SIAM REVIEW
Vol. 47, No. 3, pp. 492–515

Convergence of Polynomial Restart Krylov Methods for Eigenvalue Computations*

Christopher A. Beattie[†]
Mark Embree[‡]
D. C. Sorensen[‡]

Abstract. Krylov subspace methods have led to reliable and effective tools for resolving large-scale, non-Hermitian eigenvalue problems. Since practical considerations often limit the dimension of the approximating Krylov subspace, modern algorithms attempt to identify and condense significant components from the current subspace, encode them into a polynomial filter, and then restart the Krylov process with a suitably refined starting vector. In effect, polynomial filters dynamically steer low-dimensional Krylov spaces toward a desired invariant subspace through their action on the starting vector. The spectral complexity of nonnormal matrices makes convergence of these methods difficult to analyze, and these effects are further complicated by the polynomial filter process.

The principal object of study in this paper is the angle an approximating Krylov subspace forms with a desired invariant subspace. Convergence analysis is posed in a geometric framework that is robust to eigenvalue ill-conditioning, yet remains relatively uncluttered. The bounds described here suggest that the sensitivity of desired eigenvalues exerts little influence on convergence, provided the associated invariant subspace is well-conditioned; ill-conditioning of unwanted eigenvalues plays an essential role. This framework also gives insight into the design of effective polynomial filters. Numerical examples illustrate the subtleties that arise when restarting non-Hermitian iterations.

Key words. Krylov subspaces, Arnoldi algorithm, Lanczos algorithm, eigenvalue computations, containment gap, pseudospectra

AMS subject classification. 15A18, 15A60, 30E10, 47A15, 65F15

DOI. 10.1137/S0036144503433077

1. Introduction.

Recent improvements in algorithms and software have made large-scale eigenvalue computations increasingly routine. For example, Burroughs et al. resolve unstable flow regimes in a differentially heated cavity by calculating the three rightmost eigenvalues of matrices with dimension beyond 3 million [4]. For problems of such scale, computation of all eigenvalues and eigenvectors is both impractical and unnecessary. Instead, one restricts the matrix to a well-chosen subspace, from which approximations to eigenvalues and eigenvectors of physical interest are drawn.

*Received by the editors August 12, 2003; accepted for publication (in revised form) August 11, 2004; published electronically July 29, 2005. This material is partly based upon work supported by contract 74837-001-0349 from the Regents of University of California (Los Alamos National Laboratory) to William Marsh Rice University. Further support was provided by DOE grant DE-FG03-02ER25531 and NSF grants DMS-9972591, CCR-9988393, and ACI-0082645.

http://www.siam.org/journals/sirev/47-3/43307.html

[†]Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123 (beattie@math.vt.edu).

[‡]Department of Computational and Applied Mathematics, Rice University, 6100 Main Street—MS 134, Houston, TX 77005-1892 (embree@rice.edu, sorensen@rice.edu).

In this paper we analyze restriction onto Krylov subspaces, the approach taken by the Arnoldi and bi-orthogonal Lanczos algorithms [2], which engage the matrix only through matrix-vector multiplications. The ℓth Krylov subspace generated by the matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ and the vector $\mathbf{v}_1 \in \mathbb{C}^n$ is

$$\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1) \equiv \operatorname{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}^{\ell-1}\mathbf{v}_1\}.$$

Krylov subspace methods approximate the eigenvalues of $\mathbf{A}$ by the eigenvalues of a restriction of $\mathbf{A}$ to $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$. If we wish to understand the capacity of Krylov subspace methods to provide accurate approximations and to quantify those factors that influence convergence, we must first settle on an appropriate way to measure accuracy.

Given an approximate eigenpair $(\widehat{\lambda}, \widehat{\mathbf{u}})$ with $\|\widehat{\mathbf{u}}\| = 1$, the residual norm $\|\mathbf{A}\widehat{\mathbf{u}} - \widehat{\lambda}\widehat{\mathbf{u}}\|$ provides a natural measure of accuracy that can be easily computed. For Hermitian problems, this residual norm bounds the distance between $\widehat{\lambda}$ and a nearest eigenvalue of $\mathbf{A}$. In contrast, the eigenvalues of non-Hermitian matrices can be highly sensitive to perturbations, in which case a small residual no longer implies comparable accuracy in the approximate eigenvalue. We contend that direct study of convergence to invariant subspaces yields greater insight than can be drawn from residual norms. Such an approach facilitates analysis when the coefficient matrix is defective or otherwise far from normal. In this work, we bound convergence of the largest canonical angle between a fixed invariant subspace and a Krylov subspace as the approximating subspace is enlarged or refined via polynomial restarts. As our development deals with subspaces, rather than the eigenvalue estimates generated by any particular algorithm, it provides a general convergence framework for all Krylov eigenvalue algorithms.

Bounds of this sort are familiar in the Krylov subspace literature, beginning with Saad's 1980 article that revived interest in the Arnoldi algorithm [22]. Among that paper's contributions is a bound on the angle between a single eigenvector and a Krylov subspace in terms of a simple polynomial approximation problem in the complex plane. Jia generalized this bound to handle defective eigenvalues; his analysis uses the Jordan structure of $\mathbf{A}$ and derivatives of the approximating polynomial [13]. Various other generalizations of Saad's bound have been developed for block Krylov methods [15, 21, 23].

Recently, new bounds have been derived for single-vector Krylov subspace methods that impose no restriction on the dimension of the desired invariant subspace or diagonalizability of $\mathbf{A}$, yet still result in a conventional polynomial approximation problem [3]. While examples demonstrate that these bounds can be descriptive, their derivation involves fairly intricate arguments. Our purpose is to present simplified bounds whose development is more elementary, even suitable for classroom presentation. The resulting analysis incorporates a different polynomial approximation problem. In typical situations the new bounds are weaker at early iterations, though the asymptotic convergence rate we establish is never worse than that obtained in [3]. In certain situations where the desired eigenvalues are ill-conditioned, these new bounds improve the earlier analysis.

Our first main result bounds the distance of $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ from a desired invariant subspace of $\mathbf{A}$ as the approximating subspace dimension ℓ increases and the starting vector $\mathbf{v}_1$ remains fixed, the classic setting for convergence analysis. In theory, Krylov projection methods terminate in a finite number of steps, but for very large problems, analysis of such asymptotic behavior still has computational significance.

In most practical situations, the desired eigenvalues are not well-separated from the rest of the spectrum. This causes slow convergence, and hence the dimension of the approximating subspace must become intractably large to deliver estimates with acceptable accuracy. To limit storage requirements and computational cost, one restarts the algorithm with improved starting vectors. Polynomial restarting is a popular approach that is often very effective. Here one projects $\mathbf{A}$ onto the Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(r)})$, where the dimension ℓ remains fixed, but the starting vector is modified at each outer iteration: $\mathbf{v}_1^{(r)} = \phi_r(\mathbf{A})\mathbf{v}_1^{(r-1)}$, where $\mathbf{v}_1^{(0)} = \mathbf{v}_1$ and ϕ_r is a polynomial with $\deg(\phi_r) < \ell$. Thus $\mathbf{v}_1^{(r)} = \Phi_r(\mathbf{A})\mathbf{v}_1$, where $\Phi_r(z) = \prod_{j=1}^r \phi_j(z)$ is the product of all the restart polynomials. Though the projection space $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(r)})$ is always a subspace of the full Krylov space $\mathcal{K}_{\ell r}(\mathbf{A}, \mathbf{v}_1)$, the asymptotic convergence behavior of restarted algorithms depends critically on the selection of the polynomials ϕ_j. Our convergence analysis is based on selecting the zeros of these polynomials with respect to regions in the complex plane, a setting in which classical polynomial approximation results apply.

Ultimately, our bounds predict asymptotic convergence at a rate determined by the distance between the desired and unwanted eigenvalues. Ill-conditioning of unwanted eigenvalues can impede the convergence rate, but similar sensitivity of the desired eigenvalues plays no role in the asymptotic behavior of our bounds, provided the associated invariant subspace is well-conditioned. Starting vector bias affects the transient delay preceding convergence, but does not influence the ultimate convergence rate.

Before proceeding to our bounds, we establish notation and give basic requirements on the matrix $\mathbf{A}$, the desired invariant subspace, and the starting vector $\mathbf{v}_1$ that ensure convergence is possible. In all that follows, $\|\cdot\|$ denotes the standard vector two-norm and the matrix norm it induces.

