The Condition Number of Similarities that Diagonalize Matrices

James Demmel

Computer Science Division
University of California, Berkeley

ABSTRACT

How ill-conditioned must a matrix S be if it (block) diagonalizes a given matrix T, i.e. if $S^{-1}TS$ is block diagonal? The answer depends on how the diagonal blocks partition T's spectrum; the condition number of S is bounded below by a function of the norms of the projection matrices determined by the partitioning. In the case of two diagonal blocks we compute an S which attains this lower bound, and we describe almost best conditioned S's for dividing T into more blocks. We apply this result to bound the error in an algorithm to compute analytic funtions of matrices, for instance $\exp(T)$. Our techniques also produce bounds for submatrices that appear in the square-root-free Cholesky and in the Gram-Schmidt orthogonalization algorithms.

1. Introduction

Two measures of the ill-conditioning of the eigenvalues of a matrix T have appeared frequently in the literature. One is the condition number of a matrix S which (block) diagonalizes T under similarity (i.e. $S^{-1}TS$ is block diagonal), and the other is the norm of the projection matrix P_i belonging to the spectrum of the i-th diagonal block of $S^{-1}TS$ (if the i-th block is 1 by 1, the norm of P_i is usually denoted $1/|s_i|$). Many authors have shown that the larger the condition number of S is, or the larger the norm of P_i is, the more sensitive to perturbations are at least some of the eigenvalues of T. Bauer and Fike [3], Kahan [6], Ruhe [9], Wilkinson [13,14] and others have all contributed theorems stating this result in different ways.

Our goal in this paper is to relate these two measures of ill-conditioning. The first measure is somewhat ill-defined, since there must be many matrices S which block diagonalize T; we therefore consider the best conditioned S among all candidates. In contrast, the second measure is defined unambiguously given T and the partitioning $\Sigma = \bigcup_{i=1,m} \Sigma_i$ of T's spectrum into disjoint

sets determined by the diagonal blocks Θ_i of $S^{-1}TS$: spectrum$(\Theta_i) = \Sigma_i$. To each Σ_i belongs a projection P_i. We show that the condition number of the best S is very nearly determined by the largest $||P_i||$, where $||P_i||$ is the norm of the projection P_i. In fact we show how to compute an S whose condition number is within a factor of dim(T) of the largest $||P_i||$, and that this S is nearly best.

Kahan [6] relates the two measures when S divides T into only 2 blocks. We sharpen his results by exhibiting a best S for decomposing T into two blocks and compute its condition number exactly in terms of the norm of a projection (see (2) below). Wilkinson [13, p 89] relates the two measures when $S^{-1}TS$ is completely diagonal; we generalize his results to diagonal blocks of arbitrary sizes in Theorems 3 and 3a below.

To describe our results more formally, let $||\cdot||$ denote the 2-norm for vectors and also the matrix norm induced by the vector norm:

$$||S|| = \max_{x \neq 0} ||Sx|| / ||x|| \quad .$$

Let $\kappa(S)$ be the condition number of S with respect to $||\cdot||$:

$$\kappa(S) \equiv ||S|| \; ||S^{-1}|| \quad .$$

Let $\Sigma = \bigcup_{i=1,m} \Sigma_i$ be a given partitioning of T's spectrum into m disjoint sets. We seek the best conditioned matrix S such that

$$S^{-1}TS = \Theta = \begin{bmatrix} \Theta_1 & & \\ & \ddots & \\ & & \Theta_m \end{bmatrix} \quad \text{and} \quad \text{spectrum}(\Theta_i) = \Sigma_i \; ; \tag{1}$$

S will be the matrix that minimizes $\kappa(S)$ subject to the constraints (1).

When $m=2$ the best conditioned S will be expressed explicitly in terms of the projection matrix P belonging to Σ_1's invariant subspace. The condition number of this best S will be

$$\kappa(S) = ||P|| + \sqrt{||P||^2 - 1} \quad . \tag{2}$$

This result sharpens an estimate for $\kappa(S)$ given by Kahan [6].

To prove (2) we will need a technical result, Theorem 1, that bounds the norms of submatrices of a positive definite matrix in terms of its condition number. Theorem 1 is a slight generalization of an inequality of Wielandt [4], and the proof technique used here yields several other

inequalities (Theorem 4), one of which (65) is an inequality of Bauer [1].

When $m>2$ we show how to compute an S whose condition number is no larger than $\sqrt{m}$ times the smallest possible condition number. This is done by first splitting T into two diagonal blocks in the optimal way mentioned above, and then continuing to split each diagonal block into two smaller ones (recursively). The columns of the resulting S form an orthonormal basis for each invariant subspace of T (i.e. the first $\dim(\Theta_1)$ columns of S are an orthonormal basis spanning the invariant subspace corresponding to Σ_1, etc.). This result generalizes a result of van der Sluis [10] in which he essentially considers the case where all invariant subspaces are one-dimensional.

We also bound $\kappa(S)$ above and below in terms of the norms of the projection matrices P_i belonging to Σ_is' invariant subspaces:

$$\max_i \left(\| P_i \| + \sqrt{\| P_i \|^2 - 1} \right) \le \kappa(S) \le \sqrt{m} \cdot \sqrt{\sum_{i=1}^{m} \| P_i \|^2} \ . \tag{3}$$

The lower bound in (3) generalizes a result of Bauer [2], who considered the case of one-dimensional invariant subspaces. The upper bound is similar to a number of results [2,11,13] where again one-dimensional subspaces are considered.

