3], $\|V\|_2^2$ is bounded by

$$\|V\|_2^2 \le \min_{\Delta} \sqrt{\kappa(\Delta^* G^* G \Delta)},$$

where the minimum is over all matrices which commute with J. Thus, by taking $\Delta = D^{-1}$, we have

$$\|V\|_2^2 \le \sqrt{\kappa(D^{-1} G^* G D^{-1})} = \sqrt{\kappa(B^* B)} = \kappa(B). \tag{16}$$

By comparing the bound (14) with the bounds from [9, Th. 2.16, Cor. 2.17] and [17, Th. 4.3], we see that the left (unitary) singular vectors in the HSVD behave as well as the left singular vectors in the classical SVD. Namely, all bounds essentially depend on δB, $\sigma_{\min}(B)$ and the relative gap. On the other hand, the bound (15) for the right hyperbolic singular vectors has an additional factor $\|V\|_2^2$ over the corresponding bounds from [9, Th. 2.16, Cor. 2.17] and [17, Th. 4.3]. However, when applying (15) to the classical SVD with $J = I$ this term vanishes since V is unitary. Since V is J-orthogonal, we have $\|V\|_2^2 = \kappa(V)$. This additional factor is not unusual, since the spectral condition number of the non-unitary eigenvectors appears naturally in various other matrix perturbation bounds.

2.2 One-sided J-orthogonal Jacobi method

The one-sided or implicit J-orthogonal Jacobi method, originally proposed by Veselić [31], consists of an iterative application of the one-sided transformation

$$G_{k+1} = G_k J_k, \tag{17}$$

where $G \equiv G_0$ and J_k is a J-orthogonal Jacobi-type plane rotation. Let $A^{(i,j)}$ denote the 2×2 pivot submatrix of any square matrix A. The matrix J_k is equal to the identity matrix except for the (i,j) 2×2 submatrix $J_k^{(i,j)}$ obtained on the intersection of rows and columns i and j. It is defined by

$$J_k^{(i,j)} = \begin{cases} \begin{bmatrix} ch & sh \\ sh & ch \end{bmatrix}, & \text{for } J_{ii} = -J_{jj}, \\ \\ \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix}, & \text{for } J_{ii} = J_{jj}. \end{cases}$$

The pair (i,j) is the pivot pair. The J-orthogonality of the matrix J_k implies that $ch = \cosh\psi$, $sh = \sinh\psi$, $cs = \cos\varphi$ and $sn = \sin\varphi$ for some ψ and φ, respectively. These two types of rotations are called the *hyperbolic* and the *orthogonal rotation*, respectively. The parameter φ or ψ is chosen to annihilate the (i,j)-element of the Gram matrix $H_k = G_k^T G_k$. In other words, the transformation (17) makes the i-th and j-th columns of G_{k+1} orthogonal. More precisely, let

$$H_k^{(i,j)} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

be the (i,j) pivot submatrix of H_k. Then

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

or

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

In exact arithmetic, the sequence (17) is closely related to the two-sided J-orthogonal Jacobi method for the hyperbolic eigenvalue problem $Hx = \lambda Jx$, where $H = G^T G$. Namely, in the sequence

$$H_0 = H, \qquad H_{k+1} = J_k^T H_k J_k, \tag{18}$$

the matrix H_{k+1} obtains zeros at the positions (i,j) and (j,i). The sequence (18) converges towards a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_i)$ [31], and this convergence is quadratic [13].

One difference between orthogonal and hyperbolic rotations is that $\mathrm{Trace}\,(H_{k+1}) = \mathrm{Trace}\,(H_k)$ after orthogonal, and $\mathrm{Trace}\,(H_{k+1}) < \mathrm{Trace}\,(H_k)$ after hyperbolic rotation. Using this trace reduction argument, Veselić [31] proved that the hyperbolic tangent tends to zero as the sequence H_k converges. The second difference is that the condition of the rotation matrix J_k is in the orthogonal case one, while in the hyperbolic case it can be large. Notice that $\tanh\psi_k$ is bounded as follows: set $H_k = G_k^T G_k$ and define the scaled matrix A_k by (4). Then

$$|\tanh\psi_k| \leq \frac{\sqrt{\kappa(A_k^{(i,j)})} - 1}{\sqrt{\kappa(A_k^{(i,j)})} + 1}.$$

This, in turn, implies that in the hyperbolic case

$$\kappa(J_k^{(i,j)}) \leq \sqrt{\kappa(A_k^{(i,j)})}.$$

The convergence of the sequence (18) towards a diagonal matrix implies that the sequence (17) approaches the set of matrices with orthogonal columns. Assume that we terminate the

sequence (18) after M steps, when the final matrix H_M is sufficiently diagonal according to some chosen stopping criterion. Then the columns of G_M are sufficiently orthogonal, and the HSVD of the starting pair (G, J) is approximated as follows (c.f. (9)):

$$\begin{aligned} \sigma_i &\approx \sqrt{(H_M)_{ii}} = \|(G_M)_{\cdot i}\|_2, \qquad i = 1, \ldots n, \\ V &\approx J_0 J_1 \cdots J_{M-1}, \\ U_{\cdot,1:n} &\approx GV \,\mathrm{diag}(\sigma_i^{-1}) = G_M \,\mathrm{diag}(\sigma_i^{-1}). \end{aligned}$$

The choice of pivot pair (i, j) in the k-th step can be made according to various pivoting strategies. Here we use the commonly used row-cyclic strategy [14, §8.4.4], [20, §9.4.2]:

$$(1,2), (1,3), \ldots, (1,n), (2,3), \cdots, (2,n), (3,4), \cdots, (n-1,n).$$

We now present our algorithm:

Algorithm 1 *Implicit J-orthogonal Jacobi method for the pair (G, J). Tolerance tol is a user defined stopping criterion. V is initially the identity matrix.*

repeat
  for $i = 1$ *to* $n-1$
    for $j = i+1$ *to* n
      /* *compute* $\widehat{H} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$, *the* (i,j) *submatrix of* $G^T G$ */
        $a = \sum_{k=1}^{m} G_{ki}^2$
        $b = \sum_{k=1}^{m} G_{kj}^2$
        $c = \sum_{k=1}^{m} G_{ki} * G_{kj}$
      /* *if* $c = 0$, *the step is skipped* */
        if $c = 0$ *then go to the next step*
      /* *compute the parameter hyp:* $hyp = 1$ *for the orthogonal, and*
        $hyp = -1$ *for the hyperbolic rotation, respectively* */
        $hyp = J_{ii} * J_{jj}$
      /* *compute the* $J-$*orthogonal Jacobi rotation which diagonalizes* $\widehat{H}$ */
        $\zeta = hyp * (b - hyp * a)/(2c)$
        $t = sign(\zeta)/(|\zeta| + \sqrt{\zeta^2 + hyp})$
        $h = \sqrt{1 + hyp * t^2}$
        $ch = 1/h$
        $sh = t/h$
        $sh1 = -hyp * sh$
      /* *update columns* i *and* j *of* G */
        for $k = 1$ *to* m
          $tmp = G_{ki}$
          $G_{ki} = ch * tmp + sh1 * G_{kj}$
          $G_{kj} = sh * tmp + ch * G_{kj}$
        endfor
      /* *update columns* i *and* j *of* V */
        for $k = 1$ *to* n
          $tmp = V_{ki}$
          $V_{ki} = ch * tmp + sh1 * V_{kj}$
          $V_{kj} = sh * tmp + ch * V_{kj}$

endfor
endfor
endfor
until convergence (all $|c|/\sqrt{ab} \leq tol$*)*
/ the computed hyperbolic singular values are* $\sigma_i = (\sum_{k=1}^{m} G_{ki}^2)^{1/2}$ **/*
/ the corresponding computed left singular vectors are the normalized columns of the final* G **/*

Notice that if G is square [9, 10, 18], the one-sided method can be applied from the right either to G *or* G^T, since for $J = I$ the matrices G^TG and GG^T have the same eigenvalues and simply related eigenvectors. For $J \neq I$, however, only application to G from the right or to G^T from the left makes sense.

Algorithm 1 gives only the simplest version of the method, in order to make the subsequent error analysis clearer. In practice, however, we frequently use several enhancements which reduce the operation count:

- keeping and updating the diagonal of the Gram matrix in a separate vector,
- fast rotations,
- fast self-scaling rotations.

Updating the diagonal elements of the Gram matrix in a separate vector makes the computation of parameters a and b unnecessary, thus saving $4m$ operations in each step. Using fast rotations of the form

$$J_k^{(i,j)} = \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix}$$

saves another $2m$ multiplications in updating G and $2n$ multiplications in updating V. Fast self-scaling or dynamically scaling rotations, originally introduced in [1], are used in order to avoid possible underflows when using fast rotations.

The algorithms which use the above enhancements are described in detail and analyzed in [23, §3.3, §3.4]. The error bounds for the solutions obtained by these algorithms differ only in constants from the bounds which we derive for Algorithm 1 in subsequent sections.

2.3 Error analysis

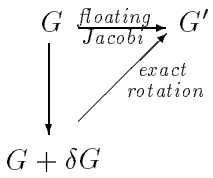
In this section we give error analysis of one step of Algorithm 1. We use the standard model of the finite precision floating-point arithmetic. The floating-point result $fl(\odot)$ of the operation $\odot$ is given by [34]

$$\begin{aligned} fl(x \odot y) &= (x \odot y)(1 + \varepsilon_{\odot}) \\ fl(\sqrt{x}) &= \sqrt{x}(1 + \varepsilon_{\surd}) \end{aligned}$$

where $\odot$ represents any of the four basic arithmetic operations, '$+$', '$-$', '$\times$' or '$\div$'. Here $\varepsilon_{\odot}$ ($\varepsilon_{\surd}$) depends on x, y and $\odot$ (on x), but we always have $|\varepsilon_{\odot}|, |\varepsilon_{\surd}| \leq \varepsilon$, where $\varepsilon \ll 1$ is the machine precision.

Numerically subscripted ε's (like ε_1, ε_2, etc.) will denote independent quantities bounded in absolute value by ε. All other sub- or superscripted ε's will be defined in the proof.

