Approximation theory and numerical linear algebra

L. N. Trefethen

Department of Mathematics,
Massachusetts Institute of
Technology,
Cambridge, Massachusetts, USA

Abstract Many large matrix problems are best solved by iteration, and the convergence of matrix iterations is usually connected with classical questions of polynomial or rational approximation. This paper discusses three examples: (1) the conjugate gradient iteration for symmetric positive definite matrices, (2) Strang's preconditioned conjugate gradient iteration for Toeplitz matrices, and (3) polynomial iterations for nonsymmetric matrices with complex eigenvalues. This last example raises a fundamental problem: what happens to iterative methods in linear algebra, and to the role of approximation theory, when the matrices become highly non-normal? Part of the solution may lie in approximation on a set of "approximate eigenvalues". *

Key words: approximation, linear algebra, conjugate gradient iteration, Toeplitz, Hankel, normal, resolvent

1. Introduction

Many large-scale numerical computations require the solution of large systems of equations $Ax = b$, and very often, the best known methods are iterative. This has become especially true since 1970, thanks to the development of multigrid and preconditioned conjugate gradient methods. It is impossible to know what the state of the art may be a generation from now, but for the present, iterative methods of linear algebra occupy a central position in scientific computing.

The purpose of this paper is to explore some connections, mostly already known, between matrix iterative methods and classical problems of polynomial and rational approximation. Since the 1950s, experts in numerical linear algebra have made use of approximation ideas to analyze rates of convergence [16, 27, 57, 58, 60]. The arguments are natural, often elegant, and deserve to be more widely appreciated among approximation theorists. We shall consider three examples:

1. Conjugate gradient iteration. The conjugate gradient iteration converges rapidly if A is a symmetric positive definite matrix with clustered eigenvalues or a low condition number. This phenomenon can be explained by considering an associated problem of polynomial approximation at the eigenvalues of A. The explanation is a beautiful one, one of the gems of modern numerical analysis.

2. Strang's preconditioned Toeplitz iteration. Recently Strang has proposed a method for solving symmetric positive definite Toeplitz matrix problems by a conjugate gradient iteration with a circulant preconditioner. The iteration often converges extremely fast, and the explanation turns out to be a surprisingly deep connection with a certain problem of approximation by rational functions on the complex unit disk. This example is much more specialized than the last one, but is offered for its novelty and nontriviality.

3. Polynomial iterations for nonsymmetric matrices. A less well understood problem is what to do with nonsymmetric matrices with complex eigenvalues. One approach is to construct an iteration based explicitly on a problem of polynomial approximation at these eigenvalues, assuming that something is known about their location. We describe here a particular version of this idea, recently explored by Fischer and Reichel and by Tal-Ezer, in which the approximation problem is solved by interpolation in Fejér points determined by means of conformal mapping.

Approximation theory has many other links with numerical linear algebra besides these. For example:

- The Chebyshev iteration for a symmetric positive definite matrix is connected with polynomial approximation on an interval $[a, b]$ containing the spectrum [21].
- The same is true of the design of "polynomial preconditioners" for accelerating conjugate gradient iterations [30].
- For nonsymmetric matrices, the Chebyshev iteration is connected with polynomial approximation on ellipses in the complex plane [15, 34, 35].
- The iterative solution of symmetric indefinite systems is connected with simultaneous polynomial approximation on two disjoint real intervals $[a, b]$, $[c, d]$ with $b < 0 < c$ [12, 43].
- The alternating direction implicit (ADI) iteration is connected with a problem of rational approximation on a real interval [58].
- Various connections have been described between the Lanczos iteration, continued fractions, and Padé approximation [4, 9, 22]. The history of the LR and QR algorithms is also linked with Padé approximation [22].
- The convergence of multigrid iterations is connected with problems of trigonometric approximation in Fourier space [3, 26].

The list could go on. These citations are by no means comprehensive; references

*subsequently renamed pseudo-eigenvalues

Bibliographic citation: see last page

to earlier work can be found therein.

In the final section we discuss an important general problem which is an outgrowth of topic *(3)*: what is the proper way to treat matrices that are not normal — that is, whose eigenvectors are not orthogonal? In such cases there is a gap of size $\kappa(V)$ — the condition number of the matrix of eigenvectors — between approximation at the eigenvalues and convergence of a matrix iteration, and as we show by several examples, in practical problems $\kappa(V)$ can be enormous. We propose that in these cases, approximations should be designed to be accurate on a set of "approximate eigenvalues" rather than just on the set of exact eigenvalues.

2. Conjugate gradient iteration

Let A be a real symmetric positive definite matrix of dimension n, let b be a real n-vector, and let ϕ be the quadratic form

$$\phi(x) = \tfrac{1}{2}x^TA\,x - x^Tb, \qquad x \in \mathbb{R}^n. \tag{1}$$

It is readily calculated that the gradient of $\phi(x)$ is $\nabla\phi(x) = Ax - b$, and therefore, $Ax = b$ is satisfied if and only if x is a stationary point of ϕ. Since A is positive definite, the only stationary point is the global minimum, and we are left with the following equivalence:

$$Ax = b \quad \Longleftrightarrow \quad \phi(x) = \inf_{y\in\mathbb{R}^n} \phi(y). \tag{2}$$

We shall denote this solution vector x by x^*.

The conjugate gradient iteration amounts to an iterative minimization of $\phi(x)$ based on a cleverly chosen sequence of search directions [11, 20, 29, 33, 48]. For each k, let K_k denote the *Krylov subspace* of $\mathbb{R}^n$ spanned by the *Krylov vectors* $b, Ab, \dots, A^{k-1}b$:

$$K_k \;=\; \langle b, Ab, \dots, A^{k-1}b\rangle \;=\; \langle Ax^*, A^2x^*, \dots, A^kx^*\rangle, \tag{3}$$

and let x_k be the unique minimizer of ϕ in this subspace:

$$\phi(x_k) = \inf_{y\in K_k} \phi(y) \qquad (x_0 = 0). \tag{4}$$

If $\|\cdot\|_A$ denotes the norm

$$\|y\|_A = \sqrt{y^TA\,y}, \tag{5}$$

and e_k denotes the error in x_k,

$$e_k = x^* - x_k \qquad (e_0 = x^*), \tag{6}$$

then an easy calculation shows that $\|e_k\|_A$ is also minimized at each step:

$$\|e_k\|_A = \inf_{e\in x^*-K_k} \|e\|_A. \tag{7}$$

Since $K_1 \subseteq K_2 \subseteq \cdots \subseteq \mathbb{R}^n$, the values of $\phi(x_k)$ and $\|e_k\|_A$ must decrease monotonically:

$$\phi(x_1) \geq \phi(x_2) \geq \cdots \geq \phi(x^*), \qquad \|e_1\|_A \geq \|e_2\|_A \geq \cdots \geq 0, \tag{8}$$

and the iteration must converge in at most n steps to some limit (in the absence of rounding errors). Since A^{-1} is equal to some polynomial in A (by the Cayley-Hamilton theorem), the limit vector x_k must be x^*.

So far, this idea of optimization in nested subspaces is a general one that may not seem a particularly promising basis for an algorithm. Yet the algorithm turns out to be excellent because of two remarkable properties, both related to the fact that the family of nested subspaces we have chosen is the Krylov sequence. The first is that the optimal vectors $\{x_k\}$ can be computed speedily by the following simple iteration:

```
Conjugate gradient (CG) iteration

$x_0 := 0,\ r_0 := b,\ \beta_0 := 0,\ p_0 := 0$
For $k := 1, 2, \dots$
    $p_k := r_{k-1} + \beta_{k-1}p_{k-1}$                 (search direction)
    $\alpha_k := r_{k-1}^Tr_{k-1}/p_k^TA\,p_k$              (distance along search direction)
    $x_k := x_{k-1} + \alpha_k p_k$                       (approximate solution vector)
    $r_k := r_{k-1} - \alpha_kA\,p_k$                     (residual $b - Ax_k$)
    $\beta_k := r_k^Tr_k/r_{k-1}^Tr_{k-1}$
```

(We shall not reproduce the derivation; see [20].) Since each step involves just $\sim n^2$ floating-point operations (one matrix-vector multiplication $A\,p_k$), the solution $x_n = x^*$ of $Ax = b$ can in principle be found in $\sim n^3$ operations, or potentially less if A is sparse.

