Electronic Transactions on Numerical Analysis.
Volume 65, pp. 26–62, 2026.
© 2026, the Author(s), licensed via CC BY 4.0.
ISSN 1068–9613.
DOI: 10.1553/etna_vol65s26

ETNA
Kent State University and
Johann Radon Institute (RICAM)

# ON CONVERGENCE AND ACCURACY OF THE $J$-JACOBI METHOD UNDER THE DE RIJK PIVOT STRATEGY*

VJERAN HARI† AND VEDRAN NOVAKOVIĆ‡

**Abstract.** This paper proves global convergence of the elementwise $J$-Jacobi method for $J$-Hermitian matrices under the de Rijk pivot strategy and briefly considers the asymptotic convergence of the method. Also considered is the accuracy of a new code for hyperbolic rotations of order two that employs only correctly rounded operations. The numerical tests demonstrate the advantage in the convergence speed of the $J$-Jacobi method under the de Rijk pivot strategy over the same method under the row-cyclic strategy for both the two-sided and the one-sided (implicit) variant of the method.

**Key words.** eigenvalue problem, $J$-Jacobi method, de Rijk pivot strategy, global convergence, high relative accuracy, hyperbolic singular value decomposition

**AMS subject classifications.** 65F15

## 1. Introduction.

This paper considers the elementwise $J$-Jacobi method of Veselić [49] for solving the eigenvalue problem of $J$-Hermitian matrices. It solves one theoretical problem related to global convergence, and it shows how to increase the accuracy of the method.

The $J$-Jacobi method solves the generalized eigenvalue problem $Ax = \lambda Jx$, where $A$ is a complex Hermitian matrix of order $n$, $J$ is a real diagonal matrix of signs, and where for some real number $\mu$, the Hermitian matrix $A - \mu J$ is positive (or negative) definite. In short, the method solves the definite $J$-Hermitian eigenvalue problem.

The importance of the method stems from the fact that the definite generalized eigenvalue problem $Ax = \lambda Bx$, with indefinite Hermitian matrices $A$ and/or $B$, can be reduced to the $J$-Hermitian eigenvalue problem via the Hermitian indefinite factorization of Bunch and Parlett [7]. Many real-world problems lead to the generalized eigenvalue problem, one example being that in [42]. Even some important quadratic eigenvalue problems, such as those for overdamped pencils [38], can be reduced to the $J$-Hermitian eigenvalue problem [49]. Perhaps the most important application of the $J$-Jacobi method lies in the area of the accurate computation of the eigenvalues and eigenvectors of indefinite Hermitian matrices. A brief description of this relatively novel approach can be found in the introductions of the papers [21] and [22]. However, pioneering work and a deep analysis of the underlying matrix theory, which includes the development and research of the elementwise and block $J$-Jacobi methods, has been carried out by Veselić, Slapničar, Truhar, and others (see [5, 21, 22, 28, 43, 44, 45, 46, 47, 48, 50]).

This paper aims to resolve some open problems related to the elementwise $J$-Jacobi method. This is especially important in the context of block $J$-Jacobi methods, which are well suited for contemporary parallel computers [30, 41]. We view the elementwise $J$-Jacobi method as an excellent candidate for the core (also called pointwise or unblocked) algorithm of the block $J$-Jacobi method [21, 22]. This is due to its accuracy as well as its speed on nearly diagonal $J$-Hermitian matrices. Note that the core algorithm operates on matrices of small order.

[*] Received July 11, 2025. Accepted December 11, 2025. Published online on February 2, 2026. Recommended by Ken Hayami. This work has been fully supported by Croatian Science Foundation under the project IP-2020-02-2240.

[†] Corresponding author. University of Zagreb, Faculty of Science, Department of Mathematics, Bijenička cesta 30, HR-10000 Zagreb, Croatia, hari@math.hr, ORCID: 0000-0001-7283-0458.

[‡] Independent researcher (unaffiliated), Vankina ulica 15, HR-10020 Zagreb, Croatia; (venovako@venovako.eu), ORCID: 0000-0003-2964-9674.

One open problem related to the elementwise $J$-Jacobi method is the global convergence under the de Rijk pivot strategy [9]. There are several incentives for considering this pivot strategy. The most important one is that it nearly minimizes the number of cycles (also called sweeps) needed to complete the diagonalization over the set of cyclic strategies. This property might be the reason why it is used in the implementation of the Jacobi singular value decomposition (SVD) method in LAPACK [2]. Although the de Rijk pivot strategy is not cyclic, it is tightly connected to the row-cyclic strategy, and in the later stages of the process it reduces to the row-cyclic strategy [19]. The de Rijk strategy has been extensively studied in [19], where the global and quadratic convergence of the standard elementwise and the block Jacobi method have been proved. Here we directly use some bounds from [19].

The second open problem is related to enhancing the accuracy of the $J$-Jacobi method. This is important because the high relative accuracy is a trademark of the method. Recent advances in implementing elementary functions with the correct rounding of the result, e.g., in the context of the CORE-MATH project [40], enable us to compute, in common data types, trigonometric [32] and hyperbolic $2 \times 2$ complex rotations more accurately than with the established formulas.

The two-sided $J$-Jacobi method for the definite $J$-Hermitian eigenproblem naturally induces its implicit, one-sided counterpart for computing the hyperbolic singular value decomposition [6], or the HSVD for short. The one-sided $J$-Jacobi method for computing the HSVD of a full column-rank matrix $G$ implicitly executes the two-sided method on the pair $(G^*G, J)$ by transforming only the columns of $G$, unlike the slower and less stable two-sided $J$-Kogbetliantz HSVD method [34]. Since the HSVD has applications beyond its role in solving Hermitian indefinite eigenproblems, for example in radar tracking [23] in the signal processing [24] domain, we also briefly describe and numerically test the one-sided $J$-Jacobi method for the HSVD.

The paper is organized as follows. In Section 2 we present the algorithm that diagonalizes a $2 \times 2$ Hermitian or $J$-Hermitian matrix. We also present some facts and estimates related to the $J$-Hermitian eigenproblem and to the method. We define a “stable” $J$-Jacobi method that is less prone to instabilities in the floating-point environment. Almost all of these results are taken from the original paper [49]. We also describe the de Rijk pivot strategy. In Section 3 we prove global convergence of the $J$-Jacobi method under the de Rijk pivot strategy. To this end we consider a new pivot strategy that is equivalent to the de Rijk strategy. We first prove all the auxiliary results for this new pivot strategy and then complete the global convergence proof under the de Rijk strategy. We also prove global convergence of the stable method under the de Rijk strategy. At the end of this section we briefly discuss the ultimate quadratic convergence of the method. In Section 4 we switch our attention to floating-point arithmetic and to the accuracy of the elementwise $J$-Jacobi method by presenting new formulas for the angle-restricted hyperbolic rotations, with relative error bounds for the hyperbolic tangents, cosines, and sines. A noticeable improvement in the relative errors over those obtained with established formulas is illustrated by an exhaustive test. We describe the numerical tests for the two-sided and the one-sided $J$-Jacobi method in Section 5 and conclude the paper with a discussion of future work in Section 6.

## 2. The $J$-Jacobi method under the de Rijk pivot strategy.

In this section, we briefly describe the algorithm of the method and then the de Rijk pivot strategy. We begin with some basic facts related to the $J$-Hermitian eigenvalue problem $Ax = \lambda Jx$.

We assume that $A - \mu J$ is positive definite, where $A$ is a Hermitian matrix of order $n$, $J = \mathrm{diag}(I_m, -I_{n-m})$, $1 \leq m < n$, and $\mu$ is a real number. The number $\mu$ is called a *definitizing shift*. If a nonsingular matrix $C$ satisfies $C^*JC = J$, then it is called $J$-*unitary*. $J$-unitary matrices form a multiplicative group. Since $A - \mu J$ is positive definite, there exists a $J$-unitary matrix $C$ such that $C^*AC = \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m, \lambda_{m+1}, \dots, \lambda_n)$ is real and (see [49])

$$\delta_0 = \lambda_m + \lambda_{m+1} > 0. \tag{2.1}$$

Here, $\lambda_1, \dots, \lambda_m, -\lambda_{m+1}, \dots, -\lambda_n$ are the eigenvalues of the generalized eigenvalue problem $Ax = \lambda Jx$ arranged in non-increasing order. They are also the eigenvalues of the matrix $JA$, and we say that they are the eigenvalues of the pair $(A, J)$. In [49] it has been shown that the set of all definitizing shifts is the open interval $(-\lambda_{m+1}, \lambda_m)$ and that (see [49, Theorem 2.1])

$$a_{rr} + a_{ss} > \delta_0, \qquad 1 \le r \le m < s \le n, \tag{2.2}$$

$$\left| \frac{2a_{rs}}{a_{rr} + a_{ss}} \right| \le \left[ 1 - \frac{\delta_0^2}{(a_{rr} + a_{ss})^2} \right]^{1/2}, \qquad 1 \le r \le m < s \le n. \tag{2.3}$$

Originally, in [49], the relations (2.1)–(2.3) were proved for a symmetric matrix $A$, but it can be easily verified that they also hold when $A$ is complex Hermitian. These relations are essential for analyzing the algorithm and for the convergence proof.

If $A - \mu J$ is negative definite, then $(-A) - (-\mu)J$ is positive definite and $-A$ is Hermitian. This shows that we only need to consider the positive definite $J$-Hermitian eigenvalue problem.

### 2.1. The $J$-Jacobi algorithm.

Here we recall basic formulas and relations linked to the elementwise $J$-Jacobi method, which we briefly call the $J$-Jacobi method.

The $J$-Jacobi method gradually diagonalizes the matrix $A$ using $J$-unitary plane transformations. The iterative process has the form

$$A^{(k+1)} = [U^{(k)}]^* A^{(k)} U^{(k)}, \qquad k \ge 1, \tag{2.4}$$

where $A^{(1)} = A$ and each $U^{(k)}$ is a $J$-unitary plane matrix. This notation means that $[U^{(k)}]^* J U^{(k)} = J$, $k \ge 1$, and that each $U^{(k)}$ differs from the identity matrix in one principal submatrix of order 2. Such a principal submatrix is called the pivot submatrix of $U^{(k)}$, and it is usually denoted by $\hat{U}^{(k)}$. The role of $\hat{U}^{(k)}$ is to diagonalize $\hat{A}^{(k)}$, the pivot submatrix of $A^{(k)}$, via the congruence transformation

$$[\hat{U}^{(k)}]^* \hat{A}^{(k)} \hat{U}^{(k)} = [\hat{U}^{(k)}]^* \begin{bmatrix} a_{ii}^{(k)} & a_{ij}^{(k)} \\ a_{ji}^{(k)} & a_{jj}^{(k)} \end{bmatrix} \hat{U}^{(k)} = \begin{bmatrix} a_{ii}^{(k+1)} & \\ & a_{jj}^{(k+1)} \end{bmatrix}, \qquad k \ge 1. \tag{2.5}$$

The indices $i$ and $j$ are *pivot indices*, and $(i, j)$, $i < j$, is a *pivot pair*. The pivot pair determines which off-diagonal elements of $A^{(k)}$ are nullified. Obviously, the pivot indices are functions of $k$, $i = i(k)$, $j = j(k)$, where $k$ counts the steps of the iterative process (2.4). The element $a_{ij}^{(k)}$ is the *pivot element* in step $k$.

As shown in [49], [21, Sect. 2.3], and [5, Sect. 7.1], the matrix $\hat{U}^{(k)}$ has the form

$$\hat{U}^{(k)} = \hat{\Phi}^{(k)}(\phi_k) \hat{R}^{(k)}(\theta_k), \tag{2.6}$$

where

$$\hat{\Phi}^{(k)}(\phi_k) = \operatorname{diag}(e^{\imath\phi_k}, 1) \qquad \text{or} \qquad \hat{\Phi}^{(k)}(\phi_k) = \operatorname{diag}(1, e^{-\imath\phi_k}), \tag{2.7}$$

and

$$\hat{R}^{(k)}(\theta_k) = \begin{cases} \begin{bmatrix} \cosh(\theta_k) & \sinh(\theta_k) \\ \sinh(\theta_k) & \cosh(\theta_k) \end{bmatrix} & \text{if } 1 \le i \le m < j \le n, \\[2ex] \begin{bmatrix} \cos(\theta_k) & -\sin(\theta_k) \\ \sin(\theta_k) & \cos(\theta_k) \end{bmatrix} & \text{elsewhere.} \end{cases} \tag{2.8}$$

The role of the phase $\phi_k$ is to make the pivot submatrix $\hat{A}^{(k)}$ real and symmetric,

$$[\hat{\Phi}^{(k)}(\phi_k)]^* \hat{A}^{(k)} \hat{\Phi}^{(k)}(\phi_k) = \begin{bmatrix} a_{ii}^{(k)} & |a_{ij}^{(k)}| \\ |a_{ji}^{(k)}| & a_{jj} \end{bmatrix}, \qquad a_{ji}^{(k)} = [a_{ij}^{(k)}]^*,$$

which implies $\phi_k = \arg(a_{ij}^{(k)})$. Alternatively, one can write (see Section 4)

$$\hat{\Phi}^{(k)}(\phi_k) = \operatorname{diag}(e^{-\imath\phi_k}, 1) \quad \text{or} \quad \hat{\Phi}^{(k)}(\phi_k) = \operatorname{diag}(1, e^{\imath\phi_k}), \qquad \phi_k = \arg(a_{ji}^{(k)}).$$

The role of the angle $\theta_k$ is to nullify $|a_{ij}^{(k)}|$. This implies

$$\tanh(2\theta_k) = \frac{-2|a_{ij}^{(k)}|}{a_{ii}^{(k)} + a_{jj}^{(k)}}, \qquad 1 \le i \le m < j \le n, \tag{2.9}$$

$$\tan(2\theta_k) = \frac{2|a_{ij}^{(k)}|}{a_{ii}^{(k)} - a_{jj}^{(k)}}, \qquad 1 \le i < j \le m \ \text{ or } \ m+1 \le i < j \le n. \tag{2.10}$$

The relations (2.1)–(2.3) imply that $|\tanh(2\theta_k)| < 1$.

Let us consider the case $1 \le i \le m < j \le n$. We have (see [28, 49])

$$a_{ii}^{(k+1)} = a_{ii}^{(k)} + \tanh(\theta_k)\,|a_{ij}^{(k)}|, \tag{2.11}$$

$$a_{jj}^{(k+1)} = a_{jj}^{(k)} + \tanh(\theta_k)\,|a_{ij}^{(k)}|, \tag{2.12}$$

and consequently,

$$a_{ii}^{(k+1)} + a_{jj}^{(k+1)} = (a_{ii}^{(k)} + a_{jj}^{(k)})\kappa_k,$$

where

$$\kappa_k = 1 - \tanh(\theta_k)\tanh(2\theta_k) = \sqrt{1 - \tanh^2(2\theta_k)} > 0\,.$$

Here, we used (2.9). Note that

$$\text{trace}(A^{(k)}) - \text{trace}(A^{(k+1)}) = a_{ii}^{(k)} + a_{jj}^{(k)} - (a_{ii}^{(k+1)} + a_{jj}^{(k+1)}) = (a_{ii}^{(k)} + a_{jj}^{(k)})(1 - \kappa_k).$$

Together with (2.2), this implies

$$\begin{aligned} \tanh(\theta_k)\tanh(2\theta_k) &= \frac{\text{trace}(A^{(k)}) - \text{trace}(A^{(k+1)})}{a_{ii}^{(k)} + a_{jj}^{(k)}} \\ &\le \frac{\text{trace}(A^{(k)}) - \text{trace}(A^{(k+1)})}{\delta_0}. \end{aligned} \tag{2.13}$$

Thus, during the “hyperbolic steps”, the trace of $A^{(k)}$ is not increased, and it is decreased if and only if $a_{ij}^{(k)} \ne 0$.

Now let us consider the “trigonometric steps”, which occur when the pivot element lies within the diagonal blocks of order $m$ and $n-m$ of $A^{(k)}$. In this case, the transformation is the same as for the standard Jacobi method for Hermitian matrices. Then we have $-\pi/4 \le \theta_k \le \pi/4$, and

$$a_{ii}^{(k+1)} = a_{ii}^{(k)} + \tan(\theta_k)\,|a_{ij}^{(k)}|, \tag{2.14}$$

$$a_{jj}^{(k+1)} = a_{jj}^{(k)} - \tan(\theta_k)\,|a_{ij}^{(k)}|, \tag{2.15}$$

and consequently,

$$a_{ii}^{(k+1)} + a_{jj}^{(k+1)} = a_{ii}^{(k)} + a_{jj}^{(k)} \qquad \text{and} \qquad \text{trace}(A^{(k+1)}) = \text{trace}(A^{(k)}).$$

Thus, during the entire process, trace$(A^{(k)})$ is a non-increasing function of $k$.

In the convergence analysis we use a measure that is usually called the *departure from the diagonal form*, or off-norm. We have

$$\text{off}(X) = \|X - \text{diag}(X)\|_F,$$

where $X$ is a square matrix and $\|\cdot\|_F$ is the Frobenius norm. The spectral and infinity norms are denoted by $\|\cdot\|_2$ and $\|\cdot\|_\infty$, respectively.

Let $\mu$ be a real number such that $A - \mu J$ is positive definite. Since congruence transformations of a Hermitian matrix do not change the number of positive eigenvalues, each matrix $A^{(k)} - \mu J$ is positive definite, and we have

$$\begin{aligned} 0 < \|A^{(k)} - \mu J\|_F &= \left[\lambda_1^2(A^{(k)} - \mu J) + \cdots + \lambda_n^2(A^{(k)} - \mu J)\right]^{1/2} \\ &\leq \lambda_1(A^{(k)} - \mu J) + \cdots + \lambda_n(A^{(k)} - \mu J) = \text{trace}(A^{(k)} - \mu J) \\ &= \text{trace}(A^{(k)}) - \mu\,\text{trace}(J) = \text{trace}(A^{(k)}) - (2m-n)\mu, \qquad k \geq 1. \end{aligned} \tag{2.16}$$

Here, $\lambda_1(A^{(k)} - \mu J), \ldots, \lambda_n(A^{(k)} - \mu J)$ denote the eigenvalues of the matrix $A^{(k)} - \mu J$. In addition, we have

$$0 \leq \text{off}(A^{(k)}) \leq \|A^{(k)} - \mu J\|_F \leq \text{trace}(A^{(k)}) - (2m-n)\mu, \qquad k \geq 1. \tag{2.17}$$

From (2.17) we see that trace$(A^{(k)})$ is bounded below by $(2m-n)\mu$. Since trace$(A^{(k)})$ is non-increasing, it is convergent. Hence, relation (2.13) implies

$$\tanh(\theta_k)\tanh(2\theta_k) \to 0, \qquad \text{as } k \to \infty, \quad \text{over the set of hyperbolic steps.} \tag{2.18}$$

Since

$$\tanh(\theta_k)\tanh(2\theta_k) = \frac{2\tanh^2(\theta_k)}{1 + \tanh^2(\theta_k)} \geq \tanh^2(\theta_k),$$

we obtain, using (2.2), (2.9), and (2.18),

$$\frac{2|a_{ij}^{(k)}|}{\delta_0} \leq \frac{2|a_{ij}^{(k)}|}{a_{ii}^{(k)} + a_{jj}^{(k)}} = -\tanh(2\theta_k) \to 0, \qquad \text{as } k \to \infty, \tag{2.19}$$

over the set of hyperbolic steps. In this way we have shown that

$$a_{i(k)j(k)}^{(k)} \to 0, \qquad \text{as } k \to \infty, \quad \text{over the set of hyperbolic steps.} \tag{2.20}$$

If the eigenvectors of the pair $(A, J)$ are required, then the product $U^{(1)}U^{(2)} \cdots U^{(k)}$ has to be computed, which may not converge to a $J$-unitary matrix. However, if $A^{(k)}$ converges to a diagonal matrix, then for sufficiently large $k$, the columns of this product will be good approximations of the eigenvectors of the pair $(A, J)$.

