Jacobi's method is more accurate than QR

James Demmel[1]
Courant Institute
New York, NY 10012

Krešimir Veselić
Lehrgebiet Mathematische Physik
Fernuniversität Hagen
D-5800 Hagen, West Germany

Abstract

We show that Jacobi's method (with a proper stopping criterion) computes small eigenvalues of symmetric positive definite matrices with a uniformly better relative accuracy bound than QR, divide and conquer, traditional bisection, or any algorithm which first involves tridiagonalizing the matrix. In fact, modulo an assumption based on extensive numerical tests, we show that Jacobi's method is optimally accurate in the following sense: if the matrix is such that small relative errors in its entries cause small relative errors in its eigenvalues, Jacobi will compute them with nearly this accuracy. In other words, as long as the initial matrix has small relative errors in each component, even using infinite precision will not improve on Jacobi (modulo factors of dimensionality). We also show the eigenvectors are computed more accurately by Jacobi than previously thought possible. We prove similar results for using one-sided Jacobi for the singular value decomposition of a general matrix.

[1]The first author would like to acknowledge the financial support of the NSF via grants DCR-8552474 and ASC-8715728, and the support of DARPA via grant F49620-87-C-0065. Part of this work was done while the first author was visiting the Fernuniversität Hagen, and he acknowledges their support as well. He is also a Presidential Young Investigator.

Contents

1 Introduction

Jacobi's method and QR iteration are two of the most common algorithms for solving eigenvalue and singular value problems. Both are backward stable, and so compute all eigenvalues and singular values with an absolute error bound equal to $p(n)\varepsilon \|H\|_2$, where $p(n)$ is a slowly growing function of the dimension n of the matrix H, ε is the machine precision, and $\|H\|_2$ is the spectral norm of the matrix. Thus, large eigenvalues and singular values (those near $\|H\|_2$) are computed with high relative accuracy, but tiny ones may not have any relative accuracy at all. Indeed, it is easy to find symmetric positive definite matrices where QR returns negative eigenvalues. This error analysis does not distinguish Jacobi and QR, and so one might expect Jacobi to compute tiny values with as little relative accuracy as QR.

In this paper we show that Jacobi (with a proper stopping criterion) computes eigenvalues of positive definite symmetric matrices and singular values of general matrices with a uniformly better relative error bound than QR, or any other method which initially tridiagonalizes (or bidiagonalizes) the matrix. This includes divide and conquer algorithms, traditional bisection, Rayleigh quotient iteration, and so on. We also show that Jacobi computes eigenvectors and singular vectors with better error bounds than QR.

In fact, for the symmetric positive definite eigenproblem, we show that Jacobi is optimally accurate in the following sense. Suppose the initial matrix entries have small relative uncertainties, perhaps from prior computations. The eigenvalues will then themselves have inherent uncertainties, independent of which algorithm is used to compute them. We show that the eigenvalues computed by Jacobi have error bounds which are nearly as small as these inherent uncertainties. In other words, as long as the initial data is slightly uncertain, even using infinite precision cannot improve on Jacobi (modulo factors of n). For the singular value decomposition, we can prove a similar but necessarily somewhat weaker result.

These results depend on new perturbation theorems for eigenvalues and eigenvectors (or singular values and singular vectors) as well as a new error analysis of Jacobi, all of which are stronger than their classical counterparts. They also depend on an empirical observation for which we have overwhelming numerical evidence but somewhat weaker theoretical understanding.

First we discuss the new perturbation theory for eigenvalues, contrasting the standard error bounds with the new ones. Let H be a positive definite symmetric matrix, and δH a small perturbation of H in the sense that $|\delta H_{ij}/H_{ij}| \le \eta/n$ for all i and j. Then $\|\delta H\|_2 \le \eta \|H\|_2$. Let λ_i and λ_i' be the i-th eigenvalues of H and $H + \delta H$, respectively (numbered so that $\lambda_1 \le \cdots \le \lambda_n$). Then the standard perturbation theory [14] says

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \le \frac{\eta \|H\|_2}{\lambda_i} \le \eta \|H\|_2 \cdot \|H^{-1}\|_2 = \eta\kappa(H) \tag{1.1}$$

where $\kappa(H) \equiv \|H\|_2 \cdot \|H^{-1}\|_2$ is the condition number of H. We prove the following stronger result: Write $H = DAD$ where $D = \mathrm{diag}\,(H_{ii}^{1/2})$ and $A_{ii} = 1$. By a theorem of Van der Sluis [16, 6] $\kappa(A)$ is less than n times $\min_{\hat{D}} \kappa(\hat{D} H \hat{D})$, i.e. it nearly minimizes the condition number of H over all possible diagonal scalings. Then we show that:

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq \eta\kappa(A) \tag{1.2}$$

i.e. the error bound $\eta\kappa(H)$ is replaced by $\eta\kappa(A)$. Clearly, it is possible that $\kappa(A) \ll \kappa(H)$ (and it is always true that $\kappa(A) \leq n\kappa(H)$), so the new bound is always at least about as good and can be much better than the old bound.

In the case of the singular values of a general matrix G, we similarly replace the conventional relative error bound $\eta\kappa(G)$ with $\eta\kappa(B)$, where $G = BD$, D chosen diagonal so the columns of B have unit two-norm. This implies $\kappa(B) \leq n^{1/2} \min_{\hat{D}} \kappa(G\hat{D})$, and as before it is possible that $\kappa(B) \ll \kappa(G)$.

The effects of rounding errors in Jacobi are bounded as follows. We can weaken the assumption of small componentwise relative error $|\delta H_{ij}/H_{ij}| \leq \eta/n$ in the perturbation theory to $|\delta H_{ij}|/(H_{ii}H_{jj})^{1/2} \leq \eta/n$ without weakening bound (1.2). This more general perturbation bounds the rounding errors introduced by applying *one* Jacobi rotation, so that one Jacobi rotation causes relative errors in the eigenvalues bounded by $O(\varepsilon)\kappa(A)$. (In contrast, QR, or any algorithm which first tridiagonalizes the matrix, only computes eigenvalues with relative error bound $O(\varepsilon)\kappa(H)$.)

To bound the errors from *all* the Jacobi rotations we proceed as follows: Let $H_0 = D_0 A_0 D_0$ be the original matrix, and $H_m = D_m A_m D_m$ where H_m is obtained from H_{m-1} by applying a single Jacobi rotation, D_m is diagonal, and A_m has unit diagonal. The desired error bound is proportional to $\kappa(A_0)$, i.e. it depends only on the original matrix. But our analysis only says that at step m we get an error bounded by something proportional to $\kappa(A_m)$. Thus the error bound for all the Jacobi steps is proportional to $\max_m \kappa(A_m)$. So for Jacobi to attain optimal accuracy, $\max_m \kappa(A_m)/\kappa(A_0)$ must be modest in size. In extensive random numerical tests, its maximum value was less than 1.82. Wang [21] has recently found isolated examples where it is almost 8. Our theoretical understanding of this behavior is incomplete and providing it remains an open problem.

We must finally bound the errors introduced by Jacobi's stopping criterion. To achieve accuracy proportional to $\kappa(A)$, we have had to modify the standard stopping criterion. Our modified stopping criterion has been suggested before [20, 5, 3, 19], but without our explanation of its benefits. The standard stopping criterion may be written

$$\text{if } |H_{ij}| \leq tol \cdot \max_{kl} |H_{kl}|, \text{ set } H_{ij} = 0$$

whereas the new one is

$$\text{if } |H_{ij}| \leq tol \cdot (H_{ii}H_{jj})^{1/2}, \text{ set } H_{ij} = 0$$

(here *tol* is a small threshold value, usually machine precision).

Now we consider the eigenvectors and singular vectors. Here and throughout the paper whenever we refer to an eigenvector we assume its eigenvalue is simple. Again let H be a positive definite symmetric matrix with eigenvalues λ_i and unit eigenvectors v_i. Let δH be a small componentwise relative perturbation as before, and let λ_i' and v_i' be the eigenvalues and eigenvectors of $H + \delta H$. Then the standard perturbation theory [14] says that v_i' can be chosen such that

$$\|v_i - v_i'\| \leq \frac{\eta}{absgap_{\lambda_i}} + O(\eta^2) \tag{1.3}$$

where the *absolute gap for eigenvalues* is defined as

$$absgap_{\lambda_i} \equiv \min_{j \neq i} \frac{|\lambda_i - \lambda_j|}{\|H\|_2} \tag{1.4}$$

We prove a generally stronger result which replaces this bound with

$$\|v_i - v_i'\| \leq \frac{(n-1)^{1/2}\kappa(A) \cdot \eta}{relgap_{\lambda_i}} + O(\eta^2) \tag{1.5}$$

where the *relative gap for eigenvalues* is defined as

$$relgap_{\lambda_i} \equiv \min_{j \neq i} \frac{|\lambda_i - \lambda_j|}{|\lambda_i \cdot \lambda_j|^{1/2}} \tag{1.6}$$

The point is that if H has two or more tiny eigenvalues, their absolute gaps are necessarily small, but their relative gaps may be large, so that the corresponding eigenvectors are really well-conditioned. We prove an analogous perturbation theorem for singular vectors of general matrices. We also prove a perturbation theorem which shows that even tiny components of eigenvectors and singular vectors may be well-conditioned. Again, we show Jacobi is capable of computing the eigenvectors and singular vectors to their inherent accuracies, but QR is not.

To illustrate, consider the symmetric positive definite matrix $H = DAD$ where

$$H = \begin{bmatrix} 10^{40} & 10^{29} & 10^{19} \\ 10^{29} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & .1 \\ .1 & .1 & 1 \end{bmatrix} \quad \text{and} \quad D = \text{diag}\,(10^{20}, 10^{10}, 1)$$

Here $\kappa(H) \approx 10^{40}$ and $\kappa(A) \approx 1.33$. Thus, η relative perturbations in the matrix entries only cause 4η relative perturbations in the eigenvalues according to the new theorem, and $3 \cdot 10^{40} \cdot \eta$ relative perturbations according to the conventional theorem. Also, the absolute gaps for the eigenvalues of H are $absgap_{\lambda 1,2,3} \approx 10^{-20}, 10^{-20}, 1$, whereas the relative gaps $relgap_{\lambda 1,2,3}$ are all approximately 10^{10}. Thus the new theory predicts errors in v_1 and v_2 of norm $2 \cdot 10^{-10}\eta$, whereas the old theory predicts errors of $10^{20}\eta$. Jacobi will attain these new error bounds, but in general QR will not. For this example, QR computes two out of the three eigenvalues as negative, whereas H is positive definite. In contrast, Jacobi computes all the eigenvalues to nearly full machine precision. In fact for this example we can show Jacobi computes all components of all eigenvectors to nearly full relative accuracy, even though they vary by 21 orders of magnitude; again QR will not even get the signs of many small components correct.

One might object to this example on the grounds that by reversing the order of the rows and columns before tridiagonalizing and applying QR, one computes the correct eigenvalues. However, one can easily find similar matrices (see section 7) where Jacobi gets accurate eigenvalues and QR gets at least one zero or negative eigenvalue no matter how the rows and columns are ordered.

We also show that bisection and inverse iteration (with appropriate pivoting, and applied to the original positive definite symmetric matrix) are capable of attaining the same error bounds as Jacobi. Of course bisection and inverse iteration on a dense matrix are not competitive in speed with Jacobi, unless only one or a few eigenvalues are desired and good starting guesses are available. We use these methods to verify our numerical tests.

This work is an extension of work in [2], where analogous results were proven for matrices which are called *scaled diagonally dominant* (s.d.d.). The positive definite matrix $H = DAD$ is s.d.d. if $\|A - I\|_2 < 1$. This work replaces the assumption that A is diagonally dominant with mere positive definiteness, extending the results of [2] to all positive definite symmetric matrices, as well as to the singular value decomposition of general matrices.

This work does not contradict the results of [7, 2] where it was shown how a variation of QR could compute the singular values of a bidiagonal matrix or the eigenvalues of a symmetric positive definite tridiagonal matrix with high relative accuracy. This is because reducing a dense matrix to bidiagonal or tridiagonal form can cause large relative errors in its singular values or eigenvalues independent of the accuracy of the subsequent processing. In contrast, the results in this paper are for dense matrices.

We also discuss an accelerated version of Jacobi for the symmetric positive definite eigenproblem with an attractive speedup property: The more its accuracy exceeds that attainable by QR or other traditional methods, the faster it converges. See also [20] where earlier references for Jacobi methods on positive definite matrices as well as for one sided methods can be found.

We use the following terminology to distinguish among different versions of Jacobi. "Two-sided Jacobi" refers to the original method applying Jacobi rotations to the left and right of a symmetric matrix. "One-sided Jacobi" refers to computing the SVD by applying Jacobi rotation from one side only.

The rest of this paper is organized as follows. Section 2 presents the new perturbations theorems. Section 3 discusses two-sided Jacobi for the symmetric positive definite eigenproblem. Section 4 discusses one-sided Jacobi for the singular value decomposition. It also presents the accelerated version of Jacobi just mentioned. Section 5 discusses bisection and inverse iteration. Section 6 discusses bounds on $\max_m \kappa(A_m)/\kappa(A_0)$. Section 7 contains numerical experiments. Section 8 presents our conclusions and discussion of open problems.

2 Perturbation Theory

In this section we prove new perturbation theorems for eigenvalues and eigenvectors of symmetric positive definite matrices, and for singular values and singular vectors of general matrices. In the first subsection we consider eigendecompositions of symmetric positive definite matrices. In the second subsection we discuss the optimality of these bounds. In the third subsection we consider the singular value decomposition of general matrices. In the fourth subsection, we discuss the optimality of this second set of bounds.

2.1 Symmetric Positive Definite Matrices

The next two lemmas have also been proved in [2]:

Lemma 2.1 *Let H and K be symmetric matrices with K positive definite. Let the pencil $H - \lambda K$ have eigenvalues λ_i. Let δH and δK be symmetric perturbations and let λ_i' be the (properly ordered) eigenvalues of $(H + \delta H) - \lambda(K + \delta K)$. Suppose that*

$$|x^T \delta H x| \le \eta_H \cdot |x^T H x| \quad \text{and} \quad |x^T \delta K x| \le \eta_K \cdot |x^T K x|$$

for all vectors x and some $\eta_H < 1$ and $\eta_K < 1$. Then either $\lambda_i = \lambda_i' = 0$ or

$$\frac{1 - \eta_H}{1 + \eta_K} \le \frac{\lambda_i'}{\lambda_i} \le \frac{1 + \eta_H}{1 - \eta_K}$$

for all i.

Proof. The condition on δK implies that $K + \delta K$ is positive definite too, so the perturbed pencil is definite. The condition on δH implies that it and H have the same null space, implying that $\lambda_i = \lambda_i' = 0$ if one of them equals zero. Now we consider the case $\lambda_i > 0$; for the other eigenvalues consider $-H$ and $-\delta H$. The Courant-Fischer minimax theorem [14] expresses λ_i as

$$\lambda_i = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{x^T H x}{x^T K x}$$

where the minimum is over all i-dimensional subspaces $\mathbf{S}^i$. Let the spaces $\mathbf{S}_0^i$ and $\mathbf{S}_1^i$ satisfy

$$\lambda_i = \max_{x \in \mathbf{S}_0^i} \frac{x^T H x}{x^T K x} \quad \text{and} \quad \lambda_i' = \max_{x \in \mathbf{S}_1^i} \frac{x^T (H + \delta H) x}{x^T (K + \delta K) x}$$

Then

$$\lambda_i' = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{x^T (H + \delta H) x}{x^T (K + \delta K) x} \le \max_{x \in \mathbf{S}_0^i} \frac{x^T (H + \delta H) x}{x^T H x} \cdot \frac{x^T K x}{x^T (K + \delta K) x} \cdot \frac{x^T H x}{x^T K x} \le \frac{1 + \eta_H}{1 - \eta_K} \lambda_i$$

and similarly

$$\lambda_i = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{x^T H x}{x^T K x} \le \max_{x \in \mathbf{S}_1^i} \frac{x^T H x}{x^T (H + \delta H) x} \cdot \frac{x^T (K + \delta K) x}{x^T K x} \cdot \frac{x^T (H + \delta H) x}{x^T (K + \delta K) x} \le \frac{1 + \eta_K}{(1 - \eta_H)} \lambda_i'$$

completing the proof. ∎

Lemma 2.2 *Let $H = \Delta_H^T A_H \Delta_H$ and A_H be symmetric matrices. H and A_H need not have the same dimensions, and Δ_H may be an arbitrary full rank conforming matrix. Similarly, let $K = \Delta_K^T A_K \Delta_K$ and A_K be symmetric positive definite matrices, where K and A_K need not have the same dimensions and Δ_K may be an arbitrary full rank conforming matrix. Let $\delta H = \Delta_H^T \delta A_H \Delta_H$ be a perturbation of H such that $|x^T \delta A_H x| \leq \eta_H |x^T A_H x|$ for all x where $\eta_H < 1$. Similarly, let $\delta K = \Delta_K^T \delta A_K \Delta_K$ be a perturbation of K such that $|x^T \delta A_K x| \leq \eta_K |x^T A_K x|$ for all x where $\eta_K < 1$. Let λ_i be the i-th eigenvalue of $H - \lambda K$ and λ_i' the i-th eigenvalue of $(H + \delta H) - \lambda(K + \delta K)$. Then either $\lambda_i = \lambda_i' = 0$ or*

$$\frac{1-\eta_H}{1+\eta_K} \leq \frac{\lambda_i'}{\lambda_i} \leq \frac{1+\eta_H}{1-\eta_K}$$

PROOF. Note that for all vectors x

$$|x^T \delta H x| = |x^T \Delta_H^T \delta A_H \Delta_H x| \leq \eta_H |x^T \Delta_H^T A_H \Delta_H x| = \eta_H |x^T H x|$$

and

$$|x^T \delta K x| = |x^T \Delta_K^T \delta A_K \Delta_K x| \leq \eta_K |x^T \Delta_K^T A_K \Delta_K x| = \eta_K |x^T K x|$$

Now apply Lemma 2.1. ∎

Theorem 2.3 *Let $H = DAD$ be a symmetric positive definite matrix, and $D = \mathrm{diag}\,(H_{ii}^{1/2})$ so $A_{ii} = 1$. Let $\delta H = D \delta A D$ be a perturbation such that $\|\delta A\|_2 \equiv \eta < \lambda_{\min}(A)$. Let λ_i be the i-th eigenvalue of H and λ_i' be the i-th eigenvalue of $H + \delta H$. Then*

$$\left|\frac{\lambda_i - \lambda_i'}{\lambda_i}\right| \leq \frac{\eta}{\lambda_{\min}(A)} \leq \kappa(A) \cdot \eta \tag{2.4}$$

In particular, if $|\delta H_{ij}/H_{ij}| \leq \eta/n$, then $\|\delta A\|_2 \leq \eta$ and the bound (2.4) applies.

PROOF. Note that for all nonzero vectors x

$$\left|\frac{x^T \delta H x}{x^T H x}\right| = \left|\frac{x^T \Delta^T \delta A \Delta x}{x^T \Delta^T A \Delta x}\right| = \left|\frac{y^T \delta A y}{y^T A y}\right| \leq \frac{\eta}{\lambda_{\min}(A)}$$

Lemma 2.2 yields the desired bound, using $K = I$ and $\delta K = 0$. It remains to prove that $|\delta H_{ij}/H_{ij}| \leq \eta/n$ implies $\|\delta A\|_2 \leq \eta$. But $A_{ii} = 1$ and A positive definite imply that no entry of A is larger than 1 in absolute value. (Note that this means $\kappa(A)$ is at most n times larger than $1/\lambda_{\min}(A)$.) Therefore $|\delta A_{ij}| = |\delta H_{ij}/H_{ij} \cdot A_{ij}| \leq \eta/n$ and so $\|\delta A\|_2 \leq \eta$ as desired. ∎

Proposition 2.13 in the next subsection shows that the bound of Theorem 2.3 is nearly attained for at least one eigenvalue. However, other eigenvalues may be much less sensitive than this most sensitive one. The next proposition provides individual eigenvalue bounds which may be much tighter:

Proposition 2.5 *Let $H = DAD$ be as in Theorem 2.3, with eigenvalues λ_i and unit eigenvectors v_i. Let $H + \delta H = D(A + \delta A)D$ have eigenvalues λ_i'. Let $\|\delta A\|_2 \equiv \eta \ll \lambda_{\min}(A)$. Then the bound*

$$\frac{|\lambda_i - \lambda_i'|}{\lambda_i} \leq \frac{\eta \|D v_i\|_2^2}{\lambda_i} + O(\eta^2) \tag{2.6}$$

is attainable by the diagonal perturbation $\delta A_{jj} = \eta \,\mathrm{sign}(v_i(j))$.

PROOF. Bound (2.6) is derived from the standard first order perturbation theory which says $\lambda_i(H+\delta H) = \lambda_i(H)+v_i^T\delta H v_i+O(\|\delta H\|_2^2)$, and substituting $|v_i^T\delta H v_i| = |v_i^T D\delta A D v_i| \le \|Dv_i\|_2^2\|\delta A\|_2$. The inequality $|v_i^T D\delta A D v_i| \le \|Dv_i\|_2^2\|\delta A\|_2$ is clearly attained for the diagonal choice of δA in the statement of the proposition. ∎

We stated Lemmas 2.1 and 2.2 separately in order to emphasize their generality. For example, suppose H and K are finite element matrices, Δ_H and Δ_K the fixed assembly matrices of 0's and 1's, and A_H and A_K the block diagonal matrices of individual elements. If δA_H and δA_K are also block diagonal, perturbing each separate element in the sense of $x^T\delta A_H x/x^T A_H x$ and $x^T\delta A_K x/x^T A_K x$ being small for all nonzero x, then the relative perturbations of the eigenvalues of the assembled matrix H will also be small. In particular, consider the matrix arising from modeling a series of masses $m_1,\ldots,m_n$ on a line connected by simple linear springs with spring constants $k_0,\ldots,k_n$ (the ends of the extreme springs are fixed). The natural frequencies of vibration of this system are the square roots of the eigenvalues of the pencil $M-\lambda K$ where $M = \text{diag}\,(m_1,\ldots,m_n)$ and K is tridiagonal with diagonal $k_0+k_1, k_1+k_2,\ldots,k_{n-1}+k_n$ and offdiagonal $-k_1,\ldots,-k_{n-1}$. Lemma 2.2 implies that an η relative perturbation in any spring constant or mass makes at most η relative changes in the eigenvalues of $M-\lambda K$. Unfortunately, this property does not extend to the matrix assembled in floating point, since if $k_{i-1} \ll k_i \gg k_{i+1}$ so that $k_i + k_{i\pm 1}$ rounds to k_i, the computed K will be indefinite instead of positive definite, meaning that some eigenvalues are completely changed.

One may also prove a version of Lemma 2.1 in an infinite dimensional setting [12, section VI.3].

