

Approximate Diagonalization

(version 20)

E B Davies

14 April 2006

1 Introduction

Let A be a non-self-adjoint $n \times n$ matrix and suppose that one wants to evaluate $x = f(A)b$ or solve $f(A)x = b$ for a large number of different analytic functions f rapidly, without caring too much about high accuracy. If A is diagonalizable, i.e. $A := SDS^{-1}$ where D is diagonal, then one can solve the first problem by writing $x := Sf(D)S^{-1}b$, where $f(D)$ is evaluated by applying the function f to the diagonal entries of D , which coincide with the eigenvalues of A . The second problem may be solved in a similar manner.

This procedure may not be appropriate if A is highly non-self-adjoint, because the eigenvalues of A can be highly unstable under small perturbations, such as those associated with rounding errors in computation, and the matrix S may have an extremely large condition number $\kappa(S) := \|S\| \|S^{-1}\|$. In the most extreme case, when A has a non-trivial Jordan form, the method breaks down entirely.

In this paper we describe an approach which involves using an approximate diagonalization of A . We emphasize that this does not mean that it is close to a true diagonalization, but rather that it has many of the features of a true diagonalization, and the amount of error associated with using it can be estimated. We start by describing the idea, and formulate a conjecture about its efficiency. Much of the remainder of the paper is devoted to providing theoretical and numerical evidence in support of the conjecture.

In Section 5 we use the ideas developed to throw some light on the difficulties of computing fractional powers of highly singular matrices.

2 Definitions

Throughout this paper we assume that f is an analytic function defined on a neighbourhood of $\text{Spec}(A)$ and that $f(A)$ is defined by means of the holomorphic

functional calculus as in [1, sect. 1.5]. In order to be able to define $f(A)$ stably for highly non-self-adjoint A one has to impose some conditions on the analytic function f . If $f(z) = (z-a)^{-1}$ where a is not close to $\text{Spec}(A)$, one may nevertheless have $a \in \text{Spec}(\tilde{A})$ where $\|A - \tilde{A}\|$ is very small. In such situations $f(A)$ cannot be defined stably for reasons discussed in [7].

We assume henceforth that γ is a simple closed curve with length $|\gamma|$ and that $\text{Spec}(A) \subseteq U$, where U is the region inside γ . We also assume that $f \in \mathcal{A}$, where \mathcal{A} is the space of functions that are analytic on $\gamma \cup U$. We write $\|f\|_\infty$ for the maximum value of $|f(z)|$ for $z \in \gamma$, or equivalently for $z \in (\gamma \cup U)$.

Lemma 1 *Suppose, in addition to the above assumptions, that the resolvent operators $R(z, A)$ satisfy $\|R(z, A)\| \leq c$ for all $z \in \gamma$ and that $f \in \mathcal{A}$. Then*

$$\|f(A)\| \leq \frac{c}{2\pi} |\gamma| \|f\|_\infty.$$

If $\|A - \tilde{A}\| \leq 1/(2c)$ and $\tilde{f} \in \mathcal{A}$ then

$$\|f(A) - \tilde{f}(\tilde{A})\| \leq c^2 \pi^{-1} |\gamma| \|f\|_\infty \|A - \tilde{A}\| + c \pi^{-1} |\gamma| \|f - \tilde{f}\|_\infty.$$

Proof. The first bound depends on a routine estimate of the formula

$$f(A) = \frac{1}{2\pi i} \int_\gamma f(z) R(z, A) dz.$$

We next assume that $z \in \gamma$ and use the bound

$$\begin{aligned} \|R(z, \tilde{A})\| &= \|R(z, A) (I - (A - \tilde{A})R(z, A))^{-1}\| \\ &\leq \frac{\|R(z, A)\|}{1 - \|(A - \tilde{A})R(z, A)\|} \\ &\leq 2c \end{aligned}$$

to derive

$$\begin{aligned} \|R(z, \tilde{A}) - R(z, A)\| &= \|R(z, \tilde{A})(\tilde{A} - A)R(z, A)\| \\ &\leq 2c^2 \|A - \tilde{A}\|. \end{aligned}$$

The second formula now follows by a routine estimate of the identity

$$\begin{aligned} f(A) - \tilde{f}(\tilde{A}) &= \frac{1}{2\pi i} \int_\gamma f(z) (R(z, A) - R(z, \tilde{A})) dz \\ &\quad + \frac{1}{2\pi i} \int_\gamma (f(z) - \tilde{f}(z)) R(z, \tilde{A}) dz. \end{aligned}$$

Example 2 There are two obvious ways of ensuring that the resolvent bound of Lemma 1 is satisfied. The first uses the stability of the numerical range $\text{Num}(A)$ under small perturbations. If $\text{Num}(A) \subseteq U$ and $\text{dist}(\gamma, \text{Num}(A)) \geq 1/c$ then the bound $\|R(z, A)\| \leq c$ is valid for all $z \in \gamma$ by [7, chap. 17] or [1, sect. 9.3].

Alternatively given $c > 0$ one may define γ to be the pseudospectral contour $\{z : \|R(z, A)\| = c\}$. The shape of the contour, which may have several components, can be determined numerically by using the Eigtool software; see [8].