The second remarkable property of the conjugate gradient iteration is that for many matrices A, there is no need to take so many steps: x_k may approximate x^* to machine precision for $k \ll n$. For example, it is not unusual for a matrix problem of dimension 10,000 to be solved to the required precision in 50 or 100 steps. Thus the CG iteration often takes closer to $O(n^2)$ than $O(n^3)$ operations, or even less if A is sparse.[1] It is this phenomenon that we wish to explain by means of polynomial approximation. The underlying ideas go back to the beginnings of the

[1] The mathematical fact that the CG iteration converges rapidly for well-conditioned matrices was known from the beginning, but for twenty years its practical importance was not fully appreci-

subject, and a concise presentation can be found in the book by Luenberger [33]; see also [10, 11, 16, 31].

Let P_k denote the set of polynomials of degree at most k. By (3) and (6), we have

$$x_k = q(A)b, \qquad e_k = p(A)e_0 \tag{9}$$

for some polynomials

$$q \in P_{k-1}, \qquad p \in P_k, \quad p(0) = 1, \tag{10}$$

with $p(z) = 1 - zq(z)$. By (4) and (7), these polynomials are optimal in the following senses:

$$\begin{aligned} \phi(x_k) &= \inf_{q \in P_{k-1}} \phi(q(A)b), \\ \|e_k\|_A &= \inf_{p \in P_k,\, p(0)=1} \|p(A)e_0\|_A. \end{aligned} \tag{11}$$

To put it in words, the polynomial $q(A)$ behaves as much as possible like A^{-1}, as measured by the function ϕ, and $p(A)$ behaves as much as possible like the zero matrix, subject to the constraint $p(0) = 1$, as measured by the A-norm of the error.

With the use of an eigenvector expansion, these statements about matrices become statements about scalars. Let $\{v_j\}$ be an orthonormal set of eigenvectors of A corresponding to eigenvalues $\{\lambda_j\}$, and write

$$e_0 = \sum_{j=1}^{n} a_j v_j, \qquad \|e_0\|_A^2 = \sum_{j=1}^{n} a_j^2 \lambda_j. \tag{12}$$

Then

$$p(A)e_0 = \sum_{j=1}^{n} a_j p(\lambda_j) v_j,$$

which implies

$$\|p(A)e_0\|_A^2 = \sum_{j=1}^{n} a_j^2 \lambda_j p^2(\lambda_j). \tag{13}$$

From (9)–(13) we now obtain the following fundamental theorem on convergence of the conjugate gradient iteration. This bound is independent of the dimension n, and indeed generalizes to positive definite self-adjoint operators on a Hilbert space [10, 11].

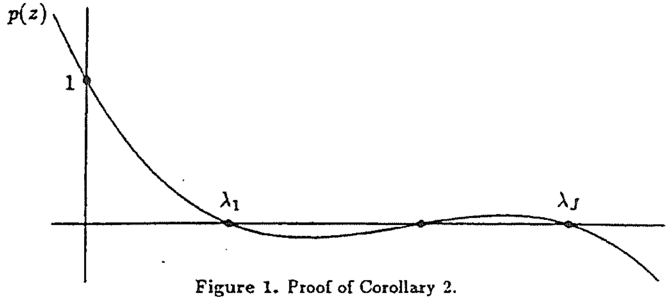

Figure 1. Proof of Corollary 2.

Theorem 1. The CG iteration satisfies

$$\frac{\|e_k\|_A}{\|e_0\|_A} \le \inf_{p \in P_k,\, p(0)=1} \left\{ \sup_{\lambda \in \Sigma} |p(\lambda)| \right\}, \tag{14}$$

where Σ denotes the spectrum of A. □

This theorem associates the CG iteration with the following approximation problem: what is the minimal magnitude that a polynomial $p \in P_k$ can attain on the spectrum of A, subject to the constraint $p(0) = 1$? It is not quite correct to say that the CG iteration implicitly finds the exact solution to this approximation problem, since after all, the exact polynomial $p(A)$ implicit in the iteration depends on e_0, while the approximation problem does not. What is correct is that the error reduction after k steps is guaranteed to be at least this good, and conversely, there are right-hand sides for which it is no better.

Depending on Σ, various consequences can be derived from Theorem 1. One extreme occurs if A happens to have a small number of distinct eigenvalues:

Corollary 2. Suppose A has J distinct eigenvalues $\{\lambda_j\}$. Then the CG iteration converges in at most J steps.

Proof Consider the polynomial $p(z) = \prod_{j=1}^{J}(1 - z/\lambda_j)$ (Figure 1). □

Another well-known corollary is obtained if one assumes merely that the condition number of A, defined by $\kappa = \lambda_{\max}/\lambda_{\min}$, is not too large [10]:[2]

Corollary 3. Suppose A has condition number κ. Then the CG iteration converges at least geometrically, as follows:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \qquad (0 \le k < \infty). \tag{15}$$

ated. (Credit for the reawakening is usually given to Reid in 1971 [41].) This seems inexplicable to us now, but one must remember that the scale of computations in the 1950s was very small; for the matrices of those days, 50 or 100 steps looked no better than $O(n)$. Also, the idea of preconditioning was not widely known until the 1970s [8].

[2] Corollary 3 is often cited, but I have been unable to track down its original appearance in print. The factor of 2 is omitted in some texts, but the resulting inequality is invalid except in the limit $\kappa = \infty$.

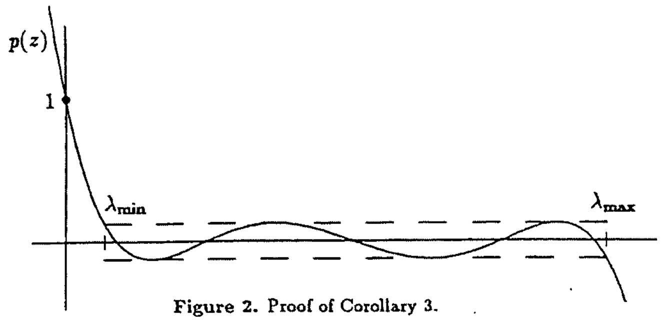

Figure 2. Proof of Corollary 3.

Proof The polynomial $p \in P_k$ that attains the minimum maximal modulus on $[\lambda_{\min}, \lambda_{\max}]$, subject to the constraint $p(0) = 1$, is a shifted and rescaled Chebyshev polynomial:

$$p(x) = \frac{T_k(\gamma - 2x/(\lambda_{\max} - \lambda_{\min}))}{T_k(\gamma)}, \qquad \gamma = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{\kappa + 1}{\kappa - 1}$$

(Figure 2). The supremum of $|p(x)|$ on $[\lambda_{\min}, \lambda_{\max}]$ is $|p(\lambda_{\min})| = |p(\lambda_{\max})| = |T_k(\gamma)|^{-1}$, which tends to be small essentially because Chebyshev polynomials grow faster than any other polynomials, suitably normalized, outside $[-1, 1]$. It can be shown that the right-hand side of (15) is an upper bound for this quantity. □

Corollary 3 implies that the number of iterations required to reduce the error to a specified level ϵ is approximately $\sqrt{\kappa} \log \epsilon$. The corresponding figure for the steepest descent iteration is $\kappa \log \epsilon$, which is much worse unless κ is small.

The significance of Theorem 1 goes far beyond Corollaries 2 and 3: whenever the spectrum of A is favorable for polynomial approximation, the CG iteration will automatically perform well. In particular, the convergence will be rapid if the eigenvalues cluster at one or several points. We turn now to an example of this kind.

3. Strang's preconditioned Toeplitz iteration

We have just seen that if the symmetric positive definite matrix A has clustered eigenvalues, the conjugate gradient iteration will converge quickly. For many problems $Ax = b$, even though the eigenvalues of A are not favorably distributed, a symmetric positive definite matrix C can be found for which the equivalent problem $C^{-1}Ax = C^{-1}b$ does have a favorable eigenvalue distribution. Of course, $C = A$ would be one such matrix, but the point is to pick C so that $C^{-1}y$ ($y \in \mathbb{R}^n$) is easily computable. Such a matrix C is called a *preconditioner*, and the CG iteration applied to $C^{-1}Ax = C^{-1}b$ is known as a *preconditioned conjugate gradient iteration*. (For implementation details, see [8, 20, 48].) The search for effective preconditioners is a central theme of numerical computation nowadays.

