FROM POTENTIAL THEORY TO MATRIX ITERATIONS IN SIX STEPS*

TOBIN A. DRISCOLL[†], KIM-CHUAN TOH[‡], AND LLOYD N. TREFETHEN[§]

To the memory of Peter Henrici

Abstract. The theory of the convergence of Krylov subspace iterations for linear systems of equations (conjugate gradients, biconjugate gradients, GMRES, QMR, Bi-CGSTAB, and so on) is reviewed. For a computation of this kind, an estimated asymptotic convergence factor $\rho \leq 1$ can be derived by solving a problem of potential theory or conformal mapping. Six approximations are involved in relating the actual computation to this scalar estimate. These six approximations are discussed in a systematic way and illustrated by a sequence of examples computed with tools of numerical conformal mapping and semidefinite programming.

Key words. potential theory, matrix iterations, Krylov subspaces, polynomial approximation, conformal mapping, semidefinite programming

AMS subject classifications. 30E10, 31A99, 41A10, 65F10

PII. S0036144596305582

1. Introduction. Many of the large-scale matrix problems of computational science and engineering are solved by Krylov subspace iterations. The most famous of these methods is the conjugate gradient iteration for symmetric positive definite matrices. Nonsymmetric generalizations of conjugate gradients are numerous and include iterations such as biconjugate gradients, GMRES, CGS, Bi-CGSTAB, QMR, and TFQMR. For introductions to these methods see [1], [3], [25], [29], [34], [40], [47], [53], [63], [77]. The importance of these iterations nowadays is enormous, and it continues to grow as computers become faster, matrices get larger, and the $O(N^3)$ operation counts associated with many direct methods become ever more unacceptable.

How fast does an iteration for solving a system of equations $Ax = b$ converge? (If the problem has been preconditioned, as is usually necessary for effective convergence, then by $Ax = b$ we mean the already preconditioned system.) Starting from an initial guess x_0, such an iteration generates a sequence of approximations $x_1, x_2, \ldots$ that converge toward the solution $A^{-1}b$. Convergence is typically measured by the ratio

$$(1) \qquad \frac{\|r_n\|}{\|r_0\|}, \qquad n = 1, 2, 3, \ldots,$$

where $r_0 = b - Ax_0$ is the initial residual and $r_n = b - Ax_n$ is the residual at step n. (One may also be concerned with ratios $\|e_n\|/\|e_0\|$ of errors rather than residuals, where $e_n = A^{-1}b - x_n$, but as r_n and e_n are related by the formula $r_n = Ae_n$ and as r_n can be measured in the course of an iteration whereas e_n cannot, we shall consider only residuals.) For simplicity we shall assume throughout this paper that $\| \cdot \|$ is the

†Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO, 80309-0526 (tad@sturm.colorado.edu).

‡Department of Mathematics, National University of Singapore, Singapore (mattohkc@leonis.nus.sg).

§Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, UK (LNT@comlab.ox.ac.uk).

2-norm (square root of sum of squares). The question of most pressing importance in any matrix iteration is, how fast do the numbers (1) decrease towards 0 ?

There is an elegant, fundamental idea that gives an approximate answer to this question. Let the spectrum of A be approximated by a compact set S in the complex plane $\mathbf{C}$, with $0 \notin S$. Sets often considered in practice include intervals, ellipses, polygons, isolated points, and combinations of these. Now ask, how small can a polynomial $p(z)$ of degree n be on S if it is normalized by $p(0) = 1$? Specifically, let P_n denote the set of polynomials of degree at most n with $p(0) = 1$, and let the norm $\| \cdot \|_S$ associated with S be defined by

$$(2) \qquad \|p\|_S \;=\; \max_{z \in S} |p(z)|.$$

Then how small are the quantities

$$(3) \qquad E_n(S) \;=\; \min_{p \in P_n} \|p\|_S, \qquad n = 1, 2, 3, \dots \; ?$$

A salient property of the sequence $\{E_n(S)\}$ is that it decreases geometrically with n at some rate $\rho \leq 1$:

$$(4) \qquad \rho \;=\; \lim_{n \to \infty} (E_n(S))^{1/n} \;\leq\; 1.$$

This limiting value ρ, which we shall call the *estimated asymptotic convergence factor,* always exists, and unless S completely surrounds the origin in the sense of separating it from the point at infinity, it is strictly less than 1. Moreover, as we shall describe, its value can be derived from potential theory, or equivalently by the consideration of Green's functions, and if S is connected, the derivation reduces to a conformal map of the exterior of S to the exterior of a disk.

Implicitly or explicitly, the number ρ is the starting point of most attempts to understand the behavior of the sequence (1). Specifically, in many cases the estimate

$$(5) \qquad \frac{\|r_n\|}{\|r_0\|} \;\approx\; \rho^n$$

is a reasonable approximation. The reason for the link is that Krylov subspace iterations also involve polynomials—polynomials of matrices. The iterates produced by these methods typically lie in the *Krylov subspaces* generated by the initial vector and the matrix A,

$$(6) \qquad r_n \;\in\; r_0 + \langle Ar_0, A^2 r_0, \dots, A^n r_0 \rangle,$$

where the angle brackets indicate the subspace spanned by the vectors inside. This is equivalent to the statement that r_n can be written in the form

$$(7) \qquad r_n \;=\; p(A)r_0$$

for some $p \in P_n$. (The error e_n satisfies $e_n = p(A)e_0$ for the same polynomial p.) And thus we see that the convergence of a Krylov subspace iteration can be expected to be rapid if it manages to find polynomials $p(A)$ that are effective at approximately annihilating r_0.

The aim of this paper is to introduce the reader to the subject of convergence of Krylov subspace iterations, using the approximation (5) as our reference point. Little

of what we say is mathematically new. For example, a shorter survey covering some of the same ground as this one has been written by Greenbaum [32], and an extensive analysis based on a different point of view can be found in the book by Nevanlinna [55]. However, our presentation has some unusual features. One is that from start to finish, everything is illustrated by computed examples based upon potential theory or approximation theory problems in the complex plane. The computation of these examples has been made possible by software developed by us for Schwarz–Christoffel conformal mapping [9] and for finding polynomials $p(z)$ that minimize norms $\|p(A)\|$ [74]. The other is that our tour is organized in an unusual way. We have observed that to get from the right-hand side of (5) to the left-hand side, one can proceed in *six steps,* breaking the "$\approx$" into a sequence of six approximations. All of these approximations have been considered before in the literature. However, discussions of this subject have not been structured explicitly in this way before.

Readers looking for an overview may turn to the summary at the end, where, among other things, we list some of the many things not done in this paper.

2. Derivation of ρ via potential theory. Let $S \subseteq \mathbf{C}$ be a compact set with $0 \notin S$, and let ρ be the estimated asymptotic convergence factor (4) for S. To avoid technical complications, we assume that S consists of a finite collection of simply connected components, each of which is either an isolated point or a continuum with a piecewise smooth boundary.

The derivation of ρ via potential theory can be explained in a few lines, if we do not attempt to be rigorous. Consider a polynomial $p(z) = \prod_{k=1}^{n}(z - z_k)$. The absolute value of $p(z)$ is

$$|p(z)| \;=\; \prod_{k=1}^{n} |z - z_k|.$$

Our goal is to minimize $|p(z)|/|p(0)|$ on S [17]. By the maximum modulus principle, an equivalent problem is to minimize the same quantity on ∂S. Taking the logarithm gives

$$\log |p(z)| \;=\; \sum_{k=1}^{n} \log |z - z_k|.$$

This function can be thought of as the potential in the complex plane associated with electric point charges of amplitude -1 at each of the points $\{z_k\}$. (The potential associated with a negative point charge in 3-space is $-|z - z_k|^{-1}$; for an infinite line of charge, whose intersection with the plane is a point, the formula becomes $\log |z - z_k|$.) Our aim is to locate these charges so as to minimize the quantity

$$(8) \qquad \log |p(z)| - \log |p(0)| \;=\; \log \prod_{k=1}^{n} \left| 1 - \frac{z}{z_k} \right|$$

on ∂S.

For finite n, this minimization problem is difficult, but we can make it simpler by taking the limit $n \to \infty$. First we rescale the problem, defining

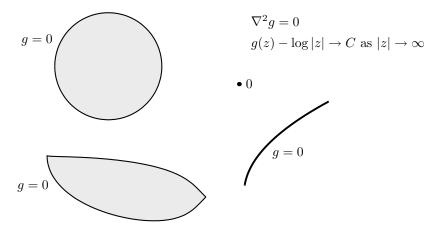$$g(z) \;=\; n^{-1} \sum_{k=1}^{n} \log |z - z_k| + C.$$

FIG. 1. *Determination of the estimated asymptotic convergence factor ρ in the case where the set S_0 has several connected components. In this sketch S_0 consists of an arc, a disk, and another region with interior. Think of these as electrical conductors that are electrically connected to one another. Inject a charge of magnitude -1, so that the potential g adjusts to a constant value along all of the boundaries, and add a constant C to g so as to make this value 0 (g is the Green's function for S_0). The number ρ is determined by the value of the potential at the origin: $\rho = \exp(-g(0))$.*

This is the potential corresponding to point charges of strength $-n^{-1}$ at each of the points $\{z_k\}$; the arbitrary constant C will be fixed in a moment. In the limit $n \to \infty$, we now imagine a negative unit charge distributed in a continuous fashion in the complex plane, and our aim is to minimize $\max_{z \in S} g(z) - g(0)$.

In passing to the limit $n \to \infty$, a complication arises concerning treatment of isolated points of S. It requires just a single root to make a polynomial zero at an isolated point. As $n \to \infty$, this corresponds to a zero proportion of the total charge, and thus we cannot expect a continuous charge distribution to handle isolated points of S properly. One way to handle this matter is as follows. Let S_0, assumed to be nonempty, denote the subset of S in which any isolated points (necessarily finite in number) have been removed. Denote the boundary of S by ∂S and the boundary of S_0 by ∂S_0. We shall apply our potential theory to S_0, and the results will be valid in the nth-root asymptotic sense (4) for S too.

Figure 1 illustrates a typical setting for our discussion. The set S_0 consists of the union of an arc and two regions with interior. The set S might contain in addition some isolated points, not shown in the figure.

Our goal is to minimize $\max g(z) - g(0)$ on S_0. Now in the case of a continuous charge distribution, it is plausible that the minimum should be achieved by a charge distribution that has the property that $g(z)$ is constant for all $z \in \partial S_0$. The idea is that if $g(z)$ were not constant on ∂S_0, then its maximum could be reduced by redistributing the charge away from low-potential regions. In fact, it can be shown that this idea is correct. Moreover, the minimum is achieved by a charge distribution in which all the charge is confined to the boundary ∂S_0. (The process of finding a charge distribution on ∂S_0 that is equivalent outside S_0 to a continuous charge distribution on S_0 is known in potential theory as *balayage* [69], [78].)