Now we turn to eigenvectors. A weaker version of the following theorem also appeared in [2]:

Theorem 2.7 *Let $H = DAD$ be as in Theorem 2.3. Define $H(\epsilon) = D(A+\epsilon E)D$, where E is any matrix with unit two-norm. Let $\lambda_i(\epsilon)$ be the i-th eigenvalue of $H(\epsilon)$, and assume $\lambda_i(0)$ is simple so that the corresponding unit eigenvector $v_i(\epsilon)$ is well defined for sufficiently small ϵ. Then*

$$\|v_i(\epsilon)-v_i(0)\|_2 \le \frac{(n-1)^{1/2}\epsilon}{\lambda_{\min}(A)\cdot relgap_{\lambda_i}} + O(\epsilon^2) \le \frac{(n-1)^{1/2}\kappa(A)\epsilon}{relgap_{\lambda_i}} + O(\epsilon^2)$$

PROOF. Let $v_k(0)$ be abbreviated by v_k. From [9] we have

$$v_i(\epsilon) = v_i + \epsilon\sum_{k\ne i}\frac{v_k^T DEDv_i}{\lambda_i-\lambda_k}\cdot v_k + O(\epsilon^2)$$

Let $y_k = Dv_k$, so that

$$v_i(\epsilon) = v_i + \epsilon\sum_{k\ne i}\frac{y_k^T E y_i}{\lambda_i-\lambda_k}\cdot v_k + O(\epsilon^2) \tag{2.8}$$

The pair (λ_i, y_i) is an eigenpair of the pencil $A-\lambda D^{-2}$. Thus

$$\lambda_k = \lambda_k y_k^T D^{-2} y_k = y_k^T A y_k \ge \lambda_{\min}(A)\|y_k\|_2^2$$

and so $\|y_k\|_2 \le (\lambda_k/\lambda_{\min}(A))^{1/2}$. Letting $z_k = y_k/\|y_k\|_2$ lets us write

$$v_i(\epsilon) = v_i + \epsilon \sum_{k \neq i} \frac{\xi_{ik} \cdot z_k^T E z_i}{(\lambda_i - \lambda_k)/(\lambda_k \lambda_i)^{1/2}} \cdot v_k + O(\epsilon^2)$$

where $|\xi_{ik}| = \|y_k\|_2 \|y_i\|_2/(\lambda_k\lambda_i)^{1/2} \le 1/\lambda_{\min}(A)$. Taking norms yields the result. ∎

Proposition 2.14 in the next subsection will show that the bound in Theorem 2.7 is nearly attainable for all v_i.

As in Corollary 3 in [2], it is possible to derive a nonasymptotic result from Theorem 2.7:

Corollary 2.9 *Let $H = DAD$ be as in Theorem 2.3. Suppose $\delta \equiv \|\delta A\|_2/\lambda_{\min}(A)$ satisfies*

$$\delta < \frac{1}{4} \quad \text{and} \quad \frac{3 \cdot 2^{-1/2} \cdot \delta}{1-\delta} < relgap_{\lambda_i}$$

Let v_i be the i-th unit eigenvector of $H = DAD$. Then the i-th unit eigenvector v_i' of $H' = D(A + \delta A)D$ can be chosen so that

$$\|v_i - v_i'\|_2 \le \frac{(n-1)^{1/2}\delta}{(1-4\delta)((1-\delta) relgap_{\lambda_i} - 3 \cdot 2^{-1/2}\delta)}$$

Proof. Let $H(\epsilon) = D(A + \epsilon \cdot \delta A/\|\delta A\|_2)D$. Let $\lambda_i(\epsilon)$ be the i-th eigenvalue of $H(\epsilon)$, and abbreviate $\lambda_i(0)$ by λ_i. Let $relgap_{\lambda_i}(\epsilon)$ denote the relative gap of the i-th eigenvalue of $H(\epsilon)$, and $relgap_\lambda(a,b) \equiv |a-b|/(ab)^{1/2}$. The idea is that if ϵ is small, then $\lambda_i(\epsilon)$ can only change by a small relative amount, and so $relgap_{\lambda_i}(\epsilon)$ can only change by a small absolute or relative amount. Note that $\lambda_{\min}(A)$ can decrease by as much as $\|\delta A\|_2$. Then by Theorem 2.3 we can bound $relgap_{\lambda_i}(\epsilon)$ below by

$$\begin{aligned} relgap_{\lambda_i}(\epsilon) &= \min_{k \neq i} \frac{|\lambda_i(\epsilon) - \lambda_k(\epsilon)|}{(\lambda_i(\epsilon)\lambda_k(\epsilon))^{1/2}} \ge \min_{k \neq i} \frac{|\lambda_i - \lambda_k| - \delta(1-\delta)^{-1}(\lambda_i + \lambda_k)}{(\lambda_i\lambda_k)^{1/2}(1 + \delta(1-\delta)^{-1})} \\ &\ge (1-\delta) \min_{k \neq i} \left(relgap_\lambda(\lambda_i, \lambda_k) - \frac{\delta}{1-\delta} \cdot \frac{\lambda_i + \lambda_k}{(\lambda_i\lambda_k)^{1/2}} \right) \end{aligned}$$

We consider two cases, $relgap_\lambda(\lambda_i, \lambda_k) \ge 2^{-1/2}$ and $relgap_\lambda(\lambda_i, \lambda_k) < 2^{-1/2}$. The first case corresponds to λ_i and λ_k differing by at least a factor of 2, whence

$$\frac{\lambda_i + \lambda_k}{(\lambda_i\lambda_k)^{1/2}} \le 3 \cdot relgap_\lambda(\lambda_i, \lambda_k)$$

The second case corresponds to λ_i and λ_k differing by at most a factor of 2, whence

$$\frac{\lambda_i + \lambda_k}{(\lambda_i\lambda_k)^{1/2}} \le 3 \cdot 2^{-1/2}$$

Altogether we have

$$relgap_{\lambda_i}(\epsilon) \ge (1-\delta)\left(1 - \frac{3\delta}{1-\delta}\right)\left(relgap_{\lambda_i} - \frac{3 \cdot 2^{-1/2} \cdot \delta}{1-\delta}\right)$$

Now integrate the bound of Theorem 2.7 from $\epsilon = 0$ to $\epsilon = \|\delta A\|_2$ to get the desired result. ∎

In complete analogy to [2], we may also prove

Proposition 2.10 *Let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of H and $h_1 \le \cdots \le h_n$ be its diagonal entries in increasing order. Then*

$$\lambda_{\min}(A) \le \frac{\lambda_i}{h_i} \le \lambda_{\max}(A)$$

In other words, the diagonal entries of H can differ from the eigenvalues only by factors bounded by $\kappa(A)$.

PROOF. See the proof of Proposition 2 in [2].

Proposition 2.11 *Let $H = DAD$ with eigenvalues λ_i. Let d_i be the diagonal entries of D. Let v_i be the i-th eigenvector of H normalized so that its i-th component $v_i(i) = 1$. Then*

$$|v_i(j)| \le \overline{v}_i(j) \equiv (\kappa(A))^{3/2} \cdot \min(\left(\frac{\lambda_i}{\lambda_j}\right)^{1/2}, \left(\frac{\lambda_j}{\lambda_i}\right)^{1/2})$$

We also have

$$|v_i(j)| \le (\kappa(A))^{3/2} \cdot \min(\frac{d_i}{d_j}, \frac{d_j}{d_i})$$

In other words, the eigenvectors are scaled analogously to the diagonal of H.

PROOF. See the proof of Proposition 6 in [2].

Proposition 2.12 *Let $H(\epsilon)$ and $v_i(\epsilon)$ be as in Theorem 2.7, and $\overline{v}_i(j)$ be as in Proposition 2.11. Then*

$$|v_i(\epsilon)(j) - v_i(0)(j)| \le \frac{(2n-2)^{1/2}}{\lambda_{\min}(A) \cdot \min(relgap_{\lambda_i}, 2^{-1/2})} \cdot \epsilon \cdot \overline{v}_i(j) + O(\epsilon^2)$$

In other words, each component of each eigenvector is perturbed by a small amount relative to its upper bound of $\overline{v}_i(j)$ of Proposition 2.11. Thus small components of eigenvectors may be determined with as much relative accuracy as large components. Note that $relgap_{\lambda_i}$ exceeds $2^{-1/2}$ only when λ_i differs from its nearest neighbor by at least a factor of 2.

PROOF. See the proof of Theorem 7 in [2].

We illustrate these results with two examples. First we consider the matrix $H = DAD$ of the introduction:

$$H = \begin{bmatrix} 10^{40} & 10^{29} & 10^{19} \\ 10^{29} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & .1 \\ .1 & .1 & 1 \end{bmatrix} \quad \text{and} \quad D = \text{diag}\,(10^{20}, 10^{10}, 1)$$

To six correct figures, H's eigenvalue matrix Λ and eigenvector matrix V (normalized to have the largest entry of each eigenvector equal to 1) are

$$\Lambda = \text{diag}\,(1.00000 \cdot 10^{40}, 9.90000 \cdot 10^{19}, 9.81818 \cdot 10^{-1})$$

and

$$V = \begin{bmatrix} 1.00000 & -1.00000 \cdot 10^{-11} & -9.09091 \cdot 10^{-22} \\ 1.00000 \cdot 10^{-11} & 1.00000 & -9.09091 \cdot 10^{-12} \\ 1.00000 \cdot 10^{-21} & 9.09091 \cdot 10^{-12} & 1.00000 \end{bmatrix}$$

One may compute that $\kappa(H) \approx 10^{40}$ and $\kappa(A) \approx 1.33$. Thus, according to Theorem 2.3, changing each entry of H in its 7th decimal place or beyond would not change Λ in the figures shown. The refined error bounds of Proposition 2.5 are essentially the same in this case. One can further verify the assertion of Proposition 2.10 that the ratios of the eigenvalues to the diagonal entries of H are bounded between $.9 = \lambda_{\min}(A)$ and $1.2 = \lambda_{\max}(A)$. One may also compute that the relative gaps $relgap_{\lambda_i}$ for all three eigenvalues are approximately 10^{10}. Thus, according to Theorem 2.7, 7th figure changes in H would not change its eigenvectors by more than 10^{-16} in norm. In fact, the eigenvectors are even more accurately determined than this. Let $\bar{V} = \{\bar{v}_i(j)\}$ be the matrix of upper bounds of entries of V as defined in Proposition 2.11:

$$\bar{V} \approx \begin{bmatrix} 1.5 & 1.5 \cdot 10^{-10} & 1.5 \cdot 10^{-20} \\ 1.5 \cdot 10^{-10} & 1.5 & 1.5 \cdot 10^{-10} \\ 1.5 \cdot 10^{-20} & 1.5 \cdot 10^{-10} & 1.5 \cdot 10^{-20} \end{bmatrix}$$

Then according to Proposition 2.12, 7th figure changes in H cause changes in at most the 5th digits of all the entries of V. In other words, for this examples all the eigenvalues and all the components of all the eigenvectors are determined to nearly full relative precision by the data. Later, we will show Jacobi can compute them with this accuracy. In contrast, QR does not even get the signs of the two small eigenvalues or many components of the eigenvectors correct.

The second example serves to illustrate the difference between Theorem 2.3 and the refined bounds of Proposition 2.5. Let $H = DAD$ where D is the same as before and

$$A = \begin{bmatrix} 1 & 1-\mu & 1-\mu \\ 1-\mu & 1 & 1-\mu \\ 1-\mu & 1-\mu & 1 \end{bmatrix}$$

where $\mu = 10^{-6}$. The eigenvalues of H are 10^{40}, $2 \cdot 10^{14}$ and $1.5 \cdot 10^{-6}$. Now $\kappa(A) \approx 10^{6}$, so according to Theorem 2.3, an η relative change in the matrix entries will cause as much as a $10^{6}\eta$ relative change in the eigenvalues. In contrast, the refined bounds predict a relative change of η in 10^{40} and $10^{6}\eta$ in the two smaller eigenvalues. Thus, the largest eigenvalue is just as insensitive as predicted by standard norm based perturbation theory.

2.2 Optimality of the Bounds for Symmetric Positive Definite Matrices

In this section we show that the bounds of the last section are attainable. In other words, the only symmetric positive definite matrices whose eigenvalues are determined to high relative accuracy by the matrix entries are those $H = DAD$ where A is well conditioned.

In particular, we give explicit small componentwise relative perturbations which attain the eigenvalue bounds; it suffices to choose a diagonal perturbation. We have (necessarily) slightly weaker results for the optimality of our eigenvector bounds.

We begin by showing that the assumption $\|\delta A\|_2 < \lambda_{\min}(A)$ of the last section is essential to having relative error bounds at all. If this bound were violated, $A + \delta A$ (and so $H + \delta H$)

could become indefinite, implying that all relative accuracy in at least one eigenvalue is completely lost. In contrast to standard perturbation theory, however, which assumes a bound on $\|\delta H\|_2$ instead of $\|\delta A\|_2$, one cannot say which eigenvalue will lose relative accuracy first. In the conventional case, as $\|\delta H\|_2$ grows, it is the smallest eigenvalues which lose accuracy first, the larger ones remaining accurate. As $\|\delta A\|_2$ grows, however, *any* eigenvalue in the spectrum (except the very largest) may lose its relative accuracy first. The following example illustrates this:

$$H = \begin{bmatrix} 10^{20} & & & \\ & 1 & .99 & \\ & .99 & 1 & \\ & & & 10^{-20} \end{bmatrix}, \; A = \begin{bmatrix} 1 & & & \\ & 1 & .99 & \\ & .99 & 1 & \\ & & & 1 \end{bmatrix}, \; D = \mathrm{diag}\,(10^{10}, 1, 1, 10^{-10})$$

Note that $\lambda_{\min}(A) = .01$. As $\|\delta A\|_2$ approaches .01, the eigenvalues near 10^{20}, 1.99 and 10^{-20} retain their accuracy, but the one near .01 can lose all its relative accuracy.

We next show that the relative error bound of Theorem 2.1 can be nearly attained for at least one eigenvalue simply by making appropriate small relative perturbations to the diagonal of H.

Proposition 2.13 *Let $H = DAD$ be symmetric positive definite, with $D = \mathrm{diag}\,(H_{ii}^{1/2})$ diagonal and $A_{ii} = 1$. Let $\delta A = \eta I$, $0 < \eta < \lambda_{\min}(A)$, and $H + \delta H = D(A + \delta A)D$. Then for some i we have*

$$\frac{\lambda_i(H+\delta H)}{\lambda_i(H)} \geq \left(1 + \frac{\eta}{\lambda_{\min}(A)}\right)^{1/n} \approx 1 + \frac{\eta}{n\lambda_{\min}(A)}$$

PROOF. We have

$$\prod_i \lambda_i(H) = \det(DAD) = \det(D^2)\det(A) = \det(D^2)\prod_i \lambda_i(A)$$

and

$$\prod_i \lambda_i(H+\delta H) = \det(D(A+\eta I)D) = \det(D^2)\det(A+\eta I) = \det(D^2)\prod_i(\lambda_i(A)+\eta)$$

Therefore

$$\prod_i \frac{\lambda_i(H+\delta H)}{\lambda_i(H)} = \prod_i \frac{\lambda_i(A)+\eta}{\lambda_i(A)} \geq 1 + \frac{\eta}{\lambda_{\min}(A)}$$

implying that at least one factor $\lambda_i(H+\delta H)/\lambda_i(H)$ must exceed $(1+\eta/\lambda_{\min}(A))^{1/n}$. This last expression is approximately $1 + \eta/(n\lambda_{\min}(A))$ when $\eta \ll \lambda_{\min}(A)$. ∎

The example at the beginning of this section showed that the error bound of Theorem 2.3 and the last proposition may only be attained for one eigenvalue. Proposition 2.5 of the last subsection showed that for asymptotically small $\|\delta A\|_2$, the maximum perturbation in each eigenvalue may be attained only with small diagonal perturbations of A.

After we show that the rounding errors introduced by Jacobi are of the form $\|\delta A\|_2 = O(\varepsilon)$ in the next section, Propositions 2.13 and 2.5 will show that Jacobi (modulo the assumption on $\max_m \kappa(A_m)/\kappa(A_0)$) computes all the eigenvalues with optimal accuracy, provided that only the diagonal entries of H have small relative errors. The same optimality property is true of bisection.

Now we consider eigenvectors. Here our results are necessarily weaker, as the following example shows. Suppose H is diagonal with distinct eigenvalues. Then small relative perturbations to the matrix entries leave H diagonal and its eigenvalue matrix (the identity matrix) unchanged. Therefore, the only way we can hope to attain the bounds of Theorem 2.7 is to use perturbations δA which are possibly dense, even if H is not. Furthermore, a block diagonal example like the first one in this section shows that the attainable eigenvector perturbations will not necessarily grow with $\kappa(A)$. Thus, the best we can prove is

Proposition 2.14 *Let $H = DAD$, λ_i, v_i, δH, δA and η be as in Proposition 2.5. Let v_i' be the unit eigenvectors of $H + \delta H$. Then one can choose δA, $\|\delta A\|_2 \equiv \eta \ll \lambda_{\min}(A)$, so that*

$$\|v_i - v_i'\|_2 \geq \frac{\eta}{\lambda_{\max}(A) relgap_{\lambda i}} + O(\eta^2)$$

Proof. Consider expression (2.8) for $v_i - v_i'$ (there δA is written ϵE). By using a Householder transformation, one can prove there exists a symmetric δA such that $y_k^T \delta A y_i = \|y_k\|_2 \|y_i\|_2 \|\delta A\|_2$ for arbitrary y_k and y_i. Since $\lambda_k = y_k^T A y_k \leq \|y_k\|^2 \lambda_{\max}(A)$, we can find δA to make $y_k^T \delta A y_i \geq (\lambda_i \lambda_k)^{1/2} \|\delta A\|_2 / \lambda_{\max}(A)$. Choosing k so that λ_k is closest to λ_i completes the proof. ∎

2.3 Singular Value Decomposition

The results on singular values and singular vectors are analogous to the results for eigenvalues and eigenvectors in the first subsection. Just as we derived perturbation bounds for eigenvalues from a more general result for generalized eigenvalues of pencils, we will start with a perturbation bound for generalized singular values and then specialize to standard singular values.

Let G_1 and G_2 be matrices with the same number of columns, G_2 of full column rank, and otherwise both arbitrary. We define the *i-th generalized singular value $\sigma_i(G_1, G_2)$ of the pair* (G_1, G_2) as the square root of the i-th eigenvalue of the definite pencil $G_1^T G_1 - \lambda G_2^T G_2$ [9]. If we let G_2 be the identity, $\sigma_i(G_1, G_2)$ is the same as the standard singular value $\sigma_i(G_1)$ of G_1.

Lemma 2.15 *Let G_1 and G_2 be matrices with the same number of columns, G_2 of full column rank, and otherwise both arbitrary. Let δG_j be a perturbation of G_j such that*

$$\|\delta G_j x\|_2 \leq \eta_j \|G_j x\|_2$$

for all x and some $\eta_j < 1$. Let σ_i be the i-th generalized singular value of (G_1, G_2) and σ_i' be the i-th generalized singular value of $(G_1 + \delta G_1, G_2 + \delta G_2)$. Then either $\sigma_i = \sigma_i' = 0$ or

$$\frac{1 - \eta_1}{1 + \eta_2} \leq \frac{\sigma_i'}{\sigma_i} \leq \frac{1 + \eta_1}{1 - \eta_2}$$

PROOF. From the Courant-Fischer minimax theorem [14] we have

$$\sigma_i = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{\|G_1 x\|_2}{\|G_2 x\|_2}$$

where the minimum is over all i-dimensional subspaces $\mathbf{S}^i$. The rest of the proof is analogous to that of Lemma 2.1. ∎

Lemma 2.16 *Let G_1 and G_2 be as in Lemma 2.15. Let $G_j = B_j \Delta_j$ where Δ_j has full rank and is otherwise arbitrary. Let $\delta G_j = \delta B_j \Delta_j$ be a perturbation of G_j such that $\|\delta B_j x\|_2 \leq \eta_j \|B_j x\|_2$ for all x and some $\eta_j < 1$. Let σ_i and σ_i' be the i-th generalized singular values of (G_1, G_2) and $(G_1 + \delta G_1, G_2 + \delta G_2)$, respectively. Then either $\sigma_i = \sigma_i' = 0$ or*

$$\frac{1-\eta_1}{1+\eta_2} \leq \frac{\sigma_i'}{\sigma_i} \leq \frac{1+\eta_1}{1-\eta_2}$$

The proof is analogous the proof of Lemma 2.2.

Theorem 2.17 *Let $G = BD$ be a general full rank matrix, and D chosen diagonal so that the columns of B have unit two-norm (i.e. D_{ii} equals the two-norm of the i-th column of G). Let $\delta G = \delta B D$ be a perturbation of G such that $\|\delta B\|_2 \equiv \eta < \sigma_{\min}(B)$. Let σ_i and σ_i' be the i-th singular values of G and $G + \delta G$, respectively. Then*

$$\frac{|\sigma_i - \sigma_i'|}{\sigma_i} \leq \frac{\eta}{\sigma_{\min}(B)} \leq \kappa(B) \cdot \eta \tag{2.18}$$

where $\kappa(B) = \sigma_{\max}(B)/\sigma_{\min}(B) \leq n^{1/2}/\sigma_{\min}(B)$, and n is the number of columns of G. In particular, if $|\delta G_{ij}/G_{ij}| \leq \eta/n$, then $\|\delta B\|_2 \leq \eta$ and the bound (2.18) applies.

The proof is analogous to that of Theorem 2.3.

Just as the bounds of Theorem 2.3 were not attainable by all eigenvalues, neither are the bounds of Theorem 2.17 attainable for all singular values. Analogous to Proposition 2.5, we may derive tighter bounds for individual singular values.

Proposition 2.19 *Let $G = BD$ be as in Theorem 2.17, with singular values σ_i, right unit singular vectors v_i and left unit singular vectors u_i. Let $G + \delta G = (B + \delta B)D$ have singular values σ_i', where $\|\delta B\|_2 \equiv \eta \ll \sigma_{\min}(B)$. Then the bound*

$$\frac{|\sigma_i - \sigma_i'|}{\sigma_i} \leq \frac{\eta \|D v_i\|_2}{\sigma_i} + O(\eta^2) \tag{2.20}$$

is attainable by the perturbation $\delta B = \eta u_i (D v_i)^T / \|D v_i\|_2$.