If $A$ is real, then the formulas above are simplified. The matrix $\hat{\Phi}^{(k)}$ is no longer present. The angle $\theta_k$ is defined by (2.9) or (2.10), with $|a_{ij}^{(k)}|$ replaced by $a_{ij}^{(k)}$. The same replacement has to be made in the formulas for updating the diagonal elements. All other relations remain valid.

As noted in [49], the $J$-Jacobi method never breaks down in exact arithmetic, although large hyperbolic angles $\theta_k$ can appear before the pivot elements become sufficiently small. In such a case, the pivot submatrix $\hat{U}^{(k)}$, and consequently the plane transformation matrix $U^{(k)}$, have a large condition number, which increases the condition number of the product $U^{(1)}U^{(2)}\cdots U^{(k)}$. As shown in [49], the condition number of $U^{(1)}U^{(2)}\cdots U^{(k)}$ will ultimately be determined only by the initial pair $(A, J)$. However, a temporary increase of this quantity can lead to instability and loss of accuracy in the output data. A remedy for this potential instability consists of bounding the angle $\theta_k$ so that the absolute value of each hyperbolic tangent $\tanh(\theta_k)$ is bounded by some suitable number $t_{\max}$ smaller than $1$. We call the method that is "stabilized" in this way *the stable $J$-Jacobi method*. In [49] the choice $t_{\max} = 0.5$ has been suggested. For our implementation, we have taken $t_{\max} = 0.8$ and provide an explanation for this choice in Section 5.

We end this section with a remark. The pivot submatrix of $U^{(k)}$ can have the form $\hat{\Phi}^{(k)}(\phi_k)\hat{R}^{(k)}(\theta_k)[\hat{\Phi}^{(k)}(\phi_k)]^*$, where $\hat{R}^{(k)}$ and $\hat{\Phi}^{(k)}$ are given by relations (2.8) and (2.7), respectively. If this new form of $\hat{U}^{(k)}$ is denoted by $\hat{V}^{(k)}$, then $\hat{V}^{(k)} = \hat{U}^{(k)}[\hat{\Phi}^{(k)}(\phi_k)]^*$, i.e.,

$$\hat{V}^{(k)}(\theta_k) = \begin{cases} \begin{bmatrix} \cosh(\theta_k) & e^{-\imath\phi_k}\sinh(\theta_k) \\ e^{\imath\phi_k}\sinh(\theta_k) & \cosh(\theta_k) \end{bmatrix} & \text{if } 1 \le i \le m < j \le n, \\ \begin{bmatrix} \cos(\theta_k) & -e^{-\imath\phi_k}\sin(\theta_k) \\ e^{\imath\phi_k}\sin(\theta_k) & \cos(\theta_k) \end{bmatrix} & \text{elsewhere.} \end{cases} \tag{2.21}$$

Using relation (2.5) we have

$$\begin{aligned} \begin{bmatrix} a_{ii}^{(k+1)} & \\ & a_{jj}^{(k+1)} \end{bmatrix} &= \hat{\Phi}^{(k)}(\phi_k) \begin{bmatrix} a_{ii}^{(k+1)} & \\ & a_{jj}^{(k+1)} \end{bmatrix} [\hat{\Phi}^{(k)}(\phi_k)]^* \\ &= \hat{\Phi}^{(k)}(\phi_k)[\hat{U}^{(k)}]^*\hat{A}^{(k)}\hat{U}^{(k)}[\hat{\Phi}^{(k)}(\phi_k)]^* \\ &= [\hat{V}^{(k)}]^*\hat{A}^{(k)}\hat{V}^{(k)}, \qquad k \ge 1. \end{aligned}$$

This shows that $\hat{V}^{(k)}$ also diagonalizes the pivot submatrix $\hat{A}^{(k)}$ and serves the same purpose as $\hat{U}^{(k)}$. Note that $\hat{V}^{(k)}(\theta_k)$ can be equivalently represented by the tangents instead of the sines (i.e., the sines of $\theta_k$ are not strictly required) as

$$\check{V}^{(k)}(\theta_k) = \begin{cases} \begin{bmatrix} 1 & e^{-\imath\phi_k}\tanh(\theta_k) \\ e^{\imath\phi_k}\tanh(\theta_k) & 1 \end{bmatrix} \cdot \cosh(\theta_k) & \text{if } 1 \le i \le m < j \le n, \\ \begin{bmatrix} 1 & -e^{-\imath\phi_k}\tan(\theta_k) \\ e^{\imath\phi_k}\tan(\theta_k) & 1 \end{bmatrix} \cdot \cos(\theta_k) & \text{elsewhere.} \end{cases} \tag{2.22}$$

The form (2.22) simplifies transformations of row and column pairs. In the real case, when $\phi_k = 0$, each new element is formed by a single $\mathrm{fma}$ operation (fused multiply-add operation), followed by a multiplication by the cosine (if the cosine is not the unity). Otherwise, a complex analogue of the $\mathrm{fma}$ operation $x \cdot y + z$ can be employed, but at present a correct rounding of the components of the result is difficult to achieve. We use[1]

$$\begin{aligned}\mathsf{fma}(x,y,z) &= \mathrm{fma}(\Re(x),\Re(y),\mathrm{fma}(-\Im(x),\Im(y),\Re(z)))\\ &\quad + \imath\,\mathrm{fma}(\Re(x),\Im(y),\mathrm{fma}(\Im(x),\Re(y),\Im(z))),\end{aligned}$$

followed by a multiplication of the real and imaginary components of the result by the cosine. In this way, the transformations are not only faster than those implied by (2.21) but also independent of the compiler's complex arithmetic and thus numerically reproducible (see also [31]). Here, $\Re(w)$ and $\Im(w)$ denote the real and imaginary parts of $w$.

**2.2. The de Rijk pivot strategy.** Here, we describe the de Rijk pivot strategy. To this end we use almost the same notation as in [19].

The selection of pivot pairs is defined by a *pivot strategy*. We can identify it with a function $\mathsf{I} : \mathcal{N} \to \mathcal{P}_n$, where

$$\mathcal{N} = \{1,2,\ldots\}, \qquad \mathcal{P}_n = \{(r,t);\ 1 \le r < t \le n\}. \tag{2.23}$$

Here, $\mathcal{P}_n$ contains pairs of indices that address the elements in the upper triangle of the matrix $A$. Hence, it has

$$N = n(n-1)/2$$

pairs. If $\mathsf{I}$ is periodic, then $\mathsf{I}$ is called a *periodic pivot strategy*. Let $P$ be the period of $\mathsf{I}$. If

$$P = N\ (P \ge N) \qquad \text{and} \qquad \{\mathsf{I}(k) : k = 1,\ldots,P\} = \mathcal{P}_n,$$

then $\mathsf{I}$ is called a *cyclic* (*quasi-cyclic*) pivot strategy. Then the first cycle (quasi-cycle) consists of the first $P$ steps of the method. More about pivot strategies can be found in [15, 25, 39], and especially in [20, Section 3].

The most commonly used cyclic pivot strategy is the *row-cyclic* strategy. It is defined by the *row-wise ordering* of the set $\mathcal{P}_n$, i.e., by the sequence of pairs

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

Suppose $A^{(N+1)}$ is obtained at the end of the first cycle of the $J$-Jacobi method under the row-cyclic strategy. Then we have

$$\begin{aligned}A^{(N+1)} &= [U^{(1)}U^{(2)}\cdots U^{(N)}]^* A^{(1)} U^{(1)}U^{(2)}\cdots U^{(N)}\\ &= [U^{(1:n-1)}U^{(n:2n-3)}\cdots U^{(N)}]^* A^{(1)} U^{(1:n-1)}U^{(n:2n-3)}\cdots U^{(N)}\\ &= [U_{1,2:n}U_{2,3:n}\cdots U_{n-1,n}]^* A^{(1)} U_{1,2:n}U_{2,3:n}\cdots U_{n-1,n},\end{aligned}$$

where

$$U^{(r_1:r_2)} = U^{(r_1)}\cdots U^{(r_2)}, \qquad 1 \le r_1 \le r_2 \le n, \tag{2.24}$$
$$U_{r,r+1:n} = U_{r,r+1}\cdots U_{rn}, \qquad 1 \le r \le n-1. \tag{2.25}$$

In the special case when $r_2 = r_1$, we have $U^{(r_1:r_1)} = U^{(r_1)}$. Continuing the process, in cycle $t$, one simply adds $(t-1)N$ to the superscripts of the terms in the above relations. We will

[1] As in a widely used implementation in the `cuComplex.h` header from the NVIDIA's CUDA Toolkit, https://developer.nvidia.com/cuda-toolkit.

also use the notation $U^{[t]}_{r,r+1:n}$, $1 \leq r \leq n-1$, where $U^{[t]}_{r,r+1:n} = U^{[t]}_{r,r+1} \cdots U^{[t]}_{rn}$ and the matrices $U^{[t]}_{rs}$ belong to cycle $t$ of the method.

The cycle $t$ is comprised of the steps $k \in \mathcal{C}_t$, where

$$\mathcal{C}_t = \{(t-1)N+1, (t-1)N+2, \ldots, tN\}.$$

The de Rijk pivot strategy [9] is a modified row-cyclic strategy. It uses, additionally, (at most) $n-1$ transposition matrices within each cycle. We describe it by considering one full cycle of the row-cyclic process (2.4), say the first cycle. We use the same notation as for the row-cyclic strategy.

REMARK 2.1. Obviously, we misuse the terms "cycle" and "cyclic" because the de Rijk strategy is not cyclic. Since the transposition matrices are plane matrices, each "cycle" of the method actually uses at most $N+n-1$ plane transformations. One may say that it is more like a quasi-cyclic method with period not larger than $N+n-1$. However, in every later "quasi-cycle", the transposition matrices may be different from those in the current one. Namely, for any given $r$, $1 \leq r \leq n-1$, the value of $r'$ in the transposition matrix $I_{rr'}$ depends on the "quasi-cycle". If $r' = r$, then $I_{rr'}$ is omitted. Thus, the current quasi-cycle can have fewer than $N+n-1$ steps. Here, $I_{rr'} = [e_1, \ldots, e_{r'}, \ldots, e_r, \ldots, e_n]$, $r < r'$, where $I_n = [e_1, \ldots, e_n]$ is the identity matrix.

Let us consider the first cycle of the method. To this end, we simplify the notation by using only subscripts or superscripts. The first cycle can be described as follows:

$$A^{(N+1)} = A^{[1]} = [U^{[1]}]^* A^{(1)} U^{[1]}, \tag{2.26}$$

where

$$U^{[1]} = P_1 U^{(1:n-1)} P_2 U^{(n:2n-3)} \cdots P_{n-2} U^{(n-2:n-1)} P_{n-1} U^{(N)}. \tag{2.27}$$

In (2.27), we have used the notation from (2.24). Using (2.25), we obtain

$$U^{[1]} = P_1 U_{1,2:n} P_2 U_{2,3:n} \cdots \; P_{n-2} U_{n-2,n-1:n} P_{n-1} U_{n-1,n}. \tag{2.28}$$

Here, $P_i$, $1 \leq i \leq n-1$, are the transposition matrices that swap the rows and columns of the current matrix. They will be defined later. In each subsequent cycle, these permutation matrices can differ from those in the previous cycle.

Let

$$s_\ell = 1 + 2 + \cdots + \ell = \frac{\ell(\ell+1)}{2}, \qquad 1 \leq \ell \leq n-1, \qquad s_0 = 0. \tag{2.29}$$

Since $s_{n-1} = 1 + 2 + \cdots + n - 1 = N$, from (2.29) we obtain $N - s_{n-1} = 0$ and

$$N - s_\ell = N - (1 + 2 + \cdots + \ell) = (n-1) + (n-2) + \cdots + (\ell+1), \qquad 1 \leq \ell < n-1.$$

Consequently, we have

$$N - s_{n-r} + 1 = \begin{cases} 1 & r = 1, \\ 1 + (n-1) + (n-2) + \cdots + (n-(r-1)) & 2 \leq r \leq n-1. \end{cases}$$

In step $k = N - s_{n-r} + 1$, the transformation $A^{(k+1)} = [P_r U^{(k)}]^* A^{(k)} [P_r U^{(k)}]$ takes place. Obviously, if $P_i = I_n$ for all $1 \leq i \leq n-1$, then in the considered cycle, the de Rijk strategy coincides with the row-cyclic one. Considering $P_r U^{(k)}$ to be a single transformation matrix has an advantage since then one "cycle" consists of exactly $N$ transformations. In the convergence analysis we will also consider $P_rU^{(k)}$ corresponding to two transformations, the first using $P_r$ and the second using $U^{(k)}$.

Let us now define the permutation matrices. We have

$$P_r = I_{rr'}, \qquad r' \geq r, \qquad 1 \leq r \leq n-1, \tag{2.30}$$

where $I_{rr'}$ is the transposition matrix. To define $r'$, we have to take into account that we want the matrix $J$ to be invariant under the congruence transformation with $I_{rr'}$. *This means that in the case* $1 \leq r \leq m-1$ *(*$m+1 \leq r \leq n-1$*), we must have* $1 \leq r \leq r' \leq m$ *(*$m+1 \leq r \leq r' \leq n$*).*

The subscript $r'$ is defined as follows:

$$a_{r'r'}^{(N-s_{n-r}+1)} = \begin{cases} \max_{r\leq \ell \leq m} a_{\ell\ell}^{(N-s_{n-r}+1)} & 1 \leq r \leq m-1, \\ a_{mm}^{(N-s_{n-m}+1)} & r = m, \\ \max_{r\leq \ell \leq n} a_{\ell\ell}^{(N-s_{n-r}+1)} & m+1 \leq r \leq n-1. \end{cases} \tag{2.31}$$

One can see that the rows and columns $r$ and $r'$ of $A^{(N-s_{n-r}+1)}$ are swapped just before the annihilation of the elements in the $r$th row and column begins. From (2.31) and (2.30), we see that this swapping makes the $(r,r)$-element larger than or equal to the elements at positions $(\ell,\ell)$, where $1 \leq r < \ell \leq m$ or $m+1 \leq r < \ell \leq n$. If $r' = r$, then we have $P_r = I_{rr} = I_n$, and no swap occurs. We also have $P_m = I_{mm'} = I_n$.

We see that the *de Rijk strategy tries to order the first* $m$ *and the last* $n-m$ *diagonal elements in non-increasing order during the process.*

## 3. Global convergence.

Here, we prove global convergence of the $J$-Jacobi method under the de Rijk strategy. We first consider the complex method. Later, we consider the stable $J$-Jacobi method and finally the real methods.

We first show that the process (2.4) under this strategy can be rearranged to facilitate the convergence analysis. To this end, we need the notion of equivalent pivot strategies.

### 3.1. A pivot strategy equivalent to the de Rijk strategy.

Every cyclic pivot strategy is defined by an ordering of the set $\mathcal{P}_n$ from (2.23). If

$$\mathsf{O} = (i_1,j_1),(i_2,j_2),\ldots,(i_N,j_N)$$

is an ordering of $\mathcal{P}_n$, then the associated cyclic strategy $\mathsf{I} = \mathsf{I}_{\mathsf{O}}$ is defined by $(i(k),j(k)) = (i_r,j_r)$ provided that $k-1 \equiv r-1 \pmod{N}$, $1 \leq r \leq N$.

We say that the pairs $(r,s)$ and $(r',s')$ from $\mathcal{P}_n$ are *disjoint* or *commuting* if the sets $\{r,s\}$ and $\{r',s'\}$ are disjoint. One can easily see that if the pairs $(r,s)$ and $(r',s')$ from $\mathcal{P}_n$ are commuting, then the plane transformations defined by those pairs also commute.

Let $\mathcal{T}$ be any subset of $\mathcal{P}_n$ containing $\tau$, $2 \leq \tau \leq N$, elements. Let $\mathbf{O}(\mathcal{T})$ be the set of all finite sequences of pairs from $\mathcal{T}$ such that each sequence from $\mathbf{O}(\mathcal{T})$ contains all pairs from $\mathcal{T}$. Let $O \in \mathbf{O}(\mathcal{T})$, $O = (i_1,j_1),(i_2,j_2)\ldots(i_t,j_t)$, $t \geq \tau$. An admissible transposition on $O$ is any transposition of two adjacent terms,

$$(i_r,j_r),(i_{r+1},j_{r+1}) \longmapsto (i_{r+1},j_{r+1}),(i_r,j_r), \qquad 1 \leq r \leq t-1,$$

provided that $(i_r,j_r)$ and $(i_{r+1},j_{r+1})$ are disjoint.

The following definition is due to Hansen (see [13]).

DEFINITION 3.1. *Two sequences $O, \tilde{O} \in \mathbf{O}(\mathcal{T})$ are equivalent if one can be obtained from the other by a set of admissible transpositions. In this case we write $O \sim \tilde{O}$. If $\mathcal{T} = \mathcal{P}_n$ and $O \sim \tilde{O}$, then the associated quasi-cyclic pivot strategies $\mathsf{I}_O$ and $\mathsf{I}_{\tilde{O}}$ are equivalent, and we write $\mathsf{I}_O \sim \mathsf{I}_{\tilde{O}}$.*

An easy inspection shows that $\sim$ is an equivalence relation on $\mathbf{O}(\mathcal{T})$. Note that each cyclic strategy is also considered to be quasi-cyclic.

LEMMA 3.2. *Let $O = (i_1, j_1), \ldots, (i_t, j_t)$ and $\tilde{O} = (\tilde{i}_1, \tilde{j}_1), \ldots, (\tilde{i}_t, \tilde{j}_t)$ be two sequences from $\mathbf{O}(\mathcal{T})$, $\mathcal{T} \subseteq \mathcal{P}_n$. Let $U_{i_r j_r}$ and $\tilde{U}_{\tilde{i}_r \tilde{j}_r}$, $r = 1, \ldots, t$, be generated by the process (2.4) following the orderings $O$ and $\tilde{O}$, respectively, so that*

$$\begin{aligned} A^{(r+1)} &= (U_{i_1 j_1} \cdots U_{i_r j_r})^* A (U_{i_1 j_1} \cdots U_{i_r j_r}), \\ \tilde{A}^{(r+1)} &= (\tilde{U}_{\tilde{i}_1 \tilde{j}_1} \cdots \tilde{U}_{\tilde{i}_r \tilde{j}_r})^* A (\tilde{U}_{\tilde{i}_1 \tilde{j}_1} \cdots \tilde{U}_{\tilde{i}_r \tilde{j}_r}) \end{aligned}$$

*holds for $r = 1, \ldots, t$. If $O \sim \tilde{O}$, then*

$$U_{pq} = \tilde{U}_{pq}, \qquad \text{for all } (p,q) \in \mathcal{T}, \tag{3.1}$$

*and*

$$A^{(t+1)} = \tilde{A}^{(t+1)}. \tag{3.2}$$

*If $(p, q)$ is repeated in $O$, i.e., if*

$$(p,q) = (i_{1'}, j_{1'}) = \cdots = (i_{k'}, j_{k'}) \quad \text{and} \quad (p,q) = (\tilde{i}_{1''}, \tilde{j}_{1''}) = \cdots = (\tilde{i}_{k''}, \tilde{j}_{k''}),$$

*for some $1' < 2' < \cdots < k'$ and $1'' < 2'' < \cdots < k''$, then relation (3.1) takes the form*

$$U_{i_{1'} j_{1'}} = \tilde{U}_{\tilde{i}_{1''} \tilde{j}_{1''}}, \ldots, U_{i_{k'} j_{k'}} = \tilde{U}_{\tilde{i}_{k''} \tilde{j}_{k''}},$$

*while (3.2) remains the same.*

*Proof.* It suffices to prove Lemma 3.2 under the assumption that $\tilde{O}$ results from $O$ by applying only one admissible transformation. However, the proof for this special case is almost identical to the proof of [13, Theorem 1]. □

From Definition 3.1, we see that two cyclic pivot strategies are equivalent if their defining orderings are such. This implies that for two equivalent cyclic strategies, the matrices obtained after each cycle are the same and that during the same cycle, the plane transformations that nullify the element at the same position are identical.