Recently Strang has devised a highly effective preconditioner for problems in which A is Toeplitz — that is, constant along diagonals:

$$A = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_1 & & & \\ & & \ddots & \\ & & & a_1 \\ a_{n-1} & \cdots & a_1 & a_0 \end{bmatrix}. \qquad (16)$$

Toeplitz matrices occur in a variety of applications, especially in signal processing and control theory, and existing direct techniques for dealing with them include the Levinson-Trench-Zohar $O(n^2)$ algorithms and a variety of $O(n \log^2 n)$ algorithms such as the recent one by Ammar and Gragg [2]. Strang's idea was to treat Toeplitz systems iteratively by a preconditioned conjugate gradient iteration in which C is chosen to be *circulant*:

$$C = \begin{bmatrix} a_0 & a_1 & \diagdown & a_2 & a_1 \\ a_1 & & & & a_2 \\ \diagdown & & \ddots & & \diagdown \\ a_2 & & & & a_1 \\ a_1 & a_2 & \diagdown & a_1 & a_0 \end{bmatrix}. \qquad (17)$$

A circulant matrix is a Toeplitz matrix in which the diagonals "wrap around," so that the entry a_1 on the first superdiagonal reappears in the lower-left corner, for example. In the present case, as suggested by the dashes, the main diagonal and the first $n/2 - 1$ superdiagonals of C are the same as those of A, but the remaining $n/2$ superdiagonals have been overwritten to achieve the wrap-around.

The idea behind this choice of C is that multiplication by a circulant matrix is equivalent to convolution with a periodic vector, and as a result, circulant systems of equations can be solved in $O(n \log n)$ operations by the Fast Fourier Transform. It follows that each step of a Toeplitz CG iteration with preconditioner C can be executed in $O(n \log n)$ operations, which brings us halfway to an efficient algorithm.

As for the other half, does C precondition this Toeplitz problem effectively? One might expect not, since the corner entries of C and A differ considerably; there is little chance that $C^{-1}A$ will be close to the identity. Yet it turns out that for many Toeplitz matrices A, this preconditioned conjugate gradient iteration converges in 10 or 20 steps. A numerical check reveals that $C^{-1}A$ has a few stray eigenvalues, typically, but that the rest cluster strongly at 1; $C^{-1}A$ is thus close to a low-rank perturbation of the identity. By Theorem 1, the CG iteration is outstanding in such circumstances.

We shall now present the remarkable explanation for this favorable eigenvalue distribution described in a recent paper by Chan and Strang [6]. The essence of the matter is a problem of complex approximation by rational functions on the unit disk. See Chan and Strang for the many details omitted here.

To begin with, assume for simplicity that A has dimension $n = \infty$ (!), which is natural in many applications where A may originate as a finite-rank approximation to an operator. That is, we are going to study the spectrum of the matrices $C_n^{-1}A_n$ in the limit $n \to \infty$, assuming that a fixed sequence of entries $\{a_k\}_{k=0}^{\infty}$ has been prescribed. As usual in the study of Toeplitz matrices, consider the Laurent series

$$f(z) = \sum_{k=-\infty}^{\infty} a_k z^k, \tag{18}$$

whose coefficients $\{a_k\}$ are the entries of A, with $a_k = a_{|k|}$ for $k < 0$. Assume further that these coefficients are absolutely summable,

$$\sum_{k=-\infty}^{\infty} |a_k| < \infty, \tag{19}$$

so that f belongs to the "Wiener algebra" of continuous functions on the complex unit circle $|z| = 1$.

Since A is symmetric and positive definite, it can be shown that $f(z)$ is real and satisfies $f(z) > 0$ for $|z| = 1$. Consider the "spectral factorization" of $f(z)$,

$$f(z) = w(z)w(z^{-1}) \qquad (|z| = 1), \tag{20}$$

where $w(z)$ is a nonzero analytic function in $|z| < 1$. Next, define $v(z)$ by

$$v(z) \;=\; \sum_{k=1}^{\infty} v_k z^k \;=\; \text{analytic (degree } \geq 1\text{) part of } w(z)/w(z^{-1}), \tag{21}$$

and finally, let H be the infinite Hankel matrix

$$H = \begin{bmatrix} v_1 & v_2 & v_3 & \\ v_2 & v_3 & & \\ v_3 & & \ddots & \\ & & & \end{bmatrix}. \tag{22}$$

(A Hankel matrix is constant along counter-diagonals.) Chan and Strang establish the following connection between the singular values of H (= absolute values of the eigenvalues) and the eigenvalues of $C_n^{-1}A_n$:

Lemma 4. [6] There is a two-to-one correspondence between the singular values of H and the eigenvalues of $C_n^{-1}A_n$: as $n \to \infty$, each singular value σ of H is approached by two eigenvalues of $C_n^{-1}A_n$,

$$\lambda_{\pm} = \frac{1}{1 \pm \sigma}. \quad \square \tag{23}$$

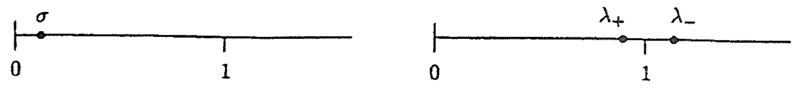

Figure 3. Singular value σ of H and corresponding eigenvalues $\lambda_{\pm}$ of $C_n^{-1}A_n$ (in the limit $n = \infty$).

Figure 3 illustrates this result. We conclude that as the singular values of H decrease to 0, the eigenvalues of $C_n^{-1}A_n$ cluster at $\lambda = 1$. If the decrease is rapid, the clustering will be rapid too, and the CG iteration will converge quickly.

This is where rational approximation unexpectedly enters the picture. According to the theory of "AAK" or "CF" approximation, the singular values $\sigma_0 \geq \sigma_1 \geq \cdots \geq 0$ of H are bounded as follows:

Lemma 5. [1, 53]. The singular values σ_k of H satisfy

$$\sigma_k \leq E_k(v), \tag{24}$$

where $E_k(v)$ denotes the minimal error in sup-norm approximation of $v(z)$ on $|z| = 1$ by rational functions of type (k, k). $\square$

Therefore if $v(z)$ can be efficiently approximated by rational functions, the singular values σ_k must decrease rapidly as $k \to \infty$. In practice, the inequality is often close to an equality, so that the relationship goes approximately in both directions.

Here is the logic in outline:

The entries of A correspond to a smooth function $f(z)$
$\Downarrow$
The function $v(z)$ of (21) is smooth too
$\Downarrow$
$v(z)$ can be well approximated by rational functions on $|z| = 1$
$\Downarrow$ (Lemma 5)
The singular values σ_k of H decrease rapidly to 0
$\Downarrow$ (Lemma 4)
The eigenvalues $\lambda_{\pm k}$ of $C_n^{-1}A_n$ cluster strongly at $\lambda = 1$
$\Downarrow$ (Theorem 1)
The preconditioned CG iteration converges rapidly.

Depending on the precise smoothness assumptions on f, this argument can be fashioned into various theorems. One extreme case is based upon Corollary 2:

Theorem 6. Suppose f is a rational function of type (μ, ν). Then the preconditioned Toeplitz CG iteration converges in at most $1 + 2\max\{\mu, \nu\}$ steps (in the limit $n = \infty$).

Proof If a rational function $f(z)$ is real and positive on $|z| = 1$, then by the Schwarz reflection principle, its poles and zeros lie in pairs symmetric with respect to $|z| = 1$. This implies that the spectral factorization (20) of f is

$$f(z) = w(z)w(z^{-1}) = \frac{p(z)}{q(z)}\,\frac{p(z^{-1})}{q(z^{-1})},$$

where p and q are polynomials, zero-free in $|z| \le 1$, of degrees μ and ν respectively. This equation implies

$$\frac{w(z)}{w(z^{-1})} = \frac{p(z)}{q(z)}\,\frac{q(z^{-1})}{p(z^{-1})},$$

and by (21), the function $v(z)$ is the analytic (degree ≥ 1) part of this expression. By the AAK/CF theory, the number of nonzero singular values of H is equal to the smallest integer N for which $v(z)$ is a rational function of type (N, N) [53] . Careful consideration of the separate cases $\mu \le \nu$ and $\mu > \nu$ shows that this number is $N = \max\{\mu, \nu\}$. From the two-to-one correspondence of Lemma 4, we conclude that $C_n^{-1}A_n$ has at most $1 + 2\max\{\mu, \nu\}$ distinct eigenvalues (in the limit $n = \infty$), including $\lambda = 1$, and the proof is completed with an application of Corollary 2. □

A weaker smoothness assumption, though still strong, is that f should be analytic on $|z| = 1$:

Theorem 7. [6]. Suppose f is analytic in a neighborhood of $|z| = 1$. Then the errors in the preconditioned Toeplitz CG iteration (in the limit $n = \infty$) satisfy

$$\frac{\|e_k\|_A}{\|e_0\|_A} \le r^{k^2} \qquad (0 \le k < \infty) \tag{25}$$

for some $r < 1$.