We say that three matrices S, D, B provide an approximate diagonalization of A if D is diagonal, S is invertible, B is small and $A = SDS^{-1} + B$; we assume that $\|A\| \leq 1$ whenever necessary for reasons stated below. We say that S, B is a permitted pair for A if S is invertible and $D := S^{-1}(A - B)S$ is diagonal. The accuracy of the approximate diagonalization is measured by the quantity

$$\sigma(A, S, B, \varepsilon) := \kappa(S)\varepsilon + \|B\|$$

where $\varepsilon \in (0, 1)$ is a preassigned degree of accuracy of the computations, for example $\varepsilon := 10^{-16}$. The term $\kappa(S)\varepsilon$ measures errors associated with the condition number of S , and would vanish if the computations could have infinite precision, i.e. if $\varepsilon = 0$. The term $\|B\|$ represents the amount that A has been perturbed with the intention of reducing the first type of error. By adding the two errors and then minimizing over all permitted pairs, one obtains the smallest overall error that is possible when diagonalizing A approximately, namely

$$\underline{\sigma}(A, \varepsilon) := \inf_{B, S} \sigma(A, S, B, \varepsilon).$$

The non-zero entries of D are the eigenvalues of $A - B$, and are of order 1 in all the cases considered below. Many of our theorems below can be viewed as providing support for the

Conjecture For every positive integer n there exists c_n such that

$$\underline{\sigma}(A, \varepsilon) \leq c_n \varepsilon^{1/2}$$

for every $n \times n$ matrix A such that $\|A\| \leq 1$ and for every $\varepsilon \in (0, 1)$.

Since one can only evaluate $\underline{\sigma}(A, \varepsilon)$ exactly in simple cases, we attempt to obtain a high quality upper bound on it by choosing B, S appropriately. The rate of convergence of $\underline{\sigma}(A, \varepsilon)$ to 0 as $\varepsilon \rightarrow 0$ depends on whether A is diagonalizable or not. Note that one obtains an approximate diagonalization for another matrix \tilde{A} from that for A by keeping the same S, D and putting $\tilde{B} := B + (\tilde{A} - A)$. Therefore

$$|\underline{\sigma}(A, \varepsilon) - \underline{\sigma}(\tilde{A}, \varepsilon)| \leq \|A - \tilde{A}\|$$

and our definition is computationally stable. Further computational questions can be asked, for example about the errors arising when evaluating S^{-1} for a fairly singular choice of S , but the methods described here allow one to replace S^{-1} by T provided

$$\frac{\|T - S^{-1}\|}{\|S^{-1}\|} = O(\varepsilon).$$

We observe that

$$\underline{\sigma}(VAV^{-1}, \varepsilon) \leq \kappa(V)\underline{\sigma}(A, \varepsilon)$$

for all invertible matrices V ; thus the order of magnitude of $\underline{\sigma}(A, \varepsilon)$ is not changed if one passes from A to VAV^{-1} where $\kappa(V)$ is of order 1. If A is normal then one may diagonalize it exactly with S unitary and $B = 0$, so $\underline{\sigma}(A, \varepsilon) = \varepsilon$.

A feature of our definitions of σ and $\underline{\sigma}$ is that they do not scale under the map $A \rightarrow \lambda A$ when λ is large. As λ increases $\text{Spec}(\lambda A)$ and $\text{Num}(\lambda A)$ expand, so the contour γ and the algebra \mathcal{A} must be changed. We therefore impose the condition $\|A\| \leq 1$ whenever necessary.

The function $\underline{\sigma}(A, \varepsilon)$ is closely related to $\mu(A, \delta)$ defined for all $\delta > 0$ by

$$\mu(A, \delta) := \inf\{\kappa(S) : A = SDS^{-1} + B, \text{ where } D \text{ is diagonal and } \|B\| \leq \delta\}.$$

Lemma 3 *If $c > 0$, $\alpha > 0$ and $\mu(A, \delta) \leq c\delta^{-\alpha}$ for all $\delta > 0$ then*

$$\underline{\sigma}(A, \varepsilon) \leq 2(c\varepsilon)^{1/(\alpha+1)}.$$

Proof. Given A, S, D, B as above we have

$$\sigma(A, S, B, \varepsilon) \leq \mu(A, \delta)\varepsilon + \delta \leq c\varepsilon\delta^{-\alpha} + \delta$$

for all $\delta > 0$. The lemma follows by applying the following general fact: if f (resp. g) are non-negative, monotonically decreasing (resp. increasing) functions on (a, b) and $f(\xi) = g(\xi)$ for some $\xi \in (a, b)$ then

$$f(\xi) \leq \inf\{f(x) + g(x) : x \in (a, b)\} \leq 2f(\xi).$$

Theorem 4 *Suppose that $\|A\| \leq 1$ and that $f(z)$ is analytic on $\{z : |z| \leq r\}$ for some $r > 1$. If $\underline{\sigma}(A, \varepsilon) < (r - 1)/2$ then*

$$\underline{\sigma}(f(A), \varepsilon) < \underline{\sigma}(A, \varepsilon) \max\left\{1, \frac{2r\|f\|_{r,\infty}}{(r-1)^2}\right\}$$