REMARK 3.3. Note that the above assertions hold in exact arithmetic but not necessarily in floating-point arithmetic. For example, let $(i', j')$ and $(i'', j'')$ be disjoint, and let two equivalent sequences $O$ and $\tilde{O}$ be such that $O = \ldots, (i', j'), (i'', j''), \ldots$ and $\tilde{O} = \ldots, (i'', j''), (i', j'), \ldots$, i.e., the sequences differ by an admissible transposition. Assume that $(i'', j')$—or, similarly, $(i', j'')$—is present in $O$ and $\tilde{O}$ after both $(i', j')$ and $(i'', j'')$. Before $(i'', j')$ (or $(i', j'')$) becomes the pivot pair, the element of the iteration matrix at this position is transformed at least twice following either $O$ or $\tilde{O}$, but in a different order. This may cause different rounding errors to be accumulated in the element in question.

A typical example of two equivalent cyclic strategies is the row- and column-cyclic strategies. The latter is defined by the column-wise ordering of $\mathcal{P}_n$:

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

Let us consider the first "cycle" of the $J$-Jacobi method defined by relations (2.26)–(2.28). From relation (2.27) or (2.28), we can read the sequence of pairs associated with the de Rijk strategy. We denote it by $O_R$,

$$\begin{aligned} O_R = &(1,1'),(1,2:n),(2,2'),(2,3:n),\ldots,(m-1,(m-1)'),(m-1,m:n),\\ &(m,m+1:n),(m+1,(m+1)'),(m+1,m+2:n),\ldots,\\ &(n-1,(n-1)'),(n-1,n). \end{aligned}$$

Here, we use the notation $(r,p:q)=(r,p),(r,p{+}1),\ldots,(r,q)$, $p<q$, and $(r,p:p)=(r,p)$. If a pair $(s,t)$ commutes with all pairs from $(r,p:q)$, then we say that $(s,t)$ commutes with $(r,p:q)$ and vice versa.

Note that the pairs $(r,r')$ are not linked through nullifying the off-diagonal elements. They are linked through swapping the rows and columns $r$ and $r'$ provided that $r'>r$. We make the following assumption:

*If $r=r'$ for some $r$, then the pair $(r,r')$ is removed from $O_R$.*

In this way, each pair from $O_R$ is an element of $\mathcal{P}_n$. Let us partition the matrices $A^{(k)}$, $k\geq 1$, obtained by the $J$-Jacobi method under the de Rijk strategy according to $J=\operatorname{diag}(I_m,I_{n-m})$,

$$A=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix},\qquad A^{(k)}=\begin{array}{c}\begin{bmatrix}A_{11}^{(k)} & A_{12}^{(k)}\\ A_{21}^{(k)} & A_{22}^{(k)}\end{bmatrix}\end{array}\begin{array}{l} m\\ n-m\end{array},\qquad k\geq 1.$$

Now we define a modification of the de Rijk pivot strategy that will be used in the global convergence proof. For simplicity, we call it the *modified de Rijk strategy*. First, we describe it in words.

We apply the $J$-Jacobi method to $A$ in the following order:

(1) it uses the de Rijk strategy for the block $A_{11}$,
(2) it nullifies the elements of the block $A_{12}$ using the row-cyclic strategy,
(3) it uses the de Rijk strategy for the block $A_{22}$.

Let us denote the associated sequence of pairs by $\tilde{O}_R$. We have

$$\begin{aligned} \tilde{O}_R = &(1,1'),(1,2:m),(2,2'),(2,3:m),\ldots,(m-1,(m-1)'),(m-1,m),\\ &(1,m+1:n),(2,m+1:n),\ldots,(m,m+1:n),\\ &(m+1,(m+1)'),(m+1,m+2:n),\ldots,(n-1,(n-1)'),(n-1,n). \end{aligned}$$

If $r=r'$ for some $r$, then the same assumption is applied to $\tilde{O}_R$.

We denote the pivot strategies linked to $O_R$ and $\tilde{O}_R$ by $\mathsf{I}_R$ and $\tilde{\mathsf{I}}_R$, respectively. Note that the sequences $O_R$ and $\tilde{O}_R$ are not necessarily orderings of $\mathcal{P}_n$ because the pairs $(r,r')$ with $r<r'$ are repeated in $O_R$ and $\tilde{O}_R$.

LEMMA 3.4. *We have*

$$\tilde{O}_R \sim O_R.$$

*Proof.* Let us prove that $\tilde{O}_R$ results from $O_R$ by a set of admissible transpositions. To this end, let us consider the sequence $O_R$.

For each $r$, $2\leq r\leq m-1$, the pair $(r,r')$ commutes with $(r-1,m+1:n)$. So, it can be moved just behind $(r-1,m)$. In this way we obtain

$$\begin{aligned} O_R \sim &(1,1'),(1,2:m),(2,2'),(1,m+1:n),(2,3:m),(3,3'),(2,m+1:n),\ldots,\\ &(m-1,(m-1)'),(m-2,m+1:n),(m-1,m),(m,m+1:n),\\ &(m+1,(m+1)'),(m+1,m+2:n),\ldots,(n-1,(n-1)'),(n-1,n). \end{aligned}$$

Next, note that for each $r$, $1 \leq r \leq m-1$, the sequence of pairs $(r, m+1:n)$ commutes with both $(s, s+1:m)$ and $(s, s')$, for $s > r$. Applying the appropriate transpositions of pairs for $r = 2, \ldots, m-1$, we obtain the sequence $\widetilde{O}_R$. □

Concerning $\mathsf{I}_R$ and $\widetilde{\mathsf{I}}_R$, we have to make an additional test. Note that the values of $r'$ in $\mathsf{I}_R$ and $\widetilde{\mathsf{I}}_R$ are chosen independently of each other. So, we have to test whether the values of $r'$ in $(r, r')$ are the same for $\mathsf{I}_R$ and $\widetilde{\mathsf{I}}_R$.

This is obvious for $r = 1$. Now, the proof can use mathematical induction with respect to $r$, $1 \leq r \leq m-1$. In the induction step, we use the fact that the plane transformations from $U_{r-1,m+1:n}$ do not change any diagonal element from the position $(r, r)$ to $(m, m)$. Therefore, $(r, r')$ has to be the same in both pivot strategies because they apply the same algorithm to obtain $r'$ when $r$ is given. This algorithm uses only the diagonal elements $a_{rr}^{(N-s_{n-r}+1)}, \ldots, a_{mm}^{(N-s_{n-r}+1)}$.

Now, using Lemma 3.2 with $O_R$ and $\widetilde{O}_R$ instead of $O$ and $\widetilde{O}$, we conclude that $A^{(N+1)} = \widetilde{A}^{(N+1)}$, or in another notation $A^{[1]} = \widetilde{A}^{[1]}$, where $A^{[1]}$ ($\widetilde{A}^{[1]}$) is obtained from $A$ after one full cycle under the de Rijk (modified de Rijk) pivot strategy.

EXAMPLE 3.5. To confirm our conclusions numerically, taking into account Remark 3.3, we have performed a numerical test. Using the MPFR library [12] with $\mathfrak{p}$ bits of precision, we compute the matrices $A^{[1]}$ and $\widetilde{A}^{[1]}$ by applying the $J$-Jacobi method to the same matrix pair $(A, J)$ under the pivot strategies $\mathsf{I}_R$ and $\widetilde{\mathsf{I}}_R$. We choose $A$ to be a Hermitian positive definite matrix of order $n = 20$, and we set $m = 10$. Note that $A - 0 \cdot J$ is positive definite, which ensures that the $J$-Jacobi method can be applied to the matrix pair $(A, J)$. We take $A$ as the ill-conditioned [1] symmetric Pascal matrix $S_n$, where $s_{ij} = \binom{i+j-2}{j-1}$.

Let $t_{pq}$ be either $\tan\theta_{pq}$ or $\tanh\theta_{pq}$, depending on the kind of transformation computed for the pivot pair $(p, q)$ in the first cycle of the algorithm with the de Rijk strategy. When the modified de Rijk strategy is used instead, the notation changes to $t'_{pq}$. For a fixed $\mathfrak{p}$, the generated $t_{pq}$ and $t'_{pq}$ are written out as quadruple precision values, each with 36 decimal digits after and one before the dot, as well as the elements of the iteration matrices $A^{[1]}$ and $\widetilde{A}^{[1]}$ after the first cycle of the de Rijk and the modified de Rijk strategies, respectively. The relative errors $\rho_t^{[\mathfrak{p}]} = \max_{p<q}(|t_{pq} - t'_{pq}|/|t_{pq}|)$ and $\rho_A^{[\mathfrak{p}]} = \|A^{[1]} - \widetilde{A}^{[1]}\|_F / \|A^{[1]}\|_F$ are computed from those outputs, for several values of $\mathfrak{p}$. Table 3.1 confirms that the two strategies produce effectively indistinguishable transformations (i.e., $\theta_{pq}$) and iteration matrices after a full cycle, for $\mathfrak{p}$ sufficiently large.

TABLE 3.1
*Numerical differences between the de Rijk and the modified de Rijk strategies after the first cycle.*

| $\mathfrak{p}$ | $\rho_t^{[\mathfrak{p}]}$ | $\rho_A^{[\mathfrak{p}]}$ | $\mathfrak{p}$ | $\rho_t^{[\mathfrak{p}]}$ | $\rho_A^{[\mathfrak{p}]}$ |
|---|---|---|---|---|---|
| 64 | $1.86762 \cdot 10^{-13}$ | $3.27395 \cdot 10^{-24}$ | 73 | 0 | $4.14221 \cdot 10^{-27}$ |
| 69 | $1.88331 \cdot 10^{-15}$ | $2.42271 \cdot 10^{-25}$ | 81 | 0 | 0 |

**3.2. The global convergence proof.** Recall that we have shown $A^{[1]} = \widetilde{A}^{[1]}$. By a simple induction argument, we have

$$A^{[t]} = \widetilde{A}^{[t]}, \qquad t \geq 1, \tag{3.3}$$

where $A^{[t]}$ and $\tilde{A}^{[t]}$ are obtained after a completion of cycle $t$. Recall that $A^{[t]} = A^{((t-1)N+1)}$, $\tilde{A}^{[t]} = \tilde{A}^{((t-1)N+1)}$, $t \geq 1$. Let

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \qquad \tilde{A}^{(k)} = \begin{bmatrix} \tilde{A}_{11}^{(k)} & \tilde{A}_{12}^{(k)} \\ \tilde{A}_{21}^{(k)} & \tilde{A}_{22}^{(k)} \end{bmatrix} \begin{matrix} m \\ n-m \end{matrix}, \qquad k \geq 1,$$

where the sequence $(\tilde{A}^{(k)}, k \geq 1)$ is obtained by applying the $J$-Jacobi method to $A$ under the modified de Rijk strategy.

Our aim is to show that

$$\lim_{t\to\infty} \operatorname{off}(\tilde{A}^{[t]}) = 0. \tag{3.4}$$

From (3.4) and (3.3), we see that (3.4) holds with $A^{[t]}$ instead of $\tilde{A}^{[t]}$. Afterwards, we show that $\operatorname{off}(A^{(k)})$ tends to zero as $k$ increases.

To prove (3.4), we consider a sequence of matrices that is obtained by splitting each cycle related to $\tilde{O}_R$ into three parts. Let

$$\tilde{O}_R = [\tilde{O}_1\, \tilde{O}_2\, \tilde{O}_3],$$

where

$$\begin{aligned} \tilde{O}_1 &= (1,1'), (1, 2:m), (2,2'), (2,3:m), \ldots, (m-1,(m-1)'), (m-1,m), \\ \tilde{O}_2 &= (1, m+1:n), (2, m+1:n), \ldots, (m, m+1:n), \\ \tilde{O}_3 &= (m+1,(m+1)'), (m+1, m+2:n), \ldots, (n-1,(n-1)'), (n-1,n). \end{aligned}$$

By $M_i$ we denote number of pairs in $\tilde{O}_i$, $1 \leq i \leq 3$. We have

$$M_1 = \frac{m(m-1)}{2}, \qquad M_2 = m(n-m), \qquad M_3 = \frac{(n-m)(n-m-1)}{2}.$$

Let us consider a cycle $t$, $t \geq 1$, of the $J$-Jacobi method under the *modified de Rijk strategy*. We denote by $\tilde{A}_1^{[t]}$, $\tilde{A}_2^{[t]}$, $\tilde{A}_3^{[t]}$, respectively, the iterated matrix obtained after completing the batch of transformations which nullify the elements of $\tilde{A}_{11}^{[t]}$, $\tilde{A}_{12}^{((t-1)N+M_1+1)}$, $\tilde{A}_{22}^{((t-1)N+M_1+M_2+1)}$. We have

$$\tilde{A}_1^{[t]} = \tilde{A}^{((t-1)N+M_1+1)}, \quad \tilde{A}_2^{[t]} = \tilde{A}^{((t-1)N+M_1+M_2+1)}, \quad \tilde{A}_3^{[t]} = \tilde{A}^{(tN+1)} = \tilde{A}^{[t+1]}.$$

Now consider the sequence $\mathcal{S}_1$,

$$\mathcal{S}_1 = \operatorname{off}(\tilde{A}),\ \operatorname{off}(\tilde{A}_1^{[1]}),\ \operatorname{off}(\tilde{A}_2^{[1]}),\ \operatorname{off}(\tilde{A}_3^{[1]}),\ \operatorname{off}(\tilde{A}_1^{[2]}),\ \operatorname{off}(\tilde{A}_2^{[2]}),\ \operatorname{off}(\tilde{A}_3^{[2]}),\ \operatorname{off}(\tilde{A}_1^{[3]}), \ldots$$

If the sequence $\mathcal{S}_1$ converges to zero, then relation (3.4) holds because $(\operatorname{off}(\tilde{A}^{[t]}), t \geq 1)$ is a subsequence of $\mathcal{S}_1$.

*The sequence $\mathcal{S}_1$ converges to zero if and only if the same is true for the following sequence $\mathcal{S}_2$*:

$$\mathcal{S}_2 = \operatorname{off}(\tilde{A}_3^{[1]}),\ \operatorname{off}(\tilde{A}_1^{[2]}),\ \operatorname{off}(\tilde{A}_2^{[2]}),\ \operatorname{off}(\tilde{A}_3^{[2]}),\ \operatorname{off}(\tilde{A}_1^{[3]}),\ \operatorname{off}(\tilde{A}_2^{[3]}),\ \operatorname{off}(\tilde{A}_3^{[3]}), \ldots$$

This is obvious since $\mathcal{S}_2$ is the 3-tail of $\mathcal{S}_1$.

To simplify notation for the subsequent analysis, let the sequence $(H^{(k)}, k \geq 1)$ be the $(M_1 + M_2)$-tail of $(\tilde{A}^{(k)}, k \geq 1)$, that is,

$$H^{(k)} = \tilde{A}^{(k+M_1+M_2)}, \qquad k \geq 1, \qquad H = H^{(1)} = \tilde{A}_2^{[1]}. \tag{3.5}$$

The sequence $\mathcal{S}_2$ is linked to the $J$-Jacobi method applied to $H$ under the pivot strategy defined by the rule:

(1) it uses the de Rijk strategy for the block $H_{22}$,
(2) it uses the de Rijk strategy for the block $H_{11}$,
(3) it nullifies the elements of the block $H_{12}$ using the row-cyclic strategy,
and it repeats this pattern in all subsequent steps.

We use the notation

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, \qquad H^{(k)} = \begin{array}{c} \begin{bmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{21}^{(k)} & H_{22}^{(k)} \end{bmatrix} \end{array} \begin{array}{l} m \\ n-m \end{array}, \qquad k \geq 1.$$

For $t \geq 1$, we have $H^{[t]} = H^{((t-1)N+1)}$, and $H_{11}^{[t]}$, $H_{12}^{[t]}$, $H_{21}^{[t]}$, $H_{22}^{[t]}$ are the blocks of $H^{[t]}$. Similar as before, the matrices $H_1^{[t]}$, $H_2^{[t]}$, $H_3^{[t]}$ are obtained after completing certain batches of transformations. We have $H_1^{[t]} = H^{((t-1)N+M_3+1)}$, $H_2^{[t]} = H^{((t-1)N+M_3+M_1+1)}$, $H_3^{[t]} = H^{[t+1]}$. With this notation, the sequence $\mathcal{S}_2$ takes the form

$$\mathcal{S}_2 = \mathrm{off}(H_1^{[1]}),\ \mathrm{off}(H_2^{[1]}),\ \mathrm{off}(H_3^{[1]}),\ \mathrm{off}(H_1^{[2]}),\ \mathrm{off}(H_2^{[2]}),\ \mathrm{off}(H_3^{[2]}),\ \mathrm{off}(H_1^{[3]}), \ldots$$

To prove that $\mathcal{S}_2$ converges to zero, we use several lemmas.

LEMMA 3.6. *We have*

$$\lim_{k\to\infty} \mathrm{off}(H_{11}^{(k)}) = 0 \qquad \textit{and} \qquad \lim_{k\to\infty} \mathrm{off}(H_{22}^{(k)}) = 0. \tag{3.6}$$

*Proof.* The matrix $H_{22}^{((t-1)N+M_3+1)}$ is a result of applying one cycle of the elementwise Jacobi method for Hermitian matrices to $H_{22}^{[t]}$ under the de Rijk pivot strategy. Using [19, Theorem 5.4], we have

$$\begin{aligned} \mathrm{off}(H_{22}^{((t-1)N+M_3+1)}) &\leq \left[ 1 - \prod_{\substack{k = (t-1)N+1 \\ i(k) \neq j(k)-1}}^{(t-1)N+M_3} \cos^2(\theta_k) \right]^{1/2} \mathrm{off}(H_{22}^{((t-1)N+1)}) \\ &\leq \mu_2 \mathrm{off}(H_{22}^{[t]}), \end{aligned}$$

where

$$\mu_2 = \sqrt{1 - 2^{-\frac{M_3-(n-m-1)}{2}}} < 1.$$

Here, we used the fact that the angles $\theta_k$ are in the interval $[-\pi/4, \pi/4]$.

In a similar way, we conclude that

$$\begin{aligned} \mathrm{off}(H_{11}^{((t-1)N+M_3+M_1+1)}) &\leq \left[ 1 - \prod_{\substack{k = (t-1)N+M_3+1 \\ i(k) \neq j(k)-1}}^{(t-1)N+M_3+M_1} \cos^2(\theta_k) \right]^{1/2} \mathrm{off}(H_{11}^{((t-1)N+M_3+1)}) \\ &\leq \mu_1 \mathrm{off}(H_{11}^{((t-1)N+M_3+1)}), \end{aligned}$$

where

$$\mu_1 = \sqrt{1 - 2^{-\frac{M_1-(m-1)}{2}}} < 1.$$

Note that the Jacobi steps nullifying the elements of $H_{22}^{((t-1)N+1)}$ ($H_{11}^{((t-1)N+M_3+1)}$) do not change the elements of $H_{11}^{((t-1)N+1)}$ ($H_{22}^{((t-1)N+M_3+1)}$). The same holds for the permutational transformations using $I_{rr'}$. Therefore, we have

$$H_{11}^{((t-1)N+M_3+1)} = H_{11}^{((t-1)N+1)} = H_{11}^{[t]}, \qquad H_{22}^{((t-1)N+M_3+M_1+1)} = H_{22}^{((t-1)N+M_3+1)}.$$

Combining the above relations, we obtain

$$\text{off}\,(H_{22}^{((t-1)N+M_3+M_1+1)}) \le \mu_2\,\text{off}\,(H_{22}^{[t]}), \qquad t \ge 0, \tag{3.7}$$

$$\text{off}\,(H_{11}^{((t-1)N+M_3+M_1+1)}) \le \mu_1\,\text{off}\,(H_{11}^{[t]}), \qquad t \ge 0. \tag{3.8}$$

Let us consider the impact of the batch of $M_2$-hyperbolic transformations on the values of $\text{off}\,(H_{11}^{((t-1)N+M_3+M_1+1)})$ and $\text{off}\,(H_{22}^{((t-1)N+M_3+M_1+1)})$. We first consider the case when $U^{(k)}$ is as in relation (2.21), that is, $U^{(k)} = V^{(k)}$.