Proof Let $u(z)$ be the harmonic function in $|z| \le 1$ which has boundary values $u(z) = \log(\sqrt{f(z)})$ on $|z| = 1$, and let $v(z)$ be the harmonic conjugate of $u(z)$ in $|z| \le 1$ normalized by $v(0) = 0$. Then the spectral factorization of f is

$$f(z) = w(z)w(z^{-1}) \quad \text{with} \quad w(z) = e^{u(z)+iv(z)},$$

and since $f(z)$ is analytic in a neighborhood of $|z| = 1$, $u(z)$ and $v(z)$ are harmonic in a neighborhood of $|z| \le 1$, so $w(z)$ is analytic and nonzero in a neighborhood of $|z| \le 1$. Therefore $w(z)/w(z^{-1})$ is analytic in a neighborhood of $|z| = 1$, which implies that the best rational approximants to its analytic part $v(z)$ converge at least geometrically as $k \to \infty$. By Lemmas 4 and 5, it follows that the eigenvalues $\lambda_{\pm k}$ of $C_n^{-1}A_n$ (in the limit $n = \infty$) approach $\lambda = 1$ geometrically as $k \to \infty$.

We now face a special case of the approximation problem of Theorem 1: how small can a polynomial $p \in P_k$ be, subject to $p(0) = 1$, on a set of points that clusters geometrically at $\lambda = 1$? The upper bound (25) comes from choosing p to have zeros at the outermost eigenvalues $\lambda_{\pm 1}, \ldots, \lambda_{\pm k/2}$. □

Even with a far weaker smoothness assumption, we still obtain super-geometric convergence (cf. Corollary 1 to Theorem 1.4.1 of [10]):

Theorem 8. Suppose f is any function in the Wiener class (i.e. with absolutely convergent Laurent series). Then for *any* $\epsilon > 0$, the errors in the preconditioned Toeplitz CG iteration (in the limit $n = \infty$) satisfy

$$\frac{\|e_k\|_A}{\|e_0\|_A} \le C\epsilon^{k} \qquad (0 \le k < \infty) \tag{26}$$

for some constant C.

Proof Define $u(z)$, $v(z)$, and $w(z)$ as in the last proof. If f is in the Wiener class, then $u(z) = \log\sqrt{f(z)}$ is certainly Dini-continuous on $|z| = 1$, which is enough to ensure that $v(z)$ and therefore $w(z)/w(z^{-1})$ are continuous on $|z| = 1$ also. By Hartman's Theorem [39], it follows that the infinite Hankel matrix H is compact, which implies that its singular values decrease to 0, however slowly. The same conclusion can be reached in a more pedestrian fashion by combining the Weierstrass approximation theorem with Lemma 5. By Lemma 4, it follows that the eigenvalues of $C_n^{-1}A_n$ cluster at $\lambda = 1$. Given ϵ, the proof is now completed by choosing $p(z)$ to interpolate 0 at a fixed set of outermost eigenvalues of $C_n^{-1}A_n$ chosen so that the remaining eigenvalues lie in a sufficiently small interval about 1; the remaining degrees of freedom provide the geometric convergence asserted in (26). □

The convergence rates of Theorems 6–8 are impressive, but lest it be obscured by our rather casual assumption "$n = \infty$", let us emphasize the conclusion that underlies all of these results: for finite Toeplitz matrices, the number of iterations is bounded independently of the dimension n as $n \to \infty$, so long as the entries $\{a_k\}$ correspond to a continuous function $f(z)$ with a modest degree of regularity. In such cases Strang's Toeplitz iteration is truly an $O(n \log n)$ algorithm.

It would be excellent to absorb these theorems in a general characterization of the rate of convergence of the iteration as a function of the smoothness of f, but this is impossible at present, for not enough is known about rational approximation. In general, rational functions are far more powerful approximators than polynomials, a fact which was discovered by Newman in 1964 [38] and has been widely generalized since then [19]. Unfortunately, Theorems 7 and 8 do not take advantage of this phenomenon. For example, the hypothesis of analyticity in Theorem 7 is necessary and sufficient to guarantee geometric convergence of *polynomial* approximations on $|z| = 1$, which is what we used to obtain the estimate (25); for rational approximations, it is much too strong. Precise conditions in the rational case are not known, although progress on these questions has been made recently.

4. Polynomial iterations for nonsymmetric matrices

So far, our matrices have been symmetric and positive definite. Now we turn to the problem of nonsymmetric matrices A with complex eigenvalues.

Many iterative methods have been devised for nonsymmetric problems, which go by names such as ORTHOMIN, ORTHODIR, GCR, GMRES, and LSQR; see

[14] for a survey. Most of these are related in one way or another to the conjugate gradient iteration, and often they minimize a norm of the error or the residual, exactly or approximately, over a sequence of subspaces. Unfortunately, no method has emerged which has the elegance and power of the conjugate gradient iteration for symmetric positive definite problems.[3] In particular, the good fortune of a 3-term recurrence relation is often lost, so that at the kth step of an iteration, one has to form linear combinations of $O(k)$ vectors. This makes it more important than ever to find methods that converge in a small number of iterations.

This section will describe a particular class of methods which are based on the assumption that the spectrum Σ of A is contained in a known subset K of the complex plane, which may have been determined adaptively:

$$\Sigma \subseteq K \subset C.$$

(Since A is nonsingular, we assume $0 \notin K$.) Perhaps it is an exceptional problem where an accurate estimate K can be obtained at low cost, but if it can, what use can be made of the information?

This is an old question, to which many researchers have contributed. Varga and his colleagues have described *semi-iterative methods*, in which a simple matrix iteration (e.g. Jacobi or Gauss-Seidel) is accelerated by the formation of linear combinations of the iterates [13, 25, 57]. Others have used the term *polynomial iteration* to describe closely related methods. Omitting details, one ends up in either case, as in Section 2, with a sequence of iterates x_k and errors $e_k = x^* - x_k$,

$$x_k = q(A)b, \qquad e_k = p(A)e_0 \tag{27}$$

for some polynomials

$$q \in P_{k-1}, \qquad p \in P_k, \quad p(0) = 1, \tag{28}$$

with $p(z) = 1 - zq(z)$. (Some methods work with residuals rather than the errors e_k.) However, whereas the conjugate gradient iteration chooses optimal polynomials p and q implicitly and automatically, here we must construct them explicitly.

Thus as in Section 2, we again face two problems of approximation by matrix polynomials: find $q \in P_{k-1}$ such that $q(A) \approx A^{-1}$, or find $p \in P_k$ with $p(0) = 1$ such that $p(A) \approx 0$. To convert these problems from matrices to scalars, it is again natural to consider an eigenvector decomposition. Suppose that A is diagonalizable, and let V be a matrix of normalized eigenvectors and Λ a corresponding diagonal matrix of eigenvalues:

$$A = V\Lambda V^{-1}. \tag{29}$$

[3] Of course one can always apply the conjugate gradient iteration to the normal equations $A^TAx = A^Tb$, implicitly or explicitly, and this is done in LSQR and some other methods. The disadvantage is the possibly large condition number (or unfavorable eigenvalue clustering) of A^TA.