where

$$\|f\|_{r,\infty} := \max\{|f(z)| : |z| \leq r\}.$$

Proof. If $\tilde{A} := SDS^{-1}$ then $\|A - \tilde{A}\| = \|B\| \leq \sigma(A, S, B, \varepsilon)$ and we can define \tilde{B} by

$$f(A) = f(\tilde{A}) + \tilde{B} = Sf(D)S^{-1} + \tilde{B}.$$

If $r > 1$ and γ is the circle $\{z : |z| = r\}$ then $\|R(z, A)\| \leq (r - 1)^{-1}$ for all $z \in \gamma$. Lemma 1 implies that if $\sigma(A, S, B, \varepsilon) < (r - 1)/2$ then

$$\|\tilde{B}\| \leq \|B\| \frac{2r\|f\|_{r,\infty}}{(r-1)^2}.$$

Therefore

$$\sigma(f(A), S, \tilde{B}, \varepsilon) = \kappa(S)\varepsilon + \|\tilde{B}\| \leq \sigma(A, S, B, \varepsilon) \max\left\{1, \frac{2r\|f\|_{r,\infty}}{(r-1)^2}\right\}.$$

The theorem now follows by taking the infimum over all permitted S, B .

Given a basis of pseudoeigenvectors $\{\phi_1, \dots, \phi_n\}$ of A one may be able to use the following theorem to construct an approximate diagonalization. In applications the vectors r_j should be small, so that $\|B\|$ is small. All vectors are regarded as column vectors.

Theorem 5 Let $\{\phi_1, \dots, \phi_n\}$ be a linearly independent set in \mathbf{C}^n such that

$$A\phi_j = \lambda_j\phi_j + r_j$$

for all $j \in \{1, \dots, n\}$, where $\lambda_j \in \mathbf{C}$ and $r_j \in \mathbf{C}^n$. Let R, S be the $n \times n$ matrices $R := [r_1 \dots r_n]$ and $S := [\phi_1 \dots \phi_n]$. Then $A = SDS^{-1} + B$, where $B = RS^{-1}$ and D is the diagonal matrix with entries $\lambda_1, \dots, \lambda_n$. Moreover

$$\|S\|^2 \leq \sum_{j=1}^n \|\phi_j\|^2$$

and

$$\|B\|^2 \leq \|S^{-1}\|^2 \sum_{j=1}^n \|r_j\|^2$$

Proof. If $\{e_1, \dots, e_n\}$ is the standard orthonormal basis in \mathbf{C}^n then

$$\begin{aligned} S^{-1}ASe_j &= S^{-1}A\phi_j \\ &= S^{-1}(\lambda_j\phi_j + r_j) \\ &= S^{-1}(\lambda_jSe_j + BSe_j) \\ &= \lambda_j e_j + S^{-1}BSe_j \\ &= (D + S^{-1}BS)e_j \end{aligned}$$

for all j . Therefore $S^{-1}AS = D + S^{-1}BS$; this is equivalent to $A = SDS^{-1} + B$.

If $v \in \mathbf{C}^n$ then

$$\begin{aligned} \|Sv\|^2 &= \left\| \sum_{j=1}^n v_j \phi_j \right\|^2 \\ &\leq \left(\sum_{j=1}^n \|\phi_j\|^2 \right) \left(\sum_{j=1}^n |v_j|^2 \right) \\ &= \left(\sum_{j=1}^n \|\phi_j\|^2 \right) \|v\|^2. \end{aligned}$$

This implies the stated bound on $\|S\|$. The bound on $\|B\|$ has a similar proof.

If A is highly non-self-adjoint, one should expect that the norm of S^{-1} will increase as one chooses ϕ_j for which the residuals r_j get smaller.

Our next result is a partial converse of Theorem 5, and is useful when $\|B\|$ is small and $\kappa(S)$ is not too big.

Theorem 6 Let $A = SDS^{-1} + B$ where D is diagonal with diagonal entries $\lambda_1, \dots, \lambda_n$, and let $\phi_j := Se_j$ for all j where $\{e_1, \dots, e_n\}$ is the standard basis of \mathbf{C}^n . Then

$$A\psi_j = \lambda_j\psi_j + r_j \tag{1}$$

for all j , where $\psi_j := \phi_j / \|\phi_j\|$ and $\|r_j\| \leq \|B\|$.

Proof. Putting $r_j := B\phi_j$, we have

$$\begin{aligned} A\phi_j &= SDe_j + B\phi_j \\ &= \lambda_j Se_j + B\phi_j \\ &= \lambda_j \phi_j + B\phi_j. \end{aligned}$$

We obtain (1) by putting $R_j := B\phi_j/\|\phi_j\|$. The bound on $\|r_j\|$ following immediately.

3 Evidence Supporting the Conjecture

We start by proving our Conjecture for Jordan matrices.

Lemma 7 *Let J denote the $n \times n$ Jordan matrix*

$$J_{r,s} := \begin{cases} 1 & \text{if } s = r + 1, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Then

$$0 \leq \mu(J, \delta) \leq \delta^{-1+1/n} \leq \delta^{-1} \quad (3)$$

for all $\delta \in (0, 1)$ and

$$0 \leq \underline{\sigma}(J, \varepsilon) \leq 2\varepsilon^{n/(2n-1)} \leq 2\varepsilon^{1/2} \quad (4)$$

for all $\varepsilon \in (0, 1)$.