Theorem 1 *Let the matrix* G' *be obtained from the matrix* G *by applying one step of Algorithm 1 in floating-point arithmetic with precision* ε*. Then the following diagram commutes:*



The top arrow indicates that G' is obtained from G by applying one J-orthogonal Jacobi rotation in floating-point arithmetic. The diagonal arrow indicates that G' is obtained from $G+\delta G$ by applying one J-orthogonal plane rotation in exact arithmetic. Thus, the pairs (G', J) and $(G+\delta G, J)$ have identical hyperbolic singular values and simply related singular vectors. δG is bounded as follows: let $G = BD$ be scaled according to (6), and write $\delta G = \delta B D$. Let $a = \sum_k G_{ki}^2$, $b = \sum_k G_{kj}^2$ and $c = \sum_k G_{ki}G_{kj}$. Notice that $D_{ii} = \sqrt{a}$ and $D_{jj} = \sqrt{b}$. Further, let $\widehat{a} = fl(\sum_k G_{ki}^2)$, $\widehat{b} = fl(\sum_k G_{kj}^2)$ and $\widehat{c} = fl(\sum_k G_{ki}G_{kj})$ be the computed values of a, b and c, respectively. Let

$$\widehat{A}^{(i,j)} = \begin{bmatrix} 1 & \widehat{c}/\sqrt{\widehat{a}\,\widehat{b}} \\ \widehat{c}/\sqrt{\widehat{a}\,\widehat{b}} & 1 \end{bmatrix},$$

and let $\widehat{\kappa} = \sqrt{\kappa(\widehat{A}^{(i,j)})}$. If $\widehat{A}^{(i,j)}$ is positive definite and $\max\{\widehat{\kappa}^2, m, 10\}\,\varepsilon \leq 0.01$, then

$$\|\delta B\|_2 \leq \|\delta B\|_F \leq C\,\varepsilon,$$

where [1]

$$C = \begin{cases} 13 & \text{in the orthogonal case,} \\ \widehat{\kappa}^2 + 11\,\widehat{\kappa} + 27 & \text{in the hyperbolic case for } \max\{\widehat{a},\widehat{b}\}/\min\{\widehat{a},\widehat{b}\} < 2, \\ 85 & \text{in the hyperbolic case for } \max\{\widehat{a},\widehat{b}\}/\min\{\widehat{a},\widehat{b}\} \geq 2 \end{cases}$$

PROOF. The orthogonal case was analyzed in several works. The values of C obtained in these works are the following: in [9, Th. 4.1] $C = 72$, in [23, Th. 3.3.3] $C = 26$, and in [18, Th. 4.2] $C = 13$, the last proof also being the simplest.

We continue with the proof of the hyperbolic case.

If $\widehat{c} = 0$, then, according to Algorithm 1, nothing is done in this step, and the theorem holds trivially.

From now on we assume that $\widehat{c} \neq 0$. Also, we assume without loss of generality that $\widehat{a} \geq \widehat{b}$ (the proof for the case $\widehat{a} < \widehat{b}$ is analogous). Let

$$\bar{\zeta} = -\frac{\widehat{a}+\widehat{b}}{2\widehat{c}}, \qquad \bar{t} = \frac{\text{sign}\,(\bar{\zeta})}{|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1}}. \tag{19}$$

By the positive definiteness of the matrix $\widehat{A}^{(i,j)}$, simple arithmetic shows that

$$|\bar{\zeta}| \;\geq\; \frac{\widehat{\kappa}^2+1}{\widehat{\kappa}^2-1} > 1, \tag{20}$$

$$|\bar{t}| \;\leq\; \frac{\widehat{\kappa}-1}{\widehat{\kappa}+1} < 1. \tag{21}$$

[1] Notice that C is defined through the quantity $\widehat{\kappa}$, which is defined by the computed quantities $\widehat{a}$, $\widehat{b}$ and $\widehat{c}$. This is convenient since these quantities are readily available during the computation.

Indeed, since $\widehat{a}+\widehat{b} \geq 2\sqrt{\widehat{a}\,\widehat{b}}$, we have

$$\frac{\widehat{\kappa}^2+1}{\widehat{\kappa}^2-1} = \frac{\frac{\sqrt{\widehat{a}\,\widehat{b}}+|\widehat{c}|}{\sqrt{\widehat{a}\,\widehat{b}}-|\widehat{c}|}+1}{\frac{\sqrt{\widehat{a}\,\widehat{b}}+|\widehat{c}|}{\sqrt{\widehat{a}\,\widehat{b}}-|\widehat{c}|}-1} = \frac{\sqrt{\widehat{a}\,\widehat{b}}}{|\widehat{c}|} \leq \frac{\widehat{a}+\widehat{b}}{2|\widehat{c}|} = |\bar{\zeta}|,$$

which proves (20). Inserting (20) into (19) gives (21).

Let $\widehat{t}$ be the computed value of $\bar{t}$. Let

$$\widetilde{ch} = 1/\sqrt{1-\widehat{t}^2}, \qquad \widetilde{sh} = \widehat{t}/\sqrt{1-\widehat{t}^2},$$

define the exact rotation which takes $G+\delta G$ to G'. More precisely, in the sequel G' will be computed by using the error analysis, and δG will be computed by using G' and the above definition of $\widetilde{ch}$ and $\widetilde{sh}$. Later we shall need the obvious inequalities

$$\widetilde{sh}^2 \leq |\widetilde{sh}|\,\widetilde{ch} \leq \widetilde{ch}^2 = \frac{1}{1-\widehat{t}^2}. \tag{22}$$

Let $\widehat{ch}$ and $\widehat{sh}$ denote the computed quantities $\widetilde{ch}$ and $\widetilde{sh}$, respectively, that is

$$\widehat{ch} = fl\left(\frac{1}{fl(\sqrt{fl(1-\widehat{t}^2)}\,)}\right), \qquad \widehat{sh} = fl\left(\frac{\widehat{t}}{fl(\sqrt{fl(1-\widehat{t}^2)}\,)}\right). \tag{23}$$

Notice an important fact that we can start our analysis from $\widehat{t}$, instead of from the exact value t. This is due to the fact that we are analyzing one-sided method – the difference between $\widehat{t}$ and t, as the proof of this theorem shows, does not affect the accuracy of the method.

Suppose that we can write (23) as

$$\widehat{sh} = (1+\varepsilon_{sh})\widetilde{sh}, \qquad \widehat{ch} = (1+\varepsilon_{ch})\widetilde{ch}. \tag{24}$$

Then

$$\begin{aligned} G'_{ki} &= fl(\widehat{ch} * G_{ki} + \widehat{sh} * G_{kj}) \\ &= [(1+\varepsilon_1)\,\widehat{ch}\,G_{ki} + (1+\varepsilon_2)\,\widehat{sh}\,G_{kj}](1+\varepsilon_3) \\ &= (1+\varepsilon_1)(1+\varepsilon_3)(1+\varepsilon_{ch})\,\widetilde{ch}\,G_{ki} + (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_{sh})\,\widetilde{sh}\,G_{kj} \\ &= \widetilde{ch}\,G_{ki} + \widetilde{sh}\,G_{kj} + E_{ki}, \end{aligned}$$

where E_{ki} contains all ε-terms, and, similarly,

$$G'_{kj} = fl(\widehat{sh} * G_{ki} + \widehat{ch} * G_{kj}) = \widetilde{sh}\,G_{ki} + \widetilde{ch}\,G_{kj} + E_{kj}.$$

The columns $E_{\cdot i}$ and $E_{\cdot j}$ are bounded by

$$\begin{aligned} \|E_{\cdot i}\|_2 &\leq |\varepsilon'_1|\,\widetilde{ch}\,\|G_{\cdot i}\|_2 + |\varepsilon'_2|\,|\widetilde{sh}|\,\|G_{\cdot j}\|_2, \\ \|E_{\cdot j}\|_2 &\leq |\varepsilon'_3|\,|\widetilde{sh}|\,\|G_{\cdot i}\|_2 + |\varepsilon'_4|\,\widetilde{ch}\,\|G_{\cdot j}\|_2. \end{aligned}$$

If $|\varepsilon_{sh}|, |\varepsilon_{ch}| \leq (0.5\,\widehat{\kappa}+4)\varepsilon$, which will be justified later, here

$$|\varepsilon'_1|, |\varepsilon'_4| \leq |\varepsilon_{ch}| + 2.02\varepsilon, \qquad |\varepsilon'_2|, |\varepsilon'_3| \leq |\varepsilon_{sh}| + 2.02\varepsilon. \tag{25}$$

For example, since the assumption of the theorem implies

$$\widehat{\kappa}\,\varepsilon = \sqrt{\widehat{\kappa}^2\varepsilon^2} \le \sqrt{0.01\,\varepsilon} \le \sqrt{0.01 \cdot 0.001} < 0.0032, \tag{26}$$

we have

$$\begin{aligned} |\varepsilon_1'| &\le |\varepsilon_1 + \varepsilon_3 + \varepsilon_{ch} + \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_{ch} + \varepsilon_3\varepsilon_{ch} + \varepsilon_1\varepsilon_3\varepsilon_{ch}| \\ &\le 2\varepsilon + |\varepsilon_{ch}| + 0.001\varepsilon + 2(0.5 \cdot 0.0032 + 4 \cdot 0.001)\varepsilon + 0.001 \cdot (0.5 \cdot 0.0032 + 4 \cdot 0.001)\varepsilon \\ &\le |\varepsilon_{ch}| + 2.02\varepsilon. \end{aligned}$$

Thus,

$$\begin{aligned} \begin{bmatrix} G_{\cdot i}' & G_{\cdot j}' \end{bmatrix} &= \begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \\ &= \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} E_{\cdot i} & E_{\cdot j} \end{bmatrix} \begin{bmatrix} \widetilde{ch} & -\widetilde{sh} \\ -\widetilde{sh} & \widetilde{ch} \end{bmatrix} \right) \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix} \\ &= \left(\begin{bmatrix} G_{\cdot i} & G_{\cdot j} \end{bmatrix} + \begin{bmatrix} \delta G_{\cdot i} & \delta G_{\cdot j} \end{bmatrix} \right) \begin{bmatrix} \widetilde{ch} & \widetilde{sh} \\ \widetilde{sh} & \widetilde{ch} \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} \|\delta G_{\cdot i}\|_2 &\le \widetilde{ch}\,\|E_{\cdot i}\|_2 + |\widetilde{sh}|\,\|E_{\cdot j}\|_2 \\ &\le (|\varepsilon_1'|\,\widetilde{ch}^2 + |\varepsilon_3'|\,\widetilde{sh}^2)\,\|G_{\cdot i}\|_2 + (|\varepsilon_2'| + |\varepsilon_4'|)\widetilde{ch}\,|\widetilde{sh}|\,\|G_{\cdot j}\|_2 \\ &\le \left(|\varepsilon_1'|\,\widetilde{ch}^2 + |\varepsilon_3'|\,\widetilde{sh}^2 + (|\varepsilon_2'| + |\varepsilon_4'|)\widetilde{ch}\,|\widetilde{sh}| \sqrt{\frac{b}{a}} \right) \sqrt{a}, \end{aligned} \tag{27}$$

and

$$\begin{aligned} \|\delta G_{\cdot j}\|_2 &\le |\widetilde{sh}|\,\|E_{\cdot i}\|_2 + \widetilde{ch}\,\|E_{\cdot j}\|_2 \\ &\le (|\varepsilon_4'|\,\widetilde{ch}^2 + |\varepsilon_2'|\,\widetilde{sh}^2)\,\|G_{\cdot j}\|_2 + (|\varepsilon_1'| + |\varepsilon_3'|)\widetilde{ch}\,|\widetilde{sh}|\,\|G_{\cdot i}\|_2 \\ &\le \left(|\varepsilon_4'|\,\widetilde{ch}^2 + |\varepsilon_2'|\,\widetilde{sh}^2 + (|\varepsilon_1'| + |\varepsilon_3'|)\widetilde{ch}\,|\widetilde{sh}| \sqrt{\frac{a}{b}} \right) \sqrt{b}. \end{aligned} \tag{28}$$

Notice that, since $\sqrt{a} = D_{ii} = \|G_{\cdot i}\|_2$ and $\sqrt{b} = D_{jj} = \|G_{\cdot j}\|_2$, dividing (27) and (28) by $\sqrt{a}$ and $\sqrt{b}$ gives bounds for $\|\delta B_{\cdot i}\|_2$ and $\|\delta B_{\cdot j}\|_2$, respectively.