Our result bears on the accuracy to which analytic functions of a matrix (such as the exponential) can be computed. A typical algorithm to compute $\exp(T)$, for example, will first find an S to reduce T to block diagonal form as in (1), exponentiate the blocks Θ_i and transform back:

$$\exp(T) = \exp(S\Theta S^{-1}) = S\exp(\Theta)S^{-1} \ . \tag{4}$$

Our bound for the error in computing $\exp(T)$ by this method includes $\kappa(S)$ as a factor. Thus, the smaller $\kappa(S)$ is, the more accurately can $\exp(T)$ be computed.

Part 2 of this paper states the main theorem and uses it to display the best S for decomposing T into 2 blocks. Part 3 discusses breaking T into $m>2$ blocks. Part 4 applies the results to an error bound for computing an analytic function of a matrix. Part 5 applies a variation of the main theorem to bound the matrices obtained from doing square root free Cholesky, and from the Gram-Schmidt orthogonalization process. Part 6 has the proof of the main theorem and some related results.

2. How to Decompose T into 2 blocks

Let

$$H = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

be a Hermitian positive definite matrix, partitioned so that A is n by n, B is m by n, and C is m by m. Let $\kappa = ||H||\ ||H^{-1}||$ be the condition number of H.

Theorem 1: If H and κ are defined as above, then

$$||(A^{-1/2})^* B C^{-1/2}|| \leq \frac{\kappa - 1}{\kappa + 1} \tag{5}$$

or, equivalently,

$$\kappa \geq \frac{1 + ||(A^{-1/2})^* B C^{-1/2}||}{1 - ||(A^{-1/2})^* B C^{-1/2}||}\ , \tag{6}$$

where $X^{-1/2}$ can be any matrix such that $X^{-1/2}(X^{-1/2})^* = X^{-1}$. Furthermore, this bound is sharp. In fact, given any m by n matrix Z such that $||Z|| < 1$, it is possible to construct an H such that $(A^{-1/2})^* B C^{-1/2} = Z$ and both sides of inequality (5) are equal.

This theorem will be proved in Part 6. Let us use it here to analyze the decomposition of T into 2 diagonal blocks.

Let Σ denote the spectrum of T, a set of points in the complex plane, and let Δ denote any proper subset of Σ. There must be many similarity matrices S such that

$$S^{-1} T S = \begin{bmatrix} E & 0 \\ 0 & F \end{bmatrix} \tag{7}$$

where E and F are square matrices whose spectra are the disjoint point sets Δ and Σ–Δ, respectively. (We change notation here temporarily to avoid a proliferation of subscripts.) These requirements do not determine S, E, and F uniquely, but they do impose constraints on the matrix S of the similarity. One constraint is a lower bound under the condition number $||S||\ ||S^{-1}||$ of S. Using Theorem 1 we shall show that

$$\inf_S \| S \| \ \| S^{-1} \| = \| P \| + \sqrt{\| P \|^2 - 1} \tag{8}$$

where $P^2 = P$ is the projection onto T's invariant subspaces belonging to Δ. Alternatively, P can be replaced by the complementary projection $1-P$ onto T's invariant subspaces belonging to $\Sigma-\Delta$ without changing the bound above because $\| 1-P \| = \| P \|$ [7]. What characterizes P besides the equations

$$P^2 = P \qquad \text{and} \qquad PT = TP$$

is the identification of P with Δ instead of some other part of T's spectrum. Thus

$$P = \int_\Gamma (\zeta I - T)^{-1} \, d\zeta/(2\pi i)$$

where Γ is any closed contour with Δ strictly inside and $\Sigma-\Delta$ strictly outside [8]. Of course, we would not compute P from this integral; another better way to compute P and show how it is related to S is as follows.

By Schur's Theorem [5], we may reduce T to upper triangular form by a unitary matrix Q

$$Q^* T Q = \begin{bmatrix} E' & G \\ 0 & F' \end{bmatrix} \tag{9}$$

where E' is similar to E and F' to F. Since

$$\kappa(S) = \| S \| \ \| S^{-1} \| = \| Q^* S \| \ \| S^{-1} Q \|$$

we may assume without loss of generality that T is initially upper triangular. Thus, we seek an S such that

$$S^{-1} \begin{bmatrix} E' & G \\ 0 & F' \end{bmatrix} S = \begin{bmatrix} E & 0 \\ 0 & F \end{bmatrix} \tag{10}$$

and a corresponding projection P which projects onto the invariant subspaces corresponding to the spectrum Δ of E' and E.

The matrices S and P can be exhibited as follows. Define R by solving $G = RF' - E'R$; this equation can be rearranged to form a triangular system of linear equations whose solution is R with its entries rearranged to form a vector. Then S must be of the form:

$$S = \begin{bmatrix} 1 & R \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & W \end{bmatrix} . \tag{11}$$

where V and W are arbitrary nonsingular matrices of the same dimensions as E' and F' respectively. Why must S be of this form? It is straightforward to verify that

$$S^{-1}TS = \begin{bmatrix} V^{-1}E'V & 0 \\ 0 & W^{-1}F'W \end{bmatrix}$$

and so the first $\dim(E')$ columns of S span the right invariant subspace of T corresponding to Δ. Any set of columns spanning the same subspace would serve equally well as columns of S, which means precisely that V may be any nonsingular matrix. The same comments apply to the last $\dim(F')$ columns of S.