Proof. We define B by

$$B_{r,s} := \begin{cases} -\delta & \text{if } r = n \text{ and } s = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and T by

$$T_{r,s} := \begin{cases} \delta^{-r/n} & \text{if } r = s, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

A direct calculation shows that

$$T(J - B)T^{-1} = \delta^{1/n}U$$

where U is the circulant and unitary matrix with entries

$$U_{r,s} := \begin{cases} 1 & \text{if } s = r + 1, \\ 1 & \text{if } r = n \text{ and } s = 1, \\ 0 & \text{otherwise.} \end{cases}$$

If F is the finite Fourier transform, which has a unitary matrix, then

$$FUF^{-1} = D$$

where D is the diagonal matrix with entries $\lambda_r = e^{2\pi ir/n}$, for $1 \leq r \leq n$. Putting $S := FT$ we finally obtain

$$J = S^{-1}DS + B.$$

Since $\|T\| = \delta^{-1}$ and $\|T^{-1}\| = \delta^{1/n}$ we deduce that $\kappa(S) = \delta^{-1+1/n}$. This implies (3). The corresponding upper bound on $\underline{\sigma}$ is obtained by applying Lemma 3.

Example 8 We compare the above theoretical result with what can be obtained numerically. We defined J by (2) with $n = 25$ and evaluated $f(\delta) := \delta^{1-1/n}\kappa(S)$ for 200 randomly generated matrices B with norms equal to δ for a range of values of δ . The matrices S and D were defined by using the Matlab command `[S,D]=eig(A-B)`. In Table 1, $\min(f(\delta))$ is the minimum value of $f(\delta)$ obtained and $\text{med}(f(\delta))$ is the median value. We also took a sample of 2000 such matrices B and found that all the values of $\min(f(\delta))$ remained larger than 2. The similarity of the numerical results to what was proved in Lemma 7 suggests that both are close to the optimal bound.

δ	$\min(f(\delta))$	$\text{med}(f(\delta))$
10^{-1}	3.67	17.91
10^{-2}	4.15	22.09
10^{-3}	3.29	22.25
10^{-4}	3.78	25.57
10^{-5}	3.88	25.95
10^{-6}	3.32	24.04
10^{-7}	3.83	20.51
10^{-8}	3.72	24.99

Table 1 Computation of condition numbers in Example 8

The following corollary does not prove the Conjecture because the constant obtained depends on the matrix involved, and not just on the dimension. It is known that finding the Jordan canonical form is an inherently unstable problem [4, p. 390], [5].

Corollary 9 *For every $n \times n$ matrix A there exists a constant c_A such that*

$$\underline{\sigma}(A, \varepsilon) \leq c_A \varepsilon^{1/2}$$

for all $\varepsilon > 0$.

Proof. By Lemma 3 it is sufficient to prove that for every $\delta > 0$ there exists an approximate diagonalization $A = SDS^{-1} + B$ with $\|B\| < \delta$ and $\kappa(S) < c_A \delta^{-1}$. This statement is invariant under similarity transformations (although the constant c changes) so it is sufficient to prove it when A is written in the Jordan canonical form. We can then deal with each Jordan block using the methods of Lemma 7.

We next prove the Conjecture for triangular Toeplitz matrices.

Theorem 10 *Given complex constants $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ such that $\sum_{r=0}^{n-1} |\alpha_r| \leq 1$, let A denote the $n \times n$ triangular Toeplitz matrix*

$$A_{r,s} := \begin{cases} \alpha_{s-r} & \text{if } s \geq r, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\underline{\sigma}(A, \varepsilon) \leq 2\varepsilon^{n/(2n-1)} \leq 2\varepsilon^{1/2}$$

for all $\varepsilon \in (0, 1)$.

Proof. If we define B by

$$B_{r,s} := \begin{cases} -\delta\alpha_{s-r+n} & \text{if } s < r, \\ 0 & \text{otherwise,} \end{cases}$$

where $\delta \in (0, 1)$ is to be determined, then $\|B\| \leq \delta$. If we define T by (5) then a direct calculation shows that

$$C := T(A - B)T^{-1}$$

is the circulant matrix with entries

$$C_{r,s} := \alpha_{s-r}\delta^{(s-r)/n}$$

where we replace $s - r$ by $s - r + n$ if the former expression is negative. If F is the finite Fourier transform then

$$D := FCF^{-1}$$

is a diagonal matrix. Putting $S := FT$ as before we obtain $\kappa(S) = \delta^{-1+1/n}$ and

$$A = S^{-1}DS + B.$$

Putting $\delta := \varepsilon^{n/(2n-1)}$ we obtain

$$\underline{\sigma}(A, \varepsilon) < 2\varepsilon^{n/(2n-1)}.$$

We have not been able to prove the Conjecture for general $n \times n$ matrices, and Theorem 15 below is the closest that we have got to it. We originally proved it under the assumption that the eigenvalues of A were collinear. The general case depends on the following theorem of Friedland [3]. Its proof depends on using the degree mod 2 of a smooth map between manifolds of equal dimension, and it would be valuable to obtain a constructive version. This may not be easy, because the number of normal ‘extensions’ N of Q varies from 1 to ∞ (inclusive) depending on Q .