2. Decomposition of Krylov Spaces with Respect to Eigenspaces of $\mathbf{A}$.

Suppose the matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ has N distinct eigenvalues, $\{\lambda_j\}$, $j = 1, \ldots, N$. We wish to compute $L < N$ of these eigenvalues, $\lambda_1, \ldots, \lambda_L$, which we shall call the *good* eigenvalues. The remaining eigenvalues, the *bad* eigenvalues, are viewed as undesirable only to the extent that they are not of immediate interest, and we do not wish to expend any effort to compute them. We impose no assumptions regarding eigenvalue multiplicity; in particular, both good and bad eigenvalues may be defective.

We aim to understand how a Krylov space might converge to an invariant subspace associated with the good eigenvalues. To do this, we need to explain how $\mathbb{C}^n$ is decomposed into such subspaces. Our focus naturally arrives at the complementary maximal invariant subspaces associated with the good and bad eigenvalues:

$$\mathcal{X}_g \equiv \bigoplus_{j=1}^{L} \mathrm{Ker}(\mathbf{A} - \lambda_j\mathbf{I})^{n_j} \quad \text{and} \quad \mathcal{X}_b \equiv \bigoplus_{j=L+1}^{N} \mathrm{Ker}(\mathbf{A} - \lambda_j\mathbf{I})^{n_j},$$

where n_j denotes the ascent of λ_j. When $\mathbf{A}$ is diagonalizable, $\mathcal{X}_g$ and $\mathcal{X}_b$ are simply the span of all eigenvectors corresponding to the good and bad eigenvalues; for defective matrices, $\mathcal{X}_g$ and $\mathcal{X}_b$ will include all generalized eigenvectors of higher grade as well. In either case,

$$\mathbb{C}^n = \mathcal{X}_g \oplus \mathcal{X}_b.$$

How well can $\mathcal{X}_g$ be approximated by vectors drawn from the Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$, and how does this relate the dimension k and properties of $\mathbf{A}$ and $\mathbf{v}_1$? In this section we characterize those good invariant subspaces (within $\mathcal{X}_g$) that can be captured with Krylov subspaces, adapting the discussion from [3]. Throughout we assume that the starting vector $\mathbf{v}_1$ is fixed.

Since the dimension of $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ is bounded by n, there exists a smallest positive integer s such that

$$\mathcal{K}_s(\mathbf{A}, \mathbf{v}_1) = \operatorname{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \mathbf{A}^2\mathbf{v}_1, \dots\} =: \mathcal{K}(\mathbf{A}, \mathbf{v}_1).$$

This *maximal Krylov subspace*, $\mathcal{K}(\mathbf{A}, \mathbf{v}_1)$, is evidently an invariant subspace of $\mathbf{A}$. However, if any good eigenvalue is derogatory (i.e., has geometric multiplicity greater than 1), then $\mathcal{X}_g \not\subseteq \mathcal{K}(\mathbf{A}, \mathbf{v}_1)$ and no Krylov subspace generated by $\mathbf{v}_1$ will capture all of $\mathcal{X}_g$. To see this, note that since $\mathbf{A}^s\mathbf{v}_1 \in \operatorname{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \mathbf{A}^2\mathbf{v}_1, \dots, \mathbf{A}^{s-1}\mathbf{v}_1\}$, there exists a polynomial, $\mu(z) = z^s - \gamma_{s-1}z^{s-1} - \cdots - \gamma_1 z - \gamma_0$, such that $\mu(\mathbf{A})\mathbf{v}_1 = \mathbf{0}$. This μ is the *minimal polynomial of* $\mathbf{A}$ *with respect to* $\mathbf{v}_1$, i.e., the monic polynomial μ of lowest degree such that $\mu(\mathbf{A})\mathbf{v}_1 = \mathbf{0}$.

Now, write $\mathbf{K} = [\mathbf{v}_1 \ \mathbf{A}\mathbf{v}_1 \ \cdots \ \mathbf{A}^{s-1}\mathbf{v}_1] \in \mathbb{C}^{n\times s}$ and note that

$$\mathbf{A}\mathbf{K} = \mathbf{K}\mathbf{A}_s,$$

where $\mathbf{A}_s$ has the companion matrix form

$$\mathbf{A}_s = \begin{pmatrix} & & & \gamma_0 \\ 1 & & & \gamma_1 \\ & \ddots & & \vdots \\ & & 1 & \gamma_{s-1} \end{pmatrix};$$

unspecified entries are zero. Since $\mathbf{A}_s$ is a companion matrix, it cannot be derogatory; hence $\mathcal{K}(\mathbf{A}, \mathbf{v}_1) = \operatorname{Range}(\mathbf{K})$ cannot contain any invariant subspace associated with a derogatory eigenvalue [25]. Can it come close?

What does it mean for a Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ to come close to a fixed invariant subspace as the dimension ℓ increases? We seek a framework to discuss the proximity of subspaces to one another. The intuitive notion of the angle between subspaces is unambiguous only for pairs of one-dimensional subspaces; we require some way of measuring the distance between subspaces of different dimensions. The *containment gap* between the subspaces $\mathcal{W}$ and $\mathcal{V}$ is defined as

$$\delta(\mathcal{W}, \mathcal{V}) \equiv \max_{\mathbf{w}\in\mathcal{W}} \min_{\mathbf{v}\in\mathcal{V}} \frac{\|\mathbf{w} - \mathbf{v}\|}{\|\mathbf{w}\|}.$$

Note that $\delta(\mathcal{W}, \mathcal{V})$ is the sine of the largest canonical angle between $\mathcal{W}$ and the closest subspace of $\mathcal{V}$ with the same dimension as $\mathcal{W}$. If $\dim \mathcal{V} < \dim \mathcal{W}$, then $\delta(\mathcal{W}, \mathcal{V}) = 1$, while $\delta(\mathcal{W}, \mathcal{V}) = 0$ if and only if $\mathcal{W} \subseteq \mathcal{V}$. See [14, sect. IV.2.1] and [26, sect. II.4] for further details.

2.1. Characterization of the Maximal Reachable Invariant Subspace.

Let $\widetilde{\mu}$ denote the minimal annihilating polynomial of $\mathbf{A}$, i.e., the monic polynomial $\widetilde{\mu}$ of lowest degree such that $\widetilde{\mu}(\mathbf{A}) = \mathbf{0}$. (Note that $\widetilde{\mu}$ must contain μ as a factor.) We decompose $\mathbb{C}^n$ into good and bad invariant subspaces using the following construction of Gantmacher [11, sect. VII.2]. Factor $\widetilde{\mu}$ as the product of two monic polynomials, $\widetilde{\mu}(z) = \widetilde{\alpha}_g(z)\widetilde{\alpha}_b(z)$, where $\widetilde{\alpha}_g$ and $\widetilde{\alpha}_b$ have the good and bad eigenvalues as roots, respectively, and are the lowest degree polynomials that satisfy

$$\widetilde{\alpha}_g(\mathbf{A})\mathcal{X}_g = \{\mathbf{0}\} \quad \text{and} \quad \widetilde{\alpha}_b(\mathbf{A})\mathcal{X}_b = \{\mathbf{0}\}.$$

A partial fraction expansion provides two polynomials, $\beta_g(z)$ and $\beta_b(z)$, such that

$$\frac{1}{\widetilde{\alpha}_g(z)\widetilde{\alpha}_b(z)} \;=\; \frac{\beta_g(z)}{\widetilde{\alpha}_g(z)} + \frac{\beta_b(z)}{\widetilde{\alpha}_b(z)}.$$

Rearranging and substituting $\mathbf{A} \hookrightarrow z$ yields $\mathbf{I} = \widetilde{\alpha}_g(\mathbf{A})\beta_b(\mathbf{A}) + \widetilde{\alpha}_b(\mathbf{A})\beta_g(\mathbf{A})$.

Now, define $\mathbf{P}_g \equiv \widetilde{\alpha}_b(\mathbf{A})\beta_g(\mathbf{A})$ and $\mathbf{P}_b \equiv \widetilde{\alpha}_g(\mathbf{A})\beta_b(\mathbf{A})$, so that $\mathbf{P}_g + \mathbf{P}_b = \mathbf{I}$. Noting that $\widetilde{\alpha}_g(\mathbf{A})\widetilde{\alpha}_b(\mathbf{A}) = \mathbf{0}$, one may verify the following:

$$\begin{array}{llll} \mathbf{P}_g = \mathbf{P}_g^2, & \mathbf{A}\mathbf{P}_g = \mathbf{P}_g\mathbf{A}, & \mathcal{X}_g = \mathrm{Range}(\mathbf{P}_g), & \mathcal{X}_b = \mathrm{Ker}(\mathbf{P}_g); \\ \mathbf{P}_b = \mathbf{P}_b^2, & \mathbf{A}\mathbf{P}_b = \mathbf{P}_b\mathbf{A}, & \mathcal{X}_b = \mathrm{Range}(\mathbf{P}_b), & \mathcal{X}_g = \mathrm{Ker}(\mathbf{P}_b). \end{array}$$

Hence $\mathbf{P}_g$ and $\mathbf{P}_b$ are spectral projections onto the good and bad invariant subspaces, $\mathcal{X}_g$ and $\mathcal{X}_b$.