Now observe that

$$P = \begin{bmatrix} 1 & R \\ 0 & 0 \end{bmatrix} = P^2 \ . \tag{12}$$

Since

$$PT = TP = \begin{bmatrix} E' & E'R \\ 0 & 0 \end{bmatrix} ,$$

P projects onto the invariant subspace corresponding to E'. Note that $\| P \|^2 = 1 + \| R \|^2$.

Now we estimate $\kappa(S)$:

$$\begin{aligned} \kappa^2(S) &= \kappa(S^*S) \\ &= \kappa\left(\begin{bmatrix} V^*V & V^*RW \\ W^*R^*V & W^*(I + R^*R)W \end{bmatrix} \right) . \end{aligned} \tag{13}$$

We can invoke Theorem 1 with $A^{-1/2} = V^{-1}$, $B = V^*RW$, and $C^{-1/2} = W^{-1}(I + R^*R)^{-1/2}$, so that $(A^{-1/2})^*BC^{-1/2} = R(I + R^*R)^{-1/2}$, to find

$$\kappa^2(S) \geq \frac{1 + \| R(I + R^*R)^{-1/2} \|}{1 - \| R(I + R^*R)^{-1/2} \|} \ . \tag{14}$$

Now we need to compute $\| R(I + R^*R)^{-1/2} \|$. Assuming without loss of generality that $(I + R^*R)^{-1/2}$ is the Hermitian square root, we obtain

$$\begin{aligned} \| R(I + R^*R)^{-1/2} \|^2 &= \| (I + R^*R)^{-1/2} R^*R (I + R^*R)^{-1/2} \| \\ &= \| R^*R(I + R^*R)^{-1} \| \end{aligned} \tag{15}$$

$$= \frac{\| R \|^2}{1 + \| R \|^2}$$

$$= \frac{\| P \|^2 - 1}{\| P \|^2} \ ,$$

Substituting (15) in (14) gives

$$\kappa(S) \geq \sqrt{\| P \|^2 - 1} + \| P \| \ . \tag{16}$$

which proves the lower bound claimed in (8).

We now show that the lower bound in (8) can be attained. According to (64) in the proof of Theorem 1, by setting

$$W = [I + R^*R]^{-1/2} \quad \text{and} \quad V = I \tag{17}$$

where W can be any matrix such that $WW^* = (I + R^*R)^{-1}$, we obtain S for which the previous inequality becomes equality. Thus have we perfected Parlett's improvement of an estimate for inf $\kappa(S)$ due to Kahan [6]. Note that W may be taken to be upper triangular by letting it be the Cholesky factor of $(I + R^*R)^{-1}$, where the Cholesky factorization is done starting from the lower right corner of the matrix instead of the upper left. This choice of W maintains the upper triangular form of $S^{-1}TS$, a fact we will use in Part 3. With this choice of V and W the first $\dim(E')$ columns of S are orthonormal, as are the remaining $\dim(F')$ columns.

The best S is not unique. By applying the variation on Theorem 1 given by equation (70) in Theorem 4, for example, we get the following alternatives for V and W:

$$W = I \quad \text{and} \quad V = [\,(\frac{\kappa^4 + 1}{2\kappa^2})\, I - RR^*]^{1/2}$$

where V can be any matrix such that VV^* is the matrix in brackets. For this choice of V and W, the columns of S are no longer orthnormal, but the first $\dim(E')$ rows of S are multiples of orthonormal vectors, and the remaining $\dim(F')$ rows are orthonormal.

3. How to Decompose T Into $m > 2$ Blocks

Note that the choice of V and W above in (17) causes S^*S to have the form

$$S^*S = \begin{bmatrix} I & R(I + R^*R)^{-1/2} \\ (R(I + R^*R)^{-1/2})^* & I \end{bmatrix} . \tag{18}$$

Since the upper left corner of $S*S$ is the identity, the columns of S which span the invariant subspace corresponding to E (the first $\dim(E)$ columns) are orthonormal. Similarly, the last $\dim(F)$ columns of S are orthonormal. This raises the following question: if we pick an orthonormal basis for each invariant subspace we want to display, how far from optimally conditioned is this choice of S?

First we will show that this choice of S has a condition number no larger than $\sqrt{m}$ times optimal (where m is the number of invariant subspaces); second we will bound $\kappa(S)$ above and below in terms of the $||P_i||$; and third we will show how to compute this S given T in upper triangular form. Finally we will discuss a different choice of S (also discussed in the literature [11,13]) which is harder to compute and has slightly different bounds on its condition number.

Theorem 2: Let S be a matrix that block diagonalizes T as shown:

$$S^{-1}TS = \begin{bmatrix} (S^{-1})_1 * \\ \cdot \\ (S^{-1})_m * \end{bmatrix} \begin{bmatrix} T_{11} & \cdot & T_{1m} \\ & \cdot & \cdot \\ & & T_{mm} \end{bmatrix} \begin{bmatrix} | & & | \\ S_1 & \cdots & S_m \\ | & & | \end{bmatrix} \tag{19}$$

$$= \begin{bmatrix} \Theta_1 & & \\ & \cdot & \\ & & \Theta_m \end{bmatrix},$$

where S_i denotes the columns of S spanning the right invariant subspace of T belonging to Σ_i, and $(S^{-1})_i *$ denotes the rows of S^{-1} spanning the corresponding left invariant subspace.