Theorem 11 (Friedland) *For every upper triangular $n \times n$ matrix Q there exists a strictly lower triangular matrix L such that $N := Q + L$ is normal.*

Example 12 A direct construction of the matrix L in the theorem is not elementary even in the case of 3×3 matrices. If the eigenvalues of the upper triangular matrix Q are collinear, then we may construct L as follows. We first write Q in the form $Q = cI + e^{i\theta}(D + U)$, where $c \in \mathbf{C}$, $\theta \in \mathbf{R}$, D is a real diagonal matrix and U is strictly upper triangular. If we define $L := e^{i\theta}U^*$ then $Q + L = cI + e^{i\theta}H$ where H is self-adjoint, and this implies that $Q + L$ is normal.

Lemma 13 *If $N := Q + L$ is normal where Q and L are upper triangular and strictly lower triangular respectively, then $\nu(Q) = \nu(L)$ where*

$$\nu(A) := \sum_{r,s} |r - s| |A_{r,s}|^2.$$

Proof. We have

$$\begin{aligned} \nu(Q) - \nu(L) &= \sum_{r,s} (s - r) |N_{s,r}|^2 \\ &= \operatorname{tr}[N^*EN] - \operatorname{tr}[N^*NE] \\ &= \operatorname{tr}[(NN^* - N^*N)E] \\ &= 0 \end{aligned}$$

where

$$E_{r,s} := \begin{cases} r & \text{if } r = s, \\ 0 & \text{otherwise.} \end{cases}$$

Further comparisons between the size of Q and L can be obtained by replacing E by

$$E_{r,s} := \begin{cases} f(r) & \text{if } r = s, \\ 0 & \text{otherwise,} \end{cases}$$

where f is any monotonic function on $\{1, \dots, n\}$.

Lemma 14 *The inequality $\nu(A) \leq n^2 \|A\|^2$ holds for all $n \times n$ matrices A . If the diagonal entries of A all vanish then $\|A\|^2 \leq \nu(A)$.*

Proof. We have

$$\nu(A) \leq n \sum_{r,s} |A_{r,s}|^2 = n \sum_{s=1}^n \|Ae_s\|^2 \leq n^2 \|A\|^2$$

where $\{e_s\}_{s=1}^n$ is the standard basis of C^n .

If the diagonal entries of A vanish then the second inequality follows from

$$\|A\|^2 \leq \|A\|_{HS}^2 \leq \nu(A)$$

where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm.

Theorem 15 *If A is an $n \times n$ matrix A satisfying $\|A\| \leq 1$ then*

$$\underline{\sigma}(A, \varepsilon) \leq (1 + n)\varepsilon^{2/(n+1)}$$

for all $\varepsilon \in (0, 1)$. In particular the Conjecture holds with

$$\underline{\sigma}(A, \varepsilon) < 4\varepsilon^{1/2}$$

for every 3×3 matrix A satisfying $\|A\| \leq 1$.

Proof. By Schur's Lemma there exists a unitary matrix U such that $P := U^{-1}AU$ is upper triangular. If $0 < \delta < 1$ and

$$V_{r,s} := \begin{cases} \delta^r & \text{if } r = s, \\ 0 & \text{otherwise,} \end{cases}$$

then $Q := V^{-1}PV$ is upper triangular and

$$\nu(Q) \leq \delta^2 \nu(P) \leq \delta^2 n^2 \|P\|^2 \leq \delta^2 n^2.$$

By Friedland's theorem there exists a strictly lower triangular matrix L such that $Q + L$ is normal and $\nu(L) = \nu(Q)$. A direct calculation establishes that

$$\|VLV^{-1}\|^2 \leq \nu(VLV^{-1}) \leq \delta^2 \nu(L) = \delta^2 \nu(Q) \leq \delta^4 n^2.$$

Therefore $B := -UVLV^{-1}U^{-1}$ satisfies $\|B\| \leq \delta^2 n$. Hence

$$V^{-1}U^{-1}(A - B)UV = Q + L = WDW^{-1}$$

where D is diagonal and W is unitary. Putting $S := UVW$ we obtain $A = SDS^{-1} + B$ where $\kappa(S) = \kappa(V) = \delta^{1-n}$. Therefore

$$\underline{\sigma}(A, \varepsilon) \leq \delta^{1-n} \varepsilon + n\delta^2.$$

The result now follows by putting $\delta := \varepsilon^{1/(n+1)}$.

4 Random Perturbations

The above methods of constructing B and S are too simple to prove the Conjecture for $n > 3$. In this section we describe a randomized approximate diagonalization method (RADM), suggested to us by L N Trefethen, which provides numerical evidence in support of the Conjecture. Numerically it is remarkably effective.

