Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices

(Appeared in SIAM J. Numer. Anal., v. 27, n. 3, pp. 762-791, 1990)

Jesse Barlow
Department of Computer Science
The Pennsylvania State University
University Park, PA 16802

James Demmel
Courant Institute
251 Mercer Str.
New York, NY 10012

Abstract

When computing eigenvalues of symmetric matrices and singular values of general matrices in finite precision arithmetic we in general only expect to compute them with an error bound proportional to the product of machine precision and the norm of the matrix. In particular, we do not expect to compute tiny eigenvalues and singular values to high relative accuracy. There are some important classes of matrices where we can do much better, including bidiagonal matrices, scaled diagonally dominant matrices, and scaled diagonally dominant definite pencils. These classes include many graded matrices, and all symmetric positive definite matrices which can be consistently ordered (and thus all symmetric positive definite tridiagonal matrices). In particular, the singular values and eigenvalues are determined to high relative precision independent of their magnitudes, and there are algorithms to compute them this accurately. The eigenvectors are also determined more accurately than for general matrices, and may be computed more accurately as well. This work extends results of Kahan and Demmel for bidiagonal and tridiagonal matrices.

Keywords: Graded matrices, Singular value decomposition, symmetric eigenproblem, perturbation theory, error analysis

AMS(MOS) subject classifications: 65F15, 15A60

Acknowledgements: The first author was supported by the Air Force Office of Scientific Research under grant no. AFOSR-88-0161 and the Office of Naval Research under grant no. N00014-80-0517. The second author was supported by the National Science Foundation (grants NSF-DCR-8552474 and NSF-ASC-8715728). Part of this work was performed while the first author was visiting the Courant Institute and the second author was visiting the IBM Bergen Scientific Center.

1. Introduction

When computing the eigenvalues of symmetric matrices and singular values of general matrices in finite precision arithmetic one generally only expects to compute them with an error bound $f(n)\varepsilon\|A\|$, where $f(n)$ is a modestly growing function of the matrix dimension n, ε is the machine precision, and $\|A\|$ is the 2-norm of the matrix A. This follows as a result of standard theorems which state:

(1.1) A perturbation δA in the matrix A cannot change its eigenvalues (singular values) by more than $\|\delta A\|$ [12].

(1.2) The standard algorithm for computing eigenvalues (singular values) of A computes the exact eigenvalues (singular values) of $A+\delta A$, $\|\delta A\| \le f(n)\varepsilon\|A\|$, where $f(n)$ is a modestly growing function of n and ε is the machine precision [12].

These error bounds imply that tiny eigenvalues and singular values (tiny compared to $\|A\|$) cannot generally be computed to high relative accuracy, since the error bound $f(n)\varepsilon\|A\|$ may be much larger than the desired quantity. In fact, if each matrix entry is uncertain in its least significant digits, the tiny eigenvalues and singular values may not even be determined accurately by the data.

Sometimes, however, the eigenvalues and singular values are determined much more accurately than error bounds like $f(n)\varepsilon\|A\|$ would indicate. This was shown for singular values of bidiagonal matrices in [9], where it was proven that small relative perturbations in the bidiagonal entries only cause small relative perturbations in the singular values, independent of their magnitudes. It was also shown how to compute all the singular values to high relative accuracy. In this paper we extend these results to eigenvalues of symmetric scaled diagonally dominant matrices and scaled diagonally dominant definite pencils. (Henceforth we will abbreviate "scaled diagonally dominant" by s.d.d.) A symmetric s.d.d. matrix is any matrix of the form $\Delta A \Delta$, where A is symmetric and diagonally dominant in the usual sense, and Δ is an arbitrary nonsingular diagonal matrix. A pencil $H-\lambda M$ is s.d.d. definite if H and M are symmetric s.d.d. and M is positive definite. Examples of s.d.d. matrices are the "graded" matrices

$$A_0 = \begin{bmatrix} 10 & 10 & & & \\ 10 & 10^2 & 10^2 & & \\ & 10^2 & 10^3 & 10^3 & \\ & & 10^3 & 10^4 & 10^4 \\ & & & 10^4 & 10^5 \end{bmatrix} \quad \text{and} \quad A_1 = \begin{bmatrix} 1 & 10 & & & \\ 10 & -10^3 & 10^4 & & \\ & 10^4 & 10^6 & 10^4 & \\ & & 10^4 & -10^3 & 10 \\ & & & 10 & 1 \end{bmatrix}.$$

Note that A_0 is graded in the usual sense, but not diagonally dominant in the usual sense. A_1 is neither diagonally dominant in the usual sense, nor graded in the usual sense, since the diagonal entries are positive and negative, and not sorted. Thus we see that the usual diagonal dominance implies being s.d.d., but not the converse. In fact, the set of s.d.d. matrices includes all symmetric positive definite matrices which can be consistently ordered, a class which includes all symmetric positive definite tridiagonal matrices. Dense matrices may be s.d.d. as well.

Another example arises 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 s.d.d. definite pencil $H-\lambda M$, where M is the diagonal mass matrix $\mathrm{diag}(m_1, \ldots, m_n)$ and H is the tridiagonal stiffness matrix with diagonal k_0+k_1, k_1+k_2 ,..., $k_{n-1}+k_n$ and offdiagonal $-k_1$,..., $-k_{n-1}$. Note that the matrix $M^{-1/2}HM^{-1/2}$, which has the same eigenvalues as $H-\lambda M$, is symmetric s.d.d.

In particular, we will show that small relative perturbations in the entries of an s.d.d. matrix only cause small relative perturbations in the eigenvalues and singular values, independent of their magnitudes. This is a much tighter perturbation bound than the classical one provided by (1.1) above. Our proof of this result generalizes and unifies results in [9] for bidiagonal matrices alone and in [14] for symmetric tridiagonal s.d.d. matrices alone.

Given that the matrix entries determine all eigenvalues or singular values to high relative accuracy, one would naturally like to compute them that accurately as well. We present algorithms based on bisection which attain this accuracy; in the case of bidiagonal or symmetric positive definite tridiagonal matrices QR iteration (suitably modified) can be shown to attain high accuracy as well. It is not yet known whether algorithms based on divide and conquer [5, 11, 13] can be made to work in some of these situations too.

One may also ask if the singular vectors and eigenvectors of s.d.d. matrices and pencils are determined any more accurately than for general matrices. To state the standard perturbation bound for eigenvectors of symmetric matrices and singular vectors of general matrices, we need to define the *gap*: if λ_i is an eigenvalue (singular value) of A then $gap(\lambda_i) \equiv \min_{j \neq i} |\lambda_i - \lambda_j|$. In other words, it is the absolute distance between λ_i and the remainder of the spectrum.

(1.3) Let y be a unit eigenvector of $A + \delta A$, $\alpha = y^T A y$ the Rayleigh quotient, λ_i the eigenvalue of A closest to α, and z_i its unit eigenvector. Let $\theta(z_i, y)$ be the acute angle between y and z_i. Then $\sin \theta(y, z_i) \leq 4\|\delta A\| / gap(\lambda_i)$ [17, p. 222].

In other words, the error as measured by the angle is proportional to the reciprocal of the gap; if the gap is small (λ_i is in a cluster of eigenvalues), the corresponding eigenvector is poorly determined. As before, the standard algorithms guarantee $\|\delta A\| \leq f(n)\varepsilon\|A\|$, so eigenvectors of eigenvalues poorly separated with respect to $\|A\|$ (i.e. $\|A\| / gap(\lambda_i)$ is large) will generally not be computed accurately. Analogous results hold for singular vectors of general matrices.

For eigenvectors of s.d.d. matrices, a stronger perturbation theorem is true. Briefly, we can replace the gap in (1.3) with the *relative gap*, $\min_{j \neq i} |\lambda_i - \lambda_j| / |\lambda_j \lambda_i|^{1/2}$. Thus, as long as λ_i is *relatively* well separated from its neighbors, its corresponding eigenvector is determined to high relative accuracy. This is a much stronger result than (1.3), as the following example shows. Suppose the eigenvalues are 1, $2 \cdot 10^{-10}$ and 10^{-10}. Then the gap for the smallest eigenvalue is $gap(10^{-10}) = 10^{-10}$, but the relative gap is .707. Thus (1.3) predicts a loss of 10 decimal digits, whereas the finer analysis predicts nearly full accuracy.

We also show that a suitable variation of inverse iteration can be used to compute the eigenvectors to this accuracy. We conjecture that other methods based on divide and conquer can attain this accuracy as well, but this has not been proven.

Similar results can be proven for singular vectors of bidiagonal matrices and eigenvectors of s.d.d. definite pencils; the result for singular vectors partially settles an open question in [9].

The rest of this paper is organized as follows. Section 2 contains definitions. Section 3 discusses some simple generalizations of Gershgorin's theorem applicable to s.d.d. matrices. Section 4 uses the minimax characterization of eigenvalues to present simple perturbation lemmas. In section 5 this lemma is applied to singular values and in section 6 to eigenvalues. Section 7 discusses perturbation theory for eigenvectors and singular vectors. Section 8 shows that the condition numbers for the eigenvectors provide good estimates for the reciprocal of the distance to the nearest matrix with multiple eigenvalues. Section 9 discusses algorithms for the bidiagonal singular value decomposition, section 10 discusses algorithms for the symmetric tridiagonal eigenproblem, and section 11 discusses algorithms for the dense symmetric eigenproblem (both matrices and pencils). The new algorithm for the symmetric positive definite tridiagonal eigenproblem will be included in the LAPACK linear algebra library [8]. Section 12 applies our results to a matrix arising from a differential operator. Section 13 summarizes the available algorithms and the current state of research, and discusses future work.

2. Definitions and Basic Lemmas

In this paper we will deal exclusively with real (usually symmetric) matrices. Extensions to complex (usually Hermitian) matrices will be obvious. $\|\cdot\|$ will denote the 2-norm.

Decompose the matrix A as $A = D + N$ where D is diagonal and N has a zero diagonal. We will call a matrix A *γ-diagonally dominant with respect to a norm* $|||\cdot|||$ if $|||N||| \leq \gamma \min_i |D_{ii}|$,

where $0 \le \gamma < 1$. Suppose that A is γ-diagonally dominant with respect to either the 1-norm or infinity-norm. Then the well known Gershgorin's Theorem says that the eigenvalues of A lie in the union of the Gershgorin disks B_i, where B_i is centered at D_{ii} and has radius at most $\gamma |D_{ii}|$. In particular, if some B_i is disjoint from the other disks, it contains exactly one eigenvalue and D_{ii} is an approximation to this eigenvalue with relative error at most γ.

Now let $A = D + N$ and $|D_{ii}| = 1$ i.e. A has ± 1's on the diagonal. Let Δ_1 and Δ_2 be arbitrary nonsingular diagonal matrices. Then we call $H \equiv \Delta_1 A \Delta_2$ *γ-scaled diagonally dominant (γ-s.d.d.) with respect to a norm* $|||\cdot|||$ if A is γ-diagonally dominant with respect to $|||\cdot|||$. If H is symmetric, we insist that A be symmetric as well in which case $\Delta_1 = \Delta_2$ can be chosen in only one way: $\Delta_{1ii} = |H_{ii}|^{1/2}$. Note that a matrix may only be diagonally dominant in the scaled sense, as the following example shows:

$$A = \begin{bmatrix} 1 & .1 \\ .1 & 1 \end{bmatrix} \quad , \quad \Delta_1 = \Delta_2 = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} \quad , \quad H \equiv \Delta_1 A \Delta_2 = \begin{bmatrix} 1 & 10 \\ 10 & 10000 \end{bmatrix}$$

Here, A is γ-diagonally dominant with $\gamma = .1$ (with respect to the 1-norm, 2-norm or infinity-norm), H is γ-s.d.d. with the same γ, but H is not diagonally dominant in the nonscaled sense for any $\gamma < 1$.

Our definition of scaling implies nothing about the monotonicity of the diagonal entries of H; for example H is γ-s.d.d. if

$$A = \begin{bmatrix} 1 & .1 & .1 \\ .1 & -1 & .1 \\ .1 & .1 & 1 \end{bmatrix} \quad , \quad \Delta_1 = \Delta_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & .01 \end{bmatrix} \quad , \quad H \equiv \Delta_1 A \Delta_2 = \begin{bmatrix} 1 & 10 & .001 \\ 10 & -10000 & .1 \\ .001 & .1 & .0001 \end{bmatrix} .$$

If H is symmetric with positive diagonal entries, being γ-s.d.d. with respect to the 2-norm is closely related to another well known property: consistent ordering [18]. Consistent ordering is defined as follows: Let $A = I + N = I + L + U$, where L is strictly lower triangular and U is strictly upper triangular. Then A is *consistently ordered* if the eigenvalues of $\alpha L + \alpha^{-1} U$ are independent of $\alpha \neq 0$. Now suppose there is a permutation matrix P such that A in $P^T H P = \Delta A \Delta = \Delta(I + N)\Delta$ is consistently ordered. Then we claim H is positive definite if and only if it is s.d.d. To prove this, note that by choosing $\alpha = 1$ and $\alpha = -1$, we see the eigenvalues of $N = L + U$ occur in $\pm$ pairs, including $\pm \|N\| = \pm \gamma$. Now note that H is positive definite if and only if $I + N$ is positive definite (by Sylvester's theorem), and that the smallest eigenvalue of $I + N$ is $1 - \|N\| = 1 - \gamma$. Therefore, the theorems in this paper apply to many matrices arising from discretized differential equations [18]; see section 12 for an example.

We will call a symmetric pencil $H - \lambda M$ *γ-scaled diagonal dominant definite (γ-s.d.d. definite) with respect to a norm* $|||\cdot|||$ if H and M are γ-s.d.d. symmetric with respect to $|||\cdot|||$ and M is positive definite. If H is positive definite as well, we call $H - \lambda M$ *γ-s.d.d. positive definite.*

If T is any symmetric matrix, $\|T\| = \max_i |\lambda_i(T)| \le |||T|||$ for any operator norm or the Frobenius norm $|||\cdot|||$. Therefore, all the theorems in this paper which are proven for diagonal dominance with respect to the 2-norm automatically hold for diagonal dominance with respect to any operator norm or the Frobenius norm.

The minimax characterization of the eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ of a definite pencil $H - \lambda M$ ($H = H^T$, $M = M^T$, M positive definite) is [17]

$$\lambda_i = \min_{\mathbf{S}^i} \max_{\substack{x \in \mathbf{S}^i \\ \|x\| = 1}} \frac{x^T H x}{x^T M x} \quad . \tag{2.1}$$

where $\mathbf{S}^i$ varies over all i-dimensional subspaces of $\mathbf{R}^n$ and x varies over all unit vectors in $\mathbf{S}^i$ (x could vary over all nonzero vectors, but we will find it convenient to restrict to unit vectors). There is an obvious simplification if M is the identity matrix (the standard eigenproblem).