Our first result decomposes the maximal Krylov subspace into two Krylov subspaces with projected starting vectors.

Lemma 2.1.

$$\mathcal{K}(\mathbf{A}, \mathbf{v}_1) = \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \oplus \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1).$$

Proof. Since $\mathbf{P}_g\mathbf{v}_1 \in \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \subseteq \mathcal{X}_g$ and $\mathbf{P}_b\mathbf{v}_1 \in \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1) \subseteq \mathcal{X}_b$ with $\mathcal{X}_g \cap \mathcal{X}_b = \{\mathbf{0}\}$, for any $\mathbf{x} = \psi(\mathbf{A})\mathbf{v}_1 \in \mathcal{K}(\mathbf{A}, \mathbf{v}_1)$ we have

$$\mathbf{x} = \psi(\mathbf{A})\big(\mathbf{P}_g + \mathbf{P}_b\big)\mathbf{v}_1 = \big(\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1 + \psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\big) \in \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \oplus \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1).$$

To demonstrate the opposite containment, suppose $\mathbf{x} \in \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \oplus \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1)$. Then there exist polynomials ψ_g and ψ_b such that

$$\begin{aligned} \mathbf{x} &= \psi_g(\mathbf{A})\mathbf{P}_g\mathbf{v}_1 + \psi_b(\mathbf{A})\mathbf{P}_b\mathbf{v}_1 \\ &= \big(\psi_g(\mathbf{A})\widetilde{\alpha}_b(\mathbf{A})\beta_g(\mathbf{A}) + \psi_b(\mathbf{A})\widetilde{\alpha}_g(\mathbf{A})\beta_b(\mathbf{A})\big)\mathbf{v}_1 \\ &\in \mathcal{K}(\mathbf{A}, \mathbf{v}_1). \qquad \square \end{aligned}$$

The next corollary immediately follows from the fact that $\mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \subseteq \mathcal{X}_g$ and $\mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1) \subseteq \mathcal{X}_b$.

Corollary 2.2.

$$\mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) = \mathcal{K}(\mathbf{A}, \mathbf{v}_1) \cap \mathcal{X}_g.$$

Thus $\mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1)$ is a distinguished subspace, called the *maximal reachable invariant subspace* for the starting vector $\mathbf{v}_1$. It is the largest invariant subspace of $\mathcal{X}_g$ to which our Krylov subspace can possibly converge; we denote it by

$$\mathcal{U}_g \equiv \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) \subseteq \mathcal{X}_g.$$

Ideally, $\mathcal{U}_g = \mathcal{X}_g$, but we have already seen that if any good eigenvalue is derogatory, no Krylov subspace generated from a single starting vector can fully capture $\mathcal{X}_g$, and then $\mathcal{U}_g \neq \mathcal{X}_g$. (Curiously, eigenvalues that are defective but nonderogatory avoid this problem.) Also note that if the starting vector $\mathbf{v}_1$ lacks a component in any good generalized eigenvector of maximal grade, then again $\mathcal{U}_g \neq \mathcal{X}_g$. The following lemma [3, 25] identifies an explicit barrier to how close a Krylov subspace can come to $\mathcal{X}_g$. This barrier is independent of the approximating subspace dimension and starting vector.

LEMMA 2.3. *If $\mathcal{U}_g$ is a proper subset of $\mathcal{X}_g$, then*

$$\delta(\mathcal{X}_g, \mathcal{K}(\mathbf{A}, \mathbf{v}_1)) \geq \frac{1}{\|\mathbf{P}_g\|}.$$

Proof. Let $\mathcal{U}_b \equiv \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1)$ denote the complementary maximal reachable invariant subspace, and recall that Lemma 2.1 allows any $\mathbf{v} \in \mathcal{K}(\mathbf{A}, \mathbf{v}_1)$ to be written as $\mathbf{v} = \mathbf{v}_g + \mathbf{v}_b$ for some $\mathbf{v}_g \in \mathcal{U}_g$ and $\mathbf{v}_b \in \mathcal{U}_b$. Since $\mathcal{U}_g$ is a proper subset of $\mathcal{X}_g$, there exists some nonzero $\mathbf{z} \in \mathcal{X}_g \setminus \mathcal{U}_g$ such that $\mathbf{z} \perp \mathcal{U}_g$. For any $\mathbf{v}_g \in \mathcal{U}_g$, we have

$$\|\mathbf{z} - \mathbf{v}_g\|^2 = \|\mathbf{z}\|^2 + \|\mathbf{v}_g\|^2,$$

and so $\|\mathbf{z} - \mathbf{v}_g\| \geq \|\mathbf{z}\|$. Thus,

$$\begin{aligned}
\delta(\mathcal{X}_g, \mathcal{K}(\mathbf{A}, \mathbf{v}_1)) &= \max_{\mathbf{u} \in \mathcal{X}_g} \min_{\mathbf{v} \in \mathcal{K}(\mathbf{A}, \mathbf{v}_1)} \frac{\|\mathbf{u} - \mathbf{v}\|}{\|\mathbf{u}\|} \\
&\geq \min_{\mathbf{v} \in \mathcal{K}(\mathbf{A}, \mathbf{v}_1)} \frac{\|\mathbf{z} - \mathbf{v}\|}{\|\mathbf{z}\|} = \min_{\mathbf{v}_g \in \mathcal{U}_g, \mathbf{v}_b \in \mathcal{U}_b} \frac{\|\mathbf{z} - \mathbf{v}_g - \mathbf{v}_b\|}{\|\mathbf{z}\|} \\
&\geq \min_{\mathbf{v}_g \in \mathcal{U}_g, \mathbf{v}_b \in \mathcal{U}_b} \frac{\|\mathbf{z} - \mathbf{v}_g - \mathbf{v}_b\|}{\|\mathbf{z} - \mathbf{v}_g\|} = \min_{\mathbf{v}_g \in \mathcal{U}_g, \mathbf{v}_b \in \mathcal{U}_b} \frac{\|\mathbf{z} - \mathbf{v}_g - \mathbf{v}_b\|}{\|\mathbf{P}_g(\mathbf{z} - \mathbf{v}_g - \mathbf{v}_b)\|} \\
&\geq \min_{\mathbf{x} \in \mathbb{C}^n} \frac{\|\mathbf{x}\|}{\|\mathbf{P}_g\mathbf{x}\|} = \frac{1}{\|\mathbf{P}_g\|}. \qquad \square
\end{aligned}$$

One might hope that polynomial restarts would provide a mechanism to reach vectors in $\mathcal{X}_g \setminus \mathcal{U}_g$, but this is not the case, as for any polynomial Φ, $\mathcal{K}(\mathbf{A}, \Phi(\mathbf{A})\mathbf{v}_1) \subseteq \mathcal{K}(\mathbf{A}, \mathbf{v}_1)$. In light of this, our analysis will focus on the gap convergence to the maximal reachable invariant subspace $\mathcal{U}_g$. Since $\mathcal{U}_g \subseteq \mathcal{K}(\mathbf{A}, \mathbf{v}_1)$, a sufficiently large Krylov subspace will exactly capture $\mathcal{U}_g$, but typically such a Krylov space is prohibitively large. Our analysis will describe a gap convergence rate that is typically descriptive well before exact termination.

3. Convergence of Polynomial Restart Methods.

We address two closely related, fundamental questions:

> What is the gap $\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1))$ between $\mathcal{U}_g$ and the Krylov space as the dimension ℓ increases?

The answer to this first question depends on the eigenvalue distribution and nonnormality of $\mathbf{A}$, as well as on the distribution of $\mathbf{v}_1$ with respect to $\mathcal{U}_g$. This analysis informs our approach to the second question:

> Given a polynomial Φ that describes a restart filter, how does the gap $\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \widehat{\mathbf{v}}_1))$ depend on $\widehat{\mathbf{v}}_1 = \Phi(\mathbf{A})\mathbf{v}_1$, and how can we optimize the asymptotic behavior of this gap as additional restarts are performed?

