SIAM J. MATRIX ANAL. APPL.
Vol. 16, No. 2, pp. 343–358, April 1995

MATRIX POWERS IN FINITE PRECISION ARITHMETIC*

NICHOLAS J. HIGHAM† AND PHILIP A. KNIGHT‡

Abstract. If A is a square matrix with spectral radius less than 1 then $A^k \to 0$ as $k \to \infty$, but the powers computed in finite precision arithmetic may or may not converge. We derive a sufficient condition for $fl(A^k) \to 0$ as $k \to \infty$ and a bound on $\|fl(A^k)\|$, both expressed in terms of the Jordan canonical form of A. Examples show that the results can be sharp. We show that the sufficient condition can be rephrased in terms of a pseudospectrum of A when A is diagonalizable, under certain assumptions. Our analysis leads to the rule of thumb that convergence or divergence of the computed powers of A can be expected according as the spectral radius computed by any backward stable algorithm is less than or greater than 1.

Key words. matrix powers, rounding errors, Jordan canonical form, nonnormal matrices, pseudospectrum

AMS subject classifications. primary 65F99, 65G05

1. Introduction.

Many numerical processes depend for their success upon the powers of a matrix tending to zero. A fundamental example is stationary iteration for solving a linear system $Ax = b$, in which a sequence of vectors is defined by $Mx_{k+1} = Nx_k + b$, where $A = M - N$ and M is nonsingular. The errors $e_k = x - x_k$ satisfy $e_k = (M^{-1}N)^k e_0$, so the iteration converges for all x_0 if $(M^{-1}N)^k \to 0$ as $k \to \infty$. Many theorems are available about the convergence of stationary iteration, but virtually all of them are concerned with exact arithmetic (for exceptions see [12], [13] and the references therein). While the errors in stationary iteration are not precisely modelled by the errors in matrix powering, as matrix powers are not formed explicitly, the behaviour of the computed powers $fl((M^{-1}N)^k)$ can be expected to give some insight into the behaviour of stationary iteration (indeed, the basic error recurrences in [12] and [13] involve powers of $M^{-1}N$ acting on vectors of rounding errors).

In [18, Chap. 20], Ostrowski proves a theorem about a product of perturbed matrices $A + \Delta A_i$ that he states "assures the theoretical *stability of the convergence* of A^μ to 0 with respect to rounding off" as $\mu \to \infty$ for any matrix A with spectral radius $\rho(A) < 1$. Although Ostrowski's theorem is correct, its interpretation with respect to computed powers is not as simple as this statement implies, because for any finite precision arithmetic, no matter how accurate, there are matrices that are sensitive enough to perturbations to cause the theoretically convergent sequence of powers to diverge. To illustrate this point, Fig. 1.1[1] plots the 2-norms of the first 200 powers of a 14×14 nilpotent matrix C_{14} discussed by Trefethen and Trummer [23] (see §3 for details). The plot confirms the statement of these authors that the matrix is not power-bounded in floating point arithmetic, even though its 14th power should be zero. The powers for our plot were computed in MATLAB, which has unit

* Received by the editors September 22, 1993; accepted for publication (in revised form) by Charles Van Loan, March 25, 1994.

† Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov). The work of this author was supported by Science and Engineering Research Council grant GR/H52139.

‡ Department of Mathematics, University of Manchester, Manchester, M13 9PL, England. Present address: Department of Mathematics, University of Strathclyde, Glasgow, G1 1XH, Scotland (p.a.knight@strath.ac.uk). This author was supported by a Science and Engineering Research Council Research Studentship.

[1] As in all our plots of norms of powers, k on the x-axis is plotted against $\|fl(A^k)\|_2$ on the y-axis.

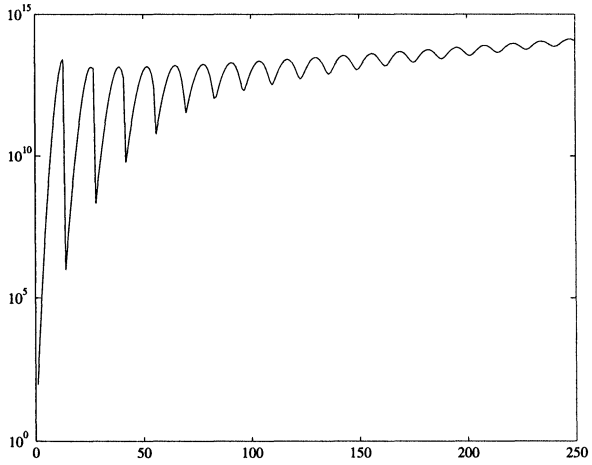

FIG. 1.1. *Diverging powers of a nilpotent matrix,* C_{14}.

roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$. Reichel and Trefethen [19] also give an example of a matrix that is nilpotent in theory but not power-bounded in practice. In this paper we determine conditions on a matrix A that ensure that the computed powers converge to zero.

In §2 we examine the behaviour of matrix powers in exact arithmetic. In particular, we review a number of bounds on the norms of powers. In §3 we use the Jordan canonical form of A to bound $\|fl(A^k)\|$ and to determine a sufficient condition for $fl(A^k) \to 0$ as $k \to \infty$. We also show that for certain matrices our bounds are tight. Finally, in §4 we rephrase our sufficient condition in terms of a pseudospectrum of A, under certain assumptions, including that A is diagonalizable; the modified result is not any sharper than the original, but offers an alternative viewpoint that is intuitively attractive.

In our analysis we use the standard model for floating point arithmetic:

$$fl(x \pm y) = x(1+\alpha) \pm y(1+\beta), \qquad |\alpha|, |\beta| \le u,$$
$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \le u, \quad \operatorname{op} = *, /,$$

where u is the unit roundoff. This model is valid for machines that do not use a guard digit in addition and subtraction.

We will use the Frobenius norm, $\|A\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$, and the p-norms $\|A\|_p = \max_{x \neq 0} \|Ax\|_p / \|x\|_p$, where $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ and $1 \le p \le \infty$. From §3 onwards we will drop the subscripts on $\|\cdot\|_p$, since all the norms from that point on are p-norms.