3. Generalizations of Gershgorin's Theorem

It turns out the eigenvalues of the s.d.d. matrix $H=\Delta_1 A \Delta_2$ lie in Gershgorin circles whose centers and radii are both scaled by $\Delta_1\Delta_2$:

Proposition 1: *Let $H=\Delta_1 A\Delta_2$ ($A_{ii}=\pm 1$) be a (possibly nonsymmetric) γ-s.d.d. matrix with respect to the infinity-norm or 1-norm, with $\gamma < 1$. Then the eigenvalues of H lie in disks B_i, where B_i is centered at H_{ii} and has radius at most $\gamma |H_{ii}|$. If B_i is disjoint from the other disks, it contains exactly one eigenvalue and H_{ii} is an approximation to that eigenvalue with relative error at most γ .*

Proof: Suppose without loss of generality that H is γ-s.d.d. with respect to the infinity-norm; otherwise consider H^T. The scalar λ is an eigenvalue of H if and only if $H-\lambda I$ is singular, which is in turn true if and only if $A-\lambda\Delta_1^{-1}\Delta_2^{-1}$ is singular. Let x be a right null vector of $A-\lambda\Delta_1^{-1}\Delta_2^{-1}$, and suppose x_j has absolute value at least as large as any other component of x. Then we may rearrange the equation

$$\sum_k A_{jk}x_k - \lambda x_j \Delta_{1,jj}^{-1}\Delta_{2,jj}^{-1} = 0$$

to obtain

$$\lambda = \Delta_{1,jj}\Delta_{2,jj}\,(A_{jj} + \sum_{k\neq j} A_{jk}\,\frac{x_k}{x_j}) \quad .$$

Since $A_{jj}\Delta_{1,jj}\Delta_{2,jj} = H_{jj}$ and $|\sum_{k\neq j}^{n} A_{jk}\,\frac{x_k}{x_j}| \leq \gamma$, λ must lie in a ball of radius $\gamma|H_{ii}|$ centered at H_{ii}. The usual Gershgorin argument shows that if this disk is isolated, it contains exactly one eigenvalue. □

This theorem implies that at least if a Gershgorin disk is isolated so that small relative changes in the matrix entries do not effect its isolation, then the eigenvalue it contains cannot change by a factor of more than $(1+\gamma)/(1-\gamma)$. If we assume H is symmetric, we need not assume the disks are isolated to obtain this result:

Proposition 2: *Let H be a γ-s.d.d. symmetric matrix with respect to the 2-norm. Let h_i be its diagonal entries in increasing order $h_1 \leq \cdots \leq h_n$, and λ_i its eigenvalues, also in increasing order. Then*

$$1-\gamma \leq \frac{\lambda_i}{h_i} \leq 1+\gamma \quad .$$

Proof: Assume without loss of generality that $H_{ii}=h_i$, by reordering the rows and columns of H if necessary. Then by (2.1)

$$\lambda_i = \min_{\mathbf{S}^i}\max_{\substack{x\in\mathbf{S}^i \\ \|x\|=1}} x^T H x \leq \max_{\substack{x\in\mathbf{S}_0^i \\ \|x\|=1}} x^T H x = \max_{\|\hat{x}\|=1} \hat{x}^T H^{(i)}\hat{x}$$

where $\mathbf{S}_0^i$ is the space spanned by the first i standard basis vectors, and $H^{(i)}$ is the leading principal i by i submatrix of H. If $h_i < 0$ then $\lambda_i \leq (1-\gamma)h_i$ by simple norm inequalities. If $h_i > 0$ and all $h_j > 0$ for $j \leq i$, simple norm inequalities again imply $\lambda_i \leq (1+\gamma)h_i$. If $h_i > 0$ and some $h_j < 0$ for $j < i$, we also have $\lambda_i \leq (1+\gamma)h_i$ but we must argue as follows:

$$\hat{x}^T H^{(i)}\hat{x} \equiv [\hat{x}_1^T\ ,\ \hat{x}_2^T]\begin{bmatrix}\Delta_1 & 0\\ 0 & \Delta_2\end{bmatrix}\cdot\left(\begin{bmatrix}-I_j & 0\\ 0 & I_k\end{bmatrix} + N^{(i)}\right)\cdot\begin{bmatrix}\Delta_1 & 0\\ 0 & \Delta_2\end{bmatrix}\cdot\begin{bmatrix}\hat{x}_1\\ \hat{x}_2\end{bmatrix}$$

$$\leq -\|\Delta_1\hat{x}_1\|^2 + \|\Delta_2\hat{x}_2\|^2 + \gamma(\|\Delta_1\hat{x}_1\|^2 + \|\Delta_2\hat{x}_2\|^2)$$

$$\leq (1+\gamma)\|\Delta_2 \hat{x}_2\|^2 \leq (1+\gamma)h_i \ .$$

Applying the same process to $-H$ yields the complementary inequalities $\lambda_i \geq (1+\gamma)h_i$ for $h_i < 0$ and $\lambda_i \geq (1-\gamma)h_i$ for $h_i > 0$. □

Finally, we may extend the result to s.d.d. symmetric definite pencils:

Proposition 3: *Let $H-\lambda M$ be a symmetric γ-s.d.d. definite pencil with respect to the 2-norm. Let r_i be the sorted ratios of diagonal entries H_{ii}/M_{ii}, where $r_1 \leq \cdots \leq r_n$, and λ_i the eigenvalues, also in increasing order. Then*

$$\frac{1-\gamma}{1+\gamma} \leq \frac{\lambda_i}{r_i} \leq \frac{1+\gamma}{1-\gamma} \ .$$

Proof: Assume as in the proof of Proposition 2 that $r_i = H_{ii}/M_{ii}$, by reordering rows and columns if necessary. Write $M = \Delta A \Delta$ where Δ is diagonal and A is diagonally dominant with ones on its diagonal. Then the pencil $\Delta^{-1} H \Delta^{-1} - \lambda A \equiv R - \lambda A$ has the same eigenvalues as $H-\lambda M$, but now the diagonal entries $R_{ii} = r_i$. Then

$$\lambda_i = \min_{\mathbf{S}^i} \max_{\substack{x \in \mathbf{S}^i \\ \|x\|=1}} \frac{x^T R x}{x^T A x} \leq \max_{\substack{x \in \mathbf{S}_0^i \\ \|x\|=1}} \frac{x^T R x}{x^T A x} = \max_{\|\hat{x}\|=1} \frac{\hat{x}^T R^{(i)} \hat{x}}{\hat{x}^T A^{(i)} \hat{x}}$$

where $\mathbf{S}_0^i$ is the space spanned by the first i standard basis vectors, and $R^{(i)}$ and $A^{(i)}$ are the leading principal i by i submatrices of R and A, respectively. Note that $1-\gamma \leq \hat{x}^T A^{(i)} \hat{x} \leq 1+\gamma$ for all unit vectors $\hat{x}$, since A equals the identity matrix plus a matrix of norm at most γ. Then by Proposition 2, we have $\lambda_i \leq (1+\gamma)/(1-\gamma) r_i$ if $r_i > 0$ and $\lambda_i \leq (1-\gamma)/(1+\gamma) r_i$ if $r_i < 0$. Applying the same process to $R + \lambda A$ yields the other inequalities. □

4. Perturbation Lemmas Based on the Minimax Theorem

Let $\lambda_1(H,M) \leq \cdots \leq \lambda_n(H,M)$ denote the eigenvalues of the definite pencil $H-\lambda M$. Given the minimax characterization in (2.1), the following lemma is simple to prove:

Lemma 1: *Suppose δH has the property that for all nonzero x*

$$g_l \leq \frac{x^T(H+\delta H)x}{x^THx} \leq g_u \quad ,$$

where $0< g_l \leq g_u$. Then

$$g_l \leq \frac{\lambda_i(H+\delta H,M)}{\lambda_i(H,M)} \leq g_u$$

for all i. In other words, if the Rayleigh quotients $x^T(H+\delta H)x$ and x^THx differ by at most a certain factor for all x, then the eigenvalues of $H+\delta H-\lambda M$ and $H-\lambda M$ differ by at most that same factor.

Proof: Let $\lambda_i \equiv \lambda_i(H,M)$ and Let $\lambda'_i \equiv \lambda_i(H+\delta H,M)$. We consider only $\lambda_i \geq 0$; the case $\lambda_i < 0$ is analogous. Let the spaces $\mathbf{S}_0^i$ and $\mathbf{S}_1^i$ satisfy

$$\lambda_i = \max_{x\in\mathbf{S}_0^i} x^THx/x^TMx \quad \text{and} \quad \lambda'_i = \max_{x\in\mathbf{S}_1^i} x^T(H+\delta H)x/x^TMx \quad .$$

Then

$$\lambda'_i = \min_{\mathbf{S}^i} \max_{x\in\mathbf{S}^i} \frac{x^T(H+\delta H)x}{x^TMx} \leq \max_{x\in\mathbf{S}_0^i} \frac{x^T(H+\delta H)x}{x^THx}\,\frac{x^THx}{x^TMx} \leq g_u\lambda_i$$

and similarly

$$\lambda_i = \min_{\mathbf{S}^i} \max_{x\in\mathbf{S}^i} \frac{x^THx}{x^TMx} \leq \max_{x\in\mathbf{S}_1^i} \frac{x^THx}{x^T(H+\delta H)x}\,\frac{x^T(H+\delta H)x}{x^TMx} \leq g_l^{-1}\lambda'_i$$

completing the proof. □

Lemma 1 can also be generalized to infinite dimensional operators [15, Thm VI.3.9].

There is an obvious analogous result if if both H and M are perturbed simultaneously:

Lemma 2: *Suppose δH and δM have the property that for all nonzero x*

$$g_{lH} \leq \frac{x^T(H+\delta H)x}{x^THx} \leq g_{uH} \quad \text{and} \quad g_{lM} \leq \frac{x^T(M+\delta M)x}{x^TMx} \leq g_{uM} \quad ,$$

where $0< g_{lH} \leq g_{uH}$ and $0< g_{lM} \leq g_{uM}$. Then

$$\frac{g_{lH}}{g_{uM}} \leq \frac{\lambda_i(H+\delta H,M+\delta M)}{\lambda_i(H,M)} \leq \frac{g_{uH}}{g_{lM}}$$

for all i.

Lemma 3: *Let H be symmetric γ-s.d.d. with respect to the 2-norm. Write $H=\Delta A\Delta$ where Δ is diagonal and A has ± 1's on its diagonal. Let λ be an eigenvalue of H, y the corresponding unit eigenvector, and $x=\Delta y/\|\Delta y\|$ the corresponding unit eigenvector of $A-\lambda\Delta^{-2}$. Then*

$$1-\gamma \leq |x^TAx| \leq 1+\gamma \quad .$$

Proof: Write $A=E+N$, where E is diagonal with ± 1's on the diagonal, N has zero diagonal, and $\|N\| \leq \gamma < 1$. The upper bound on $|x^TAx|$ follows immediately from taking norms: $|x^TAx| \leq \|E\| + \|N\| \leq 1+\gamma$. If $E=I$, then $|x^TAx| = |1+x^TNx| \geq 1-\gamma$. Assume then without loss of generality that

$$E = \begin{bmatrix} I_j & 0 \\ 0 & -I_k \end{bmatrix}$$

where I_l denotes an l-by-l identity matrix. We will prove the theorem only for $\lambda < 0$; for the positive λ consider $-H$. Partition

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad , \quad \Delta = \begin{bmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} N_1^T \\ N_2^T \end{bmatrix}$$

conformally with E. Then $Ax = \lambda\Delta^{-2}x$ may be rewritten

$$\begin{aligned} x_1 + N_1^T x &= \lambda\Delta_1^{-2}x_1 \\ -x_2 + N_2^T x &= \lambda\Delta_2^{-2}x_2 \end{aligned} \quad .$$

Solving the first equation above for x_1 yields

$$x_1 = (\lambda\Delta_1^{-2} - I)^{-1}N_1^T x \quad .$$

Note that $\lambda\Delta_1^{-2} - I$ is diagonal with diagonal entries less than -1. Now

$$x^T Ax = x^T Ex + x^T Nx = x_1^T x_1 - x_2^T x_2 + x^T Nx = 2x_1^T x_1 - 1 + x^T Nx$$

since $x_1^T x_1 + x_2^T x_2 = 1$. Combining the last two displayed equations, and using the fact that

$$-x^T N_1(\lambda\Delta_1^{-2} - I)^{-1}N_1^T x \geq x^T N_1(\lambda\Delta_1^{-2} - I)^{-2}N_1^T x \geq 0$$

yields

$$\begin{aligned} x^T Ax &= 2x^T N_1(\lambda\Delta_1^{-2} - I)^{-2}N_1^T x + (x_1^T \ , \ x_2^T)\begin{bmatrix} N_1^T x \\ N_2^T x \end{bmatrix} - 1 \\ &= 2x^T N_1(\lambda\Delta_1^{-2} - I)^{-2}N_1^T x + x_2^T N_2^T x + x^T N_1(\lambda\Delta_1^{-2} - I)^{-1}N_1^T x - 1 \\ &\leq x_2^T N_2^T x + x^T N_1(\lambda\Delta_1^{-2} - I)^{-2}N_1^T x - 1 \\ &= x_2^T N_2^T x + x_1^T(\lambda\Delta_1^{-2} - I)^{-1}N_1^T x - 1 \\ &= (x_1^T(\lambda\Delta_1^{-2} - I)^{-1} \ , \ x_2^T)Nx - 1 \\ &\leq \left\| \begin{bmatrix} (\lambda\Delta_1^{-2} - I)^{-1}x_1 \\ x_2 \end{bmatrix} \right\| \cdot \|Nx\| - 1 \\ &\leq \left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\| \cdot \|Nx\| - 1 \\ &\leq \gamma - 1 \end{aligned} \qquad \square.$$

In the next two sections we will use these results to derive perturbation theorems for eigenvalues and singular values.

5. A Perturbation Theorem for Singular Values.

Using Lemma 2, we prove the following theorem, which is a slight strengthening of a result of Kahan [9]:

Theorem 1: *Let B be an n by n bidiagonal matrix:*

$$B = \begin{bmatrix} a_1 & b_1 & & \\ & \ddots & \ddots & \\ & & \ddots & b_{n-1} \\ & & & a_n \end{bmatrix} .$$

We assume the a_i and b_i are nonzero since otherwise B splits into independent subproblems. Let $B+\delta B$ be a perturbed bidiagonal matrix with entries $\alpha_i a_i$ in place of a_i and $\beta_i b_i$ in place of b_i. Then the singular values $\sigma_1 \le \cdots \le \sigma_n$ of B and $\sigma'_1 \le \cdots \le \sigma'_n$ satisfy

$$g_l \le \frac{\sigma'_i}{\sigma_i} \le g_u$$

where g_l and g_u are defined as follows. Define the finite set S of positive numbers by

$$S \equiv \{ \left| \frac{\beta_j \cdots \beta_k}{\alpha_{j+1} \cdots \alpha_k} \right| : 1 \le j \le k \le n-1 \} \cup \{ \left| \frac{\alpha_j \cdots \alpha_k}{\beta_j \cdots \beta_{k-1}} \right| : 1 \le j \le k \le n \} .$$

Note that S contains $|\beta_j|$, $1 \le j < n$, and $|\alpha_j|$, $1 \le j \le n$. Let min S *and* max S *denote the minimum and maximum entries of S, respectively. Then*

$$g_u = \max S \quad \text{and} \quad g_l = \min S .$$

Corollary 1: *Let B and $B+\delta B$ be bidiagonal with singular values $\sigma_1(B) \le \cdots \le \sigma_n(B)$ and $\sigma_1(B+\delta B) \le \cdots \le \sigma_n(B+\delta B)$ respectively. If for all nonzero entries B_{ij}, $\tau^{-1} \le |(B+\delta B)_{ij}/B_{ij}| \le \tau$ for some $\tau \ge 1$, then*

$$\frac{1}{\tau^{2n-1}} \le \frac{\sigma_i(B+\delta B)}{\sigma_i(B)} \le \tau^{2n-1} .$$

Thus, relative perturbations of at most τ in the entries of B cause relative perturbations of at most τ^{2n-1} in its singular values. If $\tau = 1+\eta$ is close to 1, so is $\tau^{2n-1} \approx 1+(2n-1)\eta$.