One goal of restarting is to mimic the performance of a full iteration (i.e., no restarts), but with restricted subspace dimensions. If we consider $\Phi \equiv \prod_{j=1}^{r} \phi_j$, where each ϕ_j is a polynomial associated with restarting a Krylov subspace at the jth stage, a quantification of the gap $\delta(\mathcal{U}_g, \mathcal{K}(\mathbf{A}, \Phi(\mathbf{A})\mathbf{v}_1))$ will lead to a convergence rate for the restarting scheme.

3.1. Convergence Bounds for Krylov Subspaces with No Restarts.

We shall begin by discussing the distance of a Krylov space of dimension ℓ from the reachable subspace $\mathcal{U}_g$, and then introduce the consequences for restarting. We use the notation

$\mathcal{P}_k$ to denote the space of polynomials of degree at most k, and throughout assume that $\mathbf{v}_1$ is such that $m \equiv \dim\, \mathcal{U}_g > 0$. Critical to our discussion is $\alpha_g \in \mathcal{P}_m$, the *minimal polynomial of* $\mathbf{A}$ *with respect to* $\mathbf{P}_g\mathbf{v}_1$, i.e., the monic polynomial of lowest degree such that $\alpha_g(\mathbf{A})\mathbf{P}_g\mathbf{v}_1 = \mathbf{0}$.

For each $\psi \in \mathcal{P}_{m-1}$, define the set of kth-degree $\mathcal{U}_g$*-interpolants*,

$$\mathcal{P}_k^\psi \equiv \{\phi \in \mathcal{P}_k : \ \phi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1 = \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\}.$$

Each $\phi \in \mathcal{P}_k^\psi$ interpolates ψ at the good eigenvalues. Lemma 3.1 provides a full characterization of $\mathcal{P}_k^\psi$, which we then apply to obtain bounds on the containment gap. The sets $\mathcal{P}_k^\psi$ were employed in [25, Cor. 5.5] to prove a version of our Lemma 3.2.

LEMMA 3.1. *If* $k < \deg(\psi)$, $\mathcal{P}_k^\psi$ *is empty; if* $\deg(\psi) \le k \le m-1$, $\mathcal{P}_k^\psi = \{\psi\}$*; if* $k \ge m$, $\mathcal{P}_k^\psi$ *comprises all polynomials* $\phi \in \mathcal{P}_k$ *of the form*

$$\phi(z) = \psi(z) - \widehat{\phi}(z)\alpha_g(z)$$

for some $\widehat{\phi} \in \mathcal{P}_{k-m}$.

Proof. Suppose $\phi \in \mathcal{P}_k^\psi$. Then $\phi - \psi$ is an annihilating polynomial of $\mathbf{P}_g\mathbf{v}_1$: $(\phi(\mathbf{A}) - \psi(\mathbf{A}))\mathbf{P}_g\mathbf{v}_1 = \mathbf{0}$. Since α_g is a minimum-degree annihilating polynomial, either $\deg(\phi-\psi) < m$ and hence $\phi - \psi \equiv 0$, or $\phi - \psi = \widehat{\phi}\alpha_g$ with $\deg(\widehat{\phi}) = \deg(\phi - \psi) - m \ge 0$. Therefore, if $k < \deg(\psi)$, then $\deg(\phi-\psi) < m$ and $\deg(\phi) < \deg(\psi)$, so $\mathcal{P}_k^\psi$ must be empty. If $\deg(\psi) \le k \le m-1$, then $\phi \equiv \psi$ must hold and thus $\mathcal{P}_k^\psi = \{\psi\}$. Finally, when $k \ge m$, $\phi(z) - \psi(z) = \widehat{\phi}(z)\alpha_g(z)$ must hold for some $\widehat{\phi} \in \mathcal{P}_{k-m}$. □

LEMMA 3.2. *For any* $\ell \ge m = \dim \mathcal{U}_g$,

$$\begin{aligned}\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)) &\le \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{\ell-1}^\psi} \frac{\|\phi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\\ &= \max_{\psi\in\mathcal{P}_{m-1}} \min_{\widehat{\phi}\in\mathcal{P}_{\ell-m-1}} \frac{\big\|\big[\psi(\mathbf{A}) - \widehat{\phi}(\mathbf{A})\alpha_g(\mathbf{A})\big]\mathbf{P}_b\mathbf{v}_1\big\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}, \qquad (3.1)\end{aligned}$$

with the convention that $\mathcal{P}_{-1} = \{0\}$.

Proof. For any $\mathbf{x} \in \mathcal{U}_g = \mathcal{K}(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1) = \mathcal{K}_m(\mathbf{A}, \mathbf{P}_g\mathbf{v}_1)$, there is a unique polynomial of degree $m-1$ or less such that $\mathbf{x} = \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1$. Now, $\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ implies $\mathbf{v} = \phi(\mathbf{A})\mathbf{v}_1$ for some $\phi \in \ \mathcal{P}_{\ell-1}$. Thus

$$\begin{aligned}(3.2)\quad \delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)) &= \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{\ell-1}} \frac{\|\phi(\mathbf{A})\mathbf{v}_1 - \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\\ &= \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{\ell-1}} \frac{\|\phi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1 + [\phi(\mathbf{A}) - \psi(\mathbf{A})]\,\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\\ &\le \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{\ell-1}^\psi} \frac{\|\phi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}.\end{aligned}$$

The formulation (3.1) follows from Lemma 3.1. □

The estimate (3.1) will lead to a readily interpreted bound, similar in structure to the main result of [3]. Toward this end, we restrict minimization over $\widehat{\phi} \in \mathcal{P}_{\ell-m-1}$ to polynomials of the form $\widehat{\phi}(z) = \psi(z)p(z)$, where $\psi \in \mathcal{P}_{m-1}$ is the polynomial being maximized over, and $p \in \mathcal{P}_{\ell-2m}$ is an arbitrary polynomial. This then gives

$$\min_{\widehat{\phi}\in\mathcal{P}_{\ell-m-1}} \big\|\big[\psi(\mathbf{A}) - \widehat{\phi}(\mathbf{A})\alpha_g(\mathbf{A})\big]\mathbf{P}_b\mathbf{v}_1\big\| \le \min_{p\in\mathcal{P}_{\ell-2m}} \big\|\big[\psi(\mathbf{A}) - \psi(\mathbf{A})p(\mathbf{A})\alpha_g(\mathbf{A})\big]\mathbf{P}_b\mathbf{v}_1\big\|.$$

To simplify the right-hand side further, we utilize $\mathbf{\Pi}_b$, the orthogonal projection onto the complementary maximal reachable invariant subspace, $\mathcal{U}_b = \mathcal{K}(\mathbf{A}, \mathbf{P}_b\mathbf{v}_1)$. Note that $\mathbf{\Pi}_b\mathbf{P}_b = \mathbf{P}_b$, since $\mathrm{Range}(\mathbf{\Pi}_b) = \mathrm{Range}(\mathbf{P}_b)$, and $\mathbf{\Pi}_b = \mathbf{P}_b$ if and only if $\mathcal{U}_b \perp \mathcal{U}_g$. Keeping in mind that $\mathbf{A}$ and $\mathbf{P}_b$ commute, we find

$$\begin{aligned}
\min_{p\in\mathcal{P}_{\ell-2m}} \| [\psi(\mathbf{A}) \,-\, \psi(\mathbf{A})p(\mathbf{A})\alpha_g(\mathbf{A})]\, \mathbf{P}_b\mathbf{v}_1\| \qquad& \\
&= \min_{p\in\mathcal{P}_{\ell-2m}} \| [\mathbf{I} - p(\mathbf{A})\alpha_g(\mathbf{A})]\, \mathbf{\Pi}_b\, \psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\| \\
&\le \min_{p\in\mathcal{P}_{\ell-2m}} \|[\mathbf{I} - p(\mathbf{A})\alpha_g(\mathbf{A})]\mathbf{\Pi}_b\|\, \|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\| \\
&\le \min_{p\in\mathcal{P}_{\ell-2m}} \Big(\kappa(\Omega_b) \max_{z\in\Omega_b} |1 - p(z)\alpha_g(z)|\Big)\, \|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|. \qquad (3.3)
\end{aligned}$$

Here Ω_b is any compact subset of the complex plane containing all the bad eigenvalues while excluding all the good. The constant $\kappa(\Omega_b)$, introduced in [3], is the smallest positive number such that the inequality

$$\|f(\mathbf{A})\,\mathbf{\Pi}_b\| \;\le\; \kappa(\Omega_b) \max_{z\in\Omega_b} |f(z)|$$

holds uniformly for all functions f analytic on Ω_b. This constant, together with the choice of Ω_b itself, will be our key mechanism for describing the effects of nonnormality on convergence: $\kappa(\Omega_b) \ge 1$ for all nontrivial Ω_b, and $\kappa(\Omega_b) > 1$ is only possible when $\mathbf{A}$ is nonnormal.[1] In our bounds, enlarging Ω_b generally decreases $\kappa(\Omega_b)$ (provided the new Ω_b includes no additional eigenvalues), but also requires maximization in (3.3) over a larger set, degrading the convergence rate. Flexibility in the choice of Ω_b will allow us to describe convergence for general non-Hermitian problems without recourse to a diagonalizing similarity transformation or the Jordan canonical form. Precise details are addressed in section 3.2.

