The inherent inaccuracy of implicit tridiagonal QR

James W. Demmel *

University of California
Berkeley, California 94720

Abstract

Recently Demmel and Veselic showed that Jacobi's method has a tighter relative error bound for the computed eigenvalues of a symmetric positive definite matrix than does QR iteration. Here we show the weaker error bound of QR as implemented in LAPACK's SSTEQR or EISPACK's IMTQL is unavoidable. We do this by presenting a particular symmetric positive definite tridiagonal matrix for which QR must fail, given any reasonable shift strategy.

1 Introduction

Let T be an n by n symmetric positive definite matrix. Write $T = DAD$ where $D = \text{diag}\,(T_{11}^{1/2}, ..., T_{nn}^{1/2})$ and A is symmetric positive definite with unit diagonal. Let $\text{cond}(T) = \|T\|_2 \cdot \|T^{-1}\|_2$ denote the condition number of T, and ϵ denote the machine precision (which we take to be appropriate for IEEE double precision arithmetic: $2^{-52} \approx 2 \cdot 10^{-16}$ [1]). Demmel and Veselic [5] showed that relative perturbations of size η in the entries of T can cause relative perturbations of size at most about $\eta \cdot \text{cond}(A)$ in its eigenvalues. They also showed it is possible to use Jacobi's method with an appropriate stopping criterion to compute the eigenvalues of T with this relative accuracy. In contrast, the conventional error analysis of either Jacobi or tridiagonalization followed by QR provides an error bound of only $O(\epsilon)\text{cond}(T)$. If D has diagonal entries of widely varying magnitudes, then $\text{cond}(A)$ can be a great deal smaller than $\text{cond}(T)$ (and never much larger in any event), so the new relative error bound for Jacobi can be much better than the conventional one.

Here we show that the worse error bound for tridiagonal QR iteration is unavoidable. We do this by considering a particular 3 by 3 symmetric positive definite tridiagonal matrix

$$T_0 = \begin{bmatrix} 1 & 1.5 \cdot 10^{-16} & 0 \\ 1.5 \cdot 10^{-16} & 10^{-32} & 1.5 \cdot 10^{-16} \\ 0 & 1.5 \cdot 10^{-16} & 1 \end{bmatrix} \tag{1.1}$$

*Computer Science Division and Department of Mathematics, demmel@cs.berkeley.edu. The author was supported by NSF grant ASC-9005933 and DARPA grant DAAL03-91-C-0047 via a subcontract from the University of Tennessee. This work was performed during a visit to the Institute for Mathematics and its Applications at the University of Minnesota.

$$= \quad D_0 A_0 D_0 = \text{diag }(1, 10^{-16}, 1) \cdot \begin{bmatrix} 1 & .15 & 0 \\ .15 & 1 & .15 \\ 0 & .15 & 1 \end{bmatrix} \cdot \text{diag }(1, 10^{-16}, 1)$$

The eigenvalues of T_0 are 1, 1, and $.955 \cdot 10^{-32}$ to 16 digits of accuracy. We will show that no matter what reasonable choice of shifts one uses, QR will (in the absence of rounding coincidences) fail to compute the smallest eigenvalue with high relative accuracy. In contrast, since $\text{cond}(A_0) < 1.6$, all the eigenvalues are determined to high relative accuracy by the data and will be computed to nearly full machine precision by Jacobi.

Our example and proof work only for the implicit QR algorithm as implemented in LAPACK subroutine SSTEQR [2] or EISPACK subroutines IMTQL1 and IMTQL2 [7]. A different proof would be needed for the the root-free QR algorithm in EISPACK (TQLRAT) or LAPACK (SSTERF).

Indeed, provided one interprets "tridiagonal QR" sufficiently liberally, one can find a very accurate version for positive definite matrices. In [3], it was pointed out that the three step algorithm

1. Cholesky factorize $T = LL^T$.
2. Compute the singular values $\sigma_i(L)$ using the bidiagonal QR algorithm in [4].
3. Form $\lambda_i(T) = \sigma_i^2(L)$.

computes the eigenvalues of T with relative accuracy $O(\epsilon)\text{cond}(A)$. Thus, one cannot expect a general result of the form "Jacobi is more accurate than QR" independent of the implementation of QR.

2 Main Result

The algorithm we will analyze is in Figure 1.