2. Matrix powers in exact arithmetic. We begin by discussing the behaviour of matrix powers in the absence of rounding errors. In exact arithmetic the limiting behaviour of the powers of $A \in \mathbf{C}^{n \times n}$ is determined by A's eigenvalues. If the spectral radius $\rho(A) < 1$ then $A^k \to 0$ as $k \to \infty$; if $\rho(A) > 1$, $A^k \to \infty$ as $k \to \infty$. If $\rho(A) = 1$ then $\|A^k\| \to \infty$ if A has a defective eigenvalue λ such that $|\lambda| = 1$; A^k does not converge if A has a nondefective eigenvalue $\lambda \neq 1$ such that $|\lambda| = 1$ (although the norms of the powers may converge); otherwise, the only eigenvalue of modulus 1 is the

nondefective eigenvalue 1, and A^k converges to a nonzero matrix. These statements are easily proved using the Jordan canonical form

$$A = XJX^{-1} \in \mathbf{C}^{n\times n}, \tag{2.1a}$$

where X is nonsingular and

$$J = \operatorname{diag}(J_1, J_2, \ldots, J_s), \qquad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbf{C}^{n_i\times n_i}, \tag{2.1b}$$

where $n_1 + n_2 + \cdots + n_s = n$. We will call a matrix for which $A^k \to 0$ as $k \to \infty$ (or equivalently, $\rho(A) < 1$) a *convergent matrix.*

The norm of a convergent matrix can be arbitrarily large, as is shown trivially by the example

$$A_2(\alpha) = \begin{bmatrix} \lambda & \alpha \\ 0 & \lambda \end{bmatrix}, \qquad |\lambda| < 1, \quad \alpha \gg 1. \tag{2.2}$$

While the spectral radius determines the asymptotic rate of growth of matrix powers, the norm influences the initial behaviour of the powers. The interesting result that $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$ for any norm (see [14, p. 299], for example) confirms the asymptotic role of the spectral radius. An important quantity is the "hump" $\max_k \|A^k\|/\|A\|$, which can be arbitrarily large for a convergent matrix, as can be seen from $A_3(\alpha)$, the 3×3 analogue of the matrix in (2.2), for which $\|A_3(\alpha)^2\|/\|A_3(\alpha)\| = O(\alpha)$. Figure 2.1 shows an example of the hump phenomenon: the plot is for $A_3(2)$ with $\lambda = 3/4$; here, $\|A_3(2)\|_2 = 3.57$. The shape of the plot is typical of that for a convergent matrix with norm bigger than 1. Note that if A is normal (so that in (2.1a) J is diagonal and X can be taken to be unitary) we have $\|A^k\|_2 = \|\operatorname{diag}(\lambda_i^k)\|_2 = \|A\|_2^k = \rho(A)^k$, so the problem of bounding $\|A^k\|$ is of interest only for nonnormal matrices. The hump phenomenon arises in various areas of numerical analysis. For example, it is discussed for matrix powers in the context of stiff differential equations by D. J. Higham and Trefethen [8], and by Moler and Van Loan [17] for the matrix exponential e^{At} with $t \to \infty$.

In the rest of this section we briefly survey bounds for $\|A^k\|$. First, however, we comment on the condition number $\kappa(X) = \|X\|\|X^{-1}\|$ that appears in various bounds in this paper. The matrix X in the Jordan form (2.1a) is by no means unique [3, pp. 220–221], [6]: if A has distinct eigenvalues (hence J is diagonal) then X can be replaced by XD, for any nonsingular diagonal D, while if A has repeated eigenvalues then X can be replaced by XT, where T is a block matrix with block structure conformal with that of J and which contains some arbitrary upper trapezoidal Toeplitz blocks. We adopt the convention that $\kappa(X)$ denotes the minimum possible value of $\kappa(X)$ over all possible choices of X. In general it is difficult to determine this optimal value. However, for any nonsingular X we have the bound

$$\kappa_F(X) \geq \sum_i \|x_i\|_2 \|y_i\|_2,$$

where $X = [x_1, \ldots, x_n]$ and $X^{-1} = [y_1, \ldots, y_n]^H$, with equality if there is a nonzero α such that $\|x_i\|_2 = \alpha\|y_i\|_2$ for all i [21, Thm. 4.3.5]. If A has distinct eigenvalues then

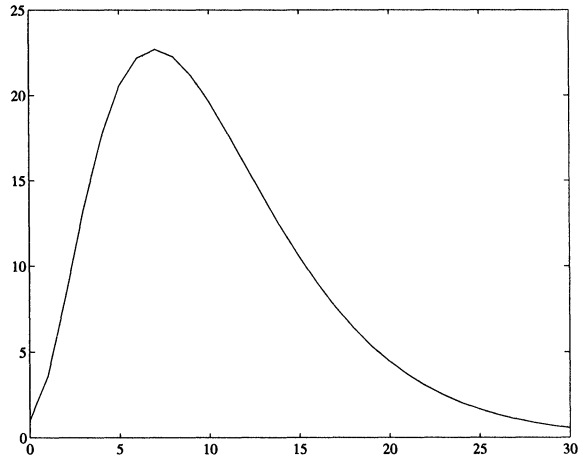

FIG. 2.1. *A typical hump for a convergent, nonnormal matrix.*

this lower bound is the same for all X in the Jordan form and is the minimum value of $\kappa_F(X)$. An alternative approach for matrices with distinct eigenvalues is to insist that the columns of X have unit 2-norm, for this gives a 2-norm condition number within a factor $n^{1/2}$ of the minimum, in view of a result by van der Sluis on diagonal scalings [24, Thm. 3.5]. However, we will see in §3 that to appreciate fully the various instability phenomena, we must consider defective problems.

If A is diagonalizable then, from (2.1a), we have the bound

$$\|A^k\|_p \leq \kappa_p(X)\rho(A)^k, \tag{2.3}$$

for any p-norm. (Since $\rho(A) \leq \|A\|$ for any norm, we also have the lower bound $\rho(A)^k \leq \|A^k\|_p$.) This bound is unsatisfactory for two reasons. First, by choosing A to have well-conditioned large eigenvalues and ill-conditioned small eigenvalues we can make the bound arbitrarily pessimistic. Second, it models norms of powers of convergent matrices as monotonically decreasing sequences, which is qualitatively incorrect if there is a large hump.

The Jordan canonical form can also be used to bound the norms of the powers of a defective matrix. If XJX^{-1} is the Jordan canonical form of $\delta^{-1}A$ then

$$\|A^k\|_p \leq \kappa_p(X)(\rho(A) + \delta)^k, \tag{2.4}$$

for all $\delta > 0$. This is a special case of a result of Ostrowski [18, Thm. 20.1], and a proof is straightforward: We can write $\delta^{-1}A = X(\delta^{-1}D + M)X^{-1}$, where $D = \operatorname{diag}(\lambda_i)$ and M is the off-diagonal part of the Jordan form. Then $A = X(D + \delta M)X^{-1}$, and (2.4) follows by taking norms. An alternative way of writing this bound is

$$\|A^k\|_p \leq \kappa_p(X)\kappa_p(D)(\rho(A) + \delta)^k,$$

where $A = XJX^{-1}$ and $D = \operatorname{diag}(\delta^{n-1}, \delta^{n-2}, \ldots, 1)$. Note that this is not the same X as in (2.4): multiplying A by a scalar changes $\kappa(X)$ when A is not diagonalizable. Both bounds suffer from the same problems as the bound (2.3) for diagonalizable matrices.