By choosing the columns constituting S_i to be any orthonormal basis of the (right) invariant subspace belonging to Σ_i, S will have a condition number no larger than $\sqrt{m}$ times the smallest possible:

$$\kappa(S) \leq \sqrt{m} \cdot \kappa(S_{\text{OPTIMAL}}) \ . \tag{20}$$

Said another way, choose S so that $S*S$ has identity matrices (of various sizes) as diagonal blocks.

Proof: This proof is a simple generalization of the proof that by diagonally scaling a positive definite matrix to have unit diagonal, its condition number is within a factor of the dimension of the matrix of the lowest condition number achievable by diagonal scaling [10]. We generalize diagonal scaling for unit diagonal to be block diagonal scaling for block unit diagonal, i.e. to have identity matrices (of various sizes) on the diagonal. We show that such a block diagonal scaling

produces a matrix whose condition number is within a factor of the number of diagonal blocks of the lowest possible condition number.

Assume the columns of S form orthonormal bases of T's right invariant subspaces and let D be a block diagonal nonsingular matrix whose blocks D_{ii} are the same size as T_{ii}. Then any S' which decomposes T as shown in (19) can be written $S' = SD$. Now

$$\sqrt{m}\,\kappa(SD) = \sqrt{m}\,\frac{\max_{w\neq 0}\frac{||\,Sw\,||}{||\,D^{-1}w\,||}}{\min_{z\neq 0}\frac{||\,Sz\,||}{||\,D^{-1}z\,||}} \geq \frac{||\,D^{-1}z_*\,||}{||\,D^{-1}w_*\,||}\,\frac{\sqrt{m}\;||\,Sw_*\,||}{\sigma_{\min}(S)}\quad , \tag{21}$$

where z_* is chosen so that $||\,z_*\,|| = 1$ and $||\,Sz_*\,|| = \sigma_{\min}(S) =$ the smallest singular value of S, and w_* is chosen so $||\,w_*\,|| = 1$ and $||\,D^{-1}w_*\,|| = \sigma_{\min}(D^{-1})$. With this choice of w_* the factor $||\,D^{-1}z_*\,||\,/||\,D^{-1}w_*\,||$ is at least one. Since D is block diagonal, w_* can be chosen to have nonzero components corresponding to only one block of D. Thus, $||\,Sw_*\,||^2 = ||\,w_*{}^*S^*Sw_*\,|| = ||\,w_*{}^*w_*\,|| = 1$. Since $\sigma_{\max}(S) =$ the largest singular value of S satisfies

$$\sigma_{\max}(S) = ||\,S\,|| \leq \sqrt{\sum_{i=1}^{m} ||\,S_i\,||^2} = \sqrt{\sum_{i=1}^{m} 1} = \sqrt{m}\quad ,$$

we get

$$\sqrt{m}\,\kappa(SD) \geq \frac{\sigma_{\max}(S)}{\sigma_{\min}(S)} = \kappa(S)\quad . \tag{22}$$

Since (22) is true for any D, it is true in particular when $SD = S_{OPTIMAL}$. Q.E.D.

In the case of $m=2$ we expressed $\kappa(S_{OPTIMAL})$ in terms of $||\,P\,||$, where P was the projection matrix corresponding to the invariant subspace belonging to Σ_1. We can also bound $\kappa(S)$ here in terms of the $||\,P_i\,||$, where P_i is the projection matrix belonging to Σ_i.

Theorem 3: Let T, S and P_i be defined as above. Then

$$\max_i\,(||\,P_i\,|| + \sqrt{||\,P_i\,||^2 - 1}) \leq \kappa(S) \leq \sqrt{m}\cdot\sqrt{\sum_{i=1}^{m} ||\,P_i\,||^2}\quad , \tag{23}$$

or weakened slightly,

$$\max_i ||\,P_i\,|| \leq \kappa(S) \leq m\cdot\max_i ||\,P_i\,||\quad . \tag{24}$$

Proof: This proof is based a similar result of Wilkinson's [13, p. 89] when all invariant subspaces are one dimensional. First we will prove the lower bound and then the upper bound.

From (8) we know that any S (not just the one defined above) which displays the invariant subspace belonging to Σ_i has a condition number bounded from below:

$$\kappa(S) \geq || P_i || + \sqrt{|| P_i ||^2 - 1} \ . \tag{25}$$

Since (25) is true for all i, the lower bound follows easily.

We compute the upper bound as follows:

$$\kappa(S) = || S || \ || S^{-1} || \leq \sqrt{m} \ || S^{-1} || \tag{26}$$

since $|| S || \leq \sqrt{m}$ (as mentioned in the proof of Theorem 2). Using the notation of (19) it is easy to verify that

$$P_i = S_i \ (S^{-1})_{i,*} \ . \tag{27}$$

Since S_i consists of orthonormal columns, (27) yields

$$|| P_i || = || (S^{-1})_{i,*} || \ . \tag{28}$$

Thus

$$|| S^{-1} || \leq \sqrt{\sum_{i=1}^{m} || (S^{-1})_{i,*} ||^2} = \sqrt{\sum_{i=1}^{m} || P_i ||^2} \tag{29}$$

and the upper bound follows. Q.E.D.

The lower bound in Theorem 3 has been proven by Bauer [2] in the case of one-dimensional invariant subspaces.