Substituting (3.3) into the right-hand side of (3.1) gives our primary result for Krylov methods without restarts.

Theorem 3.3. *For all $\ell \ge 2m$,*

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le \left(\max_{\psi\in\mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\right)\Big(\kappa(\Omega_b)\Big) \min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} \big|1 - \alpha_g(z)p(z)\big|. \qquad (3.4)$$

Compare this bound to the main result of [3]:

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le C_0\left(\max_{\psi\in\mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\right)\Big(\kappa(\Omega_b)\,\kappa(\Omega_g)\Big) \min_{q\in\mathcal{P}_{\ell-m}} \frac{\max_{z\in\Omega_b} |q(z)|}{\min_{z\in\Omega_g} |q(z)|}, \qquad (3.5)$$

where the compact set $\Omega_g \subseteq \mathbb{C} \setminus \Omega_b$ contains all the good eigenvalues, and $C_0 = 1$ if $\mathcal{U}_b \perp \mathcal{U}_g$; otherwise $C_0 = \sqrt{2}$. The constant $\kappa(\Omega_g)$ is defined analogously to $\kappa(\Omega_b)$.

The starting vector only affects the first parenthesized term, common to both bounds. This constant can take any value in $[0, \infty)$; it is small when $\mathbf{v}_1$ is strongly oriented toward $\mathcal{U}_g$. Regardless, we will see that the starting vector does not affect the predicted asymptotic convergence *rate* in either bound.

The bounds (3.4) and (3.5) differ in several interesting ways. First, they involve different polynomial approximation problems. The new problem amounts to fixing

[1] In the language of dilation theory, Ω_b is a *K-spectral set* for $K = \kappa(\Omega_b)$; see [18].

the value of the approximating polynomial $q \in \mathcal{P}_{\ell-m}$ from (3.5) to be 1 at all the good eigenvalues: if $q \in \mathcal{P}_{\ell-m}$ with $q(\lambda) = 1$ for all good eigenvalues λ (with matching multiplicities), then q must have the form $q(z) = 1 - \alpha_g(z)p(z)$ for some $p \in \mathcal{P}_{\ell-2m}$. In the special case that Ω_g consists only of the good eigenvalues, then

$$\begin{aligned}(3.6)\qquad \min_{q\in\mathcal{P}_{\ell-m}} \frac{\max\{|q(z)| : z \in \Omega_b\}}{\min\{|q(z)| : z \in \Omega_g\}} &\leq \min_{p\in\mathcal{P}_{\ell-2m}} \frac{\max\{|1-\alpha_g(z)p(z)| : z \in \Omega_b\}}{\min\{|1-\alpha_g(z)p(z)| : z \in \Omega_g\}} \\ &= \min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} |1-\alpha_g(z)p(z)|.\end{aligned}$$

When there is only a single good eigenvalue λ, and it is simple, then $m = 1$ and assigning $p(\lambda) = 1$ amounts to scaling p. Thus equality holds in (3.6), and the two polynomial approximation problems are identical. (In this case, one would always take $\Omega_g = \{\lambda\}$, giving $\kappa(\Omega_g) = 1$.) For larger m, the new bound (3.4) can be somewhat worse than (3.5). Note that gap convergence can commence as soon as the Krylov subspace dimension ℓ reaches $m = \dim \mathcal{U}_g$. The approximation problem in (3.5) captures this fact, while the new result (3.4) enforces a delay of m further iterations. The examples in section 4.2 demonstrate how this extra delay can cause the quality of our new bound to degrade as m increases, though the predicted convergence rate does not suffer. Another notable difference between (3.4) and (3.5) is the second parenthetical constant in each bound: (3.4) avoids the factor $\kappa(\Omega_g) \geq 1$.

3.2. The Size of $\kappa(\Omega)$. What governs the size of the constants $\kappa(\Omega_b)$ and $\kappa(\Omega_g)$? We present several upper bounds derived in [3]. First, take Ω to be a set of nondefective eigenvalues, and let the columns of $\mathbf{U}$ be an eigenvector basis for the corresponding invariant subspace. Then

$$(3.7)\qquad \kappa(\Omega) \leq \|\mathbf{U}\|\|\mathbf{U}^+\|,$$

where $\mathbf{U}^+$ is the pseudoinverse of $\mathbf{U}$. When $\mathbf{A}$ is Hermitian (or otherwise normal), one can always select an orthogonal basis of eigenvectors, and thus $\kappa(\Omega) = 1$.

On the other hand, nonnormal matrices can have poorly conditioned eigenvector bases (or lack a complete basis altogether). In such situations, $\|\mathbf{U}\|\|\mathbf{U}^+\|$ will be large, and convergence bounds incorporating (3.7) are often pessimistic. The problem typically stems not from a poor bound in (3.7), but from the fact that Ω is too small. Thus we seek bounds for larger Ω. One natural approach is to consider the ε-pseudospectrum of $\mathbf{A}$, defined as

$$\Lambda_\varepsilon(\mathbf{A}) \equiv \{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{A})^{-1}\| \geq \varepsilon^{-1}\},$$

with the convention that $\|(z\mathbf{I} - \mathbf{A})^{-1}\| = \infty$ if z is an eigenvalue of $\mathbf{A}$; see, e.g., [27]. If Ω_ε is a set whose boundary is a finite union of Jordan curves enclosing some components of $\Lambda_\varepsilon(\mathbf{A})$ for a fixed $\varepsilon > 0$, then a standard contour integral argument leads to the bound

$$(3.8)\qquad \kappa(\Omega_\varepsilon) \leq \frac{\mathcal{L}(\partial\Omega_\varepsilon)}{2\pi\varepsilon},$$

where $\mathcal{L}(\partial\Omega_\varepsilon)$ denotes the boundary length of Ω_ε. The ability to adjust ε provides flexibility in our ultimate convergence bounds.

The bounds (3.4) and (3.5) can differ significantly when $\kappa(\Omega_g) \gg 1$. If the good eigenvalues are ill-conditioned (more precisely, if the associated eigenvectors form an ill-conditioned or defective basis for $\mathcal{U}_g$), $\kappa(\Omega_g)$ can be large unless Ω_g extends well beyond the immediate vicinity of the good eigenvalues. However, in taking Ω_g large to reduce $\kappa(\Omega_g)$, the asymptotic convergence rate degrades, since the optimal polynomials in (3.5) are small on Ω_b, while remaining large on Ω_g. Thus, when the good eigenvalues are poorly conditioned, (3.5) can actually improve upon the old bound, as illustrated in section 4.3.

3.3. Convergence Bounds for Restarted Krylov Subspaces.

Having established bounds for the basic case of full orthogonalization, we now address a more pressing issue for practical computations, the potential for attaining gap convergence through polynomial restarting. In particular, we will revise the previous estimates by replacing the starting vector $\mathbf{v}_1$ by $\widehat{\mathbf{v}}_1 \equiv \Phi(\mathbf{A})\mathbf{v}_1$, where Φ is the product of all the previous restart polynomials. We shall assume that the dimension of our restarted Krylov subspace is fixed at $\ell = 2m$. In this case, (3.2) takes the form

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \widehat{\mathbf{v}}_1)) = \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{2m-1}} \frac{\|\phi(\mathbf{A})\Phi(\mathbf{A})\mathbf{v}_1 - \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}. \tag{3.9}$$