We have arrived at the following physical picture. Think of S_0 as a collection of conductors in the plane that are electrically connected. Inject a quantity -1 of charge into this system, and let it find an equilibrium. The charge will distribute itself along the boundary ∂S_0 in such a way that the potential $g(z)$ it generates is constant there. Add a constant C to this potential so that this constant value becomes 0. The asymptotic convergence factor is now $\rho = \exp(-g(0))$.
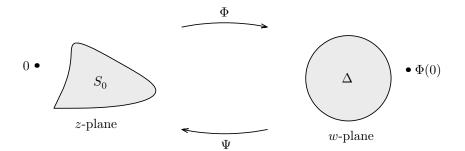
FIG. 2. *Determination of ρ in the case where S_0 is connected. All that is required is a conformal map Φ of the exterior of S_0 to the exterior of the unit disk Δ. The estimated asymptotic convergence factor is $\rho = 1/|\Phi(0)|$.*

Of course, these results can be stated mathematically, without recourse to physical interpretations. The function $g(z)$ is the *Green's function* associated with S_0, the unique function defined in the exterior of S_0 that satisfies $\nabla^2 g = 0$ outside S_0, $g(z) \to 0$ as $z \to \partial S_0$, and $g(z) - \log|z| \to C$ as $|z| \to \infty$ for some constant C. The quantity e^{-C} is the *logarithmic capacity* of S_0 (or equivalently of S), and C is *Robin's constant* [45], [78]. We can state the fundamental results about ρ as follows.

THEOREM 1. *Let $g(z)$ be the Green's function for the set S_0. The estimated asymptotic convergence factor of S as defined by* (1) *and* (2) *is*

$$(9) \qquad\qquad \rho = \exp(-g(0)).$$

Moreover, for each n and any $p \in P_n$ we have

$$(10) \qquad\qquad \|p\|_S \geq \rho^n.$$

The lower bound (10) has an intuitive explanation. A polynomial of finite degree n corresponds to the constraint that a quantity of charge must be distributed in n points rather than allowed to spread around in a continuous fashion. The resulting maximum potential on S can accordingly be worse than that for the continuous limit, but not better. Rigorous proofs of this and other results mentioned here are not difficult; one of the key tools is the maximum principle. For more information about potential theory and polynomial minimization, see [27], [42], [45], [69], [78], [81], [82]. For more about the connection with matrix iterations, see [13], [14], [15], [20], [22], [55], [56], [59].

3. If S_0 is connected: Derivation of ρ via conformal mapping. The picture just presented is conceptually simple, though calculating Green's functions may be a challenge. In the case where S_0 is connected, the problem becomes still simpler.

What happens in this case is that the potential $g(z)$ can be viewed as a level function of a conformal map. As illustrated in Fig. 2, suppose S_0 is a simply connected compact subset of $\mathbf{C}$, not containing 0, with a piecewise smooth boundary. Since S_0 is connected, the exterior of S_0 is a simply connected set with respect to the extended complex plane $\mathbf{C} \cup \{\infty\}$. It follows that a harmonic function $h(z)$ can be defined in the exterior of S_0 that is a harmonic conjugate of $g(z)$ and is single-valued except for increasing by $2\pi i$ with each counterclockwise circuit around S_0. Consider now the complex function

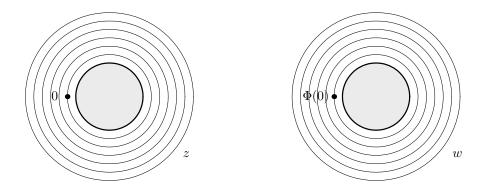$$(11) \qquad\qquad \Phi(z) = \exp(g(z) + ih(z))$$

FIG. 3. *Simplest example of a set S: a disk in the z-plane disjoint from the origin. The map to the exterior of the unit disk in the w-plane is just a translation and a scalar multiplication. The estimated asymptotic convergence factor is $\rho = (\kappa - 1)/(\kappa + 1)$, the ratio of the radius of the disk to the distance to the origin.*

defined in the exterior of S_0. This function is analytic, with nonvanishing derivative, and it satisfies $|\Phi(z)| = \exp(g(z)) \to 1$ for $z \to \partial S_0$ and $|\Phi(z)| \sim |z|e^C$ as $|z| \to \infty$. In other words, $\Phi(z)$ is a conformal map of the exterior of S_0 onto the exterior of the unit disk Δ, with $\Phi(\infty) = \infty$. Just as $h(z)$ is determined up to an additive constant, $\Phi(z)$ is determined up to a multiplicative constant of modulus 1, corresponding to an arbitrary rotation in the w-plane. For our purposes there is no need to specify this constant of rotation.

Let Ψ denote the inverse of Φ, mapping the exterior of Δ conformally onto the exterior of S_0. Thus Ψ maps the unit circle $|w| = 1$ onto ∂S_0, and it maps concentric circles $|w| = r > 1$ in the w-plane onto simple closed curves enclosing ∂S_0 in the z-plane. These are the *level curves* associated with Φ, defined by the condition $|\Phi(z)| = r$ for various values of r.

By assumption, the point 0 is exterior to S_0. Thus its image $\Phi(0)$ is exterior to Δ. And now, combining (9) and (11), we obtain a formula for ρ.

THEOREM 2. *Let S_0 be connected, and let $\Phi(z)$ be a conformal map of the exterior of S_0 to the exterior of the unit disk Δ with $\Phi(\infty) = \infty$. The estimated asymptotic convergence factor of S is*

$$(12) \qquad\qquad \rho \;=\; \frac{1}{|\Phi(0)|}.$$

In words, if S_0 is connected, the estimated asymptotic convergence factor for a matrix iteration depends on how far the origin is from S_0—provided that this distance is measured by level curves associated with the exterior conformal map.

4. Examples of regions S and convergence factors ρ. We shall illustrate these results by three examples of the constant ρ associated with various sets S. Subsequent sections will refine the picture by looking at the "six steps."

The first example is mathematically trivial, but of enduring importance for applications. Suppose $S = S_0$ is the disk $|z - z_0| \leq R$ for some $R < |z_0|$. To map the exterior of S conformally onto the exterior of the unit disk as in Theorem 2, all that is required is a translation and a scalar multiplication. A suitable pair of maps is $\Phi(z) = (z - z_0)/R$, $\Psi(w) = Rw + z_0$. Thus we see that when S is a disk of radius R about z_0, the estimated asymptotic factor is $\rho = R/|z_0|$. Rapid convergence can be expected if the disk is small relative to its distance to the origin. See Fig. 3.
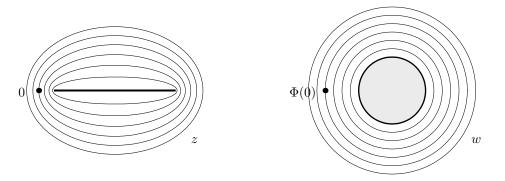
FIG. 4. *Second-simplest example of a set S: an interval of the real axis, chosen here to have the same condition number κ as in Fig. 3. Thanks to the sharp angle of ∂S at the end of the interval, the distinguished point $\Phi(0)$ in the w-plane now lies much further from the unit disk. The estimated asymptotic convergence factor changes to $\rho = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$, an improvement from $1 - O(1/\kappa)$ to $1 - O(1/\sqrt{\kappa})$ for large κ.*

For comparison with other examples, it is informative to restate this result in terms of the "condition number" of the disk S: the ratio κ of its largest to smallest points. Specifically, suppose $\rho < 1$ is given. Then any disk with ratio $R/|z_0| = \rho$ can be defined by $z_0 = \alpha(\kappa+1)/2$ and $R = |\alpha|(\kappa-1)/2$ for some constant α. The formula for ρ as a function of κ becomes

$$(13) \qquad\qquad \rho \;=\; \frac{\kappa - 1}{\kappa + 1} \qquad\qquad \text{(disk)}.$$

Rapid convergence can be expected if κ is not too large. If it is large, then we have $\rho \sim 1 - 2/\kappa$, and we must expect to take $\sim \kappa/2$ iterations to reduce the residual norm by a factor e.

This first example illustrates the most basic property one might look for in an effective preconditioner. A preconditioner is likely to be effective if (though not only if) it moves the spectrum of the original problem, or at least all of the spectrum except a few isolated points, into a disk whose radius is reasonably small relative to the distance from its center to the origin.

Our second example is also elementary. If $S = S_0$ is an interval, the conformal map of its exterior onto the exterior of the unit disk is a Joukowski map, carrying circles to ellipses (Fig. 4). To go directly to the κ notation, suppose $S = [1, \kappa]$ for some $\kappa > 1$. Then a suitable pair of maps is $\Phi(z) = (2z - \kappa - 1 + 2\sqrt{z^2 - (\kappa + 1)z + \kappa})/(\kappa - 1)$, $\Psi(w) = [(w + w^{-1})(\kappa - 1)/2 + \kappa + 1]/2$. Calculating $\Phi(0)$ as in (12) gives the estimated asymptotic convergence factor

$$(14) \qquad\qquad \rho \;=\; \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \qquad\qquad \text{(interval)}.$$

This factor is famous as associated with the convergence rate of the conjugate gradient iteration for a matrix with condition number κ. Note that whereas $\rho \sim 1 - 2/\kappa$ for large κ in (13), $\rho \sim 1 - 2/\sqrt{\kappa}$ in (14), and this number is much further from 1. For large κ, errors can be reduced by a factor e in $\sim \sqrt{\kappa}/2$ rather than $\sim \kappa/2$ iterations.

These two examples are standard and well known. Now, let us consider a more complicated example, for which the conformal map was computed numerically using the Schwarz–Christoffel Toolbox [9]. The Schwarz–Christoffel Toolbox is a publicly
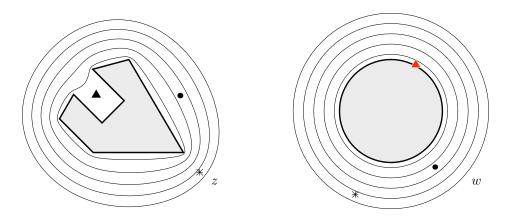
FIG. 5. *Conformal map of the exterior of an irregular octagon S in the z-plane, indicating the images in the w-plane of three distinguished points. The points lie at similar distances from S, but their images lie at very different distances from the unit disk.*
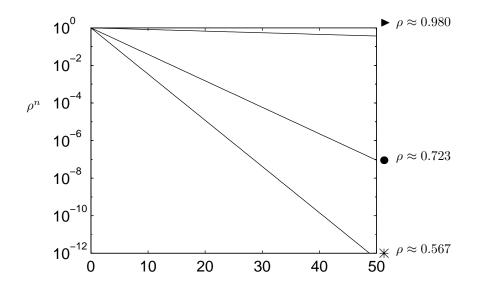

FIG. 6. *Estimated asymptotic convergence rates associated with the origin located at each of the three points of Fig. 5. The convergence is very slow for the case ▶, corresponding to a matrix A with spectrum approximately surrounding the origin.*

available Matlab package that enables the user to draw a polygon and determine the corresponding conformal map at the touch of a button.

Figures 5 and 6 present an estimated spectrum that illustrates several types of behavior at once. Here $S = S_0$ is the closed region bounded by a polygon with both salient and reentrant corners. The figure shows a collection of concentric circles in the w-plane and the corresponding level curves surrounding S in the z-plane. Three points in the z-plane have been selected, with their conformal images marked in the w-plane. The point marked by $*$ lies near a salient corner of S, and its image under $\Phi(z)$ is accordingly far from the unit circle. It follows that if the origin in the
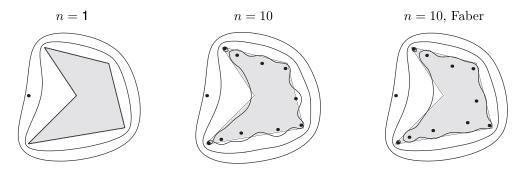
$n = 1$ $n = 10$ $n = 10$, Faber

FIG. 7. *For an irregular pentagon S in the z-plane, the first two plots compare the level curves (lemniscates) of optimal polynomials p of degrees $n = \infty$ (the limit of a continuous charge distribution or Green's function) and $n = 10$. The third shows the corresponding picture for a near-optimal polynomial, the appropriately scaled Faber polynomial. From the fact that the inner lemniscates do not quite enclose all of S, we see that the two finite-degree polynomials do not quite achieve the lower bound ρ^n.*

z-plane is located at $*$, the iteration will converge fast ($\rho \approx 0.567$). By contrast, the point marked by $\blacktriangleright$ lies near a reentrant corner of S, and its image under $\Phi(z)$ lies exceedingly close to unit circle. If the origin in the z-plane is located at $\blacktriangleright$, the iteration will converge slowly ($\rho \approx 0.980$). The point marked by $\bullet$, lying close to a straight side, is associated with behavior in-between ($\rho \approx 0.723$).