PROOF. Bound (2.20) is derived from the standard first order perturbation theory which says $\sigma_i(G + \delta G) = \sigma_i(G) + u_i^T \delta G v_i + O(\|\delta G\|_2^2)$, and substituting $|u_i^T \delta G v_i| = |u_i^T \delta B D v_i| \leq \|D v_i\|_2 \|\delta B\|_2$. This last inequality is clearly attained by the choice of δB in the statement of the proposition. ∎

Now we consider the singular vectors. For simplicity we assume G is square. We use the fact that if $G = U\Sigma V^T$ is the singular value decomposition of G, then $2^{-1/2} \cdot \begin{bmatrix} V & V \\ U & -U \end{bmatrix}$ is the eigenvector matrix of the symmetric matrix $\begin{bmatrix} 0 & G^T \\ G & 0 \end{bmatrix}$ [9]. Therefore we can use perturbation theory for eigenvectors of symmetric matrices to do perturbation theory for singular vectors of general matrices.

We also need to define the gaps for the singular vector problem. The *absolute gap for singular values* is

$$absgap_{\sigma i} \equiv \min_{k \neq i} \frac{|\sigma_i - \sigma_k|}{\|G\|_2}$$

i.e. essentially the same as the absolute gap for eigenvalues. However the *relative gap for singular values*

$$relgap_{\sigma i} \equiv \min_{k \neq i} \frac{|\sigma_i - \sigma_k|}{\sigma_i + \sigma_k}$$

is somewhat different from the relative gap for eigenvalues.

The standard perturbation theorem for singular vectors is essentially the same as for eigenvectors. Let G have right (or left) unit singular vectors v_i, and let $G + \delta G$ have right (or left) unit singular vectors v_i'. Let $\eta = \|\delta G\|_2 / \|G\|_2$. Then

$$\|v_i - v_i'\|_2 \leq \frac{\eta}{absgap_{\sigma i}} + O(\eta^2)$$

We improve this as follows:

Theorem 2.21 *Let $G = BD$ be as in Theorem 2.17. Define $G(\epsilon) = (B + \epsilon E)D$ where E is any matrix with unit two-norm. Let $\sigma_i(\epsilon)$ be the i-th singular value of $G(\epsilon)$, and assume $\sigma_i(0)$ is simple so that the corresponding right unit singular vector $v_i(\epsilon)$ and left unit singular vector $u_i(\epsilon)$ are well defined for sufficiently small ϵ. Then*

$$\max(\|v_i(\epsilon) - v_i(0)\|_2, \|u_i(\epsilon) - u_i(0)\|_2) \leq \frac{(n - .5)^{1/2}\epsilon}{\sigma_{\min}(B) \cdot relgap_{\sigma i}} + O(\epsilon^2) \leq \frac{(n - .5)^{1/2}\kappa(B)\epsilon}{relgap_{\sigma i}} + O(\epsilon^2)$$

PROOF. Let $v_k(0)$ be abbreviated by v_k and $u_k(0)$ by u_k. Define

$$\hat{G} = \begin{bmatrix} 0 & G^T \\ G & 0 \end{bmatrix}, \; \hat{E} = \begin{bmatrix} 0 & E^T \\ E & 0 \end{bmatrix}, \; \hat{D} = \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix}, \; \hat{B} = \begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix}, \; x_i^{\pm} = \frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$$

so that $\hat{G} = \hat{D}\hat{B}\hat{D}$. $\hat{G}$ has eigenvalues $\pm\sigma_i$ corresponding to eigenvectors $x_i^{\pm}$. Let $x_i^{\pm}(\epsilon)$ be the eigenvectors of

$$\begin{bmatrix} 0 & D(B^T + \epsilon E^T) \\ (B + \epsilon E)D & 0 \end{bmatrix} = \hat{D}(\hat{B} + \epsilon\hat{E})\hat{D}$$

Then from [9] we have

$$x_i^{\pm}(\epsilon) = x_i^{\pm} + \epsilon \sum_{\pm k \neq \pm i} \frac{x_i^{\pm\,T} \hat{D}\hat{E}\hat{D} x_k^{\pm}}{\pm\sigma_i \mp \sigma_k} \cdot x_k^{\pm} + O(\epsilon^2) \tag{2.22}$$

Now we compute

$$|x_i^{\pm\,T}\hat{D}\hat{E}\hat{D}x_k^{\pm}| = \frac{|u_i^T E D v_k + v_i^T D E^T u_k|}{2} \le \frac{\|Dv_k\|_2 + \|Dv_i\|_2}{2}$$

We also have

$$\sigma_i = x_i^{+\,T}\hat{G}x_i^{+} = u_i^T B D v_i = \|BDv_i\|_2 \ge \sigma_{\min}(B)\cdot\|Dv_i\|_2$$

so

$$|x_i^{\pm\,T}\hat{D}\hat{E}\hat{D}x_k^{\pm}| \le \frac{\sigma_i + \sigma_k}{2\cdot\sigma_{\min}(B)}$$

Taking norms in (2.22) yields the result. ∎

Corollary 2.23 *Let $G = BD$ be as in Theorem 2.17. Suppose $\delta \equiv \|\delta B\|_2/\sigma_{\min}(B)$ satisfies*

$$\frac{\delta}{1-\delta} < relgap_{\sigma i}$$

Let v_i and u_i be the unit right and left singular vectors of G, respectively, and let v_i' and u_i' be the unit right and left singular vectors of $G' = (B + \delta B)D$, respectively. Then

$$\max(\|v_i - v_i'\|_2, \|u_i - u_i'\|_2) \le \frac{(n-.5)^{1/2}\delta}{(1-\delta)((1-\delta)relgap_{\sigma i} - \delta)}$$

Proof. The proof is analogous to the proof of Corollary 2.9. ∎

There are analogs to Propositions 2.10 through 2.12 of the last section, gotten by considering $H = G^T G$:

Proposition 2.24 *Let $G = BD$ be as in Theorem 2.17. Let $\sigma_1 \le \cdots \le \sigma_n$ be the singular values of G and $d_1 \le \cdots \le d_n$ the diagonal entries of D in increasing order. Then*

$$\sigma_{\min}(B) \le \frac{\sigma_i}{d_i} \le \sigma_{\max}(B)$$

Proposition 2.25 *Let $G = BD$ be as in Theorem 2.17 with singular values $\sigma_1 \le \cdots \le \sigma_n$. Let v_i be the i-th right singular vector of G, normalized so that its i-th component $v_i(i) = 1$. Then*

$$|v_i(j)| \le \bar{\bar{v}}_i(j) \equiv (\kappa(B))^3 \cdot \min(\frac{\sigma_i}{\sigma_j}, \frac{\sigma_j}{\sigma_i})$$

We also have

$$|v_i(j)| \le (\kappa(B))^3 \cdot \min(\frac{d_i}{d_j}, \frac{d_j}{d_i})$$

Proposition 2.26 *Let $G(\epsilon)$ and $v_i(\epsilon)$ be as in Theorem 2.21, and $\bar{\bar{v}}_i(j)$ as in Proposition 2.25. Then*

$$|v_i(\epsilon)(j) - v_i(0)(j)| \le \frac{2(n-1)^{1/2}}{\sigma_{\min}^2(B)\cdot relgap_{\sigma i}} \cdot \epsilon \cdot \bar{\bar{v}}_i(j) + O(\epsilon^2)$$

There are analogs to all the results in this section for matrices $G = DB$ scaled from the left instead of the right. Thus, one can choose to scale either the rows or the columns of G to have unit two-norms, whichever one minimizes the condition number. It is natural to ask if one can do better by considering two sided diagonal scaling $D_1 G D_2$; to date we have been unable to formulate a reasonable perturbation theory. To see why, note that if G is triangular it can be made as close to the identity matrix as desired by two sided scaling, even though its singular values can be quite sensitive.

2.4 Optimality of the Bounds for the Singular Value Decomposition

The results in this section are analogous to but necessarily weaker than the results of subsection 2.2. In particular, it is no longer the case that the perturbation bounds for the singular values can be attained by small relative perturbations in the matrix entries.

First consider the restriction $\|\delta B\|_2 < \sigma_{\min}(B)$. Just as in the symmetric positive definite case, this is necessary so that $B + \delta B$ remains nonsingular. When $B + \delta B$ becomes singular, at least one singular value will necessarily lose all relative accuracy. The same kind of block diagonal example as in subsection 2.2 also shows that only one singular value may have its sensitivity depend on $\kappa(B)$, and it might be anywhere in the spectrum (except the very largest singular value).

In order to prove an analogue of Proposition 2.13, we must permit perturbations δB of B which are small in norm but may make large relative changes in tiny entries of B (a similar perturbation was needed to prove that the bound in Proposition 2.19 was attainable):

Proposition 2.27 *Let $G = BD$ with D diagonal and the columns of B having unit norm. Then there exists a δB with $\|\delta B\|_2 = \eta < \sigma_{\min}(B)$ such that for $G + \delta G = (B + \delta B)D$ we have for at least one i*

$$\frac{\sigma_i(G+\delta G)}{\sigma_i(G)} \geq (1 + \frac{\eta}{\sigma_{\min}(B)})^{1/n} \approx 1 + \frac{\eta}{n\sigma_{\min}(B)}$$

If we restrict δB so that $|\delta B_{ij}/B_{ij}| \leq \eta$, then such a perturbation δB may not exist.

Proof. The proof is very similar to that of Proposition 2.13. Let X be a rank one matrix of minimal 2-norm such that $B + X$ is singular, and let $\delta B = -\eta X$. Then as in Proposition 2.13 we discover that

$$\prod_i \frac{\sigma_i(G+\delta G)}{\sigma_i(G)} = 1 + \frac{\eta}{\sigma_{\min}(B)}$$

and so at least one term $\sigma_i(G + \delta G)/\sigma_i(G)$ exceeds $(1 + \eta/\sigma_{\min}(B))^{1/n}$. To see that small componentwise relative perturbations are not sufficient, consider the matrix

$$G = B = \begin{bmatrix} 1 & 1 \\ -\epsilon & \epsilon \end{bmatrix}$$

with $\epsilon \ll 1$. The condition number of B is approximately $1/\epsilon$, and relative perturbations of size η in its entries cannot change its singular values by more than a factor of about $(1 \pm \eta)^2$. ∎

As in Proposition 2.14, our lower bound on the attainable perturbations in the singular vectors requires a dense δB and does not grow with $\kappa(B)$.

Proposition 2.28 *Let $G = BD$, σ_i, u_i, v_i, $\delta G = \delta BD$ and η be as in Proposition 2.19. Let u_i' and v_i' be the unit left and right singular vectors of $G + \delta G$, respectively. Then one can choose δB, $\|\delta B\|_2 \equiv \eta \ll \sigma_{\min}(B)$, so that*

$$\max(\|u_i - u_i'\|_2, \|v_i - v_i'\|_2) \geq \frac{\eta}{2^{3/2}\sigma_{\max}(B) relgap_{\sigma i}} + O(\eta^2)$$

PROOF. Consider expression (2.22) for the perturbation $(u_i - u_i', v_i - v_i')$ (there δB is written ϵE). Choose k so $|\sigma_i - \sigma_k|$ is minimized. Let $\delta B = \eta u_i (D v_k)^T / \| D v_k \|_2$ if $\sigma_k > \sigma_i$ and $\delta B = \eta u_k (D v_i)^T / \| D v_i \|_2$ otherwise. The rest of the proof is a straightforward computation. ∎

3 Two-sided Jacobi

In this section we prove that two-sided Jacobi in floating point arithmetic applied to a positive definite symmetric matrix computes the eigenvalues and eigenvectors with the error bounds of section 2.

In this introduction we present the algorithm and our model of floating point arithmetic. In subsection 3.1, we derive error bounds for the computed eigenvalues. In subsection 3.2, we derive error bounds for the computed eigenvectors.

Let $H_0 = D_0 A_0 D_0$ be the initial matrix, and $H_m = D_m A_m D_m$ where H_m is obtained from H_{m-1} by applying a single Jacobi rotation. Here D_m is diagonal and A_m has unit diagonal as before. All the error bounds in this section contain the factor $\max_m \kappa(A_m)$, whereas the perturbation bounds of section 2 are proportional to $\kappa(A_0)$. Therefore, our claim that Jacobi solves the eigenproblem as accurately as predicted in section 2 depends on the ratio $\max_m \kappa(A_m)/\kappa(A_0)$ being modest in size. Note that convergence of H_m to diagonal form is equivalent to the convergence of A_m to the identity, or $\kappa(A_m)$ to 1. Thus we expect $\kappa(A_m)$ to be less than $\kappa(A_0)$ eventually.

We have overwhelming numerical evidence that $\max_m \kappa(A_m)/\kappa(A_0)$ is modest in size; in section 7, the largest value this ratio attained in random testing was 1.82. Our theoretical understanding of why this ratio is so small is somewhat weaker; we present our theoretical bounds on this ratio in section 6.

The essential difference between our algorithm and standard two-sided Jacobi is the stopping criterion: according to Theorem 2.3, we must set H_{ij} to zero only if $H_{ij}/(H_{ii}H_{jj})^{1/2}$ is small, not just if $H_{ij}/\max_{kl}|H_{kl}|$ is small. This stopping criterion has been suggested before [20, 5, 3, 19], but without our explanation of its benefits. Otherwise, our algorithm is a simplification of the standard one introduced by Rutishauser [14]. We have chosen a simple version of the algorithm, omitting enhancements like delayed updates of the diagonals and fast rotations, to make the error analysis clearer (an error analysis of these enhancements is future work).

Algorithm 3.1 *Two-sided Jacobi for the symmetric positive definite eigenproblem. tol is a user defined stopping criterion. The matrix V whose columns are the computed eigenvectors initially contains the identity.*

repeat
    for all pairs $i < j$
        /* *compute the Jacobi rotation which diagonalizes* $\begin{bmatrix} H_{ii} & H_{ij} \\ H_{ji} & H_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix}$ /*
        $\zeta = (b-a)/(2c)$
        $t = sign(\zeta)/(|\zeta| + \sqrt{1+\zeta^2})$
        $cs = 1/\sqrt{1+t^2}$
        $sn = cs * t$
        /* *update the 2 by 2 submatrix* */
        $H_{ii} = a - c * t$
        $H_{jj} = b + c * t$
        $H_{ij} = H_{ji} = 0$
        /* *update the rest of rows and columns i and j* */

```
        for k = 1 to n except i and j
            tmp = H_ik
            H_ik = cs * tmp - sn * H_jk
            H_jk = sn * tmp + cs * H_jk
            H_ki = H_ik
            H_kj = H_jk
        endfor
        /* update the eigenvector matrix V */
        for k = 1 to n
            tmp = V_ki
            V_ki = cs * tmp - sn * V_kj
            V_kj = sn * tmp + cs * V_kj
        endfor
    endfor
until convergence (all |H_ij|/(H_ii H_jj)^(1/2) <= tol )
```

Our model of arithmetic is a variation on the standard one: the floating point result $fl(\cdot)$ of the operation $(\cdot)$ is given by

$$
\begin{aligned}
fl(a \pm b) &= a(1+\varepsilon_1) \pm b(1+\varepsilon_2) \\
fl(a \times b) &= (a \times b)(1+\varepsilon_3) \\
fl(a/b) &= (a/b)(1+\varepsilon_4) \\
fl(\sqrt{a}) &= \sqrt{a}(1+\varepsilon_5)
\end{aligned}
\tag{3.1}
$$

where $|\varepsilon_i| \le \varepsilon$, and $\varepsilon \ll 1$ is the machine precision. This is somewhat more general than the usual model which uses $fl(a \pm b) = (a \pm b)(1+\varepsilon_1)$ and includes machines like the Cray which do not have a guard digit. This does not greatly complicate the error analysis, but it is possible that the computed rotation angle may be off by a factor of 2, whereas with a guard digit the rotation angle is always highly accurate. This may adversely affect convergence, but as we will see it does not affect the one-step error analysis.

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

3.1 Error Bounds for Eigenvalues Computed by Two-sided Jacobi

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

Theorem 3.2 *Let H_m be the sequence of matrices generated by Algorithm 3.1 in finite precision arithmetic with precision ε; that is H_{m+1} is obtained from H_m by applying a single Jacobi rotation. Then the following diagram*

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

commutes in the following sense: The top arrow indicates that H_{m+1} is obtained from H_m by applying one Jacobi rotation in floating point arithmetic. The diagonal arrow indicates that H_{m+1} is obtained from H'_m by applying one Jacobi rotation in exact arithmetic; thus H_{m+1} amd H'_m are exactly similar. The vertical arrow indicates that $H'_m = H_m + \delta H_m$. δH_m is bounded as follows. Write $\delta H_m = D_m \delta A_m D_m$. Then

$$\|\delta A_m\|_2 \le (182(2n-4)^{1/2} + 104)\varepsilon \tag{3.3}$$

In other words, if $\|\delta A_m\|_2 < \lambda_{\min}(A_m)$, one step of Jacobi satisfies the assumptions needed for the error bounds of section 2.

Corollary 3.4 *Assume Algorithm 3.1 converges, and that H_M is the final matrix whose diagonal entries we take as the eigenvalues. Write $H_m = D_m A_m D_m$ with D_m diagonal and A_m with ones on the diagonal for $0 \le m \le M$. Let λ_j be the j-th eigenvalue of H_0 and λ'_j be the j-th diagonal entry of H_M. Then to first order in ε the following error bound holds:*

$$\frac{|\lambda_j - \lambda'_j|}{\lambda_j} \le (\varepsilon \cdot M \cdot (182(2n-4)^{1/2} + 104) + n \cdot tol) \cdot \max_{0 \le m \le M} \kappa(A_m) \tag{3.5}$$

Remark. In numerical experiments presented in section 7, there was no evidence that the actual error bounded in (3.5) grew with increasing n or M.

Proof of Corollary 3.4. Bound (3.5) follows by substituting the bound (3.3) and the stopping criterion into Theorem 2.3. ∎

Remark. A similar bound can be obtained based on the error bound in Proposition 2.5.

Proof of Theorem 3.2. The proof of the commuting diagram is a tedious computation. Write the 2 by 2 submatrix of H_{mm} being reduced as

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv \begin{bmatrix} d_i^2 & z d_i d_j \\ z d_i d_j & d_j^2 \end{bmatrix}$$

where we assume without loss of generality that $a \ge b$ and $c > 0$. By positive definiteness $0 < z \le \bar{z} \equiv (\kappa(A_m) - 1)/(\kappa(A_m) + 1) < 1$. Let a' and b' be the new values of H_{ii} and H_{jj} computed by the algorithm, respectively. Let $x \equiv d_j/d_i \le 1$. We consider two cases, $x \le \bar{x} \equiv (\sqrt{5} - 1)/2 \approx .62$, and $x > \bar{x}$.

First consider $x \le \bar{x}$. Systematic application of formulas (3.1) shows that

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

where $\tilde{b} \equiv (1+\varepsilon_1)b/(1+\varepsilon_2) \equiv (1+\varepsilon_b)b$, $|\varepsilon_b| \le 2\varepsilon$. Thus $\zeta = (1+\varepsilon_\zeta)(\tilde{b} - a)/(2c)$ where $|\varepsilon_\zeta| \le 3\varepsilon$.

Let $t(c)$ denote the true value of t (i.e. without rounding error) as a function of a, $\tilde{b}$ and c. Using (3.1) again one can show $t = (1+\varepsilon_t)t(c)$ where $|\varepsilon_t| \le 7\varepsilon$.

Next

$$
\begin{aligned}
b' &= fl(b+ct) = (1+\varepsilon_5)b + (1+\varepsilon_6)(1+\varepsilon_7)ct \\
&= \frac{(1+\varepsilon_2)(1+\varepsilon_5)}{1+\varepsilon_1}(\tilde{b} + \frac{(1+\varepsilon_1)(1+\varepsilon_6)(1+\varepsilon_7)(1+\varepsilon_t)}{(1+\varepsilon_2)(1+\varepsilon_5)}ct(c)) \\
&\equiv (1+\varepsilon_{b'})(\tilde{b} + (1+\varepsilon_{ct(c)})ct(c)) \qquad (3.6)
\end{aligned}
$$

where $|\varepsilon_{ct(c)}| \le 12\varepsilon$ and $|\varepsilon_{b'}| \le 3\varepsilon$. Since $|t(c)|$ is an increasing function of c, we can write $(1+\varepsilon_{ct(c)})ct(c) = (1+\varepsilon_c)c \cdot t((1+\varepsilon_c)c)$, for some ε_c where $|\varepsilon_c| \le |\varepsilon_{ct(c)}| \le 12\varepsilon$.

Now we can define $\tilde{c} \equiv (1+\varepsilon_c)c$, and $\tilde{\zeta}$, $\tilde{t}$, $\tilde{cs}$ and $\tilde{sn}$ as the true values of the untilded quantities computed without rounding error starting from a, $\tilde{b}$ and $\tilde{c}$. $\tilde{cs}$ and $\tilde{sn}$ will define the exact Jacobi rotation $J_m \equiv \begin{bmatrix} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs} \end{bmatrix}$ which transforms H'_m to H_{m+1} in the commutative diagram in the statement of the theorem: $J_m^T H'_m J_m = H_{m+1}$.

Now we begin constructing δH_m. δH_m will be nonzero only in rows and columns i and j. First we compute its entries outside the 2 by 2 (i,j) submatrix. Using (3.1) one can show $cs = (1+\varepsilon_{cs})\tilde{cs}$ and $sn = (1+\varepsilon_{sn})\tilde{sn}$ where $|\varepsilon_{cs}| \le 22\varepsilon$ and $|\varepsilon_{sn}| \le 30\varepsilon$. Now let H'_{ik} and H'_{jk} denote the updated quantities computed by the algorithm. Then

$$
\begin{aligned}
H'_{ik} &= fl(cs * H_{ik} - sn * H_{jk}) \\
&= (1+\varepsilon_{10})(1+\varepsilon_8)csH_{ik} - (1+\varepsilon_9)(1+\varepsilon_{11})snH_{jk} \\
&= (1+\varepsilon_{10})(1+\varepsilon_8)(1+\varepsilon_{cs})\tilde{cs}H_{ik} - (1+\varepsilon_9)(1+\varepsilon_{11})(1+\varepsilon_{sn})\tilde{sn}H_{jk} \\
&\equiv \tilde{cs}H_{ik} - \tilde{sn}H_{jk} + \epsilon(H'_{ik}) \qquad (3.7)
\end{aligned}
$$

Similarly

$$
\begin{aligned}
H'_{jk} &= fl(sn * H_{ik} + cs * H_{jk}) \\
&= (1+\varepsilon_{14})(1+\varepsilon_{12})(1+\varepsilon_{sn})\tilde{sn}H_{ik} + (1+\varepsilon_{13})(1+\varepsilon_{15})(1+\varepsilon_{cs})\tilde{cs}H_{jk} \\
&\equiv \tilde{sn}H_{ik} + \tilde{cs}H_{jk} + \epsilon(H'_{jk}) \qquad (3.8)
\end{aligned}
$$

Now $x = d_j/d_i$ implies

$$
\tilde{\zeta} = \frac{b-a}{2\tilde{c}} = \frac{d_j^2 - d_i^2}{2\tilde{z}d_i d_j} = \frac{x^2-1}{2\tilde{z}x}
$$

where $\tilde{z} \equiv z(1+\varepsilon_c)$. Then $x \le \bar{x}$ implies

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

Also $|\tilde{sn}| \le |\tilde{t}|$, so this last expression is an upper bound on $|\tilde{sn}|$ as well. Substituting this bound on $\tilde{sn}$, $\tilde{cs} \le 1$, $|H_{ik}| \le d_i d_k \bar{z}$ and $|H_{jk}| \le d_j d_k \bar{z}$ into (3.7) and (3.8) yields

$$
\begin{aligned}
|\epsilon(H'_{ik})| &\le 56\varepsilon d_i d_k \bar{z} \\
|\epsilon(H'_{jk})| &\le 56\varepsilon d_j d_k \bar{z}/(1-\bar{x}^2)
\end{aligned}
$$

Thus

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

where

$$
\begin{aligned}
|\delta H_{ik}| &\leq 112\varepsilon d_i d_k \bar{z}/(1-\bar{x}^2) \\
|\delta H_{jk}| &\leq 112\varepsilon d_j d_k \bar{z}/(1-\bar{x}^2)
\end{aligned}
$$

Now we construct the 2 by 2 submatrix Δ of δH_m at the intersection of rows and columns i and j. We will construct it of three components $\Delta = \Delta_1 + \Delta_2 + \Delta_3$.