From relation (2.19) or (2.20), we conclude that each hyperbolic plane transformation $U^{(k)}$ tends to the identity matrix $I_n$ as $k$ increases. Since the hyperbolic transformations appear one after another, we can write

$$W_t = U_{1,m+1:n}^{[t]} U_{2,m+1:n}^{[t]} \cdots U_{m,m+1:n}^{[t]} = I_n + E_t, \qquad t \ge 1, \qquad \lim_{t\to\infty} E_t = 0. \tag{3.9}$$

Here, $W_t$ is $J$-unitary, and $E_t$ can be viewed as a perturbation matrix. We easily obtain

$$H^{[t+1]} = W_t^* H^{((t-1)N+M_3+M_1+1)}) W_t = H^{((t-1)N+M_3+M_1+1)}) + F_t, \qquad t \ge 1, \tag{3.10}$$

where

$$\begin{aligned} F_t = H^{((t-1)N+M_3+M_1+1)} E_t + E_t^* H^{((t-1)N+M_3+M_1+1)} \\ + E_t^* H^{((t-1)N+M_3+M_1+1)} E_t, \qquad t \ge 1. \end{aligned} \tag{3.11}$$

To bound $\|H^{(k)}\|_F$, we use relation (2.16), which holds for any $J$-Jacobi process, and the fact that $\text{trace}(H^{(k)})$ is non-increasing with $k$. We obtain

$$\begin{aligned} \|H^{(k)}\|_F &\le \text{trace}(H^{(k)}) + (n-2m)\mu + \|\mu J\|_F \\ &\le \text{trace}(H) + (n + \sqrt{n} - 2m)|\mu|, \qquad k \ge 1. \end{aligned} \tag{3.12}$$

Combining (3.11), (3.12), and (3.9), one obtains

$$\lim_{t\to\infty} F_t = 0. \tag{3.13}$$

The relations (3.8), (3.7), (3.10), and (3.13) imply

$$\text{off}\,(H_{11}^{[t+1]}) \le \mu_1\,\text{off}\,(H_{11}^{[t]}) + \nu_1^{[t]}, \quad t \ge 0, \qquad \lim_{t\to\infty} \nu_1^{[t]} = 0, \tag{3.14}$$

$$\text{off}\,(H_{22}^{[t+1]}) \le \mu_2\,\text{off}\,(H_{22}^{[t]}) + \nu_2^{[t]}, \quad t \ge 0, \qquad \lim_{t\to\infty} \nu_2^{[t]} = 0. \tag{3.15}$$

Now, applying [14, Lemma 2.2] to the sequences generated by (3.14) and (3.15), we obtain

$$\lim_{t\to\infty} \text{off}\,(H_{11}^{[t]}) = 0 \qquad \text{and} \qquad \lim_{t\to\infty} \text{off}\,(H_{22}^{[t]}) = 0. \tag{3.16}$$

To prove (3.16) for the case when $U^{(k)}$ has the form (2.6), with $\hat{\Phi}^{(k)}$ from (2.7), we note that the changes occur in relations (3.9), (3.10), and (3.11). In (3.9) we have a unitary diagonal matrix $\Phi_t$ instead of $I_n$. In (3.10) we have $\Phi_t^* H^{((t-1)N+M_3+M_1+1)}\Phi_t$ instead of $H^{((t-1)N+M_3+M_1+1)}$. In (3.11) the first (second) term on the right-hand side has the factor $\Phi_t^*$ ($\Phi_t$). In any case, the relations (3.13), (3.14), (3.15), and consequently the relation (3.16), hold.

To prove assertion (3.6) of the lemma, it is sufficient to show that

$$\text{off}(H_{11}^{(k)}) \le 1.1\,\text{off}(H_{11}^{[t]}), \qquad \text{off}(H_{22}^{(k)}) \le 1.1\,\text{off}(H_{22}^{[t]}), \qquad k \in \mathcal{C}_t, \tag{3.17}$$

holds for a sufficiently large $t$. For $k$, $(t-1)N+1 \le k \le (t-1)N+M_3+M_1+1$, we have $\text{off}(H_{11}^{(k+1)}) \le \text{off}(H_{11}^{(k)})$ and $\text{off}(H_{22}^{(k+1)}) \le \text{off}(H_{22}^{(k)})$. Thus, we consider $k \in \mathcal{C}_t'$, where

$$\mathcal{C}_t' = \{k;\ (t-1)N+M_3+M_1+1 \le k \le tN\}, \qquad t \ge 1. \tag{3.18}$$

The set $\mathcal{C}_t' \subset \mathcal{C}_t$ consists of hyperbolic steps in cycle $t$. Let

$$W_t^{(k)} = U^{((t-1)N+M_3+M_1+1)}U^{((t-1)N+M_3+M_1+2)}\cdots U^{(k)}, \qquad k \in \mathcal{C}_t'.$$

If all $U^{(k)}$ have the form (2.6), with $\hat{\Phi}^{(k)}$ from (2.7), then using (2.19) we obtain

$$W_t^{(k)} = \Phi_t^{(k)} + E_t^{(k)}, \qquad \lim_{t\to\infty} E_t^{(k)} = 0,$$

where each $\Phi_t^{(k)}$ is unitary and diagonal. Then we have

$$H^{(k+1)} = [W_t^{(k)}]^* H^{((t-1)N+1)} W_t^{(k)} = [\Phi_t^{(k)}]^* H^{[t]} \Phi_t^{(k)} + F_t^{(k)}, \qquad \lim_{t\to\infty} F_t^{(k)} = 0,$$

and the assertion related to relation (3.17) is implied by the last relation.

If all $U^{(k)}$ have the form (2.21), then the proof is even simpler because the matrix $\Phi_t^{(k)}$ is replaced by the identity. □

LEMMA 3.7. *We have*

$$\lim_{k\to\infty} \|H_{12}^{(k)}\|_F = 0.$$

*Proof.* We consider cycle $t$, $t \ge 1$, of the $J$-Jacobi process for $H$. The value of $t$ will be specified later. In the first part of the proof, we consider the first cycle, i.e., we assume $t = 1$.

Note that the unitary transformations $U^{(k)}$ and the transpositions $I_{rr'}$ do not change the Frobenius norm of $H_{12}$. Therefore, we have $\|H_{12}^{(M_3+M_1+1)}\|_F = \|H_{12}\|_F$. We consider the evolution of an element of $H_{12}^{(M_3+M_1+1)}$ from the moment when it becomes zero up to the end of the cycle. To this end, we use the notation

$$U_{ij} = U_{ij}^{(k_{ij})}, \qquad 1 \le i \le m,\ m+1 \le j \le n,$$
$$k_{ij} = M_3 + M_1 + (i-1)(n-m) + j - m,$$

and

$$\hat{U}_{ij}^{(k_{ij})} = \begin{bmatrix} u_{ii}^{(k_{ij})} & u_{ij}^{(k_{ij})} \\ u_{ji}^{(k_{ij})} & u_{jj}^{(k_{ij})} \end{bmatrix}, \qquad 1 \le i \le m,\ \ m+1 \le j \le n.$$

One can verify that $h_{ij}^{(k_{ij})}$ is the pivot element at the $(i,j)$-position. Let us consider the evolution of the element at the $(p,q)$-position, $1 \le p \le m$, $m+1 \le q \le n$. Set $\mathsf{n}_{pq} = k_{pq}$,

$$\mathsf{n}_{pq} = M_3 + M_1 + (p-1)(n-m) + q - m.$$

Thus, $h_{pq}^{(\mathsf{n}_{pq})}$ is the pivot element in step $\mathsf{n}_{pq}$. If $(p,q)=(m,n)$, then we have the equality $h_{pq}^{(\mathsf{n}_{mn}+1)}=h_{mn}^{[1]}=0$.

If $q<n$, then the left-hand transformations defined by the pivot positions $(p,q+1),\ldots$, $(p,n)$ affect the element at the $(p,q)$-position. We have $h_{pq}^{(\mathsf{n}_{pq}+1)}=0$, and

$$\begin{aligned}
h_{pq}^{(\mathsf{n}_{pq}+2)} &= \bar{u}_{pp}^{(k_{p,q+1})}h_{pq}^{(\mathsf{n}_{pq}+1)}+\bar{u}_{q+1,p}^{(k_{p,q+1})}h_{q+1,q}^{(\mathsf{n}_{pq}+1)},\\
&\cdots\\
h_{pq}^{(\mathsf{n}_{pq}+n-q)} &= \bar{u}_{pp}^{(k_{p,n-1})}h_{pq}^{(\mathsf{n}_{pq}+n-q-1)}+\bar{u}_{n-1,p}^{(k_{p,n-1})}h_{n-1,q}^{(\mathsf{n}_{pq}+n-q-1)},\\
h_{pq}^{(\mathsf{n}_{pq}+n-q+1)} &= \bar{u}_{pp}^{(k_{pn})}h_{pq}^{(\mathsf{n}_{pq}+n-q)}+\bar{u}_{np}^{(k_{pn})}h_{nq}^{(\mathsf{n}_{pq}+n-q)}.
\end{aligned}$$

Here, for a complex number $x$ the complex conjugate is denoted by $\bar{x}$. One can easily verify that $h_{rq}^{(\mathsf{n}_{pq}+r-q)}=h_{rq}^{(\mathsf{n}_{pq}+1)}$, for all $q+1\le r\le n$. Hence, we obtain

$$\begin{aligned}
h_{pq}^{(\mathsf{n}_{pq}+2)} &= 0\cdot\bar{u}_{pp}^{(k_{p,q+1})}+h_{q+1,q}^{(\mathsf{n}_{pq}+1)}\bar{u}_{q+1,p}^{(k_{p,q+1})},\\
h_{pq}^{(\mathsf{n}_{pq}+3)} &= h_{pq}^{(\mathsf{n}_{pq}+2)}\bar{u}_{pp}^{(k_{p,q+2})}+h_{q+2,q}^{(\mathsf{n}_{pq}+1)}\bar{u}_{q+2,p}^{(k_{p,q+2})},\\
&\cdots\\
h_{pq}^{(\mathsf{n}_{pq}+n-q)} &= h_{pq}^{(\mathsf{n}_{pq}+n-q-1)}\bar{u}_{pp}^{(k_{p,n-1})}+h_{n-1,q}^{(\mathsf{n}_{pq}+1)}\bar{u}_{n-1,p}^{(k_{p,n-1})},\\
h_{pq}^{(\mathsf{n}_{pq}+n-q+1)} &= h_{pq}^{(\mathsf{n}_{pq}+n-q)}\bar{u}_{pp}^{(k_{pn})}+h_{nq}^{(\mathsf{n}_{pq}+1)}\bar{u}_{np}^{(k_{pn})}.
\end{aligned}$$

If $p<m$, then we also have the right-hand side transformations defined by the pivot pairs $(p+1,q),\ldots,(m,q)$ that change the element at position $(p,q)$. From the equality $h_{pq}^{(\mathsf{n}_{pq}+n-m)}=h_{pq}^{(\mathsf{n}_{pq}+n-q+1)}$, we obtain in a similar way as above that

$$\begin{aligned}
h_{pq}^{(\mathsf{n}_{pq}+n-m+1)} &= h_{pq}^{(\mathsf{n}_{pq}+n-q+1)}u_{qq}^{(k_{p+1,q})}+h_{p,p+1}^{(\mathsf{n}_{pq}+n-m)}u_{p+1,q}^{(k_{p+1,q})},\\
h_{pq}^{(\mathsf{n}_{pq}+2(n-m)+1)} &= h_{pq}^{(\mathsf{n}_{pq}+n-m+1)}u_{qq}^{(k_{p+2,q})}+h_{p,p+2}^{(\mathsf{n}_{pq}+2(n-m))}u_{p+2,q}^{(k_{p+2,q})},\\
&\cdots\\
h_{pq}^{(\mathsf{n}_{pq}+(m-p)(n-m)+1)} &= h_{pq}^{(\mathsf{n}_{pq}+(m-p-1)(n-m)+1)}u_{qq}^{(k_{mq})}+h_{pm}^{(\mathsf{n}_{pq}+(m-p)(n-m))}u_{mq}^{(k_{mq})}.
\end{aligned}$$

Note that $h_{pq}^{(N+1)}=h_{pq}^{(\mathsf{n}_{pq}+(m-p)(n-m)+1)}$. Combining the two sets of equations above, we obtain

$$\begin{aligned}
h_{pq}^{(N+1)} = h_{pm}^{(\mathsf{n}_{pq}+(m-p)(n-m))}u_{mq}^{(k_{mq})} &+ h_{p,m-1}^{(\mathsf{n}_{pq}+(m-p-1)(n-m))}u_{m-1,q}^{(k_{m-1,q})}u_{qq}^{(k_{mq})}+\cdots\\
&+ h_{p,p+1}^{(\mathsf{n}_{pq}+n-m)}u_{p+1,q}^{(k_{p+1,q})}u_{qq}^{(k_{p+2,q})}\cdots u_{qq}^{(k_{mq})}\\
&+ h_{nq}^{(\mathsf{n}_{pq}+1)}\bar{u}_{np}^{(k_{pn})}u_{qq}^{(k_{p+1,q})}\cdots u_{qq}^{(k_{mq})}+\cdots\\
&+ h_{q+1,q}^{(\mathsf{n}_{pq}+1)}\bar{u}_{q+1,p}^{(k_{p,q+1})}\bar{u}_{pp}^{(k_{p,q+2})}\cdots\bar{u}_{pp}^{(k_{pn})}u_{qq}^{(k_{p+1,q})}\cdots u_{qq}^{(k_{mq})}.
\end{aligned}$$

Let

$$\mathsf{c}^{[1]}=\max_{M_3+M_1+1\le k\le N}\cosh(\theta_k),\qquad \mathsf{s}^{[1]}=\max_{M_3+M_1+1\le k\le N}|\sinh(\theta_k)|. \tag{3.19}$$

Using the Cauchy-Schwarz inequality twice and relation (3.19), we obtain, for $1 \leq p \leq m$, $m+1 \leq q \leq n$,

$$\begin{aligned}
|h_{pq}^{(N+1)}| \leq{}& \left[|h_{p,p+1}^{(\mathsf{n}_{pq}+n-m)}|^2 + \cdots + |h_{pm}^{(\mathsf{n}_{pq}+(m-p)(n-m))}|^2\right]^{\frac{1}{2}} \\
&\quad \times \left[|u_{p+1,q}^{(k_{p+1,q})}|^2 + \cdots + u_{mq}^{(k_{mq})}|^2\right]^{\frac{1}{2}} [\mathsf{c}^{[1]}]^{\frac{m-p-1}{2}} \\
&+ \left[|h_{q+1,q}^{(\mathsf{n}_{pq}+1)}|^2 + \cdots + |h_{nq}^{(\mathsf{n}_{pq}+1)}|^2\right]^{\frac{1}{2}} \\
&\quad \times \left[|\bar{u}_{q+1,p}^{(k_{p,q+1})}|^2 + \cdots + |\bar{u}_{np}^{(k_{pn})}|^2\right]^{\frac{1}{2}} [\mathsf{c}^{[1]}]^{\frac{n-3}{2}} \\
\leq{}& [\mathsf{c}^{[1]}]^{\frac{m-p-1}{2}} \sqrt{\frac{m-p}{2}} \max_{1\leq k\leq N} \operatorname{off}(H_{11}^{(k)}) \cdot \sqrt{m-p} \max_{p+1\leq i\leq m} |u_{iq}^{(k_{iq})}| \\
&+ [\mathsf{c}^{[1]}]^{\frac{n-3}{2}} \sqrt{\frac{n-q}{2}} \operatorname{off}(H_{22}^{(\mathsf{n}_{pq}+1)}) \cdot \sqrt{n-q} \max_{q+1\leq i\leq n} |u_{ip}^{(k_{pi})}| \\
\leq{}& [\mathsf{c}^{[1]}]^{\frac{m-2}{2}} \frac{m-p}{\sqrt{2}} \max_{1\leq k\leq N} \operatorname{off}(H_{11}^{(k)}) \max_{p+1\leq i\leq m} |u_{iq}^{(k_{iq})}| \\
&+ [\mathsf{c}^{[1]}]^{\frac{n-3}{2}} \frac{n-q}{\sqrt{2}} \operatorname{off}(H_{22}^{(\mathsf{n}_{pq}+1)}) \max_{q+1\leq i\leq n} |u_{ip}^{(k_{pi})}|.
\end{aligned}$$

Here, we also used $\cosh(\theta_k) \geq 1$ and

$$m-p+n-(q+2)+1 \leq m-p+n-q-1 \leq m-1+n-(m+1)-1 \leq n-3.$$

To bound $\|H_{12}^{(N+1)}\|_F^2$ we use the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, $a \geq 0$, $b \geq 0$, and apply some rough estimates. Let $a$ and $b$ equal the final terms containing $\operatorname{off}(H_{11}^{(k)})$ and $\operatorname{off}(H_{22}^{(\mathsf{n}_{pq}+1)})$, respectively. We obtain

$$\begin{aligned}
\|H_{12}^{(N+1)}\|_F^2 \leq{}& m(n-m)\left[(m-1)^2 \max_{1\leq k\leq N} \operatorname{off}^2(H_{11}^{(k)}) + (n-m-1)^2 \max_{1\leq k\leq N} \operatorname{off}^2(H_{22}^{(k)})\right] \\
&\times [\mathsf{c}^{[1]}]^{n-3}[\mathsf{s}^{[1]}]^2 \\
\leq{}& m(n-m)(n-2)^2 \max_{1\leq r\leq 2} \left\{\max_{1\leq k\leq N} \operatorname{off}^2(H_{rr}^{(k)})\right\} [\mathsf{c}^{[1]}]^{n-3}[\mathsf{s}^{[1]}]^2.
\end{aligned}$$

Since $m(n-m) \leq n^2/4$, we obtain

$$\|H_{12}^{(N+1)}\|_F \leq \frac{n(n-2)}{2} \max\left\{\max_{1\leq k\leq N} \operatorname{off}(H_{11}^{(k)}),\ \max_{1\leq k\leq N} \operatorname{off}(H_{22}^{(k)})\right\} [\mathsf{c}^{[1]}]^{\frac{n-3}{2}} [\mathsf{s}^{[1]}].$$

Now consider cycle $t$, $t \geq 1$. Instead of the last inequality, we have

$$\|H_{12}^{[t]}\|_F \leq \frac{n(n-2)}{2} \max\left\{\max_{k\in\mathcal{C}_t} \operatorname{off}(H_{11}^{(k)}),\ \max_{k\in\mathcal{C}_t} \operatorname{off}(H_{22}^{(k)})\right\}[\mathsf{c}^{[t]}]^{\frac{n-3}{2}}[\mathsf{s}^{[t]}],\ \ t \geq 1, \tag{3.20}$$

where

$$\mathsf{c}^{[t]} = \max_{k\in\mathcal{C}_t'} \cosh(\theta_k), \qquad \mathsf{s}^{[t]} = \max_{k\in\mathcal{C}_t'} |\sinh(\theta_k)|.$$

Here, $\mathcal{C}_t'$ is defined in (3.18). By (2.19) we have $\mathsf{s}^{[t]} \to 0$, $\mathsf{c}^{[t]} \searrow 1$, and consequently, $[\mathsf{c}^{[t]}]^{\frac{n-2}{2}}[\mathsf{s}^{[t]}] \to 0$, as $t \to \infty$. Hence, from (3.20) and Lemma 3.6 we get $\lim\limits_{t\to\infty} \|H_{12}^{[t]}\|_F = 0$.