Another bound in terms of the Jordan canonical form (2.1) of A is given by Gautschi [4]. For convergent matrices, it can be written in the form

$$\|A^k\|_F \le c\, k^{p-1} \rho(A)^k, \tag{2.5}$$

where $p = \max\{n_i : \lambda_i \ne 0\}$ and c is a constant depending only on A (c is not defined explicitly in [4]). The factor k^{p-1} makes this bound somewhat more effective at predicting the shapes of the actual curve than (2.4), but again c can be unsuitably large.

Another way to estimate $\|A^k\|$ is to introduce a measure of nonnormality. Consider the Schur decomposition $Q^H A Q = D + N$, where N is strictly upper triangular, and let S represent the set of all such N. The nonnormality of A can be measured by Henrici's departure from normality [7]

$$\Delta(A, \|\cdot\|) \equiv \Delta(A) = \min_{N \in S} \|N\|.$$

For the Frobenius norm, Henrici shows that $\|N\|_F$ is independent of the particular Schur form and that

$$\Delta_F(A) = \left(\|A\|_F^2 - \sum_i |\lambda_i|^2\right)^{1/2} \le \left(\frac{n^3 - n}{12}\right)^{1/4} \|A^H A - A A^H\|_F^{1/2}.$$

László [15] has recently shown that $\Delta_F(A)$ is within a constant factor of the distance from A to the nearest normal matrix:

$$\Delta_F(A)/\sqrt{n} \le \nu(A) \le \Delta_F(A),$$

where $\nu(A) = \min\{\|E\|_F : A + E \text{ is normal}\}$. Henrici uses the departure from normality to derive the 2-norm bounds

$$\|A^k\|_2 \le \begin{cases} \displaystyle\sum_{i=0}^{n-1} \binom{k}{i} \rho(A)^{k-i} \Delta_2(A)^i, & \rho(A) > 0, \\ \Delta_2(A)^k, & \rho(A) = 0 \text{ and } k < n. \end{cases} \tag{2.6}$$

Empirical evidence suggests that the first bound in (2.6) can be very pessimistic. However, for normal matrices both the bounds are equalities. A bound of the same form as the first bound in (2.6), but with $\|A\|_2$ replacing $\Delta_2(A)$ and with an extra factor $2^{(n-1)/2}$, is obtained from a bound of Stafney in [20, Thm. 2.1] for $\|p(A)\|$, where p is a polynomial.

Another bound involving nonnormality is given by Golub and Van Loan [5, Lem. 7.3.2]. They show that, in the above notation,

$$\|A^k\|_2 \le (1+\theta)^{n-1} \left(\rho(A) + \frac{\Delta_F(A)}{1+\theta}\right)^k$$

for any $\theta \ge 0$. This bound is an analogue of (2.4) with the Schur form replacing the Jordan form. Again, there is equality when A is normal (if we set $\theta = 0$).

To compare bounds based on the Schur form with ones based on the Jordan form we need to compare $\Delta(A)$ with $\kappa(X)$. If A is diagonalizable then [16, Thm. 4]

$$\kappa_2(X) \ge \left(1 + \frac{\Delta_F(A)^2}{\|A\|_F^2}\right)^{1/2},$$

and it can be shown by a 2×2 example that $\min_X \kappa_2(X)$ can exceed $\Delta_F(A)/\|A\|_F$ by an arbitrary factor [2, §8.1.2], [1, §4.2.7].

Another tool that can be used to bound the norms of powers is the pseudospectrum of a matrix [22]. The ϵ-pseudospectrum of $A \in \mathbf{C}^{n\times n}$ is defined for a given $\epsilon > 0$ to be the set

$$\Lambda_\epsilon(A) = \{\, z : z \text{ is an eigenvalue of } A + E \text{ for some } E \text{ with } \|E\|_2 \le \epsilon \,\},$$

and it can also be represented, in terms of the resolvent $(zI - A)^{-1}$, as

$$\Lambda_\epsilon(A) = \{\, z : \|(zI - A)^{-1}\|_2 \ge \epsilon^{-1} \,\}.$$

As Trefethen notes [22], by using the Cauchy integral representation of A^k (which involves a contour integral of the resolvent) one can show that

$$\|A^k\|_2 \le \epsilon^{-1} \rho_\epsilon(A)^{k+1}, \tag{2.7}$$

where the ϵ-pseudospectral radius

$$\rho_\epsilon(A) = \max\{\, |z| : z \in \Lambda_\epsilon(A) \,\}. \tag{2.8}$$

This bound is very similar in flavour to (2.4). The difficulty is transferred from estimating $\kappa(X)$ to choosing ϵ and estimating $\rho_\epsilon(A)$.

Finally, we mention that the Kreiss matrix theorem provides a good estimate of $\sup_{k\ge 0} \|A^k\|$ for a general $A \in \mathbf{C}^{n\times n}$, albeit in terms of an expression that involves the resolvent and is not easy to compute:

$$r(A) \le \sup_{k\ge 0} \|A^k\|_2 \le n\, e\, r(A),$$

where $r(A) = \sup\{\, (|z| - 1)\|(zI - A)^{-1}\|_2 : |z| > 1 \,\}$ and $e = \exp(1)$. Details and references are given by Wegert and Trefethen [25].

3. Bounds for finite precision arithmetic. The formulae $A \cdot A^k$ or $A^k \cdot A$ can be implemented in several ways, corresponding to different loop orderings in each individual product, but as long as each product is formed using the standard formula $(AB)_{ij} = \sum_k a_{ij} b_{kj}$, all these variations satisfy the same rounding error bounds. We do not analyse here the use of the binary powering technique, where, for example, A^9 is formed as $A((A^2)^2)^2$, alternate multiplication on the left and right: $fl(A^k) = fl(Afl(A^{k-2})A)$, or the use of fast matrix multiplication techniques such as Strassen's method, since none of these methods is equivalent to repeated multiplication in finite precision arithmetic.