The lessons of Figs. 5 and 6 apply generally to matrix iterations for $Ax = b$. If the spectrum "surrounds" the origin, approximately speaking, then the convergence is likely to be slow. Like many of the phenomena encountered in the study of conformal maps, this effect is exponentially strong; if the origin lies several channel widths deep inside a channel in the spectrum, then ρ is likely to be almost equal to 1, so that in practice, no useful convergence is achieved. The extreme situation occurs if the spectrum completely surrounds the origin in the sense of disconnecting it from the point at infinity. Then, by the maximum modulus principle, no polynomial $p \in P_n$ can satisfy $|p(z)| < 1$ everywhere on the spectrum, and there is no convergence at all.

5. Step 1: $n \to \infty$ vs. finite n. We now begin our discussion of the six steps that can be taken to get from ρ^n to $\|r_n\|/\|r_0\|$. We consider them in a natural order, whose logic is summarized by a string of inequalities in Table 1 at the end.

The first step is the difference between nth root asymptotic behavior as $n \to \infty$ and behavior for finite n, that is, between equations (4) and (3). According to Theorem 1, the nth root behavior as $n \to \infty$ of minimal polynomials on the set S is controlled by a Green's function or, equivalently, a continuous charge distribution that achieves a constant potential on ∂S_0. For finite n, however, the charge is constrained to lie in n points. In general, it cannot distribute itself so as to achieve a constant potential on ∂S. We thus have the inequality

(STEP1) $$\rho^n \;\leq\; \min_{p \in P_n} \|p\|_S,$$

whose generalization without the minimum over p was stated already in (10).

Figure 7 illustrates the phenomenon in question. The region $S = S_0$ is a pentagon in the z-plane with the origin located nearby on the left. The first plot shows level curves of the map Φ from the exterior of S to the exterior of the unit disk; the levels plotted are $|\Phi(z)| = 1$ (the boundary of S), 1.2, and 1.4. This corresponds to the
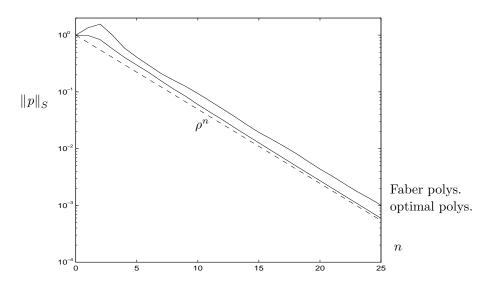
FIG. 8. *Convergence curves associated with the approximations of Fig. 7. For this region S, the gap between ρ^n and $E_n(S)$ is not so great as to be very important in practice.*

limit $n \to \infty$; the polynomial zeros or electric charges are spread around ∂S in a continuous fashion and the potential $g(z)$ is constant there. This plot was generated by computing the inverse conformal map $\Psi(w)$ by the Schwarz–Christoffel Toolbox, then plotting images under Ψ of circles in the w-plane.

The second plot of Fig. 7 is the corresponding picture for $n = 10$. Here the optimal polynomial $p \in P_{10}$ of (3) has been computed using the software described in [74]. Note that the ten zeros are distributed near the boundary ∂S, though not exactly on it. Note also how they avoid the reentrant corner. (The study of zeros of minimal and near-minimal polynomials in $\mathbf{C}$ is a highly developed topic [66], [69].) The level curves plotted are the lemniscates $|p(z)| = \rho^{10}$, $(1.2\rho)^{10}$, $(1.4\rho)^{10}$. Away from S, the level curves look almost the same as in the first figure, but the maximum on S is greater: we have $\rho^{10} = 0.0494$ but $E_{10}(S) = 0.0599$.

The third plot of Fig. 7 corresponds to a polynomial we shall not discuss in this paper, but which is standard in complex approximation theory. The degree n *Faber polynomial* for the region S is the analytic part of $(\Phi(z))^n$, that is, the polynomial obtained if terms of negative degree are dropped from the Laurent series for $(\Phi(z))^n$; the polynomial is then renormalized so that $p(0) = 1$. (This is not the standard normalization.) Our computation is again based on the Schwarz–Christoffel Toolbox. Faber polynomials are not optimal in the sense of (3), but as the figure illustrates, they may come reasonably close, and they are much more tractable analytically. The theory of Faber polynomials is presented in [7], [27], and [65], and among the papers that employ them for designing matrix iterations are [12], [39], and [70].

Figure 8 shows norms as a function of n corresponding to the three plots of Fig. 7. On the log scale, ρ^n is a straight line. The other two curves lie a small distance above that line. For this example, we have $E_n(S) \sim \rho^n$ as $n \to \infty$ for the optimal polynomials but not for the Faber polynomials. For a connected region S with a smooth boundary, the Faber polynomials too would have errors asymptotic to ρ^n. If S had been an interval $[1, \kappa]$ for some $\kappa > 1$, on the other hand, even the optimal polynomials would have exhibited a gap in the limit $\rho \to \infty$: in that case we have $E_n(S) \sim 2\rho^n$ as $n \to \infty$.
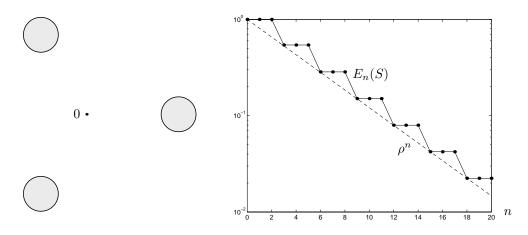
FIG. 9. *On the left, a region S with three-fold symmetry about the origin. On the right, the corresponding plot of minimax norms $E_n(S)$ as a function of n. The sawtooths do not go away as $n \to \infty$; $E_n(S)/\rho^n$ does not converge to 1.*

Another way in which a gap between $\|p\|_S$ and ρ^n may arise is if S has multiple components. In Fig. 9, S consists of three disks located symmetrically with respect to the origin. The effect on the convergence curve is easily deduced. For any S, the minimax polynomial p of (3) is unique. By symmetry, on the other hand, it is clear that if $p(z)$ is a minimax polynomial for this set S, then so are $p(\omega z)$ and $p(\omega^2 z)$, where $\omega = \exp(2\pi i/3)$. We conclude that p must contain only powers $1, z^3, z^6, \ldots$. (Where are the missing point charges in (8) when n is not a multiple of 3? At $z = \infty$.) This implies that the associated errors E_n of (3) must satisfy $E_0 = E_1 = E_2$, $E_3 = E_4 = E_5$, and so on, which explains the staircase pattern in the figure.

In applications, a perfectly symmetrical region like this would be unlikely to appear. (If it did, we would probably find a way to eliminate the symmetry and reduce the dimension.) More general sets S that are disconnected, however, also lead to lack of convergence of E_n/ρ^n as $n \to \infty$. An excellent account of this subject has been given by Widom [82].

Our discussion up to now may seem to suggest that the distinction between $n \to \infty$ and finite n, though interesting mathematically, is not important in practice. However, there is a context where it may be very important: in the effect of "outlier" eigenvalues on the convergence of a matrix iteration.

Suppose we have a matrix whose spectrum is known to lie in a continuum S_0, except for one or more isolated points. This is a common situation in practice. The presence of isolated points in the larger set S increases the quantities $E_n(S)$, in general, compared with what they would be for S_0 alone. On the other hand, according to Theorems 1 and 2, the isolated points have no effect on the asymptotic convergence factor ρ. The ready explanation of this is that any isolated eigenvalue can be annihilated by a single zero of p; the rest of the zeros of p can then be devoted to achieving minimal norm on the rest of S.

This explanation is correct, but it suggests a sharper version of the same idea that is incorrect and, in fact, a common misconception. It is tempting to think that each outlier eigenvalue "costs just one step" in a matrix iteration. This is not true. As a rule, outliers of magnitude bigger than the rest of the spectrum cost about one step, but outliers that are much smaller than the rest of the spectrum may cost many steps. Indeed, the price one must pay to annihilate an outlier grows without bound as it approaches the origin.
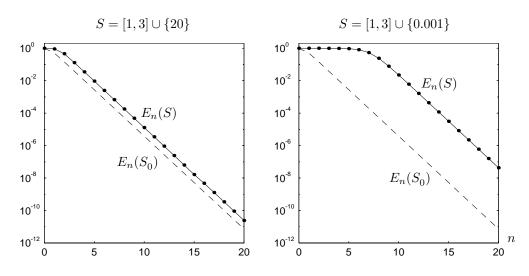
$$S = [1,3] \cup \{20\} \qquad\qquad\qquad S = [1,3] \cup \{0.001\}$$

FIG. 10. $E_n(S)$ and $E_n(S_0)$ vs. n for the sets $S = [1,3] \cup \{20\}$ and $S = [1,3] \cup \{0.001\}$. The outlier near ∞ costs just one iteration, but the outlier near 0 costs about 6 iterations.

Figure 10 illustrates this phenomenon, plotting $E_n(S)$ against n for the sets

$$S \;=\; [1,3] \cup \{20\}, \qquad S \;=\; [1,3] \cup \{0.001\}.$$

In the first case, the outlier is far from the origin, and it delays the convergence by almost exactly one step. In the second case, the outlier is very close to the origin, and it delays the convergence by a little more than six steps. This phenomenon is easily explained. These two outliers can be annihilated by polynomial factors $1 - z/20$ and $1 - 1000z$, respectively. The first of these factors has size ≈ 1 on $S_0 = [1,3]$, but the second has size greater than 1000 there. With $\kappa = 3$ and $\rho = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1) \approx 0.268$ for this problem, by (14), we have $\rho^5 \approx 1.4 \times 10^{-3}$, and this explains why it takes a little more than five extra steps to undo the damage done by the factor $1 - 1000z$, in addition to the sixth step associated with that factor itself.

Step 1 can be summarized as follows. Optimal finite-degree polynomials on a set S may have norms $E_n(S)$ somewhat larger than the estimate ρ^n suggested by potential theory. In the case of outlier eigenvalues of a matrix that lie much closer to the origin than the rest of the spectrum, the discrepancy may be great enough to be of practical importance. Such outliers typically introduce a delay of a number of steps near the beginning of an iteration, before the iteration "gets going."