We can use the splitting algorithm of Part 2 to compute such an S. Assume without loss of generality that T is initially upper triangular. Build S as a product $\prod S^{(i)}$ where $(\prod_{i=1}^{n} S^{(i)})^{-1} T (\prod_{i=1}^{n} S^{(i)})$ is block diagonal and upper triangular for all n and $S^{(n+1)}$ is chosen as in (17) to split one or more of the diagonal blocks of $(\prod_{i=1}^{n} S^{(i)})^{-1} T (\prod_{i=1}^{n} S^{(i)})$. After some easy computation we may verify that $S^* S = (\prod S^{(i)})^* (\prod S^{(i)})$ has identity matrices (of various sizes) as diagonal blocks.

The other choice of S discussed in the literature is scaled so that the i-th diagonal block of $S*S$ is $||P_i||^{1/2}$ times an identity matrix of appropriate size. With this choice of S the i-th diagonal block of $(S*S)^{-1}$ has the same norm as the corresponding block of $S*S$, namely $||P_i||^{1/2}$. In fact, in the case where all invariant subspaces are one-dimensional Smith [11] shows that this choice of S is optimally scaled with respect to the condition number

$$\kappa_F(S) \equiv ||S||_F \, ||S^{-1}||_F$$

where $||\cdot||_F$ is the Frobenius norm:

$$||S||_F \equiv \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |S_{i,j}|^2} \ .$$

With this choice of S, Theorem 2 is weakened slightly to become:

Theorem 2a: With S chosen so that the i-th diagonal block of $S*S$ is $||P_i||^{1/2}$ times an identity matrix, we have

$$\kappa(S) \leq m \cdot \kappa(S_{OPTIMAL}) \ . \tag{30}$$

Proof: Similar to Theorem 2.

Theorem 3, on the other hand, becomes slightly stronger:

Theorem 3a: With S chosen as in Theorem 2a, we can bound $\kappa(S)$ as follows:

$$\max_i \left(||P_i|| + \sqrt{||P_i||^2 - 1}\right) \leq \kappa(S) \leq \sum_{i=1}^{m} ||P_i|| \ . \tag{31}$$

Proof: Similar to Theorem 3.

The upper bound of Theorem 3a generalizes a result of Wilkinson [13, p 89] for one dimensional invariant subspaces. Note that the "spectral condition numbers" $1/|s_i|$ used by Wilkinson and others [11,13] are just $||P_i||$ when the invariant subspaces are one-dimensional. When $\sum_{i=1}^{m} ||P_i||$ is large the upper bound in (31) is comparable with the upper bound on $\kappa(S_{OPTIMAL})$ given by Bauer [2, Theorem VII] in the case of one-dimensional invariant subspaces.

This choice of S is more difficult to compute than the S of Theorems 2 and 3 because of the need to compute norms of the P_i (if the invariant subspaces are all one or two dimensional this is not hard, of course).

4. Computing a Function of a Matrix

In this section we want to show why a well conditioned block diagonalizing matrix S is better than an ill-conditioned one for computing a function of a matrix T. Assuming $f(T)$ is an analytic function of T, we compute $f(T)$ as follows:

$$f(T) = f(S\Theta S^{-1}) = Sf(\Theta)S^{-1} = S \begin{bmatrix} f(\Theta_1) & & \\ & \ddots & \\ & & f(\Theta_m) \end{bmatrix} S^{-1} \ . \tag{32}$$

The presumption is that it is easier to compute f of the small blocks Θ_i than of all of T. We will not ask about the error in computing $f(\Theta_i)$ but rather the error in computing $\Theta = S^{-1}TS$ and $f(T) = Sf(\Theta)S^{-1}$. In general, we are interested in the error in computing the similarity transformation $X = SYS^{-1}$.

We assume for this analysis that we compute with single precision floating point with relative precision ϵ. That is, when $\ $ is one of the operations $+, -, *$ or $/$, the relative error in computing $fl(a \ \ b)$ is bounded by ϵ:

$$fl(a \ \ b) = (a \ \ b)(1+e) \quad \text{where} \quad |e| \leq \epsilon \ . \tag{33}$$

Using (33) it is easy to show

Lemma 1: Let A and B be real n by n matrices, where $n\epsilon < .1$. Let $|A|$ denote the matrix of absolute entries of A: $|A|_{ij} = |A_{ij}|$. Then to first order in ϵ the error in computing the matrix product AB is bounded as follows:

$$|fl(AB) - AB| \leq n\epsilon |A| \ |B| \ . \tag{34}$$

Proof: See [12].

Computing $X = SYS^{-1}$ requires two matrix products: $Z = fl(SY)$ and $X = fl(ZS^{-1})$, where we assume S and S^{-1} are known exactly. Applying Lemma 1 to these two products yields

Lemma 2: If $n\epsilon < .1$, then to first order in ϵ

$$||fl(SYS^{-1}) - SYS^{-1}|| \leq 3n^{5/2}\kappa(S)||Y||\epsilon \ . \tag{35}$$

Proof: Straightforward.

Assuming this bound is realistic, it is clear that picking S to keep $\kappa(S)$ small is advantageous. The error in computing similarity transformations of matrices is discussed in more detail in

Wilkinson [13, chap 3].