We assume that Φ has M distinct roots $\tau_j \in \mathbb{C}\backslash\Omega_g$, and we shall let Ψ be the unique polynomial of degree $M-1$ that interpolates $1/\alpha_g$ at these roots, so that $\Psi(\tau_j) = 1/\alpha_g(\tau_j)$ for $1 \le j \le M$. Now, consider the polynomial

$$1 - \Psi(z)\alpha_g(z).$$

This polynomial is of degree at most $M+m-1$ and has a root at each of the τ_j. Hence, this polynomial must be of the form

$$\widehat{\phi}(z)\Phi(z) \equiv 1 - \Psi(z)\alpha_g(z)$$

for some $\widehat{\phi} \in \mathcal{P}_{m-1}$. Thus, for any given polynomial $\psi \in \mathcal{P}_{m-1}$,

$$\begin{aligned}
\min_{\phi\in\mathcal{P}_{2m-1}} & \frac{\|\phi(\mathbf{A})\Phi(\mathbf{A})\mathbf{v}_1 - \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|} \\
&\le \frac{\|\psi(\mathbf{A})\widehat{\phi}(\mathbf{A})\Phi(\mathbf{A})\mathbf{v}_1 - \psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|} \\
&= \frac{\left\|\psi(\mathbf{A})\widehat{\phi}(\mathbf{A})\Phi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1 + \left[\psi(\mathbf{A})\widehat{\phi}(\mathbf{A})\Phi(\mathbf{A}) - \psi(\mathbf{A})\right]\mathbf{P}_g\mathbf{v}_1\right\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|} \\
&= \frac{\left\|\psi(\mathbf{A})\widehat{\phi}(\mathbf{A})\Phi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1 + \psi(\mathbf{A})\left[\widehat{\phi}(\mathbf{A})\Phi(\mathbf{A}) - \mathbf{I}\right]\mathbf{P}_g\mathbf{v}_1\right\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|} \\
&= \frac{\|\left[\mathbf{I} - \Psi(\mathbf{A})\alpha_g(\mathbf{A})\right]\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1 - \psi(\mathbf{A})\Psi(\mathbf{A})\alpha_g(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|} \\
&= \frac{\|\left[\mathbf{I} - \Psi(\mathbf{A})\alpha_g(\mathbf{A})\right]\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}.
\end{aligned}$$

By the same argument preceding the statement of Theorem 3.3, one has

$$\|\left[\mathbf{I} - \Psi(\mathbf{A})\alpha_g(\mathbf{A})\right]\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\| \le \kappa(\Omega_b) \max_{z\in\Omega_b} |1 - \Psi(z)\alpha_g(z)|\, \|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|,$$

and using this inequality in (3.9) gives

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \widehat{\mathbf{v}}_1)) \leq \left(\max_{\psi \in \mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_b \mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g \mathbf{v}_1\|} \right) \kappa(\Omega_b) \max_{z \in \Omega_b} |1 - \Psi(z)\alpha_g(z)| . \tag{3.10}$$

This analysis is particularly applicable to the implicitly restarted Arnoldi (IRA) method [24], implemented in the ARPACK library [16] and MATLAB's `eigs` command. At the end of every IRA outer iteration, with appropriate choice of the restart dimension, we obtain a $2m$-step Arnoldi factorization. This factorization gives a basis for $\mathcal{K}_{2m}(\mathbf{A}, \widehat{\mathbf{v}}_1)$ with $\widehat{\mathbf{v}}_1 = \Phi(\mathbf{A})\mathbf{v}_1$, where Φ is the product of all of the filter polynomials ϕ_j that have been applied at previous IRA outer iterations. Since we are free to choose the roots of Φ (i.e., the interpolation points τ_j that define Ψ), we should be able to make the quantity

$$\max_{z \in \Omega_b} \left|1 - \Psi(z)\alpha_g(z)\right|$$

arbitrarily small as the degree of Ψ increases, i.e., with further outer iterations.

3.4. Establishing the Asymptotic Convergence Rate. What do bounds (3.4) and (3.10) imply about the asymptotic behavior of Krylov subspace eigenvalue algorithms? In particular, we wish to know how quickly the approximation

$$\min_{p \in \mathcal{P}_{\ell-2m}} \max_{z \in \Omega_b} \left|1 - \alpha_g(z)p(z)\right|$$

goes to zero with increasing ℓ, and, for restarted iterations, how to select the polynomial Ψ to minimize

$$\max_{z \in \Omega_b} \left|1 - \Psi(z)\alpha_g(z)\right|.$$

We begin by recalling a basic result from classical approximation theory (see, e.g., [10, 29]). Consider the behavior of

$$\min_{p \in \mathcal{P}_k} \max_{z \in \Omega_b} \left|f(z) - p(z)\right| \tag{3.11}$$

as $k \to \infty$, where f is some function analytic on Ω_b. First, suppose Ω_b is the unit disk, $\Omega_b = \{|z| \leq 1\}$, and let z_0 be the singularity of f with smallest modulus. Expand f in a Taylor series about $z = 0$ and approximate the optimal degree-k polynomial by the first k terms of the series. From the Taylor remainder formula we conclude that

$$\limsup_{k \to \infty} \min_{p \in \mathcal{P}_k} \max_{|z| \leq 1} \left|f(z) - p(z)\right|^{1/k} \leq \frac{1}{|z_0|}.$$

In fact, one can replace the inequality with equality, for although there are usually better choices for p than the Taylor polynomial, no such choice does better asymptotically. Thus, we say that (3.11) converges at the asymptotic rate $1/|z_0|$. The further the singularity z_0 is from Ω_b, the faster the convergence rate.

Now let Ω_b be any connected set whose boundary $\partial\Omega_b$ is a Jordan curve. The Riemann mapping theorem ensures the existence of a conformal map $\mathcal{G}$ taking the exterior of Ω_b to the exterior of the unit disk with $\mathcal{G}(\infty) = \infty$ and $\mathcal{G}'(\infty) > 0$. We will use the map $\mathcal{G}$ to reduce the present Ω_b to the simpler unit disk case. In particular, the convergence rate now depends on the modulus of the image of the singularities of

f. We set $f(z) = 1/\alpha_g(z)$, so the singularities of f are simply the good eigenvalues of $\mathbf{A}$. In particular, define

$$\rho \equiv \big(\min_{j=1,\ldots,L} |\mathcal{G}(\lambda_j)| \big)^{-1}.$$

We may then apply a result from classical approximation theory to characterize the asymptotic quality of polynomial approximations to $1/\alpha_g$ [10, 29].

Theorem 3.4.

$$\limsup_{k\to\infty} \min_{p\in\mathcal{P}_k} \max_{z\in\Omega_b} \left| \frac{1}{\alpha_g(z)} - p(z) \right|^{1/k} = \rho.$$

The image of the circle $\{|z| = \rho^{-1}\}$ forms a curve $\mathcal{C} \equiv \mathcal{G}^{-1}(\{|z| = \rho^{-1}\})$ exterior to Ω_b. This critical curve contains at least one good eigenvalue, with all bad and no good eigenvalues in its interior. An example of this mapping is given in Figure 4.1 of the next section. Moving a good eigenvalue anywhere on $\mathcal{C}$ has no effect on the convergence rate. For the approximation problem in (3.4), we have

$$\min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} \big|1 - \alpha_g(z)p(z)\big| \le \alpha_0 \min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} \big|1/\alpha_g(z) - p(z)\big|,$$

where $\alpha_0 \equiv \max_{z\in\Omega_b} |\alpha_g(z)|$. Thus Theorem 3.4 implies

$$\begin{aligned}\limsup_{\ell\to\infty} \min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} &\big|1 - \alpha_g(z)p(z)\big|^{1/\ell} \\ &\le \Big(\lim_{\ell\to\infty} \alpha_0^{1/\ell} \Big) \limsup_{\ell\to\infty} \left[\min_{p\in\mathcal{P}_{\ell-2m}} \max_{z\in\Omega_b} \big|1/\alpha_g(z) - p(z)\big| \right]^{1/\ell} \\ &= \rho.\end{aligned}$$

Of course, asymptotic results for Krylov iterations without restarts must be put in the proper perspective: $\mathcal{U}_g \subseteq \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ for some finite ℓ, implying $\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) = 0$. Our primary goal is to obtain an asymptotic result for restarted iterations, where by restricting the subspace dimension we generally do not obtain exact convergence. Instead, we strive to drive $\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \widehat{\mathbf{v}}_1))$ to zero by judiciously choosing the restart polynomial Φ, where $\widehat{\mathbf{v}}_1 = \Phi(\mathbf{A})\mathbf{v}_1$. In particular, we wish to mimic the optimization in Theorem 3.4 by constructing Φ to interpolate $1/\alpha_g$ at *asymptotically optimal* points in Ω_b. The following are some well-known choices for these points:

- Fejér points of order k: $\{\mathcal{G}^{-1}(z) : z^k = 1\}$;
- Fekete points of order k: $\{z_1, \ldots, z_k\} \subseteq \Omega_b$ that maximize $\prod_{j\neq k} |z_j - z_k|$;
- Leja points: Given $\{z_1, \ldots, z_{k-1}\}$, set $z_k \equiv \arg\max_{z\in\Omega_b} \prod_{j=1}^{k-1} |z - z_j|$.