6. Step 2: Estimated vs. actual spectrum. Our second step is logically trivial but of great importance in practice. Among other things, it can be viewed as responsible for the phenomenon of superlinear convergence often observed in the later stages of a matrix iteration [4], [55], [79], [83]. This is the gap between the estimated spectrum S of the matrix A and its actual spectrum, which we denote by $\Lambda(A)$. Symbolically,

(STEP2) $$\min_{p \in P_n} \|p\|_S \;\gtrsim\; \min_{p \in P_n} \|p\|_{\Lambda(A)}.$$

The mathematics of (STEP2) is immediate. If S is different from the spectrum $\Lambda(A)$, then in general the minimal polynomial norms on these two sets will be different. Depending on the relationship of S and $\Lambda(A)$, the inequality could go in either
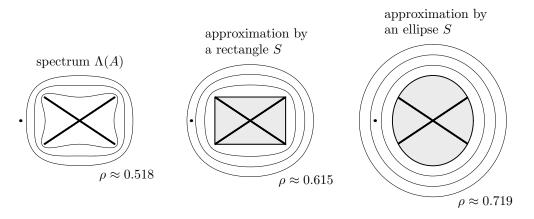
FIG. 11. *Illustration of approximation of the spectrum $\Lambda(A)$ by a simpler region S. Polygons and ellipses are among the regions most often used for such approximations. If S strictly contains $\Lambda(A)$, then the approximation can be expected to increase the asymptotic convergence factor ρ.*

direction. In practice, it is most common that $\Lambda(A)$ is a subset of S, at least approximately, in which case the minimal norms will usually be smaller on $\Lambda(A)$ than on S. This is the meaning (or, rigorously speaking, lack of meaning!) of the expression "$A \underset{\sim}{\geq} B$" in (STEP2) and elsewhere in this paper: $A \geq B$ and $A \leq B$ are both possible, but the former is typical. The symbol "$\underset{\sim}{\leq}$" is used analogously.

The spectrum $\Lambda(A)$ of any matrix A is discrete, but from the beginning of this subject in the 1950s, it has been a standard procedure in some circumstances to approximate $\Lambda(A)$ by a continuum. In a sense this is not even always an oversimplification of the truth. Most large matrices of computational interest arise via discretization of continuous problems. For example, A may be the preconditioned matrix A_h corresponding to a particular choice of a mesh size h for a discretization of a partial differential equation. In such a situation, though the spectrum of A_h is discrete for any fixed choice of h, the spectrum of the family $\{A_h\}$ may fill a continuous region S in $\mathbf{C}$. If one wants bounds on convergence rates that are independent of h, it may be appropriate to work with S rather than with the spectrum of any individual matrix A_h. This notion of the spectrum of a family of matrices is discussed in [2], [28], [61].

A related observation is that Krylov subspace iterations are applicable not just to matrices but to bounded linear operators. This fact has also been appreciated since the 1950s [8], [41] and is considered at an advanced level in the book by Nevanlinna [55].

Let us first consider an example of a general kind. Suppose the matrix or operator A has a spectrum $\Lambda(A)$ that fills the cross in the complex plane displayed in the first plot of Fig. 11. We might naturally look for the norms $\{E_n(\Lambda(A))\}$ and convergence factor ρ associated with this region. Some equipotential curves corresponding to the appropriate conformal map are shown in the figure; the convergence factor is $\rho \approx 0.518$.

Now suppose that for one reason or another, we decide to approximate the cross by a simpler region $S \subseteq \mathbf{C}$. Typically we might enclose $\Lambda(A)$ by S, in which case the "$\underset{\sim}{\geq}$" of (STEP2) becomes a "$\geq$." One choice for S might be the enclosing rectangle illustrated in the middle plot. For this region, the asymptotic convergence factor worsens by 35% (as measured by logarithms), to $\rho \approx 0.615$. Another choice might be the ellipse illustrated in the third plot. Here the asymptotic convergence factor worsens by an additional 47%, to $\rho \approx 0.719$.

In the study of matrix iterations, polygons and ellipses arise commonly as estimated regions S. For example, there is a large class of *adaptive* or *hybrid iterations* that proceed by first estimating the spectrum $\Lambda(A)$ by some set S, then iterating in some fashion based on that approximation S; sometimes the estimate S is refined as the iteration proceeds. The first such method, proposed by Manteuffel in the 1970s, took S to be an ellipse, which has the advantage that an optimal iteration for S can be derived analytically from Chebyshev polynomials [51]. Subsequent refinements on this idea have often taken S to be a polygon, which some have suggested should be transformed explicitly by a numerical Schwarz–Christoffel map [16], [52], [67], [70]. For a summary of the various strategies that had been proposed up to 1992, see [54]. Thus Fig. 11 is very much in the spirit of practical matrix computations, and it illustrates the price that may be paid by overestimating the spectrum. In more extreme cases, where the convex hull of $\Lambda(A)$ contains the origin, a nonconvex estimate S will certainly be needed or there will be no convergence at all.

Hybrid matrix iterations are also related to the idea of *polynomial preconditioning* of a problem $Ax = b$ by a polynomial $p(A)$ with coefficients derived from spectral or other information [1], [21], [63].

Now let us switch tacks. Instead of generalities about approximating a spectrum $\Lambda(A)$ by a region S, let us turn to the specific context in which this issue arises most often, alluded to at the beginning of this section. This is the business of approximating a discrete spectrum by a continuum.

Such approximations are very familiar. Indeed, perhaps the most famous result in matrix iterations is that the conjugate gradient iteration converges in $O(\sqrt{\kappa})$ steps to a fixed precision, as mentioned earlier in connection with (14). Such a statement may be sharp for an operator or a family of matrices, but for an individual matrix, it is necessarily an approximation. It will be sharp only insofar as $\Lambda(A)$ "fills the interval" by which it is approximated.

Typically, a Krylov subspace iteration behaves as if the spectrum is continuous in the early stages, then begins to accelerate in the later stages. There is no mystery about this behavior. The iteration depends upon polynomials $p \in P_n$ that are small throughout the spectrum $\Lambda(A)$. When n is small, there may be no way to achieve this except by being small throughout a wide neighborhood of $\Lambda(A)$, i.e., a continuum. As n increases, however, there may be enough degrees of freedom for p to take the strategy of approximately annihilating certain eigenvalues of A while being larger in-between. This phenomenon is related to the convergence of approximations to eigenvalues of A computed by iterative methods for eigenvalue problems, notably the Lanczos and Arnoldi iterations [6], [29], [62].

Complicated examples can generate very complicated behavior. We shall content ourselves with just one simple example. Figure 12 shows E_n as a function of n for the set $\Lambda(A) = \{1, 2, 3, \ldots, 100\}$ and the associated continuum approximation $S = [1, 100]$. The upper curve shows the norms $E_n(S)$, and the lower, downward-bending curve corresponds to the norms $E_n(\Lambda(A))$. At first, there is little difference between $E_n(S)$ and $E_n(\Lambda(A))$. After 20 or so iterations, however, the polynomial degree becomes high enough that some of the discrete points of $\Lambda(A)$ begin to be resolved. Individual eigenvalues begin to be annihilated in the fashion of outliers; the iteration is free to concentrate its efforts on a smaller subset of $[1, 100]$. The convergence accelerates, and at step 100, off the scale of this plot, $E_n(\Lambda(A))$ will be exactly zero.

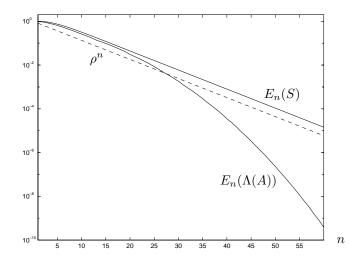Plots analogous to Fig. 12 have been presented previously in [30].

FIG. 12. *Illustration of superlinear convergence in the later stages of an iteration, as the iteration begins to take advantage of the fact that the spectrum is not continuous but discrete. Here $S = [1, 100]$ and $\Lambda(A) = \{1, 2, 3, \dots, 100\}$.*

We must point out that given a spectrum $\Lambda(A)$, there is no unique choice of a set S that can be used to approximate it. In the example of Fig. 12, for example, suppose we were to take $S = \{1\} \cup \{2\} \cup \{3\} \cup [4, 100]$ instead of $S = [1, 100]$. This would not affect the curve of $E_n(\Lambda(A))$, but it would affect how we interpret it. We would now describe some of the downward bend of this curve as caused by the finite n phenomenon of Fig. 10 rather than by the gap between S and $\Lambda(A)$.

7. Step 3: Normal vs. nonnormal A. A matrix is *normal* if it has a complete set of orthogonal eigenvectors, or equivalently, if it can be unitarily diagonalized: $A = UDU^*$, where U is unitary and $D = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$. From this formula it follows that for a normal matrix, $\|p(A)\|$ is equal to the maximum of $|p(z)|$ on the spectrum of $\Lambda(A)$. For general matrices, however, all we have is the inequality

(STEP3) $$\min_{p \in P_n} \|p(z)\|_{\Lambda(A)} \ \le \ \min_{p \in P_n} \|p(A)\|.$$

This is the third in our sequence of approximations.

The degree of nonnormality of a matrix—a notion we shall not attempt to define precisely—can vary arbitrarily. In particular, most estimates such as (STEP3) that are based on the eigenvalues of a matrix can fail to be sharp to an arbitrary degree, for certain matrices [75]. As an elementary example, consider the norms of powers $\|A^n\|$ for an $N \times N$ Jordan matrix of the form

(15) $$A \ = \ \begin{pmatrix} 0 & 2 & & & \\ & 0 & 2 & & \\ & & 0 & 2 & \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}.$$

For $n < N$ we have $\|A^n\| = 2^n$, even though z^n is identically zero on the spectrum of A (the origin). Nor does this discrepancy depend on A being nondiagonalizable.
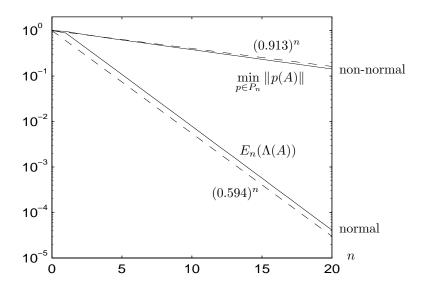
FIG. 13. *Convergence curves for the tridiagonal Toeplitz matrix* (16) *and for a normal matrix with the same eigenvalues. Nonnormality slows down the convergence by about a factor of five.*

For example, if we replace the zeros on the diagonal of (15) by distinct numbers of magnitude less than 1, the matrix becomes diagonalizable, but $\|A^n\|$ still grows exponentially for $n < N$, while $\|z^n\|_{\Lambda(A)}$ shrinks exponentially.

Over the years, analysis of eigenvalues has been the basis of most studies of convergence of matrix iterations. The justification for this kind of analysis has been that although non-normality may introduce certain transient effects, the asymptotic behavior as $n \to \infty$ of a function such as A^n or $p(A)$ is determined by eigenvalues. This reasoning is flawed, however, since matrix iterations are concerned with $n \ll N$. A better explanation of the success of eigenvalue analysis in practice is the fact that in most applications, the degree of nonnormality is mild.

For another example where it is not mild, consider the matrix of the form

$$
(16) \qquad A \;=\; \begin{pmatrix} 0.5 & 0.7 & & & \\ -0.3 & 0.5 & 0.7 & & \\ & -0.3 & 0.5 & 0.7 & \\ & & -0.3 & 0.5 & 0.7 \\ & & & -0.3 & 0.5 \end{pmatrix}
$$

of dimension $N = 50$. Since the entries are constant along diagonals, A is, like (15), a *Toeplitz matrix*. Its eigenvalues are a set of 50 distinct numbers in the complex interval $[\frac{1}{2} - 0.917i, \frac{1}{2} + 0.917i]$. A computation based on Theorem 2 reveals that the estimated asymptotic convergence factor associated with this interval is $\rho \approx 0.594$. Figure 13, however, shows that the actual convergence of minimal polynomials for this matrix is about five times slower, at a rate approximately $(0.913)^n$. Only when we plot the convergence curve for a normal matrix with the same eigenvalues do we see behavior like $(0.594)^n$.