Corollary 2: *Let B and $B+\delta B$ be bidiagonal. If $|(B+\delta B)_{ij}| = \tau |B_{ij}|$ for all i and j, then $\sigma_i(B+\delta B) = \tau \sigma_i(B)$. This simply says that if you multiply each entry of a bidiagonal matrix by $\pm\tau$, you multiply all its singular values by τ as well.*

Proof of Theorem 1: For notational convenience rename the entries of B so that

$$B = \begin{bmatrix} s_1 & s_2 & & & \\ & s_3 & s_4 & & \\ & & \ddots & \ddots & \\ & & & \ddots & s_{2n-2} \\ & & & & s_{2n-1} \end{bmatrix}$$

and so that $B+\delta B$ has entries $\gamma_i s_i$ in place of s_i. We may assume without loss of generality that all the s_i are real and positive, since this may be achieved by pre- and postmultiplying B by unitary diagonal matrices. We may also assume the γ_i are real and positive for the same reason. We also use the well known fact that the eigenvalues of

$$C = \begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix}$$

are plus and minus the singular values of B. Furthermore, by reordering the rows and columns of C in the order 1, $n+1$, 2, $n+2$, . . . , n, $2n$, we see C is similar to

$$E = \begin{bmatrix} 0 & s_1 & & \\ s_1 & \cdot & \cdot & \\ & \cdot & \cdot & s_{2n-1} \\ & & s_{2n-1} & 0 \end{bmatrix} .$$

Thus, the singular values of B are the absolute values of the eigenvalues of the pencil $E-\lambda I$. We are interested in comparing these eigenvalues with the eigenvalues of $E+\delta E-\lambda I$, where $E+\delta E$ has entries $\gamma_i s_i$ in place of s_i. We will chose a diagonal matrix D such that $D(E+\delta E)D = E$, so that these eigenvalues will be the eigenvalues of the pencil $E-\lambda D^2$; we will then apply Lemma 2 to conclude that the eigenvalues will change at most by a factor in the range $\min_i D_{ii}^{-2}$ to $\max_i D_{ii}^{-2}$.

Actually, we will find two diagonal matrices D_u and D_l satisfying $D(E+\delta E)D=E$ but corresponding to different choices of D_{11} in order to minimize $\max_i D_{lii}$ and maximize $\min_i D_{uii}$; this will in turn maximize g_l and minimize g_u where (by Lemma 2)

$$g_l \equiv \frac{1}{\max_i D_{lii}^2} \le \frac{\sigma'_i}{\sigma_i} \le \frac{1}{\min_i D_{uii}^2} \equiv g_u .$$

First consider D_l. We want to choose D_{l11} to minimize $\max_i D_{lii}$. $D(E+\delta E)D=E$ implies the diagonal entries of D_l must satisfy the recurrence $D_{l,i+1,i+1} = (\gamma_i D_{lii})^{-1}$. This means the diagonal entries of D_l are either of the form

$$D_{l11} \cdot \frac{\gamma_1 \gamma_3 \cdots \gamma_{2j-1}}{\gamma_2 \gamma_4 \cdots \gamma_{2j}} \tag{5.1}$$

or

$$\frac{1}{D_{l11}\gamma_1} \cdot \frac{\gamma_2 \gamma_4 \cdots \gamma_{2j}}{\gamma_3 \gamma_5 \cdots \gamma_{2j+1}} . \tag{5.2}$$

Choosing D_{l11} to minimize the maximum of these terms is equivalent to choosing D_{l11} to minimize $\max(s_1 D_{l11}\ ,\ s_2 D_{l11}^{-1})$, where s_1 is the maximum coefficient of D_{l11} in (5.1) and s_2 is the maximum coefficient of D_{l11}^{-1} in (5.2); this minimum is easily seen to be $(s_1 s_2)^{1/2}$. Some tedious algebraic manipulation leads to the expression for g_l in the statement of the theorem. Choosing D_{u11} to maximize $\min_i D_{uii}$ is analogous. □

A similar proof yields the following analogous result: Let A be an arbitrary matrix, and D_1 and D_2 arbitrary nonsingular diagonal matrices, which we may assume are real and nonnegative. Then the singular values σ_i of A and σ'_i of $D_1 A D_2$ satisfy

$$\min_i D_{1i} \cdot \min_j D_{2j} \le \frac{\sigma'_i}{\sigma_i} \le \max_i D_{1i} \cdot \max_j D_{2j} .$$

6. Perturbation Theorems for Eigenvalues

In this section we use the perturbation lemma of section 3 to prove perturbation theorems for eigenvalues of symmetric γ-s.d.d. matrices and γ-s.d.d. positive definite pencils. Theorems 2 and 3 apply only to positive definite matrices and pencils, and are stronger than Proposition 4 and Theorem 4 which apply to indefinite matrices as well. In fact, when the matrix or pencil is positive definite, Δ may be an *arbitrary* matrix, not just diagonal. Our results for definite pencils with both positive and negative eigenvalues are rather weaker than the results for s.d.d. indefinite symmetric matrices.

Theorem 2: *Let $H=\Delta^T A\Delta=\Delta^T(I+N)\Delta$ be a positive definite matrix where $\|N\|=\gamma<1$, and Δ is an* arbitrary *nonsingular matrix. (If Δ is diagonal this is equivalent to H being a symmetric positive definite γ-s.d.d. matrix with respect to the 2-norm.) Let δH be a symmetric perturbation of H with $\|\Delta^{-T}\delta H\Delta^{-1}\| \equiv \eta < 1-\gamma$. Then the eigenvalues $\lambda_1\le \cdots \le \lambda_n$ of H and $\lambda'_1\le \cdots \le \lambda'_n$ of $H+\delta H$ satisfy*

$$1 - \frac{\eta}{1-\gamma} \le \frac{\lambda'_i}{\lambda_i} \le 1 + \frac{\eta}{1-\gamma} \quad .$$

Thus, if Δ is diagonal, relative perturbations of at most η in the entries of H cause only relative perturbations of at most $n\eta/(1-\gamma)$ in the eigenvalues.

Proof: We have

$$\frac{x^T(H+\delta H)x}{x^THx} = \frac{x^T\Delta^T(A+\Delta^{-T}\delta H\Delta^{-1})\Delta x}{x^T\Delta^TA\Delta x} = \frac{y^T(A+\Delta^{-T}\delta H\Delta^{-1})y}{y^TAy} \tag{6.1}$$

where $y=\Delta x$. The values taken by the quantity in (6.1) as x varies over all nonzero vectors are the same as the values taken on as y varies over all unit vectors. Since $1-\gamma\le y^TAy$ if y is a unit vector,

$$1-\frac{\eta}{1-\gamma} \le \frac{y^T(A+\Delta^{-T}\delta H\Delta^{-1})y}{y^TAy} \le 1+\frac{\eta}{1-\gamma} \quad .$$

The result now follows from Lemma 1. □

Theorem 3: *Let $H=\Delta_H^TA_H\Delta_H=\Delta_H^T(I+N_H)\Delta_H$ and $M=\Delta_M^TA_M\Delta_M=\Delta_M^T(I+N_M)\Delta_M$ be positive definite matrices where $\|N_H\|\le\gamma<1$, $\|N_M\|\le\gamma<1$, and Δ_H and Δ_M are* arbitrary *nonsingular matrices. (If Δ_H and Δ_M are diagonal this is equivalent to $H-\lambda M$ being a γ-s.d.d. positive definite pencil with respect to the 2-norm.) Let δH and δM be symmetric perturbations of H and M such that $\|\Delta_H^{-T}\delta H\Delta_H^{-1}\| \le \eta < 1-\gamma$ and $\|\Delta_M^{-T}\delta M\Delta_M^{-1}\| \le \eta < 1-\gamma$. Then the eigenvalues $\lambda_1\le \cdots \le \lambda_n$ of $H-\lambda M$ and $\lambda'_1\le \cdots \le \lambda'_n$ of $(H+\delta H)-\lambda(M+\delta M)$ satisfy*

$$\frac{1-\gamma-\eta}{1-\gamma+\eta} \le \frac{\lambda_i}{\lambda'_i} \le \frac{1-\gamma+\eta}{1-\gamma-\eta}$$

Thus, if Δ_H and Δ_M are diagonal, relative perturbations of at most η in the entries of H and M cause only relative perturbations of at most $2n\eta/(1-\gamma-2n\eta)$ in the eigenvalues of $H-\lambda M$.

Proof: Apply the technique in the proof of Theorem 2 to both H and M and apply Lemma 2. □

When H is indefinite, our results are weaker. In the case of H alone, we must assume Δ is diagonal to attain a bound like that of Theorem 2:

Proposition 4: *Let H be an n-by-n symmetric γ-s.d.d. matrix with respect to the 2-norm. Thus $H = \Delta A \Delta$ where Δ is diagonal, $A = E + N$ where E is diagonal with ± 1's on the diagonal, N has a zero diagonal and $\|N\| \le \gamma < 1$. Let δH be a symmetric perturbation of H with $\|\Delta^{-1}\delta H \Delta^{-1}\| \equiv \eta < (1-\gamma)/n$. Assume also that $H + \delta H$ is γ-s.d.d. Then the eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ of H and $\lambda'_1 \le \cdots \le \lambda'_n$ of $H + \delta H$ satisfy*

$$1 - \frac{n\eta}{1-\gamma} \le \frac{\lambda'_i}{\lambda_i} \le (1 - \frac{n\eta}{1-\gamma})^{-1} \quad .$$

Thus, relative perturbations of at most η in the entries of H can cause relative perturbations of at most $n^2\eta/(1-\gamma)$ in the eigenvalues.

Proof: We will prove the theorem only for the negative eigenvalues; for the positive ones consider $-H$. We cannot apply Lemma 1 directly here because $x^T Ax / x^T x$ is not bounded away from 0 for all x as in the proof of Proposition 4. By Lemma 3, however, it is so bounded if we restrict x to lie in an eigenspace of $A - \lambda\Delta^{-2}$ corresponding to only negative (or only positive) eigenvalues; this will be sufficient for the proof.

To this end, let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues and $x_1, \ldots, x_n$ the corresponding unit right eigenvectors of $A - \lambda\Delta^{-2}$; let the primed quantities $\lambda'_1 \le \cdots \le \lambda'_n$ and x'_i correspond to $A + \Delta^{-1}\delta H \Delta^{-1} - \lambda\Delta^{-2}$. By Lemma 3 $x_i^T A x_i \le \gamma - 1 < 0$ if $\lambda_i < 0$.

Now let $X_i = [x_1, \ldots, x_i]$ and $\Lambda_i = diag(\lambda_1, \ldots, \lambda_i)$, so $AX_i = \Delta^{-2} X_i \Lambda_i$. Since the columns of X_i are eigenvectors of $A - \lambda\Delta^{-2}$, the columns of $\Delta^{-1}X_i$ are eigenvectors of H and so are orthogonal. Thus

$$X_i^T A X_i = X_i^T \Delta^{-2} X_i \Lambda_i = diag(x_i^T \Delta^{-2} x_i \lambda_i)$$

is diagonal with diagonal entries bounded above by $\gamma - 1$. Let z be an arbitrary unit vector; then

$$z^T X_i^T A X_i z = z^T diag(x_i^T \Delta^{-2} x_i \lambda_i) z \le \gamma - 1 \quad .$$

Now we use the characterization

$$\lambda_i = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{x^T A x}{x^T \Delta^{-2} x}$$

where the minimum is attained for $\mathbf{S}^i = \mathbf{S}_0^i = span(X_i)$. Then

$$\lambda'_i = \min_{\mathbf{S}^i} \max_{x \in \mathbf{S}^i} \frac{x^T(A + \Delta^{-1}\delta H \Delta^{-1})x}{x^T \Delta^{-2} x} \le \max_{x \in \mathbf{S}_0} \frac{x^T(A + \Delta^{-1}\delta H \Delta^{-1})x}{x^T A x} \cdot \frac{x^T A x}{x^T \Delta^{-2} x}$$

$$= \max_{\|z\|=1} (1 + \frac{z^T X_i^T \Delta^{-1}\delta H \Delta^{-1} X_i z}{z^T X_i^T A X_i z}) \cdot \frac{z^T X_i^T A X_i z}{z^T X_i^T \Delta^{-2} X_i z} \quad .$$

Now $|z^T X_i^T \Delta^{-1}\delta H \Delta^{-1} X_i z| \le i\eta$ and $|z^T X_i^T A X_i z| \ge 1 - \gamma$ so

$$\lambda'_i \le (1 - \frac{i\eta}{1-\gamma})\lambda_i \quad \text{or} \quad \frac{\lambda'_i}{\lambda_i} \ge 1 - \frac{i\eta}{1-\gamma}$$

Swapping the roles of A and $A + \Delta^{-1}\delta H \Delta^{-1}$ we obtain

$$1 - \frac{i\eta}{1-\gamma} \le \frac{\lambda'_i}{\lambda_i} \le (1 - \frac{i\eta}{1-\gamma})^{-1}$$

as desired. □

The factor n in the bound of Proposition 4 is an overestimate, and can be removed by modifying the conditions of the proposition just slightly:

Theorem 4: *Let $H=\Delta A\Delta$ be an n-by-n symmetric γ-s.d.d. matrix with respect to the 2-norm. Here Δ is diagonal and A has ± 1's on the diagonal. Let δH be a symmetric perturbation with $\|\Delta^{-1}\delta H\Delta^{-1}\| \equiv \eta$. Assume that $H+\xi\delta H$ is γ-s.d.d. for all $0\le\xi\le 1$. Then letting $\lambda_1\le\cdots\le\lambda_n$ be the eigenvalues of H and $\lambda'_1\le\cdots\le\lambda'_n$ be the eigenvalues of $H+\delta H$, we have*

$$\exp(\frac{-\eta}{1-\gamma}) \le \frac{\lambda'_i}{\lambda_i} \le \exp(\frac{\eta}{1-\gamma}) \quad . \tag{6.2}$$

Proof: Let $E = \Delta^{-1}\delta H\Delta^{-1}/\|\Delta^{-1}\delta H\Delta^{-1}\|$ be a matrix of norm 1, $H(\zeta) = \Delta(A+\zeta E)\Delta$, and $\lambda_1(\zeta)\le\cdots\le\lambda_i(\zeta)$ be the eigenvalues of $H(\zeta)$. Suppose first that $\lambda_i(0)$ is simple. Let x_i be the unit eigenvector corresponding to $\lambda_i(0)$. Then from standard eigenvalue perturbation theorem [15, 19], we know

$$\lambda_i(\zeta) = \lambda_i(0) + \zeta x_i^T\Delta E\Delta x_i + O(\zeta^2)$$

Therefore

$$\frac{\lambda_i(\zeta)}{\lambda_i(0)} = 1 + \zeta\frac{x_i^T\Delta E\Delta x_i}{x_i^T\Delta A\Delta x_i} + O(\zeta^2) = 1 + \zeta\frac{y_i^T E y_i}{y_i^T A y_i} + O(\zeta^2)$$

where $\|y_i\|$ may be taken to be one. By Lemma 3, $|y_i^T A y_i| \ge 1-\gamma$, and so

$$1 - \frac{\zeta}{1-\gamma} + O(\zeta^2) \le \frac{\lambda_i(\zeta)}{\lambda_i(0)} \le 1 + \frac{\zeta}{1-\gamma} + O(\zeta^2) \quad . \tag{6.3}$$

Assume now that $\lambda_i(\zeta)$ is simple for all $0\le\zeta\le\eta$; then (6.3) implies that $|d\log\lambda_i(x)/dx| \le (1-\gamma)^{-1}$. Integrating from 0 to η yields (6.2). By [11, Theorem II.6.1], the eigenvalues are all real analytic, even when they are multiple. Thus, if there are only finitely many ζ where $\lambda_i(\zeta)$ is multiple, we can apply the above argument in the intermediate intervals.