5. Applications of a Variation of Theorem 1

It is more convenient here to use a slight variation on Theorem 1, stated as (66) in Theorem 4:

$$|| A^{-1}B|| \leq \frac{1}{2}(\sqrt{\kappa} - 1/\sqrt{\kappa}) \ .$$

Application 1: Cholesky without square roots. The square root free Cholesky algorithm decomposes a positive definite Hermitian matrix H into the product of a unit lower triangular matrix L, a nonnegative diagonal matrix D, and L^*:

$$H = LDL^*$$

We wish to bound the entries of L. Consider the following partitioning of the decomposition:

$$H = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} L_1 & \\ R & L_2 \end{bmatrix} \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \begin{bmatrix} L_1^* & R^* \\ & L_2^* \end{bmatrix} . \tag{36}$$

From (36) we see

$$L_1 D_1 R^* = B$$

or

$$\begin{aligned} R^* &= (L_1 D_1)^{-1} B \\ &= L^*_1 (L_1 D_1 L^*_1)^{-1} B \\ &= L^*_1 A^{-1} B \ . \end{aligned}$$

Since L^*_1 is unit upper triangular, the last row of R^* and the last row of $A^{-1}B$ are identical. But the last row of R^* is the conjugate transpose of a subdiagonal column of L. Thus

$$\begin{aligned} || \text{ subdiagonal column of } L || &= || \text{ last column of corresponding } A^{-1}B || \\ &\leq || \text{ all of corresponding } A^{-1}B || \ , \end{aligned}$$

and so Theorem 4 implies

$$\| \text{subdiagonal column of } L \| \leq \frac{1}{2}(\sqrt{\kappa} - 1/\sqrt{\kappa}) \ .$$

A 2 by 2 example suggested by the proof of Theorem 4 (see (66) and (70)) shows this bound is achievable.

This bound is tighter than the simpler bound

$$|L_{ij}| \leq \sqrt{(H_{ii} - D_{ii})/D_{jj}} \leq \sqrt{(\Lambda - \lambda)/\lambda} = \sqrt{\kappa - 1} \ , \tag{37}$$

which is derived by considering the i,i-th entries of both sides of $H = LDL^*$:

$$L_{ij}^2 D_{jj} + D_{ii} + \text{positive terms} = H_{ii} \ .$$

This result can also be used to get a lower bound on $\kappa(H)$ given its Cholesky decomposition.

A similar application to Gauss-Jordan elimination appears in [1].

Application 2: Gram-Schmidt Orthogonalization Process. The Gram-Schmidt process takes a set of independent vectors $v_i \in \mathbf{C}^n$, $1 \leq i \leq m$, and produces a set of orthonormal vectors $q_i \in \mathbf{C}^n$, $1 \leq i \leq m$, where q_i is a linear combination of v_1 through v_i and orthogonal to v_1 through v_{i-1} for $i > 1$. We wish to bound the coefficients of q_1 to q_{i-1} (or v_1 to v_{i-1}) in the expression for q_i. We do this by showing Gram-Schmidt to be equivalent to square-root-free Cholesky performed on a certain matrix, and use Application 1.

The Gram-Schmidt process expresses q_i as a linear combination of v_i and q_1 through q_{i-1}. Let V be the n by m matrix whose columns are the vectors v_i and let Q be the n by m matrix with columns q_i. Then the Gram-Schmidt process may be expressed succinctly as

$$V = QD^{1/2}U \ , \tag{38}$$

where U is an n by n unit upper triangular matrix and D is an n by n nonnegative diagonal matrix. The entries of U are the coefficients we seek to bound. Multiplying both sides of (38) on the left by their transposes, we obtain

$$V^*V = U^*DU \ . \tag{39}$$

U is the factor of V^*V obtained by doing square root free Cholesky. Thus, from Application 2 we see

$$\| \text{superdiagonal column of } U \| \leq \frac{1}{2}\left(\sqrt{\kappa(V^*V)} - 1/\sqrt{\kappa(V^*V)} \right) \ , \tag{40}$$

which is the desired bound.

If we wanted to express q_i as a linear combination of v_1 through v_i instead of v_i and q_1 through q_{i-1}, we would express the Gram-Schmidt process as

$$V\hat{U}\hat{D}^{-1/2} = Q \quad . \tag{41}$$

What is a bound for the columns of $\hat{U}$? Multiply both sides of (41) on the the left by their transposes to obtain

$$\hat{D}^{-1/2}\hat{U}*V*V\hat{U}\hat{D}^{-1/2} = Q*Q = I \quad , \tag{42}$$

or

$$(V*V)^{-1} = \hat{U}\hat{D}^{-1}\hat{U}* \quad . \tag{43}$$

$\hat{U}$ is the factor of $(V*V)^{-1}$ obtained by doing square root free Cholesky starting at the lower right corner of $(V*V)^{-1}$ instead of the upper left corner as is usual. Thus, from Application 2 we see

$$\begin{aligned} ||\text{ superdiagonal column of } \hat{U}|| &\le \frac{1}{2}\left(\sqrt{\kappa((V*V)^{-1})} - 1/\sqrt{\kappa((V*V)^{-1})}\right) \\ &= \frac{1}{2}\left(\sqrt{\kappa(V*V)} - 1/\sqrt{\kappa(V*V)}\right) \quad , \end{aligned} \tag{44}$$

since $\kappa(M) = \kappa(M^{-1})$ for all M. Thus, we get the same bound on the columns of $\hat{U}$ as on the columns of U.