Then $q(A) = Vq(\Lambda)V^{-1}$, which implies

$$A^{-1} - q(A) \;=\; V(\Lambda^{-1} - q(\Lambda))\,V^{-1}, \tag{30}$$

and therefore, making use of the assumption $\Sigma \subseteq K$,

$$\begin{aligned} \|A^{-1} - q(A)\| \;&\le\; \kappa(V)\,\|\Lambda^{-1} - q(\Lambda)\| \\ &=\; \kappa(V)\,\|z^{-1} - q(z)\|_\Sigma \;\le\; \kappa(V)\,\|z^{-1} - q(z)\|_K \end{aligned} \tag{31}$$

if $\|\cdot\|$ is any p-norm ($1 \le p \le \infty$), where $\kappa(V) = \|V\|\,\|V^{-1}\|$ is the condition number of V. Here $\|f\|_\Sigma$ and $\|f\|_K$ are abbreviations for $\sup_{z\in\Sigma}|f(z)|$ and $\sup_{z\in K}|f(z)|$, respectively. Similarly,

$$\begin{aligned} \|p(A)\| \;&\le\; \kappa(V)\,\|p(\Lambda)\| \\ &=\; \kappa(V)\,\|p(z)\|_\Sigma \;\le\; \kappa(V)\,\|p(z)\|_K. \end{aligned} \tag{32}$$

For any fixed nondefective matrix A, $\kappa(V)$ is a finite constant, and thus (31) and (32) imply that there is at most a constant gap between our matrix approximation problems and the corresponding scalar approximation problems, which can be formulated as follows:

Problem Q. Find $q \in P_{k-1}$ to minimize $\|z^{-1} - q(z)\|_K$.

Problem P. Find $p \in P_k$, satisfying $p(0) = 1$, to minimize $\|p(z)\|_K$.

(Problems Q and P are not very different; one can be converted to the other by introducing a sup-norm weighted by $|z|$ or $|z|^{-1}$.)

Polynomial approximation in the complex plane is a well understood subject, though not as straightforward as real approximation on the real line [18, 45, 59]. How shall we (approximately) solve Problem Q or P in the context of our linear algebra problem $Ax = b$? In certain special cases, as mentioned in the Introduction, the solution is well known. For symmetric positive definite matrices, K can be taken to be a real interval, and we are led to Chebyshev iteration [20, 21, 57]. For symmetric indefinite matrices K can be taken as a pair of disjoint real intervals [12, 43]. For some nonsymmetric problems good results can be obtained by a Chebyshev iteration based on a choice of K as an ellipse centered on the real axis [15, 34, 35]. For more general problems, Gutknecht has described a method based on Pick-Nevanlinna interpolation [24], and various authors have considered methods based on least-squares approximation and/or orthogonal polynomials [16, 23, 44, 46, 47]. I shall now describe the method of *interpolation in Fejér points*, which has been investigated recently by Reichel and Fischer [17, 40] and Tal-Ezer [50].

For simplicity, assume that K is simply connected (although the method can be generalized via Green's functions to the multiply-connected case [40]), and let $f(z)$ be a conformal map of the exterior of the unit disk $|z| \le 1$ onto the exterior of K, with $f(\infty) = \infty$. If k interpolation points are needed, let $z_1, \ldots, z_k$ be the

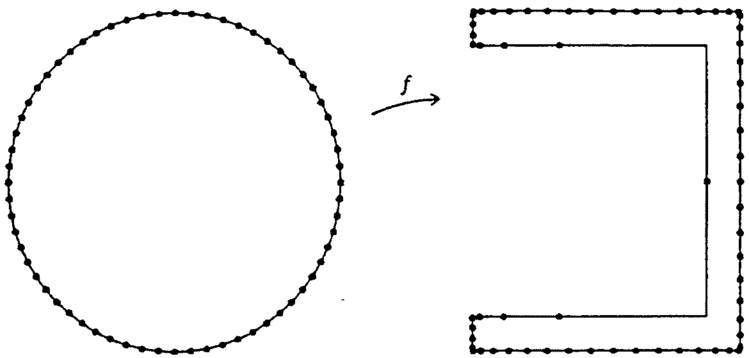

Figure 4. Fejér points for polynomial interpolation ($k=64$). The density is zero at reentrant corners and infinite at salient corners.

kth roots of unity on $|z|=1$, and let $w_1,\ldots,w_k$ be their images under f on the boundary of K. (The map f is guaranteed to extend continuously to the boundary if, for example, K is a Jordan region.) These *Fejér points* are shown in Figure 4 for a region K in the shape of a polygonal U.

Interpolation in Fejér points, though not an optimal approximation strategy for any fixed k, is guaranteed to produce the asymptotically optimal order of convergence as $k\to\infty$:

Theorem 9. [18, 59] For each $k\geq 1$, let $q\in P_{k-1}$ be defined by interpolation of z^{-1} in $k-1$ Fejér points on K. Then for any $\epsilon>0$, there are constants C_1, C_2, and $\rho<1$ such that

$$\|z^{-1}-q(z)\|_K \;\leq\; C_1\,(1+\epsilon)^k \inf_{q'\in P_{k-1}} \|z^{-1}-q'(z)\|_K \;\leq\; C_2\rho^k \tag{33}$$

for all sufficiently large k. Similarly, if $p\in P_k$ is defined by interpolation of 0 with $p(0)=1$, we get the estimate

$$\|p(z)\|_K \;\leq\; C_1\,(1+\epsilon)^k \inf_{p'\in P_k,\,p'(0)=1} \|p'(z)\|_K \;\leq\; C_2\rho^k.\ \square \tag{34}$$

By combining (31) and (33), or (32) and (34), we obtain bounds on the accuracy of the matrix iteration based on interpolation in Fejér points.[4]

There is still a large piece missing in our numerical algorithm: it depends on the conformal map f, and the computation of conformal maps is not trivial. For general smooth domains K, effective algorithms for computing f have been devised, but no software is available [28, 54]. If K is a polygon with a reasonably small number of vertices, on the other hand, f can be obtained by a Schwarz-Christoffel transformation computed by the Fortran package SCPACK [52], and that is how Figure 4 was generated. Since f is a map of the exterior of a polygon, its computation would appear to require a modification of SCPACK, which was designed for interior maps. In most cases of practical interest, however, A has real entries, so that Σ and presumably K are symmetric about the real axis. To take advantage of this symmetry one can compute the map of the exterior of the unit disk in the upper half-plane to the exterior of K in the upper half-plane, and then complete the map by the reflection principle. This map of half-planar regions is essentially an interior Schwarz-Christoffel map of the usual sort (with a vertex at infinity), so unmodified SCPACK is applicable after all.

In certain applications by Fischer and Reichel and by Tal-Ezer and his collaborators, the rather complicated iterative method described in this section has performed dramatically well [50, 51]. But its limitations must also be emphasized: accurate estimates of the spectrum Σ of A are not always available, and the conformal map f may be hard to compute. In the final section we will now discuss a more basic and more interesting limitation.

5. Non-normal matrices

The eigenvalues of a nonsymmetric matrix are in general complex, but there is a more fundamental problem: they may be irrelevant! In a sense eigenvalues are never *exactly* the right thing to look at, when A is not normal;[5] the larger the condition number $\kappa(V)$ of the matrix of eigenvectors, the more wrong they may be (see (31) and (32)). In practical applications $\kappa(V)$ is sometimes huge, especially if A is a member of a family of matrices obtained by a process of discretization.

To begin with an extreme example, consider the defective matrix

$$A=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},\qquad A^{-1}=\begin{pmatrix}1 & -1\\ 0 & 1\end{pmatrix},$$

with $\kappa(V)=\infty$, and suppose our goal is to approximate A^{-1} by a polynomial $q(A)$. Since the spectrum is the single point $\Sigma=\{1\}$, the polynomial $q(z)=1$ matches z^{-1} exactly for $z\in\Sigma$, and thus one might be tempted to consider the approximation

$$q(A)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\approx A^{-1}.$$

[4] Both Reichel & Fischer and Tal-Ezer point out that in the practical implementation of this iteration, the interpolation points cannot be taken in arbitrary order. For numerical stability, the ordering must correspond to an approximately uniform sampling of the boundary, which can be obtained by numbering the boundary points sequentially in binary and then reversing the bits.

[5] A normal matrix is one that possesses a complete orthogonal system of eigenvectors. Equivalently, A is normal if and only if $A^HA=AA^H$, where A^H is the conjugate transpose. Hermitian, skew-Hermitian, unitary, and circulant matrices fall in this category.

Of course this matrix is utterly incorrect in the upper-right corner. Approximation at the eigenvalues is not enough.