We suppose, without loss of generality, that the columns of A^m are computed one at a time, the jth as $fl(A(A(\ldots(Ae_j)\ldots)))$, where e_j is the jth unit vector. Standard error analysis shows that the jth computed column of A^m satisfies

$$fl(A^m e_j) = (A + \Delta A_1)(A + \Delta A_2)\ldots(A + \Delta A_m)e_j, \tag{3.1}$$

where

$$|\Delta A_i| \le c_n u |A|, \tag{3.2}$$

with c_n a constant of order n. (The inequality and absolute value are taken componentwise.) This bound holds for both real and complex matrices. It follows that

$$|fl(A^m e_j)| \le (1 + c_n u)^m |A|^m e_j,$$

and so a sufficient condition for convergence of the computed powers is that

$$\rho(|A|) < \frac{1}{1+c_n u}.$$

This result is useful in certain special cases: $\rho(|A|) = \rho(A)$ if A is triangular or has a checkerboard sign pattern (since then $|A| = DAD^{-1}$ where $D = \mathrm{diag}(\pm 1)$); if A is normal then $\rho(|A|) \leq \sqrt{n}\rho(A)$ (this bound being attained for a Hadamard matrix); and in Markov processes, where the a_{ij} are transition probabilities, $|A| = A$. However, in general $\rho(|A|)$ can exceed $\rho(A)$ by an arbitrary factor.

To obtain sharper and more informative results it is necessary to use more information about the matrix. Although the Jordan form is usually avoided by numerical analysts because of its sensitivity to perturbations, it is convenient to work with in this application and leads to informative results.

We point out that, because the analysis below is based on (3.1), our proofs of sufficient conditions for $fl(A^m) \to 0$ yield, with only trivial changes, sufficient conditions for $fl(A^m b) \to 0$, for any vector b. These conditions do not, however, exploit any special relations between A and b (such, as for example, b being an eigenvector of A).

3.1. Nilpotent matrices. We begin by considering nilpotent matrices, that is, those whose spectral radius is zero. The fact that nth power of an $n \times n$ nilpotent matrix is zero simplifies the analysis. The following theorem gives a bound on the norm of a computed power, together with a condition for the limit of the powers to be zero.

THEOREM 3.1. *Let $A \in \mathbf{C}^{n\times n}$ be a nilpotent matrix with the Jordan canonical form (2.1). A sufficient condition for $fl(A^m) \to 0$ as $m \to \infty$ is*

$$d_n u \kappa(X)\|A\| < 1 \tag{3.3}$$

for some p-norm, where u is the unit roundoff and d_n is a modest constant that depends only on n. Furthermore, if, for some $k \geq 1$ and $\theta > 1$,

$$\theta d_n u \kappa(X)^{\frac{k+1}{k}}\|A\| < 1, \tag{3.4}$$

then, with $t = \max_i n_i$,

$$\|fl(A^{rt})\| \leq n^{1-\frac{1}{p}} \frac{\kappa(X)^{1-\frac{r}{k}}\theta^{-r}}{1-\theta^{-1}\kappa(X)^{-\frac{1}{k}}} = O(\theta^{-r}), \quad r \geq k. \tag{3.5}$$

Proof. Taking norms in (3.1) we have

$$\|fl(A^m e_j)\| \leq \|(A+\Delta A_1)(A+\Delta A_2)\dots(A+\Delta A_m)\|.$$

Using the inequality

$$\|A\| \leq n^{1-\frac{1}{p}} \max_j \|Ae_j\| \tag{3.6}$$

from [10], we have

$$\|fl(A^m)\| \leq n^{1-\frac{1}{p}}\|(A+\Delta A_1)(A+\Delta A_2)\dots(A+\Delta A_m)\|.$$

Expanding this product and collecting together terms with the same number of ΔA_i factors we obtain the bound

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^m)\| \le \|A^m\| &+ \sum_{i=1}^{m} \|A^{i-1}\Delta A_i A^{m-i}\| \\ &+ \sum_{i=1}^{m-1}\sum_{j=1}^{m-i} \|A^{i-1}\Delta A_i A^{j-1}\Delta A_{i+j} A^{m-i-j}\| + \cdots \\ &+ \|\Delta A_1 \Delta A_2 \ldots \Delta A_m\|. \end{aligned}$$

From (3.2) we find, using (3.6) and an analogous result involving duality, that $\|\Delta A_i\| \le c_n' u\|A\|$, where $c_n' = n^{\min(1/p,1-1/p)}c_n$. Since $A^l = X^{-1}J^lX$ we have

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^m)\| \le \|A^m\| &+ \kappa(X)^2 c_n' u\|A\| \sum_{i=1}^{m} \|J^{i-1}\|\|J^{m-i}\| \\ &+ \kappa(X)^3 (c_n' u\|A\|)^2 \sum_{i=1}^{m-1}\sum_{j=1}^{m-i} \|J^{i-1}\|\|J^{j-1}\|\|J^{m-i-j}\| + \cdots \\ &+ (c_n' u\|A\|)^m. \end{aligned} \tag{3.7}$$

Now let $m = rt$, where $r \ge 1$. Since A is nilpotent, $A^t = J^t = 0$, and since every term in the first $r-1$ summations in (3.7) contains a factor $\|J^i\|$ with $i \ge t$, all these terms disappear. Furthermore, in the remaining summations we need only count terms in which all the exponents of J are less than t (again, the other terms disappear). Overall, we have, using the fact that $\|J^i\| = 1$ $(0 \le i < t)$,

$$n^{\frac{1}{p}-1}\|fl(A^{rt})\| \le \kappa(X) \sum_{j=r}^{rt} \left(tc_n' u\kappa(X)\|A\|\right)^j.$$

Now suppose that (3.4) holds with $d_n = tc_n'$, for some $\theta > 1$. Then, for $r \ge k$,

$$\begin{aligned} n^{\frac{1}{p}-1}\|fl(A^{rt})\| &\le \kappa(X) \sum_{j=r}^{rt} \left(\theta\kappa(X)^{\frac{1}{k}}\right)^{-j} \\ &\le \kappa(X)^{1-\frac{r}{k}}\theta^{-r} \sum_{j=0}^{rt-r} \left(\theta\kappa(X)^{\frac{1}{k}}\right)^{-j} \\ &\le \frac{\kappa(X)^{1-\frac{r}{k}}\theta^{-r}}{1-\theta^{-1}\kappa(X)^{-\frac{1}{k}}}. \end{aligned}$$

This gives the second part of the theorem. The first part follows immediately by choosing $\theta = 1+\epsilon$, with ϵ an arbitrarily small positive number, and taking the limit as $k \to \infty$. □

In practice we may have a computed matrix $\widehat{A} \approx A$ that is not exactly nilpotent. As long as $\|A - \widehat{A}\| \le c_n u\|A\|$, we can absorb the error $A - \widehat{A}$ into the terms ΔA_i in the proof, and so by applying the theorem to A we will obtain conclusions valid for $\widehat{A}$.