If the $n \times n$ matrix A cannot be diagonalized or can only be diagonalized by means of a matrix S whose condition number is extremely large, then one can instead diagonalize the matrix $A - B$ where B is a small random perturbation. We found experimentally that for a variety of strictly upper triangular $n \times n$ matrices A (none of which can be diagonalized) with $n = 100$ and $\varepsilon := 10^{-16}$ one has

$\sigma(A, \varepsilon) \leq 3 \times 10^{-7}$. In each case we minimized over 100 randomly chosen B such that $\|B\| = 10^{-8}$. On the other hand for a series of 100 matrices of the form $A = \text{rand}(n)$ with $n = 100$ and $B = 0$ we found that $50 \leq \kappa(S) \leq 1000$ in every case; our methods are not necessary for such matrices. In the computations below the random perturbation was of the form $B = s * \text{randn}(n)$ where s is a small constant. However, we got the same results with small random perturbations $B = s * \text{randn}(n, 1) * \text{randn}(1, n)$ of rank one.

One can use RADM to evaluate $f(A)$. Our conclusions from a range of such problems, some described below, is that RADM is less accurate than standard Matlab algorithms when the latter exist and the condition number of the matrix is small. For very singular matrices the two methods have comparable accuracy. For many functions one cannot apply Matlab's `funm` algorithm, described in [2], but RADM still yields a result whose accuracy can be confirmed by repeating the computation with another choice of the random perturbation.

Example 16 We consider the $n \times n$ matrix

$$A_{r,s} := \begin{cases} r/n & \text{if } s = r + 1, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

We put $n := 100$ and defined

```
B1=randn(n)
B=10^(-r)*B1/norm(B1)
[S,D] = eig(A-B)
```

in the notation of Matlab. We computed $\sigma(A, S, B, \varepsilon)$ and $\log_{10}(\kappa(S))$ for $\varepsilon = 10^{-16}$ and $1 \leq r \leq 15$. Table 1 shows that $\sigma(A, S, B, \varepsilon)$ took its minimum value for $\|B\| \sim 10^{-7}$ but that the condition number of S increased steadily as r increases. The minimum value of σ is of order $\varepsilon^{1/2}$.

We also carried out a computation in which the entries r/n of A were replaced by randomly chosen numbers. The conclusions were similar.

r	$\sigma(A, S, B, \varepsilon)$	$\log_{10}(\kappa(S))$
1	0.1	2.2784
2	0.01	3.723
3	0.001	4.3007
4	0.0001	5.3355
5	$1e - 005$	6.169
6	$1.0016e - 006$	7.1996
7	$1.2316e - 007$	8.3647
8	$2.0592e - 007$	9.2921
9	$1.7551e - 006$	10.244
10	$2.0103e - 005$	11.303
11	0.00015367	12.187
12	0.0015981	13.204
13	0.019479	14.29
14	0.19699	15.294
15	1.7837	16.251

Table 2 Computation of condition numbers in Example 16

5 Fractional Powers

The definition of the square root of an $n \times n$ matrix A is not as straightforward as it appears. If A has n distinct non-zero eigenvalues then it has exactly 2^n square roots, which commute pairwise. On the other hand the matrices 0 and 1 have a continuum of non-commuting square roots. If $A^n = 0$ but $A^{n-1} \neq 0$ then A has no square root, but A^2 has a continuum of commuting square roots, namely $A + cA^{n-1}$ for any choice of c . If A has n distinct non-zero eigenvalues but two (or more) of these are equal to machine precision then it may have a large number of pairwise non-commuting approximate square roots. One may avoid these ambiguities by using the holomorphic functional calculus to define $A^{1/2}$, and choosing the branch of $z^{1/2}$ that has a cut along the negative real axis.

Example 17 Let A be the $n \times n$ matrix

$$(A)_{r,s} := \begin{cases} r/n & \text{if } s = r + 1, \\ c & \text{if } s = r, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where $0 < c < 1$. We computed $A^{1/2}$ for various values of c when $n := 20$ by two methods, using the program listed in the Appendix, and present the results in Table 3. We compared $\|E^2 - A\|$, where E is the square root computed by using RADM, with $\|F^2 - A\|$, where $F = \text{sqrtm}(A)$ in the notation of Matlab.

If one only looks at the third column of Table 3 one sees that the algorithm `sqrtm` is not accurate for $c < 0.1$. The situation for RADM is not as straightforward.

Column 2 wrongly suggests that RADM is reasonably accurate for all values of c investigated. However, Column 4 shows that $\|E\|$ increases rapidly as c decreases, while Column 5 presents the ratio of two values of $\|E\|$ computed with two different random perturbations B_i , both satisfying $\|B_i\| = 10^{-8}$. One sees that the two values of $\|E\|$ obtained are quite different if $c < 0.2$, even though $\|E^2 - A\| = O(10^{-8})$ in both cases. This is possible because the map $A \rightarrow A^{1/2}$ varies very rapidly for small values of c in spite of Lemma 1. If one only wants an approximate square root of A then RADM works well for all c investigated, but it does not produce a good approximation to the exact square root of A for $c < 0.2$.