An explanation of this behavior is suggested in Fig. 14. The spectrum of a matrix, $\Lambda(A)$, is the set of points in $\mathbf{C}$ where the *resolvent* of A, $(zI - A)^{-1}$, does not exist. Let us adopt the convention that the norm of the resolvent at such points
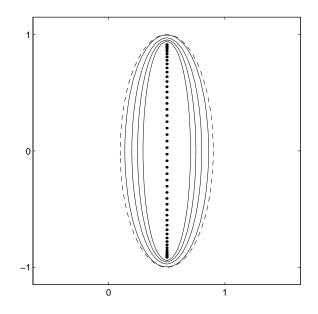
FIG. 14. *Partial explanation of Fig. 13. The solid dots are the eigenvalues of A, filling an interval that corresponds to $\rho \approx 0.594$. The dashed ellipse is the image of the unit circle under the symbol $-0.3w^{-1} + 0.5 + 0.7w$ of this Toeplitz matrix, corresponding to $\rho \approx 0.913$. The other curves are the boundaries of the ϵ-pseudospectra $\Lambda_\epsilon(A)$ for $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$.*

is infinite; $\|(zI - A)^{-1}\| = \infty$. The matrix (16) has the property that in regions of $\mathbf{C}$ where $\|(zI - A)^{-1}\|$ is finite, it may still be exceedingly large. For each $\epsilon \geq 0$, the ϵ-*pseudospectrum* of A is the subset of the complex plane defined by

$$(17) \qquad \Lambda_\epsilon(A) = \{ z \in \mathbf{C} : \|(zI - A)^{-1}\| \geq \epsilon^{-1} \}.$$

The matrix (16) has ϵ-pseudospectra that extend far beyond the spectrum, even when ϵ is very small [60]. For $\epsilon = 10^{-5}$, for example, $\Lambda_\epsilon(A)$ extends a distance about 0.4 away from $\Lambda(A)$. The pseudospectra of a matrix can also be defined in another equivalent fashion:

$$(18) \qquad \Lambda_\epsilon(A) = \{ z \in \mathbf{C} : z \in \Lambda(A + E) \text{ for some } E \text{ with } \|E\| \leq \epsilon\}.$$

Thus in Fig. 14, a perturbation of norm 10^{-5} could move some of the eigenvalues a distance about 0.4.

Banded Toeplitz matrices such as (16) represent extreme examples of nonnormality in the sense that the ϵ-pseudospectra depend only weakly on ϵ. For a different kind of example, consider the 30×30 matrix of the form

$$(19) \qquad A = \begin{pmatrix} 1 & \alpha & & & \\ & 1 & \alpha/2 & & \\ & & 1 & \alpha/3 & \\ & & & 1 & \ddots \\ & & & & \ddots \end{pmatrix},$$

where α is a parameter. The spectrum of this matrix is the single point $\{1\}$, corresponding to an estimated asymptotic convergence factor $\rho = 0$. Figure 15 shows,
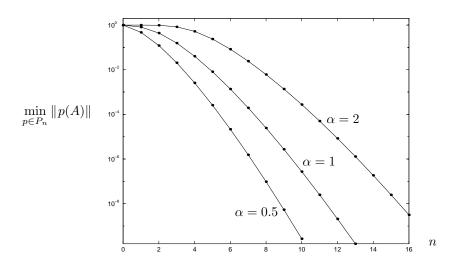
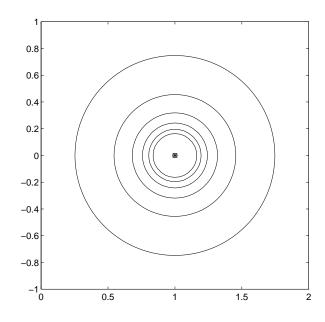FIG. 15. *Convergence curves for the matrix* (19) *for three values of* α.

FIG. 16. *Boundaries of the ϵ-pseudospectra of the matrix* (19) *with $\alpha = 1$ for $\epsilon = 10^{-0.5}, 10^{-1}$, $10^{-1.5}, \ldots, 10^{-3}$. The pseudospectra shrink significantly with ϵ.*

however, that a Krylov subspace iteration for this matrix takes many steps to converge, starting slowly and then accelerating. Some insight into this behavior is suggested by the ϵ-pseudospectra of Fig. 16, which shrink substantially as ϵ decreases.

A connection can be made between pseudospectra and convergence of Krylov subspace iterations via the contour integral $p(A) = (1/2\pi i) \int_\Gamma (zI - A)^{-1} p(z) \, dz$, valid for any polynomial p and any simple closed integration contour Γ enclosing $\Lambda(A)$ [53], [71], [75]. If $p(z)$ is small on a region where $\|(zI - A)^{-1}\|$ is not too large, then taking absolute values in this formula may give a reasonable bound on $\|p(A)\|$. For example,

if Γ is taken as the boundary of $\Lambda_\epsilon(A)$ for some ϵ, we get

$$(20) \qquad \|p(A)\| \ \leq \ \frac{1}{2\pi} L \epsilon^{-1} \|p(z)\|_{\Lambda_\epsilon(A)},$$

where L is the arc length of the boundary of $\Lambda_\epsilon(A)$. This estimate suggests that if A is far from normal, then the convergence of a Krylov subspace iteration may be determined approximately by the size of minimal polynomials on the pseudospectra rather than the spectra. Often this approximation is about right, though not always, as (20) is only an inequality. What is clear is that in extreme cases of nonnormality, the spectrum may give no useful information. A general theorem to this effect has been proved by Greenbaum, Ptak, and Strakoš [36].

A user who suspects that a matrix iteration is converging slowly because of nonnormality may investigate this possibility in various ways. Computation of pseudospectra by the obvious method of evaluating resolvent norms on a grid may be too expensive, but several acceleration devices are available, including projection onto a lower-dimensional invariant subspace [73], preliminary reduction to triangular form followed by inverse or Lanczos iteration [50], or the use of Arnoldi iteration [73]. Speedups by factors of hundreds may sometimes be obtained by applying these devices in combination. An alternative approach is to perturb the matrix A at random, say, by a perturbation of relative norm 10^{-3}, and see if the effect on the eigenvalues is pronounced. If the dominant eigenvalues prove to be highly sensitive to perturbations, there is a risk that analysis of convergence based on eigenvalues alone may be misleading.

8. Step 4: $\|p(A)\|$ vs. $\|p(A)r_0\|/\|r_0\|$. Krylov subspace iterations work with vectors generated from the initial vector r_0, not with matrices. Indeed, it is this property that makes them so powerful. Yet the standard ideas for investigating convergence are matrix ones, independent of r_0. This is our fourth step:

$$(\text{STEP4}) \qquad \min_{p \in P_n} \|p(A)\| \ \geq \ \min_{p \in P_n} \frac{\|p(A)r_0\|}{\|r_0\|}.$$

We shall see that this inequality is responsible for the phenomenon of sublinear convergence often observed in the early stages of a matrix iteration [55].

The mathematics of (STEP4) is elementary. From the definition of the matrix norm, we have $\|p(A)\| \geq \|p(A)r_0\|/\|r_0\|$, for any $r_0 \neq 0$, and (STEP4) follows by minimizing over $p \in P_n$. We have written "min" instead of "inf," for it can be shown that for any A and n, there exists a unique polynomial $p \in P_n$ that minimizes $\|p(A)\|$, known as the *ideal GMRES polynomial* for the given matrix A and step n [38]. Such polynomials can be computed numerically by methods of semidefinite programming [71,74], and computations of this kind are the basis of many of the figures of this paper.

Krylov subspace iterations approximately minimize the right-hand side of (STEP4), not the left-hand side. Accordingly, to understand their behavior, we need to know how these two may differ. An important part of the answer can be stated immediately. *To achieve a reduction in $\|r_0\|$, a matrix $p(A)$ need not reduce all of the eigencomponents of r_0, just some of them.*

An example makes the point. In Fig. 17, polynomials have been computed corresponding to the 202×202 diagonal matrix

$$(21) \qquad A \ = \ \mathrm{diag}(1.00, 1.01, 1.02, \ldots, 2.99, 3.00, \quad 0.001).$$
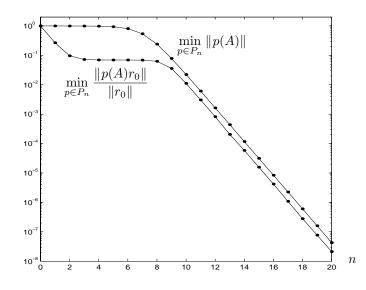
FIG. 17. *The difference between minimizing* $\|p(A)\|$ *and minimizing* $\|p(A)r_0\|/\|r_0\|$ *for the matrix* (21) *and a random starting vector* r_0. *The outlier eigenvalue causes initial stagnation in the upper curve, as illustrated already in Fig.* 10(b). *It has negligible effect at first on the lower curve, however, since* $\|p(A)r_0\|/\|r_0\|$ *can be reduced considerably by polynomials p that pay no attention to the outlier. Eventually, further progress requires attention to the outlier, and the stagnation phase appears.*

This set of eigenvalues is densely distributed in $[1,3]$, with an additional outlier at 0.001; the spectrum is a finite analogue of the set $[1,3] \cup \{0.001\}$ considered in Fig. 10(b). The upper curve of Fig. 17, corresponding to the left-hand side of (STEP4), is essentially the same as the upper curve in Fig. 10(b). The lower curve corresponds to the right-hand side of (STEP4) and a random choice of r_0.

Here is what is going on in Fig. 17. An initial random r_0 with $\|r_0\| = 1$ has roughly equal "energy" in each of the 202 eigencomponents. By this we mean that if r_0 is expanded in the basis of eigenvectors (which in this case happens to be the canonical basis $\{e_j\}$, since A is diagonal), then a typical expansion coefficient is of order $(202)^{-1/2}$. Now suppose a polynomial p is found that approximately annihilates 201 of these 202 components, without changing very much the component corresponding to the outlier. The result is that the norm is reduced from 1 to order $(202)^{-1/2}$. This is just what we see in the figure. At first, polynomials $p \in P_n$ are found that are small on $[1,3]$ but approximately 1 at 0.001, and this suffices to bring the norm down by an order of magnitude. Then the iteration stagnates; further progress cannot be achieved without taking the outlier into consideration. Eventually n is large enough that polynomials $p \in P_n$ can be found that are small at the outlier too; and from here on, the convergence proceeds essentially as in Fig. 10(b).

The reader may wonder, why did we use a continuum $[1,3]$ for Fig. 10(b) but a discrete set of 201 points in that interval for Fig. 17? The answer is that the effect just illustrated does not make sense in the continuum limit. For an $N \times N$ matrix, we expect a coefficient approximately $N^{-1/2}$ associated with the outlier eigenvalue, making the level of the plateau in the lower curve of Fig. 17 approximately $N^{-1/2}$. In the limit $N \to \infty$, the outlier would have a zero proportion of the total energy and the appearance of the plateau would be deferred forever.