The rest of the proof uses the last lines in the proof of Lemma 3.6, in particular, starting from relation (3.18) until the end of the proof. □

THEOREM 3.8. *Let $A$ be a Hermitian matrix of order $n$ such that the matrix pair $(A, J)$, $J = \mathrm{diag}(I_m, -I_{n-m})$, $1 \leq m \leq n-1$, is positive definite. Let the sequence of matrices $(A^{(k)}, k \geq 1)$ be obtained by applying the $J$-Jacobi method to $A$ under the de Rijk pivot strategy. Then the sequence $(A^{(k)}, k \geq 1)$ converges to a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_m, -\lambda_{m+1}, \ldots, -\lambda_n$ are the eigenvalues of $JA$. The same is true provided that $A$ is real symmetric.*

*Proof.* Let the sequence of matrices $(\tilde{A}^{(k)}, k \geq 1)$ be obtained by applying the $J$-Jacobi method to $A$ under the pivot strategy $\tilde{\mathsf{I}}_R$. Let the sequence $(H^{(k)}, k \geq 1)$ be defined by relation (3.5). Then the Lemmas 3.6 and 3.7 imply

$$\lim_{k\to\infty} \mathrm{off}(H^{(k)}) = 0. \tag{3.21}$$

Relation (3.21) is equivalent to $\lim_{k\to\infty} \mathrm{off}(\tilde{A}^{(k)}) = 0$. Thus, relation (3.4) is proven. From relation (3.3) we have $A^{[t]} = \tilde{A}^{[t]}$, $t \geq 1$. In this way we have shown

$$\lim_{t\to\infty} \mathrm{off}(A^{[t]}) = 0. \tag{3.22}$$

If we show that

$$\mathrm{off}(A^{(k)}) \leq 1.1\, \mathrm{off}(A^{[t]}), \qquad k \in \mathcal{C}_t, \tag{3.23}$$

holds for a sufficiently large $t$ and for $\mathrm{off}(A^{[t]}) > 0$, then relation (3.22) implies

$$\lim_{k\to\infty} \mathrm{off}(A^{(k)}) = 0. \tag{3.24}$$

The unitary transformations cannot increase $\mathrm{off}(A^{(k)})$. Hence, to prove (3.23), it is sufficient to consider only the hyperbolic steps under the pivot strategy $\mathsf{I}_R$. We show that inequality (3.23) holds for $k \in \mathcal{C}''_t$, where

$$\mathcal{C}''_t = \{k \in \mathcal{C}_t;\ k \text{ counts hyperbolic steps}\}, \qquad t \geq 1,$$

when $t$ is sufficiently large.

For a given $t$ and $k \in \mathcal{C}''_t$, we have $A^{(k+1)} = [U^{(k)}]^* A^{(k)} U^{(k)}$. We can write

$$\mathrm{off}(A^{(k+1)}) = \mathrm{off}(A^{(k)}) + \varepsilon^{(k)} = \mathrm{off}(A^{((t-1)N+M_1+1)}) + \epsilon_t^{(k)}, \qquad k \in \mathcal{C}''_t.$$

Since the hyperbolic angle $\theta_k$ tends to zero as $k$ increases, we have, for $k \in \mathcal{C}''_t$,

$$\epsilon_t^{(k)} = \varepsilon^{((t-1)N+M_1+1)} + \varepsilon^{((t-1)N+M_1+2)} + \cdots + \varepsilon^{(k)} \to 0, \qquad \text{as } t \to \infty.$$

This proves that (3.23) holds for a sufficiently large $t$. Thus, (3.24) holds.

It remains to show that the diagonal elements of $A^{(k)}$ converge. Although the matrix $A$ here is not positive definite as it is in [21], the proof is identical to the corresponding part of the proof of [21, Theorem 3.7]. For the sake of completeness, we provide the details.

Note that $JA$ and each $JA^{(k)}$, $k \geq 1$, have the same eigenvalues as the problem $Ax = \lambda Jx$. Let us arrange the eigenvalues of $JA$ non-increasingly:

$$\begin{aligned}\lambda_1 = \lambda_2 = \cdots = \lambda_{s_1} > \lambda_{s_1+1} = \cdots = \lambda_{s_2} > \cdots > \lambda_{s_{p-1}+1} = \cdots = \lambda_{s_p}& \\ > -\lambda_{s_p+1} = \cdots = -\lambda_{s_{p+1}} > \cdots > -\lambda_{s_{\omega-1}+1} = \cdots = -\lambda_{s_\omega}&.\end{aligned}$$

Here, $s_p = m$. Obviously, $n_r = s_r - s_{r-1}$ is the multiplicity of $\lambda_{s_r}$, $1 \le r \le \omega$. Let us denote the minimum gap between two different eigenvalues by $3\delta$,

$$3\delta = \min\left\{ \min_{1\le r\le p-1}(\lambda_{s_r} - \lambda_{s_{r+1}}), \lambda_{s_p} + \lambda_{s_{p+1}}, \min_{p+1\le r\le \omega-1}(\lambda_{s_{r+1}} - \lambda_{s_r}) \right\}. \tag{3.25}$$

For each $r$, $1 \le r \le p$ ($p+1 \le r \le \omega$), let $\mathcal{D}_r$ be the disk of radius $\delta$ with center at $\lambda_{s_r}$ ($-\lambda_{s_r}$). Relation (3.25) implies that the disks are disjoint and that the minimum distance between two disks equals $\delta$.

Since (3.24) holds, there exists a $k_0 \ge 1$ such that

$$\|A^{(k)} - \operatorname{diag}(A^{(k)})\|_\infty \le \frac{\delta}{2n}, \qquad k \ge k_0. \tag{3.26}$$

This condition means that all Geršgorin disks of $JA^{(k)}$ have radii not larger than $\delta/(2n)$. Now, the theory of Geršgorin disks then implies that each disk $\mathcal{D}_r$ contains exactly $n_r$ Geršgorin disks of $JA^{(k)}$ forming one connected component of the union of all Geršgorin disks. In particular, each $\mathcal{D}_r$ contains exactly $n_r$ diagonal elements. Since the radii of the Geršgorin disks tend to 0 as $k$ increases, the diagonal elements of $JA^{(k)}$ approach the eigenvalues of $JA$. Hence, it remains to show that at step $k$, $k \ge k_0$, the diagonal elements of $JA^{(k)}$ cannot jump from one disk $\mathcal{D}_r$ to another.

To prove this, note that (3.26) implies $|a^{(k)}_{i(k)j(k)}| \le \delta/(2n)$, $k \ge k_0$. Since $|\tanh(\theta_k)| < 1$ ($|\tan(\theta_k)| \le 1$), the formulas (2.11) and (2.12) ((2.14) and (2.15)) imply that neither $a^{(k)}_{i(k)i(k)}$ nor $-a^{(k)}_{j(k)j(k)}$ (neither $\rho_k a^{(k)}_{i(k)i(k)}$ nor $\rho_k a^{(k)}_{j(k)j(k)}$, $\rho_k \in \{-1, 1\}$) can move to another disk. This completes the proof in the case of a complex Hermitian matrix $A$.

If $A$ is real, then the simpler, real $J$-Jacobi method [49] is employed. It is easy to verify that all the lemmas and the theorem hold as well for the real method. □

### 3.3. Global convergence of the stable $J$-Jacobi method.

The stable method was investigated for the case of a real symmetric matrix $A$ by Veselić in [49]. Here, we briefly consider the complex method.

For the stable method, we denote the iterated matrix, the transformation matrix, the angle $\theta_k$, and the phase $\phi_k$ by $\mathsf{A}^{(k)} = (\mathsf{a}^{(k)}_{rs})$, $\mathsf{U}^{(k)}$, $\vartheta_k$ and $\varphi_k$, respectively. The stable method is defined as follows:

If for a given $k$ we have $1 \le i(k) \le m < j(k) \le n$ and $|\tanh(\theta_k)| \le t_{\max}$, then set $\varphi_k = \phi_k$ and $\vartheta_k = \theta_k$. Otherwise, set $\varphi_k = \phi_k$, $\vartheta_k = -\tanh^{-1}(t_{\max})$.

If $k$ is such that $1 \le i(k) < j(k) \le m$ or $m+1 \le i(k) < j(k) \le n$, then we have a unitary plane rotation $\mathsf{U}^{(k)}$ with $\vartheta_k = \theta_k$ and $\varphi_k = \phi_k$, i.e., $\mathsf{U}^{(k)} = U^{(k)}$.

In [49] the value $t_{\max} = 0.5$ has been suggested as a good choice for practical computation. In Section 4 we use $t_{\max} = 0.8$. However, the convergence proof below does not assume any specific value of $t_{\max}$.

The convergence of the stable method is implied by combining Theorem 3.8 with [49, Lemma 2.2].

COROLLARY 3.9. *Let the eigenvalue problem $Ax = \lambda Jx$ be positive definite. Then the stable $J$-Jacobi method is globally convergent under the de Rijk pivot strategy. The same is true for the real stable method.*

*Proof.* We first show that the sequence $(\operatorname{trace}(\mathsf{A}^{(k)}), k \ge 1)$ is non-increasing and convergent. The proof is a slight modification of [49, Lemma 2.2].

First, note that $\operatorname{trace}(\mathsf{A}^{(k)})$ is invariant under a similarity transformation with a diagonal unitary matrix. Therefore, it is irrelevant whether we assume the transformation from relations (2.7), (2.8), or from (2.21). The phase $\phi_k$ will not be present in the analysis.

We know that the trace function is invariant under unitary transformations. For any hyperbolic step $k$, we have (see [49, Lemma 2.2])

$$\begin{aligned} \operatorname{trace}(\mathsf{A}^{(k)}) - \operatorname{trace}(\mathsf{A}^{(k+1)}) &= (\mathsf{a}_{i(k)i(k)}^{(k)} + \mathsf{a}_{j(k)j(k)}^{(k)}) \left(1 - \frac{1-\tanh(2\theta_k)\tanh(2\vartheta_k)}{\sqrt{1-\tanh^2(2\vartheta_k)}}\right) \\ &\geq (\mathsf{a}_{i(k)i(k)}^{(k)} + \mathsf{a}_{j(k)j(k)}^{(k)}) \tanh(\vartheta_k)\tanh(2\vartheta_k). \end{aligned} \tag{3.27}$$

Since the matrix $\mathsf{A}^{(k)} - \mu J$ is positive definite and since $1 \leq i(k) \leq m < j(k) \leq n$, we obtain from relation (2.2) that $a_{i(k)i(k)}^{(k)} + a_{j(k)j(k)}^{(k)} > \delta_0 > 0$. Therefore, (3.27) implies that the sequence $(\operatorname{trace}(\mathsf{A}^{(k)}),\ k \geq 1)$ is non-increasing. Since it is bounded from below by $(2m-n)\mu$, it is convergent.

Consequently, we have $\vartheta_k \to 0$ as $k$ increases over the hyperbolic steps. Hence, for $k_0$ large enough (which may depend on $A$ and $t_{\max}$), we have $\vartheta_k = \theta_k$, $k \geq k_0$. Thus, for $k \geq k_0$, the stable method reduces to the standard one, and the assertion of the corollary is implied by Theorem 3.8. Obviously, the proof holds for the real $J$-Jacobi method as well. □

We end this section by briefly considering the asymptotic convergence.

**3.4. A brief analysis concerning asymptotic convergence.** We know that for $k$ large enough, the stable method reduces to the standard one. Hence, we consider only the latter one.

From [19] we know that the permutations and the unitary transformations, under the de Rijk pivot strategy, gradually order the diagonal elements within the diagonal blocks of order $m$ and $n-m$ non-increasingly. The hyperbolic $J$-unitary transformations approach the set of unitary diagonal matrices. Therefore, for a sufficiently large $k$, they will not change the affiliation of the diagonal elements of $JA^{(k)}$ to the eigenvalues of $JA$. This is shown in the proof of Theorem 3.8.

In the case of simple eigenvalues of $JA$, the de Rijk pivot strategy will ultimately be reduced to the row-cyclic one. Then we can apply [11, Theorem 3.7], which states that for the $J$-Jacobi method under the row-cyclic strategy it holds that

$$\operatorname{off}(A^{(N+1)}) \leq \frac{\operatorname{off}^2(A)}{\delta},$$

provided that $\operatorname{off}(A) \leq \delta/(m(n-m))$.

Now, consider the case of multiple eigenvalues. The de Rijk pivot strategy will ultimately order the diagonal elements so that those converging to the same eigenvalue will occupy successive positions on the diagonal. Then the transposition matrices (which are part of the de Rijk strategy) can swap only those diagonal elements that converge to the same eigenvalue. We are confident that this fact can be used in the proof of [11, Theorem 3.7], so that after some adaptation it holds for the de Rijk pivot strategy.

Another important approach to the asymptotic analysis of the $J$-Jacobi method is the one of using scaled iteration matrices. It assumes that $A$ is positive definite. Instead of $(\operatorname{off}(A^{(k)}),\ k \geq 1)$, this approach considers the sequence $(\operatorname{off}([D^{(k)}]^{-1/2}A^{(k)}[D^{(k)}]^{-1/2}),\ k \geq 1)$, where $D^{(k)}$ is the diagonal part of $A^{(k)}$. Since in the later stage of the process, the de Rijk strategy is reduced (or in the case of multiple eigenvalues, almost reduced) to the row-cyclic strategy, the quadratic convergence result [28, Theorem 5.2] will hold (following some adaptation in the proof) for the de Rijk strategy.

## 4. The accurate computation of $\hat{V}^{(k)}$.

The relations (2.9) and (2.10), as well as the form of $\hat{V}^{(k)}$ from (2.21) that they imply, can be computed using only correctly rounded arithmetical operations. Along with the basic ones (addition/subtraction, multiplication, division, and the square root), the set of such operations has been recently expanded[2] by the CORE-MATH project [40] to include, among others, floating-point implementations of the functions $\operatorname{rsqrt}(x) = 1/\sqrt{x}$ and $\operatorname{hypot}(x, y) = \sqrt{x^2 + y^2}$.

For the trigonometric case, the relatively accurate computation of $\tan(\theta_k)$, $\cos(\theta_k)$, and $\sin(\theta_k)$, and thus $\hat{V}^{(k)}$, barring inexact underflow of any partial result, has been described in [32]. The hyperbolic case is handled similarly, as discussed in the following, with the assumption that *neither inexact underflow nor overflow* occur.

Since we make estimates for a given step $k$, we can simplify the notation by removing the subscript $k$, as in $\theta_k$, and the superscript $(k)$, as in $a_{ji}^{(k)}$ or $\hat{V}^{(k)}$. As before, we use the notation $\Re(x)$ and $\Im(x)$ for the real and imaginary parts of $x$.

Let $\phi = \arg(a_{ji})$. If $a_{ji} \neq 0$, then $\cos(\phi) = \Re(a_{ji})/|a_{ji}|$ and $\sin(\phi) = \Im(a_{ji})/|a_{ji}|$ (else, $\phi = 0$). The magnitude of $a_{ji}$ is computed with only a single rounding of its exact value by the correctly rounded cr_hypot function as

$$\underline{|a_{ji}|} = \operatorname{cr\_hypot}(\Re(a_{ji}), \Im(a_{ji})) = \sqrt{\Re(a_{ji})^2 + \Im(a_{ji})^2}(1 + \epsilon_1), \qquad |\epsilon_1| \leq \varepsilon, \tag{4.1}$$

with $\varepsilon$ being the machine precision. The underlined expressions here and in the following represent the computed floating-point values of their exact non-underlined counterparts. The elements $a_{ii}$, $a_{jj}$, and $a_{ji}$ are considered exact for the purposes of this section, i.e., $\underline{a_{ii}} = a_{ii}$, $\underline{a_{jj}} = a_{jj}$, and $\underline{a_{ji}} = a_{ji}$.

From (2.9), the value $\tanh(\theta)$ can be obtained as

$$\tanh(\theta) = \frac{\tanh(2\theta)}{1 + \sqrt{1 - \tanh^2(2\theta)}}, \tag{4.2}$$

where the expression under the square root, instead of treating it traditionally as a difference of squares, can be computed with a single rounding as

$$\operatorname{fma}(-\underline{\tanh(2\theta)}, \underline{\tanh(2\theta)}, 1),$$

using the standard operation $\operatorname{fma}(x, y, z) = (x \cdot y + z)(1 + \epsilon_{\mathrm{fma}})$, $|\epsilon_{\mathrm{fma}}| \leq \varepsilon$. Here, $\varepsilon$ is the unit roundoff. A similar simplification of the argument of the square root is applicable to the hyperbolic cosine,

$$\cosh(\theta) = \frac{1}{\sqrt{1 - \tanh^2(\theta)}}, \tag{4.3}$$

while the square root and the ensuing division can be merged into a single operation. Then we have $\operatorname{cr\_rsqrt}(x) = (1 + \epsilon_{\mathrm{rsqrt}})/\sqrt{x}$, with $|\epsilon_{\mathrm{rsqrt}}| \leq \varepsilon$, and

$$\underline{\cosh(\theta)} = \operatorname{cr\_rsqrt}(\operatorname{fma}(-\underline{\tanh(\theta)}, \underline{\tanh(\theta)}, 1)).$$

With $\sinh(\theta) = \tanh(\theta) \cdot \cosh(\theta)$, the computation of $\hat{V}$ is now completed.

The relative errors in $\underline{\tanh(\theta)}$, $\underline{\cosh(\theta)}$, and $\underline{\sinh(\theta)}$ computed in this way can only be bounded by imposing a further assumption on the maximal magnitude of $\tanh(\theta)$ (or

[2] See https://gitlab.inria.fr/core-math for the implementation.

$\tanh(2\theta)$). We assume that $|\tanh(\theta)| \le 4/5$ (or, equivalently, $|\tanh(2\theta)| \le 40/41$), except in the following Lemma 4.1, which bounds the relative error in $\underline{\tanh(2\theta)}$.

LEMMA 4.1. *We have*

$$\underline{\tanh(2\theta)} = \frac{-2|a_{ji}|(1+\epsilon_1)}{(a_{ii}+a_{jj})(1+\epsilon_2)}(1+\epsilon_3) = \tanh(2\theta)(1+\epsilon_d), \tag{4.4}$$

*where* $\max\{|\epsilon_1|, |\epsilon_2|, |\epsilon_3|\} \le \varepsilon$. *The expression* $1+\epsilon_d = (1+\epsilon_1)(1+\epsilon_3)/(1+\epsilon_2)$ *is bounded as*

$$\frac{(1-\varepsilon)^\gamma}{1+\varepsilon} \le 1+\epsilon_d \le \frac{(1+\varepsilon)^\gamma}{1-\varepsilon}, \tag{4.5}$$

*with* $\gamma = 1$ *if* $a_{ji} \in \mathbb{R}$ *and* $\gamma = 2$ *otherwise.*

*Proof.* Since scaling of a floating-point value by a power of two is exact unless inexact underflow or overflow occurs, (4.4) follows directly from (4.1). If $a_{ji} \in \mathbb{R}$, then its absolute value is taken exactly, so $\epsilon_1 = 0$. By minimizing the numerator and maximizing the denominator in the expression for $1+\epsilon_d$, the first inequality in (4.5) is obtained while the second one follows by maximizing the numerator and minimizing the denominator. □

The next step is to bound the relative error in $\underline{\tanh(\theta)}$, what is achieved with the help of the following sequence of four lemmas:

LEMMA 4.2. *We have*

$$1 - (\underline{\tanh(2\theta)})^2 = (1-\tanh^2(2\theta))(1+\epsilon_4),$$

*where, using* $\epsilon_d$ *from* (4.4),

$$|\epsilon_4| \le \frac{1519}{81}|\epsilon_d'|, \qquad \epsilon_d' = (2+\epsilon_d)\epsilon_d. \tag{4.6}$$

*Proof.* Let $y = 1 - \tanh^2(2\theta)$. Then $1 \ge y \ge 1 - (40/41)^2 = 81/1681$. From Lemma 4.1, we obtain

$$1 - (\underline{\tanh(2\theta)})^2 = 1 - \tanh^2(2\theta)(1+\epsilon_d)^2 = y - \tanh^2(2\theta)\epsilon_d',$$

since $(1+\epsilon_d)^2 = 1+\epsilon_d'$, with $\epsilon_d'$ from (4.6).