Our intention above was to illuminate the problems involved in computing square roots rather than to advocate the use of a particular method, but if one wishes to use RADM it is recommended that one should check that two successive applications with different random perturbations yield the same answer to within $O(10^{-8})$.

c	$\ E^2 - A\ $	$\ F^2 - A\ $	$\ E\ $	$\ E_1\ /\ E_2\ $
0.8	$3.0453e - 008$	$1.8228e - 016$	1.2743	1
0.6	$5.5242e - 008$	$1.9563e - 016$	1.1984	1
0.4	$1.5797e - 008$	$5.6362e - 016$	5.3628	1
0.3	$6.2193e - 008$	$1.3148e - 014$	103.7	1.0001
0.2	$2.6138e - 008$	$3.2596e - 012$	$2.0706e + 004$	0.97515
0.1	$2.0012e - 008$	$6.3565e - 008$	$3.3292e + 006$	0.50668
0.05	$2.8685e - 008$	0.0074873	$3.0604e + 006$	0.55402
0.02	$2.2364e - 008$	93768	$9.1859e + 006$	0.71102
0.01	$1.453e - 008$	$5.5561e + 008$	$2.2661e + 007$	0.89547

Table 3 Computation of square roots in Example 17

Let A be an $n \times n$ matrix whose numerical range is contained in $\{z : \operatorname{Re}(z) \geq 0\}$ and contains some points very close to 0. Suppose that one wishes to compute A^t for $0 \leq t \leq 1$. The formula

$$A^t = e^{t \log(A)}$$

is not recommended because $\log(A)$ may have a very large norm, and it is undefined if 0 is an eigenvalue of A . An accuracy of 10^{-8} is more than sufficient for plotting the graph of $f(t) := \|A^t\|$, and RADM provides a way of doing this with a minimum of effort. Many other applications of a similar character can easily be devised.

There are two other possible methods of evaluating A^t . If $A = I - B$ where $\|B\| \leq 1$ then

$$\operatorname{Spec}(A) \subseteq \operatorname{Num}(A) \subseteq \{z : \operatorname{Re}(z) \geq 0\}.$$

If $s > 0$ then one may define A^s by

$$A^s := \frac{1}{2\pi i} \int_{\gamma} z^s (zI - A)^{-1} dz$$

where γ is the boundary of the region

$$\{re^{i\theta} : 0 < r < 5/2 \text{ and } -3\pi/2 < \theta < 3\pi/2\}.$$

The integral is norm convergent for all $s > 0$, but it may develop a singularity at $z = 0$ as $s \rightarrow 0+$, so it is not always useful for small s .

Alternatively one might use the expansion

$$A^s = I - \sum_{r=1}^{\infty} c_{r,s} B^r$$

where

$$c_{r,s} := (-1)^{r+1} s(s-1)\dots(s-r+1)/r!$$

The series is norm convergent for all $s \in (0, 1)$ because $c_{r,s} \geq 0$ and $\sum_{r=1}^{\infty} c_{r,s} = 1$. However, the convergence of the series is very slow for small $s > 0$, so it is not numerically useful for such s . Both of these problems are apparent if 1 is an eigenvalue of B , but they also occur if the pseudospectra of B are significant near 1, even if B has no spectrum near 1.

Example 18 We used RADM to compute the r th root C_r of the matrix (7), with $c := 0.5$, $n = 20$ and $r = 1, \dots, 10$, using a small modification of the program in the Appendix. Other real powers may be treated in exactly the same way. In the final column of Table 3, $C_{r,1}$ and $C_{r,2}$ are two independent computations of C_r both obtained using RADM. The small size of the entries in this column indicate that the results are all reliable to $O(10^{-8})$.

r	$\ (C_r)^r - A\ $	$\ C_r\ $	$\ C_{r,1}\ /\ C_{r,2}\ - 1$
1	$3.4677e - 007$	1.33558842	$-2.1906e - 009$
2	$1.4319e - 007$	1.35606917	$8.9048e - 009$
3	$3.8861e - 008$	1.57711707	$4.0503e - 008$
4	$1.0367e - 007$	1.59766857	$4.9686e - 008$
5	$7.8846e - 008$	1.55852362	$8.1706e - 008$
6	$4.6441e - 008$	1.50671663	$2.3817e - 009$
7	$8.0153e - 008$	1.45657560	$5.0527e - 008$
8	$7.5740e - 008$	1.41197706	$-4.5141e - 008$
9	$2.2498e - 007$	1.37341090	$5.3190e - 008$
10	$7.8768e - 008$	1.34032556	$1.1992e - 007$

Table 4 Computation of r th roots in Example 18

We finally remark that if greater accuracy is needed, then one may use the above procedure to obtain the starting point for a Newton type iteration.

We also used RADM to compute $\|A^t\|$ for the matrix (7) with $n = 100$ and $c = 0.6$. We put $t := 2^{-7}r$ where r is a positive integer, $v_1 := \|A^t\|$ computed using

RADM, and $v_2 := \|B^r\|$ where $B := A^{1/128}$ is computed by repeated applications of Matlab's `sqrtm` operator. The two methods give the same answer to within 0.04 for all $t \in (0, 2)$, i.e. a relative accuracy of 10^{-4} . This may seem rather low, but it is more than enough for graph-drawing needs. Both methods computed the norm of A and A^2 correctly. Most of the CPU time was used computing the matrix norms, but excluding that RADM is substantially faster because it involves one application of `eig` as opposed to seven applications of `sqrtm`. For values of c much smaller than 0.6 the comparison cannot be made because `sqrtm` is not reliable. One might also compute B directly using the new algorithm of Guo and Higham [6]. Figure 1 shows the graph of the norm, and is typical of problems in which pseudospectral behaviour is important.