Consider the formula $a' = fl(a - c * t)$ for the i, i entry of H_{m+1}. Applying (3.1) systematically, we see

$$
\begin{aligned}
a' &= (1+\varepsilon_{18})a - (1+\varepsilon_{17})(1+\varepsilon_{16})ct \\
&= (1+\varepsilon_{18})a - (1+\varepsilon_{17})(1+\varepsilon_{16})(1+\varepsilon_t)ct(c) \\
&= (1+\varepsilon_{18})a - \frac{(1+\varepsilon_{17})(1+\varepsilon_{16})(1+\varepsilon_t)\tilde{c}t(\tilde{c})}{1+\varepsilon_{ct(c)}} \\
&\equiv (1+\varepsilon_{18})a - (1+\varepsilon'_{ct(c)})\tilde{c}t(\tilde{c})
\end{aligned}
$$

where $|\varepsilon'_{ct(c)}| \leq 21\varepsilon$. Since $a > 0$ and $\tilde{c}t(\tilde{c}) < 0$, we get

$$
a' = \left(1 + \frac{\varepsilon_{18}a - \varepsilon'_{ct(c)}\tilde{c}t(\tilde{c})}{a - \tilde{c}t(\tilde{c})}\right)(a - \tilde{c}t(\tilde{c})) \equiv (1+\varepsilon_{a'})(a - \tilde{c}t(\tilde{c}))
$$

where $|\varepsilon_{a'}| \leq 21\varepsilon$.

Now let

$$
\Delta_1 = \begin{bmatrix} 0 & \varepsilon_c c \\ \varepsilon_c c & \varepsilon_b b \end{bmatrix} = \begin{bmatrix} 0 & \tilde{c} - c \\ \tilde{c} - c & \tilde{b} - b \end{bmatrix}
$$

From earlier discussion we see

$$
J_m^T \left(\begin{bmatrix} a & c \\ c & b \end{bmatrix} + \Delta_1 \right) J_m = \begin{bmatrix} a - \tilde{c}t(\tilde{c}) & 0 \\ 0 & \tilde{b} + \tilde{c}t(\tilde{c}) \end{bmatrix}
$$

Next let

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

Thus

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

Finally, let

$$\Delta_3 = J_m \begin{bmatrix} 0 & 0 \\ 0 & b'(1 - ((1+\varepsilon_{a'})/(1+\varepsilon_b)) \end{bmatrix} J_m^T \equiv \begin{bmatrix} \tilde{sn}^2 \varepsilon_{b''} b & \tilde{cs}\tilde{sn}\varepsilon_{b''} b \\ \tilde{cs}\tilde{sn}\varepsilon_{b''} b & \tilde{cs}^2 \varepsilon_{b''} b \end{bmatrix}$$

where $|\varepsilon_{b''}| \le |\varepsilon_{a'}| + |\varepsilon_b| \le 24\varepsilon$. Then

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

as desired. This completes the construction of δH_m. We may bound

$$\|\delta A_m\|_2 \le \left(\frac{112(2n-4)^{1/2}\bar{z}}{1-\bar{x}^2} + 104 \right) \varepsilon \tag{3.9}$$

Now we consider the second case, when $x > \bar{x}$. The only thing that changes in the previous analysis is our analysis of δH_{ik} and δH_{jk}, since $\tilde{sn}$ is no longer small. Instead we substitute the bounds $|\tilde{sn}| \le 1$, $|\tilde{cs}| \le 1$, $|H_{ik}| \le d_i d_k \bar{z} \le d_j d_k \bar{z}/\bar{x}$ and $|H_{jk}| \le d_j d_k \bar{z}$ into (3.7) and (3.8) to get

$$\begin{aligned} |\epsilon(H'_{ik})| &\le 56\varepsilon d_i d_k \bar{z} \\ |\epsilon(H'_{jk})| &\le 56\varepsilon d_i d_k \bar{z} \end{aligned}$$

whence

$$\begin{aligned} |\delta H_{ik}| &\le 112\varepsilon d_i d_k \bar{z} \\ |\delta H_{jk}| &\le 112\varepsilon d_j d_k \bar{z}/\bar{x} \end{aligned}$$

and

$$\|\delta A_m\|_2 \le \left(\frac{112(2n-4)^{1/2}\bar{z}}{\bar{x}} + 104 \right) \varepsilon \tag{3.10}$$

Finally, we note our choice of $\bar{x}$ makes the upper bounds in (3.9) and (3.10) both equal, with $1/(1-\bar{x}^2) = 1/\bar{x} < 1.62$, proving the theorem. ∎

Remark. The quantity $182(2n-4)^{1/2}$ in the theorem may be multiplied by $\max_{m,i\ne j} |A_{m,ij}| < 1$. Thus if the A_m are strongly diagonally dominant, the part of the error term which depends on n is suppressed.

Commutative diagrams like the one in the theorem, where performing one step of the algorithm in floating point arithmetic is equivalent to making small relative errors in the matrix and then performing the algorithm exactly, occur elsewhere in numerical analysis. For example, such a diagram describes an entire sweep of the zero-shift bidiagonal QR algorithm [7], and is the key to the high accuracy achieved by that algorithm. Also, if one modifies one line in the traditional zero-shift symmetric tridiagonal QR algorithm so it is performed in higher precision, such a diagram describes an entire sweep of that algorithm as well [13]. Unfortunately, this ability to have a high accuracy sweep does not often translate into overall accuracy, because in tridiagonal QR $\kappa(A_m)$ frequently grows greatly for that algorithm before converging to 1.

3.2 Error Bounds for Eigenvectors Computed by Two-sided Jacobi

The next two theorems justify our accuracy claims for eigenvectors computed by two-sided Jacobi.

Theorem 3.11 *Let $V = [v_1, \ldots, v_n]$ be the matrix of unit eigenvectors computed by Algorithm 3.1 in finite precision arithmetic with precision ε. Let $U = [u_1, \ldots, u_n]$ be the true eigenvector matrix. Let $\bar{\kappa} \equiv \max_m \kappa(A_m)$ be the largest $\kappa(A_m)$ of any iterate. Then the error in the computed eigenvectors is bounded in norm by*

$$\|v_i - u_i\|_2 \leq \frac{(n-1)^{1/2}(n \cdot tol + M \cdot (182(2n-4)^{1/2} + 104)\varepsilon)\bar{\kappa}}{relgap_{\lambda_i}} + 46M\varepsilon \qquad (3.12)$$

Remark In numerical experiments presented in section 7, there was no evidence that the actual error bounded in (3.12) grew with increasing n or M.

PROOF. Let $H_0, \ldots, H_M$ be the sequence of matrices generated by the Jacobi algorithm, where H_M satisfies the stopping criterion. Let J_m be the exact Jacobi rotation which transforms H'_m to H_{m+1} in the commuting diagram of Theorem 3.2: $J_m^T H'_m J_m = H_{m+1}$.

We will use the approximation that $relgap_{\lambda_i}$ is the same for all H_m, even though it changes slightly. This contributes an $O(\varepsilon^2)$ term to the overall bound (which we ignore), but could be accounted for using the bounds of Theorem 3.2.

Initially, we will compute error bounds for the columns of $J_0 \cdots J_{M-1}$, ignoring any rounding errors occurring in computing their product. Then we will incorporate these rounding errors.

We will prove by induction that the i-th column v_{mi} of $V_m \equiv J_m \cdots J_{M-1}$ is a good approximation to the true i-th eigenvector u_{mi} of H_m. In particular, we will show that to first order in ε

$$\|u_i - v_{0i}\|_2 \leq \frac{(n-1)^{1/2}(n \cdot tol + M \cdot (182(2n-4)^{1/2} + 104)\varepsilon)\bar{\kappa}}{relgap_{\lambda_i}}$$

The basis of the induction is as follows. $V_M = I$ is the eigenvector matrix for H_M, which is considered diagonal since it satisfies the stopping criterion. Thus the norm error in v_{Mi} follows from plugging the stopping criterion into Theorem 2.7:

$$\|u_{Mi} - v_{Mi}\|_2 \leq \frac{(n-1)^{1/2} \cdot n \cdot tol \cdot \bar{\kappa}}{relgap_{\lambda_i}}$$

For the induction step we assume that

$$\|u_{m+1,i} - v_{m+1,i}\|_2 \leq \frac{(n-1)^{1/2}(n \cdot tol + (M-m-1) \cdot (182(2n-4)^{1/2} + 104)\varepsilon)\bar{\kappa}}{relgap_{\lambda_i}}$$

and try to extend to m. Consider the commuting diagram of Theorem 3.2. Accordingly, the errors in $V_m = J_m V_{m+1}$ considered as eigenvectors of H'_m are just the errors in V_{m+1} premultiplied by J_m. This does not increase them in 2-norm, since J_m is orthogonal. Now we change H'_m to H_m. This increases the norm error in v_{mi} by an amount bounded by plugging the bound for $\|\delta A_m\|_2$ into Theorem 2.7: $(n-1)^{1/2}\bar{\kappa}(182(2n-4)^{1/2} + 104)\varepsilon / relgap_{\lambda_i}$. This proves the induction step.

Finally, consider the errors from accumulating the product of slightly wrong values of J_m in floating point arithmetic. From the proof of Theorem 3.2, we see the relative errors in the entries of J_m are at most 30ε, and from the usual error analysis of a product of 2 by 2 rotations, we get $32\sqrt{2}M\varepsilon < 46M\varepsilon$ for the norm error in the product of M rotations. This completes the proof of bound (3.12). ∎

Now we consider the errors in the individual components of the computed eigenvectors $|u_i(j) - v_i(j)|$. From Proposition 2.12 we see that we can hope to bound this quantity by $O(\varepsilon)\bar{\kappa}\bar{v}_i(j)/\min(relgap_{\lambda_i}, 2^{-1/2})$, where

$$\bar{v}_i(j) \equiv \bar{\kappa}^{3/2} \min(\left(\frac{\lambda_i}{\lambda_j}\right)^{1/2} \left(\frac{\lambda_j}{\lambda_i}\right)^{1/2}) \tag{3.13}$$

is a modified upper bound for the eigenvector component $v_i(j)$ as in Proposition 2.11. In other words, we may have high relative accuracy even in the tiny components of the computed eigenvectors; this is the case in the example in the introduction and at the end of subsection 2.1. Our proof of this fact will not be as satisfactory as the previous result, because it will contain a "pivot growth" factor which probably grows at most linearly in M but for which we can only prove an exponential bound. In numerical experiments presented in section 7, there was no evidence that this factor grew with increasing n or M.

We will use $\bar{v}_i(j)$ as defined in (3.13) for each H_m, even though the values of λ_i and λ_j vary slightly from step to step. This error will contribute an $O(\varepsilon^2)$ term to the overall bound (we are ignoring such terms) but could be incorporated using the bounds of Corollary 3.4.

Theorem 3.14 *Let V, U, and $\bar{\kappa}$ be as in Theorem 3.11, and $\bar{v}_i(j)$ be as in (3.13). Then we can bound the error in the individual eigencomponents by*

$$|u_i(j) - v_i(j)| \le p(M,n) \cdot \frac{(tol + \varepsilon)\cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} \tag{3.15}$$

Here $p(M,n)$ is a "pivot growth" factor which probably grows at most linearly in M and in $n^{3/2}$ although all we can prove is an exponential bound.

Proof. The proof is similar to that of Theorem 3.11. One difference is that we use Proposition 2.12 instead of Theorem 2.7 to bound the errors in the eigenvectors. Another difference, which introduces the growth factor $p(M)$, is that we need to use the scaling of the entries of J_m to see how small eigenvector components have small errors; not being able to use the orthogonality of J_m introduces $p(M)$.

As in the proof of Theorem 3.11, let $V_m = J_m \cdots J_{M-1}$, where $J_m^T H_m' J_m = H_{m+1}$. Set $V_M = I$. The proof has three parts. In the first part we will show the i-th column of V_0 is a good approximation to the eigenvectors of H_0 in the sense of the theorem. In the second part we will show that the (i,j) entry of $J_0 \cdots J_m$ is bounded by a modest multiple of $\bar{v}_i(j)$. In the third part we will show the rounding errors committed in computing $J_0 \cdots J_m$ in floating point are small compared to $\bar{v}_i(j)$.

For the first part of the proof we will use induction to prove that the i-th column v_{mi} of V_m is a good approximation to the true i-th eigenvector u_{mi} of H_m. This will show

$$|u_i(j) - v_{0i}(j)| \le \rho_0 \frac{(tol + \varepsilon)\cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} \tag{3.16}$$

where ρ_0 is a constant (part of the "pivot growth" factor) we need to estimate. The base of the induction follows from plugging the stopping criterion into the bound of Proposition 2.12, yielding

$$|u_{Mi}(j) - v_{Mi}(j)| \le \frac{(2n-2)^{1/2} \cdot n \cdot tol \cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} \le \rho_M \frac{(tol + \varepsilon) \cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})}$$

where $\rho_M \equiv n(2n-2)^{1/2}$. The induction step will assume that

$$|u_{m+1,i}(j) - v_{m+1,i}(j)| \le \rho_{m+1} \frac{(tol + \varepsilon) \cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})}$$

which we will try to extend to m. Consider the commuting diagram of Theorem 3.2. Accordingly, the errors in the columns of $V_m = J_m V_{m+1}$ considered as eigenvectors of H'_m are just the errors in V_{m+1} premultiplied by J_m; let e_{mi} denote this error for the i-th column of V_m. Suppose J_m rotates in rows and columns k and l; then e_{mi} is identical to $u_{m+1,i} - v_{m+1,i}$ except for $e_{mi}(k)$ and $e_{mi}(l)$. We may assume without loss of generality that $k < l$ and $d_k \ge d_l$ (d_k^2 and d_l^2 are the diagonal entries of H_m). As in the proof of Theorem 3.2, there are two cases, when $x \equiv d_l/d_k \le \bar{x} \equiv (\sqrt{5}-1)/2$, and $x > \bar{x}$.

In the first case, $x \le \bar{x}$, we know as in the proof of Theorem 3.2 that $\tilde{sn}$, the sine in the rotation J_m, is bounded in magnitude by $x/(1-\bar{x}^2)$. Write $|\tilde{sn}| \le c_m (\lambda_l/\lambda_k)^{1/2}$ instead, where c_m is a modest constant. We can do this because $d_r \approx \lambda_r^{1/2}$ from Proposition 2.10. This lets us bound

$$\begin{aligned}
\begin{bmatrix} |e_{mi}(k)| \\ |e_{mi}(l)| \end{bmatrix} &= \left| J_m \begin{bmatrix} u_{m+1,i}(k) - v_{m+1,i}(k) \\ u_{m+1,i}(l) - v_{m+1,i}(l) \end{bmatrix} \right| \\
&\le \begin{bmatrix} |u_{m+1,i}(k) - v_{m+1,i}(k)| + |\tilde{sn}(u_{m+1,i}(l) - v_{m+1,i}(l))| \\ |\tilde{sn}(u_{m+1,i}(k) - v_{m+1,i}(k))| + |u_{m+1,i}(l) - v_{m+1,i}(l)| \end{bmatrix} \\
&\le \frac{\rho_{m+1}(tol+\varepsilon)\bar{\kappa}^{5/2}}{\min(relgap_{\lambda_i}, 2^{-1/2})} \cdot \\
&\quad \begin{bmatrix} \min(\left(\frac{\lambda_i}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_i}\right)^{1/2}) + c_m \left(\frac{\lambda_l}{\lambda_k}\right)^{1/2} \min(\left(\frac{\lambda_i}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_i}\right)^{1/2}) \\ c_m \left(\frac{\lambda_l}{\lambda_k}\right)^{1/2} \min(\left(\frac{\lambda_i}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_i}\right)^{1/2}) + \min(\left(\frac{\lambda_i}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_i}\right)^{1/2}) \end{bmatrix} \\
&\le \frac{\rho_{m+1}(tol+\varepsilon)\bar{\kappa}^{5/2}}{\min(relgap_{\lambda_i}, 2^{-1/2})} \cdot \begin{bmatrix} (1+c_m)\min(\left(\frac{\lambda_i}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_i}\right)^{1/2}) \\ (1+c_m)\min(\left(\frac{\lambda_i}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_i}\right)^{1/2}) \end{bmatrix} \\
&= (1+c_m)\frac{\rho_{m+1}(tol+\varepsilon)\bar{\kappa}}{\min(relgap_{\lambda_i}, 2^{-1/2})} \cdot \begin{bmatrix} \bar{v}_i(k) \\ \bar{v}_i(l) \end{bmatrix}
\end{aligned} \tag{3.17}$$

Now consider case 2, $x > \bar{x}$. Now λ_k and λ_l are reasonably close together. Thus, we may bound $|\tilde{sn}|$ simply by 1 in the derivation (3.17). This leads to the same bound with a possibly different c_m; we take the final c_m as the maximum of these two values. This bounds the error in the columns of V_m considered as eigenvectors of H'_m.

Now we change H'_m to H_m. This increases the bound for $|u_{mi}(j) - v_{mi}(j)|$ by an amount bounded by plugging the bound for $\|\delta A_m\|_2$ from Theorem 3.2 into Proposition 2.12: $(2n-2)^{1/2}(182(2n-4)^{1/2}+104)\cdot\varepsilon\cdot\bar\kappa\cdot\bar v_i(j)/\min(relgap_{\lambda_i}, 2^{-1/2})$. This completes the induction with

$$
\begin{aligned}
|u_{mi}(j) \;&-\; v_{mi}(j)| \\
&\leq ((1+c_m)\rho_{m+1} + (2n-2)^{1/2}(182(2n-4)^{1/2}+104)) \cdot \frac{(tol+\varepsilon)\cdot\bar\kappa\cdot\bar v_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} \\
&\equiv \rho_m \cdot \frac{(tol+\varepsilon)\cdot\bar\kappa\cdot\bar v_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} \qquad (3.18)
\end{aligned}
$$

Here

$$\rho_m = (1+c_m)\rho_{m+1} + (2n-2)^{1/2}(182(2n-4)^{1/2}+104) \ , \ \rho_M = n(2n-2)^{1/2} \qquad (3.19)$$

satisfies an exponential error bound, but it is clear from the derivation that linear growth is far more likely than exponential growth. This completes the first part of the proof.

In the second part of the proof we will show that the (i,j) entry of $\tilde V_m \equiv J_0\cdots J_m$ is bounded by a modest multiple of $\bar v_i(j)$. To do this we will prove by induction that

$$|\tilde v_{mi}(j)| \leq \tau_m \bar v_i(j) \qquad (3.20)$$

where $\tilde V_m = [\tilde v_{m1},\ldots,\tilde v_{mn}]$ and τ_m is a constant (part of the "pivot growth" factor) we need to estimate. The base of the induction is for $m=-1$, i.e. the null product, which we set equal to the identity matrix. This clearly satisfies (3.20) with $\tau_{-1}=1$. Now we assume (3.20) is true for $m-1$ and try to extend it to m. Suppose J_m rotates in rows and columns k and l. Postmultiplying $\tilde V_{m-1}$ by J_m only changes it in columns k and l. Assume as before that $k<l$ and $x = d_l/d_k \leq 1$. There are two cases as before, $x\leq\bar x$ and $x>\bar x$.

First consider the case $x\leq\bar x$. We may bound the (j,k) and (j,l) entries of $\tilde V_m$ as follows:

$$
\begin{aligned}
|[\tilde v_{m,j}(k), \tilde v_{m,j}(l)]| &= |[\tilde v_{m-1,j}(k), \tilde v_{m-1,j}(l)]\cdot\begin{bmatrix} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs}\end{bmatrix}| \\
&\leq \tau_{m-1}\bar\kappa^{3/2}[\min(\left(\frac{\lambda_j}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_j}\right)^{1/2}), \min(\left(\frac{\lambda_j}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_j}\right)^{1/2})]\cdot \\
&\quad \begin{bmatrix} 1 & c_m\left(\frac{\lambda_l}{\lambda_k}\right)^{1/2} \\ c_m\left(\frac{\lambda_l}{\lambda_k}\right)^{1/2} & 1\end{bmatrix} \\
&\leq \tau_{m-1}\bar\kappa^{3/2}(1+c_m)[\min(\left(\frac{\lambda_j}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_j}\right)^{1/2}), \min(\left(\frac{\lambda_j}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_j}\right)^{1/2})] \\
&= \tau_{m-1}(1+c_m)[\bar v_k(j), \bar v_l(j)] \\
&\equiv \tau_m[\bar v_k(j), \bar v_l(j)] \qquad (3.21)
\end{aligned}
$$

In the second case, $x>\bar x$, we get a similar bound. Here $\lambda_k\approx\lambda_l$, and we can simply bound $|\tilde{sn}|\leq 1$. This yields a slightly different c_m; for the final c_m we again take the maximum of the two. This ends the second part of the proof with

$$\tau_m = (1 + c_m)\tau_{m-1} \quad , \quad \tau_{-1} = 1 \tag{3.22}$$

Even though this only yields an exponential upper bound for τ_M, it is clear from the derivation that linear growth is far more likely than exponential growth.