Using the definition of $y$, we find an $\epsilon_4$ such that

$$y(1+\epsilon_4) = y - \tanh^2(2\theta)\epsilon_d' = y + (y-1)\epsilon_d'.$$

Subtracting $y > 0$ from these equalities gives $y\epsilon_4 = (y-1)\epsilon_d'$. This implies

$$\epsilon_4 = \frac{y-1}{y}\epsilon_d' = -\left(\frac{1}{y} - 1\right)\epsilon_d'.$$

By taking the absolute value of $\epsilon_4$ and the lower bound for $y$, we obtain

$$|\epsilon_4| \le \left(\frac{1681}{81} - 1\right)|\epsilon_d'| = \frac{1600}{81}|\epsilon_d'|,$$

which was to be proven. □

LEMMA 4.3. *We have*

$$\mathrm{sqrt}(\mathrm{fma}(-\underline{\tanh(2\theta)}, \underline{\tanh(2\theta)}, 1)) = \sqrt{1-\tanh^2(2\theta)}(1+\epsilon_7),$$

*where*

$$1+\epsilon_7 = \sqrt{(1+\epsilon_4)(1+\epsilon_5)}(1+\epsilon_6), \qquad \max\{|\epsilon_5|, |\epsilon_6|\} \le \varepsilon.$$

*Proof.* From the definition of fma and Lemma 4.2, it follows that

$$\begin{aligned} \mathrm{fma}(-\underline{\tanh(2\theta)}, \underline{\tanh(2\theta)}, 1) &= (1-(\underline{\tanh(2\theta)})^2)(1+\epsilon_5) \\ &= (1-\tanh^2(2\theta))(1+\epsilon_4)(1+\epsilon_5). \end{aligned} \tag{4.7}$$

Taking the floating-point square root, sqrt, of (4.7) concludes the proof. □

LEMMA 4.4. *Let* $x = 1+\sqrt{1-\tanh^2(2\theta)}$. *Then,*

$$1+\mathrm{sqrt}(\mathrm{fma}(-\underline{\tanh(2\theta)}, \underline{\tanh(2\theta)}, 1)) = x(1+\epsilon_8),$$

*where*

$$\frac{9}{50}|\epsilon_7| \le |\epsilon_8| \le \frac{1}{2}|\epsilon_7|.$$

*Proof.* From Lemma 4.3 it follows that

$$1+\mathrm{sqrt}(\mathrm{fma}(-\underline{\tanh(2\theta)}, \underline{\tanh(2\theta)}, 1)) = 1+(x-1)(1+\epsilon_7) = x+\epsilon_7(x-1).$$

Now, $\epsilon_8$ has to be found such that

$$x(1+\epsilon_8) = x+\epsilon_7(x-1).$$

Subtraction of $x > 0$ from the left- and right-hand sides gives

$$\epsilon_8 = \left(1-\frac{1}{x}\right)\epsilon_7. \tag{4.8}$$

From $1 \ge 1-\tanh^2(2\theta) \ge 81/1681 = 9^2/41^2$, it follows that $50/41 \le x \le 2$. Substituting these bounds for $x$ in (4.8) and taking the absolute value of $\epsilon_8$ completes the proof. □

LEMMA 4.5. *For the denominator in* (4.2) *we have*

$$\underline{1+\sqrt{1-\tanh^2(2\theta)}} = (1+\sqrt{1-\tanh^2(2\theta)})(1+\epsilon_{10}), \qquad \begin{aligned} &1+\epsilon_{10} = (1+\epsilon_8)(1+\epsilon_9), \\ &|\epsilon_9| \le \varepsilon. \end{aligned}$$

*Proof.* The proof follows from Lemma 4.4. The factor $1+\epsilon_9$ comes from rounding the final addition of unity. □

Theorem 4.6 gives the relative error in $\underline{\tanh(\theta)}$.

THEOREM 4.6. *We have*

$$\underline{\tanh(\theta)} = \tanh(\theta)(1+\epsilon_t), \qquad 1+\epsilon_t = \frac{1+\epsilon_d}{1+\epsilon_{10}}(1+\epsilon_{11}), \quad |\epsilon_{11}| \le \varepsilon.$$

*Proof.* The proof is implied by the Lemmas 4.2, 4.3, 4.4, and 4.5. □

Now we can determine the relative errors in $\underline{\cosh(\theta)}$ and $\underline{\sinh(\theta)}$. Recall that we assume $|\tanh(\theta)| \le 4/5$.

LEMMA 4.7. *We have*

$$1-(\underline{\tanh(\theta)})^2=(1-\tanh^2(\theta))(1+\epsilon_{12}),\qquad |\epsilon_{12}|\le\frac{16}{9}|\epsilon_t'|,\quad \epsilon_t'=(2+\epsilon_t)\epsilon_t.$$

*Proof.* This is proven similarly to Lemma 4.2, using $z=1-\tanh^2(\theta)\ge 9/25$ instead of $y$. □

THEOREM 4.8. *We have*

$$\underline{\cosh\theta}=\cosh\theta(1+\epsilon_c),\qquad 1+\epsilon_c=\frac{1+\epsilon_{14}}{\sqrt{(1+\epsilon_{12})(1+\epsilon_{13})}},\qquad \max\{|\epsilon_{13}|,|\epsilon_{14}|\}\le\varepsilon.$$

*Proof.* From (4.3), Lemma 4.7, and the definition of $\mathrm{fma}$, it follows that

$$\begin{aligned}\mathrm{fma}(-\underline{\tanh(\theta)},\underline{\tanh(\theta)},1)&=(1-(\underline{\tanh(\theta)})^2)(1+\epsilon_{13})\\&=(1-\tanh^2(\theta))(1+\epsilon_{12})(1+\epsilon_{13}).\end{aligned}\tag{4.9}$$

Note that $\mathrm{cr\_rsqrt}$ is correctly rounded. Taking $\mathrm{cr\_rsqrt}$ of (4.9) concludes the proof, similarly to Lemma 4.3. □

It remains to bound the relative error in $\underline{\sinh(\theta)}$. In computing $\underline{\sinh(\theta)}$, we use the formula $\sinh(\theta)=\tanh(\theta)\cdot\cosh(\theta)$.

THEOREM 4.9. *We have*

$$\underline{\sinh\theta}=\sinh\theta(1+\epsilon_s),\qquad 1+\epsilon_s=(1+\epsilon_t)(1+\epsilon_c)(1+\epsilon_{15}),\qquad |\epsilon_{15}|\le\varepsilon.$$

*Proof.* The proof is implied by Theorem 4.6 and Theorem 4.8. □

This completes the error analysis if $a_{ji}$ is real. Otherwise, for some $\epsilon_{16}$ and $\epsilon_{17}$ such that $\max\{|\epsilon_{16}|,|\epsilon_{17}|\}\le\varepsilon$, it holds that

$$\begin{aligned}\underline{\Re(e^{\imath\phi})}&=\frac{\Re(a_{ji})(1+\epsilon_{16})}{|a_{ji}|(1+\epsilon_1)}=\Re(e^{\imath\phi})(1+\epsilon_{\Re}'),\qquad 1+\epsilon_{\Re}'=\frac{1+\epsilon_{16}}{1+\epsilon_1},\\ \underline{\Im(e^{\imath\phi})}&=\frac{\Im(a_{ji})(1+\epsilon_{17})}{|a_{ji}|(1+\epsilon_1)}=\Im(e^{\imath\phi})(1+\epsilon_{\Im}'),\qquad 1+\epsilon_{\Im}'=\frac{1+\epsilon_{17}}{1+\epsilon_1}.\end{aligned}\tag{4.10}$$

THEOREM 4.10. *We have*

$$\underline{\Re(e^{\imath\phi})\sinh\theta}=\Re(e^{\imath\phi})\sinh\theta(1+\epsilon_{\Re}),\qquad \underline{\Im(e^{\imath\phi})\sinh\theta}=\Im(e^{\imath\phi})\sinh\theta(1+\epsilon_{\Im}),$$

*where*

$$1+\epsilon_{\Re}=(1+\epsilon_{\Re}')(1+\epsilon_s)(1+\epsilon_{18}),\qquad 1+\epsilon_{\Im}=(1+\epsilon_{\Im}')(1+\epsilon_s)(1+\epsilon_{19}),$$

*with* $\max\{|\epsilon_{18}|,|\epsilon_{19}|\}\le\varepsilon$.

*Proof.* The proof follows from (4.10) and Theorem 4.9. □

Upper bounds for $|\epsilon_d|$, $|\epsilon_t|$, $|\epsilon_c|$, $|\epsilon_s|$, $|\epsilon_{\Re}|$, and $|\epsilon_{\Im}|$ can be obtained using symbolic computation in terms of $\varepsilon$ and $\gamma$ for a set of floating-point datatypes of interest.

Alongside double precision, implementations of $\mathrm{cr\_hypot}$ and $\mathrm{cr\_rsqrt}$ exist in half, single, and quadruple precisions, and for the Intel's 80-bit extended type [8]. The 32-bit, 64-bit, and 80-bit data types, natively supported on the testing hardware, are represented by their precisions $\varepsilon_{32}=2^{-24}$, $\varepsilon_{64}=2^{-53}$, and $\varepsilon_{80}=2^{-64}$, assuming rounding to the nearest.

Table 4.1 provides the upper bounds, ub, for the relative errors in the complex and the real cases, computed using a Wolfram Language script[3] with Wolfram Engine 14.2.1 on Linux and 113 digits of precision and rounded upwards to nine digits after the decimal point on output,

TABLE 4.1
*Precision-dependent upper bounds for the relative errors (with* $\operatorname{ub}|\epsilon_{\Im}| = \operatorname{ub}|\epsilon_{\Re}|$*).*

| | precision ($\mathbb{C}$) | | precision ($\mathbb{R}$) | |
|---|---|---|---|---|
| | single | double, extended | single | double, extended |
| $\operatorname{ub}\lvert\epsilon_d\rvert$ | 3.000000239 | 3.000000001 | 2.000000120 | 2.000000001 |
| $\operatorname{ub}\lvert\epsilon_t\rvert$ | 35.379749082 | 35.379629630 | 24.503140676 | 24.503086420 |
| $\operatorname{ub}\lvert\epsilon_c\rvert$ | 64.397757398 | 64.397119342 | 45.061344394 | 45.061042525 |
| $\operatorname{ub}\lvert\epsilon_s\rvert$ | 100.777648228 | 100.776748972 | 70.564555029 | 70.564128944 |
| $\operatorname{ub}\lvert\epsilon_{\Re}\rvert$ | 103.777666487 | 103.776748972 | $= \operatorname{ub}\lvert\epsilon_s\rvert$ | |

similarly to the method for computing the bounds in the trigonometric case in [32]. All values in Table 4.1 are multiples of the respective machine precisions ($\varepsilon_{32}$, $\varepsilon_{64}$, and $\varepsilon_{80}$).

We note that by using a special (and pretty complicated) rounding error analysis from [29], almost all bounds from Table 4.1 can be further reduced.

Table 4.2 can be consulted for a comparison of the theoretical upper bounds for the relative errors with the maximal relative errors observed for $31 \cdot 2^{30}$ single precision positive definite $2 \times 2$ real and complex matrices defined by pseudorandom values of $a_{ii}$, $a_{ji}$, and $a_{jj}$, sampled from the interval $[0, 1]$. The relative errors, computed in extended precision, expressed in multiples of $\varepsilon_{32}$, and rounded upwards, also contain $|\epsilon_{\mathrm{det}}|$, the maximal observed departure of $\det(\hat{V})$ from unity.

TABLE 4.2
*Maximal observed relative errors for a set of single precision positive definite* $2 \times 2$ *matrices.*

| maximal observed relative errors in multiples of $\varepsilon_{32}$ | | | | | | |
|---|---|---|---|---|---|---|
| $\max\lvert\epsilon_{\mathrm{det}}^{\mathbb{C}}\rvert$ | $\max\lvert\epsilon_c^{\mathbb{C}}\rvert$ | $\max\lvert\epsilon_{\Re}^{\mathbb{C}}\rvert$ | $\max\lvert\epsilon_{\Im}^{\mathbb{C}}\rvert$ | $\max\lvert\epsilon_{\mathrm{det}}^{\mathbb{R}}\rvert$ | $\max\lvert\epsilon_c^{\mathbb{R}}\rvert$ | $\max\lvert\epsilon_s^{\mathbb{R}}\rvert$ |
| 11.96683 | 21.98160 | 33.25813 | 33.99575 | 4.48249 | 14.99693 | 23.56537 |

Let $\nu$ be the largest finite positive floating-point value. If

$$\hat{a} = \max\{|a_{ii}|, |a_{ji}|, |a_{jj}|\} \le \nu/2,$$

then neither the numerator nor the denominator in the expression (2.9) defining $\tanh(2\theta)$ can overflow. As in [32], if the pivot submatrix is scaled by the largest power of two for which it still holds that $\lfloor \lg \hat{a} \rfloor < \lfloor \lg \nu \rfloor$ (where $\lg = \log_2$), then no overflow can occur in the computation of $\hat{V}$, and the possibility of inexact underflows is diminished. Observe that the left-hand side of (4.9) is a normal floating-point value since $\tanh(2\theta)$, and thus also $\tanh\theta$, must be a value strictly less than unity for the computation to proceed. The largest floating-point value below unity is the immediate floating-point predecessor of $1$ (let it be called $1^-$ here), and the rounded value of $1 - (1^-)^2$ is normal, as well as its inverse square root. Even if $\tanh\theta$ is bounded above only by unity (not by $4/5$), the computation of $\cosh\theta$ cannot suffer from overflow.

At present, $\mathrm{cr\_hypot}$ and $\mathrm{cr\_rsqrt}$ in all precisions, as well as $\mathrm{fma}$ in extended, consist of a non-trivial series of processor instructions. It might be argued that the older formulas for

[3] Available at https://github.com/venovako/AccJac/blob/master/var/rejv2.wls.

$\underline{\tanh(\theta)}$ and $\underline{\cosh(\theta)}$, using only the basic operations, as in

$$\begin{aligned} \underline{\tanh\theta} &= \underline{\tanh(2\theta)}/\operatorname{sqrt}((1-\underline{\tanh(2\theta)})\cdot(1+\underline{\tanh(2\theta)})),\\ \underline{\cosh\theta} &= 1/\operatorname{sqrt}((1-\underline{\tanh\theta})\cdot(1+\underline{\tanh\theta})) \end{aligned} \tag{4.11}$$

are faster to evaluate. However, the new formulas are generally more accurate in the worst case than the old ones. This is shown to hold in single precision for all interesting $|\tanh(2\theta)|$ by the following test procedure that would be almost intractable in higher precisions (e.g., double) but is efficient in single.

Let $t_2 = |\underline{\tanh(2\theta)}|$ be considered given (and thus exact) in single precision. Then, $\underline{t} = \underline{\tanh(\theta)}$, $\underline{c} = \underline{\cosh(\theta)}$, and $\underline{s} = \underline{\sinh(\theta)}$ can be computed from $t_2$ using the old formulas, while the results of the new ones are denoted by $\underline{t}'$, $\underline{c}'$, and $\underline{s}'$, respectively. Starting from $t_2$, the computations can be repeated in a higher precision, with the results $t$, $c$, and $s$ representing more accurate approximations of the exact values of $\tanh(\theta)$, $\cosh(\theta)$, and $\sinh(\theta)$, respectively. For this, the new formulas using the MPFR library [12] with 1024 bits of precision are chosen.

All single precision values of $t_2$ that are large enough to make $\underline{t}$ and $\underline{t}'$ different from $t_2/2$ are iterated over. Given the exponent of $t_2$, i.e., $\lfloor \lg t_2 \rfloor$, all significands for that exponent are taken one after the other, unless $\lfloor \lg t_2 \rfloor = -1$, which is when the iteration stops below the cutoff of $40/41$ since otherwise $\underline{\tanh(\theta)}$, $\underline{\cosh(\theta)}$, and $\underline{\sinh(\theta)}$ are to be set to the rounded values of $4/5$, $5/3$, and $4/3$, respectively. Figure 4.1 below displays the results, where $\rho_{\text{old}}(x) = |x-\underline{x}|/|x\varepsilon|$ and $\rho_{\text{new}}(x) = |x-\underline{x}'|/|x\varepsilon|$, for $x \in \{t, c, s\}$.

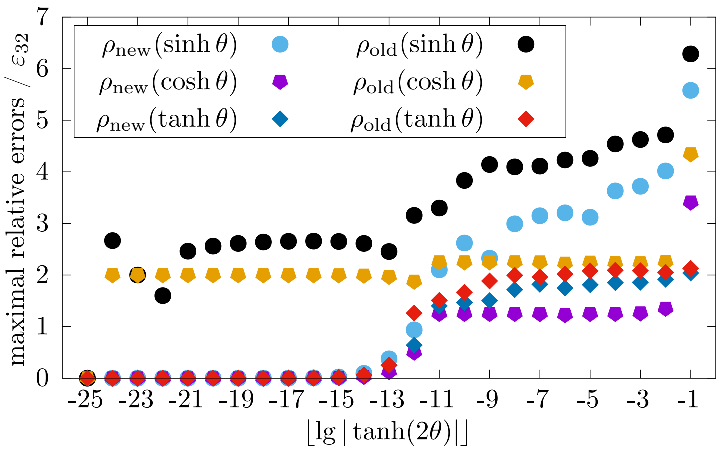

FIG. 4.1. *The maximal relative errors as multiples of* $\varepsilon_{32}$ *in* $\underline{\tanh(\theta)}$, $\underline{\cosh(\theta)}$, *and* $\underline{\sinh(\theta)}$, *as computed from a given* $|\tanh(2\theta)|$, *by the old and the new formulas in single precision.*

The new formulas are clearly more accurate in the worst case than the old ones. Specifically, for a small $|\tanh(2\theta)|$, such that $\varepsilon \le |\tanh(2\theta)| < \sqrt{\varepsilon}$, the differences of squares in (4.11) induce a perturbation of unity that is more inaccurate than the result of the corresponding $\mathrm{fma}$ operations (i.e., the exact unity). In higher precisions, a random sampling of $|\tanh(2\theta)|$ with a given exponent should be used instead of the described procedure to get a comparable picture.

For completeness, (4.12) summarizes the method from [32] for computing $\hat{V}$ in the trigonometric case more accurately than the traditional way as

$$\begin{aligned}\underline{\tan(\theta)} &= \underline{\tan(2\theta)}/(1+\mathrm{cr\_hypot}(1,\underline{\tan(2\theta)})),\\ \underline{\cos(\theta)} &= 1/\,\mathrm{cr\_hypot}(1,\underline{\tan(\theta)}).\end{aligned}\tag{4.12}$$

In certain cases, a possibly faster and/or more accurate computation might be

$$\underline{\cos(\theta)} = \mathrm{cr\_rsqrt}(\mathrm{fma}(\underline{\tan(\theta)},\underline{\tan(\theta)},1)),$$

but (4.12) suggests the most accurate calculation possible for $\underline{\cos(\theta)}$ to be given by

$$\underline{\cos(\theta)} = \mathrm{cr\_rhypot}(\underline{\tan(\theta)},1),$$

where, for a correctly rounded reciprocal hypotenuse function, it should hold that

$$\mathrm{cr\_rhypot}(x,y) = (1+\epsilon_{-1/2})/\sqrt{x^2+y^2},$$

with $|\epsilon_{-1/2}| \le \varepsilon$, if such a function can be implemented performantly in the future.