To a pure mathematician, what went wrong is obvious: since $\lambda = 1$ is a defective eigenvalue of multiplicity 2, $q(z)$ should have been chosen to match both z^{-1} and its derivative. But to a numerical analyst, this remedy is unappealing, for it violates the principle that a good algorithm should be insensitive to small perturbations. Here, an infinitesimal perturbation might separate the eigenvalues and make A non-defective, suggesting approximation again of function values only. Then $q(z)$ would remain a good approximation to z^{-1} on Σ (though no longer exact), but $q(A)$ would still be a bad approximation to A^{-1}. Thus even for nondefective matrices, approximation at the eigenvalues is not enough.

We propose that a better way to treat highly non-normal matrices is to approximate function values only, not derivatives, but replace Σ by a larger region Σ_ϵ of "approximate eigenvalues." Here is the definition:

Definition. Let A be a square matrix of dimension n, and let $\epsilon \geq 0$ be arbitrary. Then Σ_ϵ, the set of *ϵ-approximate eigenvalues* of A, is the set of numbers $z \in C$ that satisfy any of the following equivalent conditions:

(i) z is an eigenvalue of $A + \Delta$ for some matrix Δ with $\|\Delta\| \leq \epsilon$;
(ii) A has an *ϵ-approximate eigenvector* $u \in C^n$ with $\|(A - z)u\| \leq \epsilon$, $\|u\| = 1$;
(iii) $\sigma_n(zI - A) \leq \epsilon$;
(iv) $\|(zI - A)^{-1}\| \geq \epsilon^{-1}$.

In these assertions $\|\cdot\|$ is the 2-norm, σ_n denotes the smallest singular value, and $(zI - A)^{-1}$ is known as the *resolvent.* The proof that the four conditions are equivalent is a routine exercise in the use of the singular value decomposition.

The sets Σ_ϵ form a nested family, and Σ_0 is the same as Σ. For any ϵ, Σ_ϵ contains all the numbers at a distance $\leq \epsilon$ from Σ,

$$\Sigma_\epsilon \supseteq \{z \in C : \operatorname{dist}(z, \Sigma) \leq \epsilon\}, \qquad (35)$$

with equality if A is normal. If A is not normal, Σ_ϵ may be much larger. In the 2×2 example above, Σ_ϵ is a disk about $z = 1$ of radius $\sim \epsilon^{1/2}$ as $\epsilon \to 0$, which becomes $\sim \epsilon^{1/J}$ if A is generalized to a Jordan block of dimension J. When J is large, $\epsilon^{1/J}$ is close to 1 even when ϵ is as small as machine precision.

Figure 5 illustrates the idea of approximate eigenvalues for two upper-triangular Toeplitz matrices of the form

$$F = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & \text{\Large 0} & \\ & & 0 & 1 & & \\ & & & 0 & 1 & \\ & \text{\Large 0} & & & 0 & 1 \\ & & & & & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} 0 & 1 & 1 & & & \\ & 0 & 1 & 1 & \text{\Large 0} & \\ & & 0 & 1 & 1 & \\ & & & 0 & 1 & 1 \\ & \text{\Large 0} & & & 0 & 1 \\ & & & & & 0 \end{pmatrix},$$

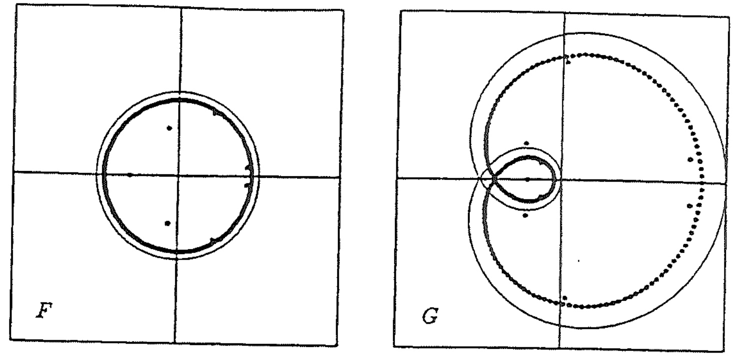

Figure 5. Some ϵ-approximate eigenvalues of two nilpotent matrices F and G of dimension $n = 200$ ($\epsilon = 10^{-8}$). The solid curves are the images of the unit circle under $f(z) = z$ and $g(z) = z + z^2$.

but of dimension $n = 200$; the first is our Jordan block and the second is a kind of generalized Jordan block. (These matrices are again defective, but that could be changed by an arbitrarily small perturbation.) Both F and G are mathematically nilpotent: in each case the spectrum is $\Sigma = \{0\}$. The figure, however, shows eigenvalues of perturbed matrices

$$\tilde{F} = F + \Delta, \quad \tilde{G} = G + \Delta, \qquad \|\Delta\| = 10^{-8},$$

where Δ is a dense random matrix with independent normally distributed elements. The effect of the perturbation on the eigenvalues is enormous and anything but random. If we let f and g denote the "symbols" of these Toeplitz matrices,

$$f(z) = z, \qquad g(z) = z + z^2,$$

and Γ_F and Γ_G the images of the unit circle $|z| = 1$ under f and g, then most of the eigenvalues of $\tilde{F}$ and $\tilde{G}$ evidently lie close to Γ_F and Γ_G.[6] By definition, they are all contained in the corresponding ϵ-approximate eigenvalue regions Σ_ϵ for $\epsilon = 10^{-8}$, and in fact Σ_ϵ becomes exactly Γ_F or Γ_G, together with the regions they enclose, in any limit $\epsilon \to 0$, $n \to \infty$ with $n\epsilon \geq \text{const.} > 0$.

These pictures, coupled with conditions *(i)* and *(ii)* above, give a first indication of why Σ_ϵ may be a more appropriate domain than Σ on which to approximate z^{-1}. If $z \in \Sigma_\epsilon$ for some $\epsilon \ll 1$, then A will behave approximately as if z were an eigenvalue, and if ϵ is as small as machine precision the difference will very likely be undetectable. Under such circumstances, any algorithm which requires a knowledge of Σ in principle had better be supplied with a larger set Σ_ϵ in practice.

[6] If F and G were bi-infinite Toeplitz matrices, their spectra would be Γ_F and Γ_G exactly, and if they were semi-infinite, their spectra would be the regions enclosed by Γ_F and Γ_G.

A more precise statement can be obtained with the aid of condition *(iv)*. It is well known that $A^{-1} - q(A)$ can be represented by the resolvent integral

$$A^{-1} - q(A) = \frac{1}{2\pi i}\oint_\Gamma (z^{-1} - q(z))(zI - A)^{-1}\, dz, \qquad (36)$$

where Γ is any positively oriented closed contour or union of contours that encloses Σ but not $z = 0$ (since z^{-1} has a pole there) [7, 32]. If we define

$$\begin{aligned} \delta &= \|z^{-1} - q(z)\|_\Gamma, \\ L &= \frac{1}{2\pi} \times \text{arc length of } \Gamma, \\ R &= \sup_{z\in\Gamma} \|(zI - A)^{-1}\|, \\ \sigma_{\min} &= R^{-1} = \inf_{z\in\Gamma} \sigma_n(zI - A), \end{aligned}$$

then (36) leads to the estimate

$$\|A^{-1} - q(A)\| \le LR\delta = \frac{L}{\sigma_{\min}}\delta. \qquad (37)$$

In particular, if Γ is taken as the boundary of Σ_ϵ for some $\epsilon > 0$, then $\sigma_{\min} = \epsilon$ and we get

$$\|A^{-1} - q(A)\| \le \frac{L}{\epsilon}\delta. \qquad (38)$$

Equations (37) and (38) reflect a basic tradeoff in the approximation of matrix functions by methods of complex analysis: should the domain of approximation be chosen to lie close to the spectrum of A, or not so close? As Γ contracts to Σ, L decreases to 0 but R increases to ∞. If A is normal, the two effects cancel and small contours are best because of the factor δ: z^{-1} can be approximated more accurately on a small region. If A is far from normal, however, then R may be much larger, and the advantage shifts to contours further away.[7] Whether or not A happens to be exactly defective is unimportant.