In order to bound the above inequalities we have to consider two cases, depending whether $\sqrt{a/b}$ in (28) is bounded away from infinity or not. More precisely, we shall consider cases $\widehat{a}/\widehat{b} < 2$ and $\widehat{a}/\widehat{b} \ge 2$, respectively. In each case we will compute bounds for $|\varepsilon_i'|$, $\widetilde{ch}^2$, $\widetilde{sh}^2$, $\widetilde{ch}\,|\widetilde{sh}|$, $\sqrt{b/a}$ and $\sqrt{a/b}$, and insert those bounds into (27) and (28).

Case 1. Let $\widehat{a}/\widehat{b} < 2$. In order to bound $\widetilde{ch}^2$, $\widetilde{sh}^2$, $\widetilde{ch}\,|\widetilde{sh}|$ in terms of $\widehat{\kappa}$, we must bound $\widehat{t}$ in terms of $\bar{t}$. From (19) and the assumption $10\,\varepsilon \le 0.01$, it follows that

$$\widehat{\zeta} = fl(\bar{\zeta}) = fl\left(-\frac{\widehat{a} + \widehat{b}}{2\widehat{c}} \right) = (1 + \varepsilon_\zeta)\bar{\zeta}, \qquad |\varepsilon_\zeta| \le 3.005\,\varepsilon.$$

Further,

$$\begin{aligned} fl(\widehat{\zeta}^2 - 1) &= [(1 + \varepsilon_5)\widehat{\zeta}^2 - 1](1 + \varepsilon_6) \\ &= (1 + \varepsilon_5)(1 + \varepsilon_6)(1 + \varepsilon_\zeta)^2\bar{\zeta}^2 - (1 + \varepsilon_6) \\ &= (1 + \varepsilon_s)(\bar{\zeta}^2 - 1). \end{aligned}$$

Solving the last equality for ε_s, taking absolute value, and using the assumption $10\,\varepsilon \le 0.01$, gives

$$|\varepsilon_s| \le \frac{\varepsilon + 8.04\,\bar{\zeta}^2\varepsilon}{\bar{\zeta}^2 - 1} \le \frac{9.04\,\bar{\zeta}^2}{\bar{\zeta}^2 - 1}\varepsilon.$$

We continue with the error analysis: solving the last equality in

$$fl(\sqrt{\widehat{\zeta}^2 - 1}\,) = (1+\varepsilon_7)\sqrt{(1+\varepsilon_s)(\bar{\zeta}^2 - 1)} = (1+\varepsilon_u)\sqrt{\bar{\zeta}^2 - 1},$$

for ε_u, taking absolute value, and using the assumption on ε, gives

$$\begin{aligned}
|\varepsilon_u| &= |(1+\varepsilon_7)\sqrt{1+\varepsilon_s} - 1| \le (1+\varepsilon)\sqrt{1+|\varepsilon_s|} - 1 \\
&\le (1+\varepsilon)\sqrt{1+|\varepsilon_s| + \frac{|\varepsilon_s|^2}{4}} - 1 = (1+\varepsilon)\left(1 + \frac{|\varepsilon_s|}{2}\right) - 1 \\
&\le \varepsilon + \frac{|\varepsilon_s|}{2}(1+\varepsilon) \le \varepsilon + \frac{9.04}{2}\,\frac{\bar{\zeta}^2\,\varepsilon}{\bar{\zeta}^2 - 1}\,1.001 \le \frac{5.53\,\bar{\zeta}^2}{\bar{\zeta}^2 - 1}\varepsilon.
\end{aligned}$$

Further,

$$\begin{aligned}
fl(|\widehat{\zeta}| + \sqrt{\widehat{\zeta}^2 - 1}\,) &= [(1+\varepsilon_\zeta)|\bar{\zeta}| + (1+\varepsilon_u)\sqrt{\bar{\zeta}^2 - 1}\,](1+\varepsilon_8) \\
&= (1+\varepsilon_v)(|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1})
\end{aligned} \tag{29}$$

where

$$\begin{aligned}
|\varepsilon_v| &\le \frac{|(\varepsilon_\zeta + \varepsilon_8 + \varepsilon_\zeta\varepsilon_8)|\bar{\zeta}| + (\varepsilon_8 + \varepsilon_u + \varepsilon_8\varepsilon_u)\sqrt{\bar{\zeta}^2 - 1}|}{|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1}} \\
&\le 4.02\,\frac{|\bar{\zeta}|}{|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1}}\,\varepsilon + (\varepsilon + 1.001\,|\varepsilon_u|)\,\frac{\sqrt{\bar{\zeta}^2 - 1}}{|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1}} \\
&\le 4.02\,\varepsilon + \frac{6.54\,|\bar{\zeta}|}{\sqrt{\bar{\zeta}^2 - 1}}\varepsilon.
\end{aligned}$$

Since the right hand side is the decreasing function for $|\bar{\zeta}| > 1$, by using (20), we have

$$\frac{|\bar{\zeta}|}{\sqrt{\bar{\zeta}^2 - 1}} \le \frac{\widehat{\kappa}^2 + 1}{2\,\widehat{\kappa}} \le 0.5\,\widehat{\kappa} + 0.5,$$

and

$$|\varepsilon_v| \le 4.02\,\varepsilon + 6.54\,(0.5\,\widehat{\kappa} + 0.5)\,\varepsilon \le 10.56\,\widehat{\kappa}\,\varepsilon.$$

Further,

$$\begin{aligned}
\widehat{t} &= fl\left(\frac{\text{sign }(\widehat{\zeta})}{|\widehat{\zeta}| + \sqrt{\widehat{\zeta}^2 - 1}}\right) = \frac{\text{sign }(\bar{\zeta})}{(1+\varepsilon_v)(|\bar{\zeta}| + \sqrt{\bar{\zeta}^2 - 1})}(1+\varepsilon_9) \\
&= (1+\varepsilon_t)\bar{t},
\end{aligned} \tag{30}$$

where, by using (26),

$$|\varepsilon_t| \le \frac{|\varepsilon_9| + |\varepsilon_v|}{1 - |\varepsilon_v|} \le \frac{11.56\,\widehat{\kappa}\,\varepsilon}{1 - 10.56 \cdot 0.0032} \le 12\,\widehat{\kappa}\,\varepsilon.$$

Using this, (21), (26), the assumption $\max\{\widehat{\kappa}^2, 10\}\varepsilon \le 0.01$, and $\widehat{\kappa} > 1$, we have

$$\begin{aligned}
\frac{1}{1-\widehat{t}^2} &= \frac{1}{1 - \bar{t}^2(1+\varepsilon_t)^2} \\
&\le \frac{1}{1 - \left(\frac{\widehat{\kappa}-1}{\widehat{\kappa}+1}\right)^2 (1+\varepsilon_t)^2} \\
&\le \frac{(\widehat{\kappa}+1)^2}{4\,\widehat{\kappa} - \widehat{\kappa}\,(24\,\widehat{\kappa}^2\varepsilon + 48\,\widehat{\kappa}\,\varepsilon + 24\,\varepsilon + 12^2\,\widehat{\kappa}^3\varepsilon^2 + 2\cdot 12^2\,\widehat{\kappa}^2\varepsilon^2 + 12^2\,\widehat{\kappa}\,\varepsilon^2)} \\
&\le \frac{(\widehat{\kappa}+1)^2}{3.574\,\widehat{\kappa}} \le 0.28\,\left(\widehat{\kappa} + 2 + \frac{1}{\widehat{\kappa}}\right) \\
&\le 0.28\,\widehat{\kappa} + 0.84. \qquad (31)
\end{aligned}$$

The required bounds for $\widetilde{ch}^2$, $\widetilde{sh}^2$ and $\widetilde{ch}\,|\widetilde{sh}|$ terms in (27) and (28) follow from this and (22).

Now we have to bound ε_{sh} and ε_{ch} from (24) and insert those bounds into (25). Since $\widehat{sh}$ and $\widehat{ch}$ are defined in terms of $\widehat{t}$ (c.f. (23)), we start the analysis from there. By using (31), we have

$$fl(1-\widehat{t}^2) = [1 - (1+\varepsilon_{10})\,\widehat{t}^2](1+\varepsilon_{11}) = (1+\varepsilon_w)(1-\widehat{t}^2),$$

where

$$|\varepsilon_w| \le \frac{2.01\,\varepsilon}{1-\widehat{t}^2} \le (0.563\,\widehat{\kappa} + 1.689)\,\varepsilon.$$

This, in turn, implies

$$\widehat{h} = fl(\sqrt{1-\widehat{t}^2}) = (1+\varepsilon_{12})\sqrt{(1+\varepsilon_w)(1-\widehat{t}^2)} = (1+\varepsilon_h)\sqrt{1-\widehat{t}^2},$$

where

$$|\varepsilon_h| \le |\varepsilon_{12}| + \frac{|\varepsilon_w|}{2} + |\varepsilon_{12}|\,\frac{|\varepsilon_w|}{2} \le (0.282\,\widehat{\kappa} + 1.847)\,\varepsilon.$$

Therefore,

$$\widehat{ch} = fl\left(\frac{1}{\sqrt{1-\widehat{t}^2}}\right) = \frac{1}{(1+\varepsilon_h)\sqrt{1-\widehat{t}^2}}(1+\varepsilon_{13}) = (1+\varepsilon_{ch})\widetilde{ch},$$

where

$$|\varepsilon_{ch}| \le \frac{|\varepsilon_{13}| + |\varepsilon_h|}{1 - |\varepsilon_h|} \le (0.29\,\widehat{\kappa} + 2.86)\,\varepsilon.$$

The same estimate holds for $|\varepsilon_{sh}|$ since $fl(\widehat{t}) = \widehat{t}$,

$$|\varepsilon_{sh}| \le (0.29\,\widehat{\kappa} + 2.86)\,\varepsilon.$$

The above bounds for $|\varepsilon_{sh}|$ and $|\varepsilon_{ch}|$ justify, in turn, the assumption made in deriving (25). Inserting these bounds into (25) gives

$$|\varepsilon_i'| \le (0.29\,\widehat{\kappa} + 5)\,\varepsilon, \qquad i = 1,2,3,4. \qquad (32)$$

These inequalities bound the ε_i' terms in (27) and (28).