In all cases, these points fall on the boundary of Ω_b. Given $\mathcal{G}$, the Fejér points are simplest to compute, while the Leja points are the most straightforward to implement in software [1], as increasing the approximating polynomial degree simply adds new Leja points without altering the previous ones. In contrast, all Fejér and Fekete points typically change as the order increases. The following classical result can be found in [10, sect. II.3] and the related papers [9, 19].

Theorem 3.5. *Let $q_M \in \mathcal{P}_{M-1}$ be a polynomial that interpolates $1/\alpha_g$ at order-M Fejér, Fekete, or Leja points for Ω_b. Then*

$$\limsup_{M\to\infty} \max_{z\in\Omega_b} \left| \frac{1}{\alpha_g(z)} - q_M(z) \right|^{1/M} = \rho.$$

This interpolation result immediately gives an asymptotic convergence bound on the right of inequality (3.10) for restarted Krylov methods.

Corollary 3.6. *Let Ψ_M interpolate $1/\alpha_g(z)$ at the order-M Fejér, Fekete, or Leja points for Ω_b. Then*

$$\limsup_{M\to\infty}\left[\max_{z\in\Omega_b}|1-\Psi_M(z)\alpha_g(z)|\right]^{1/M}=\rho.$$

Thus, the restarted iteration recovers the asymptotic convergence rate ρ. In practice, we have $M=\nu m$ after ν outer iterations, each of which is restarted with a degree-m polynomial. Every outer iteration should, in the asymptotic regime, decrease the residual by the factor ρ^m. (In practice, one is not restricted to degree-m polynomials—we simply fixed this degree to simplify the derivation. Increasing the dimension beyond m has no effect on our convergence analysis.)

If we convert the lim sup statement into a direct bound on $\delta(\mathcal{U}_g,\mathcal{K}(\mathbf{A},\widehat{\mathbf{v}}_1))$, we obtain

$$\begin{aligned}\delta(\mathcal{U}_g,\mathcal{K}_\ell(\mathbf{A},\widehat{\mathbf{v}}_1)) &\le \left(\max_{\psi\in\mathcal{P}_{m-1}}\frac{\|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|}\right)\ \kappa(\Omega_b)\ \max_{z\in\Omega_b}|1-\Psi(z)\alpha_g(z)|\\ &\le C_1\,C_2\,C_r\,r^M\end{aligned}$$

for any $r>\rho$, where $C_1=\max_{\psi\in\mathcal{P}_{m-1}}\|\psi(\mathbf{A})\mathbf{P}_b\mathbf{v}_1\|/\|\psi(\mathbf{A})\mathbf{P}_g\mathbf{v}_1\|$ and $C_2=\kappa(\Omega_b)$. The constant $C_r=C_r(\alpha_g,\Omega_b)$ accounts for transient effects in the polynomial approximation problem [10, sect. II.2]. The constant C_2 incorporates the nonnormality of $\mathbf{A}$ acting only on $\mathcal{U}_b$; nonnormality associated with both good and bad eigenvalues influences the constant C_1, which describes the bias in the starting vector toward $\mathcal{U}_g$.

In summary, a restarted iteration can recover the same asymptotic convergence rate predicted for full orthogonalization iteration with a fixed Ω_b. This comforting conclusion hides several subtleties. First, the restarted iteration locks in a fixed Ω_b through its construction of the restart polynomial Φ. Without restarts, on the other hand, one is free to choose Ω_b to optimize the bound (3.4) for each iteration. At early stages, a large Ω_b may yield a small $\kappa(\Omega_b)$ but a slow rate; later in the iteration, a reduced Ω_b can give a sufficiently improved rate to compensate for the corresponding increase in $\kappa(\Omega_b)$. Second, the restarted iteration must somehow determine the set Ω_b. Precise a priori information is rare, so Ω_b must be found adaptively. This has been successfully implemented for Hermitian problems using Leja points [1, 5], and similar ideas have been advanced for general matrices [12]. In practice, a different approach not explicitly derived from potential theory, called *exact shifts* [24], has proved to be very effective. As seen in the experiments of section 4.4, exact shifts can effectively determine the region Ω_b.

4. Examples.

We now demonstrate the accuracy of the convergence bound (3.4) and the use of related potential-theoretic tools in a variety of circumstances, with matrices ranging from Hermitian to far from normal. Our examples complement those provided in [3, sect. 6]. This section closes with a performance comparison of restarting strategies for non-Hermitian examples with various Ω_b. In all cases, the Krylov subspaces are generated from the starting vector $\mathbf{v}_1=(1,\,1,\,\ldots,1)^T$.

4.1. Schematic Illustration.

Our first example illustrates use of the tools of section 3.4 to predict the asymptotic convergence rate of full Krylov iterations. We construct $\mathbf{A}$ to be a normal matrix whose spectrum comprises 1000 bad eigenvalues that randomly cover an arrow-shaped region in the complex plane with uniform probability, together with the three rightmost good eigenvalues, $\{-\frac{1}{4},0,\frac{1}{4}\}$, well-separated from

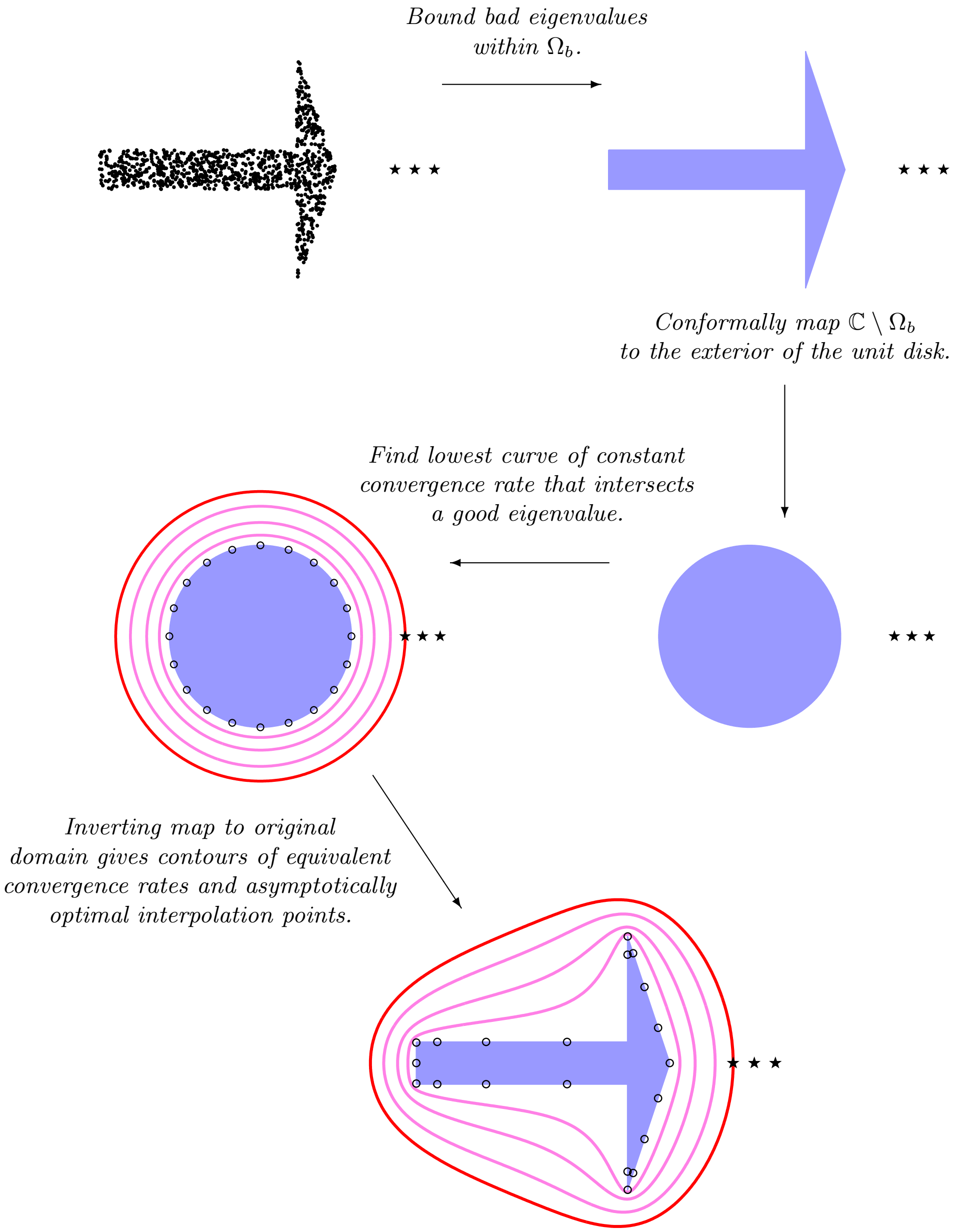
Bound bad eigenvalues within Ω_b.
Conformally map $\mathbb{C} \setminus \Omega_b$ to the exterior of the unit disk.
Find lowest curve of constant convergence rate that intersects a good eigenvalue.
Inverting map to original domain gives contours of equivalent convergence rates and asymptotically optimal interpolation points.