All of these observations carry over to more general matrix approximation problems in which z^{-1} and A^{-1} are replaced by arbitrary functions $f(z)$ and $f(A)$, so long as Γ lies in a region where $f(z)$ is analytic. An important example is the approximation of matrix exponentials $f(A) = e^{tA}$, which arises in the numerical solution of differential equations [20, 37, 50]. A further generalization would be to permit approximations $q(z)$ other than polynomials. The following theorem restates the results above in this more general context:

Theorem 10. Let A be a square matrix with spectrum Σ, and let $f(z)$ and $q(z)$ be analytic functions defined in a region K containing Σ. Let Γ be a positively oriented closed contour in K that encloses Σ_ϵ; let L, R, and $\sigma_{\min}$ be defined as above; and define $\delta = \|f(z) - q(z)\|_\Gamma$. Then

$$\|f(A) - q(A)\| \le LR\delta = \frac{L}{\sigma_{\min}}\delta. \qquad (39)$$

In particular, if Γ is the boundary of Σ_ϵ for some $\epsilon > 0$, then

$$\|f(A) - q(A)\| \le \frac{L}{\epsilon}\delta. \ \square \qquad (40)$$

For comparison, here is the estimate analogous to (31):

$$\|f(A) - q(A)\| \le \kappa(V)\,\delta. \qquad (41)$$

This inequality has the advantage that it is independent of Γ, but (39) and (40) have the potentially more important advantage that they are independent of $\kappa(V)$. If A is a member of a family whose eigenvector matrices have unbounded condition numbers, (39) and (40) may possibly provide uniformly valid bounds if Γ is chosen suitably, but (41) cannot.

Although Theorem 10 is stated as an inequality, in practice it gives guidance in both directions. In particular we single out the following rule of thumb:

> *A scalar approximation $q(z) \approx f(z)$ of accuracy δ is of no use for constructing a matrix approximation $q(A) \approx f(A)$ unless it is valid at least on the approximate eigenvalue domain Σ_δ.*

Here is an example to illustrate these ideas. The function

$$h(z) = \frac{1 + z/4}{1 - z/2} = 1 + \frac{3}{4}\left(z + \frac{1}{2}z^2 + \frac{1}{4}z^3 + \cdots\right)$$

maps the unit disk conformally onto the disk D of radius 1 and center 3/2, with $h(0) = 1$. The corresponding $n \times n$ upper-triangular Toeplitz matrix

$$H = \begin{pmatrix} 1 & \frac{3}{4} & \frac{3}{8} & & \cdots \\ & 1 & \frac{3}{4} & \frac{3}{8} & \\ & & 1 & \frac{3}{4} & \frac{3}{8} \\ & 0 & & \ddots & \\ & & & & \end{pmatrix} \qquad (42)$$

has spectrum $\Sigma = \{1\}$ but ϵ-approximate spectrum $\Sigma_\epsilon \approx D$ when ϵ is small and n is large. This situation is represented in Figure 6.

[7] This kind of balancing argument underlies the Kreiss Matrix Theorem and its applications in finite-difference computations [42].

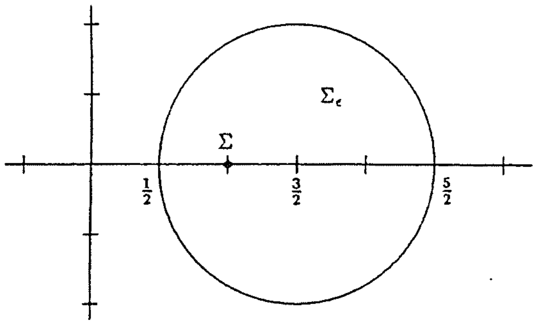

Figure 6. Exact and approximate spectra of the upper-triangular matrix H (ϵ small, n large). The approximate spectrum controls the behavior of a matrix iteration; see Figure 7 below.

Now suppose we solve $Hx = b$ by the one-step Richardson iteration

$$x_{k+1} := x_k + (b - Hx_k)/d, \qquad (x_0 = 0) \tag{43}$$

for some constant d, with corresponding error equation

$$e_{k+1} = (1 - H/d)e_k \qquad (e_k = x^* - x_k). \tag{44}$$

Implicitly we are approximating $f(H) = 0$ by $p_k(H)$ with $p_k(z) = (1 - z/d)^k$. Figure 7 shows the resulting convergence history for dimension $n = 200$ and two values of d.[8] The upper curve represents the "correct" parameter $d = 1$ based on the exact spectrum $\Sigma = \{1\}$, but instead of the instantaneous convergence one would observe if the matrix were normal, the figure shows steady geometric divergence at the rate $(3/2)^k$. The error decreases temporarily around step $k = n$, but rounding errors prevent convergence and soon it is growing again. On the other hand with the "incorrect" parameter $d = 3/2$ in the lower curve, based on the approximate spectrum $\Sigma_\epsilon \approx D$, the iteration converges geometrically at the rate $(2/3)^k$ down to the level of machine precision. Both of these rates are what one would observe for a normal matrix with spectrum D.[9]

This matrix H was manufactured to prove a point, but equally non-normal matrices occur in the wild. One example is the matrix associated with the Gauss-Seidel iteration for the standard discretization of the Poisson equation. For the

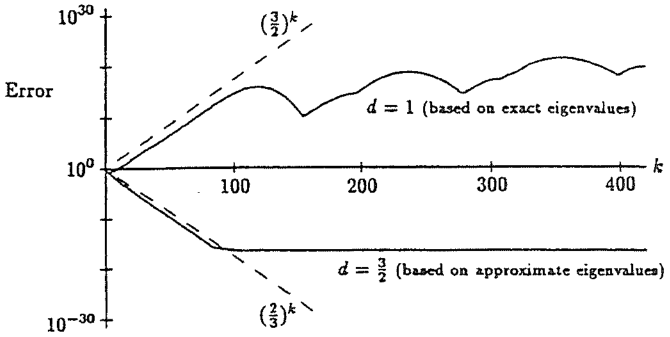

Figure 7. Convergence history of the one-step Richardson iteration (43) for the matrix H ($n = 200$) with two choices of the parameter d.

simplest one-dimensional geometry the eigenvalues are real and positive, but the approximate eigenvalues fill a complex region in the shape of a tennis racket defined by the function $\frac{1}{2}z/(1 - \frac{1}{2}z^{-1})$. As it happens, in this case non-normality does not affect the convergence rate, because the dominant eigenvalue is essentially the same with or without perturbations.

Non-normality has a pronounced effect on convergence, however, in the solution of certain partial differential equations by spectral methods [5]. For example, suppose the model problem

$$u_t = u_x, \qquad x \in [-1, 1], \quad t \geq 0, \quad u(1, t) = 0 \tag{45}$$

is solved by an explicit finite-difference formula in t coupled with a discretization in x consisting of interpolation in N Gauss-Legendre points by a global polynomial $p_N(x)$ satisfying $p_N(1) = 0$ followed by differentiation of $p_N(x)$. This spectral differentiation process is equivalent to multiplication by an $N \times N$ matrix D, which proves to be highly non-normal. The eigenvalues of D are of size $O(N)$, suggesting that the method will be subject to a stability restriction of the form $\Delta t \leq CN^{-1}$ for some constant C [49], but the *approximate* eigenvalues are of size $O(N^2)$ and the actual (Lax-)stability restriction is apparently $\Delta t = O(N^{-2})$ [55, 56]. The stability of spectral methods is poorly understood at present, and perhaps one reason is that the matrices involved tend to be so far from normal.

Although the observations of this section are based on standard mathematics, their application to practical problems has apparently not been explored. It remains to be seen what the implications may be concerning the place of approximation theory in numerical linear algebra. On one hand, the idea of approximating on regions other than exact spectra suggests a new role in which approximation theory

[8] The right-hand side was $b = (1, 1, \ldots, 1)^T$, and the matrix H was first transformed by a random orthogonal similarity transformation to ensure the occurrence of rounding errors. This Richardson iteration is the same as a Chebyshev iteration in which the ellipse degenerates to a point.

[9] An interesting question is whether adaptive methods of estimating eigenvalues, such as those employed by Elman, et al. and by Manteuffel [15, 35], tend to come close to exact or approximate eigenvalues in cases where the two are very different. If the latter is true or can be made true by a suitable choice of adaptive scheme, then adaptively estimated eigenvalues may sometimes yield better matrix iterations than exact ones.

may prove useful. On the other hand, the possibility of extreme non-normality should perhaps stand as a warning that beautiful tricks like the conjugate gradient iteration are inevitably tied to special cases — that in general, matrices are truly more complicated than scalars, and approximation theory can provide only some of the answers.