**5. Numerical experiments.** The numerical experiments were performed using the GNU Fortran and C compilers, version 13.4, on an Intel Xeon Phi 7210 CPU, running at 1.30 GHz. The source code is available at the Github repositories `https://github.com/venovako/AccJac` and `https://github.com/venovako/libpvn`[4]. Since the algorithms under test are sequential by nature, a subset of the tests was repeated on an Intel Core i7-4770K CPU, running at 3.50 GHz, using the GNU compilers, version 15.2, to collect a realistic timing, marked by (B) below. All algorithms have been designed to give unconditionally reproducible outputs in single and double precision on modern architectures, although the relative errors computed from those results in parallel might subtly differ from one system to another, depending on the maturity of quadruple precision support and on the OpenMP implementation.

First we look at the rate of the reduction of $\mathrm{off}(A)$ throughout the $J$-Jacobi process, using the modified de Rijk and the row-cyclic serial strategies, alongside the Mantharam–Eberlein [26] parallel strategy (`ME`), for a matrix $A = G^TG$ with $n = 512$ and $m = 256$, where $G$ has pseudorandom entries in the interval $[0,1]$. Figure 5.1 shows that the modified de Rijk strategy in this (but not necessarily every) case exhibits a faster rate than the row-cyclic one, which is similar to `ME`.

Since the (modified) de Rijk strategy differs from the row-cyclic one in the gradual ordering of $\mathrm{diag}(A)$ such that the diagonals of the diagonal $m\times m$ and $(n-m)\times(n-m)$ blocks are eventually sorted non-increasingly, we introduce the full non-increasing sorting of those diagonals before each cycle. As Figure 5.2 shows, this speeds up the convergence with all three strategies and makes the off-norm reduction rate of the row-cyclic one more similar to that of the modified de Rijk one. Such sorting is beneficial enough to be included in all the following tests. Also, the (modified) de Rijk strategy is applicable to all matrix orders, while for `ME` it is required that $n = 2^l$ for some $l$ (or, for the generalization [30] of `ME`, without a proof of convergence, it has to hold that $n = o\cdot 2^l$ for an odd $o \le 21$).

Next, we shift our attention from the two-sided $J$-Jacobi generalized Hermitian eigensolver to the implicit, one-sided $J$-Jacobi HSVD. Let a Hermitian indefinite non-singular matrix $H$ be given in a factored form, as $GJG^*$, where $G$ is of full column rank, and let its

[4] The tests were performed using the code from the "testing3" tags in both repositories, but the latest development sources are otherwise recommended for a practical application.

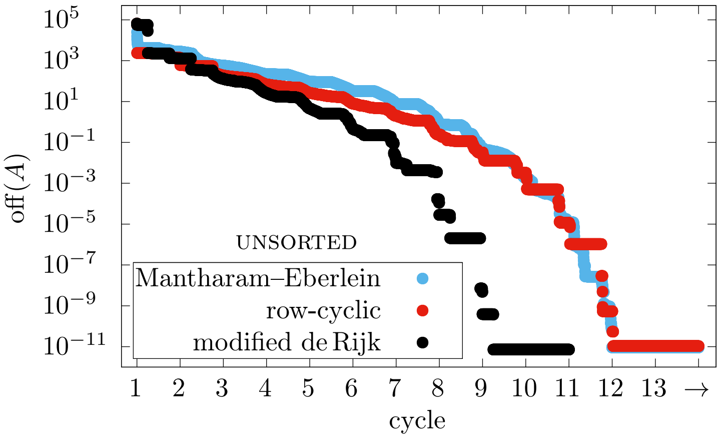

FIG. 5.1. *Reduction of* off$(A)$ *for a real double precision matrix* $A$ *with* $n = 512$ *and* $m = 256$*, throughout the* $J$*-Jacobi process, using three strategies without ordering of the diagonal before each cycle. The first data point is the initial off-norm. In each cycle,* off$(A)$ *was computed after every* $n/2$ *steps and after the first and the last step. The stopping of the longest-running case is denoted by* →.

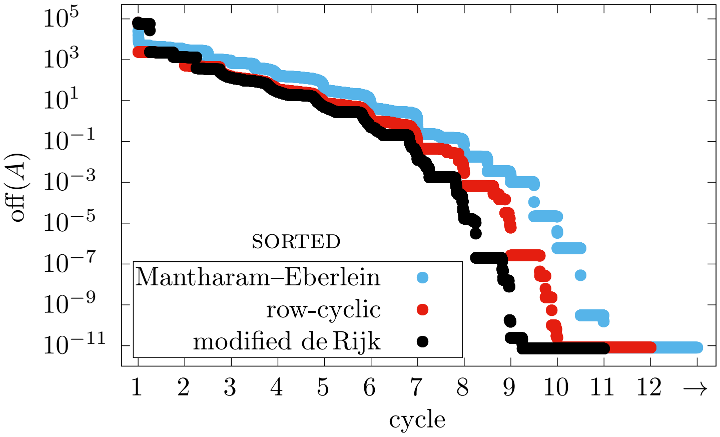

FIG. 5.2. *Reduction of* off$(A)$ *for the same matrix* $A$ *and with the same frequency of the off-norm computation as in Figure 5.1, throughout the* $J$*-Jacobi process, using the same strategies alongside the ordering of the diagonal non-increasingly in the* $m \times m$ *and* $(n-m) \times (n-m)$ *diagonal blocks before each cycle.*

HSVD be $G = U\Sigma V^{-1}$, where $U$ is unitary, $\Sigma$ is diagonal with positive diagonal elements, and $V$ is $J$-unitary (so $V^{-1} = JV^*J$). Then,

$$H = U\Sigma^2 JU^*, \qquad HU = U\Lambda, \qquad \Lambda = \Sigma^2 J,$$

i.e., $\lambda_{ii} = \sigma_{ii}^2 j_{ii}$ are the eigenvalues, while the $u_i$ are the corresponding eigenvectors of $H$, for $1 \leq i \leq n$. Thus, the HSVD of $G$ solves the associated Hermitian indefinite eigenproblem for $H$. The one-sided $J$-Jacobi HSVD mutually orthogonalizes the columns of $G$ by implicitly annihilating the off-diagonal elements of $G^*G$ with respect to $J$, leaving at the point of convergence a matrix $G^{[K]}$ with nearly orthogonal columns that approximates $U\Sigma$. The stopping criterion [10] is taken to be

$$|g_j^* g_i| = |g_i^* g_j| < \|g_i\|_F \|g_j\|_F \cdot \mathfrak{e} \qquad \text{for all } \ (i,j),\ 1 \le i < j \le n, \quad \mathfrak{e} = \varepsilon\sqrt{n}, \tag{5.1}$$

which has also been backported to the two-sided $J$-Jacobi method as

$$|a_{ij}| < \sqrt{a_{ii}}\sqrt{a_{jj}} \cdot \mathfrak{e}, \tag{5.2}$$

i.e., the transformation defined by the pivot pair $(i,j)$ is not applied if (5.1) (respectively, (5.2)) holds for that $(i,j)$. The process is terminated when an empty cycle is reached.

The transformation matrices have the form (2.22), i.e., $\check{V}^{(k)}$ is used. The norms of the columns of $G^{(k)}$ are kept in a vector $\Sigma^{(k)}$ that is sorted as described before each cycle. This sorting induces a permutation of the columns of the iteration matrix, which is not applied, but instead the columns are addressed via a permutation vector. The column swaps in the (modified) de Rijk strategy are realized by updating the same indirect indexing vector. The column norms are updated [10] with $\mathrm{fma}$ as

$$\begin{aligned} \|g_i^{(k+1)}\|_F &= \sqrt{\|g_i^{(k)}\| + \mathfrak{t}_i^{(k)}(|\mathfrak{s}^{(k)}|\|g_j^{(k)}\|_F)}\sqrt{\|g_i^{(k)}\|_F}, \\ \|g_j^{(k+1)}\|_F &= \sqrt{\|g_j^{(k)}\| + \mathfrak{t}_j^{(k)}(|\mathfrak{s}^{(k)}|\|g_i^{(k)}\|_F)}\sqrt{\|g_j^{(k)}\|_F} \end{aligned} \tag{5.3}$$

after each transformation and recomputed anew before the sorting that precedes each cycle using the $\mathrm{cr\_hypot}$ function as described in [33, Algorithm A]. In the trigonometric case, $\mathfrak{t}_i^{(k)} = \tan(\theta_k) = -\mathfrak{t}_j^{(k)}$, while in the hyperbolic one, $\mathfrak{t}_i^{(k)} = \tanh(\theta_k) = \mathfrak{t}_j^{(k)}$ in (5.3), where $\mathfrak{s}^{(k)}$ stands for the scaled [10] dot product $(g_j/\|g_j\|_F)^*(g_i/\|g_i\|_F)$. The elements of $G^{(k)}$ are throughout the process rescaled by a power of two, as needed, to ensure that the subsequent transformation (a hyperbolic one being more dangerous) cannot cause overflow, while simultaneously avoiding underflows when possible, in the spirit of [31]. For the stable $J$-Jacobi method, the maximal growth of any element's magnitude is at most threefold since

$$\begin{aligned} |(x \pm y \cdot e^{\pm\imath\phi}\tanh(\theta))\cosh\theta| &\le (|x| + |y| \cdot |\tanh(\theta)|)\cosh(\theta) \\ &\le \max\{|x|,|y|\}(1 + 4/5) \cdot (5/3) = 3\max\{|x|,|y|\}. \end{aligned}$$

The norms are rescaled if the iteration matrix has been rescaled. To avoid their overflow, no element should have its absolute value larger than $\nu/\sqrt{n}$. Combining these two bounds and adding a measure of safety due to roundings, the elements of the iteration matrix should be kept below $\nu/(3n)$ in magnitude (or, similarly, $\max_{i,j}\{|\Re(g_{ij})|, |\Im(g_{ij})|\} \le \nu/(5n) < \nu/(3\sqrt{2}n)$ to avoid computing the magnitude of complex numbers). No input matrix with finite elements can therefore cause overflow at any point in the process. If the number of rows of $G$ is $\mathbf{n} > n$, then (5.1) and the bounds there have to be adjusted accordingly. Thus, $\Sigma^{[K]}$ approximates the scaled singular values, $g_i^{[K]}/\sigma_{ii}^{[K]}$ approximates $u_i$, and the columns of $V^{[K]}$ approximate the right singular vectors.

Real and complex double precision test matrices of orders $n = 2^l$, $l \le 12$, were prepared as follows. First, $\Lambda^{\{l\}}$ was set as $\lambda_{ii}^{\{l\}} = 1 - (i-1)2/(n-1)$, where $1 \le i \le n$, so $-1 \le \lambda_{ii}^{\{l\}} \le 1$, using the same eigenvalues in the real and in the complex case. Then, Hermitian (symmetric) indefinite matrices $H^{\{l\}} = U^{\{l\}}\Lambda^{\{l\}}U^{\{l\}*}$ were computed in quadruple precision, with pseudorandom unitary (orthogonal) matrices $U^{\{l\}}$ as in LAPACK [2] and factored as $H^{\{l\}} = G^{\{l\}}J^{\{l\}}G^{\{l\}*}$ using the Slapničar's algorithm [43] for the Hermitian (symmetric) indefinite factorization. The factors $G^{\{l\}}$ were rounded to double precision, and $m = n/2$ by construction.

Each test produced the following datapoints: the number of cycles before convergence ($\mathbf{c}_{\mathtt{s}}^{\mathbb{F}}$), the number of rotations performed ($\mathbf{t}_{\mathtt{s}}^{\mathbb{F}}$), the execution's wall-time ($\mathbf{s}_{\mathtt{s}}^{\mathbb{F}}$), and the relative errors (residuals, computed in quadruple precision),

$$
\begin{aligned}
\rho_U^{\mathbb{F}}[\mathtt{s}] &= \|U_{\mathtt{s}}^* U_{\mathtt{s}} - I\|_F,\\
\rho_V^{\mathbb{F}}[\mathtt{s}] &= \|V_{\mathtt{s}}^* J V_{\mathtt{s}} - J\|_F,\\
\rho_G^{\mathbb{F}}[\mathtt{s}] &= \|G - U_{\mathtt{s}} \Sigma_{\mathtt{s}} V_{\mathtt{s}}^{-1}\|_F / \|G\|_F,\\
\rho_\sigma^{\mathbb{F}}[\mathtt{s}] &= \max_{1\le i\le n} (\|G v_i - u_i \sigma_{ii}\|_F / \sigma_{ii}),\\
\rho_\Lambda^{\mathbb{F}}[\mathtt{s}] &= \max_{1\le i\le n} (|\lambda_{ii} - (\Sigma_{\mathtt{s}})_{ii}^2 j_{ii}| / |\lambda_{ii}|),
\end{aligned}
\tag{5.4}
$$

where $\mathbb{F} \in \{\mathbb{C}, \mathbb{R}\}$ and $\mathtt{s} \in \{\mathtt{dR}, \mathtt{rc}\}$, with $\mathtt{dR}$ denoting the modified de Rijk strategy and $\mathtt{rc}$ the row-cyclic strategy. Only $\rho_\Lambda^{\mathbb{F}}$ cannot be easily computed for an arbitrary input when $\Lambda$, unlike here, is not known in advance. We show the residuals mostly in the complex case and only illustrate that they are similar in the real case.

Figure 5.3 demonstrates that the right singular vectors suffer a higher loss of their $J$-unitarity than the left ones of their unitarity, in both the complex and the real case. The choice of strategy does not affect the results much here.

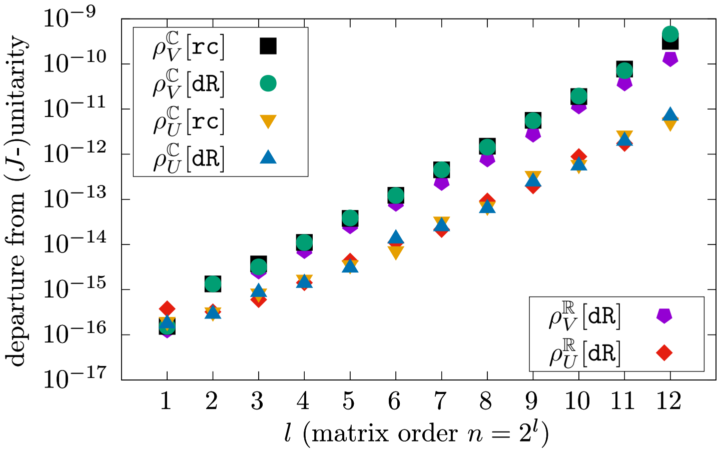

FIG. 5.3. *Departure from ($J$-)unitarity for the test matrices with the notation from* (5.4).

The other residuals are displayed in Figure 5.4. Note that $\rho_\sigma^{\mathbb{C}} > \rho_G^{\mathbb{C}}$ for $l$ large enough, i.e., some singular triplets $(u_i, \sigma_{ii}, v_i)$ might be somewhat inaccurate even if $\rho_G^{\mathbb{C}}$ is still acceptable. This is more evident with the triplets when $i$ approaches $m$ and $n$ from below if the tests are run for, e.g., the Cholesky factor of a large enough Pascal matrix $S$. The singular values are relatively accurate here with respect to $\Lambda$. The modified de Rijk strategy is, overall, slightly more accurate here, according to the considered measures, than the row-cyclic one.

The most prominent advantage of the modified de Rijk strategy over the row-cyclic one for this test set is its speed of convergence as shown in Table 5.1. It is obvious that $\mathbf{t}_{\mathtt{dR}}^{\mathbb{F}} < \mathbf{t}_{\mathtt{rc}}^{\mathbb{F}}$, for both $\mathbb{F}$ when $n$ is sufficiently large. In these rotations, the column transpositions due to the de Rijk's diagonal ordering are not included, but they are of constant complexity as explained in the context of the indirect column addressing, and they do not effectively change the iteration matrix.

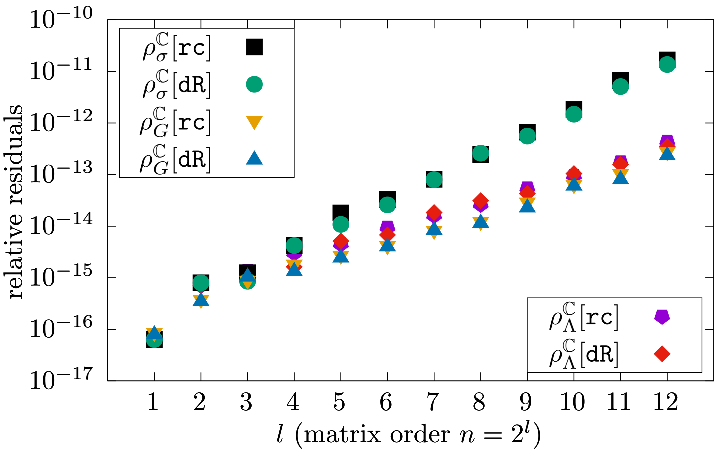

FIG. 5.4. *Various relative residuals for the complex test set with the notation from* (5.4).

TABLE 5.1
*The number of cycles and rotations for the modified de Rijk and the row-cyclic strategies.*

| $n$ | $\mathbf{c}_{\mathtt{dR}}^{\mathbb{C}}$ | $\mathbf{c}_{\mathtt{rc}}^{\mathbb{C}}$ | $\mathbf{t}_{\mathtt{dR}}^{\mathbb{C}}$ | $\mathbf{t}_{\mathtt{rc}}^{\mathbb{C}}$ | $\mathbf{c}_{\mathtt{dR}}^{\mathbb{R}}$ | $\mathbf{c}_{\mathtt{rc}}^{\mathbb{R}}$ | $\mathbf{t}_{\mathtt{dR}}^{\mathbb{R}}$ | $\mathbf{t}_{\mathtt{rc}}^{\mathbb{R}}$ |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| 4 | 5 | 5 | 19 | 19 | 5 | 5 | 20 | 20 |
| 8 | 6 | 6 | 122 | 122 | 6 | 6 | 120 | 120 |
| 16 | 7 | 7 | 608 | 636 | 7 | 7 | 573 | 579 |
| 32 | 8 | 8 | 2856 | 2931 | 8 | 8 | 2755 | 2861 |
| 64 | 10 | 9 | 12938 | 13850 | 9 | 9 | 12482 | 13474 |
| 128 | 11 | 10 | 57246 | 63235 | 10 | 11 | 56251 | 58908 |
| 256 | 12 | 12 | 256006 | 265876 | 12 | 12 | 243525 | 260808 |
| 512 | 13 | 13 | 1088631 | 1137491 | 12 | 13 | 1054765 | 1099070 |
| 1024 | 13 | 14 | 4713456 | 4848262 | 14 | 14 | 4469830 | 4739832 |
| 2048 | 15 | 17 | 19632623 | 20449091 | 15 | 16 | 19099857 | 19691953 |
| 4096 | 18 | 17 | 83016895 | 86164705 | 17 | 16 | 78953756 | 82110484 |

The number of cycles is more erratic since even a single transformation in a later cycle causes another, probably empty, cycle. Yet, when looking at the executions' wall-times, it mostly holds $\mathbf{s}_{\mathtt{dR}}^{\mathbb{F}} < \mathbf{s}_{\mathtt{rc}}^{\mathbb{F}}$ for $n \geq 32$. Table 5.2 provides the ratio of the wall-times as well as the actual wall-times for the modified de Rijk strategy on the machine (B).

The execution times can be improved by turning off the dynamic scaling of the iteration matrix when input matrices are not expected to be badly scaled and by a further manual vectorization [31]. Here, the Intel's FMA and AVX2 instruction sets with 256-bit vectors are requested from the compiler's autovectorizer.

Similarly to the diagonal updates in [37], the updates of any column $i < n$, in a fixed cycle, can be delayed and accumulated for all $(i, j)$, where $j > i$, and applied to the column $i$ after processing the pivot $(i, n)$[5]. Formally, let at the start of each cycle $c_i = 1$, and let $z_i$ be set to a zero vector of the same length as $g_i$. Then, for a pivot pair $(i, j)$ in some step $k$ with

[5] With the (modified) de Rijk and row-cyclic strategies, $(i, n)$ is the last pivot $(i, j)$ in a cycle to be transformed with the given $i$. Otherwise, the last such $j$ should be substituted for $n$ here.