To complete the proof, we need to bound the $\sqrt{b/a}$ term in (27) and the $\sqrt{a/b}$ term in (28). Systematic application of (19) and the assumption $m\,\varepsilon \le 0.01$ implies (see, for example, the classical error analysis of the scalar product in [14, § 2.4]):

$$\widehat{a} = a(1+\varepsilon_a), \qquad \widehat{b} = b(1+\varepsilon_b), \qquad |\varepsilon_a|, |\varepsilon_b| \le 1.01\, m\,\varepsilon. \tag{33}$$

This implies

$$\frac{b}{a} \le \frac{\widehat{b}}{\widehat{a}} \cdot \frac{1+1.01\,m\,\varepsilon}{1-1.01\,m\,\varepsilon} \le \frac{\widehat{b}}{\widehat{a}} \cdot \frac{1+1.01\cdot 0.01}{1-1.01\cdot 0.01} \le 1.03\,\frac{\widehat{b}}{\widehat{a}}, \tag{34}$$

and, similarly,

$$\frac{a}{b} \le \frac{\widehat{a}}{\widehat{b}} \cdot \frac{1+1.01\,m\,\varepsilon}{1-1.01\,m\,\varepsilon} \le 1.03\,\frac{\widehat{a}}{\widehat{b}}. \tag{35}$$

From (34) and the assumption $\widehat{a} \ge \widehat{b}$, we have

$$\sqrt{\frac{b}{a}} < 1.02. \tag{36}$$

By inserting this, (32), (22) and (31) into (27), we have

$$\|\delta G_{\cdot i}\|_2 \le C_1\sqrt{a}\,\varepsilon, \qquad \|\delta B_{\cdot i}\|_2 \le C_1\,\varepsilon, \tag{37}$$

where

$$C_1 = (2+2\cdot 1.02)(0.29\,\widehat{\kappa}+5)(0.28\,\widehat{\kappa}+0.84) \le 0.33\,\widehat{\kappa}^2 + 6.65\,\widehat{\kappa} + 16.97.$$

Similarly, from (35) and the assumption $\widehat{a}/\widehat{b} < 2$, we have

$$\sqrt{\frac{a}{b}} < \sqrt{1.03\cdot 2} \le 1.44.$$

By inserting this, (32), (22) and (31) into (28), we have

$$\|\delta G_{\cdot j}\|_2 \le C_2\sqrt{b}\,\varepsilon, \qquad \|\delta B_{\cdot j}\|_2 \le C_2\,\varepsilon,$$

where

$$C_2 = (2+2\cdot 1.44)(0.29\,\widehat{\kappa}+5)(0.28\,\widehat{\kappa}+0.84).$$

By comparing this with (37), we see that $C_2 \le 1.21\,C_1$, which finally gives

$$\|\delta B\|_2 \le \|\delta B\|_F \le \sqrt{C_1^2+C_2^2}\,\varepsilon \le \sqrt{1+1.21^2}\,C_1\,\varepsilon \le (0.52\,\widehat{\kappa}^2 + 10.5\,\widehat{\kappa} + 26.7)\,\varepsilon,$$

as desired.

Case 2. Let $\widehat{a}/\widehat{b} \ge 2$. This case is easier to analyze since $\bar{\zeta}$ and $\bar{t}$ from (19) are bounded away from the respective worst-case bounds (20) and (21). However, in this case there is no upper bound for $\sqrt{a/b}$ in (28). Instead, we use the identity $\widetilde{ch}\,|\widetilde{sh}| = \widetilde{ch}^2|\widehat{t}|$, which transforms (28) to

$$\|\delta G_{\cdot j}\|_2 \le \left(|\varepsilon_4'|\,\widetilde{ch}^2 + |\varepsilon_2'|\,\widetilde{sh}^2 + (|\varepsilon_1'|+|\varepsilon_3'|)\widetilde{ch}^2\,|\widehat{t}|\,\sqrt{\frac{a}{b}}\right)\sqrt{b}, \tag{38}$$

and bound the term $|\widehat{t}|\sqrt{a/b}$.

We first compute the bounds for $\widetilde{ch}$ and $|\widetilde{sh}|$. Using (19), positive definiteness of the matrix $\widehat{A}^{(i,j)}$, and the assumption $\widehat{a}/\widehat{b} \ge 2$, we have

$$|\bar{\zeta}| \ge \frac{\widehat{a}+\widehat{b}}{2\sqrt{\widehat{a}\,\widehat{b}}} = \frac{1}{2}\left(\sqrt{\frac{\widehat{a}}{\widehat{b}}} + \sqrt{\frac{\widehat{b}}{\widehat{a}}}\right) \ge \frac{1}{2}\left(\sqrt{2} + \frac{1}{\sqrt{2}}\right) \ge 1.06.$$

In the last inequality we have used the fact that $x^{1/2} + x^{-1/2}$ is a continuous function with minimum at $x = 1$.

Further, ε_ζ, ε_s and ε_u are estimated as in Case 1, while for ε_v holds (c.f. (29))

$$|\varepsilon_v| \leq 4.02\,\varepsilon + \frac{6.54 \cdot 1.06}{\sqrt{1.06^2 - 1}}\,\varepsilon \leq 24\,\varepsilon.$$

Using this, (30), and the assumption $10\,\varepsilon \leq 0.01$, we have

$$|\widehat{t}| \leq \frac{1 + 0.001}{(1 - 24 \cdot 0.001)(1.06 + \sqrt{1.06^2 - 1})} \leq 0.73.$$

Thus,

$$\widetilde{ch} \leq 1.47, \qquad |\widetilde{sh}| \leq 1.07. \tag{39}$$

Now we compute the bounds for ε_{sh} and ε_{ch}, and insert them into (25). Similarly as in Case 1, we have

$$\begin{aligned}
fl(1 - \widehat{t}^2) &= (1 + \varepsilon_w)(1 - \widehat{t}^2), \qquad |\varepsilon_w| \leq \frac{2.01\,\varepsilon}{1 - 0.73^2} \leq 4.31\,\varepsilon, \\
fl(\sqrt{1 - \widehat{t}^2}) &= (1 + \varepsilon_h)\sqrt{1 - \widehat{t}^2}, \qquad |\varepsilon_h| \leq 3.16\,\varepsilon, \\
\widehat{ch} &= (1 + \varepsilon_{ch})\widetilde{ch}, \qquad |\varepsilon_{ch}| \leq 4.18\,\varepsilon, \\
\widehat{sh} &= (1 + \varepsilon_{sh})\widetilde{sh}, \qquad |\varepsilon_{sh}| \leq 4.18\,\varepsilon.
\end{aligned}$$

The last two bounds and $\widehat{\kappa} \geq 1$ justify the assumptions made in deriving (25). Inserting these bounds into (25) gives

$$|\varepsilon_i'| \leq 6.2\,\varepsilon, \qquad i = 1, 2, 3, 4. \tag{40}$$

Using this, (39), (34) and $\widehat{b}/\widehat{a} \leq 1/2$ in the relation (27), one obtains

$$\|\delta G_{\cdot i}\|_2 \leq 34.5\sqrt{a}\,\varepsilon, \qquad \|\delta B_{\cdot i}\|_2 \leq 34.5\,\varepsilon. \tag{41}$$

Further, similarly to (34), the relations (33) also imply

$$\frac{\widehat{b}}{\widehat{a}} \leq \frac{b}{a} \cdot \frac{1 + 1.01\,m\,\varepsilon}{1 - 1.01\,m\,\varepsilon} \leq 1.03\,\frac{b}{a}.$$

Using this, (30) and the positive definiteness of the matrix $\widehat{A}^{(i,j)}$, we have

$$\begin{aligned}
|\widehat{t}| &\leq \frac{1 + \varepsilon}{(1 - |\varepsilon_v|)\,|\widehat{\zeta}|} = 1.03\,\frac{2|\widehat{c}|}{\widehat{a} + \widehat{b}} \leq 2.06\,\frac{\sqrt{\widehat{a}\,\widehat{b}}}{\widehat{a}} \\
&= 2.06\sqrt{\frac{\widehat{b}}{\widehat{a}}} \leq 2.06\sqrt{1.03}\sqrt{\frac{b}{a}} \leq 2.1\sqrt{\frac{b}{a}}.
\end{aligned}$$

Therefore, in (38) we will have

$$|\widehat{t}|\sqrt{\frac{a}{b}} \leq 2.1.$$

Inserting this, (40) and (39) into (38), we obtain

$$\|\delta G_{\cdot j}\|_2 \leq 77\sqrt{b}\,\varepsilon, \qquad \|\delta B_{\cdot j}\|_2 \leq 77\,\varepsilon.$$

Finally, from this and (41), we have

$$\|\delta B\|_2 \leq \|\delta B\|_F \leq 85\,\varepsilon,$$

and the theorem is proved. ■

2.4 Overall error bounds

The overall error bounds for the HSVD computed by Algorithm 1 are obtained by plugging the one-step error analysis of Theorem 1 into the perturbation bounds of Section 2.1.

The proof requires the following lemma due to Veselić:

Lemma 1 *Let*

$$B^T B = I + E, \qquad \|E\|_2 = \epsilon < 1,$$

where B is any real matrix with full column rank. Then there exists a matrix Q such that $Q^T Q = I$ and $\|B - Q\|_2 \leq \epsilon$.

PROOF. We make the polar decomposition $B = QP$ where $Q^T Q = I$ and P is Hermitian positive definite matrix. Since $QQ^T B = B$, 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

$$\|B - Q\|_2 = \|QP - Q\|_2 = \|P - I\|_2,$$

and the lemma is proved. ■

The error in the computed hyperbolic singular values is bounded as follows.

Theorem 2 *Let G_k, $0 \leq k \leq M$, be the sequence of matrices computed by Algorithm 1 from the starting pair (G, J). Here $G_0 \equiv G$. Assume that Algorithm 1 converges, and that (G_M, J) is the final pair which satisfies the stopping criterion. For $0 \leq k \leq M$ let $G_k = B_k D_k$ be scaled according to (6). Let σ_i be the i-th singular value of the pair (G_0, J), and let $\sigma_i' = fl(\|G_{M,\cdot i}\|_2)$ be the i-th computed singular value. Assume that the assumptions of Theorem 1 are satisfied in each step of Algorithm 1, and let C_k denote the constant C from Theorem 1 in the k-th step. If, additionally, $\max\{n \cdot tol, m\, n\, \varepsilon\} \leq 0.01$, then*

$$1 - \beta \leq \frac{\sigma_i'}{\sigma_i} \leq 1 + \beta,$$

where

$$\beta = \left[\prod_{k=0}^{M-1} \left(1 + \frac{C_k}{\sigma_{\min}(B_k)}\right)\right] (1 + 1.05\, n \cdot tol + 1.05\, m\, n\, \varepsilon)(1 + (0.51\, m + 1.01)\, \varepsilon) - 1,$$

provided $\beta < 1$.

PROOF. Let $\sigma_{M,i}$ be the hyperbolic singular values of the final pair (G_M, J). Since

$$\frac{\sigma_i'}{\sigma_i} = \frac{\sigma_{M,i}}{\sigma_i} \cdot \frac{\|G_{M,\cdot i}\|_2}{\sigma_{M,i}} \cdot \frac{\sigma_i'}{\|G_{M,\cdot i}\|_2}, \tag{42}$$

we shall compute the bound for β in three steps.