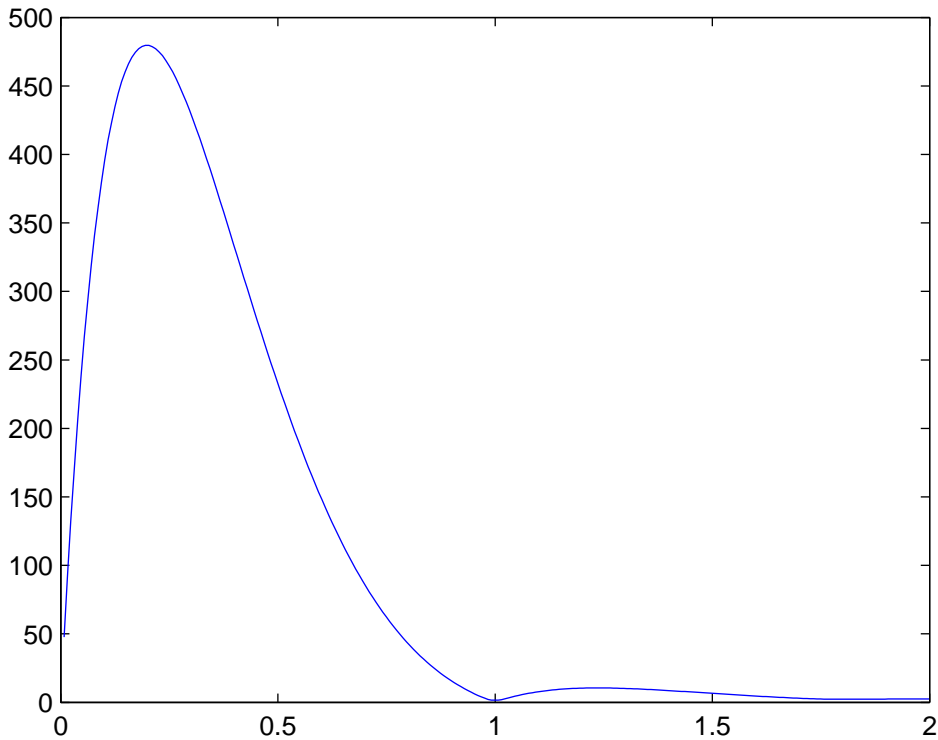


Figure 1: Graph of $\|A^t\|$ for $0 < t < 2$

6 Appendix

The following program was used to compute the square roots in Table 3.

```
n=20; % size of the matrix
X=[1:n-1]'/n;
M=9; % number of matrices considered;
Ans=zeros(M,5); % output data
c=zeros(8,1);
c(1,1)=0.8;
c(2,1)=0.6;
c(3,1)=0.4;
c(4,1)=0.3;
c(5,1)=0.2;
c(6,1)=0.1;
c(7,1)=0.05;
c(8,1)=0.02;
c(9,1)=0.01;
Ans(:,1)=c(:,1);

for r=1:M
    A=c(r,1)*eye(n)+diag(X,1); % construct the matrix to be analyzed
    B0=randn(n);
    B=(1e-8)*B0/norm(B0); %construct a small random perturbation
    [S,D]=eig(A-B);
    E1=S*sqrt(D)/S; % use RADM to compute E1=A^(1/2)
    Ans(r,2)=norm(E1^2-A); % estimate the error
    F=sqrtm(A); % use sqrtm to compute F=A^(1/2)
    Ans(r,3)=norm(F^2-A); % estimate the error
    Ans(r,4)=norm(E1);
    B0=randn(n); % repeat the computation
    B=(1e-8)*B0/norm(B0); %construct a small random perturbation
    [S,D]=eig(A-B);
    E2=S*sqrt(D)/S; % use RADM to compute E2=A^(1/2)
    Ans(r,5)=Ans(r,4)/norm(E2);
end;

format short g ;
Ans
```

Acknowledgements We should like to thank N J Higham, S Shkarin, E Shar-gorodsky and L N Trefethen for helpful criticism and encouragement. We also acknowledge support under EPSRC grant GR/R81756.

References

- [1] Davies E B: Linear Operators and Their Spectra. Camb. Univ. Press, 2007, to appear.
- [2] Davies P I and Higham N J: A Schur-Parlett algorithm for computing matrix functions. SIAM J. Matrix Anal. Appl. 25 (2003) 464-485.
- [3] Friedland S: Normal matrices and the completion problem. SIAM J. Matrix Anal. Appl. 23 (2002) 896-902.
- [4] Golub G, Van Loan C: Matrix Computations, 2nd ed. Johns Hopkins Univ. Press, Baltimore, 1989.
- [5] Gu M: Finding well-conditioned similarities to block-diagonalize nonsymmetric matrices is NP-hard. J. Complexity 11 (1995) 377-391.
- [6] Guo C-H, Higham N J: A Schur-Newton method for the matrix p th root and its inverse. Preprint, 2005.
- [7] Trefethen L N, Embree M: Spectra and Pseudospectra. Princeton Univ. Press, Princeton, 2005.
- [8] Wright T G: EigTool software available at <http://www.comlab.ox.ac.uk/pseudospectra/eigtool>

Department of Mathematics
King's College
Strand
London WC2R 2LS