It remains to consider the case where $\lambda_i(\zeta)$ is multiple for infinitely many ζ. Here we argue that this can only happen for a set of pairs of matrices H and E of measure zero, that off this set the previous argument holds, and by continuity of the eigenvalues the same bounds must hold on the set. To see that the set of H and E such that some $\lambda_i(\zeta)$ is multiple infinitely often is of measure zero, consider the discriminant of the characteristic polynomial of $H-\zeta\Delta E\Delta$; this is a polynomial in ζ and the $2n^2$ entries of H and E. $H(\zeta)$ can have multiple eigenvalues if and only if this discriminant vanishes. If it vanishes for infinitely many ζ, its coefficients (viewing it as a polynomial in ζ) must vanish identically. These coefficient are in turn polynomials in the entries of H and E, and so vanish only on a proper variety (a set of measure zero). Off this set, the discriminant has at most a finite number of zeros (bounded by its degree as a polynomial in ζ). □

Result (6.2) was claimed without proof in [14] just for the case of tridiagonal matrices perturbed on their offdiagonals.

In light of Proposition 3, it is probably possible to prove an analogous theorem for γ-s.d.d. definite pencils which have both positive and negative eigenvalues, but we have not been able to do so. However, the following technique frequently succeeds in reducing the eigenproblem for such pencils to a problem where Theorem 4 may be applied:

Algorithm 1: *Reducing a γ-s.d.d. definite pencil $H-\lambda M$ to an s.d.d. matrix Y:*

(1) Let $D_1=\mathrm{diag}(M_{ii}^{1/2})$, and compute $H_1=D^{-1}HD^{-1}$ and $M_1=D^{-1}MD^{-1}$. Now M_1 has unit diagonal and is diagonally dominant in the usual sense.

(2) Let P be a permutation matrix chosen so that $H_2 = PH_1P^T$ has its diagonal entries sorted from smallest to largest in absolute value (smallest at the top left, largest at the bottom right). Let $M_2=PM_1P^T$.

(3) Let L be the lower triangular Cholesky factor of M_2. Let $Y=L^{-1}H_2L^{-T}$. Then Y and $H-\lambda M$ have the same eigenvalues.

This is a variation on the usual reduction of a definite pencil to standard form. The point is that if M is sufficiently diagonally dominant, L will also be diagonally dominant with nearly unit diagonal, and the multiplication $L^{-1}H_2L^{-T}$ will not destroy the s.d.d. property of H_2. Thus, Y will be s.d.d. The following theorem formalizes this, but is weak in that it only guarantees diagonal dominance of Y for rather small γ, much smaller than those that work in practice:

Proposition 5: *Let $H-\lambda M$ be an n by n γ-s.d.d. definite pencil, and let Y be the output of the above reduction algorithm. Define*

$$\gamma' \equiv ((2n)^{1/2}+1)\cdot\frac{1+\gamma}{1-\gamma}\cdot[2 + ((2n)^{1/2}+1)\gamma^{1/2}]\cdot\gamma^{1/2} \quad .$$

Then if

$$\bar{\gamma} \equiv \frac{(n+1)\gamma'+\gamma}{1-\gamma'} < 1 \quad ,$$

which will be true for γ small enough, Y will be $\bar{\gamma}$-s.d.d.

The proof of Proposition 5 requires the following lemma:

Lemma 4: *Let M be an n by n γ-s.d.d. matrix with unit diagonal. Let L be its lower triangular Cholesky factor. Then $\|L-I\| \le ((2n)^{1/2}+1)\gamma^{1/2}$ and $\|L^{-1}-I\| \le ((2n)^{1/2}+1)\gamma^{1/2}/(1-\gamma)^{1/2}$. Also, $(1+\gamma)^{-1/2} \le \|L^{-1}\| \le (1-\gamma)^{-1/2}$.*

Proof: Let $X=L-I$ and $W=L^{-1}-I = -L^{-1}X$. Since

$$(1-\gamma)^{1/2} \le \lambda_{\min}^{1/2}(M) = \sigma_{\min}(L) \le L_{ii} \le \sigma_{\max}(L) = \lambda_{\max}^{1/2}(M) \le (1+\gamma)^{1/2}$$

we have $|X_{ii}| = |L_{ii}-1| \le 1-(1-\gamma)^{1/2} \le \gamma^{1/2}$. Also, we may bound the norm of the i-th subdiagonal column of X as follows:

$$1+\gamma \ge \|[L_{i,i}\ ,\ L_{i+1,i}\ ,\ \cdots\ ,\ L_{n,i}]\|^2 = \|[L_{i,i}\ ,\ X_{i+1,i}\ ,\ \cdots\ ,\ X_{n,i}]\|^2 \ge 1-\gamma + \|X_{i+1:n,i}\|^2$$

whence $\|X_{i+1:n,i}\| \le (2\gamma)^{1/2}$. Thus $\|X\| \le (2n\gamma)^{1/2}+\gamma^{1/2}$ as desired. Finally, $\|W\| \le \|L^{-1}\|\cdot\|X\| \le (1-\gamma)^{-1/2}\|X\|$. □

Proof of Proposition 5: By applying the first two steps of the reduction, assume without loss of generality that M has unit diagonal and that $|H_{ii}| \le |H_{i+1,i+1}|$ for all i. Let $L_{\cdot,i}$ denote the i-th column of L and similarly for $L_{j,\cdot}$. Also, let $G^{(i,j)}$ denote the leading i by j submatrix of G. Let $D = \mathrm{diag}(|H_{ii}|^{1/2})$ and $L^{-1}=I+W$. Then

$$Y_{ij} = L_{i,\cdot}^{-1}HL_{\cdot,j}^{-T} = H_{ij} + W_{i,\cdot}HL_{\cdot,j}^{-T} + L_{i,\cdot}^{-1}HW_{\cdot,j} - W_{i,\cdot}HW_{\cdot,j} \quad .$$

Therefore

$$|Y_{ij}-H_{ij}| \le |W_{i,\cdot}HL_{\cdot,j}^{-T}| + |L_{i,\cdot}^{-1}HW_{\cdot,j}| + |W_{i,\cdot}HW_{\cdot,j}|$$

$$\le (2\|W\|\cdot\|L^{-1}\| + \|W\|^2)\|H^{(i,j)}\| \le (2\|W\|\cdot\|L^{-1}\| + \|W\|^2)D_{ii}D_{jj}(1+\gamma) \quad .$$

Now insert the bounds of Lemma 4 to get

$$|(D^{-1}(Y-H)D^{-1})_{ij}| \le ((2n)^{1/2}+1)\cdot\frac{1+\gamma}{1-\gamma}\cdot[2 + ((2n)^{1/2}+1)\gamma^{1/2}]\cdot\gamma^{1/2} = \gamma' \quad .$$

Finally, letting $D_Y = \mathrm{diag}(|Y_{ii}|^{1/2})$

$$\|\mathrm{offdiag}(D_Y^{-1} Y D_Y^{-1})\| \leq \frac{(n+1)\gamma' + \gamma}{1-\gamma'}$$

follows from simple norm inequalities. □

In order to apply Proposition 4 or Theorem 4, however, we must argue that small relative perturbations in H and M (of the type permitted in Proposition 4 and Theorem 4) cause small perturbations of the same type in Y:

Theorem 5: *Let $H - \lambda M = D_H A_H D_H - \lambda D_M A_M D_M$ be a γ-s.d.d. definite pencil, where A_H and A_M have ± 1s on their diagonals. Let $Y = D_Y A_Y D_Y$ be the output of Algorithm 1 applied to $H - \lambda M$. Assume that Y is $\overline{\gamma}$-s.d.d. ($\overline{\gamma}$ may be smaller than the expression in Proposition 5). Now define $H(\zeta) = D_H(A_H + \zeta E_H)D_H$, and similarly $M(\zeta) = D_M(A_M + \zeta E_M)D_M$, where $\|E_H\| = \|E_M\| = 1$. Let $Y(\zeta)$ be the output of the reduction algorithm applied to $H(\zeta)$. Then for asymptotically small ζ*

$$\|D_Y^{-1}(Y(\zeta) - Y)D_Y^{-1}\| \leq \zeta \cdot \frac{n(2 \cdot n^{1/2} + 1)(1+\gamma)}{(1-\gamma)^2} + O(\zeta^2)$$

and the eigenvalues $\lambda_i(\zeta)$ of $H(\zeta) - \lambda M(\zeta)$ satisfy

$$1 - \zeta \cdot \frac{n(2 \cdot n^{1/2} + 1)(1+\gamma)}{(1-\overline{\gamma})(1-\gamma)^2} + O(\zeta^2) \leq \frac{\lambda_i(\zeta)}{\lambda_i(0)} \leq 1 + \zeta \cdot \frac{n(2 \cdot n^{1/2} + 1)(1+\gamma)}{(1-\overline{\gamma})(1-\gamma)^2} + O(\zeta^2) \quad .$$

For the proof of Theorem 5 we require the following lemma:

Lemma 5: *Let M be a γ-s.d.d. symmetric matrix with unit diagonal. Let L be its lower triangular Cholesky factor. Let $L + \delta L$ be the lower triangular Cholesky factor of the perturbed matrix $M + \delta M$. Then*

$$\|\delta L\| \leq \left(\frac{n}{1-\gamma} \right)^{1/2} \cdot \|\delta M\| + O(\|\delta M\|^2) \quad .$$

Proof: It suffices to consider M diagonal, since the Cholesky factors of M and QMQ^T (Q orthogonal) have the same norm. Equating first order terms on both sides of $M + \delta M = (L + \delta L)(L + \delta L)^T$ yields $\delta L_{ij} = M_{jj}^{-1/2} \delta M_{ij}$ $(i > j)$ and $\delta L_{ii} = .5 \cdot M_{ii}^{-1/2} \delta M_{ii}$ $(i = j)$. Taking norms yields the result. □

Proof of Theorem 5: By applying the first two steps of Algorithm 1, we may assume without loss of generality that $M_{ii} = 1$ and $|H_{ii}| \leq |H_{i+1,i+1}|$. Let $L(\zeta)$ be the Cholesky factor of $M(\zeta)$. Then the eigenvalues of $H(\zeta) - \lambda M(\zeta)$ are the eigenvalues of $Y(\zeta) = L^{-1}(\zeta) H(\zeta) L^{-T}(\zeta)$. Letting $M(\zeta) = M + \delta M$, $L(\zeta) = L + \delta L$, and $H(\zeta) = H + \delta H$, we get that to first order

$$\begin{aligned} Y(\zeta) &= (L^{-1} - L^{-1}\delta L L^{-1})(H + \delta H)(L^{-T} - L^{-T}\delta L^T L^{-T}) \\ &= L^{-1} H L^{-T} - L^{-1}\delta L L^{-1} H L^{-T} + L^{-1}\delta H L^{-T} - L^{-1} H L^{-T}\delta L^T L^{-T} \quad . \end{aligned}$$

Therefore to first order in ζ

$$\begin{aligned} |(Y(\zeta) - Y)_{ij}| &\leq D_{A,ii} D_{A,jj} (2 \cdot (1+\gamma) \cdot \|L^{-1}\|^3 \cdot n^{1/2} \cdot (1-\gamma)^{-1/2} + \|L^{-1}\|^2)\zeta \\ &\leq D_{A,ii} D_{A,jj} \frac{(2n^{1/2} + 1)(1+\gamma)}{(1-\gamma)^2} \cdot \zeta \end{aligned}$$

as desired. Applying Theorem 4 yields the final result. □

We may apply Theorem 4 to analyze the convergence criterion for the QR algorithm for eigenvalues of symmetric tridiagonal matrices [17]. In the course of running the QR algorithm on a symmetric tridiagonal matrix one must examine the matrix

$$T = \begin{bmatrix} \cdot & \cdot & & \\ \cdot & a_j & b_j & \\ & b_j & a_{j+1} & \cdot \\ & & \cdot & \cdot \end{bmatrix}$$

and decide whether b_j can be set to zero ("convergence") without making unacceptably large perturbations in the eigenvalues. Theorem 4 tells us that if T is γ-s.d.d., then setting b_j to zero makes relative errors no larger than

$$\exp\left(\frac{|b_j|}{|a_j \cdot a_{j+1}|^{1/2}} \cdot \frac{1}{1-\gamma}\right) - 1 \tag{6.4}$$

in any eigenvalue. This result is attractive because it is inexpensive and purely local; it only depends on b_j and its neighbors a_j and a_{j+1} on the diagonal of T. It does differ from the standard criterion which essentially asks if

$$\frac{|b_j|}{|a_j| + |a_{j+1}|}$$

is small; this criterion is weaker than (6.4) and does not guarantee high relative accuracy.

Unfortunately, even using (6.4) as a convergence criterion does not guarantee that QR will compute eigenvalues with high relative accuracy; there are examples which are even fairly strongly s.d.d. of QR computing eigenvalues with incorrect signs, i.e. no relative accuracy at all. We discuss this further in Section 10.