Acknowledgments

Several people have helped me with valuable advice and references; I would like in particular to mention Martin Gutknecht, Tom Manteuffel, Youcef Saad, and Hillel Tal-Ezer. The experiments of the last section were carried out with the help of Satish Reddy.

References

1. V. M. Adamjan, D. Z. Arov, and M. G. Krein. Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem. *Math. USSR Sbornik*, 15:31–73, 1971.
2. G. S. Ammar and W. B. Gragg. Superfast solution of real positive definite Toeplitz systems. *SIAM J. Matrix Anal. Applics.*, 9:61–76, 1988.
3. A. Brandt. Multi-level adaptive solutions to boundary value problems. *Math. Comp.*, 31:333–390, 1977.
4. C. Brezinski. *Padé-type Approximation and General Orthogonal Polynomials*. Birkhäuser, 1980.
5. C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods in Fluid Dynamics*. Springer, 1988.
6. R. H. Chan and G. Strang. Toeplitz equations by conjugate gradients with circulant preconditioner. *SIAM J. Sci. Stat. Comput.*, to appear.
7. F. Chatelin. *Spectral Approximation of Linear Operators*. Academic Press, 1983.
8. P. Concus, G. H. Golub, and D. P. O'Leary. A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In J. R. Bunch and D. J. Rose, eds., *Sparse Matrix Computations*, Academic Press, 1976.
9. G. Cybenko. An explicit formula for Lanczos polynomials. *Lin. Alg. Applics.*, 88:99–115, 1987.
10. J. W. Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.*, 4:10–26, 1967.
11. J. W. Daniel. *The Approximate Minimization of Functionals*. Prentice-Hall, 1971.
12. C. de Boor and J. R. Rice. Extremal polynomials with application to Richardson iteration for indefinite linear systems. *SIAM J. Sci. Stat. Comp.*, 3:47–57, 1982.
13. M. Eiermann, W. Niethammer, and R. S. Varga. A study of semiiterative methods for nonsymmetric systems of linear equations. *Numer. Math.*, 47:505–533, 1985.
14. H. C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*. PhD thesis, Res. Rep. #229, Dept. of Comp. Sci., Yale U., 1982.
15. H. C. Elman, Y. Saad, and P. E. Saylor. A hybrid Chebyshev Krylov subspace algorithm for solving nonsymmetric systems of linear equations. *SIAM J. Sci. Stat. Comput.*, 7:840–855, 1986.

16. D. K. Faddeev and V. N. Faddeeva. *Computational Methods of Linear Algebra*. W. H. Freeman, 1963.
17. B. Fischer and L. Reichel. A stable Richardson iteration method for complex linear systems. *Numer. Math.*, to appear.
18. D. Gaier. *Lectures on Complex Approximation*. Birkhäuser, 1987.
19. T. Ganelius, W. K. Hayman, and D. J. Newman. *Lectures on Approximation and Value Distribution*. Les Presses de L'Université de Montréal, 1982.
20. G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins U. Press, 1983.
21. G. H. Golub and R. S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second-order Richardson iterative methods, parts I and II. *Numer. Math.*, 3: 147–56 and 157–168.
22. W. B. Gragg. The Padé table and its relation to certain algorithms of numerical analysis. *SIAM Review*, 14:1–62, 1972.
23. W. B. Gragg and L. Reichel. On the application of orthogonal polynomials to the iterative solution of linear systems of equations with indefinite or non-Hermitian matrices. *Lin. Alg. Applics.*, 88:349–371, 1987.
24. M. H. Gutknecht. An iterative method for solving linear equations based on minimum norm Pick-Nevanlinna interpolation. In C. K. Chui, et al., eds., *Approximation Theory V*, Academic Press, 1986.
25. M. H. Gutknecht. Stationary and almost stationary iterative (k,l)-step methods for linear and nonlinear systems of equations. To appear.
26. W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, 1985.
27. L. A. Hageman and D. M. Young. *Applied Iterative Methods*. Academic Press, 1981.
28. P. Henrici. *Applied and Computational Complex Analysis, v. 3*. Wiley, 1986.
29. M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49:409–436, 1952.
30. O. G. Johnson, C. A. Micchelli, and G. Paul. Polynomial preconditioners for conjugate gradient calculations. *SIAM J. Numer. Anal.*, 20:362–376, 1983.
31. S. Kaniel. Estimates for some computational techniques in linear algebra. *Math. Comp.*, 20:369–378, 1966.
32. T. Kato. *Perturbation Theory for Linear Operators*. Springer, 1976.
33. D. Luenberger. *Introduction to Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, 1984.
34. T. A. Manteuffel. The Tchebychev iteration for nonsymmetric linear systems. *Numer. Math.*, 28:307–327, 1977.
35. T. A. Manteuffel. Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration. *Numer. Math.*, 31:183–208, 1978.
36. G. Meinardus. Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch. *Numer. Math.*, 5:14–23, 1963.
37. C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20:801–836, 1978.
38. D. J. Newman. Rational approximation to $|x|$. *Michigan Math. J.*, 11:11–14, 1964.
39. S. C. Power. *Hankel Operators on Hilbert Space*. Pitman, 1982.
40. L. Reichel. Polynomials for the Richardson iteration method for complex linear systems. Preprint 84/1, U. Hamburg, 1984.
41. J. K. Reid. On the method of conjugate gradients for the solution of large sparse systems of linear equations. In J. K. Reid, ed., *Large Sparse Sets of Linear Equations*,

Academic Press, 1971.
42. R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*, 2nd ed. Wiley, 1967.
43. Y. Saad. Iterative solution of indefinite symmetric linear systems by methods using orthogonal polynomials over two disjoint intervals. *SIAM J. Numer. Anal.*, 20:784–811, 1983.
44. Y. Saad. Least squares polynomials in the complex plane and their use for solving nonsymmetric linear systems. *SIAM J. Numer. Anal.*, 24:155–169, 1987.
45. E. B. Saff. Polynomial and rational approximation in the complex domain. *Proc. Symp. Appl. Math.*, 36:21–49, Amer. Math. Soc., 1986.
46. D. C. Smolarski and P. E. Saylor. Optimum parameters for the solution of linear equations by Richardson iteration. Unpublished paper.
47. E. L. Stiefel. Kernel polynomials in linear algebra and their numerical applications. *U.S. Nat. Bur. Stand. Appl. Math. Series*, 49:1–24, 1958.
48. G. Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, 1986.
49. H. Tal-Ezer. A pseudospectral Legendre method for hyperbolic equations with an improved stability condition. *J. Comp. Phys.*, 67:145–172, 1986.
50. H. Tal-Ezer. Polynomial approximation of functions of matrices and its application to the solution of a general system of linear equations. ICASE Report 87-63, NASA Langley Research Center, 1987.
51. H. Tal-Ezer, J. M. Carcione, and D. Kosloff. An accurate and efficient scheme for wave propagation in linear viscoelastic media. *Geophysics*, to appear.
52. L. N. Trefethen. Numerical computation of the Schwarz-Christoffel transformation. *SIAM J. Sci. Stat. Comput.*, 1:82–102, 1980.
53. L. N. Trefethen. Rational approximation on the unit disk. *Numer. Math.*, 37:297–320, 1981.
54. L. N. Trefethen, ed. *Numerical Conformal Mapping*. North-Holland, 1986.
55. L. N. Trefethen. Lax-stability vs. eigenvalue stability of spectral methods. To appear in Proceedings, 1988 Oxford Conference on Computational Fluid Dynamics.
56. L. N. Trefethen and M. R. Trummer. An instability phenomenon in spectral methods. *SIAM J. Numer. Anal.*, 24:1008–1023, 1987.
57. R. S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, 1962.
58. E. L. Wachspress. *Iterative Solution of Elliptic Systems*. Prentice-Hall, 1966.
59. J. L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*, 5th ed. Amer. Math. Soc., 1969.
60. D. M. Young. *Iterative Solution of Elliptic Systems*. Academic Press, 1971.

Algorithms for Approximation II

Based on the proceedings of the Second International Conference on Algorithms for Approximation, held at Royal Military College of Science, Shrivenham, July 1988

Edited by

J. C. Mason
Professor of Computational Mathematics,
Royal Military College of Science, Shrivenham, UK
and

M. G. Cox
Senior Principal Scientific Officer,
National Physical Laboratory, Teddington, UK

London New York
CHAPMAN AND HALL
1990