According to Theorem 1, for every $0 \leq k \leq M - 1$ we have

$$G_{k+1} = (G_k + \delta G_k)\tilde{J}_k, \tag{43}$$

where

$$\delta G_k = \delta B_k D_k, \qquad \|\delta B_k\|_2 \leq C_k\, \varepsilon.$$

Here $\widetilde{J}_k$ is the exact rotation from the commutative diagram of Theorem 1 in the k-th step. Further, for every $0 \le k \le M-1$ we can write

$$G_{k+1} = (G + \delta G^{(k)})\widetilde{J}_0 \cdot \widetilde{J}_1 \cdots \widetilde{J}_k, \tag{44}$$

that is, we interpret G_{k+1} as being obtained by a sequence of exact transformations applied to a perturbed starting matrix G. The proof is by induction on k. For $k = 0$ we simply set $\delta G^{(0)} = \delta G_0$. Now suppose that (44) holds for some $k \ge 1$. By (43) and the induction assumption we have

$$\begin{aligned} G_{k+1} &= (G_k + \delta G_k)\widetilde{J}_k \\ &= [(G + \delta G^{(k-1)})\widetilde{J}_0 \cdots \widetilde{J}_{k-1} + \delta G_k]\widetilde{J}_k \\ &= (G + \delta G^{(k)})\widetilde{J}_0 \cdots \widetilde{J}_k, \end{aligned}$$

where

$$\delta G^{(k)} = \delta G^{(k-1)} + \delta G_k (\widetilde{J}_0 \cdots \widetilde{J}_{k-1})^{-1}. \tag{45}$$

Set

$$\delta B_k = \delta G_k D_k^{-1}, \qquad \delta B^{(k)} = \delta G^{(k)} D^{-1},$$

where $D = D_0$. Then for every $0 \le k \le M-1$

$$\|\delta B^{(k)} B^\dagger\|_2 \le \prod_{l=0}^{k} \left(1 + \frac{C_l}{\sigma_{\min}(B_l)}\varepsilon\right) - 1. \tag{46}$$

The proof is by induction on k. For $k = 0$ the statement follows from Theorem 1 since

$$\|\delta B^{(0)} B^\dagger\|_2 = \|\delta B_0 B_0^\dagger\|_2 \le \frac{\|\delta B_0\|_2}{\sigma_{\min}(B_0)} \le \frac{C_0}{\sigma_{\min}(B_0)}\varepsilon.$$

Now suppose that (46) holds for some $k \ge 1$. Writing (45) for $k+1$ and post-multiplying it by $D^{-1}B^\dagger$ gives

$$\begin{aligned} \delta B^{(k+1)} B^\dagger &= \delta B^{(k)} B^\dagger + \delta G_{k+1} (\widetilde{J}_0 \cdots \widetilde{J}_k)^{-1} D^{-1} B^\dagger \\ &= \delta B^{(k)} B^\dagger + \delta B_{k+1} B_{k+1}^\dagger B_{k+1} D_{k+1} (\widetilde{J}_0 \cdots \widetilde{J}_k)^{-1} D^{-1} B^\dagger \\ &= \delta B^{(k)} B^\dagger + \delta B_{k+1} B_{k+1}^\dagger G_{k+1} (\widetilde{J}_0 \cdots \widetilde{J}_k)^{-1} D^{-1} B^\dagger \\ &= \delta B^{(k)} B^\dagger + \delta B_{k+1} B_{k+1}^\dagger (G + \delta G^{(k)}) D^{-1} B^\dagger \\ &= \delta B^{(k)} B^\dagger + \delta B_{k+1} B_{k+1}^\dagger (B\, B^\dagger + \delta B^{(k)} B^\dagger). \end{aligned}$$

Taking norms and using Theorem 1 gives

$$\|\delta B^{(k+1)} B^\dagger\|_2 \le \|\delta B^{(k)} B^\dagger\|_2 + \frac{C_{k+1}\varepsilon}{\sigma_{\min}(B_{k+1})}(1 + \|\delta B^{(k)} B^\dagger\|_2).$$

Finally, inserting the induction assumption and rearranging completes the proof of (46).

By using (46) for $k = M-1$, and setting $\delta G \equiv \delta G^{(M-1)}$ and $\delta B = \delta G D^{-1}$, we have

$$G_M = (G + \delta G)\widetilde{J}_0 \cdots \widetilde{J}_{M-1}, \tag{47}$$

where

$$\beta_M \equiv \prod_{k=0}^{M-1} \left(1 + \frac{C_k}{\sigma_{\min}(B_k)}\varepsilon\right) - 1 \ge \|\delta B B^\dagger\|_2. \tag{48}$$

Then, according to (11) and (10), we have

$$1 - \beta_M \leq \frac{\sigma_{M,i}}{\sigma_i} \leq 1 + \beta_M. \tag{49}$$

We have, therefore, proved the first part of the expression for β.

Now we have to account for two more facts:

- the columns of G_M are not exactly orthogonal; instead G_M numerically satisfies the stopping criterion,
- final singular values σ_i' are numerically computed norms of the columns of G_M.

First notice that

$$G_M = B_M D_M, \qquad D_{M,ii} = \|G_{M,\cdot i}\|_2.$$

Thus,

$$B_M^T B_M = I + E, \qquad E_{ii} = 0, \qquad E_{ij} = \frac{\sum_k G_{M,ki} G_{M,kj}}{\|G_{M,\cdot i}\|_2 \|G_{M,\cdot j}\|_2}, \quad i \neq j. \tag{50}$$

For the sake of simplicity we set[2]

$$a = \|G_{M,\cdot i}\|_2^2, \qquad b = \|G_{M,\cdot j}\|_2^2, \qquad c = \sum_k G_{M,ki} G_{M,kj}.$$

The classical error analysis of the scalar product (see, [14, § 2.4]) implies (33) and

$$|fl(c) - c| \leq 1.01\, m\, \varepsilon \sum_k |G_{M,ki}|\, |G_{M,kj}| \leq 1.01\, m\, \varepsilon \sqrt{a\, b}.$$

Since G_M numerically satisfies the stopping criterion, by this and (33) we have

$$fl\left(\frac{|c|}{\sqrt{a\, b}}\right) = \frac{|c + (fl(c) - c)|}{(1+\varepsilon_1)\sqrt{a\,(1+\varepsilon_a)\, b\,(1+\varepsilon_b)(1+\varepsilon_2)}}(1+\varepsilon_3) \leq tol.$$

Therefore, for $i \neq j$ we have

$$\begin{aligned} |E_{ij}| &= \frac{|c|}{\sqrt{a\, b}} \leq \frac{(1+\varepsilon)\sqrt{(1+1.01\, m\, \varepsilon)^2(1+\varepsilon)}}{(1-\varepsilon)}\, tol + \frac{|(fl(c) - c|}{\sqrt{a \cdot b}} \\ &\leq 1.02\, tol + 1.01\, m\, \varepsilon. \end{aligned}$$

Here we have used the assumption $\max\{m, 10\} \leq \varepsilon$. Thus,

$$\|E\|_2 \leq 1.02\, n\, tol + 1.01\, n\, m\, \varepsilon.$$

From this, (50) and Lemma 1, there exits an orthonormal matrix $\bar{B}$ such that

$$\bar{B} = B_M + \delta \bar{B}, \qquad \|\delta \bar{B}\|_2 \leq 1.02\, n\, tol + 1.01\, n\, m\, \varepsilon.$$

Set $\bar{G} = \bar{B} D_M$. Since the columns of $\bar{G}$ are orthogonal, the hyperbolic singular values of the pair $(\bar{G}, J)$ are $\bar{\sigma}_i = D_{M,ii} = \|G_{M,\cdot i}\|_2$. Thus, (11) implies

$$1 - \bar{\beta} \leq \frac{\|G_{M,\cdot i}\|_2}{\sigma_{M,i}} \leq 1 + \bar{\beta}, \tag{51}$$

[2] These a, b and c are different than the ones from Theorem 1.

where

$$\bar{\beta} \leq \frac{\|\delta\bar{B}\|_2}{\sigma_{\min}(B_M)} \leq \frac{\|\delta\bar{B}\|_2}{1-\|\delta\bar{B}\|_2} \leq 1.05\, n\, tol + 1.04\, n\, m\, \varepsilon.$$

In the last inequality we have used the assumption $\max\{n \cdot tol, m\, n\, \varepsilon\} \leq 0.01$. This completes the proof of the second part of the bound for β.

Finally, we have to account for the difference between $\|G_{M,\cdot i}\|_2$ and $\sigma_i' = fl(\|G_{M,\cdot i}\|_2)$. We have

$$\sigma_i' = fl(\|G_{M,\cdot i}\|_2) = (1+\varepsilon_4)\sqrt{\|G_{M,\cdot i}\|_2^2(1+\varepsilon_a)} = (1+\varepsilon')\,\|G_{M,\cdot i}\|_2,$$

where $|\varepsilon_a| \leq 1.01\, m\, \varepsilon$ as in (33), and, consequently, $|\varepsilon'| \leq (0.51\, m + 1.01)\, \varepsilon$. Therefore,

$$1 - (0.51\, m + 1.01)\, \varepsilon \leq \frac{\sigma_i'}{\|G_{M,\cdot i}\|_2} \leq 1 + (0.51\, m + 1.01)\, \varepsilon.$$

The theorem follows by combining this, (51) and (49) with (42). ■

We have two remarks. First, notice that the first order approximation for β reads

$$\beta = \varepsilon \sum_{k=0}^{M-1} \frac{C_k}{\sigma_{\min}(B_k)} + 1.05\, n \cdot tol + 1.05\, m\, n\, \varepsilon + (0.51\, m + 1.01)\, \varepsilon + O(\varepsilon^2), \tag{52}$$

which is the form that was used in [9]. Second, in Theorem 1 both $\|\delta B\|_2$ and $\|\delta B\|_F$ are bounded by $C\,\varepsilon$. By repeating the part of the proof of Theorem 2 between (43) and (48) for Frobenius norm, we easily see that (48) holds for the Frobenius norm, as well, that is

$$\beta_M \geq \|\delta B B^{\dagger}\|_F. \tag{53}$$

We need this result to prove our singular vector bounds.

The errors in the singular vectors are bounded as follows.

Theorem 3 *Assume Algorithm 1 converges, and that (G_M, J) is the final pair which satisfies the stopping criterion. Let $G = U\Sigma V^{-1}$ and $G_M = U'\Sigma'(V')^{-1}$ be the HSVDs of the pairs (G, J) and (G_M, J), respectively, partitioned according to (12). Let σ_i and σ_i' be the diagonal entries of Σ and Σ', respectively. Here σ_i and σ_i' may be in any, but same, order. Let $\mathcal{U}_1$ and $\mathcal{U}_1'$ be the subspaces spanned by the columns of U_1 and U_1', respectively, and let $\mathcal{V}_1$ and $\mathcal{V}_1'$ be the subspaces spanned by the columns of V_1 and V_1', respectively. For $0 \leq k \leq M$ let $G_k = B_k D_k$ be scaled according to (6), and let C_k denote the constant C from Theorem 1 in the k-th step. Finally, let* [3]

$$\beta = \prod_{k=0}^{M-1} \left(1 + \frac{C_k}{\sigma_{\min}(B_k)}\,\varepsilon\right) - 1, \qquad \psi = \frac{3\,\beta}{\sqrt{1-3\,\beta}},$$

and let $\mathrm{rg}(\Sigma_1', \Sigma_2)$ be defined according to (13). Then,

$$\|\sin\Theta(\mathcal{U}_1, \mathcal{U}_1')\|_F \leq \frac{2\,\beta}{1-\beta} \cdot \frac{1}{\mathrm{rg}(\Sigma_1', \Sigma_2)}, \tag{54}$$

$$\|\sin\Theta(\mathcal{V}_1, \mathcal{V}_1')\|_F \leq \|V\|_2^2 \left(\frac{1}{2}\psi + \sqrt{1+\frac{1}{4}\psi^2}\right) \frac{\psi}{\mathrm{rg}(\Sigma_1', \Sigma_2)}. \tag{55}$$

[3] Notice that $\beta = \beta_M$, where β_M is defined in (48).

PROOF. The first bound follows by inserting (47), (48) and (53) into (14), and the second bound follows by inserting (47), (48) and (53) into (15). ■

Let us give some remarks concerning the practical application of the above theorems. Theorem 3 is incomplete in the sense that we ignore the fact that the bounds hold for the exact singular vectors of the final pair (G_M, J), and not for the actually computed ones. More precisely, in (54) we ignore the fact that the computed left singular vectors are the normalized columns of the final matrix (G_M, J). In (55) we ignore the round-off errors which occur in the updating the columns of V in Algorithm 1, as well as the errors which are due to the fact that these updates are performed with slightly perturbed rotation matrices. It is possible to include these details, but they are technically very demanding. We decided not to do so since the bounds of Theorem 3 show well the essential behavior of the errors in the computed singular vectors, and including these details would greatly complicate the exposition.