In the third and final part of the proof we show the rounding errors in the (i, j) entry of the computed approximation to $\tilde{V}_{m-1}$ is bounded by $O(\varepsilon)\bar{v}_i(j)$. Let $\tilde{J}_m$ be the actual rotation which only approximates J_m. From the proof of Theorem 3.2, we have $cs = \tilde{cs}(1+\varepsilon_{cs})$ with $|\varepsilon_{cs}| \le 22\varepsilon$ and $sn = \tilde{sn}(1+\varepsilon_{sn})$ with $|\varepsilon_{sn}| \le 30\varepsilon$. Let $\tilde{\tilde{V}}_m = fl(\tilde{\tilde{V}}_{m-1} * \tilde{J}_m)$ be the actually computed eigenvector matrix after the m-th Jacobi rotation. The final computed eigenvector matrix is $V = \tilde{\tilde{V}}_{M-1}$. We will use induction to prove

$$|\tilde{\tilde{v}}_{m,i}(j) - \tilde{v}_{m,i}(j)| \le \chi_m \varepsilon \bar{v}_i(j) \tag{3.23}$$

where $\tilde{\tilde{V}}_m = [\tilde{\tilde{v}}_{m1}, \ldots, \tilde{\tilde{v}}_{mn}]$ and χ_m is a constant (part of the "pivot growth" factor) we need to estimate. The basis is again for $m = -1$ when $\tilde{\tilde{V}}_{-1} = \tilde{V}_{-1} = I$ and $\chi_{-1} = 0$. Now we assume (3.23) is true for $m - 1$ and try to extend it to m. As before, we assume J_m rotates in rows and columns k and l with $k < l$ and $x = d_k/d_l \le 1$. Write $\tilde{e}_{mi} \equiv \tilde{\tilde{v}}_{mi} - \tilde{v}_{mi}$. The (j, k) and (j, l) entries of $\tilde{\tilde{V}}_m$ are

$$\begin{aligned}
[\tilde{\tilde{v}}_{m,j}(k) \quad &, \quad \tilde{\tilde{v}}_{m,j}(l)] \\
&= [\tilde{\tilde{v}}_{m-1,j}(k)\tilde{cs}(1+\varepsilon_{cs})(1+\varepsilon_1)(1+\varepsilon_2) - \tilde{\tilde{v}}_{m-1,j}(l)\tilde{sn}(1+\varepsilon_{sn})(1+\varepsilon_3)(1+\varepsilon_4), \\
&\quad\ \tilde{\tilde{v}}_{m-1,j}(k)\tilde{sn}(1+\varepsilon_{sn})(1+\varepsilon_5)(1+\varepsilon_6) + \tilde{\tilde{v}}_{m-1,j}(l)\tilde{cs}(1+\varepsilon_{cs})(1+\varepsilon_7)(1+\varepsilon_8)] \\
&= [\tilde{v}_{m-1,j}(k)\tilde{cs} - \tilde{v}_{m-1,j}(l)\tilde{sn}, \tilde{v}_{m-1,j}(k)\tilde{sn} + \tilde{V}_{m-1,j}(l)\tilde{cs}] + \\
&\quad\ [24\varepsilon_9\tilde{cs}\tilde{v}_{m-1,j}(k) + 32\varepsilon_{10}\tilde{sn}\tilde{v}_{m-1,j}(l), 32\varepsilon_{11}\tilde{sn}\tilde{v}_{m-1,j}(k) + 24\varepsilon_{12}\tilde{cs}\tilde{v}_{m-1,j}(l)] + \\
&\quad\ [(1+24\varepsilon_9)\tilde{cs}\tilde{e}_{m-1,j}(k) + (1+32\varepsilon_{10})\tilde{sn}\tilde{e}_{m-1,j}(l), \\
&\quad\ (1+32\varepsilon_{11})\tilde{sn}\tilde{e}_{m-1,j}(k) + (1+24\varepsilon_{12})\tilde{cs}\tilde{e}_{m-1,j}(l)] \\
&= [\tilde{v}_{m,j}(k), \tilde{v}_{m,j}(l)] + I_1 + I_2
\end{aligned}$$

so $[\tilde{e}_{m,j}(k), \tilde{e}_{m,j}(l)] = I_1 + I_2$.

As before, there are two cases, $x \le \bar{x}$ and $x > \bar{x}$. Consider the case $x \le \bar{x}$. Using $|\tilde{sn}| \le c_m(\lambda_l/\lambda_k)^{1/2}$, $|\tilde{cs}| \le 1$, and $|\tilde{v}_{m-1,i}(j)| \le \tau_{m-1}\bar{v}_i(j)$, we get

$$\begin{aligned}
|I_1| &\le \varepsilon\tau_{m-1}\bar{\kappa}^{3/2}[(24+32c_m)\min(\left(\frac{\lambda_j}{\lambda_k}\right)^{1/2}, \left(\frac{\lambda_k}{\lambda_j}\right)^{1/2}), (24+32c_m)\min(\left(\frac{\lambda_j}{\lambda_l}\right)^{1/2}, \left(\frac{\lambda_l}{\lambda_j}\right)^{1/2})] \\
&= \varepsilon\tau_{m-1}(24+32c_m)[\bar{v}_k(j), \bar{v}_k(l)]
\end{aligned}$$

and

$$|I_2| \le \chi_{m-1}(1+c_m)\varepsilon[\bar{v}_k(j), \bar{v}_k(l)]$$

Taken together, we get

$$\chi_m = (1+c_m)\chi_{m-1} + \varepsilon\tau_{m-1}(24+32c_m) \quad , \quad \chi_{-1} = 0 \tag{3.24}$$

In the second case, $x > \bar{x}$, we get a similar bound with a possibly different c_m. Again, we take the maximum of the two. This completes the third part of the proof.

Finally, combining (3.23) and (3.16) we get

$$|v_i(j) - u_i(j)| \leq (\rho_0 + \chi_{M-1}) \frac{(tol + \varepsilon) \cdot \bar{\kappa} \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})}$$

proving the theorem with $p(M, n) = \rho_0 + \chi_{M-1}$. ∎

4 One-sided Jacobi

In this section we prove that one-sided Jacobi in floating point arithmetic applied to a general matrix computes the singular values and singular vectors with the error bounds of section 2. Here we present our algorithm; the model of arithmetic was presented in section 3. In subsection 4.1 we derive error bounds for the computed singular values. In subsection 4.2 we derive error bounds for the computed singular vectors. In subsection 4.3, we present two algorithms for the symmetric positive definite eigenproblem H, both of which involve applying one-sided Jacobi to the Cholesky factor L^T (or L) of H. The second of these algorithms cannot compute eigenvectors quite as accurately as the first, but may be much faster than either the first algorithm or two-sided Jacobi.

Let $G_0 = B_0 D_0$ be the initial matrix, and $G_m = B_m D_m$, where G_m is obtained from G_{m-1} by applying a single Jacobi rotation. Here D_m is diagonal and B_m has columns of unit norm. All the error bounds in this section contain the factor $\max_m \kappa(B_m)$, whereas the perturbation bounds in section 2 are proportional to $\kappa(B_0)$. Therefore, as in section 3, our claim that Jacobi computes the SVD as accurately as predicted in section 2 depends on the ratio $\max_m \kappa(B_m)/\kappa(B_0)$ being modest. In exact arithmetic, one-sided Jacobi on $G = BD$ is identical to two-sided Jacobi on $H = G^T G = DB^T BD = DAD$, so the question of the growth of $\kappa(B_m) = \kappa(A_m)^{1/2}$ is essentially identical to the question of the growth of $\kappa(A_m)$ in the case of two-sided Jacobi.

The essential difference between our algorithm and standard one-sided Jacobi is the stopping criterion: according to Theorem 2.17, we must stop when all $H_{ij}/(H_{ii}H_{jj})^{1/2}$ are small ($H = G^T G$), not just when $H_{ij}/\max_{kl}|H_{kl}|$ is small. This stopping criterion has been suggested before [20, 5, 3], but without our explanation of its benefits. Otherwise, our algorithm is based on the standard one introduced by Rutishauser [14]. We have chosen a simple version of the algorithm, omitting enhancements like delayed updates of the diagonals and fast rotations, to make the error analysis clearer (an error analysis of these enhancements is future work).

Algorithm 4.1 *One-sided Jacobi for the singular value problem. tol is a use defined stopping criterion. The matrix V whose columns are the computed right singular vectors initially contains the identity.*

repeat
    for all pairs $i < j$
        /* *compute* $\begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv$ *the* (i,j) *submatrix of* $G^T G$ */
        $a = \sum_{k=1}^{n} G_{ki}^2$
        $b = \sum_{k=1}^{n} G_{kj}^2$
        $c = \sum_{k=1}^{n} G_{ki} * G_{kj}$
        /* *compute the Jacobi rotation which diagonalizes* $\begin{bmatrix} a & c \\ c & b \end{bmatrix}$ */
        $\zeta = (b - a)/(2c)$
        $t = sign(\zeta)/(|\zeta| + \sqrt{1 + \zeta^2})$
        $cs = 1/\sqrt{1 + t^2}$

```
            sn = cs * t
            /* update columns i and j of G */
            for k = 1 to n
                 tmp = G_ki
                 G_ki = cs * tmp − sn * G_kj
                 G_kj = sn * tmp + cs * G_kj
            endfor
            /* update the matrix V of right singular vectors */
            for k = 1 to n
                 tmp = V_ki
                 V_ki = cs * tmp − sn * V_kj
                 V_kj = sn * tmp + cs * V_kj
            endfor
       endfor
  until convergence (all |c|/√(ab) ≤ tol)
  /* the computed singular values are the norms of the columns of the final G */
  /* the computed left singular vectors are the normalized columns of the final G */
```

4.1 Error Bounds for Singular Values Computed by One-sided Jacobi

The next theorem and its corollary justify our accuracy claims for singular values computed by one-sided Jacobi.

Theorem 4.1 *Let G_m be the sequence of matrices generated by the one-sided Jacobi algorithm in finite precision arithmetic with precision ε; that is G_{m+1} is obtained from G_m by applying a single Jacobi rotation. Then the following diagram*

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

commutes in the following sense: The top arrow indicates that G_{m+1} is obtained from G_m by applying one Jacobi rotation in floating point arithmetic. The diagonal arrow indicates that G_{m+1} is obtained from G'_m by applying one plane rotation in exact arithmetic; thus G_{m+1} and G'_m have identical singular values and left singular vectors. The vertical arrow indicates that $G'_m = G_m + \delta G_m$. δG_m is bounded as follows. Write $\delta G_m = \delta B_m D_m$, where D_m is diagonal such that B_m in $G_m = B_m D_m$ has unit columns. Then

$$\|\delta B_m\|_2 \le 72\varepsilon \tag{4.2}$$

In other words, one step of Jacobi satisfies the assumptions needed the error bounds of section 2.

Corollary 4.3 *Assume Algorithm 4.1 converges, and that G_M is the final matrix which satisfies the stopping criterion. For $0 \le m \le M$ write $G_m = B_m D_m$ with D_m diagonal and*

B_m with unit columns. Let σ_j be the j-th singular value of G_0 and σ'_j the j-th computed singular value. Then to first order in ε the following error bound holds:

$$\frac{|\sigma_j - \sigma'_j|}{\sigma_j} \le (72\varepsilon \cdot M + n^2\varepsilon + n \cdot tol) \cdot \max_{0 \le k \le M} \kappa(B_k) + n\varepsilon \tag{4.4}$$

PROOF OF COROLLARY 4.3. Bound (4.4) follows by substituting the bound (4.2) and the stopping criterion into Theorem 2.17. The $n^2\varepsilon$ term comes from the fact that $c/\sqrt{ab}$ in the stopping criterion may be underestimated by as much as $n\varepsilon$. The trailing $n\varepsilon$ comes from computing the norms of the columns of the final G matrix. ∎

Remark. A similar bound can be obtained based on the error bound in Proposition 2.19.

PROOF OF THEOREM 4.1. The proof of the commuting diagram is a tedious computation. Let a_T, b_T and c_T be the true values of $\sum_k G_{ki}^2$, $\sum_k G_{kj}^2$ and $\sum_k G_{ki}G_{kj}$. Then (in the notation of the proof of Theorem 3.2) write $a_T = d_i^2$, $b_T = d_j^2$ and $c_T = zd_id_j$. We may assume without loss of generality that $a_T \ge b_T$ and $c_T > 0$. By positive definiteness $0 < z \le \bar{z} \equiv (\kappa^2(B_m) - 1)/(\kappa^2(B_m) + 1) < 1$. Let $x \equiv d_j/d_i \le 1$. We consider two cases, $x \le \bar{x} \equiv (\sqrt{5} - 1)/2 \approx .62$, and $x > \bar{x}$.

First consider $x \le \bar{x}$. Systematic application of formulas (3.1) shows that

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

Let $\tilde{cs} \equiv (1+t^2)^{-1/2}$ and $\tilde{sn} \equiv t(1+t^2)^{-1/2}$. Then from (3.1) again we get $sn = (1+\varepsilon_{sn})\tilde{sn}$ and $cs = (1+\varepsilon_{cs})\tilde{cs}$ where $|\varepsilon_{sn}| \le 4\varepsilon$ and $|\varepsilon_{cs}| \le 3\varepsilon$. $\tilde{cs}$ and $\tilde{sn}$ define the plane rotation $J_m = \begin{bmatrix} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs} \end{bmatrix}$ which take G'_m to G_{m+1}: $G'_mJ_m = G_{m+1}$. Also, we can show $t \le (1 + O(\varepsilon))x(z + n\varepsilon)/(1 - \bar{x}^2)$ and so $|\tilde{sn}| \le (1 + O(\varepsilon))x(z + n\varepsilon)/(1 - \bar{x}^2)$.

Let G'_{ki} and G'_{kj} be the new values for these entries computed by the algorithm. Using the bounds $|\tilde{cs}| \le 1$ and $|\tilde{sn}| \le (1 + O(\varepsilon))x(z + n\varepsilon)/(1 - \bar{x}^2)$, we estimate

$$\begin{aligned} G'_{ki} &= fl(cs * G_{ki} - sn * G_{kj}) \\ &= (1+\varepsilon_1)(1+\varepsilon_2)csG_{ki} - (1+\varepsilon_3)(1+\varepsilon_4)snG_{kj} \\ &= (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_{cs})\tilde{cs}G_{ki} - (1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_{sn})\tilde{sn}G_{kj} \\ &\equiv \tilde{cs}G_{ki} - \tilde{sn}G_{kj} + E_{ki} \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} G'_{kj} &= fl(sn * G_{ki} + cs * G_{kj}) \\ &= (1+\varepsilon_5)(1+\varepsilon_6)snG_{ki} + (1+\varepsilon_7)(1+\varepsilon_8)csG_{kj} \\ &= (1+\varepsilon_5)(1+\varepsilon_6)(1+\varepsilon_{sn})\tilde{sn}G_{ki} + (1+\varepsilon_7)(1+\varepsilon_8)(1+\varepsilon_{cs})\tilde{cs}G_{kj} \\ &\equiv \tilde{sn}G_{ki} + \tilde{cs}G_{kj} + E_{kj} \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} \|E_{\cdot i}\|_2 &\le 5\varepsilon\|G_{\cdot i}\|_2 + 6\varepsilon\|G_{\cdot j}\|_2 \le 11\varepsilon d_i \\ \|E_{\cdot j}\|_2 &\le 6\varepsilon\tilde{sn}\|G_{\cdot i}\|_2 + 5\varepsilon\|G_{\cdot j}\|_2 \le \varepsilon(\frac{6(z+n\varepsilon)}{1-\bar{x}^2} + 5)d_j \end{aligned}$$

(here $G_{\cdot i}$ refers to the i-th column of G, etc.). Thus

$$\begin{aligned}
\left[\begin{array}{cc} G'_{\cdot i} & G'_{\cdot j} \end{array}\right] &= \left[\begin{array}{cc} G_{\cdot i} & G_{\cdot j} \end{array}\right] \cdot \left[\begin{array}{cc} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs} \end{array}\right] + \left[\begin{array}{cc} E_{\cdot i} & E_{\cdot j} \end{array}\right] \\
&= \left(\left[\begin{array}{cc} G_{\cdot i} & G_{\cdot j} \end{array}\right] + \left[\begin{array}{cc} E_{\cdot i} & E_{\cdot j} \end{array}\right] \cdot \left[\begin{array}{cc} \tilde{cs} & -\tilde{sn} \\ \tilde{sn} & \tilde{cs} \end{array}\right] \right) \left[\begin{array}{cc} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs} \end{array}\right] \\
&\equiv \left(\left[\begin{array}{cc} G_{\cdot i} & G_{\cdot j} \end{array}\right] + \left[\begin{array}{cc} F_{\cdot i} & F_{\cdot j} \end{array}\right] \right) \left[\begin{array}{cc} \tilde{cs} & \tilde{sn} \\ -\tilde{sn} & \tilde{cs} \end{array}\right]
\end{aligned}$$

where

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\leq \|E_{\cdot i}\|_2 + \|E_{\cdot j}\|_2 \leq \varepsilon\left(\frac{6(z+n\varepsilon)}{1-\bar{x}^2} + 16\right) d_i \\
\|F_{\cdot j}\|_2 &\leq |\tilde{sn}| \cdot \|E_{\cdot i}\|_2 + \|E_{\cdot j}\|_2 \leq \varepsilon\left(\frac{17(z+n\varepsilon)}{1-\bar{x}^2} + 5\right) d_j
\end{aligned}$$

Thus

$$\|\delta B_m\|_2 \leq \frac{\|F_{\cdot i}\|_2}{d_i} + \frac{\|F_{\cdot j}\|_2}{d_j} \leq \varepsilon\left(\frac{23(z+n\varepsilon)}{1-\bar{x}^2} + 21\right) \tag{4.7}$$

Now consider the case $x > \bar{x}$. The analysis differs from the previous one only in the fact that sn is no longer small. Using the bounds $|\tilde{sn}| \leq 1$, $|\tilde{cs}| \leq 1$ in (4.5) and (4.6) yields

$$\begin{aligned}
\|E_{\cdot i}\|_2 &\leq 5\varepsilon\|G_{\cdot i}\|_2 + 6\varepsilon\|G_{\cdot j}\|_2 \leq 11\varepsilon d_i \\
\|E_{\cdot j}\|_2 &\leq 6\varepsilon\|G_{\cdot i}\|_2 + 5\varepsilon\|G_{\cdot j}\|_2 \leq 11\varepsilon d_i
\end{aligned}$$

whence

$$\begin{aligned}
\|F_{\cdot i}\|_2 &\leq \|E_{\cdot i}\|_2 + \|E_{\cdot j}\|_2 \leq 22\varepsilon d_i \\
\|F_{\cdot j}\|_2 &\leq \|E_{\cdot i}\|_2 + \|E_{\cdot j}\|_2 \leq 22\varepsilon d_j/\bar{x}
\end{aligned}$$

and

$$\|\delta B_m\|_2 \leq \frac{\|F_{\cdot i}\|_2}{d_i} + \frac{\|F_{\cdot j}\|_2}{d_j} \leq 44\varepsilon/\bar{x} \tag{4.8}$$

Since $\bar{x}$ satisfies $1/(1-\bar{x}^2) = 1/\bar{x} < 1.62$, we see from (4.7) and (4.8) that in both cases

$$\|\delta B_m\|_2 \leq 72\varepsilon$$

proving the theorem. ∎

4.2 Error Bounds for Singular Vectors Computed by One-sided Jacobi

The next two theorems justify our accuracy claims for singular vectors computed by one-sided Jacobi.

Theorem 4.9 *Let $V = [v_1, \cdots, v_n]$ be the matrix of unit right singular vectors and $U = [u_1, \cdots, u_n]$ be the matrix of unit left singular vectors computed by Algorithm 4.1 in finite precision arithmetic with precision ε. Let $V_T = [v_{T1}, \cdots, v_{Tn}]$ and $U_T = [u_{T1}, \cdots, u_{Tn}]$ be* *the matrices of true unit right and left singular vectors, respectively. Let* $\bar{\kappa} \equiv \max_m \kappa(B_m)$ *be the largest* $\kappa(B_m)$ *of any iterate. Then the error in the computed singular vectors is bounded in norm by*

$$\max(\|u_{Ti} - u_i\|_2, \|v_{Ti} - v_i\|_2) \le \frac{(n-.5)^{1/2} \cdot \bar{\kappa} \cdot (72M \cdot \varepsilon + n \cdot tol + n^2 \cdot \varepsilon)}{relgap_{\sigma i}} + (9M + n + 1)\varepsilon \tag{4.10}$$

PROOF. The proof is similar to that of Theorem 3.11. Let $G_0, \ldots, G_M$ be the sequence of matrices generated by the Jacobi algorithm, where G_M satisfies the stopping criterion. Let J_m be the exact plane rotation which transforms G'_m to G_{m+1} in the commuting diagram of Theorem 4.1: $G'_m J_m = G_{m+1}$.

We will use the approximation that $relgap_{\sigma i}$ is the same for all G_m, even though it changes slightly. This contributes an $O(\varepsilon^2)$ term (which we ignore), but could be accounted for using the bounds of Theorem 4.1.

First we consider the left singular vectors. In exact arithmetic, these remain unchanged throughout the computation since all rotations are applied on the right. Thus, we need only plug the bounds for the stopping criterion and each $\|\delta B_m\|_2$ from Theorem 4.1 into Theorem 2.21 to get

$$\|u_{Ti} - u_i\|_2 \le \frac{(n-.5)^{1/2} \cdot \bar{\kappa} \cdot (72M \cdot \varepsilon + n \cdot tol + n^2 \cdot \varepsilon)}{relgap_{\sigma i}} + (n+1)\varepsilon$$

as claimed (the $(n+1)\varepsilon$ term comes from normalizing the i-th column of G_M at the end of the computation).

Now we consider the right singular vectors. First we will compute error bounds for the columns of $J_0 \cdots J_{M-1}$ ignoring any rounding errors occurring in computing their product. Then we will incorporate these rounding errors.

We will prove by induction that the i-th column v_{mi} of $V_m \equiv J_m \cdots J_{M-1}$ is a good approximation to the true i-th right singular vector v_{Tmi} of G_m. In particular, we wll show that to first order in ε

$$\|v_{Ti} - v_{0i}\|_2 \le \frac{(n-.5)^{1/2} \cdot \bar{\kappa} \cdot (72M \cdot \varepsilon + n \cdot tol + n^2 \cdot \varepsilon)}{relgap_{\sigma i}}$$

The basis of the induction is as follows. $V_M = I$ the the singular vector matrix for G_M, which is considered to have orthogonal columns since it passes the stopping criterion. Thus the norm error in v_{Mi} follows from plugging the stopping criterion $n \cdot tol$ (increased by $n^2\varepsilon$ since $n \cdot tol$ may be underestimated by this amount) into Theorem 2.21:

$$\|v_{TMi} - v_{Mi}\|_2 \le \frac{(n-.5)^{1/2} \cdot \bar{\kappa} \cdot (n \cdot tol + n^2 \cdot \varepsilon)}{relgap_{\sigma i}}$$

For the induction step we assume that

$$\|v_{T,m+1,i} - v_{m+1,i}\|_2 \le \frac{(n-.5)^{1/2} \cdot \bar{\kappa} \cdot (72(M-m-1) \cdot \varepsilon + n \cdot tol + n^2 \cdot \varepsilon)}{relgap_{\sigma i}}$$

and try to extend to m. Consider the commuting diagram of Theorem 4.1. Accordingly, the errors in $V_m = J_m V_{m+1}$ considered as right singular vectors of G'_m are just the errors in V_{m+1} premultiplied by J_m. This does not change their norm, since J_m is orthogonal. Now we change G'_m to G_m. This increases the norm error in v_{mi} by an amount bounded by plugging the bound for $\|\delta B_m\|_2$ into Theorem 2.21: $72\varepsilon(n-.5)^{1/2}\bar{\kappa}/relgap_{\sigma i}$. This proves the induction step.