TABLE 5.2
*The execution wall-times and their ratios for the modified de Rijk and the row-cyclic strategies.*

| $n$ | $\mathtt{s}^{\mathbb{C}}_{\mathtt{rc}}/\mathtt{s}^{\mathbb{C}}_{\mathtt{dR}}$ | $\mathtt{s}^{\mathbb{R}}_{\mathtt{rc}}/\mathtt{s}^{\mathbb{R}}_{\mathtt{dR}}$ | $\mathtt{s}^{\mathbb{C}}_{\mathtt{dR}}$ (B) | $\mathtt{s}^{\mathbb{R}}_{\mathtt{dR}}$ (B) |
|---:|---:|---:|---:|---:|
| 32 | 1.01926 | 1.02540 | 0.009 s | 0.008 s |
| 64 | 0.99564 | 1.05020 | 0.035 s | 0.026 s |
| 128 | 1.01857 | 1.06272 | 0.172 s | 0.112 s |
| 256 | 1.03750 | 1.06311 | 1.044 s | 0.605 s |
| 512 | 1.02906 | 1.07713 | 7.093 s | 3.321 s |
| 1024 | 1.04928 | 1.03172 | 55.954 s | 23.121 s |
| 2048 | 1.08360 | 1.05188 | 458.906 s | 185.698 s |
| 4096 | 0.99161 | 0.98110 | 4008.359 s | 1481.609 s |

$\mathfrak{c} = \cos(\theta)$ or $\mathfrak{c} = \cosh(\theta)$,

$$\begin{aligned} x_i &= z_i + c_i \cdot g_i, \\ z_i' &= (z_i + e^{\imath\phi}\mathfrak{t}_i \cdot g_j) \cdot \mathfrak{c}, \\ g_j' &= (g_j + e^{-\imath\phi}\mathfrak{t}_j \cdot x_i) \cdot \mathfrak{c}, \\ c_i' &= c_i \cdot \mathfrak{c}, \end{aligned} \tag{5.5}$$

where the primed quantities are updated in-place, i.e., overwritten, while $\mathfrak{t}_i$ and $\mathfrak{t}_j$ are as in (5.3). The validity of (5.5) can be shown by induction on $j = i + 1, \ldots, n$.

In (5.5), $x_i$ is a temporary vector that holds the up-to-date "view" of $g_i$ as it would have been at the start of the $(i, j)$-step without the delayed updates, while $z_i$ and $c_i$ have to be preserved throughout the cycle. From $\mathfrak{c} \geq 2^{-1/2}$ it follows that $c_i \geq 2^{-(n-i)/2}$, which might lead to underflow if $n$ is large, but the hyperbolic cosines are likely to prevent this in realistic scenarios. Finally, at the end of the $(i, n)$-step, the column $i$ is updated as $g_i' = z_i + c_i \cdot g_i$. Thus, $z_i$ accumulates the updates of $g_i$, each of which might separately affect $g_i$ little to none. On our test set we have not observed a significant improvement with the delayed updates in any accuracy measure, so more targeted test inputs should be generated for that purpose.

Finding the optimal upper bound $t_{\max}$ for the magnitude of the hyperbolic tangents remains an open and vaguely defined problem since a trade-off is involved between increasing the total number of transformations required for convergence and decreasing the departure from $J$-unitarity of the accumulated transformations. For the test matrices of larger orders, even the bound $t_{\max} = 0.8$ is reached, and this happens not in the first but in the third or the fourth cycle. Before that point, the extremal hyperbolic tangents grow slightly in magnitude and fall afterwards.

Another problem with combating the departure from $J$-unitarity of the hyperbolic rotations is that it is not monotone with respect to $|\tanh(\theta)|$, where $\tanh(\theta)$ takes only floating-point values. As a test, the relative error in $\det(\hat{V}(\theta))$ for a real hyperbolic $\hat{V}$ from (2.21), i.e., $|1 - (\cosh^2(\theta) - \sinh^2(\theta))|$, can be computed accurately (e.g., using 1024 bits of precision) from $\cosh(\theta)$ and $\sinh(\theta)$, obtained in single precision for all $\tanh(\theta)$ in a certain interval.

Figure 5.5 shows that the relative error given in multiples of $\varepsilon_{32}$ behaves monotonically for particular subsequences of the single precision values of $|\tanh(\theta)|$ when approaching $t_{\max}$ from below. The maximal attained error of $< 4.5\,\varepsilon_{32}$ is also the maximum when all non-negligible $\tanh\theta$ are considered.

Finally, let us consider the behavior of the HSVD algorithm with the modified de Rijk strategy, when $n$ and $G$ are fixed while $m$ varies. By taking $n = 4096$ and letting $m$ range from 0 to 3584 in increments of 512, we observe and summarize in Table 5.3 that, when $m$

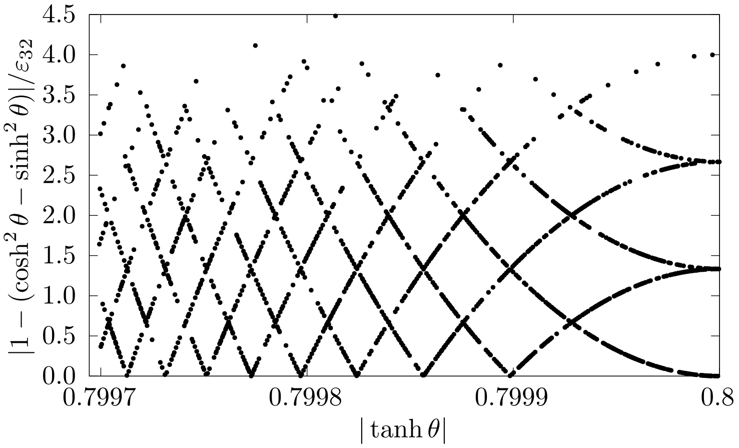

FIG. 5.5. *The relative error in* $\det(\hat{V})$ *for the last* 1352 *single precision values not over* $t_{\max}$.

gets near $n/2$ either from below or from above, i.e., when the total number of hyperbolic transformations to be performed in a cycle grows, it induces a drop in the $J$-orthogonality of the final $V$ and a slower convergence. Similar conclusions can be drawn with the row-cyclic strategy. In this sense, the choice of $m = n/2$ for the rest of the tests is justified.

TABLE 5.3
*The HSVD results obtained with the modified de Rijk strategy,* $n = 4096$, *and G fixed, for various* $m$.

| $m$ | $\mathbf{c}_{\mathtt{dR}}^{\mathbb{C}}$ | $\mathbf{t}_{\mathtt{dR}}^{\mathbb{C}}$ | $\rho_V^{\mathbb{C}}[\mathtt{dR}]$ | $\mathbf{c}_{\mathtt{dR}}^{\mathbb{R}}$ | $\mathbf{t}_{\mathtt{dR}}^{\mathbb{R}}$ | $\rho_V^{\mathbb{R}}[\mathtt{dR}]$ |
|---:|---:|---:|---:|---:|---:|---:|
| 0 | 16 | 74702603 | $3.01958 \cdot 10^{-12}$ | 16 | 72576954 | $2.80047 \cdot 10^{-12}$ |
| 512 | 17 | 79529392 | $6.20928 \cdot 10^{-11}$ | 16 | 75784726 | $2.51998 \cdot 10^{-11}$ |
| 1024 | 18 | 84337990 | $1.14484 \cdot 10^{-10}$ | 16 | 79848476 | $7.18952 \cdot 10^{-11}$ |
| 1536 | 18 | 91804903 | $2.31207 \cdot 10^{-10}$ | 17 | 85805074 | $1.18419 \cdot 10^{-10}$ |
| 2048 | 18 | 83016895 | $4.60046 \cdot 10^{-10}$ | 17 | 78953756 | $1.30368 \cdot 10^{-10}$ |
| 2560 | 18 | 92829473 | $2.11188 \cdot 10^{-10}$ | 19 | 88212777 | $9.28746 \cdot 10^{-11}$ |
| 3072 | 18 | 87312434 | $9.61531 \cdot 10^{-11}$ | 16 | 83928167 | $6.75265 \cdot 10^{-11}$ |
| 3584 | 16 | 79426059 | $6.37419 \cdot 10^{-11}$ | 15 | 77184610 | $2.46849 \cdot 10^{-11}$ |

**6. Conclusions and future work.** In this paper, the global convergence of the $J$-Jacobi method under the de Rijk pivot strategy is proven, and its asymptotic quadratic convergence is discussed. Alongside the improvements of accuracy of the trigonometric and the hyperbolic rotations, this opens the way for applications of the method in real-world scenarios, where the generalized Hermitian eigendecomposition (i.e., the two-sided variant) or the hyperbolic singular value decomposition (i.e., the one-sided variant) is required, since the numerical tests suggest that the (modified) de Rijk strategy has a faster convergence rate than the well-established row-cyclic strategy, with an at least comparable accuracy.

However, similarly to the Jacobi-SVD with the de Rijk strategy from LAPACK (when $J = I$), the proposed elementwise $J$-Jacobi method is inherently sequential, and the possibilities for improving its performance are limited. Thus, for larger problems, a careful blocking and parallelization of the blocked method is essential.

As suggested in Section 5, the departure from $J$-unitarity of the accumulated transformation matrices grows with the number of transformations applied. Therefore, the block size should not be too large, i.e., it should correspond to the size of the lower levels of the cache memory, and a consideration in this context whether to employ the block-oriented [21] or the full block [22] method is warranted.

No serial strategy adequately maps to the massive parallelism of modern accelerators such as GPUs. In the CPU world, however, the de Rijk strategy has its place at the innermost level of blocking. It might be argued that even there, a parallel strategy can be taken for the $J$-Jacobi method since the long vector data types make computing several rotations at the same time possible [31]. This is true for the standard formulas for both the hyperbolic and the trigonometric Jacobi rotations, but the more accurate ones, as described, require the correctly rounded functions that, at the time of writing, operate only on scalars and not yet on SIMD vectors. The serial de Rijk strategy is therefore an excellent choice for the core $J$-Jacobi method in an efficient block-parallel CPU algorithm for the generalized Hermitian EVD or the HSVD, which will be a part of future work.

It is an open problem to see how the de Rijk pivot strategy compares to the classical optimal pivot strategy in the context of the $J$-Hermitian eigenvalue problem. The latter strategy has recently been modified to work with blocks on parallel machines. It is known by the name dynamic ordering (see [3, 4, 35]).

In [17] it has been shown numerically that the de Rijk pivot strategy stabilizes the process and reduces number of iteration steps for the complex HZ method for solving the positive definite generalized eigenvalue problem $Ax = \lambda Bx$ with complex Hermitian matrices $A$ and $B$. This suggests that further research, which includes proving global convergence of the real [16] and complex HZ and CJ methods [17, 18] under the de Rijk pivot strategy, would be needed.

Finally, we note that the modified de Rijk strategy, as described in this paper, can be further refined to obtain a "quasi-cyclic" pivot strategy that makes the $J$-Jacobi method cubically convergent per "quasi-cycle" (cf. [27, 36]). Here, a "quasi-cycle" includes around $1.25N$ trigonometric and hyperbolic transformations and at most $2(n-4)$ transpositions.

**Acknowledgments.** The authors would like to thank the anonymous referees for their remarks that have improved the quality of this manuscript.

## REFERENCES

[1] P. Alonso, J. Delgado, R. Gallego, and J. M. Peña, *Conditioning and accurate computations with Pascal matrices*, J. Comput. Appl. Math., 252 (2013), pp. 21–26.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed. SIAM, Philadelphia, 1999.

[3] M. Bečka, G. Okša, and M. Vajteršic, *Dynamic ordering for a parallel block-Jacobi SVD algorithm*, Parallel Comput., 28 (2002), pp. 243–262.

[4] ———, *New dynamic orderings for the parallel one-sided block-Jacobi SVD algorithm*, Parallel Process. Lett., 25 (2015), Paper No. 1550003, 19 pages.

[5] E. Begović Kovač and V. Hari, *Convergence of the complex block Jacobi methods under the generalized serial pivot strategies*, Linear Algebra Appl., 699 (2024), pp. 421–458.

[6] A. W. Bojanczyk, R. Onn, and A. O. Steinhardt, *Existence of the hyperbolic singular value decomposition*, Linear Algebra Appl., 185 (1993), pp. 21–30.

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

[8] S. Corbineau and P. Zimmermann, *Correct rounding in double extended precision*, in IEEE 32nd Symposium on Computer Arithmetic (ARITH 2025), IEEE Conference Proceedings, Los Alamitos, 2025, pp. 117–124.

[9] P. P. M. de Rijk, *A one-sided Jacobi algorithm for computing the singular value decomposition on a vector computer*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 359–371.

[10] Z. Drmač, *Implementation of Jacobi rotations for accurate singular value computation in floating-point arithmetic*, SIAM J. Sci. Comput, 18 (1997), pp. 1200–1222.

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

[12] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann, *MPFR: a multiple-precision binary floating-point library with correct rounding*, ACM Trans. Math. Softw., 33 (2007), Paper No. 13, 15 pages.

[13] E. R. Hansen, *On cyclic Jacobi methods*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 449–459.

[14] V. Hari, *On the global convergence of the Eberlein method for real matrices*, Numer. Math., 39 (1982), pp. 361–369.

[15] ———, *Convergence to diagonal form of block Jacobi-type methods*, Numer. Math., 129 (2015), pp. 449–481.

[16] ———, *Globally convergent Jacobi methods for positive definite matrix pairs*, Numer. Algorithms, 79 (2018), pp. 221–249.

[17] ———, *On the global convergence of the complex HZ method*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 1291–1310.

[18] ———, *Complex Cholesky-Jacobi algorithm for PGEP*, in International Conference of Numerical Analysis and Applied Mathematics 2018, AIP Conference Proceedings 2116, AIP Publishing, Melville, 2019, Paper No. 450011, 4 pages.

[19] ———, *Global and quadratic convergence of the block Jacobi method for Hermitian matrices under the de Rijk pivot strategy*, Electron. Trans. Numer. Anal., 63 (2025), pp. 83–128.
https://etna.ricam.oeaw.ac.at/vol.63.2025/pp83-128.dir/pp83-128.pdf

[20] V. Hari and E. Begović Kovač, *Convergence of the cyclic and quasi-cyclic block Jacobi methods*, Electron. Trans. Numer. Anal., 46 (2017), pp. 107–147.
https://etna.ricam.oeaw.ac.at/vol.46.2017/pp107-147.dir/pp107-147.pdf

[21] V. Hari, S. Singer, and S. Singer, *Block-oriented $J$-Jacobi Methods for Hermitian matrices*, Linear Algebra Appl., 433 (2010), pp. 1491–1512.

[22] ———, *Full block $J$-Jacobi method for Hermitian matrices*, Linear Algebra Appl., 444 (2014), pp. 1–27.

[23] G. Yu. Kulikov and M. V. Kulikova, *Hyperbolic-SVD-based square-root unscented Kalman filters in continuous-discrete target tracking scenarios*, IEEE Trans. Automat. Control, 67 (2022), pp. 366–373.

[24] M. V. Kulikova *Hyperbolic SVD-based Kalman filtering for Chandrasekhar recursion*, IET Control Theory Appl., 13 (2019), pp. 1543–1553.

[25] F. Luk and H. Park, *On parallel Jacobi orderings*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 18–26.

[26] M. Mantharam and P. J. Eberlein, *Block recursive algorithm to generate Jacobi-sets*, Parallel Comput., 19 (1993), pp. 481-496.

[27] W. Mascarenhas, *On the convergence of the Jacobi method for arbitrary orderings*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1197–1209.

[28] J. Matejaš and V. Hari, *Quadratic convergence estimate of scaled iterates by $J$-symmetric Jacobi method*, Linear Algebra Appl., 417 (2006), pp. 434–465.

[29] ———, *The high relative accuracy of the HZ method*, Appl. Math. Comput., 433 (2022), Paper No. 127358, 21 pages. Supplementary material https://doi.org/10.1016/j.amc.2022.127358.

[30] V. Novaković, *A hierarchically blocked Jacobi SVD algorithm for single and multiple graphics processing units*, SIAM J. Sci. Comput., 37 (2015), pp. C1–C30.

[31] ———, *Vectorization of a thread-parallel Jacobi singular value decomposition method*, SIAM J. Sci. Comput., 45 (2023), pp. C73–C100.

[32] ———, *Accurate complex Jacobi rotations*, J. Comput. Appl. Math., 450 (2024), Paper No. 116003, 6 pages.

[33] ———, *Recursive vectorized computation of the vector p-norm*, Preprint on arXiv, 2025.
https://arxiv.org/abs/2509.06220

[34] V. Novaković and S. Singer, *A Kogbetliantz-type algorithm for the hyperbolic SVD*, Numer. Algorithms, 90 (2022), pp. 523–561.

[35] G. Okša, Y. Yamamoto, and M. Vajteršic, *Convergence to singular triplets in the two-sided block-Jacobi SVD algorithm with dynamic ordering*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 1238–1262.

[36] H. N. Rhee and V. Hari, *On the global and cubic convergence of a quasi-cyclic Jacobi method*, Numer. Math., 66 (1993), pp. 97–122.

[37] H. Rutishauser, *The Jacobi method for real symmetric matrices*, Numer. Math., 9 (1966), pp. 1–10.

[38] D. S Scott and R. C. Ward, *Solving symmetric-definite quadratic $\lambda$-matrix problems without factorization*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 58–67.

[39] G. Shroff and R. Schreiber, *On the convergence of the cyclic Jacobi method for parallel block orderings*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 326–346.

[40] A. Sibidanov, P. Zimmermann, and S. Glondu, *The CORE-MATH project*, in IEEE 29th Symposium on Computer Arithmetic (ARITH 2022), IEEE Conference Proceedings, Los Alamitos, 2022, pp. 26–34.

[41] S. SINGER, S. SINGER, V. NOVAKOVIĆ, D. DAVIDOVIĆ, K. BOKULIĆ, AND A. UŠĆUMLIĆ,, *Three-level parallel $J$-Jacobi algorithms for Hermitian matrices*, Appl. Math. Comput., 218 (2012), pp. 5704–5725.

[42] S. SINGER, E. DI NAPOLI, V. NOVAKOVIĆ, AND G. ČAKLOVIĆ, *The LAPW method with eigendecomposition based on the Hari–Zimmermann generalized hyperbolic SVD*, SIAM J. Sci. Comput., 42 (2020), pp. C265–C293.

[43] I. SLAPNIČAR, *Componentwise analysis of direct factorization of real symmetric and Hermitian matrices*, Linear Algebra Appl., 272 (1998), pp. 227–275.

[44] ———, *Highly accurate symmetric eigenvalue decomposition and hyperbolic SVD*, Linear Algebra Appl., 358 (2003), pp. 387–424.

[45] I. SLAPNIČAR AND K. VESELIĆ, *Perturbations of the eigenprojections of a factorized Hermitian matrix*, Linear Algebra Appl., 218 (1995), pp. 273–280.

[46] ———*A bound for the condition of a hyperbolic eigenvector matrix*, Linear Algebra Appl., 290 (1999), pp. 247–255.

[47] I. SLAPNIČAR AND N. TRUHAR, *Relative perturbation theory for hyperbolic singular value problem*, Linear Algebra Appl., 358 (2003), pp. 367–386.

[48] N. TRUHAR AND I. SLAPNIČAR, *Relative perturbation bound for invariant subspaces of graded indefinite Hermitian matrices*, Linear Algebra Appl., 301 (1999), pp. 171–185.

[49] K. VESELIĆ, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.

[50] K. VESELIĆ AND I. SLAPNIČAR *Floating-point perturbations of Hermitian matrices*, Linear Alg. Appl., 195 (1993), pp. 81–116.