Even though our proof is for the specific matrix T_0, it will be obvious from the proof that it works for a neighborhood of T_0, and indeed for many other matrices. Thus T_0 is not an isolated example. Also, since T_0 is symmetric from top-left to bottom-right, it does not matter whether one performs QR or QL; we will use QR.

To explain the idea of the proof, we contrast QR's behavior with Jacobi. Both QR and Jacobi begin with a matrix T_0 and produce a sequence of orthogonally similar matrices T_i converging to diagonal form. We may write each T_i as $D_i A_i D_i$ with D_i diagonal and A_i of unit diagonal as above. Assuming there are no "rounding coincidences", the accuracy of the computed eigenvalues will be ϵ times $\max_i \text{cond}(A_i)$. Since the sensitivity of the eigenvalues of the original problem T_0 is $\text{cond}(A_0)$, the algorithm will succeed in computing eigenvalues to their inherent accuracy only if $\max_i \text{cond}(A_i)$ is not too much larger than $\text{cond}(A_0)$.

The difference between QR and Jacobi is that with Jacobi $\max_i \text{cond}(A_i)$ is apparently never much larger $\text{cond}(A_0)$ [5] (we have extensive numerical evidence of this, although no proof), whereas we may construct examples for QR where $\max_i \text{cond}(A_i)$ exceeds $\text{cond}(A_0)$ by $1/\epsilon$, an enormous factor. The T_0 in equation (1.1) is such an example, as we will show.

We will also assume only "reasonable" shift strategies. This means we will use shifts equal or close to 0, a diagonal entry, an eigenvalue of a 2 by 2 submatrix, or an eigenvalue

Figure 1: **Implicit tridiagonal QR with shift.**

In the following $d(1:n)$ is a vector of diagonal entries of the input n by n tridiagonal matrix, $e(1:n-1)$ is the vector of its superdiagonal entries, and σ is the shift. On output d and e have been set equal to the diagonal and superdiagonal of the transformed matrix T_1, resp. $[c,s,r] = \mathrm{ROT}(g,f)$ returns $r = \sqrt{g^2+f^2}$, $c = g/r$ and $s = f/r$. All the values c, s and r are computed to high relative accuracy; the details of their computation (which may involve tests to avoid over/underflow) are unimportant.

$g = d(1) - \sigma$
$s = 1,\ c = 1,\ p = 0$
for $i = 1,\ n-1$
    $f = s * e(i)$
    $b = c * e(i)$
    $[c, s, r] = ROT(g, f)$
    if $i \neq 1$ then $e(i-1) = r$, endif
    $g = d(i) - p$
    $r = (d(i+1) - g) * s + 2 * c * b$
    $p = s * r$
    $d(i) = g + p$
    $g = c * r - b$
end
$d(n) = d(n) - p$
$e(n-1) = g$

of the whole matrix. Using an exact eigenvalue as a shift corresponds to the "perfect shift" strategy discussed in [6]. I believe the result to be true for quite arbitrary shifts strategies as well.

Let T_1 denote the matrix after one QR step.

The proof has three steps:

1. Let T be any 3 by 3 symmetric tridiagonal matrix with two eigenvalues very near 1 and one much smaller, like T_0. Let $T = DAD$ as above. We will show that $\mathrm{cond}(A)$ is small only if the diagonal of T is nearly a permutation of its eigenvalues. This means that among the set of all T similar to T_0, there are three disconnected "stability islands" where $\mathrm{cond}(A)$ is small, each one corresponding to a permutation. Hence any acceptable shift strategy is only allowed to produce matrices within these islands. Furthermore, the only "reasonable" shifts for matrices T with small $\mathrm{cond}(A)$ are near 0 or near 1.

2. We show that shifts near 0 necessarily either compute an indefinite T_1 (and so with

a totally inaccurate tiny eigenvalue), or else with $\mathrm{cond}(A_1)$ near 10^{15}, and so with eigenvalues so sensitive as to be completely untrustworthy.