7. Perturbation Theorems for Eigenvectors

In this section we discuss the sensitivity of the eigenvectors of symmetric s.d.d. matrices under the same small perturbations as in section 6. As discussed in the introduction, the standard perturbation bound (1.3) is proportional to the reciprocal of $gap(\lambda_i) = \min_{j \neq i} |\lambda_i - \lambda_j|$. Thus, if the *absolute* distance from λ_i to its nearest neighbor is small, we expect the corresponding eigenvector to be sensitive to perturbations. Our first result will be an analogous theorem which replaces the gap with the *relative gap*

$$relgap(\lambda_i) = \min_{j \neq i} \frac{|\lambda_i - \lambda_j|}{|\lambda_i \lambda_j|^{1/2}} \quad , \tag{7.1}$$

which may be large even when the usual gap is small.

Even more may be shown. Proposition 6 will show that the eigenvectors are scaled analogously to the matrix entries: if x_i is the eigenvector for λ_i, and λ_j differs from λ_i by a large factor, the j-th component of x_i will be small. Theorem 7 will show that small relative perturbations in the matrix only cause small perturbations in the eigenvector entries *relative to their upper bounds* of Proposition 6; thus some tiny eigenvector components may be determined to high relative accuracy as well. Finally, we discuss eigenvector bounds for definite pencils and singular vector bounds for bidiagonal matrices (partially settling a conjecture from [9]).

Perturbation theory and algorithms for conventionally diagonally dominant nonsymmetric matrices were developed in [1], under the assumption that the gap between eigenvalues greatly exceeded the norm γ of the offdiagonal part. Thus these results apply to general nonsymmetric matrices, but require much stronger diagonal dominance assumptions than we do and are weaker than our results.

Theorem 6: *Let $H=\Delta A\Delta$ be a γ-s.d.d. symmetric matrix with respect to the 2-norm. Let E be symmetric and have 2-norm one, and define $H(\zeta)=\Delta(A+\zeta E)\Delta$. Let $\lambda_i(\zeta)$ be the i-th eigenvalue of $H(\zeta)$, and assume $\lambda_i(0)$ is simple so that the corresponding unit eigenvector $x_i(\zeta)$ is well defined for sufficiently small ζ. Then for asymptotically small ζ*

$$\|x_i(\zeta) - x_i(0)\| \leq \frac{(n-1)\zeta}{(1-\gamma)\ relgap(\lambda_i)} + O(\zeta^2)$$

Proof: From [19] we have

$$x_i(\zeta) = x_i(0) + \zeta\sum_{k\neq i}\frac{x_k^T\Delta E\Delta x_i}{(\lambda_i - \lambda_k)}x_k + O(\zeta^2)$$

Let $y_k = \Delta x_k$. Then

$$x_i(\zeta) = x_i(0) + \zeta\sum_{k\neq i}\frac{y_k^T E y_i}{(\lambda_i - \lambda_k)}x_k + O(\zeta^2)$$

The pair (λ_k, y_k) is an eigenpair of the pencil $A - \lambda\Delta^{-2}$. Thus $y_k^T A y_k = \lambda_k y_k^T\Delta^{-2}y_k$. From Lemma 3 we have

$$(1-\gamma)\|y_k\|^2 \leq |y_k^T A y_k| = |\lambda_k|\cdot\|\Delta^{-1}y_k\|^2 = |\lambda_k| \leq (1+\gamma)\|y_k\|^2 \ .$$

Thus

$$\left(\frac{|\lambda_k|}{1+\gamma}\right)^{1/2} \leq \|y_k\| \leq \left(\frac{|\lambda_k|}{1-\gamma}\right)^{1/2} \tag{7.2}$$

If we let $z_k = y_k/\|y_k\|$ then

$$x_i(\zeta) = x_i(0) + \zeta\sum_{k\neq i}\xi_{ik}\frac{z_k^T E z_i}{(\lambda_i - \lambda_k)/|\lambda_i\lambda_k|^{1/2}} + O(\zeta^2) \tag{7.3}$$

where $(1+\gamma)^{-1} \leq |\xi_{ik}| \leq (1-\gamma)^{-1}$. If we take norms then

$$\|x_i(\zeta) - x_i(0)\| \leq \frac{(n-1)\zeta}{(1-\gamma)\ relgap(\lambda_i)} + O(\zeta^2)$$

as desired. □

Corollary 3: *Let $H(\zeta)$, $\lambda_i(\zeta)$, and $x_i(\zeta)$ be as in Theorem 6. Assume further that $H(\zeta)$ is γ-s.d.d. for all $0\leq\zeta\leq\bar{\zeta}$, that $1-\gamma-3n\bar{\zeta} > 0$, and*

$$relgap(\lambda_i) \geq \frac{3\cdot 2^{-1/2}\cdot n\cdot\bar{\zeta}}{1-\gamma-3n\bar{\zeta}} \ .$$

Then

$$\|x_i(\bar{\zeta}) - x_i(0)\| \leq \frac{3(n-1)\bar{\zeta}}{2\cdot(1-\gamma-3n\bar{\zeta})\cdot\left(relgap(\lambda_i) - \frac{3\cdot 2^{-1/2}\cdot n\cdot\bar{\zeta}}{1-\gamma-3n\bar{\zeta}}\right)} \ .$$

Proof: The idea is that if ζ is small enough, the $\lambda_k(\zeta)$ can only change by a small relative amount, so the relative gap can only change by a small absolute amount. From Proposition 4, we can bound the perturbed relative gap from below as follows:

$$relgap(\lambda_i(\zeta)) = \min_{k\neq i}\frac{|\lambda_i(\zeta)-\lambda_k(\zeta)|}{|\lambda_i(\zeta)\lambda_k(\zeta)|^{1/2}} \geq \min_{k\neq i}\frac{|\lambda_i(0)-\lambda_k(0)| - \frac{n\zeta}{1-\gamma}(|\lambda_i(0)|+|\lambda_k(0)|)}{|\lambda_i(0)\lambda_i(0)|\left(1-\frac{n\zeta}{1-\gamma}\right)^{-1}}$$

$$= (1-\frac{n\zeta}{1-\gamma})\min_{k\neq i}[(relgap(\lambda_i(0),\lambda_k(0)) - \frac{n\zeta}{1-\gamma}\frac{|\lambda_i(0)| + |\lambda_k(0)|}{|\lambda_i(0)\lambda_k(0)|^{1/2}}]$$

where $relgap(\lambda_i(0),\lambda_k(0)) \equiv |\lambda_i(0)-\lambda_k(0)|\cdot|\lambda_i(0)\lambda_k(0)|^{-1/2}$.

We consider two cases, $relgap(\lambda_i(0),\lambda_k(0)) \geq 2^{-1/2}$, and $relgap(\lambda_i(0),\lambda_k(0)) \leq 2^{-1/2}$. The first case corresponds to $\lambda_i(0)$ and $\lambda_k(0)$ differing by at least a factor of 2, whence

$$\min_{k\neq i}\frac{|\lambda_i(0)| + |\lambda_k(0)|}{|\lambda_i(0)\lambda_k(0)|^{1/2}} \leq 3\cdot relgap(\lambda_i(0),\lambda_k(0))$$

and

$$relgap(\lambda_i(\zeta),\lambda_k(\zeta)) \geq (1-\frac{n\zeta}{1-\gamma})\cdot relgap(\lambda_i(0),\lambda_k(0))\cdot(1-\frac{3n\zeta}{1-\gamma}) \quad .$$

The second case corresponds to $\lambda_i(0)$ and $\lambda_k(0)$ differing by at most a factor of 2, whence

$$\min_{k\neq i}\frac{|\lambda_i(0)| + |\lambda_k(0)|}{|\lambda_i(0)\lambda_k(0)|^{1/2}} \leq 3\cdot 2^{-1/2}$$

and

$$relgap(\lambda_i(\zeta),\lambda_k(\zeta)) \geq (1-\frac{n\zeta}{1-\gamma})\cdot(relgap(\lambda_i(0),\lambda_k(0)) - \frac{3\cdot 2^{-1/2}n\zeta}{1-\gamma}) \quad . \tag{7.4}$$

Altogether, we have

$$relgap(\lambda_i(\zeta)) \geq (1-\frac{n\zeta}{1-\gamma})\cdot(1-\frac{3n\zeta}{1-\gamma})(relgap(\lambda_i(0)) - \frac{(3\cdot 2^{-1/2}n\zeta)/(1-\gamma)}{1 - 3n\zeta/(1-\gamma)}) \quad .$$

Now integrate the bound of Theorem 6 from $\zeta=0$ to $\zeta=\bar{\zeta}$ to get the desired result. □

The next theorem shows that the components of each eigenvector are scaled analogously to the way the matrix is scaled:

Proposition 6: *Let $H=\Delta A\Delta$ be a γ-s.d.d. symmetric matrix with respect to the 2-norm. Let $H=X\Lambda X^T$ be its eigenvector decomposition, where $X=[x_1, \ldots, x_n]$ is the matrix whose columns are orthonormal eigenvectors and $\Lambda=\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $x_i(j)$ be the j-th component of x_i. Then*

$$|x_i(j)| \leq \bar{x}_i(j) \equiv \left(\frac{1+\gamma}{1-\gamma}\right)^{3/2}\cdot\min(\left|\frac{\lambda_i}{\lambda_j}\right|^{1/2}, \left|\frac{\lambda_j}{\lambda_i}\right|^{1/2})$$

We also have

$$|x_i(j)| \leq \left(\frac{1+\gamma}{1-\gamma}\right)^{3/2}\cdot\min(\frac{\Delta_{ii}}{\Delta_{jj}}, \frac{\Delta_{jj}}{\Delta_{ii}})$$

Proof: Let $y_i=\Delta x_i$ for all *i*. First we consider the case $\Delta_{ii} \leq \Delta_{jj}$. From (7.2) we have $\|y_i\| \leq (|\lambda_i|/(1-\gamma))^{1/2}$. Thus, applying Proposition 2 as well,

$$|x_i(j)| = \Delta_{jj}^{-1}|y_i(j)| \leq \Delta_{jj}^{-1}(|\lambda_i|/(1-\gamma))^{1/2} \leq \left(\left|\frac{\lambda_i}{\lambda_j}\right|\cdot\frac{1+\gamma}{1-\gamma}\right)^{1/2}$$

as desired. We may also write

$$|x_i(j)| = \Delta_{jj}^{-1}|y_i(j)| \leq \Delta_{jj}^{-1}(|\lambda_i|/(1-\gamma))^{1/2} \leq \frac{\Delta_{ii}}{\Delta_{jj}}\cdot\left(\frac{1+\gamma}{1-\gamma}\right)^{1/2}$$

to get the other inequality.

Now consider the case $\Delta_{ii} \geq \Delta_{jj}$. We will take the *j*-th components of both sides of the equality $Ay_i = \lambda_i\Delta^{-2}y_i$, and bound them as follows. The left hand side component is bounded

above in absolute value by

$$(1+\gamma)\|y_i\| \le (1+\gamma)(1-\gamma)^{-1/2}|\lambda_i|^{1/2} .$$

The right hand side component is bounded below in absolute value by

$$|\lambda_i|\Delta_{jj}^{-2}|y_i(j)| \ge (1-\gamma)|\lambda_i/\lambda_j|\cdot|y_i(j)| .$$

Thus

$$|x_i(j)| = \Delta_{jj}^{-1}|y_i(j)| \le \Delta_{jj}^{-1}\frac{(1+\gamma)|\lambda_j|}{(1-\gamma)^{3/2}|\lambda_i|^{1/2}} \le \left(\frac{1+\gamma}{1-\gamma}\right)^{3/2}\cdot\left|\frac{\lambda_j}{\lambda_i}\right|^{1/2}$$

as desired. The bound in terms of Δ_{jj}/Δ_{ii} is obtained similarly. □

Thus, if a matrix is strongly scaled (the Δ_{jj} vary greatly in magnitude), the eigenvectors will be strongly scaled, and small relative perturbations in the matrix entries will not be able to change the smaller eigenvector components much. In the next theorem, we prove something even stronger: the perturbations in the eigenvector components $x_i(j)$ will be small compared to the upper bounds $\bar{x}_i(j)$:

Theorem 7: *Let $H(\zeta)$ be as in Theorem 6. Let $x_i(\zeta)(j)$ denote the j-th component of the i-th unit eigenvector of $H(\zeta)$. Let $\bar{x}_i(j)$ be the upper bound for the j-th component of the unperturbed unit eigenvector $x_i(0)(j)$. Then*

$$|x_i(\zeta)(j) - x_i(0)(j)| \le \zeta\cdot\frac{2^{1/2}(n-1)}{(1-\gamma)\cdot\min(relgap(\lambda_i)\ ,\ 2^{-1/2})}\cdot\bar{x}_i(j) + O(\zeta^2)$$

(λ_i is the same as $\lambda_i(0)$). Note that $relgap(\lambda_i)$ exceeds $2^{-1/2}$ only when λ_i differs from its nearest neighbor by a factor greater than 2 or less than 1/2.

Proof: We start from (7.3) in the proof of Theorem 6; it implies

$$\begin{aligned}
|x_i(\zeta)(j)-x_i(0)(j)| &= \left|\zeta\sum_{k\ne i}\frac{\xi_{ij}z_i^T E z_k}{|\lambda_i-\lambda_k|/|\lambda_i\lambda_k|^{1/2})}\cdot x_k(0)(j) + O(\zeta^2)\right| \\
&\le \frac{\zeta(n-1)(1+\gamma)^{3/2}}{(1-\gamma)^{5/2}}\cdot\frac{|\lambda_i\lambda_k|^{1/2}}{|\lambda_i-\lambda_k|}\cdot\min\left(\left|\frac{\lambda_k}{\lambda_j}\right|^{1/2}\ ,\ \left|\frac{\lambda_j}{\lambda_k}\right|^{1/2}\right) + O(\zeta^2) \\
&= \frac{\zeta(n-1)(1+\gamma)^{3/2}}{(1-\gamma)^{5/2}}\cdot\min\left(\left|\frac{\lambda_i}{\lambda_j}\right|^{1/2}\frac{|\lambda_k|}{|\lambda_i-\lambda_k|}\ ,\ \left|\frac{\lambda_j}{\lambda_i}\right|^{1/2}\frac{|\lambda_i|}{|\lambda_i-\lambda_k|}\right) + O(\zeta^2) \\
&\le \frac{\zeta(n-1)(1+\gamma)^{3/2}}{(1-\gamma)^{5/2}}\cdot\min\left(\left|\frac{\lambda_i}{\lambda_j}\right|^{1/2}\ ,\ \left|\frac{\lambda_j}{\lambda_i}\right|^{1/2}\right)\frac{\max(|\lambda_i|\ ,\ |\lambda_k|)}{|\lambda_i-\lambda_k|} + O(\zeta^2) \\
&\le \frac{\zeta(n-1)(1+\gamma)^{3/2}}{(1-\gamma)^{5/2}}\cdot\min\left(\left|\frac{\lambda_i}{\lambda_j}\right|^{1/2}\ ,\ \left|\frac{\lambda_j}{\lambda_i}\right|^{1/2}\right)\frac{2^{1/2}}{\min(relgap(\lambda_i)\ ,\ 2^{-1/2})} + O(\zeta^2)
\end{aligned}$$

as desired. □

A version of Theorem 7 for nonasymptotically small ζ can be proven in the same way as Corollary 3 was derived from Theorem 6.