Another important issue are the factors $1/\sigma_{\min}(B_m)$ which appear in both theorems. Clearly, B_m changes from step to step, and so does this factor. However, as Algorithm 1 converges, $1/\sigma_{\min}(B_m) \to 1$. Also, there is strong numerical evidence in previous works [9, 10, 23] and in our numerical experiments that this factor does not grow much during the computation. The theoretical understanding of this phenomenon is weaker. Some (partial) theoretical results can be found in [9, 23]. In [23, §3.2.2], an algorithm was derived with which upper bound for $1/\sigma_{\min}(B_m)$ can be efficiently monitored.

From the above comments, and the fact that the constants in Theorems 1 and 2 come from considering worst cases, we conclude that the error in the computed hyperbolic singular values should be bounded by

$$\frac{|\sigma_i' - \sigma_i|}{\sigma_i} \leq \varepsilon \, \frac{1}{\sigma_{\min}(B)} \, f_\sigma(m,n), \tag{56}$$

where $f_\sigma(m,n)$ is a factor which moderately grows with dimensions.

In numerical experiments we focus our attention to the individual singular vectors. Let u_i be the left exact singular vector of σ_i, and let $u_i' = u_i + \delta u_i$ be the corresponding left computed singular vector. Similarly as above, from (54) we expect the error in the computed left singular vectors to be bounded by

$$\|\delta u_i\|_2 \leq \varepsilon \, \frac{1}{\sigma_{\min}(B)} \cdot \frac{1}{\mathrm{rg}(\sigma_i', \Sigma_2')} \, f_u(m,n), \tag{57}$$

where $f_u(m,n)$ is a factor which moderately grows with dimensions. Notice that this is just the bound (56) divided by the corresponding relative gap. Also, since the exact singular values σ_i are not available, here we use the relative gap which is defined by using only the computed singular values. We can bound the errors introduced in the relative gap in this manner by Theorem 2. However, this is unnecessary since the bound (57) depicts well the actual error (see Table 1).

Similarly, if v_i and $v_i' = v_i + \delta v_i$ are the exact and the computed right singular vectors of σ_i, respectively, by (55) we expect that the error is bounded by

$$\|\delta v_i\|_2 \leq \varepsilon \, \|V'\|_2^2 \, \frac{1}{\sigma_{\min}(B)} \cdot \frac{1}{\mathrm{rg}(\sigma_i', \Sigma_2')} \, f_v(m,n), \tag{58}$$

where $f_v(m,n)$ is a factor which moderately grows with dimensions. Here V' is the computed right singular vector matrix, which is readily available upon completion of Algorithm 1.

2.5 Numerical experiments

We performed series of experiments on randomly generated test pairs (G, J). For each test pair we first computed the HSVD by Algorithm 1 in double precision and assumed that to be the exact solution. Then we solved the same problem by the single precision version of Algorithm 1, and verified that the expected error bounds (56), (57) and (58) are satisfied. Our programs are written in Fortran, compiled by GNU `g77` Fortran compiler, and executed on a Pentium III 866 Linux machine. In generating test pairs we have used the LAPACK [2] random number generator `dlaran.f`.

We first describe the procedure used in generating test pairs, and the sets of parameters used. Then we show the results of our experiments. Besides results concerning the accuracy, we also show the number of cycles which were executed until the convergence.

For given dimensions m and n, we first generate random diagonal matrix D_0 whose diagonal entries' logarithm is uniformly distributed in the interval $[-\beta/2, \beta/2]$. We then form matrix $G_0 = Q_1 D_0 Q_2$ where Q_1 and Q_2 are random orthonormal matrices of dimensions $m \times n$ and $n \times n$, respectively. Further, we generate random diagonal matrix D_1 whose diagonal entries' logarithm is uniformly distributed in the interval $[-\gamma/2, \gamma/2]$. We then form the matrix $G = G_0 D_1$. Thus, $\kappa(B) \approx 10^{\beta}$, where $G = BD$ is scaled according to (6). Finally, we generate random $n \times n$ diagonal matrix J with elements in the set $\{-1, 1\}$.

We tested matrices for $m = 50, 100, 200, 400$, and for each m we used $n = m/2, m$, which gives eight classes of matrices. Further, we chose $\beta = 1, 2, 3, 4$ and $\gamma = 2, 4, 6, 8, 10, 12, 14$. This gives a total of 224 classes of matrices. In each class we constructed 60 test pairs, which totals to 13440 experiments.

The results are as follows. Here σ_i, u_i and v_i denote the singular values and vectors computed in double precision, and σ_i', u_i' and v_i' denote the singular values and vectors computed in single precision For each experiment we computed the maximal factors $f_\sigma(n)$, $f_u(n)$ and $f_v(n)$ according to (56), (57) and (58), respectively, that is

$$
\begin{aligned}
f_\sigma(n) &= \max_{i=1,\ldots,n} \frac{|\sigma_i' - \sigma_i|}{\sigma_i} \Big/ \frac{\varepsilon}{\sigma_{\min}(B)}, \\
f_u(n) &= \max_{i=1,\ldots,n} \|\delta u_i\|_2 \Big/ \left(\frac{\varepsilon}{\sigma_{\min}(B)} \cdot \frac{1}{\mathrm{rg}(\sigma_i', \Sigma_2')} \right), \\
f_v(n) &= \max_{i=1,\ldots,n} \|\delta v_i\|_2 \Big/ \left(\|V'\|_2^2 \, \frac{\varepsilon}{\sigma_{\min}(B)} \cdot \frac{1}{\mathrm{rg}(\sigma_i', \Sigma_2')} \right)
\end{aligned}
$$

The behavior of $f_\sigma(n)$, $f_u(n)$ and $f_v(n)$ is shown in Table 1.

n	50	100	200	400
mean $f_\sigma(n)$	1.82	3.30	6.23	12.2
max $f_\sigma(n)$	14.9	26.0	53.3	104.6
mean $f_u(n)$	3.67	7.92	16.3	32.6
max $f_u(n)$	26.4	59.6	139.4	333.3
mean $f_v(n)$	0.656	1.35	3.00	6.61
max $f_v(n)$	5.36	8.48	18.1	35.8

Table 1: Error factors in the computed HSVD in 13440 experiments

We see that the expectations given in (56), (57) and (58) are fully confirmed by numerical experiments. Thus, we may conclude that it is indeed the scaled matrix B, and not the starting

matrix G which governs the accuracy of the computed HSVD.

Further, in each experiment we monitored the number of cycles executed before convergence, and the spectral condition of the right singular vector matrix, $\kappa(V') = \|V'\|_2^2$. The results are in Table 2

n	50	100	200	400
mean$(cycles)$	8	9	10	11
max$(cycles)$	13	15	16	18
mean $\kappa(V')$	4.27	4.44	4.46	4.45
max $\kappa(V')$	48.4	29.4	23.5	23.3

Table 2: Number of cycles and $\kappa(V')$ in 13440 experiments

3 Symmetric eigenvalue decomposition

We consider the classical symmetric eigenvalue problem

$$Hx = \lambda x, \qquad x \neq 0, \tag{59}$$

where H is a $n \times n$ non-singular matrix. The eigenvalue decomposition of H will be denoted by

$$H = U \Lambda U^T,$$

where Λ is diagonal matrix whose diagonal entries are the eigenvalues of H, and U is orthonormal matrix whose columns are the corresponding eigenvectors. As already mentioned in the introduction, we use the following algorithm:

Algorithm 2 *Eigenvalue decomposition of a non-singular symmetric matrix H.*

1. *Factorize H as*
$$P^T H P = G_1 J G_1^T, \tag{60}$$
where P is a permutation matrix, G_1 is non-singular lower block triangular matrix with 1×1 and 2×2 diagonal blocks, and J is diagonal matrix of signs, $J_{ii} \in \{-1, 1\}$.

2. *Compute the hyperbolic singular values σ_i and the left singular vector matrix U of the pair (G, J), where $G = PG_1$, by using Algorithm 1.*

The eigenvalues of H are $\lambda_i = \sigma_i^2 J_{ii}$, and the columns of U are the corresponding eigenvectors.

The aim of this section is to show that Algorithm 2 computes the eigenvalue decomposition (59) with high relative accuracy. We first state the error bounds for the first step of the algorithm, originally proved in [23, 24]. In §3.1 we then state the relative perturbation results for the eigenvalues and eigenvectors of the problem (59). In §3.2 we give overall error bounds for the eigensolution computed by Algorithm 2, and in §3.3 we describe results of our numerical experiments which confirm the theoretical predictions.

Detailed description and the formal algorithm, as well as the error analysis of the symmetric indefinite factorization (60) are given in [24, §2 and §3]. This factorization is, in fact, a modification of the well-known Bunch–Parlett factorization [4]. The variant of the Bunch–Parlett

factorization with partial pivoting is implemented in the LAPACK routine `dsytf2.f` [2]. The factorization (60) uses the original unequilibrated diagonal pivoting from [4], which defines the permutation matrix P.

The error bound for the factorization (59) was proved in [24, Th. 3.1]: the factors $G = PG_1$ and J computed in floating-point arithmetic with precision ε are the exact factors of some perturbed matrix $H + \delta H$, that is,

$$GJG^T = H + \delta H, \qquad |\delta H| \leq 91n(|H| + |G||G|^T)\varepsilon + O(\varepsilon^2). \tag{61}$$

3.1 Relative perturbation bounds

We shall use the relative perturbation bounds for the non-singular symmetric eigenvalue problem from [32, 29]. The bounds are stated in terms of the spectral absolute value of H, $|H| = \sqrt{H^2}$. Notice that $|H|$ is, in fact, the positive definite polar factor of H. Let the scaled matrix $\widehat{A}$ be defined by

$$|H| = \widehat{D}\widehat{A}\widehat{D},$$

where $\widehat{D}$ is some non-singular diagonal matrix. Further, let $H + \delta H$ be the perturbed matrix, where

$$\delta H = \widehat{D}\delta A\widehat{D}.$$

According to [32, Th. 2.1], if δH is such that $|x^T\delta H x| \leq \eta x^T|H|x$ for all vectors x and some $\eta < 1$, then the eigenvalues of the matrices H and $H + \delta H$, λ_i and $\widetilde{\lambda}_i$, respectively, satisfy the inequalities

$$1 - \eta \leq \frac{\widetilde{\lambda}_i}{\lambda_i} \leq 1 + \eta. \tag{62}$$

Since

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

if δA is known, then (62) holds with η defined by

$$\eta = \frac{\|\delta A\|_2}{\lambda_{\min}(\widehat{A})}. \tag{63}$$

Further, if $H = GJG^T$ and $G = U\Sigma V^{-1}$ is the HSVD of the pair (G, J), then $\Sigma = |\Lambda|^{1/2}$, $U = GV|\Lambda|^{-1/2}$, and

$$GVV^TG^T = U|\Lambda|^{1/2}|\Lambda|^{1/2}U^T = U|\Lambda|U^T = |H|.$$

Thus, we may rewrite (63) as

$$\eta = \frac{\|\delta A\|_2}{\sigma^2_{\min}(\widehat{D}^{-1}GV)}. \tag{64}$$

This is convenient way to apply the bound (62), since an approximation of the matrix GV is readily available upon completion of Algorithm 2 – this is the final matrix G_M of Algorithm 1 and Theorem 2. Usual choice for the matrix $\widehat{D}$ is such that the matrix $\widehat{D}^{-1}G$ has unit rows.