To exhibit the sharpness of the bounds we give the following example, using the Chebyshev spectral differentiation matrix $C_n \in \mathbb{R}^{n\times n}$ described in [23]. The matrix

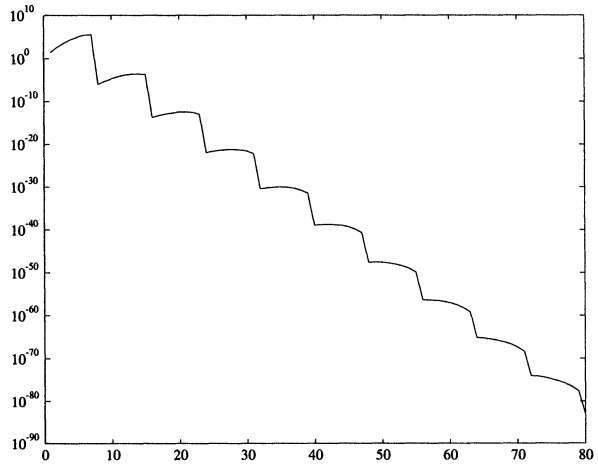

FIG. 3.1. *Converging powers of the nilpotent matrix* C_8.

C_n arises from degree $n-1$ polynomial interpolation of n arbitrary data values at n Chebyshev points, including a boundary condition at $x_0 = 1$. It is nilpotent and is similar to a single Jordan block of dimension n. We generate C_n in MATLAB using the routine **chebspec** from the test collection of Higham [9], [11].

Figure 3.1 shows the 2-norms of the computed powers of C_8 and Fig. 1.1 those of C_{14}. The powers of C_8 converge to zero, while the powers of C_{14} diverge.

To check the sharpness of the bounds in Theorem 3.1 we need an estimate of the condition number of the matrix X_n in the Jordan canonical form of C_n. We outline our approach in Appendix A. Our estimate for $\kappa_2(X_8)$ is 3.42×10^5, and $\|C_8\|_2 = 28.56$. Table 3.1 gives the order of the bound (3.5) for a number of powers, with $r = k$ and θ chosen as large as possible so that (3.4) is satisfied (we take $d_n = n$, instead of the actual value $d_n \approx n^{5/2}$ for this example, to allow for the inevitable overestimation of errors inherent in a strict rounding error bound of this type). The actual computed order is given for comparison and clearly there is reasonable agreement. According to (3.3), we require $d_n \kappa(X)\|A\|u < 1$ to guarantee that the computed powers of A converge to zero. For C_{14} we have $u\kappa_2(X)\|C_{14}\|_2 \approx 0.28$ so, allowing for d_n, (3.3) correctly does not predict convergence of the computed powers.

To emphasize that the behaviour of the computed powers is scale-dependent, we mention that the computed powers of $15C_8$ diverge. Again, this is in accord with Theorem 3.1 because $u\kappa_2(X)\|15C_8\|_2 \approx 2.7$. Finally, we note that for C_{12}, Theorem 3.1 again correctly predicts convergence of the computed powers, but the powers computed by alternate left and right multiplication and by binary powering diverge; this confirms that our analysis is not applicable to these forms of multiplication.

3.2. General matrices. Now we turn to general convergent matrices. In contrast to the theory we have developed for nilpotent matrices, we need separate theorems to describe the limiting behaviour of the matrix powers and to bound the norm for a finite exponent. In the following theorem we give a sufficient condition, based on the Jordan canonical form, for the computed powers of a matrix to converge to zero.

TABLE 3.1
Expected and actual orders of $\|fl(C_8^m)\|_2$.

Power	Bound	Actual
$m = 8$	10^{-2}	10^{-6}
$m = 16$	10^{-11}	10^{-14}
$m = 32$	10^{-27}	10^{-31}
$m = 64$	10^{-59}	10^{-66}

THEOREM 3.2. *Let* $A \in \mathbf{C}^{n\times n}$ *with the Jordan form* (2.1) *have spectral radius* $\rho(A) < 1$. *A sufficient condition for* $fl(A^m) \to 0$ *as* $m \to \infty$ *is*

$$d_n u\kappa(X)\|A\| < (1-\rho(A))^t \tag{3.8}$$

for some p-norm, where $t = \max_i n_i$ *and* d_n *is a modest constant depending only on* n.

Proof. Since any two p-norms differ by a factor at most n, we need only show convergence for one particular norm. We choose the ∞-norm.

It is easy to see that if we can find a nonsingular matrix S such that

$$\|S^{-1}AS\|_\infty + \kappa_\infty(S)\|\Delta A_i\|_\infty < 1 \tag{3.9}$$

for all i, then the product $(A+\Delta A_1)\dots(A+\Delta A_m) = S(S^{-1}AS+S^{-1}\Delta A_1 S)\dots(S^{-1}AS+S^{-1}\Delta A_m S)S^{-1} \to 0$ as $m \to \infty$. In the rest of the proof we construct such a matrix S for the ΔA_i in (3.1).

Let $P(\epsilon) = \operatorname{diag}(P_1(\epsilon),\dots,P_s(\epsilon))$ where $0 < \epsilon < 1-\rho(A)$ and

$$P_i(\epsilon) = \operatorname{diag}((1-|\lambda_i|-\epsilon)^{1-n_i}, (1-|\lambda_i|-\epsilon)^{2-n_i},\dots,1) \in \mathbb{R}^{n_i\times n_i}.$$

Now consider the matrix $P(\epsilon)^{-1}JP(\epsilon)$. Its ith diagonal block is of the form $\lambda_i I + (1-|\lambda_i|-\epsilon)N$, where the only nonzeros in N are 1s on the first superdiagonal, and so

$$\|P(\epsilon)^{-1}X^{-1}AXP(\epsilon)\|_\infty = \|P(\epsilon)^{-1}JP(\epsilon)\|_\infty \le \max_i(|\lambda_i| + 1 - |\lambda_i| - \epsilon) = 1-\epsilon.$$

Defining $S = XP(\epsilon)$, we have $\|S^{-1}AS\|_\infty \le 1-\epsilon$ and

$$\kappa_\infty(S) \le \kappa_\infty(P(\epsilon))\kappa_\infty(X) \le (1-\rho(A)-\epsilon)^{1-t}\kappa_\infty(X). \tag{3.10}$$