Finally, consider the errors from accumulating the product of slightly wrong values of J_m in floating point arithmetic. From the proof of Theorem 4.1, we see the relative errors in the entries of J_m are at most 4ε, and from the usual error analysis of a product of 2 by 2 rotations, we get $6\sqrt{2}M\varepsilon \le 9M\varepsilon$ for the norm error in the product of M rotations. This completes the proof of the theorem. ∎

Now we consider the errors in the individual components of the computed right singular vectors $|v_{Ti}(j) - v_i(j)|$. From Proposition 2.26 we see that we can hope to bound this quantity by $O(\varepsilon)\bar{\kappa}^2\bar{\bar{v}}_i(j)/relgap_{\sigma i}$, where

$$\bar{\bar{v}}_i(j) \equiv \bar{\kappa}^3 \min(\frac{\sigma_i}{\sigma_j}, \frac{\sigma_j}{\sigma_i}) \tag{4.11}$$

is a modified upper bound for the right singular vector component $v_i(j)$ as in Proposition 2.25. In other words, we may have high relative accuracy even in the tiny components of the computed right singular vectors. Our proof of this fact will not be as satisfactory as the previous result, because it will contain a "pivot growth" factor which probably grows at most linearly in M but for which we can only prove an exponential bound. In numerical experiments presented in section 7, there was no evidence that this factor grew with increasing n or M.

We will use $\bar{\bar{v}}_i(j)$ as defined in (4.11) for each G_m, even though the values of σ_i and σ_j vary slightly from step to step. This error will contribute an $O(\varepsilon^2)$ term to the overall bound (which we ignore) but could be incorporated using the bounds of Corollary 4.3.

Theorem 4.12 *Let V, V_T and $\bar{\kappa}$ be as in Theorem 4.9, and $\bar{\bar{v}}_i(j)$ be as in (4.11). Then we can bound the error in the individual components of v_i by*

$$|v_{Ti}(j) - v_i(j)| \le q(M,n) \cdot \frac{(tol+\varepsilon)\cdot\bar{\kappa}^2\cdot\bar{\bar{v}}_i(j)}{relgap_{\sigma i}} \tag{4.13}$$

PROOF. The proof is nearly identical to that of Theorem 3.14; we just outline the differences here. Let J_m be the exact plane rotation whic transforms G'_m to G_{m+1} in the commuting diagram of Theorem 4.1: $G'_m J_m = G_{m+1}$. Let $V_m = J_m \cdots J_{M-1}$ and $\tilde{V}_m = J_0 \cdots J_m$. In the first part of the proof we show the columns of V_0 have small componentwise errors in the sense of the theorem. In the second part we will show each (i,j) entry of each $\tilde{V}_m$ is bounded by a modest multiple of $\bar{\bar{v}}_i(j)$. In the third part we show the rounding errors committed in computing $\tilde{V}_m$ in floating point are componentwise small compared to $\bar{\bar{v}}_i(j)$.

For the first part, the same induction argument as in Theorem 3.14 leads to the bound

$$|v_{Tmi}(j) - v_{mi}(j)| \le \rho'_m \frac{(tol+n\varepsilon)\cdot\bar{\kappa}^2\cdot\bar{\bar{v}}_i(j)}{relgap_{\sigma i}}$$

where

$$\rho'_m = (1 + c'_m)\rho'_{m+1} + \frac{144\sqrt{n-1}}{n} \ , \ \rho'_M = 2n\sqrt{n-1}$$

Here c'_m is a small constant as in Theorem 3.14. We use Proposition 2.26 and Theorem 4.1 in place of Proposition 2.12 and Theorem 3.2 which were used in Theorem 3.14.

For the second part, let $\tilde{V}_m = [\tilde{v}_{m1}, \ldots, \tilde{v}_{mn}]$. The same induction argument as in Theorem 3.14 yields

$$|\tilde{v}_{mi}(j)| \le \tau'_m \bar{\bar{v}}_i(j)$$

where

$$\tau'_m = (1 + c'_m)\tau'_{m-1} \ , \ \tau'_{-1} = 1$$

For the third part, let $\tilde{\tilde{V}}_m = fl(\tilde{\tilde{V}}_{m-1} * \tilde{J}_m)$ be the actually computed singular vector matrix after the m-th rotation. Write $\tilde{\tilde{V}}_m = [\tilde{\tilde{v}}_{m1}, \ldots, \tilde{\tilde{v}}_{mn}]$. The same induction argument as in Theorem 3.14 yields

$$|\tilde{\tilde{v}}_{mi}(j) - \tilde{v}_{mi}(j)| \le \chi'_m \varepsilon \bar{\bar{v}}_i(j)$$

with

$$\chi'_m = (1 + c'_m)\chi'_{m-1} + \varepsilon\tau'_{m-1}(5 + 6c'_m) \ , \ \chi'_{-1} = 0$$

Altogether, we get

$$|v_{Ti}(j) - v_i(j)| \le (\rho'_0 + \chi'_{M-1}) \frac{(tol + n\varepsilon) \cdot \bar{\kappa}^2 \cdot \bar{\bar{v}}_i(j)}{relgap_{\sigma i}}$$

proving the theorem with $q(M, n) = \rho'_0 + \chi'_{M-1}$. ∎

4.3 Using Cholesky Followed by One-sided Jacobi for the Symmetric Positive Definite Eigenproblem

In this subsection we consider two algorithms for the symmetric positive definite eigenproblem H, both based on performing Cholesky on H, and using one-sided Jacobi to compute the SVD of L. The first algorithm (Algorithm 4.2) does one-sided Jacobi on L^T, returning its right singular vectors as the eigenvectors of H and the squares of its singular values as the eigenvalues of H. The second algorithm (Algorithm 4.4), originally proposed in [20], does Cholesky with complete pivoting (which is equivalent to diagonal pivoting) and then one-sided Jacobi on L, returning its left singular vectors as the eigenvectors of H and the squares of its singular values as the eigenvalues of H. The second algorithm, which we call accelerated one-sided Jacobi, is less accurate than the first because it will not always compute tiny eigenvector components with the accuracy of Theorem 3.14, although it does compute the eigenvalues as accurately, and the eigenvectors with the same norm error bound. However, it can be several times faster than either the first algorithm or two-sided Jacobi. In fact, the larger the range of numbers on the diagonal of D, the faster the second algorithm will converge. This means that the more the guaranteed accuracy of the algorithm exceeds that of QR (or any tridiagonalization based algorithm), the faster it converges.

Algorithm 4.2 *One-sided Jacobi method for the symmetric positive definite eigenproblem H.*

1. *Form the Cholesky factor L of H: $H = LL^T$.*

2. *Compute the singular values σ_i and right singular vectors v_i of L^T using one-sided Jacobi.*

3. *The eigenvalues λ_i of H are $\lambda_i = \sigma_s^2$. The eigenvectors of H are v_i.*

We show this method is as accurate as using two-sided Jacobi directly on H. The proof involves a new error analysis of Cholesky decomposition, so we begin by restating Cholesky's algorithm in order to establish notation for our error analysis:

Algorithm 4.3 *Cholesky decomposition $H = LL^T$ for an n by n symmetric positive definite matrix H.*

for $i = 1$ *to* n
  $L_{ii} = (H_{ii} - \sum_{k=1}^{i-1} L_{ik}^2)^{1/2}$
  for $j = i+1$ *to* n
    $L_{ji} = (H_{ji} - \sum_{k=1}^{i-1} L_{jk}L_{ik})/L_{ii}$
  endfor
endfor

Lemma 4.14 *Let L be the Cholesky factor of H computed using Algorithm 4.3 in finite precision arithmetic with precision ε. Then $LL^T = H+E$ where $|E_{ij}| \le (n+5)\varepsilon(H_{ii}H_{jj})^{1/2}$.*

Proof. Applying rules (3.1) for floating point arithmetic yields

$$L_{ii} = (1+\varepsilon_1)((1+\varepsilon_2)H_{ii} - \sum_{k=1}^{i-1} L_{ik}^2(1+i\varepsilon_{k+2}))^{1/2}$$

Since $\sum_{k=1}^{i-1} L_{ik}^2 \le H_{ii}$ to within a small relative error, we may write

$$H_{ii} = \sum_{k=1}^{i} L_{ik}^2 + (i+5)\varepsilon_{ii}H_{ii}$$

as desired. Next we have

$$L_{ij} = (1+\varepsilon_1)((1+\varepsilon_2)H_{ji} - \sum_{k=1}^{i-1} L_{jk}L_{ik}(1+i\varepsilon_{k+2})$$

since $|\sum_{k=1}^{i-1} L_{jk}L_{ik}| \le (H_{ii}H_{jj})^{1/2}$ to within a small relative error by the Cauchy-Schwartz inequality, we can write

$$H_{ji} = \sum_{k=1}^{i} L_{ik}L_{jk} + (i+4)\varepsilon_{ji}(H_{ii}H_{jj})^{1/2}$$

proving the result. ∎

Theorem 4.15 *Let L be the Cholesky factor of $H = DAD$ computed in floating point arithmetic using Algorithm 4.3. Let σ_i and v_{Li} be the exact singular values and right singular vectors of L^T, and λ_i and v_{Hi} be the eigenvalues and eigenvectors of H. Let $\bar{v}_i(j)$ be as in Proposition 2.11. Then*

$$\frac{|\lambda_i - \sigma_i^2|}{\lambda_i} \le (n^2 + 5n) \cdot \varepsilon \cdot \kappa(A)$$

$$\|v_{Li} - v_{Hi}\|_2 \le \frac{(n^2 + 5n)(n-1)^{1/2} \cdot \varepsilon \cdot \kappa(A)}{relgap_{\lambda_i}} + O(\varepsilon^2)$$

$$|v_{Li}(j) - v_{Hi}(j)| \le \frac{(n^2 + 5n)(2n-2)^{1/2} \cdot \varepsilon \cdot \kappa(A) \cdot \bar{v}_i(j)}{\min(relgap_{\lambda_i}, 2^{-1/2})} + O(\varepsilon^2)$$

PROOF. Plug the bound of Lemma 4.14 into Theorem 2.3, Theorem 2.7 and Proposition 2.12. ▮

Theorem 4.15 implies that the errors introduced by Cholesky are as small as those introduced by two-sided Jacobi. Write $H = DAD$ and $L_A = D^{-1}L$. Since $\|A - L_A L_A^T\|_2 \le (n^2 + 5n)\varepsilon$, $\kappa(A) \approx (\kappa(L_A))^2$ (unless both are very large). Since the columns of L_A^T have nearly unit norm, the accuracy of one-sided Jacobi applied to L^T is governed by $\kappa(L_A)$. Thus, Cholesky followed by one-sided Jacobi results in a problem whose condition number $\kappa(L_A)$ is approximately the square root of the condition number of the original problem $\kappa(A)$. Corollary 4.3 and Theorems 4.9 and 4.12 guarantee that the computed eigenvalues and eigenvectors are accurate. In exact arithmetic one-sided Jacobi on L^T is the same as two-sided Jacobi on $DAD = H = LL^T = D(L_A L_A^T)D$, so the question of how much $\kappa(L_A)$ can grow during subsequent Jacobi rotations is essentially identical to the question of the growth of $\kappa(A_m)$ during two-sided Jacobi.

Here is the second algorithm:

Algorithm 4.4 *Accelerated one-sided Jacobi method for the symmetric positive definite eigenproblem H.*

1. *Form the Cholesky factor L of H using complete pivoting. Then there is a permutation matrix P such that $P^T HP = LL^T$.*
2. *Compute the singular values σ_i and left singular vectors u_i of L using one-sided Jacobi.*
3. *The eigenvalues λ_i of H are $\lambda_i = \sigma_s^2$. The eigenvectors of H are Pu_i.*

Even if we did not do complete pivoting, Theorem 4.15 would guarantee that the squares of the true singular values of L would be accurate eigenvalues of H, and that the true left singular vectors of L would be accurate eigenvectors of $P^T HP$. Since we are computing left singular vectors of L, Theorem 4.12 does not apply, but from Corollary 4.3 we know the computed eigenvalues are accurate, and from Theorem 4.9 we know the computed eigenvectors are accurate in a norm sense. Numerical experiments in section 7 below bear out the fact that tiny eigenvector components may not always be computed as accurately by Algorithm 4.4 as Algorithm 4.2.

The advantage of complete pivoting is accelerated convergence. This is because the algorithm is in principle doing two-sided Jacobi on $L^T L$, so writing $L^T L = H' = D'A'D'$, it is $\kappa(A')$ which the algorithm must drive to 1, not $\kappa(A)$. The wider the range of numbers on the diagonal of D, the smaller $\kappa(A')$ will be. We discuss this in more detail in section 6, and content ourselves here with a small example. Let

$$A = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \quad , \quad D = \mathrm{diag}\ (1, d) \quad \text{and} \quad H = \begin{bmatrix} 1 & da \\ da & d^2 \end{bmatrix}$$

where $0 \le a < 1$, and $0 < d < 1$ (any 2 by 2 symmetric positive definite H can be scaled and permuted to be in this form). Here $\kappa(A) = (1+a)/(1-a)$, which can be made as large as desired by choosing a near 1. The matrix is already ordered for complete pivoting, and the Cholesky factor is

$$L = \begin{bmatrix} 1 & 0 \\ da & d(1-a^2)^{1/2} \end{bmatrix}$$

so $L^T L = H' = D'A'D'$ where

$$A' = \begin{bmatrix} 1 & (1+d^2a^2)^{-1/2} \\ (1+d^2a^2)^{-1/2} & 1 \end{bmatrix}$$

and $\kappa(A') \le (3+\sqrt(5))/2 \approx 2.62$ *independent of* H. In section 6 we will prove in general that $\kappa(A')$ is bounded independent of H.

5 Bisection and Inverse Iteration

Here we show bisection and inverse iteration applied to the symmetric positive definite matrix $H = DAD$ can compute the eigenvalues and eigenvectors within the accuracy bounds section of 2. Let inertia(H) denote the triple $(neg, zero, pos)$ of the number n of negative eigenvalues of H, the number z of zero eigenvalues of H, and the number p of positive eigenvalues of H. These results are extensions of Algorithms 3 and 5 in [2].

Algorithm 5.1 *Stably computing the inertia of $H - xI = DAD - xI$.*

1. *Permute the rows and columns of $A - xD^{-2}$ (which has the same inertia as $H - xI$) and partition it as*
$$\begin{bmatrix} A_{11} - xD_1^{-2} & A_{12} \\ A_{21} & A_{22} - xD_2^{-2} \end{bmatrix}$$
so that if $1 - xd^{-2}$ is a diagonal entry of $A_{11} - xD_1^{-2}$, then $xd^{-2} \geq 2n+1$, where n is the dimension of H.

2. *Compute $X = A_{22} - xD_2^{-2} - A_{21}(A_{11} - xD_1^{-2})^{-1}A_{12}$, using Cholesky to compute $(A_{11} - xD_1^{-2})^{-1}A_{12}$.*

3. *Compute* inertia$(X) = (neg, zero, pos)$ *using a stable pivoting scheme such as in [4].*

4. *The inertia of $H - xI$ is $(neg + \dim(A_{11}), zero, pos)$.*

We need to partition $A - xD^{-2}$ as above in order to make the proof convenient but it may not be necessary algorithmically.

The proof of correctness requires the following

Lemma 5.1 *Let $H = D'A'D'$ be positive definite, and let $Hx = b$ be solved by Cholesky to get an approximate solution $\hat{x}$. We do not assume A' has a unit diagonal. Let ε be the machine precision, and assume no overflow nor underflow occurs. Then to first order in ε,*

$$\|x - \hat{x}\|_2 \leq O(\varepsilon \|A'\|_2 \cdot \|A'^{-1}\|_2^2 \cdot \|D'^{-1}\|_2^2 \cdot \|b\|_2)$$

Proof. We begin by defining some convenient notation. Let $\tilde{H}$ be defined by $\tilde{H}_{ij} = (H_{ii}H_{jj})^{1/2}$. Let $|E|$ denote the matrix of absolute values of entries of E, and let inequalities like $X \leq Y$ between matrices be interpreted componentwise. Then Lemma 4.14 of the last section says that if L is the computed Cholesky factor of H, then $LL^T = H + E$ where $|E| \leq (n+5)\varepsilon\tilde{H}$. Note also by Cauchy-Schwartz that $|L| \cdot |L^T| \leq \tilde{H}$.

We begin by proving that $D'(A'+F)D'\hat{x} = b$ where $\|F\|_2 = O(\varepsilon)\|A'\|_2$. In solving $Ly = b$ with forward substitution, we actually get $(L + \delta L_1)\hat{y} = b$, where $|\delta L_{1,ij}| \leq n\varepsilon|L_{ij}|$ [22]. In solving $L^T x = \hat{y}$ we actually get $(L + \delta L_2)^T\hat{x} = \hat{y}$ where $|\delta L_{2,ij}| \leq n\varepsilon|L_{ij}|$. Altogether

$$(H + E + \delta L_1 L^T + L\delta L_2^T + \delta L_1 \delta L_2^T)\hat{x} \equiv D'(A' + F)D'\hat{x} = b$$

where

$$
\begin{aligned}
\|F\|_2 &= \|D'^{-1}(E + \delta L_1 L^T + L\delta L_2^T + \delta L_1 \delta L_2^T)D'^{-1}\|_2 \\
&\le \|D'^{-1}ED'^{-1}\|_2 + \|D'^{-1}|\delta L_1| \cdot |L^T| D'^{-1}\|_2 + \|D'^{-1}|L| \cdot |\delta L_2^T| D'^{-1}\|_2 + \|D'^{-1}|\delta L_1| \cdot |\delta L_2^T| D'^{-1}\|_2 \\
&\le (n+5)\varepsilon \|D'^{-1}\tilde{H}D'^{-1}\|_2 + n\varepsilon \|D'^{-1}\tilde{H}D'^{-1}\|_2 + n\varepsilon \|D'^{-1}\tilde{H}D'^{-1}\|_2 + n^2\varepsilon^2 \|D'^{-1}\tilde{H}D'^{-1}\|_2 \\
&\le (3n^2 + 5n + n^3\varepsilon)\varepsilon \|A'\|_2
\end{aligned}
$$

Thus

$$
\begin{aligned}
\|x - \hat{x}\|_2 &= \|D'^{-1}A'^{-1}F(A'+F)^{-1}D'^{-1}b\|_2 \le \|D'^{-1}\|_2^2 \cdot \|F\|_2 \cdot \|A'^{-1}\|_2^2 \cdot \|b\|_2 \\
&\le O(\epsilon)\|D'^{-1}\|_2^2 \cdot \|A'\|_2 \cdot \|A'^{-1}\|_2^2 \cdot \|b\|_2
\end{aligned}
$$

to first order in ε. ∎

Theorem 5.2 *Let ε be the machine precision in which Algorithm 5.1 is carried out, where we assume neither overflow nor underflow occur. Then Algorithm 5.1 computes the exact inertia of $D(A + \delta A)D - xI$, where $\|\delta A\|_2 = O(\varepsilon)$. Thus, Algorithm 5.1 can be used in a bisection algorithm to find all the eigenvalues of H to the accuracy of Theorem 2.3 or Proposition 2.5.*

PROOF. X is defined so that

$$
\begin{aligned}
&\begin{bmatrix} A_{11} - xD_1^{-2} & A_{12} \\ A_{21} & A_{22} - xD_2^{-2} \end{bmatrix} \\
&\quad = \begin{bmatrix} I & 0 \\ A_{21}(A_{11} - xD_1^{-2})^{-1} & I \end{bmatrix} \cdot \begin{bmatrix} A_{11} - xD_1^{-2} & 0 \\ 0 & X \end{bmatrix} \cdot \begin{bmatrix} I & (A_{11} - xD_1^{-2})^{-1}A_{12} \\ 0 & I \end{bmatrix}
\end{aligned}
$$

so that the inertia of $H - xI$ equals

$$
\mathrm{inertia}(A - xD^{-2}) = \mathrm{inertia}(X) + \mathrm{inertia}(A_{11} - xD_1^{-2}) = \mathrm{inertia}(X) + (\dim(A_{11}), 0, 0)
$$

by Sylvester's Theorem and the fact that $A_{11} - xD_1^{-2}$ is negative definite. The algorithm in [4] will compute the exact inertia of $X + \delta X$, where $\|\delta X\|_2 = O(\varepsilon)\|X\|_2$. Thus if we show $\|X\|_2 = O(1)$ and that X can be computed with error $O(\varepsilon)$, we will be done. By construction the diagonal entries of $A_{11} - xD_1^{-2}$ are less than or equal to $-2n$, and the offdiagonal entries of all of $A - xD^{-2}$ are bounded by 1 in absolute value. Write $-(A_{11} - xD_1^{-2}) = H_1 = D_1 A_1 D_1$ where $A_{1ii} = 1$. Then $\|A_1\|_2 \le 3/2$, $\|A_1^{-1}\|_2 \le 2$ and $\|D_1^{-1}\|_2^2 \le 1/(2n)$. By Lemma 5.1 the error in computing $(A_{11} - xD_1^{-2})^{-1}A_{12}$ is bounded by $O(\varepsilon)$. Also, by construction $\|A_{12}\|_2 = \|A_{21}\|_2 \le n$, $\|(A_{11} - xD_1^{-2})^{-1}\|_2 \le 1/n$ and $\|A_{22} - xD_2^{-2}\|_2 \le 3n + 1$, so $\|X\|_2 \le 3n + 1 + n^2/n = 4n + 1 = O(1)$ as desired. ∎

Algorithm 5.2 *Inverse iteration for computing the eigenvector x of a symmetric positive definite matrix $H = DAD$ corresponding to eigenvalue z. tol is a user-specified stopping criterion.*

1. *We assume the eigenvalue z has been computed accurately, for example using Algorithm 5.1.*

2. *Choose a starting vector* y_0*; set* $i = 0$.