In order to state the eigenvector bound, let us partition the eigenvalue decomposition $H = U\Lambda U^T$ as

$$H = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}, \tag{65}$$

where U_1 is $n \times k$ matrix, U_2 is $n \times (n-k)$ matrix, and the rest of the matrices have the corresponding dimensions. Let the perturbed matrix $\widetilde{H} = H + \delta H = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^T$ be partitioned accordingly. Similarly as in §2.1, we define the relative gap by

$$\mathrm{rg}_1(\widetilde{\Lambda}_1, \Lambda_2) = \min_{\substack{1 \le p \le k \\ k+1 \le q \le n}} \frac{|\widetilde{\lambda}_p - \lambda_q|}{\sqrt{|\widetilde{\lambda}_p \lambda_q|}}. \tag{66}$$

Let $\mathcal{U}_1$ and $\widetilde{\mathcal{U}}_1$ be the subspaces spanned by the columns of U_1 and $\widetilde{U}_1$, respectively. By using [29, Th. 6], one can easily prove that

$$\| \sin\Theta(\mathcal{U}_1, \widetilde{\mathcal{U}}_1) \|_F \le \frac{\|V\|_2^2}{\sqrt{1 - 4\,\alpha \|V\|_2^2}} \cdot \frac{\gamma_F}{\sqrt{1-\gamma}} \cdot \frac{1}{\mathrm{rg}_1(\widetilde{\Lambda}_1, \Lambda_2)}, \tag{67}$$

where

$$\gamma = \frac{\|\delta A\|_2}{\sigma_{\min}^2(\widehat{D}^{-1}G)}, \qquad \gamma_F = \frac{\|\delta A\|_F}{\sigma_{\min}^2(\widehat{D}^{-1}G)}, \qquad \alpha = \frac{\gamma_F}{2 - 3\,\gamma}.$$

3.2 Overall error bounds

The error bounds for the eigensolution computed by Algorithm 2 are obtained by adding the error bounds for the first and the second step. More precisely, the error bounds are obtained by inserting the error bound (61) into perturbation bounds (62), (64) and (67), and adding the error bounds for the HSVD from §2.4. In §2.5 we have seen that the actual errors in computed HSVD behave like the first order approximations of the bounds which were proved in §2.4. Having this in mind, for the sake of simplicity, here we shall state and prove only first order bounds.

The error in computed eigenvalues is bounded as follows:

Theorem 4 *Let λ_i' be the eigenvalues of the matrix H computed by Algorithm 2 in floating-point arithmetic with precision ε, and let λ_i be the exact eigenvalues of H in the same order. Let (G, J) be the output of the first step of Algorithm 2, and let $\widehat{D}$ be the positive definite diagonal matrix such that the matrix $\widehat{B} = \widehat{D}^{-1}G$ has unit rows. For $0 \le k \le M$ let $G_k = B_k D_k$ be the sequence of matrices generated by Algorithm 1, starting from $G = G_0$, scaled according to (6). Assume that the assumptions of Theorems 1 and 2 are satisfied in each step of Algorithm 1, and assume, additionally, that $n\,\varepsilon \le 0.001$. Let C_k denote the constant C from Theorem 1 in the k-th step of Algorithm 1. Finally, let β be defined by (52). Then*

$$1 - \eta \le \frac{\lambda_i'}{\lambda_i} \le 1 + \eta,$$

where

$$\eta = 201\, n^2 \frac{1}{\sigma_{\min}^2(\widehat{B}V)}\, \varepsilon + 2\,\beta + O(\varepsilon^2).$$

Proof. Set $\widehat{H} = GJG^T$. Then $\widehat{H} = H + \delta H$, where δH is bounded by (61). Further, by inserting

$$|H| \le |GJG^T| + |\delta H| \le |G|\,|G|^T + |\delta H|,$$

into (61) we have

$$\begin{aligned} |\delta H| &\leq 91\,n\,(|H| + |G|\,|G|^T)\,\varepsilon + O(\varepsilon^2) \\ &\leq 91\,n\,(|G|\,|G|^T + |\delta H| + |G|\,|G|^T)\,\varepsilon + O(\varepsilon^2), \end{aligned}$$

or

$$(1 - 91\,n\,\varepsilon)|\delta H| \leq 182\,n\,|G|\,|G|^T\,\varepsilon + O(\varepsilon^2).$$

Dividing this inequality by $1 - 91\,n\,\varepsilon$ and using the assumption $n\,\varepsilon \leq 0.001$ gives

$$|\delta H| \leq 201\,n\,|G|\,|G|^T\,\varepsilon + O(\varepsilon^2).$$

Set $\delta A = \widehat{D}^{-1}\delta H\,\widehat{D}^{-1}$. Then

$$|\delta A| \leq 201\,n|\widehat{B}|\,|\widehat{B}|^T\,\varepsilon + O(\varepsilon^2). \tag{68}$$

Inserting this into (62) and (64) gives

$$1 - \widehat{\eta} \leq \frac{\widehat{\lambda}_i}{\lambda_i} \leq 1 + \widehat{\eta},$$

where $\widehat{\lambda}_i$ are the eigenvalues of $\widehat{H}$, and

$$\widehat{\eta} = \frac{\|\,|\delta A|\,\|_2}{\sigma^2_{\min}(\widehat{B}V)} \leq 201\,n^2\,\frac{1}{\sigma^2_{\min}(\widehat{B}V)}\,\varepsilon + O(\varepsilon^2). \tag{69}$$

We have thus proved the first part of η.

Further, we have $\widehat{\lambda}_i = \widehat{\sigma}_i^2 J_{ii}$, where $\widehat{\sigma}_i$ are the hyperbolic singular values of the pair (G, J). Similarly, we can write $\lambda'_i = \sigma_i'^2 J_{ii}$. Since

$$\frac{\lambda'_i}{\widehat{\lambda}_i} = \frac{\sigma_i'^2}{\widehat{\sigma}_i^2},$$

squaring the bound of Theorem 2 gives

$$1 - 2\,\beta + \beta^2 \leq \frac{\lambda'_i}{\widehat{\lambda}_i} \leq 1 + 2\,\beta + \beta^2,$$

where the first order approximation for β is given by (52). The theorem now follows by combining this with (69). ■

The error in eigenvectors is bounded as follows:

Theorem 5 *Let the assumptions of Theorems 4 and 3 hold. Let the eigenvalue decompositions of the matrices $H = U\Lambda U^T$, $\widehat{H} = GJG^T = \widehat{U}\widehat{\Lambda}\widehat{U}^T$ and $H' = G_M J G_M^T = U'\Lambda' U'^T$ be partitioned according to (65). Let $\mathcal{U}_1$ and $\mathcal{U}'_1$ be the subspaces spanned by the columns of U_1 and U'_1, respectively. For $0 \leq k \leq M$ let $G_k = B_k D_k$ be scaled according to (6), and let C_k denote the constant C from Theorem 1 in the k-th step. Let β be defined as in Theorem 3, and let*

$$\gamma = 201\,n^2\,\frac{1}{\sigma^2_{\min}(\widehat{B})}\,\varepsilon + O(\varepsilon^2), \qquad \alpha = \frac{\gamma}{2 - 3\,\gamma}.$$

Let $\mathrm{rg}_1(\widehat{\Lambda}_1, \Lambda_2)$ and $\mathrm{rg}(\Lambda'_1, \widehat{\Lambda}_2) \equiv \mathrm{rg}(\Sigma'_1, \widehat{\Sigma}_2)$ be defined according to (66) and (13), respectively. Then,

$$\|\sin\Theta(\mathcal{U}_1, \mathcal{U}'_1)\|_F \leq \frac{\|V\|_2^2}{\sqrt{1 - 4\,\alpha\|V\|_2^2}} \cdot \frac{\gamma}{\sqrt{1 - \gamma}} \cdot \frac{1}{\mathrm{rg}_1(\widehat{\Lambda}_1, \Lambda_2)} + \frac{2\,\beta}{1 - \beta} \cdot \frac{1}{\mathrm{rg}(\Lambda'_1, \widehat{\Lambda}_2)} + O(\varepsilon^2).$$

PROOF. The theorem follows by inserting (68) into (67), and adding the bound (54). ■

The remarks made in §2.4 after Theorem 3 hold for Theorems 4 and 5, as well. In particular, the bound of Theorem 5 holds for the exact left singular vectors of the final pair (G_M, J), that is, for the exact eigenvectors of the matrix $G_M J G_M^T$, and not for the actually computed ones.

Also, notice that for the matrix $G = BD$ obtained by the first step of Algorithm 2, $1/\sigma_{\min}(B)$ is bounded by a function of $O(3.781^n)$ irrespective of G (see [24, Th. 6.1]). In our experiments $1/\sigma_{\min}(B)$ was never too large, which, together with the bound (16), implies that the quantities $\sigma^2_{\min}(\widehat{B}V)$ and $\sigma^2_{\min}(\widehat{B})$ from Theorems 4 and 5 do not differ by much.

From the above discussion we conclude that the expected error in the computed eigenvalues should be bounded by

$$\frac{|\lambda_i' - \lambda_i|}{\lambda_i} \leq \varepsilon \left(\frac{1}{\sigma^2_{\min}(\widehat{D}^{-1} G_M)} + \frac{1}{\sigma_{\min}(B)} \right) f_\lambda(n), \tag{70}$$

where $f_\lambda(n)$ is a factor which moderately grows with n. Here we have assumed that the matrix G_M is sufficiently good approximation of the matrix GV.

Further, let u_i be the eigenvector of λ_i, and let $u_i' = u_i + \delta u_i$ be the corresponding computed eigenvector. Similarly as above, from Theorem 5 we conclude that the error in the computed left eigenvectors should be bounded by

$$\|\delta u_i\|_2 \leq \varepsilon \frac{1}{\sigma^2_{\min}(\widehat{B})} \cdot \frac{1}{\mathrm{rg}_1(\lambda_i', \Lambda_2')} f_u(n), \tag{71}$$

where $f_u(n)$ is a factor which moderately grows with n. In (71) we also ignored the factor $\|V\|_2^2$ in the first term and the contribution of the second term of the bound of Theorem 5, which is justified by the numerical experiments in the following section (see Table 3).

3.3 Numerical experiments

Similarly as in §2.5, we performed series of experiments on randomly generated test matrices H. For each test matrix we first computed the eigenvalue decomposition by Algorithm 2 in double precision and assumed that to be the exact solution. Then we solved the same problem by the single precision version of Algorithm 2, and verified that the expected error bounds (70) and (71) are satisfied.

Test matrices were generated as follows. For given dimension n, we first generate random diagonal matrix D_0 whose diagonal entries' logarithm is uniformly distributed in the interval $[-\beta/2, \beta/2]$. We then form matrix $A_0 = Q_1 D_0 J Q_1^T$ where Q_1 is random orthonormal matrix and J is random diagonal matrix with $J_{ii} \in \{-1, 1\}$. Further, we generate random diagonal matrix D_1 whose diagonal entries' logarithm is uniformly distributed in the interval $[-\gamma/2, \gamma/2]$. We then form the matrix $H = D_1 A_0 D_1$. Since all matrices are randomly generated, this procedure generates matrix H for which usually $\kappa^2(\widehat{B}) \approx 10^\beta$ and $\kappa(H) \approx 10^{2\gamma}$. More precisely, additional row-scaling of the factor of A_0 does not influence the condition number of that factor, and the condition number of H is primarily determined by $\kappa^2(D_1)$.