6. Proof of Theorem 1

Unit vectors $x \in \mathbf{C}^n$ and $y \in \mathbf{C}^s$ satisfying

$$y*(A^{-1/2})*BC^{-1/2}x = ||\,(A^{-1/2})*BC^{-1/2}||\, \tag{45}$$

must exist. Use them to construct the unit vectors

$$z = A^{-1/2}y/||\,A^{-1/2}y\,|| \quad , \quad w = C^{-1/2}x/||\,C^{-1/2}x\,|| \quad , \tag{46}$$

and

$$s(\theta) = \begin{bmatrix} z\sin\theta \\ w\cos\theta \end{bmatrix} \quad . \tag{47}$$

We want to consider H acting on the 2-dimensional subspace in which $s(\theta)$ lies. Now

$$s^*(\theta)Hs(\theta) \leq \Lambda \tag{48}$$

implies

$$[z^*\sin\theta,\ w^*\cos\theta] \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} z\sin\theta \\ w\cos\theta \end{bmatrix} \leq \Lambda \ , \tag{49}$$

or

$$\sin^2\theta \cdot z^*Az + \cos^2\theta \cdot w^*Cw + \sin\theta\cos\theta\,(w^*B^*z + z^*Bw) \leq \Lambda \ . \tag{50}$$

To simplify notation, let $a \equiv z^*Az$ and $c \equiv w^*Cw$.

From (45) and (46) we know that

$$z^*Bw = ||\,(A^{-1/2})BC^{-1/2}||\,/(||\,A^{-1/2}y||\ \cdot\ ||\,C^{-1/2}z||\,) \tag{51}$$

$$= ||\,(A^{-1/2})BC^{-1/2}||\ \cdot\ ||\,A^{1/2}z||\ \cdot\ ||\,C^{1/2}w||\ .$$

Since $(C^{1/2})^*C^{1/2} = C$, we get $c \equiv w^*Cw = ||\,w^*(C^{1/2})^*C^{1/2}w||\ = ||\,C^{1/2}w||^{\,2}$. Similarly, $a \equiv z^*Az = ||\,A^{1/2}z||^{\,2}$, so (51) becomes

$$z^*Bw = ||\,(A^{-1/2})^*BC^{-1/2}||\ \cdot \sqrt{ac}\ . \tag{52}$$

Substituting (52) into (50) and rearranging, we obtain

$$(\frac{c+a}{2}) + (\frac{c-a}{2})\cos 2\theta + \sqrt{ac}\ ||\,(A^{-1/2})^*BC^{-1/2}||\ \sin 2\theta \leq \Lambda \tag{53}$$

Since θ was arbitrary, we can maximize the L.H.S. of (53) over θ yielding

$$(\frac{c+a}{2}) + \sqrt{(\frac{c-a}{2})^2 + ac\ ||\,(A^{-1/2})^*BC^{-1/2}||^{\,2}} \leq \Lambda \ , \tag{54}$$

or

$$||\,(A^{-1/2})^*BC^{-1/2}|| \ \leq \frac{\sqrt{(\Lambda - (c+a)/2)^2 - ((c-a)/2)^2}}{\sqrt{ac}} \tag{55}$$

$$= \frac{\sqrt{(\Lambda - a)\,(\Lambda - c)}}{\sqrt{ac}}\ .$$

Similarly, the inequality

$$\lambda \leq s^*(\theta)Hs(\theta) \tag{56}$$

implies

$$\lambda \le (\frac{c+a}{2}) + (\frac{c-a}{2})\cos 2\theta + \| (A^{-1/2})^* B C^{-1/2} \| \; \| A^{1/2} z \| \; \| C^{1/2} w \| \sin 2\theta \; . \tag{57}$$

Minimizing the R.H.S. of (57) over θ we obtain

$$\lambda \le (\frac{c+a}{2}) - \sqrt{(\frac{c-a}{2})^2 + ac \; \| (A^{-1/2})^* B C^{-1/2} \|^2} \tag{58}$$

or, rearranging,

$$\| (A^{-1/2})^* B C^{-1/2} \| \le \frac{\sqrt{(a-\lambda)(c-\lambda)}}{\sqrt{ac}} \; . \tag{59}$$

Combining (55) and (59) yields

$$\| (A^{-1/2})^* B C^{-1/2} \| \le \min\left[\sqrt{(a-\lambda)(c-\lambda)/(ac)} \, , \sqrt{(\Lambda - a)(\Lambda - c)/(ac)}\right] \; .$$

All we know about $z^*Az \equiv a$ is that $\lambda \le a \le \Lambda$, and similarly $\lambda \le c \equiv w^*Cw \le \Lambda$. Thus

$$\| (A^{-1/2})^* B C^{-1/2} \| \le \max_{\lambda \le \alpha, \gamma \le \Lambda} \min\left[\sqrt{(\alpha-\lambda)(\gamma-\lambda)/(\gamma\alpha)} \, , \sqrt{(\Lambda-\alpha)(\Lambda-\gamma)/(\gamma\alpha)}\right]. \tag{60}$$

Since $(\alpha - \lambda)/\alpha$ is an increasing function of α and $(\Lambda - \alpha)/\alpha$ is a decreasing function of α in the range $\lambda \le \alpha \le \Lambda$, we see the max in the last inequality occurs when the two arguments of the min are equal. This equality implies

$$(\alpha - \lambda)(\gamma - \lambda) = (\Lambda - \alpha)(\Lambda - \gamma) \tag{61}$$

or

$$\alpha + \gamma = \Lambda + \lambda \; . \tag{62}$$

Substituting (62) into (60) yields

$$\| (A^{-1/2})^* B C^{-1/2} \| \le \max_{\lambda \le \gamma \le \Lambda} \sqrt{(\gamma - \lambda)(\Lambda - \gamma)}/\sqrt{\gamma(\Lambda + \lambda - \gamma)} \tag{63}$$

$$= \frac{\Lambda - \lambda}{\Lambda + \lambda}$$

$$= \frac{\kappa - 1}{\kappa + 1} \; .$$

as desired.