Continuum limits make sense in this context, however, so long as we do not try to balance intervals against isolated points of finite multiplicity. Fig. 18 shows an
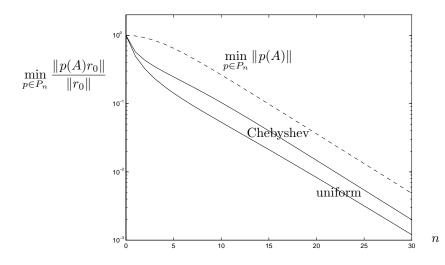
FIG. 18. *The difference between minimizing $\|p(A)\|$ and minimizing $\|p(A)r_0\|/\|r_0\|$ for the spectrum $[0.01, 1]$ and a starting vector r_0 with equal energy in all eigencomponents. Faster initial convergence is achieved with a uniform distribution of eigenvalues in that interval than with a Chebyshev distribution, though the behavior as $n \to \infty$ is the same.*

example. Here, the spectrum of A fills the interval $[0.01, 1]$ continuously. The upper curve shows the corresponding minimal norms $\|p(A)\|$, computable analytically for this simple spectrum. The lower curves show the minimal norms $\|p(A)r_0\|/\|r_0\|$ for two particular choices of initial vector r_0. In one case, the lowest curve, the energy in r_0 is uniformly distributed throughout the interval $[0.01, 1]$. In the second case, the middle curve, it is distributed according to a Chebyshev distribution, i.e., with density of type $(1 - x^2)^{-1/2}$ rescaled appropriately to $[0.01, 1]$. (Both of these curves were computed via discretizations of the interval involving 500 points.) The curve corresponding to Chebyshev points is higher than the other one because Chebyshev points are more densely clustered near the boundary $x = 0.01$ than uniform points, forcing polynomials p to take greater pains to be small near there.

Figure 18 can also be interpreted as showing results (with probability 1) for random initial vectors r_0 but for two matrices A, again in the continuum limit $N \to \infty$. In one case the spectrum is uniformly distributed in $[0.01, 1]$, and in the other it lies in a Chebyshev distribution. To put it concisely, when it comes to minimizing $\|p(A)r_0\|/\|r_0\|$ for random vectors r_0, as opposed to minimizing $\|p(A)\|$, *multiplicities matter.* More generally we may say that even when the eigenvalues effectively fill a certain region of the complex plane, the density with which they do so may affect the rate of convergence.

Sometimes algorithms are designed to take advantage of this fact. An example is presented in a recent paper of Fischer and Freund [21], where the density distribution of the eigenvalues is explicitly estimated in order to design a polynomial preconditioner for a hermitian positive definite matrix. Actually, many of the adaptive and hybrid methods alluded to in the discussion of Step 2 have the property that they adapt to densities as well as locations of eigenvalues, though in most cases this is a fortunate accident rather than a part of the authors' design concept [54].

Figures 19 and 20 show another set of curves illustrating the significance of eigenvalue multiplicities in Krylov subspace iterations. Here, the spectrum is in the shape of a cross. If the eigenvalues are distributed uniformly on the boundary, the convergence is slower than if they are distributed uniformly in the interior. We used 670
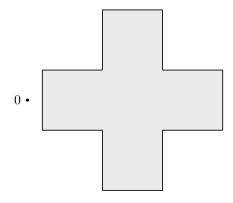
FIG. 19. *A cross located in the right half-plane. For the results of potential theory, it does not matter whether the region is the boundary alone or the boundary together with the interior.*
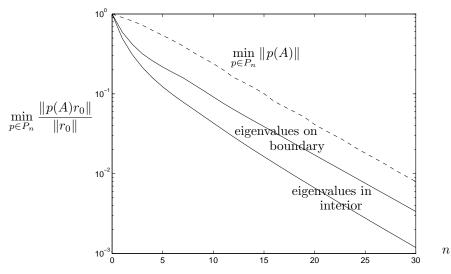

FIG. 20. *The minimal value of $\|p(A)r_0\|/\|r_0\|$ for the problem of Fig.* 19, *however, depends significantly on how the eigenvalues are distributed within the cross.*

and 3000 uniformly distributed discrete points, respectively, for these computations, but the curves would have been little different if these numbers had been infinite.

The reader may suspect that in the last few pages, we have really been discussing the distinction between the ∞-norm and a weighted 2-norm of polynomials p over a set S. If A is normal, this is true, for in that case $\|p(A)\| = \|p\|_{\Lambda(A)}$, whereas $\|p(A)r_0\|/\|r_0\|$ is given by the formula

$$(22) \qquad \frac{\|p(A)r_0\|}{\|r_0\|} = \left(\sum_{j=1}^{N} |\alpha_j|^2 |p(\lambda_j)|^2 \right)^{1/2} \bigg/ \left(\sum_{j=1}^{N} |\alpha_j|^2 \right)^{1/2},$$

where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A and α_j is the expansion coefficient of r_0 with respect to the jth eigenvector. If all of these coefficients are equal, which will be approximately true if r_0 is random, then the norm reduces to the unweighted 2-norm of $p(z)$ on $\Lambda(A)$, implying that the ratio of $\|p(A)\|$ to $\|p(A)r_0\|/\|r_0\|$ can be no

greater than $\sqrt{N}$. This observation applies to the initial stages of Figs. 18 and 20, for example, since the matrices involved in those examples were normal.

If A is normal but r_0 is not random, (22) indicates that $\|p(A)\|$ and $\|p(A)r_0\|/\|r_0\|$ may differ in other ways. A favorable r_0 might make $\|p(A)r_0\|/\|r_0\|$ arbitrarily small. If r_0 is a multiple of a single eigenvector of A, for example, then convergence may occur in one step. It can be shown that the worst-case r_0, by contrast, having all its energy appropriately concentrated at points where the minimax polynomial $p(z)$ on $\Lambda(A)$ achieves its maximum, eliminates the gap between $\|p(A)\|$ and $\|p(A)r_0\|/\|r_0\|$ entirely, making (STEP4) an equality [35], [46].

Does (STEP4) become an equality even for nonnormal matrices A, for some worst-case choice of r_0? Greenbaum and Trefethen conjectured in [38] that the answer was yes. Subsequently, a counterexample was found by Faber et al., showing that it is in general no [18]. If we maximize over r_0 in (STEP4), we get the inequality

$$(23) \qquad \min_{p \in P_n} \max_{r_0 \neq 0} \frac{\|p(A)r_0\|}{\|r_0\|} \ \geq \ \max_{r_0 \neq 0} \min_{p \in P_n} \frac{\|p(A)r_0\|}{\|r_0\|}.$$

The example of Faber et al. established that the inequality is sometimes strict; the two quantifiers do not commute. More recently, Toh has shown that for certain matrices, the ratio of the right-hand side to the left-hand side can be arbitrarily small [72].

The lack of equality in (23) never seems to be an issue in practice. Experiments with various matrices indicate that although the ratio of the two sides might be 10^{-1} or 10^{-6} in principle, it is usually more like 0.99 or exactly 1 [71]. Thus, in practice, it is reasonable to think of Step 4 as the difference between worst-case and actual initial vectors r_0, even when A is not normal. In some applications, for example with some linear iterations imbedded in larger nonlinear calculations, initial guesses will be available that are significantly better than random, making Step 4 especially significant.

9. Step 5: Minimization vs. quasi minimization. At step n, a typical Krylov subspace iteration for solving $Ax = b$ constructs an iterate x_n whose residual r_n lies in the nth Krylov subspace generated by the initial residual r_0, as indicated in (6): $r_n \in r_0 + \langle Ar_0, A^2r_0, \ldots, A^nr_0 \rangle$. Such an x_n satisfies (7) for some $p \in P_n$: $r_n = p(A)r_0$. But which such x_n do these iterations construct? The simplest principle is that it should be the (unique) choice that minimizes $\|r_n\|$, where $\|\cdot\|$ is, as always, the 2-norm. Some methods achieve this, but some do not, and this is our fifth step:

$$(\text{STEP5}) \qquad \min_{p \in P_n} \frac{\|p(A)r_0\|}{\|r_0\|} \ \leq \ \frac{\|p(A)r_0\|}{\|r_0\|}.$$

The standard algorithm that minimizes $\|r_n\|$ for general matrices is GMRES [64]. The algorithm constructs a sequence of orthonormal vectors $\{q_k\}$ that span the successive Krylov subspaces $\langle r_0, Ar_0, \ldots, A^{n-1}r_0 \rangle$ and then solves the minimization problem, involving a Hessenberg matrix, by standard orthogonal matrix operations for least-squares problems. All this seems straightforward now, but it was not so obvious a generation ago; GMRES was not developed until the 1980s. Before that, other methods were published that minimized $\|r_n\|$, notably GCR, ORTHOMIN, and ORTHODIR, but they had restrictions concerning the classes of applicable matrices and numerical stability.

There are two reasons why $\|r_n\|$ may not be minimized at step n. The first is that some methods minimize something different from $\|r_n\|$. In the conjugate gradient

(CG) iteration [43], for example, which is restricted to symmetric positive definite matrices A, it is the A-norm of the error that is minimized: $\|e_n\|_A = (e_n^T A e_n)^{1/2}$, where $e_n = x_n - A^{-1}b$. (The difference is visible in the lower curves of Figs. 22 and 23, below.) This makes sense for symmetric positive definite matrices but not for general ones. Alternatively, $\|r_n\|$ may be minimized but over a different space than the standard Krylov subspace (6). The method known as CGNR, for example, which consists of the application of CG to the normal equations $A^*Ax = A^*b$, makes use of a Krylov subspace based on A^*A instead of A. (In such cases (STEP5) is not a correct description of the nature of the approximation under consideration, since the actual r_n is not of the form $p(A)r_0$.)

Variations of this kind may significantly affect the convergence rate of an iteration. In the case of a change of norm, say from $\|r_n\|$ to $\|e_n\|_A$, the effect is typically bounded by a number such as $\sqrt{N}$ or $\sqrt{\kappa(A)}$. In the case of a change of Krylov subspace, it may be unbounded. For example, for some $N \times N$ matrices CGNR converges in one step while GMRES takes N steps; for others, GMRES converges in two steps, while CGNR takes N steps [53].

The other reason why $\|r_n\|$ is not always minimized is more interesting, for it touches a deeper issue and is a topic of current concern and imperfect understanding. For some problems, each iterate of GMRES is dauntingly expensive to compute. Consequently, other methods have been devised that require much less work per iteration, but at the cost of delivering potentially larger residuals r_n. Two classic examples that employ the Krylov subspace (6) are biconjugate gradients (BCG) [23], [48] and quasi-minimal residuals (QMR) [25], [26]. Other methods employing different Krylov subspaces include conjugate gradients squared (CGS) [68], Bi-CGSTAB [80], and transpose-free QMR (TFQMR) [24]. The behavior of these "quasi-minimization" methods is generally more complicated and less well understood than that of GMRES, but they are very fast and very popular.

What makes GMRES expensive is that to compute the new iterate x_n at step n, it manipulates a recurrence relation involving $n + 1$ vectors, potentially requiring a great deal of work and memory. This is no problem at the beginning of an iteration, when n is small, and thus GMRES is often the best method for problems that are well enough preconditioned that they converge fast. It may also be no problem, relatively speaking, in cases where matrix-vector multiplications of the form Ax are so costly as to dominate all the costs of manipulating $n + 1$ vectors at step n. For many applications, however, these conditions do not hold, and it is safe to say that the cost of GMRES typically begins to be distressing by $n = 20$ and prohibitive by $n = 100$.