3. *Compute the symmetric indefinite factorization* LDL^T *of* $P(A - zD^{-2})P^T$ *[4], where* P *is the same permutation as in Algorithm 5.1, step 1.*

4. *Repeat*

   $i = i + 1$

   Solve $(A - zD^{-2})\tilde{y}_i = y_{i-1}$ *for* $\tilde{y}_i$ *using the* LDL^T *factorization of step 3.*

   $r = 1/\|\tilde{y}_i\|_2$

   $y_i = r \cdot \tilde{y}_i$

   until $(r \leq tol)$)

5. $x = D^{-1}y_i$

Theorem 5.3 *Suppose Algorithm 5.2 terminates with* x *as the computed eigenvector of* $H = DAD$*. Then there is a diagonal matrix* $\hat{D}$ *with* $\hat{D}_{ii} = 1 + O(tol)$ *and a matrix* δA *with* $\|\delta A\|_2 = O(tol)$*, such that* $\hat{D}x$ *is the exact eigenvector of* $D(A + \delta A)D$*. Thus, the error in* x *is bounded by Theorem 2.7, Corollary 2.9 and Proposition 2.12.*

The proof is identical to the proof of Theorem 11 in [2].

6 Upper Bounds for $\max_m \kappa(A_m)/\kappa(A_0)$

As stated in sections 3 and 4, our claims about the accuracy to which Jacobi can solve the eigenproblem depend on the ratio $\max_m \kappa(A_m)/\kappa(A_0)$ being modest. Here $H_0 = D_0 A_0 D_0$ is the initial matrix, and $H_m = D_m A_m D_m$ is the sequence produced by Jacobi (H_{m+1} is obtained from H_m by applying a single Jacobi rotation, D_m is diagonal and A_m has ones on the diagonal). The reason is that the error bounds for Jacobi are proportional to $\max_m \kappa(A_m)$, and the error bounds of section 2 are proportional to $\kappa(A_0)$.

In this section we present several results explaining why $\max_m \kappa(A_m)/\kappa(A_0)$ should not be expected to grow very much. Recall that convergence of H_m to diagonal form is equivalent to the convergence of A_m to the identity matrix, or of $\kappa(A_m)$ to 1. Thus we expect $\kappa(A_m) < \kappa(A_0)$ eventually. The best situation would be monotonic convergence, but this is unfortunately not always the case.

We have not been able to completely explain the extremely good numerical results of section 7, that $\max_m \kappa(A_m)/\kappa(A_0)$ never exceeded 1.82, and averaged 1.20 in random experiments. (Wang [21] has found a sequence H_n of matrices of dimension n where this ratio grows slowly with n, reaching 8 for $n = 50$. Changing the sweep strategy eliminated this growth.) A complete theoretical explanation of this remains an open question.

We will only speak in terms of two-sided Jacobi in this section. This is no loss of generality because in exact arithmetic one-sided Jacobi on G is equivalent to two-sided Jacobi on $G^T G$.

Our first result will show that $\kappa(A_m)/\kappa(A_0)$ cannot be too large if A_m is obtained from A_0 by a sequence of Jacobi rotations in pairwise disjoint rows and columns. The second result give a cheaply computable guaranteed upper bound on $\max_m \kappa(A_m)/\kappa(A_0)$ in terms of the Hadamard measure of A_0. This bound is generally quite pessimistic unless the dimension of A is modest and $\kappa(A_0)$ is small, at most a few hundred. The third and fourth results will be for accelerated one-sided Jacobi (Algorithm 4.4). The third result shows that the wider the range of numbers on the diagonal of H, the smaller $\kappa(A_1)$ for that algorithm. This in turn makes it converge faster. In other words, the more the guaranteed accuracy of the algorithm exceeds that of QR (or any tridiagonalization based algorithm), the faster it converges. The fourth rather surprising result that $\kappa(A_1)$ is bounded by a constant depending *only* on the dimension n, not on A_0. These last two results lead us to recommend accelerated one-sided Jacobi as the algorithm of choice (unless it is important to get small eigenvector components to high accuracy; see the discussion in subsection 4.3).

Proposition 6.1 *Let H_0 be n by n. Let H_m be obtained from H_0 by applying m Jacobi rotations in pairwise nonoverlapping rows and columns (this means $m \le n/2$). Write $H_m = D_m A_m D_m$ as before. Then*

$$\frac{\kappa(A_m)}{\kappa(A_0)} \le \frac{1 + \max_{1\le k\le m} |A_{0,2k-1,2k}|}{1 - \max_{1\le k\le m} |A_{0,2k-1,2k}|} \le \min(\kappa(A_0), 2n) \tag{6.2}$$

Also

$$\frac{\kappa(A_{i+1})}{\kappa(A_i)} \le \min(\kappa(A_0), 8) \tag{6.3}$$

Furthermore, the spectrum of A_m is independent of D_0, even though the entries of A_m depend on D_0. More precisely, the spectrum of A_m coincides with the spectrum of the pencil $A_0 - \lambda A_0'$, where A_0' coincides with A_0 on every rotated element and is the identity otherwise.

PROOF. We begin by deriving a matrix pencil depending only on A_0 whose eigenvalues are the same as A_m. This will prove that the eigenvalues of A_m depend only on A_0. We assume without loss of generality that the m Jacobi rotations are in rows and columns (1,2), (3,4), ..., $(2m-1, 2m)$. This lets us write $J^T H_0 J = H_m$ where J is block diagonal with the 2 by 2 Jacobi rotations (and possibly ones) on its diagonal. Rewrite this as

$$A_m = (D_m^{-1} J^T D_0) A_0 (D_0 J D_m^{-1}) \equiv Z^T A_0 Z$$

where Z has the same block diagonal structure as J. Let A_0' be a block diagonal matrix with the same block structure as Z and J, where A_0' is identical to A_0 within its 2 by 2 blocks, and has ones on its diagonal when J does. Since $H_{m,12} = H_{m,34} = \cdots = 0$, also $A_{m,12} = A_{m,23} = \cdots = 0$. Thus A_m has 2 by 2 identity matrices on its diagonal matching the block structure of Z, J and A_0'. Thus $A_m = Z^T A_0 Z$ implies $Z^{-T} Z^{-1} = A_0'$. Therefore the eigenvalues of $A_m = Z^T A_0 Z$ are identical to those of the pencil $A_0 - \lambda Z^{-T} Z^{-1} = A_0 - \lambda A_0'$.

Now we apply the minimax theorem to bound $\lambda_{\min}(A_m)$ below by

$$\lambda_{\min}(A_m) = \min_{x \neq 0} \frac{x^T A_0 x}{x^T A_0' x} \geq \frac{\min_{||x||=1} x^T A_0 x}{\max_{||x||=1} x^T A_0' x} = \frac{\lambda_{\min}(A_0)}{1 + \max_{1 \leq k \leq m} |A_{0,2k-1,2k}|} \tag{6.4}$$

We may bound $1 + \max_{1 \leq k \leq m} |A_{0,2k-1,2k}|$ from above by both $\lambda_{\max}(A_0)$ and 2, yielding

$$\lambda_{\min}(A_m) \geq \frac{\lambda_{\min}(A_0)}{\min(2, \lambda_{\max}(A_0))} \tag{6.5}$$

Now we bound $\lambda_{\max}(A_m)$ from above. First by the minimax theorem we may write

$$\lambda_{\max}(A_m) = \max_{x \neq 0} \frac{x^T A_0 x}{x^T A_0' x} \leq \frac{\max_{||x||=1} x^T A_0 x}{\min_{||x||=1} x^T A_0' x} \leq \frac{\lambda_{\max}(A_0)}{1 - \max_{1 \leq k \leq m} |A_{0,2k-1,2k}|}$$

which when combined with (6.5) yields

$$\kappa(A_m) \leq (\kappa(A_0))^2$$

proving half of (6.2). For the other half note that $1 \leq \lambda_{\max}(A_i) \leq n$ for all i, so that $\lambda_{\max}(A_m)/\lambda_{\max}(A_0) \leq n$. Now combine this with (6.5).

Now we show $\lambda_{\max}(A_{i+1}) \leq 4\lambda_{\max}(A_i)$, which when combined with (6.5) yields (6.3). It suffices to show $\lambda_{\max}(A_1) \leq 4\lambda_{\max}(A_0)$. Write

$$A_0 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is 2 by 2. Then by the minimax theorem there exists a conformally partitioned unit vector $x^T = [x_1^T, x_2^T]$ where

$$\lambda_{\max}(A_1) = \frac{x_1^T A_{11} x_1 + 2x_1^T A_{12} x_2 + x_2^T A_{22} x_2}{x_1^T A_{11} x_1 + x_2^T x_2}$$

Write $x_1^T x_1 = \zeta$ ($0 \le \zeta \le 1$), $x_2^T x_2 = 1-\zeta$, $x_1^T A_{11} x_1 = \tau_1 \zeta$ and $x_2^T A_{22} x_2 = \tau_2(1-\zeta)$, so that

$$\begin{aligned}
\lambda_{\max}(A_1) &= \frac{\tau_1\zeta + 2x_1^T A_{12} x_2 + \tau_2(1-\zeta)}{\tau_1\zeta + 1 - \zeta} \\
&\le 2\frac{\tau_1\zeta + \tau_2(1-\zeta)}{\tau_1\zeta + 1 - \zeta}
\end{aligned}$$

The maximum of this last expression over all $0 \le \zeta \le 1$ is

$$2 + 2\tau_2 \le 2 + 2\lambda_{\max}(A_{22}) \le 4\lambda_{\max}(A_0)$$

∎

Our second bound is based on the Hadamard measure of a symmetric positive definite matrix H:

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

Proposition 6.6 *The Hadamard measure $\mathcal{H}(H)$ has the following properties:*

1. $\mathcal{H}(H) \le 1$ *and* $\mathcal{H}(H) = 1$ *if and only if* H *is diagonal.*
2. $\mathcal{H}(H) = \mathcal{H}(\tilde{D} H \tilde{D})$ *for any nonsingular diagonal* $\tilde{D}$.
3. *Let* $H = DAD$ *with* D *diagonal and* A *with unit diagonal. Then*

$$\lambda_{\min}(A) \ge \frac{\mathcal{H}(H)}{e} = \frac{\det(A)}{e}$$

   where $e = \exp(1)$.

4. *Let* H' *be obtained from* H *by applying a Jacobi rotation (in exact arithmetic) in rows and columns* i *and* j. *Then*

$$\mathcal{H}(H') = \frac{\mathcal{H}(H)}{1 - A_{ij}^2} \ge \mathcal{H}(H)$$

5. *Let* $H_0, \ldots, H_m, \ldots$ *be a sequence of symmetric positive definite matrices obtained from Jacobi's method in exact arithmetic. Let* $H_m = D_m A_m D_m$ *with* D_m *diagonal and* A_m *with unit diagonal. Then*

$$\max_m \kappa(A_m) \le \frac{n \cdot e}{\det(A_0)} = \frac{n \cdot e}{\mathcal{H}(H_0)}$$

PROOF.
1. Write the Cholesky decomposition $H = LL^T$. Then

$$H_{11} \cdots H_{nn} = \prod_{i=1}^{n} \left(\sum_{k=1}^{i} L_{ik}^2\right) \ge \prod_{i=1}^{n} L_{ii}^2 = \det(H)$$

2. $\det(\tilde{D}^2)$ factors out of the numerator and denominator of $\mathcal{H}(\tilde{D}H\tilde{D})$.
3. From (2.) $\mathcal{H}(H) = \mathcal{H}(A) = \det(A)$, so it suffices to show $\lambda_{\min}(A) \geq \det(A)/e$. Let $0 < \lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of A. Since $\lambda_1 = \det(A)/\prod_{i=2}^n \lambda_i$, we need to show $\prod_{i=2}^n \lambda_i \leq e$. Now $\sum_{i=2}^n \lambda_i \leq \mathrm{tr}(A) = n$. Since $ab \geq (a+x)(b-x)$ for all $a \geq b \geq x \geq 0$, we see $\prod_{i=2}^n \lambda_i$ is greatest when all $\lambda_i = n/(n-1)$, in which case $\prod_{i=2}^n \lambda_i = ((n-1)/n)^{n-1} \leq e$.
4. From Proposition 6.1 we have

$$\mathcal{H}(H') = \det(AA'^{-1}) = \det(A)/(1 - A_{ij}^2) = \mathcal{H}(H)/(1 - A_{ij}^2)$$

where $A' = I$ except for $A'_{ij} = A'_{ji} = A_{ij}$.
5. This is directly implied by (3.) and (4.). ∎

Thus, Part 5 of this proposition gives us a guaranteed upper bound on $\max_m \kappa(A_m)$ at a cost of about $n^3/6$ flops, compared to $2n^3$ flops per Jacobi sweep ($4n^3$ if accumulating eigenvectors). If we use the algorithm in subsection 4.3, where we must do Cholesky anyway, this upper bound comes nearly for free.

Basically, this upper bound is only useful as long as $\kappa(A_0)$ is quite small and A_0 has low dimension; otherwise it is much too large to be useful.

Our third and fourth bounds are for accelerated one-sided Jacobi (Algorithm 4.4). Recall that this algorithm begins by doing Cholesky with complete pivoting on H_0 to get $PH_0P^T = LL^T$, where P is a permutation matrix. Then it does one-sided Jacobi on L, which is equivalent (in exact arithmetic) to two-sided Jacobi on L^TL. Therefore, Algorithm 4.4 essentially starts with $L^TL = H_1 = D_1A_1D_1$. As mentioned in [20] the transition from H_0 to H_1 is, in fact, one step of the symmetric LR algorithm which usually has some non-trivial diagonalizing effect (the pivoting cares for the proper ordering). This effect will be more pronounced with growing $\kappa(H_0)$. Quite analogous effects are present if a Jacobi-SVD algorithm is preceded by the QR decomposition with column pivoting [10].

Our third result, which we state rather informally, is that the larger the range of numbers on the diagonal D^2 of H, the smaller is $\kappa(A_1)$ (this effect was also observed in [20]). We argue as follows. Let $L = DL_A$ be the factor obtained from complete pivoting. Here, L_A has rows of unit norm. Since Algorithm 4.4 does one-sided Jacobi on L, its performance depends on the condition number of DL_AD', where D' is chosen diagonal to make the columns of DL_AD' unit vectors. From van der Sluis's theorem [16] we know the condition number of DL_AD' can be at most n times DL_AD^{-1}, so it suffices to examine $\kappa(DL_AD^{-1})$. The effect of complete pivoting is essentially to reorder D so that $D_{ii} \geq D_{i+1,i+1}$, and to keep $L_{A,ii}$ as large as possible. Now $(DL_AD^{-1})_{ii} = L_{A,ii}$ is unchanged, and the subdiagonal entry $(DL_AD^{-1})_{ij} = L_{A,ij}D_{ii}D_{jj}^{-1}$ is multiplied by the factor $D_{ii}D_{jj}^{-1}$ which is between 0 and 1. The more D_{jj} exceeds D_{ii}, the smaller this factor, and the more nearly diagonal DL_AD^{-1} becomes. Since complete pivoting tries to keep the diagonal of L_A large, this improves the condition number.

Our fourth result shows that surprisingly $\max_{m\geq 1} \kappa(A_m)$ is bounded independent of H_0:

Proposition 6.7 *Let $PH_0P^T = LL^T$ be the Cholesky decomposition of the n by n matrix H_0 obtained with complete pivoting. Let $H_1 = L^TL = D_1A_1D_1$. Let $H_m = D_mA_mD_m$, $m > 1$, be obtained from two-sided Jacobi applied to H_1. Then*

1. *$\mathcal{H}(H_1) \geq \mathcal{H}(H_0)$.*

2. $\mathcal{H}(H_1) \geq 1/n!$. *This bound is attainable.*

3. $\max_{m\geq 1} \kappa(A_m) \leq e \cdot n/\mathcal{H}(H_1) \leq e \cdot n \cdot n!$

PROOF.
1. Since $\det(H_1) = \det(H_0)$, it suffices to show $\prod_i H_{0,ii} \geq \prod_i H_{1,ii}$. Assume without loss of generality that $P = I$. Then $H_{0,ii} = \sum_{k=1}^{i} L_{ik}^2$ and $H_{1,ii} = \sum_{k=i}^{n} L_{ki}^2$. Complete pivoting is equivalent to the fact that $L_{ii}^2 \geq \sum_{k=i}^{j} L_{jk}^2$ for all $j > i$. We wish to prove $\prod_{i=1}^{n} \sum_{k=1}^{i} L_{ik}^2 \geq \prod_{i=1}^{n} \sum_{k=i}^{n} L_{ki}^2$. We systematically use the fact that $ab \geq (a+x)(b-x)$ for $a \geq b \geq x \geq 0$. We illustrate the general procedure in the case of $n = 3$:

$$\begin{aligned}
(L_{11}^2)(L_{21}^2 + L_{22}^2)(L_{31}^2 + L_{32}^2 + L_{33}^2) &\geq (L_{11}^2 + L_{21}^2)(L_{22}^2)(L_{31}^2 + L_{32}^2 + L_{33}^2) \\
&\geq (L_{11}^2 + L_{21}^2 + L_{31}^2)(L_{22}^2)(L_{32}^2 + L_{33}^2) \\
&\geq (L_{11}^2 + L_{21}^2 + L_{31}^2)(L_{22}^2 + L_{32}^2)(L_{33}^2)
\end{aligned}$$

2. We have

$$\mathcal{H}(H_1) = \frac{\det(L)^2}{\prod_{i=1}^{n}(L^T L)_{ii}} = \frac{\prod_{i=1}^{n} L_{ii}^2}{\prod_{i=1}^{n}(\sum_{k=i}^{n} L_{ki}^2)} = \prod_{i=1}^{n} \frac{L_{ii}^2}{\sum_{k=i}^{n} L_{ki}^2} \geq \prod_{i=1}^{n} \frac{1}{i} = \frac{1}{n!}$$

To see that this bound is attainable, let $H = LL^T$ where $L_{ii} = \mu^{(i-1)/2}$ and $L_{ij} = (1-\mu)^{1/2}\mu^{(i-1)/2}$. Now let $\mu > 0$ become small.
3. The result follows from part 2 and Proposition 6.6, part 5. ∎

The example in part 2 of the Proposition for which the Hadamard bound is attainable unfortunately has the property that the resulting upper bound in part 3 is a gross overestimate. While the upper bound grows as $e \cdot n \cdot n!$, $\kappa(A_1)$ only grows like $n^{3/2}$. However, $\kappa(A_0)$ grows like $\mu^{-n/2}$, which can be arbitrarily larger than the bound in part 3. The choice $\mu = .5$ provides an example where the upper bound in part 3 can arbitrarily exceed both $\kappa(A_0)$ and $\max_{m\geq 1} \kappa(A_m)$ for large n.

Nonetheless, in numerical experiments the upper bound $e \cdot n/\mathcal{H}(H_1)$ on $\max_{m\geq 1} \kappa(A_m)$ never exceeded 40. We also always observed that $\kappa(A_1) \leq \kappa(A_0)$ in all cases, although we have not been able to prove it in general.

Recently Slapničar [15] has improved the $e \cdot n \cdot n!$ bound to $O(4^n)$ and shown that this improved bound is attainable; see also related results in [11].

7 Numerical Experiments

In this section we present the results of numerical experiments. Briefly, we tested every error bound of every algorithm presented in this paper, and verified that they held in all examples. In fact, the performance is better than we were able to explain theoretically, both because we could observe little or no growth in actual errors for increasing dimension, and because of the surprisingly small values attained by $\max_m \kappa(A_m)/\kappa(A_0)$ (see section 6).

These tests were performed using FORTRAN on a SUN 4/260. The arithmetic was IEEE standard double precision [1], with a machine precision of $\varepsilon = 2^{-53} \approx 10^{-16}$ and over/underflow threshold $10^{\pm 308}$.

There were essentially four algorithms tested: two-sided Jacobi (Algorithm 3.1), one-sided Jacobi (Algorithms 4.2 and 4.1), accelerated one-sided Jacobi (Algorithms 4.4 and 4.1), and bisection/inverse iteration (Algorithms 5.1 and 5.2). All were used with the stopping criterion $tol = 10^{-14}$.

Since we claim these algorithms are more accurate than any other, we tested their accuracy as follows. We considered only symmetric positive definite eigenproblems, and solved every one using every algorithm. The different answers were compared to see if they agreed to the predicted accuracy (which they did). They were also compared to the EISPACK routines tred2/tql2 [17], which implement tridiagonalization followed by QR iteration. Small eigenvalues computed by EISPACK were often negative, indicating total loss of relative accuracy.

For example, the matrix

$$H = \begin{bmatrix} 10^{40} & 10^{19} & 10^{19} \\ 10^{19} & 10^{20} & 10^{9} \\ 10^{19} & 10^{9} & 1 \end{bmatrix}$$

has all its eigenvalues computed to high relative accuracy by Jacobi, whereas QR computes at least one negative or zero eigenvalue, no matter how the rows and columns are ordered. This shows that QR cannot be made to deliver high relative accuracy on appropriately graded matrices, as suggested in [17].

The rest of this section is organized as follows: Subsection 7.1 discusses test matrix generation. Subsection 7.2 discusses the accuracy of the computed eigenvalues. Subsection 7.3 discusses the accuracy of the computed eigenvectors. Subsection 7.4 discusses the the growth of $\max_m \kappa(A_m)/\kappa(A_0)$. Subsection 7.5 discusses convergence rates; here the speed advantage of accelerated one-sided Jacobi will be apparent.

7.1 Test Matrix Generation

We generated several categories of random test matrices according to three parameters: the dimension n, κ_A, and κ_D. First we describe the algorithm used to generate a random matrix from these parameters, and then the sets of parameters used.

We tested matrices of dimension $n = 4$, 8, 16 and 50. Since testing involved solving an n by n eigenproblem after each Jacobi rotation (to evaluate $\kappa(A_m)$) and there are $O(n^2)$ Jacobi rotations required for convergence, testing costs $O(n^5)$ operations per matrix.

Given κ_A, we generated a random symmetric positive definite matrix with unit diagonal and approximate condition number κ_A as follows. We began by generating a diagonal matrix T with diagonal entries in a geometric series from 1 down to $1/\kappa_A$. Then we generated an orthogonal matrix U uniformly distributed with respect to Haar measure [18], and formed UTU^T. Finally, we computed another diagonal matrix K so that $A_0 = KUTU^TK$ had unit diagonal. This last transformation can decrease the condition number of UTU^T, but usually not by much. For 4 by 4 matrices, it decreased it by as much as a factor of 500, for 8 by 8 matrices by a factor of 20, for 16 by 16 matrices by a factor of 5 and for 50 by 50 matrices by a factor of 1.5. (This decreasing variability is at least partly due to the fact that we ran fewer tests on the larger matrices.) For a more complete discussion of the test matrix generation software, see [8].

Given κ_D, we generated a random diagonal matrix D_0 with diagonal entries whose logarithms were uniformly distributed between 0 and $\log \kappa_D$. This means the diagonal entries themselves were distributed from 1 to κ_D. The uniform distribution of the logarithm essentially means every decade is equally likely, and so generates matrices D_0 with entries of widely varying magnitudes.

The resulting random matrix was then $H_0 = D_0 A_0 D_0$.