Any 2 by 2 positive definite matrix whose diagonal entries are equal shows the the inequality of Theorem 1 is sharp.

We now show that given κ and $Z = (A^{-1/2})^* B C^{-1/2}$ such that $\| Z \| < 1$ and the inequality of the theorem is sharp, it is possible to construct an H with the given constraints. Simply choose

$$A = I \quad , \quad C = I \quad \text{and} \quad B = Z \tag{64}$$

corresponding to the (arbitrary) choice $\Lambda = 1 + \| Z \|$ and $\lambda = 1 - \| Z \|$. It is easy to verify that every inequality in the proof is sharp for this choice of A, B, and C. Q.E.D.

Theorem 4: Let H, Λ, λ, and κ be as above. Define $X^{-1/2}$ such that $X^{-1/2}(X^{-1/2})^* = X^{-1}$. Then the following inequalities are sharp:

$$\| BC^{-1} \| \leq \frac{1}{2} (\sqrt{\kappa} - 1/\sqrt{\kappa}) \tag{65}$$

$$\| A^{-1}B \| \leq \frac{1}{2} (\sqrt{\kappa} - 1/\sqrt{\kappa}) \tag{66}$$

$$\| B \| \leq \frac{1}{2} (\Lambda - \lambda) \tag{67}$$

$$\| (A^{-1/2})^* B \| \leq \sqrt{\Lambda} - \sqrt{\lambda} \tag{68}$$

$$\| BC^{-1/2} \| \leq \sqrt{\Lambda} - \sqrt{\lambda} \tag{69}$$

Proof: All the proofs are analogous to the proof of Theorem 1. To prove (65), for example (also proved in [1]), choose z and y unit vectors such that

$$z^* BC^{-1} y = \| BC^{-1} \|$$

and let

$$x = C^{-1}y / \| C^{-1}y \| \quad .$$

Consider H restricted to the two dimensional subspace in which

$$s(\theta) = \begin{bmatrix} z \sin\theta \\ x \cos\theta \end{bmatrix}$$

lies. The rest of the proof follows similarly to that of Theorem 1.

We can also show that given κ and arbitrary $R = BC^{-1}$ such that (65) is sharp, it is possible to construct an H with the given constraints. Simply choose

$$C = I \quad , \quad A = (\frac{\kappa^2 + 1}{2\kappa}) I \quad \text{and} \quad B = R \tag{70}$$

corresponding to the (arbitrary) choice $\Lambda = (\kappa + 1)/2$ and $\lambda = (\kappa + 1)/2\kappa$. It is easy to verify that every inequality in the proof is sharp for this choice of A, B, and C.

Note that Theorems 1 and 4 are still true when A, B, and C are conforming submatrices extracted from a larger H (or Q^*HQ with Q unitary) since the bounds are monotonic in κ (or Λ and λ). In particular, if A, B, and C are scalar Theorem 1 becomes an inequality of Wielandt [4].

Acknowledgement

I thank Prof. W. Kahan for suggesting the application of the theorem to the problem of finding best conditioned similarities, and both Prof. Kahan and Prof. B. Parlett for much constructive criticism regarding the presentation of the results. I also acknowledge the financial support of the U. S. Department of Energy, Contract DE-AM03-76SF00034, Project Agreement DE-AS03-79ER10358, and the Office of Naval Research, Contract N00014-76-C-0013.

Bibliography

[1] F. L. Bauer,A further generalization of the Kantorovic inequality, *Numer. Math.*, 3, pp 117-119, 1961

[2] F. L. Bauer,Optimally scaled matrices, *Numer. Math.*, 5, pp 73-87, 1963

[3] F. L. Bauer, C. T. Fike, Norms and Exclusion Theorems, *Numer. Math.*, 2, pp 137-141, 1960

[4] F. L. Bauer, A. S. Householder, Some inequalities involving the euclidean condition of a matrix, *Numer. Math.*, 2, pp 308-311, 1960

[5] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods* , Wiley, 1966

[6] W. Kahan, Conserving Confluence Curbs Ill-Condition, Technical Report 6, Computer Science Dept., University of California, Berkeley, August 4, 1972

[7] T. Kato, Estimation of Iterated Matrices, with Application to the von Neumann Condition, *Numer. Math.*, 2, pp 22-29, 1960

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

[9] A. Ruhe, Properties of a Matrix with a Very Ill-conditioned Eigenproblem, *Numer. Math.*, 15, pp 57-60, 1970

[10] A. van der Sluis, Condition Numbers and Equilibration of Matrices, *Numer. Math.*, 14, pp 14-23, 1969

[11] R. A. Smith, The Condition Numbers of the Matrix Eigenvalue Problem, *Numer. Math.*, 10, pp 232-240, 1967

[12] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice Hall, 1963

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

[14] J. H. Wilkinson, Note on Matrices with a Very Ill-Conditioned Eigenproblem, *Numer. Math.*, 19, pp 176-178, 1972