If we have a cluster of eigenvalues which are relatively well separated from the others, similar analyses to those above show that the invariant subspace they span is insensitive to perturbations, even if the individual eigenvectors are sensitive.

Now consider s.d.d. definite pencils. From Proposition 5 of the last section, we know we can often reduce such pencils to standard form without sacrificing diagonal dominance: Given $H-\lambda M$, M positive definite with lower triangular Cholesky factor L, the matrix $Y=L^{-1}HL^{-T}$ has the same spectrum as $H-\lambda M$. Also, if y is an eigenvector of Y, $x=L^{-T}y$ is an eigenvector

of $H-\lambda M$; thus Theorems 6 and 7 can be used to derive perturbation bounds on x, although we will not do so here.

Finally, we consider perturbation theory for the singular vectors of bidiagonal matrices. Let

$$B = \begin{bmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \cdot & \cdot & \\ & & & \cdot & b_{n-1} \\ & & & & a_n \end{bmatrix} . \tag{7.5}$$

be bidiagonal, where as in Theorem 1 we may assume without loss of generality that all a_i and b_i are positive. Recall that the left singular vectors of B are the eigenvectors of BB^T and the right singular vectors of B are the eigenvectors of B^TB. Since

$$BB^T = \begin{bmatrix} a_1^2+b_1^2 & b_1a_2 & & & \\ b_1a_2 & a_2^2+b_2^2 & b_2a_3 & & \\ & b_2a_3 & \cdot & \cdot & \\ & & \cdot & a_{n-1}^2+b_{n-1}^2 & b_{n-1}a_n \\ & & & b_{n-1}a_n & a_n^2 \end{bmatrix}$$

and

$$B^TB = \begin{bmatrix} a_1^2 & b_1a_1 & & & \\ b_1a_1 & a_2^2+b_1^2 & b_2a_2 & & \\ & b_2a_2 & \cdot & \cdot & \\ & & \cdot & a_{n-1}^2+b_{n-2}^2 & b_{n-1}a_{n-1} \\ & & & b_{n-1}a_{n-1} & a_n^2+b_{n-1}^2 \end{bmatrix}$$

small relative perturbations in the a_i and b_i only cause small relative perturbations in the entries of BB^T and B^TB. Therefore, we can reduce perturbation theory for the singular vectors of a bidiagonal matrix to perturbation theory for the tridiagonal matrices BB^T and B^TB.

Proposition 7: *Let B be as in (7.5). Since BB^T and B^TB are positive definite tridiagonal, and hence s.d.d., small relative changes in the entries of B cause perturbations in the singular vectors as described by Theorems 6 and 7. More specifically, let $D_L = \mathrm{diag}((BB^T)_{ii})$ and $D_R = \mathrm{diag}((B^TB)_{ii})$. Then*

$$\|D_L^{-1/2}BB^TD_L^{-1/2} - I\| \le \gamma_L \equiv 2\cdot\max\left(\max_{j<n-1} \frac{b_ja_{j+1}}{(a_j^2+b_j^2)^{1/2}(a_{j+1}^2+b_{j+1}^2)^{1/2}}\ ,\ \frac{b_{n-1}}{(a_{n-1}^2+b_{n-1}^2)^{1/2}}\right)$$

and

$$\|D_R^{-1/2}B^TBD_R^{-1/2} - I\| \le \gamma_R \equiv 2\cdot\max\left(\max_{j>1} \frac{a_jb_j}{(a_j^2+b_{j-1}^2)^{1/2}(a_{j+1}^2+b_j^2)^{1/2}}\ ,\ \frac{b_1}{(a_2^2+b_1^2)^{1/2}}\right) .$$

Both γ_L and γ_R are bounded by 2. If $a_j > 3^{1/2}b_j$, then $\gamma_L < 1$, and if $a_j > 3^{1/2}b_{j-1}$, then $\gamma_R < 1$.

Proof: A simple computation shows that the diagonal of $D_L^{-1/2}BB^TD_L^{-1/2}$ consists of ones and that the offdiagonals are $b_ja_{j+1}(a_j^2+b_{j-1}^2)^{-1/2}(a_{j+1}^2+b_j^2)^{-1/2}$ for $j<n-1$ and $b_{n-1}\cdot(a_{n-1}^2+b_{n-1}^2)^{-1/2}$ for $j=n-1$. Thus γ_L bounds the 1-norm of $D_L^{-1/2}BB^TD_L^{-1/2} - I$. A similar computation applies to B^TB. □

Thus, the sensitivity of the singular vectors to relative perturbations in the entries of B is governed by the relative gap between singular values, as conjectured in [9]. Actually, more was conjectured: it does not appear that the measure of diagonal dominance γ of BB^T or B^TB

affects the singular vector sensitivity. A proof of this stronger conjecture will appear in [6]. More precisely, a version of Theorem 6 is proved in [6] without the $1-\gamma$ factor in the denominator, and extended to nonasymptotically small perturbations as in Corollary 3. Whether Theorem 7 can be generalized without a $1/(1-\gamma)$ dependence is an open question.

8. On Condition Numbers and the Distance to the Nearest Ill-Posed Problem

In [7] it was observed that a common feature of many numerical analysis problems is that their condition numbers approximate or at least bound the reciprocal of the distance to the nearest *ill-posed* problem, i.e. problem whose condition number is infinite. The classical example of this is matrix inversion, where the condition number of a matrix T is $\kappa(T) \equiv \|T\| \cdot \|T^{-1}\|$. By scaling T we may assume without loss of generality that $\|T\| = 1$, so that $\kappa(T) = \|T^{-1}\|$. But the distance (in the $\|\cdot\|$ norm) from T to the nearest singular matrix (nearest matrix whose condition number is infinite) is $\|T^{-1}\| = 1/\kappa(T)$. Thus, the condition number is exactly the reciprocal of the distance to the nearest ill-posed problem. This phenomenon recurs throughout numerical analysis, although we usually get a somewhat weaker relationship, such as one sided bounds between the condition number and reciprocal distance.

Here we investigate this phenomenon in the case of finding the eigenvectors of a symmetric matrix. From (1.3), we see that $1/gap(\lambda_i)$ is a condition number for the i-th eigenvector of a general symmetric matrix. This condition number is infinite precisely when $gap(\lambda_i) = 0$, i.e. λ_i is a multiple eigenvalue. It is reasonable to call such an eigenproblem ill-posed because the eigenvector is no longer uniquely determined: any vector in an at least two-dimensional invariant subspace will do. It is elementary to show that the reciprocal of this condition number gives exactly the distance to the nearest ill-posed problem:

Proposition 8: *Let H be a symmetric matrix with simple eigenvalue λ_i; thus $gap(\lambda_i) = \min_{k \neq i} |\lambda_i - \lambda_k| > 0$. Then the smallest $\|\delta H\|$ such that the eigenvalue of $H + \delta H$ "corresponding to" λ_i is multiple is*

$$\min \|\delta H\| = \frac{gap(\lambda_i)}{2} \quad .$$

By "the eigenvalue corresponding to λ_i is multiple" we mean that if the continuous function $\lambda_i(\xi)$ is an eigenvalue of $H + \xi\delta H$ for all $0 \leq \xi \leq 1$, with $\lambda_i(0) = \lambda_i$, then $\lambda_i(\xi)$ is simple for $0 \leq \xi < 1$ and $\lambda_i(1)$ is multiple.

Proof: Suppose $|\lambda_j - \lambda_i| = gap(\lambda_i)$ and that x_i and x_j are corresponding unit eigenvectors. Let $\delta H = .5 \cdot (\lambda_j - \lambda_i) \cdot (x_i x_i^T - x_j x_j^T)$ to show $\min \|\delta H\| \leq gap(\lambda_i)/2$. To get the other inequality apply (1.1) to see that any smaller $\|\delta H\|$ could not move either λ_i or λ_j more than half the distance towards one another. □

It turns out that a similar relationship holds for γ-s.d.d. symmetric matrices, provided we measure distances in the scaled way used so far in this paper, and that we use $1/relgap(\lambda_i)$ as a condition number. This is interesting because it extends work in [7] to a case where the distance metric is quite skewed from the usual norm, and shows that $1/relgap(\lambda_i)$ is the most natural condition number for this problem, because it shares the same geometric properties as other condition numbers.

Proposition 9: *Let $H=\Delta A\Delta$ be an n by n γ-s.d.d. symmetric matrix with simple eigenvalue λ_i; thus $relgap(\lambda_i) = \min_{k \neq i} |\lambda_i - \lambda_k| \cdot |\lambda_i \lambda_k|^{-1/2} > 0$. Assume further that $relgap(\lambda_i) \leq 2^{-1/2}$ (this means that λ_i and its nearest neighbor differ by a factor between .5 and 2). Then the smallest $\|\delta A\|$ such that the eigenvalue of $\Delta(A+\delta A)\Delta$ "corresponding to" λ_i is multiple satisfies*

$$\frac{(1-\gamma)\,relgap(\lambda_i)}{3(\sqrt{2}\cdot n + relgap(\lambda_i))} \leq \min \|\delta A\| \leq relgap(\lambda_i)\cdot 2^{1/2}\cdot n \cdot \frac{(1+\gamma)^4}{(1-\gamma)^3} \quad .$$

For $relgap(\lambda_i) << 1$, the lower bound on $\min \|\delta A\|$ equals $relgap(\lambda_i)(1-\gamma)/(2^{1/2}\cdot 3\cdot n) + O((relgap(\lambda_i))^2)$. In other words, both the upper and lower bounds are $\Theta(relgap(\lambda_i))$.

Proof: By scaling H we may assume without loss of generality that $\Delta_{ii} = 1$. Let λ_j satisfy $relgap(\lambda_i) = |\lambda_i - \lambda_j| \cdot |\lambda_i \lambda_j|^{1/2}$, and let x_i and x_j be corresponding unit eigenvectors.

First we prove the upper bound on $\min\|\delta A\|$. From Proposition 6 we have that $|(x_i x_i^T)_{kl}| \leq (1+\gamma)^3(1-\gamma)^{-3}\Delta_k\Delta_l$. Thus $\|\Delta^{-1}x_i x_i^T \Delta^{-1}\| \leq n(1+\gamma)^3(1-\gamma)^{-3}$. Let $\delta A = (\lambda_j - \lambda_i)x_i x_i^T$. Clearly $\Delta(A+\delta A)\Delta$ has a multiple eigenvalue at λ_j as desired. Also,

$$|\lambda_j - \lambda_i| = relgap(\lambda_i)\cdot|\lambda_i\lambda_j|^{1/2} \leq 2^{1/2} relgap(\lambda_i)\cdot|\lambda_i| \leq 2^{1/2}\cdot(1+\gamma)\cdot relgap(\lambda_i) \quad ,$$

which when combined with the bound on $\|\Delta^{-1}x_i x_i^T\Delta^{-1}\|$, yields the desired upper bound.

Now we consider the lower bound. Abbreviate $\|\delta A\|$ by η. Note that if H is γ-s.d.d., then $H+\Delta\delta A\Delta$ is $(\gamma+2\eta)(1-\eta)^{-1}$-s.d.d., if $(\gamma+2\eta)(1-\eta)^{-1} < 1$. From (7.4) in Corollary 3, we see that if $(\gamma+2\eta)(1-\eta)^{-1} < 1$, then the perturbed relative gap will be at least

$$relgap(\lambda_i) - \frac{3\cdot 2^{1/2}\cdot n\eta}{1 - (\gamma+2\eta)(1-\eta)^{-1}} \quad .$$

In order for the perturbed relative gap to be zero, this lower bound will have to be nonpositive, and so

$$\eta \geq \frac{relgap(\lambda_i)}{3\cdot 2^{1/2}\cdot n}\cdot(1 - (\gamma+2\eta)(1-\eta)^{-1})$$

Solving for the smallest η satisfying this inequality yields the desired lower bound. When $relgap(\lambda_i) << 1$ so that η is small enough that $(\gamma+2\eta)(1-\eta)^{-1} \approx \gamma$, we get $\eta \approx relgap(\lambda_i)(1-\gamma)/(3\cdot 2^{1/2}\cdot n)$ as desired. □

The same results could have been obtained using the general machinery of differential inequalities in [7], but these proofs are more straightforward.

9. Algorithms for the Bidiagonal Singular Value Decomposition

In this section we discuss algorithms capable of attaining the high relative accuracy inherent in the data as described in Theorem 1. Most of this work has appeared elsewhere, but since we will need the results in the next section, we outline them here.

The three classes of algorithms we will discuss are QR, bisection, and divide and conquer. The standard QR iteration [12] as implemented in LINPACK [3] does not compute all singular values to high relative precision. It may be modified, however, to achieve this as described in [9]. Briefly, the idea is to use a zero shift in a QR sweep when a tiny singular value is present. It turns out this zero-shift QR can be implemented in a forward stable way that only introduces small relative errors in each entry of the bidiagonal matrix. Corollary 1 of Theorem 1 then guarantees the singular values are not changed significantly. The standard convergence criterion must also be changed to guarantee high relative accuracy; see [9] for details. The resulting algorithm is not only more accurate than the standard implementation but faster on the average; this is because the zero-shift QR sweep contains significantly fewer floating point operations than shifted QR.

It was conjectured in [9] that this modified QR algorithm computes the singular vectors as accurately as the "relative gap" error bounds of section 7 permit; this conjecture is supported by Proposition 7 and will be proven completely in [6] (see section 7 for discussion).

Bisection is another method that guarantees high relative accuracy. An error analysis of the Sturm sequence recurrence for counting the number of singular values of a bidiagonal matrix in an interval [14] shows that it computes the exact number of singular values for a matrix differing from the original one only by small relative perturbations in each entry; Corollary 1 of Theorem 1 then guarantees high relative accuracy again.

Divide and conquer [13] has not yet been shown to achieve high relative accuracy, at least without resorting to extended precision arithmetic in the inner loop. Achieving this accuracy is a current area of research.

10. Algorithms for the Symmetric Tridiagonal Eigenproblem

In this section we present algorithms for computing eigenvalues of γ-s.d.d. symmetric tridiagonal matrices to high relative accuracy. We note that reducing a dense γ-s.d.d. symmetric matrix to tridiagonal form will not generally preserve diagonal dominance or the accuracy of the eigenvalues; thus the algorithms in this section are suitable only when the original matrix is tridiagonal. If the original matrix is dense, the algorithm in the next section should be used.

Briefly, bisection can always be used to find the eigenvalues accurately. If the matrix is positive definite as well, Cholesky followed by the algorithm of the last section applied to the bidiagonal Cholesky factor can be used. QR does not seem to work in general, but may if the matrix is strongly diagonally dominant and monotonically graded. It is still an open question whether divide and conquer techniques [5, 11] can achieve high relative accuracy.