We tested matrices for $n = 50, 100, 200, 400$. Further, we chose $\beta = 1, 2, 3, 4$ and $\gamma = 2, 4, 6, 8, 10, 12$. This gives a total of 96 classes of matrices. In each class we constructed 100 test pairs, which totals to 9600 experiments.

The results are as follows. Here λ_i and u_i denote the eigenvalues and eigenvectors computed in double precision, and λ_i' and u_i' denote the eigenvalues and eigenvectors computed in single

precision For each experiment we computed the maximal factors $f_\lambda(n)$ and $f_u(n)$ according to (70) and (71), respectively, that is

$$f_\lambda(n) = \max_{i=1,\ldots,n} \frac{|\lambda_i' - \lambda_i|}{\lambda_i} / \left(\frac{\varepsilon}{\sigma^2_{\min}(\widehat{B}V)} + \frac{\varepsilon}{\sigma_{\min}(B)} \right),$$

$$f_u(n) = \max_{i=1,\ldots,n} \|\delta u_i\|_2 / \left(\frac{\varepsilon}{\sigma^2_{\min}(\widehat{B})} \cdot \frac{1}{\mathrm{rg}_1(\lambda_i', \Lambda_2')} \right).$$

The behavior of $f_\lambda(n)$ and $f_u(n)$ is shown in Table 3.

n	50	100	200	400
mean $f_\lambda(n)$	0.213	0.273	0.417	0.661
max $f_\lambda(n)$	6.10	4.94	6.61	9.84
mean $f_u(n)$	0.0596	0.0320	0.0176	0.00981
max $f_u(n)$	0.587	0.297	0.113	0.0581

Table 3: Error factors in 9600 experiments

We see that the expectations given in (70) and (71) are fully confirmed by numerical experiments. Even more, $f_u(n)$ appears to be decreasing with n. Thus, we may conclude that it is indeed the scaled matrices, and not the starting matrix H which governs the accuracy of the computed eigensolution.

Further, in each experiment we monitored the number of cycles executed before convergence, and the spectral condition of the matrix V. The results are in Table 4.

n	50	100	200	400
mean$(cycles)$	6	7	8	9
max$(cycles)$	8	10	11	12
mean $\kappa(V)$	7.80	13.9	25.4	47.4
max $\kappa(V)$	28.1	39.7	77.3	140.432

Table 4: Number of cycles and $\kappa(V)$ in 9600 experiments

From Table 4 we see that the convergence of Algorithm 1 is faster on pairs (G, J) obtained by the first step of Algorithm 2, than on the pairs generated in §2.5. This is due to the pivoting in the symmetric indefinite factorization (60), since the columns of obtained G have higher degree of orthogonality. Namely, as noted by several researchers (see e.g. [31]), the transition from the pair (GJG^T, I) to the pair (G^TG, J) is essentially one step of Rutishauser's LR algorithm and usually carries some non-negligible diagonalization effect.

It is also possible to modify the algorithm in order to decrease the number of cycles until convergence. Namely, the pair (G, J) can be transformed by appropriate permutation to the pair (G_1, J_1) with $J_1 = \mathrm{diag}(I_l, -I_{n-l})$ and $G_1 = [G_1' \quad G_2'']$ such that the columns of G_1' and G_1'' have decreasing norms. However, this modification only slightly decreases the number of cycles until convergence - the values in the first two rows of Table 4 are decreased by one.

We have also compared Algorithm 2 with the classical QR method as implemented in the LAPACK routine `ssyev.f` [2] and with the classical two-sided Jacobi method [22]. In almost all experiments with large $\kappa(H)$, the QR and the Jacobi method completely missed the tiny eigenvalues. This behavior is expected since the relative errors in the tiny eigenvalues computed

by both methods are bounded by $\varepsilon\kappa(H)$ [14, 20, 34]. In cases where $\kappa^2(\widehat{B}) \approx \kappa(H)$ all three methods performed equally well, as expected. More details on comparison of Algorithm 2 with the QR and Jacobi method can be found in [9, 23].

4 Conclusion

We showed that the accuracy of the hyperbolic singular value decomposition of the pair (G, J), computed by the one-sided J-orthogonal Jacobi method, depends on the spectral condition of the scaled matrix and not on the condition of G. For matrix G which is well scaled from the right, the one-sided J-orthogonal Jacobi method computes the hyperbolic singular values with high relative accuracy, and the left and right singular vectors with high normwise accuracy.

For example, if the spectral condition of the scaled matrix is $\kappa(B) = 10^3$, and we run the computation in single precision accuracy with $\varepsilon = 2^{-23} \approx 10^{-8}$, then the computed hyperbolic singular values will have 4 or 5 accurate digits. Also, if the hyperbolic singular values are well separated, that is, if there are no clusters of relatively close singular values, then the norm error in the computed left and right singular vectors will be around 10^{-5}.

We also showed that the accuracy of the eigenvalue decomposition of a symmetric indefinite matrix H, computed by the symmetric indefinite factorization $H = GJG^T$ followed by the one-sided J-orthogonal Jacobi method, depends on the spectral condition of the scaled spectral absolute value matrix $|H|$, and not on the condition of H. If $|H|$ is well scaled, or, even simpler, if the matrix G is well scaled from the left and from the right, this algorithm computes the eigenvalues with high relative accuracy, and the eigenvectors with high normwise accuracy.

For example, if the spectral condition of the scaled matrix is $\kappa(\widehat{A}) = 10^3$, or, equivalently, if $\kappa^2(\widehat{B}) = 10^3$, where $\widehat{B}$ is the matrix G scaled from the left, and we run the computation in single precision accuracy, then the computed eigenvalues will have 4 or 5 accurate digits. Also, if the eigenvalues are well separated, that is, if there are no clusters of relatively close eigenvalues, then the norm error in the computed eigenvectors will be around 10^{-5}.

Numerical experiments showed that the constants in the error bounds are indeed moderately growing functions of the dimension. Also, the two-step method computes the eigenvalue decomposition with uniformly higher accuracy than the classical methods.

I would like to thank Krešimir Veselić, Fernuniversität Hagen, James Demmel, UC Berkeley, Jesse Barlow, The Pennsylvania State University, Eberhard Pietzsch, Universität Heidelberg, Zlatko Drmač, University of Zagreb, and Xiaofeng Wang for fruitful discussions and their helpful comments. I also thank the referees for helpful comments which led to more detailed error analysis, improved error bounds and improved presentation of the results.

References

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

[2] E. Anderson et al, *LAPACK Users' Guide*, SIAM, Philadelphia, 1995.

[3] J. Barlow and J. Demmel, Computing accurate eigensystems of scaled diagonally dominant matrices, *SIAM J. Numer. Anal.*, 27:762–791, 1990.

[4] J. R. Bunch and B. N. Parlett, Direct methods for solving symmetric indefinite systems of linear equations, *SIAM J. Numer. Anal.*, 8:639–655, 1971.

[5] C. Davis and W. M. Kahan, The rotation of eigenvectors by a perturbation. III, *SIAM J. Numer. Anal.*, 7:1–46 1970.

[6] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić and Z. Drmač, Computing the singular value decomposition with high relative accuracy, *Linear Algebra Appl.*, 299:21–80, 1999.

[7] J. Demmel and W. Gragg, On computing accurate singular values and eigenvalues of acyclic matrices, *Linear Algebra Appl.*, 185:203–218, 1993.

[8] J. Demmel and W. Kahan, Accurate singular values of bidiagonal matrices, *SIAM J. Sci. Stat. Comput.*, 11:873–912, 1990.

[9] J. Demmel and K. Veselić, Jacobi's method is more accurate than QR, *SIAM J. Matrix Anal. Appl.*, 13:1204–1243, 1992.

[10] Z. Drmač, *Computing the Singular and the Generalized Singular Values*, Ph. D. thesis, Fernuniversität Hagen, Germany, 1994.

[11] Z. Drmač, Accurate computation of the product induced singular value decomposition with applications, *SIAM J. Numer. Anal.*, 35:1969–1994, 1998.

[12] Z. Drmač, A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm, *IMA J. Numer. Anal.*, 19:191–213, 1999.

[13] Z. Drmač and V. Hari, On quadratic convergence bounds for the J-symmetric Jacobi method, *Numer. Math.*, 64:147–180, 1993.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 1996.

[15] M. Gu and S. C. Eisenstat, A divide-and-conquer algorithm for the bidiagonal SVD, *SIAM J. Matrix Anal. Appl.*, 16:79–92, 1995.

[16] M. Gu and S. C. Eisenstat, A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem, *SIAM J. Matrix Anal. Appl.*, 16:172–191, 1995.

[17] R.-C. Li, Relative perturbation theory: (ii) eigenspace and singular subspace variations, *SIAM J. Matrix Anal. Appl.*, 20:471–492, 1998.

[18] R. Mathias, Accurate eigensystem computations by Jacobi method, *SIAM J. Matrix Anal. Appl.*, 16:977–1003, 1996.

[19] R. Onn, A. O. Steinhardt and A. Bojanczyk, Hyperbolic singular value decompositions and applications, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1575–1588, 1991.

[20] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Engelwood Cliffs, NJ, 1980.

[21] R. A. Rosanoff, J. F. Gloudeman and S. Levy, Numerical conditions of stiffness matrix formulations for frame structures, In *Proc. of the Second Conference on Matrix Methods in Structural Mechanics*, WPAFB, Dayton, Ohio, 1968.

[22] H. Rutishauser, The Jacobi method for real symmetric matrices, *Numer. Math.*, 9:1–10, 1966.

[23] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph. D. thesis, Fernuniversität Hagen, Germany, 1992.

[24] I. Slapničar, Componentwise analysis of direct factorization of real symmetric and Hermitian matrices, *Linear Algebra Appl.*, 272:227–275, 1998.

[25] I. Slapničar and N. Truhar, Relative perturbation theory for hyperbolic eigenvalue problem, *Linear Algebra Appl.*, 309:57–72, 2000.

[26] I. Slapničar and N. Truhar, Relative perturbation theory for hyperbolic singular value problem, submitted to *Linear Algebra Appl.*

[27] I. Slapničar and K. Veselić, A bound for the condition of a hyperbolic eigenvector matrix, *Linear Algebra Appl.*, 290:247–255, 1999.

[28] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic press, Boston, 1990.

[29] N. Truhar and I. Slapničar, Relative perturbation bounds for invariant subspaces of graded indefinite Hermitian matrices, *Linear Algebra Appl.*, 301:171–185, 1999.

[30] A. van der Sluis, Condition numbers and equilibration of matrices, *Numer. Math.*, 14:14–23, 1969.

[31] K. Veselić, A Jacobi eigenreduction algorithm for definite matrix pairs, *Numer. Math.*, 64:241–269, 1993.

[32] K. Veselić and I. Slapničar, Floating-point perturbations of Hermitian matrices, *Linear Algebra Appl.*, 195:81–116, 1993.

[33] H. Zha, A note on the existence of the hyperbolic singular value decomposition, *Linear Algebra Appl.*, 240:199–205, 1996.

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