3. We show that shifts near 1 have the same property as shifts near 0.

Lemma 1 *: Let the 3 by 3 symmetric positive definite tridiagonal matrix $T = DAD$ have two eigenvalues between $1-\eta$ and $1+\eta$, and one less than η. We consider η small; $\eta = .01$ is adequate. Then $\mathrm{cond}(A) \geq \min_i T_{ii}/(8\lambda_{\min}(T))$, and "reasonable" shifts are either less than $8\lambda_{\min}(T)\mathrm{cond}(A)$ or in the range from $1 - 8\lambda_{\min}(T)\mathrm{cond}(A) - 3\eta$ to $1+\eta$.*

Let the two superdiagonal entries of A be a_{12} and a_{23}, and let $v = \sqrt{a_{12}^2 + a_{23}^2}$. Then

$$\mathrm{cond}(A) = \frac{1+v}{1-v} \geq \frac{1}{1-v^2} = \frac{1}{\det A} = \frac{\prod_i T_{ii}}{\det(T)} \geq \frac{\prod_i T_{ii}}{\lambda_{\min}(T)(1+\eta)^2} \geq \frac{\min_i T_{ii}}{8 \cdot \lambda_{\min}(T)} \qquad (2.2)$$

The other two inequalities follow from $2 - 2\eta \leq \mathrm{tr}(T) \leq 2 + 3\eta$, $0 < T_{ii} \leq 1+\eta$, and the Cauchy interlace theorem. ∎

Lemma 2 *If $T = DAD$ is similar to T_0 and $\mathrm{cond}(A)$ is small then the only "reasonable" shifts are near 0 and near 1. In particular, shifts near 0 are at most $3 \cdot 10^{-20}\mathrm{cond}(A)$, and shifts near 1 are at least $1 - 3 \cdot 10^{-20}\mathrm{cond}(A)$.*

Proof. From Lemma 1 we see T has diagonal entries near 1, 1 and 0. Since its Frobenius norm is near $\sqrt{2}$ we see its offdiagonal entries are near 0. So by our definition of "reasonable" we see all the shifts are near 0 or near 1. ∎

Lemma 3 *If we use a "reasonable" shift near 0 on T_0 then we will either compute an indefinite T_1 (and so with a totally inaccurate tiny eigenvalue), or else $\mathrm{cond}(A) \geq 10^{15}$, indicating we expect to lose nearly all figures in an eigenvalue.*

Proof. We need to examine the inner loop of the implicit QR algorithm. ¿From the first line of the algorithm, we see that any shift $\sigma < \epsilon \cdot d(1)$ yields $g = d(1)$ after roundoff. Since $d(1)$ is near 1, any σ less than $\epsilon \approx 2 \cdot 10^{-16}$ behaves the same as $\sigma = 0$, in particular any of our "reasonable" shifts. Now we simply trace through the algorithm to see that the final entry $d(n)$ is computed as a difference $d(n) - p$. For us, $d(n)$ is near 1 and $d(n) - p$ is near 0, so there is massive cancellation. The result can only be zero or a integer multiple of ϵ. If it is nonpositive, we have lost positive definiteness. If it is a positive multiple of ϵ, Lemma 1 tells us

$$\epsilon \leq 8\lambda_{\min}(T)\mathrm{cond}(A)$$

so if $\lambda_{\min}(T)$ is accurate (near 10^{-32}), then $\mathrm{cond}(A)$ is at least $\epsilon/(8\lambda_{\min}(T) \approx 10^{15}$. ∎

Lemma 4 *If we use a "reasonable" shift near 1 on T_0 then we will either compute T_1 with a totally inaccuracy tiny eigenvalue, or else $\mathrm{cond}(A) \geq 10^{15}$, indicating we expect to lose nearly all figures in an eigenvalue.*

Proof. As in Lemma 3, we see the new $d(1)$ is computed as $g + p = d(1) + p$, where p is near -1 since the new $d(1)$ is near 0. As before, we either get $d(1) \leq 0$, losing positive definiteness, or $d(1)$ a small positive integer multiple of ϵ, making $\mathrm{cond}(A)$ at least 10^{15} as before. ∎

References

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

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide, Release 1.0.* SIAM, Philadelphia, 1992.

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

[4] J. Demmel and W. Kahan. Accurate singular values of bidiagonal matrices. *SIAM J. Sci. Stat. Comput.*, 11(5):873–912, September 1990.

[5] J. Demmel and K. Veselić. *Jacobi's Method is More Accurate than QR.* Computer Science Dept. Technical Report 468, Courant Institute, New York, NY, October 1989. (also LAPACK Working Note #15), to appear in SIAM J. Mat. Anal. Appl.

[6] A. Greenbaum and J. Dongarra. *Experiments with QL/QR methods for the symmetric tridiagonal eigenproblem.* Computer Science Dept. Technical Report CS-89-92, University of Tennessee, Knoxville, 1989. (LAPACK Working Note #17).

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