Now we set $\epsilon = \theta(1-\rho(A))$ where $0 < \theta < 1$ and we determine θ so that (3.9) is satisfied, that is, so that $\kappa_\infty(S)\|\Delta A_i\|_\infty < \epsilon$ for all i. From (3.2) and (3.10) we have

$$\kappa_\infty(S)\|\Delta A_i\|_\infty \le c_n u(1-\theta)^{1-t}(1-\rho(A))^{1-t}\kappa_\infty(X)\|A\|_\infty.$$

Therefore (3.9) is satisfied if

$$c_n u(1-\theta)^{1-t}(1-\rho(A))^{1-t}\kappa_\infty(X)\|A\|_\infty < \theta(1-\rho(A)),$$

that is, if

$$c_n u\kappa_\infty(X)\|A\|_\infty < (1-\theta)^{t-1}\theta(1-\rho(A))^t.$$

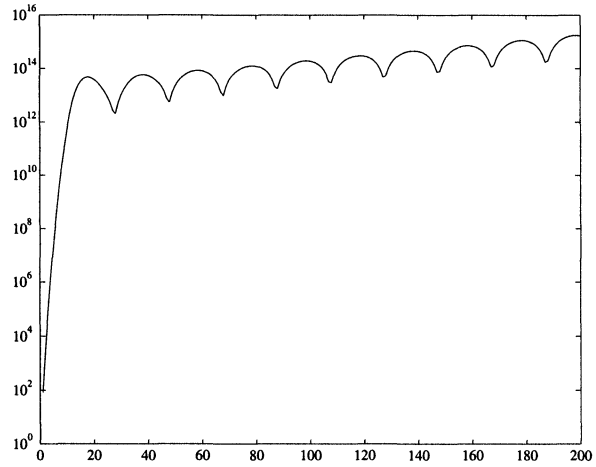

FIG. 3.2. *Diverging powers of* $C_{13} + 0.36I$.

If the integer t is greater than 1 then the function $f(\theta) = (1-\theta)^{t-1}\theta$ has a maximum on $[0,1]$ at $\theta_* = t^{-1}$ and $f(\theta_*) = (t-1)^{-1}(1-t^{-1})^t$ satisfies $(4(t-1))^{-1} \le f(\theta_*) < e^{-1}$. We conclude that for all integers $1 \le t \le n$,

$$c_n u \kappa_\infty(X)\|A\|_\infty < \frac{1}{4t}(1-\rho(A))^t$$

is sufficient to ensure that (3.9) holds. The theorem is proved with $d_n = 4tc_n$. □

If A is normal then $\|A\|_2 = \rho(A) < 1$, $t = 1$, and $\kappa_2(X) = 1$, so (3.8) takes the form

$$\rho(A) < \frac{1}{1+d_n u}.$$

This condition is also easily derived by taking 2-norms in (3.1) and (3.2).

Again, we can show the sharpness of this bound by using the Chebyshev spectral differentiation matrix C_n, this time adding multiples of the identity matrix.

Figure 3.2 shows the nonconverging 2-norms of the first 200 computed powers of $A = C_{13} + 0.36I$. Since the same matrix X takes both C_{13} and A to Jordan form, we can use the same routines as for our nilpotent examples to estimate $\kappa(X)$. Our estimate for $\kappa_2(X)\|A\|_2 u/(1-\rho(A))^{13}$ in this case is 3.05. On the other hand, the computed powers of $A = C_{13}+0.01I$ converge to zero, and $\kappa_2(X)\|A\|_2 u/(1-\rho(A))^{13} \approx 0.01$. Thus our bound (3.8) is reasonably sharp.

Figure 3.2 reveals an interesting scalloping pattern in the curve of the norms. In Figs. 1.1 and 3.1 for nilpotent matrices the norm dips whenever the power is a multiple of the dimension of the matrix. Here the norm first dips for $\|fl(A^{28})\|_2$ and then regularly after every further 20 powers, but the point of first dip and the dipping intervals can be altered by adding different multiples of the identity matrix. The reason for this behaviour is not clear.

A difficulty we have when attempting to bound $\|fl(A^m)\|$ for a finite m is that, as explained in §2, we do not have a good estimate of the true value $\|A^m\|$. If we do

have such an estimate we can prove results similar to (3.4) and (3.5), although we are not able to determine a precise bound for $\|fl(A^m)\|$ in simple form.

THEOREM 3.3. *Let $A \in \mathbf{C}^{n\times n}$ with the Jordan form (2.1) have spectral radius $\rho(A) < 1$. Let q be such that $\|A^q\| = cu$ where $c = O(1)$ and suppose that, for some $k \geq 1$ and $\theta > 1$,*

$$\theta d_n q u \mu^{\frac{k+1}{k}} \|A\| < 1,$$

where $\mu = \kappa(X)/(1-\rho(A))^{t-1}$ and $t = \max_i n_i$. Then

$$\|fl(A^{rq})\| = O(\theta^{-r}), \quad r \geq k.$$

Proof. We omit the proof of the theorem, which is very similar to the proof of Theorem 3.1. □

We conclude this section by commenting that the proof of Theorem 3.2 can be adapted to use the Schur decomposition in place of the Jordan canonical form. The modified analysis leads to the sufficient condition

$$d_n' u \|N\|_F^{n-1} \|A\|_2 < (1-\rho(A))^n \tag{3.11}$$

for $fl(A^m) \to 0$ as $m \to \infty$, where N is the strictly upper triangular part of the Schur form. This condition is weaker than (3.8) in two respects. First, it takes no account of the defectiveness of A, because it contains a power n on the right-hand side instead of $t = \max_i n_i \leq n$. Second, under the scaling $A \leftarrow \alpha A$ the left-hand side of (3.11) scales by $|\alpha|^n$, which tends to make the left-hand side of (3.11) much larger than that of (3.8) when $\|A\|_F > 1$. It is an open question how to obtain a sharp sufficient condition for convergence in terms of the Schur decomposition.

4. A pseudospectral approach. In this section we show how the pseudospectrum can be used to predict the limiting behaviour of a computed sequence of powers. Figure 4.1 shows approximations to pseudospectra for the matrices in the examples of Figs. 1.1, 3.1, and 3.2; we have approximated $\Lambda_\epsilon(A)$ with $\epsilon = c_n\|A\|_2 u$, taking $c_n = n$ for simplicity. The inner ring is an approximation to the pseudospectrum of C_8, that of C_{14} is marked by + and that of $C_{13} + 0.36I$ is marked by o. The solid curve is the unit disc.