As with the bidiagonal singular value problem, the standard implementation of QR for tridiagonal matrices does not guarantee high relative accuracy in the computed eigenvalues for symmetric γ-s.d.d. matrices, even if the change in convergence criterion suggested in section 6 is adopted. Even if a zero-shifted QR algorithm like the one used for the bidiagonal singular value decomposition is used, relative accuracy is lost (in numerical experiments on a strongly s.d.d. positive definite matrix, negative eigenvalues were computed).

However, recent work by Le and Parlett [16] gives some hope that zero-shifted tridiagonal QR may sometimes be used in the same way as the bidiagonal zero-shifted QR to compute all eigenvalues to high relative accuracy. Their work shows that the inner loop of the standard QR iteration may be modified to provide componentwise relative stability in the following sense: in floating point this modified zero-shifted QR is equivalent to making small relative perturbations in each entry of the tridiagonal matrix, performing QR *exactly* on this perturbed matrix, and again making small relative perturbations in each entry of the resulting matrix. This is a much stronger kind of stability than the usual kind described in paragraph (1.2) of the introduction. If

the original and final tridiagonals are also γ-s.d.d., Proposition 4 would imply that their eigenvalues agree to high relative accuracy. Thus, QR, combined with the stopping criterion (6.4), could be used to compute all eigenvalues to high relative accuracy, but only if all the matrices produced in the course of the iteration were γ-s.d.d.. Unfortunately, numerical experiments show this is not the case in general, and may only be true if the original matrix is very strongly diagonally dominant and monotonically scaled ($H_{ii} \geq H_{i+1,i+1}$). Thus, we do not expect to be able to generally use QR based algorithms for the γ-s.d.d. tridiagonal eigenproblem.

If we limit ourselves to positive definite matrices T, the following QR based approach will work. The following algorithm originally appeared in [10]. Recall that a tridiagonal matrix T is positive definite if and only if it has a positive diagonal and is γ-s.d.d. for some $\gamma < 1$.

Algorithm 2: *Computing the eigenvalues of a positive definite tridiagonal matrix T:*

1) Compute the Cholesky factorization $LL^T = T$ of T.
2) Find the singular values of the bidiagonal matrix L using the bidiagonal QR algorithm of section 9.
3) Square the singular values of L to get the eigenvalues of T.

To show that this method is viable, one needs to show that scaled diagonal dominance is sufficient to guarantee that the squares of the exact singular values of L are all relatively close to the eigenvalues of A; the algorithms of section 9 then guarantee that we can compute the singular values of L to high relative accuracy.

To this end we present a backward error analysis of the Cholesky decomposition of a positive definite symmetric tridiagonal matrix A. Our goal is to show that the computed Cholesky factor L is a small componentwise relative perturbation of the exact Cholesky factor of a small componentwise relative perturbation of A. We assume the usual model of floating point arithmetic $fl(a \; op \; b) = (a \; op \; b)\cdot(1+e)$, where $|e| \leq \varepsilon$ and $op \in \{+,-,\times,/\}$, and that the floating point square root function *sqrt* satisfies $sqrt(a) = (1+e)a^{1/2}$ where $|e| \leq \varepsilon$.

Proposition 10: *Let A be an n by n positive definite symmetric tridiagonal matrix with diagonal entries $a_1, \ldots, a_n$ and offdiagonal entries $b_1, \ldots, b_{n-1}$. Let L be the computed Cholesky factor from the following algorithm:*

$l_{11} = sqrt(a_1)$
for $i = 1$ *to* $n-1$
  $l_{i+1,i} = b_i / l_{ii}$
  $l_{i+1,i+1} = sqrt(a_{i+1} - l_{i+1,i}^2)$
endfor

Then barring over/underflow and attempts to take square roots of negative arguments, L is the exact Cholesky factor of $\hat{A} = A + \delta A = LL^T$, where $|\delta A_{ij}| \leq g(\varepsilon)|A_{ij}|$, and

$$g(\varepsilon) = 3\varepsilon + 3\varepsilon^2 + \varepsilon^3 + (4\varepsilon + 6\varepsilon^2 + 4\varepsilon^3 + \varepsilon^4)\cdot\frac{(1+\varepsilon)^3}{(1-\varepsilon)^4} = 7\varepsilon + O(\varepsilon^2) \qquad (10.1)$$

Proof: We construct $\hat{A} = A + \delta A$ as follows. Subscripted εs denote independent quantities bounded in norm by ε.

$$l_{11} = fl(sqrt(a_1)) = (1+\varepsilon_{11})a_1^{1/2} = ((1+\varepsilon_{11})^2 a_1)^{1/2} \equiv \hat{a}_1^{1/2}$$

$$l_{i,i-1} = fl(b_{i-1}/l_{i-1,i-1}) = (1+\varepsilon_{i1})b_{i-1}/l_{i-1,i-1} \equiv \hat{b}_{i-1}/l_{i-1,i-1}$$

$$l_{i,i} = fl(sqrt(a_i - l_{i,i-1}^2))$$

$$= (1+\varepsilon_{i2})\cdot((1+\varepsilon_{i3})(a_i-(1+\varepsilon_{i4})l_{i,i-1}^2))^{1/2}$$

$$= ((1+\varepsilon_{i2})^2(1+\varepsilon_{i3})a_i - (1+\varepsilon_{i2})^2(1+\varepsilon_{i3})(1+\varepsilon_{i4})l_{i,i-1}^2))^{1/2}$$

$$\equiv ((1+g_{i1})a_i - (1+g_{i2})l_{i,i-1}^2)^{1/2}$$

where $|g_{i1}| \le 3\varepsilon+3\varepsilon^2+\varepsilon^3 = 3\varepsilon+O(\varepsilon^2)$ and $|g_{i2}| \le 4\varepsilon+6\varepsilon^2+4\varepsilon^3+\varepsilon^4 = 4\varepsilon+O(\varepsilon^2)$. By assumption $(1+g_{i1})a_i < (1+g_{i2})l_{i,i-1}^2$ so

$$|g_{i2}l_{i,i-1}^2| \le |g_{i2}|\cdot\frac{1+g_{i1}}{1+g_{i2}}\cdot a_i \equiv g_{i3}a_i$$

where

$$|g_{i3}| \le (4\varepsilon+6\varepsilon^2+4\varepsilon^3+\varepsilon^4)\cdot\frac{(1+\varepsilon)^3}{(1-\varepsilon)^4} = 4\varepsilon+O(\varepsilon^2)$$

Thus

$$l_{i,i} = ((1+g_{i1}-g_{i3})a_i-l_{i,i+1}^2)^{1/2}$$

$$\equiv ((1+g_{i4})a_i-l_{i,i+1}^2)^{1/2}$$

$$\equiv (\hat{a}_i-l_{i,i+1}^2)^{1/2}$$

where

$$|g_{i4}| \le 3\varepsilon+3\varepsilon^2+\varepsilon^3+(4\varepsilon+6\varepsilon^2+4\varepsilon^3+\varepsilon^4)\frac{(1+\varepsilon)^3}{(1-\varepsilon)^4} = 7\varepsilon+O(\varepsilon^2)$$

as desired. □

Theorem 8: *Let A be an n by n positive definite symmetric tridiagonal matrix, L its computed Cholesky factor as in Proposition 10, and $g(\varepsilon)$ as in (10.1). Let $\gamma < 1$ be the scaled diagonal dominance of A. Let $\sigma_1 \le \cdots \le \sigma_n$ be the singular values of L and $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of A. Then*

$$(1-\frac{g(\varepsilon)}{1-\gamma}) \le \frac{\lambda_i(A)}{\sigma_i^2(L)} \le (1+\frac{g(\varepsilon)}{1-\gamma}) \quad .$$

For example, when $\varepsilon < .001$, the upper and lower bounds are bounded by $1\pm[\frac{7.04\varepsilon}{1-\gamma}]$.

Proof: Combine Corollary 1 of Theorem 1, Theorem 2 and Proposition 10. □

It is easy to find T for which Algorithm 2 computes all the eigenvalues accurately, but where the standard QR iteration [17] loses all accuracy on the smallest eigenvalues.

Since the bidiagonal SVD algorithm of section 9 can compute the singular vectors as accurately as the "relative gap" error bound permits (see the discussion of section 9), Algorithm 2 will compute the eigenvectors of T as accurately as Theorem 6 permits.

Bisection is a viable algorithm for all s.d.d. tridiagonal matrices. A similar error analysis to the one for bidiagonal Sturm sequences shows that Sturm sequences can find accurate eigenvalues of $T+\delta T$ where δT causes only small relative perturbations in each entry of T [14]. No pivoting is required, so Sturm sequence evaluation can be done in linear time with no storage beyond that needed for T. Together with Theorems 2 or 4, this implies that the computed eigenvalues all have high relative accuracy if the matrix is symmetric tridiagonal γ-s.d.d.

As in the bidiagonal case, the ability of divide and conquer algorithms [11] to achieve high relative accuracy is an open problem.

11. Algorithms for the Dense Symmetric Eigenproblem

First we present a new algorithm (or rather a new analysis of an old algorithm) for finding accurate eigenvalues of (possibly dense) symmetric s.d.d. matrices. The algorithm is based on bisection, but rather than computing the LDL^T factorization of $H - xI$ and using the number of negative D_{ii} to compute the number of eigenvalues of H less than x, we compute the LDL^T factorization of $A - x\Delta^{-2}$, where $H = \Delta A \Delta$.

Second, we show that a suitable variation of inverse iteration can compute the eigenvectors of a symmetric s.d.d. matrix to the limiting accuracy of the "relative gap" error bounds in Theorems 6 and 7.

Algorithm 3: *Stably computing the inertia of a shifted γ-scaled diagonally dominant symmetric matrix $H - xI$:*

(0) We assume as before that $A_{ii} = \pm 1$. We consider only $x > 0$; for $x < 0$ consider $-H - xI$.

(1) Permute the rows and columns of $A - x\Delta^{-2}$ and partition it as

$$\begin{bmatrix} A_{11} - x\Delta_1^{-2} & A_{12} \\ A_{21} & A_{22} - x\Delta_2^{-2} \end{bmatrix}$$

so that if $a - xd^{-2}$ is a diagonal entry of $A_{11} - x\Delta_1^{-2}$, then either $a = -1$ or $a = 1$ and $xd^{-2} \geq 2$.

(2) Compute $X = A_{22} - x\Delta_2^{-2} - A_{21}(A_{11} - x\Delta_1^{-2})^{-1} A_{12}$.

(3) Compute inertia$(X) = (n, z, p)$ using a stable pivoting scheme such as in [4]. Here n is the number of negative eigenvalues, z the number of zero eigenvalues, and p the number of positive eigenvalues of X.

(4) The inertia of $H - xI$ is $(n + \dim(A_{11}), z, p)$.

Theorem 9: *Let $H = \Delta A \Delta$ be a γ-scaled diagonally dominant symmetric matrix and $x > 0$ a real scalar. Algorithm 3 computes the exact inertia of a matrix $H + \delta H - xI$, where $\delta H = \Delta \delta A \Delta$, $\|\delta A\| = O(\varepsilon)$, ε being the machine precision. Thus, Algorithm 3 can be used in a bisection algorithm to find all the eigenvalues of H to high relative accuracy.*

Proof: The partitioning guarantees that the diagonal entries of $A_{11} - x\Delta_1^{-2}$ are all less than or equal to -1. Therefore, all the eigenvalues of $A_{11} - x\Delta_1^{-2}$ are less than or equal to $-1 + \gamma$. Since X is defined so that

$$\begin{bmatrix} A_{11} - x\Delta_1^{-2} & A_{12} \\ A_{21} & A_{22} - x\Delta_2^{-2} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_{21}(A_{11} - x\Delta_1^{-2})^{-1} & I \end{bmatrix} \cdot \begin{bmatrix} A_{11} - x\Delta_1^{-2} & 0 \\ 0 & X \end{bmatrix} \cdot \begin{bmatrix} I & (A_{11} - x\Delta_1^{-2})^{-1} A_{12} \\ 0 & I \end{bmatrix},$$

the inertia of $A - x\Delta^{-2}$ is equal to

$$\text{inertia}(A - x\Delta^{-2}) = \text{inertia}(X) + \text{inertia}(A_{11} - x\Delta_1^{-2}) = \text{inertia}(X) + (\dim(A_{11}), 0, 0)$$

by Sylvester's Theorem. The algorithm in [4] will compute the exact inertia of $X + \delta X$, where $\|\delta X\| = O(\varepsilon)\|X\|$. Thus if we show that $\|X\|$ is of order $1 - \gamma \leq \|A_{22}\| \leq 1 + \gamma$, we will be done. To this end we note that by construction $\|x\Delta_2^{-2}\| \leq 2$, $\sigma_{\min}(A_{11} - x\Delta_1^{-2}) \geq 1 - \gamma$, and $\|A_{12}\| = \|A_{21}\| \leq \gamma$. Thus

$$\|X\| \leq 1 + \gamma + 2 + \frac{\gamma^2}{1 - \gamma} \leq \frac{3}{1 - \gamma}$$

as desired. □

In the case of pencils, there is an analogous algorithm. If $H - \lambda M = \Delta_H A_H \Delta_H - \lambda \Delta_M A_M \Delta_M$ is a γ-s.d.d. definite pencil, we compute the LDL^T decomposition of $A_H - x\Delta_H^{-1}\Delta_M A_M \Delta_M \Delta_H^{-1}$ in order to count the number of eigenvalues less than x. Henceforth we will assume without loss of generality that $H - \lambda M = H - \lambda \Delta A \Delta$ with $|H_{ii}| = A_{ii} = 1$.

Algorithm 4: *Stably computing the inertia of a shifted γ-scaled diagonally dominant definite pencil $H-xM$:*

(0) As stated above, we assume without loss of generality that $M=\Delta A\Delta$ with $|H_{ii}| = A_{ii} = 1$. We also consider only $x > 0$; for $x < 0$ consider $-H-xM$.

(1) Permute the rows and columns of $H-x\Delta A\Delta$ and partition it as

$$\begin{bmatrix} H_{11}-x\Delta_1^{-1}A_{11}\Delta_1^{-1} & H_{12}-x\Delta_1^{-1}A_{12}\Delta_2^{-1} \\ H_{21}-x\Delta_2^{-1}A_{21}\Delta_1^{-1} & H_{22}-x\Delta_2^{-1}A_{22}\Delta_2^{-1} \end{bmatrix}$$

so that if $h-xd^{-2}$ is a diagonal entry of $H_{11}-x\Delta_1^{-1}A_{11}\Delta_1^{-1}$, then $xd^{-2} \geq \mu \equiv 2(1+\gamma)/(1-\gamma)$.

(2) Compute

$$X = H_{22}-x\Delta_2^{-1}A_{22}\Delta_2^{-1} - (H_{21}-x\Delta_2^{-1}A_{21}\Delta_1^{-1})\cdot(H_{11}-x\Delta_1^{-1}A_{11}\Delta_1^{-1})^{-1}\cdot(H_{12}-x\Delta_1^{-1}A_{12}\Delta_2^{-1})$$

(3) Compute inertia$(X) = (n\ ,\ z\ ,\ p)$ using a stable pivoting scheme such as in [4].