We generated random matrices with 5 possible different values of κ_A: 10, 10^2, 10^4, 10^8 and 10^{12}, 6 possible different values of κ_D: 10^5, 10^{10}, 10^{20}, 10^{30}, 10^{50} and 10^{100}, and 4 different dimensions $n = 4$, 8, 16 and 50. This makes a total of $5 \times 6 \times 4 = 120$ different classes of matrices. In each class of dimension $n = 4$ matrices, we generated 100 random matrices, in each class of $n = 8$, we generated 50 random matrices, in each class of $n = 16$, we generated 10 random matrices, and in each class of $n = 50$, we generated one random matrix. This makes a total of 4830 different test matrices.

The matrices had in some cases eigenvalues ranging over 200 orders of magnitude (when $\kappa_D = 10^{100}$). The relative gaps $relgap_\lambda$ ranged from .028 to $2 \cdot 10^{42}$.

7.2 Accuracy of the Computed Eigenvalues

There are two accuracy bounds for eigenvalues from section 2 which we tested. The first one is based on Theorem 2.3 (or Theorem 2.17 together with Theorem 4.15), which says that if λ_i' and λ_i'' are approximations of λ_i computed by two of our algorithms, then

$$Q_1 \equiv \frac{|\lambda_i' - \lambda_i''|}{\kappa(A_0)\lambda_i'}$$

should be $O(tol)$, where $tol = 10^{-14}$ is our stopping criterion. For two-sided Jacobi and one-sided Jacobi, Q_1 never exceeded $2 \cdot 10^{-15}$. For two-sided Jacobi and accelerated one-sided Jacobi, Q_1 also never exceeded $2 \cdot 10^{-15}$. Every matrix had an eigenvalue for which Q_1 exceeded $4 \cdot 10^{-18}$, showing that the bound of Theorem 2.3 is attainable, as predicted by Proposition 2.13.

In the case of bisection, we did not run a bisection algorithm to convergence for each eigenvalue, but rather took the eigenvalues λ_i' computed by two-sided Jacobi, made intervals $[(1 - tol \cdot \kappa(A_0))\lambda_i', (1 + tol \cdot \kappa(A_0))\lambda_i']$ from each one, and used bisection to verify that each interval contained one eigenvalue (overlapping intervals were merged and the counting modified in the obvious way). All intervals successfully passed this test.

The second accuracy bound is from Proposition 2.5 (or Proposition 2.19 together with Theorem 4.15) which predicts that

$$Q_2 \equiv \frac{|\lambda_i' - \lambda_i''|}{\|D_0 v_i\|_2^2}$$

should be $O(tol)$. Here v_i is the unit eigenvector computed by two-sided Jacobi. For two-sided Jacobi and one-sided Jacobi, Q_2 never exceeded $2 \cdot 10^{-14}$. For two-sided Jacobi and accelerated one-sided Jacobi, Q_2 never exceeded $9 \cdot 10^{-15}$. Every matrix had an eigenvalue for which Q_2 exceeded $5 \cdot 10^{-16}$, showing that the bound of Proposition 2.5 is attainable, as it predicts.

In the case of bisection, we again made intervals $[\lambda_i' - tol \cdot \|D_0 v_i\|_2^2, \lambda_i' + tol \cdot \|D_0 v_i\|_2^2]$ from each eigenvalue λ_i' and verified that each interval contained the proper number of eigenvalues.

Finally, we verified a slightly weakened version of Proposition 2.10, that

$$\lambda_{\min}(A_0) - tol \le \frac{\lambda_i'}{h_i} \le \lambda_{\max}(A_0) + tol$$

for the eigenvalues λ_i' computed by two-sided Jacobi. Here h_i is the i-th smallest diagonal entry of H_0. Adding and subtracting tol to the upper and lower bounds takes into account the errors in computing λ_i'.

7.3 Accuracy of the Computed Eigenvectors

There is one bound on the magnitude of the components of the eigenvectors, and two accuracy bounds, one for the norm error and one for the componentwise error.

We begin with a few details about our implementation of inverse iteration. We used the eigenvalues computed by two-sided Jacobi, and the vector of all ones as a starting vector. Convergence always occurred after just one iteration.

The componentwise bound on the magnitude of the eigenvectors is based on Proposition 2.11, which says that the components of the normalized eigenvector v_i should be bounded by

$$|v_i(j)| \le \bar{v}_i(j) \equiv (\kappa(A_0))^{3/2} \cdot \min(\left(\frac{\lambda_i}{\lambda_j}\right)^{1/2}, \left(\frac{\lambda_j}{\lambda_i}\right)^{1/2})$$

This was verified for the eigenvectors computed by all four algorithms. We note that since this bound is proportional to $\kappa(A_0)^{3/2}$, it becomes weaker as $\kappa(A_0)$ becomes larger, and indeed becomes vacuous for matrices with $\kappa(A_0)$ large and eigenvalues in a narrow range.

The norm error bounds are based on Theorem 2.7 (or Theorem 2.21 together with Theorem 4.15), which predicts that if v_i' and v_i'' are approximations of the unit eigenvector v_i computed by two of our algorithms, then

$$Q_3 \equiv \frac{\|v_i' - v_i''\|_2}{(\kappa(A_0)/relgap_{\lambda_i}) + 1}$$

should be $O(tol)$. (We add the 1 in the denominator because a single roundoff error in the largest entry can cause a norm error of ε; see Theorem 3.11 or Theorem 4.9.) For two-sided Jacobi and one-sided Jacobi, Q_3 never exceeded $3 \cdot 10^{-16}$. For two-sided Jacobi and accelerated one-sided Jacobi, Q_3 also never exceeded $2 \cdot 10^{-14}$. For two-sided Jacobi and inverse iteration, Q_3 never exceeded $8 \cdot 10^{-14}$. Every matrix had an eigenvector for which Q_3 exceeded 10^{-18} for every pair of algorithms compared, showing that the bound of Theorem 2.7 is nearly attainable, as predicted by Proposition 2.14.

The second accuracy bound is based on Proposition 2.12 (or Proposition 2.26 and Theorem 4.15), which predicts

$$Q_4 \equiv \frac{|v_i'(j) - v_i''(j)| \min(relgap_{\lambda_i}, 2^{-1/2})}{\kappa(A_0) \cdot \bar{v}_i(j)}$$

should be $O(tol)$. For two-sided Jacobi and one-sided Jacobi, Q_4 never exceeded $3 \cdot 10^{-17}$. For two-sided Jacobi and inverse iteration, Q_4 never exceeded $3 \cdot 10^{-15}$. For two-sided Jacobi and accelerated one-sided Jacobi, Q_4 was as large as .02, which is consistent with the fact that accelerated one-sided Jacobi computes the eigenvectors as left singular vectors of L, for which we only have a normwise error bound (Theorem 4.9). For the other algorithm Q_4 was only 10^{-30} for matrices with $\kappa(A_0) = 10^{12}$; this reflects the factor $\kappa(A_0)^{5/2}$ in the denominator of Q_4, a weakness of Proposition 2.11. In other words, the componentwise error bounds are generally only interesting for small to medium $\kappa(A_0)$.

7.4 Growth of $\max_m \kappa(A_m)/\kappa(A_0)$

In computing

$$Q_5 \equiv \max_m \kappa(A_m)/\kappa(A_0)$$

we note that a single computation requiring M Jacobi rotations supplied us not just with one value of Q_5 but rather $M-1$: Since every A_i can be thought of as starting a new eigenvalue computation, we may also measure $\max_{m \geq i} \kappa(A_m)/\kappa(A_i)$ for all $i < M$. Thus, all told, our 4830 different matrices represent over 900000 data points of Q_5.

The largest value of Q_5 encountered was 1.82. This was for an 8 by 8 matrix with $\kappa(A_0) = 1.4 \cdot 10^{12}$, and eigenvalues ranging over 133 orders of magnitude. 141 Jacobi rotations (a little over 5 sweeps) were required for convergence, plus 28 more steps (one more sweep) where no work is done to recognize convergence. In Figure 1, a plot is shown of $\kappa(A_i) - 1$ versus i. We plot $\kappa(A_i) - 1$ instead of $\kappa(A_i)$ in order to see the quadratic convergence of $\kappa(A_i)$ to 1. The graph appears nearly monotonic, except for a slight rise near $i = 20$. This is seen more clearly in Figure 2, which plots $\max_{m \geq i} \kappa(A_m)/\kappa(A_i)$ versus i. Here the maximal nonmonotonicity of the curve near $i = 20$ is apparent.

Recently Wang [21] found a family of examples where Q_5 was as large as 8 for matrices up to dimension 50. These matrices have 1 on the diagonal and $1-\epsilon$ on the offdiagonal, where ϵ is small. However, by using a different pivoting strategy than cyclic-by-rows, namely the parallel pivoting discussed in Proposition 6.1, this growth could be eliminated.

Now we consider the Hadamard based upper bound on Q_5 from Proposition 6.6:

$$Q_5 \leq Q_6 \equiv \frac{e \cdot n}{\mathcal{H}(H_0) \cdot \kappa(A_0)}$$

Table 1 gives the maximum values of this upper bound for different values of dimension n and $\kappa_A \approx \kappa(A_0)$. Recall that the true value of Q_5 never exceeds 1.82. As Proposition

Figure 1: $\kappa(A_i) - 1$ versus i

Figure 2: $\max_{m \geq i} \kappa(A_m)/\kappa(A_i)$ versus i

Table 1: Hadamard upper bound Q_6 on $\max_m \kappa(A_m)/\kappa(A_0)$					
n	κ_A				
	10	10^2	10^4	10^8	10^{12}
4	5.8	13	590	$6.3 \cdot 10^6$	$6.1 \cdot 10^{10}$
8	21	410	$1.1 \cdot 10^7$	$9.1 \cdot 10^{17}$	∞
16	200	$2.7 \cdot 10^5$	$1.8 \cdot 10^{15}$	∞	∞
50	$6.4 \cdot 10^5$	$8.0 \cdot 10^{16}$	∞	∞	∞

6.6 suggests, this upper bound should not depend on D_0 and indeed the values observed depended very little on D_0.

As can be seen, the Hadamard based bound is of little use except for very small matrices of modest $\kappa(A_0)$. ∞ means the value overflowed.

Now we consider accelerated one-sided Jacobi. Let us recall the notation of section 6: Let $PH_0P^T = LL^T$ be Cholesky with complete pivoting, and let $L^TL = H_1 = D_1A_1D_1$. As suggested in that section, we expect both $\kappa(A_1)$ to be smaller than $\kappa(A_0)$, and the Hadamard based upper bound

$$Q_5 \le Q_7 \equiv \max(1, \frac{e \cdot n}{\mathcal{H}(H_1) \cdot \kappa(A_0)})$$

on Q_5 to be much smaller than the one for two-sided Jacobi.

First of all $\kappa(A_1)/\kappa(A_0)$ never exceeded $\frac{6}{10}$. In fact, $\kappa(A_1)$ *never exceeded 40 for any matrix.* This is quite remarkable. This means that all essential rounding errors occurred during the initial Cholesky decomposition. Finally, the Hadamard upper bound Q_7 on Q_5 never exceeded 29. (Recently, Wang [21] has found an example where $\kappa(A_1)/\kappa(A_0)$ slightly exceeded 1; in his example $\kappa(A_0)$ was close to 1.)

7.5 Convergence Rates

We begin with a few details of how we counted the number of Jacobi rotations required for convergence. In all three algorithms (two-sided Jacobi, one-sided Jacobi and accelerated one-sided Jacobi), we stopped when the last $n(n-1)/2$ stopping tests $|H_{ij}| \cdot (H_{ii}H_{jj})^{-1/2} \le$ *tol* succeeded; this means every offdiagonal entry of H satisfies the stopping criterion. In the case of two-sided Jacobi, this means the last $n(n-1)/2$ Jacobi rotations involved almost no work. For the two one-sided Jacobis, however, evaluating the stopping criterion costs 3 inner products, so the last $n(n-1)/2$ rotation involve a significant amount of work, even if no rotations are performed. This must be kept in mind when comparing the number of rotations for two-sided and one-sided Jacobi.

We used the same standard cyclic pivot sequence for all the algorithms: (1,2), (1,3), ..., (1,n), (2,3), ..., (2,n), (3,4), ..., $(n-1, n)$.

We begin by comparing two-sided Jacobi and one-sided Jacobi. In exact arithmetic, these two algorithms are identical. In practice, they usually took the same number of steps, although one-sided Jacobi did vary from 20% faster to 50% slower than two-sided Jacobi on some examples. From now on we will only compare two-sided Jacobi to accelerated one-sided Jacobi.

Table 2: Average Number of Sweeps for Two-sided Jacobi (TsJ) and Accelerated One-sided Jacobi (AOsJ)									
κ_A	κ_D	Dimension n							
		4		8		16		50	
		TsJ	AOsJ	TsJ	AOsJ	TsJ	AOsJ	TsJ	AOsJ
10	10^5	3.7	3.0	4.9	3.7	5.7	4.4	6.4	5.0
	10^{10}	3.5	2.5	4.6	3.3	5.6	4.1	6.4	5.0
	10^{20}	3.1	2.2	4.5	2.8	5.5	3.6	6.0	4.0
	10^{30}	3.0	2.1	4.6	2.5	5.5	3.4	6.3	4.0
	10^{50}	2.8	1.9	4.4	2.3	5.5	3.1	5.8	4.0
	10^{100}	2.7	1.7	4.5	2.0	5.6	2.6	5.8	3.0
10^2	10^5	3.8	3.0	5.2	3.8	6.4	4.5	7.5	6.0
	10^{10}	3.5	2.5	5.1	3.3	6.2	4.1	7.4	5.0
	10^{20}	3.2	2.2	4.9	2.9	6.2	3.9	7.1	4.0
	10^{30}	3.0	2.0	4.8	2.6	5.8	3.3	6.8	4.1
	10^{50}	2.9	1.9	4.8	2.2	6.1	3.0	6.5	4.0
	10^{100}	2.8	1.6	4.7	2.0	6.0	2.7	6.8	3.4
10^4	10^5	4.0	2.9	5.8	3.6	7.5	4.5	9.2	6.0
	10^{10}	3.7	2.5	5.6	3.3	7.2	4.1	9.3	5.0
	10^{20}	3.2	2.2	5.3	2.9	7.2	3.7	8.5	4.9
	10^{30}	3.1	2.1	5.2	2.6	6.8	3.1	8.2	4.0
	10^{50}	2.9	1.9	5.2	2.4	6.6	3.0	8.5	4.6
	10^{100}	2.7	1.7	4.9	2.2	6.9	2.4	8.0	3.9
10^8	10^5	3.9	2.7	6.4	3.5	9.7	4.1	13.5	6.0
	10^{10}	3.6	2.3	6.3	3.2	9.4	3.8	12.4	5.0
	10^{20}	3.3	2.1	5.7	2.8	8.9	3.5	11.7	4.7
	10^{30}	3.1	2.1	5.5	2.6	8.6	3.4	12.0	4.0
	10^{50}	2.9	1.9	5.3	2.3	8.5	3.1	11.6	4.0
	10^{100}	2.9	1.7	5.1	2.0	8.7	2.6	11.6	4.0
10^{12}	10^5	3.8	2.5	6.8	3.1	10.6	4.0	16.5	6.0
	10^{10}	3.6	2.2	6.4	3.0	10.3	3.9	15.6	5.0
	10^{20}	3.4	2.1	6.0	2.7	9.8	3.5	15.3	5.0
	10^{30}	3.1	2.0	5.8	2.5	10.2	3.3	15.2	4.0
	10^{50}	2.9	1.9	5.6	2.3	9.3	3.2	13.7	3.9
	10^{100}	2.8	1.6	5.2	2.0	8.7	2.7	15.2	3.0

The most interesting phenomenon was the speed up experienced by accelerated one-sided Jacobi with respect to two-sided Jacobi. In Table 2 we present the raw data on the number of sweeps required for convergence.

There are a number of interesting trends exhibited in this table. First, AOsJ (accelerated one-sided Jacobi) never takes more than 6 sweeps to converge for any matrix, whereas TsJ (two-sided Jacobi) takes up to 16.5. In fact AOsJ is almost always faster than TsJ (in one example it took 5% longer), and can be up to 5 times faster (3.0 sweeps vs. 15.2 sweeps for $\kappa_A = 10^{12}$, $\kappa_D = 10^{100}$ and $n = 50$). Second, the number of sweeps increases with increasing κ_A for TsJ, but not for AOsJ. Third, the number of sweeps increases with increasing dimension for both TsJ and AOsJ, but much more modestly for AOsJ (from 2-3 up to 6) than for TsJ (from 3-4 up to 15). Thus, the running time for AOsJ is much less dependent on the problem size or sensitivity (as measured by κ_A) than TsJ. Fourth, the number of sweeps decreases as κ_D increases, both for TsJ and AOsJ, but much more markedly for AOsJ (up to a factor of 2) than for TsJ (usually just 1 sweep).

8 Conclusions

In this paper we have developed new perturbation theory for the eigenvalues and eigenvectors of symmetric positive definite matrices, as well as for eigenvalues of symmetric positive definite pencils. This theory assumes the perturbations are scaled analogous to the way the matrix is scaled, letting us derive much tighter bounds than in the classical theory. In particular, we get relative error bounds for the eigenvalues and individual components of the eigenvectors, which are (nearly) attainable. The bound for symmetric positive definite pencils may be applied to matrices arising in finite element modeling.

Second, we have shown both through formal error analysis and numerical experiment that Jacobi's method (with a proper stopping criterion) computes the eigenvalues and eigenvectors with these error bounds. We also show that bisection and inverse iteration (applied to the original matrix) attain these bounds. In contrast, methods based on tridiagonalization (such as QR, divide and conquer, traditional bisection, etc.) fail to attain these bounds. In particular QR can fail to attain these bounds whether or not preceded by tridiagonalization.

We have similar perturbation theorems for the singular value decomposition of a general matrix and the generalized singular values of a pair of matrices, and similar error analyses and numerical experiments for one-sided Jacobi applied to this problem. We may also use one-sided Jacobi to solve the symmetric positive definite eigenproblem.

We have discussed an accelerated version of Jacobi for the symmetric positive definite eigenproblem, which has the property that the more its accuracy exceeds that of QR (or other conventional algorithms), the faster it converges. However, it cannot compute tiny components of eigenvectors as accurately as the other versions of Jacobi, although it computes the eigenvectors with the same norm error bounds. Unless getting the tiny eigenvector components is important, we recommend this accelerated version of Jacobi for the symmetric positive definite eigenproblem.

The quantity $\max_m \kappa(A_m)/\kappa(A_0)$ was seen to be central in the analysis of Jacobi's accuracy. Numerical experiments show it to be much smaller in practice than we can explain. For the accelerated version of Jacobi we provide an inexpensive estimator of $\max_m \kappa(A_m)/\kappa(A_0)$ which works very well in practice. Explaining the excellent behavior of $\max_m \kappa(A_m)/\kappa(A_0)$ is an important open problem.

The error analyses of Jacobi dealt only with the simplest implementations. It would be worthwhile to extend these analyses to cover various enhancements introduced by Veselić, Hari, Rutishauser and others. These include delayed updates of the diagonal entries and an alternate formula for updating the offdiagonal entries [14, 20], as well as block Jacobi methods.

In future work we plan to extend these results to the symmetric positive definite generalized eigenproblem, as well as indefinite matrices. Any extension requires an appropriate perturbation theory; therefore we do not expect to be able to extend the result to all indefinite matrices, since there is no guaranteed way to compute the zero eigenvalues of a singular matrix to "high relative accuracy" without computing them exactly, a feat requiring high precision arithmetic. A class of indefinite matrices for which a suitable perturbation theory exists are the scaled diagonally dominant matrices [2]. The perturbation theory also already exists (at least for eigenvalues) for the symmetric positive definite generalized eigenproblem.

References

[1] *IEEE Standard for Binary Floating Point Arithmetic.* ANSI/IEEE, New York, Std 754-1985 edition, 1985.

[2] Jesse Barlow and James Demmel. Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices. *SIAM J. Num. Anal.*, 27(3):762–791, June 1990 York, NY, December 1988.

[3] M. Berry and A. Sameh. *Parallel algorithms for the singular value and dense symmetric eigenvalues problems.* Center for Supercomputing Research and Development Report 761, University of Illinois at Urbana-Champaign, Urbana, IL, March 1988. (to appear in J. Comp. and Appl. Math.).

[4] James Bunch and Linda Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Mathematics of Computation*, 31(137):163–179, January 1977.

[5] P. P. M. de Rijk. A one-sided Jacobi algorithm for computing the singular value decomposition on a vector computer. *SIAM J. Sci. Stat. Comp.*, 10(2):359–371, March 1989.

[6] James Demmel. The condition number of equivalence transformations that block diagonalize matrix pencils. *SIAM Journal on Numerical Analysis*, 20(3):599–610, June 1983.

[7] James Demmel and W. Kahan. Accurate Singular Values of Bidiagonal Matrices. *SIAM J. Sci. Stat. Comp.*, 11(5):873–912, September 1990.

[8] James Demmel and Alan McKenney. *A Test Matrix Generation Suite.* Computer Science Dept. Technical Report, Courant Institute, New York, NY, July 1989. (LAPACK Working Note #9).

[9] Gene Golub and Charles Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, 1983.

[10] V. Hari and K. Veselić. On Jacobi methods for singular value decompositions. *SIAM J. Sci. Stat. Comp.*, 8:741–754, 1987.

[11] N. Higham. Analysis of the Cholesky decomposition of a semi-definite matrix. in *Reliable Numerical Computation*, Clarendon Press. eds. M. G. Cox and S. Hammarling, 1990.

[12] T. Kato. *Perturbation Theory for Linear Operators.* Springer Verlag, Berlin, 2 edition, 1980.

[13] J. Le and Beresford Parlett. *On the forward instability of the QR transformation.* Center for Pure and Applied Mathematics PAM-419, University of California, Berkeley, CA, July 1988. submitted to SIAM J. Mat. Anal. Appl.

[14] Beresford Parlett. *The Symmetric Eigenvalue Problem.* Prentice Hall, Englewood Cliffs, NJ, 1980.

[15] Ivan Slapničar. *Upper bound for the condition of the scaled matrix of the symmetric eigenvalue problem.* Fern-Universität Hagen, Lehrgebiet Mathematische Physik, November 1990, Hagen, Germany.

[16] A. Van Der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14:14–23, 1969.

[17] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler. *Matrix Eigensystem Routines – EISPACK Guide.* Volume 6 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, 1976.

[18] G. W. Stewart. On efficient generation of random orthogonal matrices with an application to condition estimation. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

[19] G. W. Stewart. A method for computing the generalized singular value decomposition. in *Matrix Pencils.* Lecture Notes in Mathematics v. 973, eds B. Kågström and A. Ruhe, 1983.

[20] K. Veselić and V. Hari. A Note on a One-Sided Jacobi Algorithm. *Numerische Mathematik*, 56:627–633, 1990.

[21] Xiaofeng Wang. personal communication, 1990.

[22] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, 1965.