A heuristic argument based on (3.1) and (3.2) suggests that, if for randomly chosen perturbations ΔA_i with $\|\Delta A_i\| \leq c_n u \|A\|$, most of the eigenvalues of the perturbed matrices lie outside the unit disc, then we can expect a high percentage of the terms $A + \Delta A_i$ in (3.1) to have spectral radius bigger than one and hence we can expect the product to diverge. On the other hand, if the $c_n u\|A\|$-pseudospectrum is wholly contained within the unit disc, each $A+\Delta A_i$ will have spectral radius less than one and the product can be expected to converge. (Note, however, that if $\rho(A) < 1$ and $\rho(B) < 1$ it is not necessarily the case that $\rho(AB) < 1$.) To make this heuristic precise, we need an analogue of Theorem 3.2 phrased in terms of the pseudospectrum rather than the Jordan form.

To obtain such an analogue directly from Theorem 3.2 we need to relate $\kappa(X)$ to the pseudospectral radius $\rho_\epsilon(A)$ (see (2.8)). If we can show that

$$\rho_\epsilon(A) \geq \rho(A) + (c_n \epsilon \kappa(X))^{1/t} \tag{4.1}$$

for a particular ϵ, then $\rho_\epsilon(A) < 1$ implies $c_n\epsilon\kappa(X) < (1-\rho(A))^t$, which is a condition of the same form as (3.8). In Theorem 4.2 we show that (4.1) holds for diagonalizable

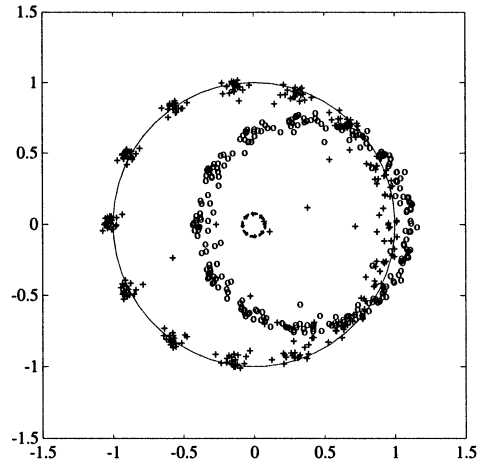

FIG. 4.1. *Pseudospectra of the three example matrices.*

matrices to first order, under a certain assumption. We need the following standard result (see, for example, [21, pp. 183–184]).

THEOREM 4.1. *Let λ be a simple eigenvalue of the matrix A, with right and left eigenvectors x and y, and let $\widetilde{A} = A + E$ be a perturbation of A. Then there is an eigenvalue $\widetilde{\lambda}$ of $\widetilde{A}$ such that*

$$\widetilde{\lambda} = \lambda + \frac{y^H E x}{y^H x} + O(\|E\|^2).$$

We can now prove the following theorem.

THEOREM 4.2. *Let $A \in \mathbf{C}^{n\times n}$ have the Jordan canonical form (2.1), with $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Suppose that $\|X\|_1 = \sum_{i=1}^n |x_{i1}|$ and $\|X^{-1}\|_\infty = \sum_{j=1}^n |y_{1j}|$, where $X^{-1} = (y_{ij})$. Then there is a perturbation $\widetilde{A} = A + E$ of A, with $\|E\| = \epsilon$ for all p-norms, such that*

$$\rho(\widetilde{A}) \geq \rho(A) + \frac{\kappa(X)\epsilon}{n^2} + O(\epsilon^2). \tag{4.2}$$

Proof. By assumption, λ_1 is a simple eigenvalue, so $x_1 = Xe_1$ and $y_1 = (e_1^T X^{-1})^H$ are the right eigenvector and the left eigenvector corresponding to λ_1. From Theorem 4.1 we know that any perturbation $\widetilde{A}$ will have an eigenvalue

$$\widetilde{\lambda} = \lambda_1 + y_1^H E x_1 + O(\|E\|^2)$$

(since $y_1^H x_1 = 1$). Define E by $|e_{ij}| = \epsilon/n$ and $\arg(e_{ij}) = \arg(y_{1i}) - \arg(x_{j1}) + \arg(\lambda_1)$. Then E has rank 1, $\|E\| = \epsilon$ for all p, and

$$\widetilde{\lambda} = e^{i\arg(\lambda_1)}(|\lambda_1| + \frac{\epsilon}{n}\|y_1^H\|_1\|x_1\|_1) + O(\epsilon^2). \tag{4.3}$$

Now for an $n \times n$ matrix B and any $1 \leq p, q \leq \infty$ [10],

$$\|B\|_p \leq n^{\left(\frac{1}{\min(p,q)} - \frac{1}{\max(p,q)}\right)}\|B\|_q,$$

and together with the conditions of the theorem this gives

$$\|y_1^H\|_1\|x_1\|_1 = \|X^{-1}\|_\infty\|X\|_1 \geq \frac{\kappa(X)}{n}.$$

The proof is completed by taking absolute values in (4.3). □

Because of the assumptions on $\|X\|_1$ and $\|X^{-1}\|_\infty$ we do not have the freedom to choose X to minimize $\kappa(X)$ in Theorem 4.2. We note, however, that if A is diagonalizable and X has columns all of the same norm, then the condition in Theorem 4.2 on the rows and columns of X and X^{-1} reduces to the requirement that the eigenvalue of largest modulus be the most ill conditioned.

Theorem 4.2 enables us to obtain the following corollary of Theorem 3.2.

COROLLARY 4.3. *Suppose that $A \in \mathbf{C}^{n\times n}$ is diagonalizable and satisfies the conditions of Theorem 4.2, and suppose that the $O(\epsilon^2)$ term in (4.2) is negligible. If $\rho_\epsilon(A) < 1$ for $\epsilon = c_n\|A\|u$, where c_n is a modest constant depending only on n, then $\lim_{m\to\infty} fl(A^m) = 0$.*

Proof. By Theorem 4.2, if $\rho_\epsilon(A) < 1$ then

$$\rho(A) + \epsilon\kappa(X)/n^2 < 1.$$

Rearranging gives

$$c_n u\kappa(X)\|A\|/n^2 < 1 - \rho(A).$$

Using Theorem 3.2 we have the required result for $c_n = n^2 d_n$, since $t = 1$. □

Suppose we compute the eigenvalues of A by a backward stable algorithm, that is, one that yields the exact eigenvalues of $A + E$, where $\|E\|_2 \leq c_n u\|A\|_2$, with c_n a modest constant. (An example of such an algorithm is the QR algorithm [5, §7.5]). Then the computed spectral radius $\widehat{\rho}$ satisfies $\widehat{\rho} \leq \rho_{c_n u\|A\|_2}(A)$. In view of Corollary 4.3 we can formulate a rule of thumb:

> The computed powers of A can be expected to converge to zero if the spectral radius computed via a backward stable eigensolver is less than 1.