(4) The inertia of $H-xM$ is $(n+\dim(A_{11})\ ,\ z\ ,\ p)$.

Theorem 10: *Let $H-\lambda M = \Delta_H A_H \Delta_H - \lambda \Delta_M A_M \Delta_M$ be a γ-s.d.d. definite pencil and $x > 0$ a real scalar. Algorithm 4 computes the exact inertia of $H+\delta H-xM$, where $\delta H = \Delta_A \delta A \Delta_A$, $\|\delta A\| = O(\varepsilon)$, ε being the machine precision. Thus, Algorithm 4 can be used in a bisection algorithm to find all the eigenvalues of $H-\lambda M$ to high relative accuracy.*

Proof: Let $Y = H_{11}-x\Delta_1^{-1}A_{11}\Delta_1^{-1}$, and define $K=x^{-1}\Delta_1 H_{11}\Delta_1 - A_{11}$, so that $Y = x\Delta_1^{-1}K\Delta_1^{-1}$. Now if $\lambda(K)$ is any eigenvalue of K, we have

$$\lambda(K) \leq \lambda_{\max}(x^{-1}\Delta_1 H_{11}\Delta_1) - \lambda_{\min}(A_{11}) \leq \mu^{-1}(1+\gamma) - (1-\gamma) = -\frac{1-\gamma}{2}$$

and $\|K^{-1}\| \leq 2/(1-\gamma)$. Thus, Y is also nonsingular, with all negative eigenvalues. Since X and Y are defined so that

$$\begin{bmatrix} H_{11}-x\Delta_1^{-1}A_{11}\Delta_1^{-1} & H_{12}-x\Delta_1^{-1}A_{12}\Delta_2^{-1} \\ H_{21}-x\Delta_2^{-1}A_{21}\Delta_1^{-1} & H_{22}-x\Delta_2^{-1}A_{22}\Delta_2^{-1} \end{bmatrix} =$$

$$\begin{bmatrix} I & 0 \\ (H_{21}-x\Delta_2^{-1}A_{21}\Delta_1^{-1})Y^{-1} & I \end{bmatrix} \cdot \begin{bmatrix} Y & 0 \\ 0 & X \end{bmatrix} \cdot \begin{bmatrix} I & Y^{-1}(H_{12}-x\Delta_1^{-1}A_{12}\Delta_2^{-1}) \\ 0 & I \end{bmatrix},$$

we have

$$\text{inertia}(H-xM) = \text{inertia}(X) + \text{inertia}(Y) = \text{inertia}(X) + (\dim(A_{11}), 0, 0)$$

by Sylvester's Theorem. The algorithm in [4] will compute the exact inertia of $X+\delta X$, where $\|\delta X\| = O(\varepsilon)\|X\|$. Thus if we show that $\|X\|$ is of order $1-\gamma \leq \|H_{22}\| \leq 1+\gamma$, we will be done. To this end we write

$$\begin{aligned} \|X\| \leq\ & \|H_{22}\| + \|x\Delta_2^{-1}A_{22}\Delta_2^{-1}\| + \|H_{21}Y^{-1}H_{12}\| + \|x\Delta_2^{-1}A_{21}\Delta_1^{-1}Y^{-1}H_{12}\| \\ & + \|H_{21}Y^{-1}x\Delta_1^{-1}A_{12}\Delta_2^{-1}\| + \|x\Delta_2^{-1}A_{21}\Delta_1^{-1}Y^{-1}x\Delta_1^{-1}A_{12}\Delta_2^{-1}\| \\ =\ & \|H_{22}\| + \|x\Delta_2^{-1}A_{22}\Delta_2^{-1}\| + \|H_{21}x^{-1}\Delta_1 K^{-1}\Delta_1 H_{12}\| + \|\Delta_2^{-1}A_{21}K^{-1}\Delta_1 H_{12}\| \\ & + \|H_{21}\Delta_1 K^{-1}A_{12}\Delta_2^{-1}\| + \|x\Delta_2^{-1}A_{21}K^{-1}A_{12}\Delta_2^{-1}\| \ . \end{aligned}$$

Using the facts that $\|A_{12}\| \leq \gamma$, $\|A_{21}\| \leq \gamma$, $\|H_{12}\| \leq \gamma$, $\|H_{21}\| \leq \gamma$, $\|K^{-1}\| \leq 2/(1-\gamma)$, $\|\Delta_1\|\cdot\|\Delta_2^{-1}\| \leq 1$, $x^{-1}\|\Delta_1\|^2 \leq \mu^{-1}$, and $x\|\Delta_2^{-1}\|^2 \leq \mu$, we get

$$\|X\| \leq 1+\gamma + (1+\gamma)\mu + \gamma^2(2/(1-\gamma))\mu^{-1} + \gamma^2(2/(1-\gamma)) + \gamma^2(2/(1-\gamma)) + \gamma^2(2/(1-\gamma))\mu$$

$\leq 14/(1-\gamma)^2$

as desired. □

Now we present a variation on inverse iteration which can compute the eigenvectors of a symmetric s.d.d. matrix to the limiting accuracy of the "relative gap" error bounds of Theorems 6 and 7. A similar algorithm applies to pencils:

Algorithm 5: *Inverse iteration for computing the eigenvector x of a symmetric s.d.d. matrix* $H=\Delta A\Delta$ *corresponding to eigenvalue z:*

(0) We assume the eigenvalue z has been computed accurately using one of the previous algorithms.

(1) Choose a unit starting vector y_0; set $i = 0$.

(2) Compute the LDL^T factorization of $P^T(A-z\Delta^{-2})P$, where P is the same permutation as in Algorithm 3.

(3) Repeat

$i = i+1$

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

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

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

until $(r = O(\varepsilon))$

(4) $x = \Delta^{-1}y_i$

Theorem 11: *Suppose Algorithm 5 terminates with x as the computed eigenvector of the symmetric s.d.d. matrix* $H=\Delta A\Delta$. *Then there is a diagonal matrix D with* $D_{ii} = 1+O(\varepsilon)$, ε = *machine precision, and a matrix* δA, $\|\delta A\| = O(\varepsilon)$, *such that Dx is an exact eigenvector of* $\Delta(A+\delta A)\Delta$. *Thus, the error in x is bounded by the results in Theorems 6 and 7.*

Sketch of Proof: Let $\tilde{y}_i$ be the computed solution of $(A-z\Delta^{-2})\tilde{y}_i = y_{i-1}$ at the last iteration of Algorithm 5. Applying the error analysis of the proof of Theorem 4, one can show that there is a diagonal matrix D, $D_{ii}=1+O(\varepsilon)$, and an E, $\|E\| = O(\varepsilon)$, such that $D(A-z\Delta^{-2}+E)D\tilde{y}_i = y_{i-1}$. Applying the result in [2], we can assume E is symmetric. Since Algorithm 5 guarantees $r = 1/\|\tilde{y}_i\| = O(\varepsilon)$, another application of the result in [2] guarantees the existence of a symmetric F, $\|F\| = O(\varepsilon)$, such that $(A-z\Delta^{-2}+E+F)Dy_i = 0$. Thus, Dy_i is an exact eigenvector of $A+E+F-\lambda\Delta^{-2}$ for $\lambda=z$, and $Dx=D\Delta^{-1}y_i$ is an exact eigenvector of $\Delta(A+\delta A)\Delta$, $\|\delta A\| = \|E+F\| = O(\varepsilon)$ as desired. □

12. Application to Differential Operators

Consider the n by n second central difference matrix

$$H_n = h^{-2}\cdot\begin{bmatrix} 2 & -1 & & \\ -1 & \cdot & \cdot & \\ & \cdot & \cdot & -1 \\ & & -1 & 2 \end{bmatrix}$$

which arises from discretizing $-d^2/dx^2$ at equally spaced grid points $x_i=ih$, $1\leq i\leq n$. One easily computes that $1-\gamma$ for H_n is $1-\cos\frac{\pi}{n+1}$, which approaches 0 as $n\rightarrow\infty$. This is to be expected since H_n approximates an unbounded operator, and so has a wider and wider range of eigenvalues as $n\rightarrow\infty$. Since the diagonal of H_n is constant, nothing is gained by writing $H_n=\Delta A\Delta$, $A_{ii}=1$, and our perturbation theory merely says that all the eigenvalues are at least as sensitive as the smallest one. Our theory becomes more interesting when considering unevenly spaced

grid points x_i. For example, let $h_i = x_i - x_{i-1}$, and suppose $h_{i+1}/h_i \equiv \beta$ for all i; this corresponds to a uniformly graded grid. Then the corresponding $H_n(\beta)$ has diagonal entries ranging from $2h_1^{-2}\beta^{-1}$ to $2h_1^{-2}\beta^{-2n+1}$. One can show $\gamma(\beta)$ for $H_n(\beta)$ satisfies $\gamma(\beta) = 2(2+\beta+\beta^{-1})^{-1/2}\gamma \leq \gamma$, so that the eigenvalues of $H_n(\beta)$ are always at least as accurately determined as the eigenvalues of H_n. In fact, if $\beta \neq 1$, $1-\gamma(\beta)$ is bounded away from 0 for all n.

13. Summary and Future Work

We have shown that there are a number of situations where tiny eigenvalues and singular values can be determined much more accurately than standard perturbation theorems and numerical algorithms can guarantee. This is true for singular values of bidiagonal matrices, eigenvalues of symmetric s.d.d. matrices and eigenvalues of s.d.d. definite pencils. In addition, eigenvectors of symmetric s.d.d. matrices corresponding to relatively isolated eigenvalues are determined accurately by the data. Open questions remain in the perturbation theory for the singular vectors of bidiagonal matrices and in perturbation theory for eigenvalues of s.d.d. pencils with positive and negative eigenvalues.

The following is a tabular summary of the current state of research into corresponding high accuracy algorithms. We consider two classes of algorithms for eigenvalues and singular values (bisection and QR), and two classes of algorithms for eigenvectors and singular vectors (inverse iteration using accurate eigenvalues/singular values, and QR). If an algorithm was presented in earlier research, a reference is given; otherwise it was discussed here for the first time. Conjectured algorithms are also indicated. (Divide and conquer is another technique for these problems. Since its ability to deliver high accuracy has not been proven in any of the cases considered in this paper, it qualifies as a "conjectured" algorithm for all of them.)

Bisection based algorithms for computing eigenvalues and singular values to high relative accuracy:

1) Singular values of bidiagonal matrices [9].
2) Eigenvalues of symmetric tridiagonal s.d.d. matrices (Theorem 4 and [14]).
3) Eigenvalues of not necessarily tridiagonal symmetric s.d.d. matrices (Algorithm 3).
4) Eigenvalues of s.d.d. definite pencils (Algorithm 4).

QR based algorithms for computing eigenvalues and singular values to high relative accuracy:

1) Singular values of bidiagonal matrices [9].
2) Eigenvalues of symmetric positive definite tridiagonal matrices (Algorithm 2 and [10]).
3) Eigenvalues of symmetric indefinite tridiagonal scaled diagonally dominant matrices: no QR based algorithm appears to work in general.

Inverse Iteration based algorithms for computing eigenvectors and singular vectors accurately:

1) Eigenvectors of symmetric s.d.d. matrices and pencils (Algorithm 5).

QR based algorithms for computing eigenvectors and singular vectors accurately:

1) Conjectured: singular vectors of bidiagonal matrices (Proposition 7 and [6]).
2) Eigenvectors of symmetric positive definite tridiagonal s.d.d. matrices (Algorithm 2 and [6]). (Conjectured: the finer error bounds of Theorem 7).
3) Eigenvectors of symmetric indefinite tridiagonal s.d.d. matrices: since the eigenvalues apparently cannot be computed accurately, neither can the eigenvectors.

In summary, various numerical algorithms are available to compute eigenvalues and eigenvectors, and singular values and singular vectors with high accuracy. Algorithm 2 for the symmetric positive definite tridiagonal eigenproblem will be incorporated in the LAPACK linear algebra library [8]. Not all algorithmic questions have been settled, however, and these will be the subject of future research.

References

[1] M. Blevins and G. W. Stewart, *Calculating the Eigenvectors of Diagonally Dominant Matrices*, J. Assc. Comp. Mach., Vol. 21, No. 2, April 1974, pp 261-271

[2] J. Bunch, J. Demmel, C. Van Loan, *The Strong Stability of Algorithms for Solving Symmetric Linear Systems*, to appear in SIAM J. Matrix Anal. Appl.

[3] J. Bunch, J. Dongarra, C. B. Moler, G. W. Stewart, *LINPACK Users' Guide*, SIAM, Philadelphia, 1979

[4] J. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., Vol. 31, No. 137, Jan 1977, pp 163-179

[5] J. J. M. Cuppen, *A Divide and Conquer method for the Symmetric Tridiagonal Eigenproblem*, Numer. Math. 36, pp. 177-195, 1981

[6] P. Deift, J. Demmel, L.C. Li, C. Tomei, *The Bidiagonal Singular Value Decomposition and Hamiltonian Mechanics*, in preparation

[7] J. Demmel, *On Condition Numbers and the Distance to the Nearest Ill-posed Problem*, Num. Math., v. 51, n. 3, pp. 251-89, July 1987

[8] J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, D. Sorensen, *Prospectus for the Development of a Linear Algebra Library for High-Performance Computers*, Argonne National Lab, Mathematics and Computer Science Division, ANL-MCS-TM-97, Sept. 1987

[9] J. Demmel and W. Kahan, *Accurate Singular Values of Bidiagonal Matrices*, to appear in SIAM J. Sci. Stat. Comp. (updated version of [10])

[10] J. Demmel and W. Kahan, *Computing Small Singular Values of Bidiagonal Matrices with Guaranteed High Relative Accuracy*, LAPACK Working Note # 3, ANL-MCS-TM-110, Argonne National Laboratory, February 1988

[11] J. J. Dongarra and D. C. Sorensen, *A Fully Parallel Algorithm for the Symmetric Eigenproblem*, SIAM J. Sci. Stat. Comput. 8, pp. 139-154, 1987

[12] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1983

[13] E. R. Jessup and D. C. Sorensen, *A Parallel Algorithm for Computing the Singular Value Decomposition of a Matrix*, ANL-MCS-TM-102, Argonne National Laboratory, December 1987

[14] W. Kahan, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Technical Report CS41, Computer Science Dept., Stanford University, 1966, revised 1968

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966

[16] J. Le and B. Parlett, *On the Forward Instability of the QR Transformation*, submitted to SIAM J. Mat. Anal. Appl.; also Report PAM-419, Center for Pure and Applied Mathematics, University of California, Berkeley, July 1988

[17] B. Parlett, *The Symmetric Eigenproblem*, Prentice Hall, 1980

[18] R. S. Varga, *Matrix Iterative Analysis*, Prentice Hall, 1962

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