One way of coping with this problem is to restart GMRES intermittently, throwing away the information accumulated in the Krylov subspace and retaining only the current iterate x_n. The other, deeper idea is to employ alternative algorithms based on three-term recurrence relations such as those mentioned above. Typically such an algorithm employs two Krylov sequences rather than one, making use of the transpose A^T or adjoint A^* as well as A itself, and enforces a biorthogonalization condition that can be implemented by three-term recurrences. However, no known methods of this type minimize $\|r_n\|$. Indeed, in response to a $100 challenge posed by Golub in the early 1980s, it was proved by Faber and Manteuffel in 1984 that within a certain class of methods based on a three-term recurrence, there is none that minimizes $\|r_n\|$ for general problems $Ax = b$ [19].

The field of quasi-minimization algorithms is very large, and we shall make no attempt to survey it. Instead, we present an example to give some flavor of the typical performance of these methods. Figure 21 shows convergence curves corresponding to
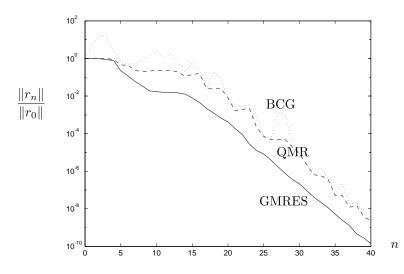
FIG. 21. *Comparison of GMRES (residual norm minimization) with QMR and BCG (quasi minimization) for a 201×201 matrix and a random initial vector r_0. The GMRES convergence curve is optimal in the sense of minimizing $\|r_n\|$ at each step, but the other methods require less work per step.*

GMRES, QMR, and BCG applied to a certain 201×201 matrix A with a random initial vector r_0. To construct A, we began with the diagonal matrix D with 200 evenly spaced entries from 1 to 9. This was modified by a similarity transformation by a random 200×200 matrix, and then a 201st row and column were added consisting of all zeros except for the entry $1/2$ on the diagonal. All told, A is thus a modestly nonnormal 201×201 matrix with spectrum in $[1, 9]$ except for an outlier at $1/2$.

The main feature to note in Fig. 21 is that the GMRES convergence curve represents a lower bound, with the QMR convergence curve lying above it and the BCG curve lying mainly, though not entirely, above that. This behavior is typical. It is also noteworthy that whereas the GMRES curve is monotonically nonincreasing, the other curves are not, though the QMR curve comes close. For practical problems the convergence of BCG is often far more erratic than this, and other methods that may be faster, such as CGS, may converge more erratically still. An analysis of the erratic nature of these convergence curves is presented in [5]; some of the mathematics involved appeared earlier in [24].

The reader may find it an interesting exercise to explain the shape of the GMRES curve of Fig. 21 based on the previous sections of this paper. The initial stagnation is due to nonnormality (Step 3). The rapid convergence around $n = 5$ is made possible by the difference between vector and matrix norms (Step 4). The stagnation around $n = 10$ is the finite n effect caused by the outlier eigenvalue (Step 1). Finally, around $n = 15$, we enter a phase of steady linear convergence at the rate $\rho^n = 2^{-n}$ appropriate to the region $S_0 = [1, 9]$. If the computation were carried a bit further, we would begin to observe acceleration caused by the discreteness of the spectrum (Step 2).

10. Step 6: Exact vs. floating-point arithmetic. Our final approximation is the one that arises throughout scientific computing. On a computer, a Krylov subspace iteration will be implemented in floating-point arithmetic, and rounding errors will be introduced. The result is that r_n will not be exactly equal to $p(A)r_0$ for any polynomial $p \in P_n$. In principle, the error could go in either direction, but

for statistical reasons, rounding errors will almost always hurt rather than help, and thus we write this step with the symbol $\lesssim$:

(STEP6)
$$\frac{\|p(A)r_0\|}{\|r_0\|} \lesssim \frac{\|r_n\|}{\|r_0\|}.$$

Rounding errors play an important role in matrix iterations. Of course, one effect they have is the obvious one of preventing convergence of $\|r_n\|/\|r_0\|$ below the level of machine precision, or sometimes machine precision times the condition number of A. But they have another effect that is different from this, which might be summed up by the idea that three-term recurrences executed in floating-point arithmetic suffer from *loss of memory.*

Methods involving three-term recurrences invariably depend upon certain orthogonality properties. For example, the CG and Lanczos iterations produce a sequence of residuals and search directions that are orthogonal in various senses, and the BCG and QMR iterations produce pairs of sequences of vectors that are mutually biorthogonal. Mathematically speaking, these orthogonality properties are consequences of identities based on, for example, the symmetry of A in the case of CG or Lanczos. In floating-point arithmetic, however, these properties typically cease to hold accurately after a number of steps have been taken. Nearby vectors in a sequence may be nearly orthogonal to one another, but they are typically far from orthogonal to vectors earlier in the sequence.

At a qualitative level, it is easy to see what causes this loss of orthogonality. Small errors are introduced at each step of a recurrence, and these may then grow exponentially as the iteration proceeds. One may take the view, for example, that the component of r_n corresponding to a particular eigenvalue of A is never exactly annihilated by a root of a polynomial $p(A)$; at best it is annihilated to machine precision. The residue that remains may grow in succeeding iterations until it becomes very large, and eventually it may have to be annihilated again. Such effects may delay convergence of matrix iterations by a number of steps. In the case of Lanczos iterations for finding eigenvalues (not discussed here), they result in the phenomenon of multiple computed copies of eigenvalues known as "ghosts." Note that such effects result from having isolated points in the spectrum; for a linear operator whose spectrum has no isolated points, or for a matrix at an early enough stage in the iteration that the spectrum does not behave as if it has isolated points, ghosts and related phenomena do not appear.

Establishing theorems to quantify these effects is not an easy matter. In some areas, a good deal is known; examples are Paige's analysis of loss of orthogonality in the Lanczos iteration [57] and Greenbaum's analysis of the convergence of conjugate gradients [31]. Another example is a result of Greenbaum to the effect that if an iteration is erratic enough that some iterates have norms $\|x_n\|$ much larger than the exact norm $\|A^{-1}b\|$, then the attainable accuracy suffers accordingly [33]. Overall, however, the completeness of the understanding of the effect of rounding errors on Krylov subspace iterations falls short of what numerical analysts have come to expect for the more classical direct methods in linear algebra [29, 44]. Readers interested in learning something of this subject should begin with Greenbaum's book [34].

Figure 22 illustrates a typical effect of rounding errors in a three-term recurrence. The matrix, of dimension $N = 24$, is taken from [37]. We start with $D = \text{diag}(\lambda_1, \ldots, \lambda_N)$, where $\lambda_j = 0.001 + 0.999(j - 1)2^{j-N}/(N - 1)$ for $j = 1, 2, \ldots, N$. We then set $A = UDU^T$, where U is a random orthogonal matrix. The matrix A is
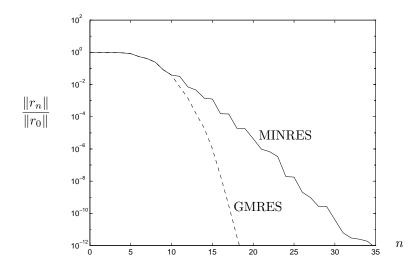
FIG. 22. *Comparison of MINRES (three-term recurrence) and GMRES ((n + 1)-term recurrence) for a* 24×24 *symmetric positive definite matrix. The curves would be identical in exact arithmetic.*

thus dense and symmetric, with eigenvalues lying in the interval $[0.001, 1]$ but very strongly clustered near 0.001.

Figure 22 shows two convergence curves for this matrix A and a random initial vector r_0. The GMRES curve is not significantly affected by rounding errors. (Exercise: why does it have the shape it does?) Since A is symmetric, however, each residual r_n is computable by a three-term recurrence instead of an $(n + 1)$-term recurrence, and a standard algorithm for this is known as MINRES [58]. The figure shows that although the MINRES curve would be identical to the GMRES curve in exact arithmetic, in floating-point arithmetic it begins to deviate very early. (The computer has machine epsilon $2^{-52} \approx 2.2 \times 10^{-16}$.) Convergence to high accuracy is not prevented, but it is substantially delayed.

Figure 23 presents the same comparison for a different pair of algorithms. One curve corresponds to CG as implemented in the usual way by a three-term recurrence. Note that since $\|r_n\|$ is plotted, whereas it is $\|e_n\|_A$ that CG minimizes, the convergence curve does not decrease monotonically. What is striking is that after step $n = 10$, the behavior begins to be strongly affected by rounding errors. The comparison curve labeled "exact CG" was computed by an $(n + 1)$-term recurrence, following [37], so that the orthogonality property of CG was maintained explicitly.

11. Summary. Our six steps represent a chain of inequalities stretching from ρ^n to $\|r_n\|/\|r_0\|$, the two sides of (5). They are collected in Table 1.

Obviously, the inequalities do not all point in the same direction. The convergence of a matrix iteration can be faster or slower than the estimate ρ^n based on potential theory, depending on what effects are dominant.

Here is a brief recapitulation of the principal effect of each of the steps.

Step 1 is the difference between nth root $n \to \infty$ asymptotics and finite n potential theory for a set S. If S has isolated points near the origin, corresponding to outlier eigenvalues of A, these may delay convergence by many steps without affecting the asymptotic rate.
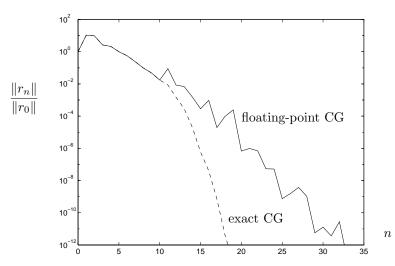
FIG. 23. *For the same matrix as in Fig.* 22, *comparison of CG implemented in floating-point and in exact arithmetic. The exact CG curve is simulated by implementing CG with an explicit* $(n+1)$*-term recurrence.*

TABLE 1
Summary of the six steps.

ρ^n		
$\leq$	$n \to \infty$ vs. finite n	1
$\min_{p \in P_n} \|p\|_S$		
$\underset{\sim}{\geq}$	Estimated vs. actual spectrum	2
$\min_{p \in P_n} \|p\|_{\Lambda(A)}$		
$\leq$	Normal vs. nonnormal A	3
$\min_{p \in P_n} \|p(A)\|$		
$\geq$	$\|p(A)\|$ vs. $\|p(A)r_0\|/\|r_0\|$	4
$\min_{p \in P_n} \|p(A)r_0\|/\|r_0\|$		
$\leq$	Minimization vs. quasi minimization	5
$\|p(A)r_0\|/\|r_0\|$		
$\underset{\sim}{\leq}$	Exact vs. floating-point arithmetic	6
$\|r_n\|/\|r_0\|$		

Step 2 is the approximation of the spectrum $\Lambda(A)$ by the continuum S. As an iteration proceeds, it typically begins to discover that $\Lambda(A)$ is not a continuum after all. Isolated eigenvalues begin to be "peeled off," and the convergence accelerates. This is sometimes called a phase of "superlinear convergence."

Step 3 is the difference between normal and nonnormal matrices. If A is normal, its spectrum determines everything, but in the nonnormal case, the convergence may be arbitrarily much slower than can be explained on the basis of the spectrum alone. Sometimes the ϵ-pseudospectra of A provide a better understanding of convergence behavior. As the iteration proceeds, smaller ϵ may become relevant, representing a second potential cause of acceleration of convergence in the later stages of the iteration.