This rule of thumb has also been discussed by Trefethen and Trummer [23] and Reichel and Trefethen [19]. In our experience the rule of thumb is fairly reliable when $\widehat{\rho}$ is not too close to 1. For the matrices used in our examples we have

$$\widehat{\rho}(C_8) = 0.073, \quad \widehat{\rho}(15C_8) = 2.7, \quad \widehat{\rho}(C_{14}) = 1.005,$$
$$\widehat{\rho}(C_{13} + 0.01I) = 0.70, \quad \widehat{\rho}(C_{13} + 0.36I) = 1.05,$$

and we observed convergence of the computed powers for C_8 and $C_{13} + 0.01I$ and divergence for the other matrices.

Appendix A. Approximating X in the jordan form of C_n. In §3 we needed an estimate of $\kappa(X)$ for the Chebyshev differentiation matrix, C, where $C = XJX^{-1}$ is the Jordan form. In this appendix we outline our approach for computing an estimate of $\kappa(X)$.

Recall that

$$C = XJX^{-1}, \tag{A.1}$$

where J is a single Jordan block whose diagonal is zero. Suppose we decompose C into Schur form via the orthogonal matrix Q (which is real since C's spectrum is real), that is,

$$C = QTQ^T,$$

where T is upper triangular with zero diagonal. If we can find an upper triangular matrix R such that $T = RJR^{-1}$ then $X = QR$ and $\kappa_2(X) = \kappa_2(R)$. We require $TR = RJ$, that is, $Tr_j = r_{j-1}$, $2 \le j \le n$, and $Tr_n = 0$, where r_j is the jth column of R.

We choose the arbitrary last column of R to be the last column of the identity matrix. The following algorithm computes R (here, we use the MATLAB colon notation).

$R(:,n) = e_n$
for $j = n-1:-1:1$
  $R(1:j,j) = T(1:j,1:j+1)R(1:j+1,j+1)$
end

It remains to compute the Schur form of C. We do not use the QR algorithm to compute the Schur form, as for nilpotent matrices it can lead to triangular matrices with elements of order 1 on the diagonal. We use the following algorithm described by Golub and Wilkinson in [6, §10], which, although computationally expensive, has good error properties.

Compute the SVD of $C_1 = C = U_1\Sigma_1 V_1^T$.
for $i = 1:n-2$
  $C_{i+1} = V_i^T C_i V_i$
  Compute the SVD $C_{i+1}(1:n-i, 1:n-i) = U_{i+1}\Sigma_{i+1}W_{i+1}^T$.
  $V_{i+1} = \mathrm{diag}(W_{i+1}, I_i)$
end
$L = V_{n-1}^T C_{n-1} V_{n-1}$
$Q = V_1 V_2 \ldots V_{n-1}$

Upon completion of the algorithm we have $C = QLQ^T$ with L lower triangular, and so we apply our first algorithm to L^T to estimate $\kappa_2(X)$ (note that the Jordan matrix for A^T is a permutation of the one for A [14, §3.2.3]).

Acknowledgments. We thank Des Higham and Nick Trefethen for their comments on the manuscript.

REFERENCES

[1] F. CHATELIN, *Eigenvalues of Matrices*, John Wiley, Chichester, 1993.
[2] F. CHATELIN AND V. FRAYSSÉ, *Qualitative computing: Elements of a theory for finite precision computation*, lecture notes, CERFACS, Toulouse, France and THOMSON-CSF, Orsay, France, June 1993.
[3] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[4] W. GAUTSCHI, *The asymptotic behaviour of powers of matrices*, Duke Math. J., 20 (1953), pp. 127–140.
[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[6] G. H. GOLUB AND J. H. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.
[7] P. HENRICI, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.

[8] D. J. HIGHAM AND L. N. TREFETHEN, *Stiffness of ODEs*, BIT, 33 (1993), pp. 285–303.
[9] N. J. HIGHAM, *Algorithm* 694: *A collection of test matrices in MATLAB*, ACM Trans. Math. Software, 17 (1991), pp. 289–305.
[10] ———, *Estimating the matrix p-norm*, Numer. Math., 62 (1992), pp. 539–555.
[11] ———, *The Test Matrix Toolbox for Matlab*, Numerical Analysis Report No. 237, University of Manchester, England, Dec. 1993.
[12] N. J. HIGHAM AND P. A. KNIGHT, *Componentwise error analysis for stationary iterative methods*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., Vol. 48, IMA Volumes in Mathematics and its Applications, Springer-Verlag, New York, 1993, pp. 29–46.
[13] ———, *Finite precision behavior of stationary iteration for solving singular systems*, Linear Algebra Appl., 192 (1993), pp. 165–186.
[14] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, 1985.
[15] L. LÁSZLÓ, *An attainable lower bound for the best normal approximation*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1035–1043.
[16] G. LOIZOU, *Nonnormality and Jordan condition numbers of matrices*, J. Assoc. Comput. Mach., 16 (1969), pp. 580–584.
[17] C. B. MOLER AND C. F. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.
[18] A. M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973; Solution of Equations and Systems of Equations, 3rd ed.
[19] L. REICHEL AND L. N. TREFETHEN, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.
[20] J. D. STAFNEY, *Functions of a matrix and their norms*, Linear Algebra Appl., 20 (1978), pp. 87–94. Correction in Linear Algebra Appl., 39 (1981), pp. 259–260.
[21] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, London, 1990.
[22] L. N. TREFETHEN, *Pseudospectra of matrices*, in Numerical Analysis 1991, Proc. 14th Dundee Conference, D. F. Griffiths and G. A. Watson, eds., Vol. 260, Pitman Research Notes in Mathematics, Longman Scientific and Technical, Essex, UK, 1992, pp. 234–266.
[23] L. N. TREFETHEN AND M. R. TRUMMER, *An instability phenomenon in spectral methods*, SIAM J. Numer. Anal., 24 (1987), pp. 1008–1023.
[24] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[25] E. WEGERT AND L. N. TREFETHEN, *From the Buffon needle problem to the Kreiss matrix theorem*, Amer. Math. Monthly, 101 (1994), pp. 132–139.