Step 4 is the difference between matrix and vector norms, which is roughly the difference between a worst-case initial vector r_0 and the actual r_0. For an arbitrary r_0, an iteration can usually manage to ignore troublesome eigenvalues in the early stages. Convergence later slows down when it is forced to take them into account. This is sometimes called a phase of "sublinear convergence."

Step 5 is the difference between methods that minimize $\|r_n\|$ at each step (notably GMRES) and those that do not (BiCG, CGS, Bi-CGSTAB, QMR, TFQMR, etc.). These quasi-minimization methods require more iterations, but each iteration may require much less work, especially in cases of slow convergence where the cost of the computation is not dominated by the cost of matrix-vector products.

Step 6 is the approximation of exact by floating-point arithmetic. The effects of rounding on matrix iterations are complicated and not fully understood, especially for methods based on three-term recurrences. Most often rounding errors delay convergence without preventing it, but the latter is possible too.

Let us emphasize some things we have not done in this paper.

We have not given any details of Krylov subspace iterations and their implementation. Some of these methods, like conjugate gradients, are elementary and are described in numerous books. Others, like QMR with look-ahead, are complicated and perhaps best used in software form rather than programmed for oneself.

We have not, for that matter, said anything about software. For a start on this subject, see [3].

We have not given examples of matrices obtained from scientific or engineering applications, preferring to construct examples artificially to illustrate each point.

As mentioned in the introduction, we have not studied errors, only residuals. For ill-conditioned matrices A, some of the effects we have described look different if one examines $\|e_n\|/\|e_0\|$ rather than $\|r_n\|/\|r_0\|$. For example, the initial Step 4 sublinear convergence effect may be weakened, since $\|e_n\|/\|e_0\|$ is more strongly dominated by troublesome eigenvalues near the origin. Relatedly, somewhat different effects may be observed if x_0 and e_0 are random than if r_0 is random.

We have not addressed the crucial problem of preconditioners. If a matrix iteration is converging slowly, our study may help to explain why, but it will probably not point the way to a cure. The cure is a better preconditioner—but finding good preconditioners is a proverbial art, not a science.

We have not considered the use of Krylov subspace iterations for purposes other than solving systems of equations. The most conspicuous other area of application is eigenvalue problems [62]. In addition, these methods are also used for further purposes such as evaluation of $\exp(tA)$ or other matrix functions [10,11].

Finally, we have certainly not presented the only valid way to analyze the convergence of Krylov subspace iterations. To some degree the ordering of our six steps could be different. For example, the matrix-vector inequality (STEP4) could be formulated for arbitrary p rather than minimal p. In some areas, entirely different approaches might be possible. For example, isolated points in the spectrum could be treated differently; not all quasi-minimization methods fit the inequality (STEP5); and our assignment of all rounding errors to the equality (STEP6) at the end of the chain is somewhat artificial.

In closing, on the other hand, let us emphasize what we have done. We have reviewed, systematically and with concrete examples, the fundamental phenomena that govern convergence of Krylov subspace matrix iterations for solving systems of equations $Ax = b$. We believe we have touched upon all the main phenomena that arise in this subject.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994.

[2] N. S. Bakhvalov, *Numerical Methods*, Mir Publishers, Moscow, 1977.

[3] R. Barrett, et al., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.

[4] 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 Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 309–332.

[5] J. Cullum and A. Greenbaum, *Relations between Galerkin and norm-minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223–247.

[6] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. 1, Birkhäuser, Boston, 1985.

[7] J. H. Curtiss, *Faber polynomials and the Faber series*, Amer. Math. Monthly, 78 (1971), pp. 577–596.

[8] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.

[9] T. A. Driscoll, *A MATLAB toolbox for Schwarz–Christoffel mapping*, ACM Trans. Math. Softw., 22 (1996), pp. 168–186. See also `http://amath.colorado.edu/appm/faculty/tad/SC-toolbox`.

[10] V. Druskin, A. Greenbaum, and L. Knizhnerman, *Using nonorthogonal Lanczos vectors in the computation of matrix functions*, SIAM J. Sci. Comp., 19 (1998), pp. 38–54.

[11] W. S. Edwards, L. S. Tuckerman, R. A. Friesner, and D. C. Sorensen, *Krylov methods for the incompressible Navier–Stokes equations*, J. Comp. Phys., 110 (1994), pp. 82–102.

[12] M. Eiermann, *On semiiterative methods generated by Faber polynomials*, Numer. Math., 56 (1989), pp. 139–156.

[13] M. Eiermann, X. Li, and R. S. Varga, *On hybrid semi-iterative methods*, SIAM J. Numer. Anal., 26 (1989), pp. 152–168.

[14] M. Eiermann and W. Niethammer, *On the construction of semiiterative methods*, SIAM J. Numer. Anal., 20 (1983), pp. 1153–1160.

[15] M. Eiermann, W. Niethammer, and R. S. Varga, *A study of semiiterative methods for nonsymmetric systems of linear equations*, Numer. Math., 47 (1985), pp. 505–533.

[16] H. Elman and R. Streit, *Polynomial iteration for nonsymmetric indefinite linear systems*, in Numerical Analysis, Lecture Notes in Math. 1230, J. P. Hennert, ed., Springer-Verlag, 1986, pp. 103–117.

[17] G. Faber, *Über Tschebyscheffsche Polynome*, J. Reine Angew. Math., 150 (1920), pp. 79–106.

[18] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[19] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[20] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, Freeman, San Francisco, 1963.

[21] B. Fischer and R. W. Freund, *On adaptive weighted polynomial preconditioning for Hermitian positive definite matrices*, SIAM J. Sci. Comp., 15 (1994), pp. 408–426.

[22] B. Fischer and L. Reichel, *A stable Richardson iteration method for complex linear systems*, Numer. Math., 54 (1988), pp. 225–242.

[23] R. Fletcher, *Conjugate gradient methods for indefinite systems,* in Proceedings of the Dundee Biennial Conference on Numerical Analysis, G. A. Watson, ed., Springer-Verlag, New York, 1975, pp. 73–89.

[24] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.

[25] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.

[26] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[27] D. Gaier, *Lectures on Complex Approximation*, Birkhäuser, Boston, 1987.

[28] S. K. Godunov and V. S. Ryabenkii, *Theory of Difference Schemes*, North-Holland, Amsterdam, 1964.

[29] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[30] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[31] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[32] A. Greenbaum, *Krylov subspace approximations to the solution of a linear system*, in Linear and Nonlinear Conjugate Gradient-Related Methods, L. Adams and J. L. Nazareth, eds., SIAM, Philadelphia, 1996.

[33] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–552.

[34] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[35] A. Greenbaum and L. Gurvits, *Max-Min properties of matrix factor norms*, SIAM J. Sci. Comp., 15 (1994), pp. 348–358.

[36] A. Greenbaum, V. Ptak, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.

[37] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.

[38] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comp., 15 (1994), pp. 359–368.

[39] M. H. Gutknecht, *An iterative method for solving linear equations based on minimum norm Pick–Nevanlinna interpolation*, in Approximation Theory V, Academic Press, C. K. Chui, et al., eds., 1986, pp. 371–374.

[40] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[41] R. M. Hayes, *Iterative methods of solving linear problems on Hilbert space*, Nat. Bur. Stand. Appl. Math. Ser., 39 (1954), pp. 71–103.

[42] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 3, Wiley, New York, 1986.

[43] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.

[44] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[45] E. Hille, *Analytic Function Theory*, Vol. 2, Chelsea, New York, 1962.

[46] W. A. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comp., 15 (1994), pp. 427–439.

[47] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[48] C. Lanczos, *Solution of systems of linear equations by minimized iteration*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 33–53.

[49] H. J. Landau, *On Szegő's eigenvalue distribution theorem and non-Hermitian kernels*, J. d'Analyse Math., 28 (1975), pp. 335–357.

[50] S. H. Lui, *Computation of pseudospectra by continuation*, SIAM J. Sci. Comput., 18 (1997), pp. 565–573.

[51] T. A. Manteuffel, *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math., 31 (1978), pp. 183–208.

[52] T. A. Manteuffel and G. Starke, *On hybrid iterative methods for nonsymmetric systems of linear equations*, Numer. Math., 73 (1996), pp. 489–506.

[53] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.

[54] N. M. Nachtigal, L. Reichel, and L. N. Trefethen, *A hybrid GMRES algorithm for nonsymmetric linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 796–825.

[55] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.

[56] G. Opfer and G. Schober, *Richardson's iteration for nonsymmetric matrices*, Linear Algebra Appl., 58 (1984), pp. 343–361.

[57] C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Lin. Alg. Appl., 34 (1980), pp. 235–258.

[58] C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–624.

[59] L. Reichel, *Polynomials by conformal mapping for the Richardson iteration method for complex linear systems*, SIAM J. Numer. Anal., 25 (1988), pp. 1359–1368.

[60] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.

[61] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, 2nd ed., Wiley, New York, 1967.

[62] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, 1992.

[63] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Co., Boston, 1996.

[64] Y. Saad and M. H. Schultz, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comp., 7 (1986), pp. 856–869.

[65] E. B. SAFF, *Polynomial and rational approximation in the complex domain*, Proc. Symp. Appl. Math., 36 (1986), pp. 21–49.

[66] E. B. SAFF, *Orthogonal polynomials from a complex perspective*, in Orthogonal Polynomials, P. Nevai, ed., Kluwer, Norwell, MA, 1990, pp. 363–393.

[67] P. E. SAYLOR AND D. C. SMOLARSKI, *Implementation of an adaptive algorithm for Richardson's method*, Linear Algebra Appl., 154–156 (1991), pp. 615–646.

[68] P. SONNEVELD, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comp., 10 (1989), pp. 36–52.

[69] H. STAHL AND V. TOTIK, *General Orthogonal Polynomials*, Cambridge University Press, Cambridge, 1992.

[70] G. STARKE AND R. S. VARGA, *A hybrid Arnoldi–Faber method for nonsymmetric systems of linear equations*, Numer. Math., 64 (1993), pp. 213–240.

[71] K. C. TOH, *Matrix Approximation Problems and Nonsymmetric Iterative Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1996.

[72] K.-C. TOH, *GMRES vs. ideal GMRES*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 30–36.

[73] K.-C. TOH AND L. N. TREFETHEN, *Calculation of pseudospectra by the Arnoldi iteration*, SIAM J. Sci. Comp., 17 (1996), pp. 1–15.

[74] K.-C. TOH AND L. N. TREFETHEN, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., to appear.

[75] L. N. TREFETHEN, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, J. C. Mason and M. G. Cox, eds., Chapman and Hall, London, 1990, pp. 336–360.

[76] L. N. TREFETHEN, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1992, pp. 234–266.

[77] L. N. TREFETHEN AND DAVID BAU, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[78] M. TSUJI, *Potential Theory in Modern Function Theory*, Dover, New York, 1959.

[79] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

[80] H. A. VAN DER VORST, *Bi-CGSTAB: A fast and smoothly convergent variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comp., 13 (1992), pp. 631–644.

[81] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 5th ed., Amer. Math. Soc., Providence, RI, 1969.

[82] H. WIDOM, *Extremal polynomials associated with a system of curves in the complex plane*, Adv. Math., 3 (1969), pp. 127–232.

[83] R. WINTHER, *Some superlinear convergence results for the conjugate gradient methods*, SIAM J. Numer. Anal., 17 (1980), pp. 14–17.