Fig. 4.1 *Schematic illustration showing calculation of the asymptotic convergence rate. The rate predicted by* (3.4) *would not change if new good eigenvalues were added on or outside the outermost curve in the final image. The small circles on the last two images show the Fejér points of order* 20 *for the exterior of the unit circle and the arrow, respectively.*

the arrow. Without loss of generality, we take $\mathbf{A}$ to be diagonal. (Section 4.4 illustrates the complexities that nonnormality and restarting introduce to this example.)

Figure 4.1 demonstrates the procedure outlined in section 3.4 for estimating the asymptotic convergence rate. The bad eigenvalues are enclosed within the region Ω_b, taken to be the arrow over which the bad eigenvalues are distributed. The ex-

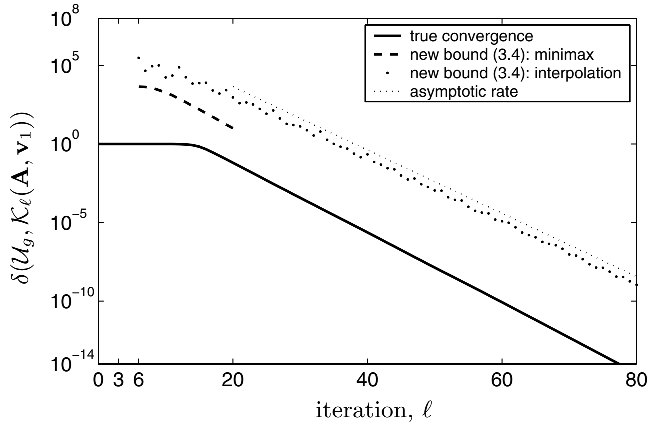

Fig. 4.2 *Gap convergence and two bounds based on* (3.4). *The better bound solves the minimax approximation problem directly, while the lesser bound approximates the optimal polynomial by interpolating* $1/\alpha_g$ *at Fejér points. Although this approximation procedure degrades the bound, it does not affect the asymptotic convergence rate.*

terior of this region is conformally mapped to the exterior of the unit disk. Since Ω_b is a polygon, we can compute this map $\mathcal{G}$ using a Schwarz–Christoffel transformation, implemented in Driscoll's SC Toolbox [6]. In this domain, the polynomial approximation problem is straightforward: the convergence rate is determined by the modulus of the singularities $\mathcal{G}(\lambda_j)$ alone. Thus, level sets of constant convergence rate are simply concentric circles. Applying $\mathcal{G}^{-1}$ to any of these level sets gives a curve of constant convergence rate exterior to the original Ω_b domain. If additional good eigenvalues were added on or beyond this critical level curve, the predicted asymptotic convergence rate would not change. In the present case, the predicted asymptotic convergence rate is approximately 0.629.[2] Driscoll, Toh, and Trefethen apply similar potential-theoretic ideas to the iterative solution of linear systems [7].

Figure 4.2 shows the bound (3.4) for this example. This figure compares true gap convergence to two versions of the new bound. For the most accurate bound, shown as a broken line, the minimax approximation problem of (3.4) is solved exactly using the COCA package [8] to compute best uniform approximation to $f(z) \equiv 1$ by polynomials of the form $\alpha_g(z)p(z)$. Using a monomial basis for $p(z)$, this procedure becomes highly ill-conditioned as the degree increases, so we only show results for early iterations. As an alternative, we illustrate an upper bound on (3.4) obtained by replacing the optimal polynomial $p \in \mathcal{P}_{\ell-2m}$ by the polynomial that interpolates $1/\alpha_g$ at the order $\ell - 2m + 1$ Fejér points of the arrow. (The order-20 Fejér points are shown as small circles in the final image of Figure 4.1.) These points were computed using the SC Toolbox, followed by high-precision polynomial interpolation in *Mathematica*. As they are asymptotically optimal interpolation points, the convergence rate obtained by the interpolation procedure must match that predicted by the conformal map and that realized by exactly solving the minimax problem in (3.4). Indeed, this is observed in Figure 4.2, though the Fejér bound is roughly two orders of magnitude larger than the optimal minimax bound.

[2] In practice, one is more likely to have only an estimate, say, for the convex hull of the spectrum. Replacing the arrow by its convex hull increases the predicted convergence rate to approximately 0.647.

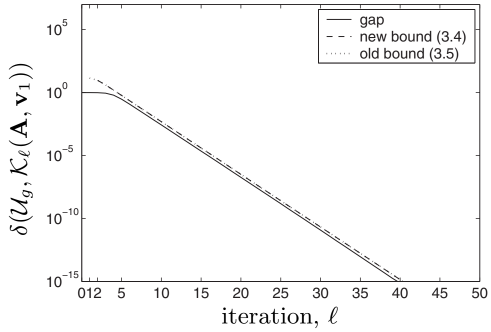

Fig. 4.3 *Comparison of convergence bounds for the Hermitian example with a single good eigenvalue. In this special case, the bounds* (3.4) *and* (3.5) *are identical.*

4.2. Hermitian Examples. We next examine the bound (3.4) applied to several Hermitian examples and compare results to the bound (3.5) from [3]. In this situation, and indeed for any normal $\mathbf{A}$, one should take Ω_g to be the set of good eigenvalues, giving $\kappa(\Omega_g) = 1$. Hence (3.5) is superior to (3.4), as we established in equation (3.6).

Let $\mathbf{A}$ be a diagonal matrix with 200 bad eigenvalues uniformly distributed on $\Omega_b = [-1, 0]$. First suppose there is one good eigenvalue, $\lambda_1 = 1/4$, so Theorem 3.3 reduces to

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le \frac{\|\mathbf{P}_b\mathbf{v}_1\|}{\|\mathbf{P}_g\mathbf{v}_1\|} \min_{p\in\mathcal{P}_{\ell-2}} \max_{z\in\Omega_b} \Big|1 - p(z)(z-\lambda_1)\Big|,$$

while the bound (3.5) reduces to

$$\delta(\mathcal{U}_g, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le \frac{\|\mathbf{P}_b\mathbf{v}_1\|}{\|\mathbf{P}_g\mathbf{v}_1\|} \min_{q\in\mathcal{P}_{\ell-1}} \frac{\max\{|q(z)| : z \in \Omega_b\}}{|q(\lambda_1)|}.$$

Here we have used the fact that $\kappa(\Omega_b) = \kappa(\Omega_g) = 1$ since $\mathbf{A}$ is Hermitian, and hence normal. As noted in section 3.1, the two bounds are identical in this $m = 1$ case:

$$\min_{q\in\mathcal{P}_{\ell-1}} \frac{\max\{|q(z)| : z \in \Omega_b\}}{|q(\lambda_1)|} = \min_{\substack{q\in\mathcal{P}_{\ell-1}\\ q(\lambda_1)=1}} \max_{z\in\Omega_b} |q(z)|.$$

(This is essentially the same polynomial approximation problem as in Saad's Proposition 2.1 [22].) Posed over $\Omega_b = [-1, 0]$, this optimization problem is solved by suitably normalized and shifted Chebyshev polynomials. Figure 4.3 illustrates the results.

How do the bounds evolve as the number of good eigenvalues increases? If additional good eigenvalues are added to the right of $\lambda_1 = 1/4$, the bounds are no longer identical. Since the Chebyshev polynomials used to approximate zero on $\Omega_b = [-1, 0]$ in the single good eigenvalue case grow monotonically in magnitude outside Ω_b, we can use those same polynomials to approximate the term

$$\min_{q\in\mathcal{P}_{\ell-m}} \frac{\max\{|q(z)| : z \in \Omega_b\}}{\min\{|q(z)| : z \in \Omega_g\}}$$

in (3.5). Since $\mathbf{A}$ is normal, one should take Ω_g to be the set of good eigenvalues. The addition of new eigenvalues to the right of 1/4 will not alter the *asymptotic*

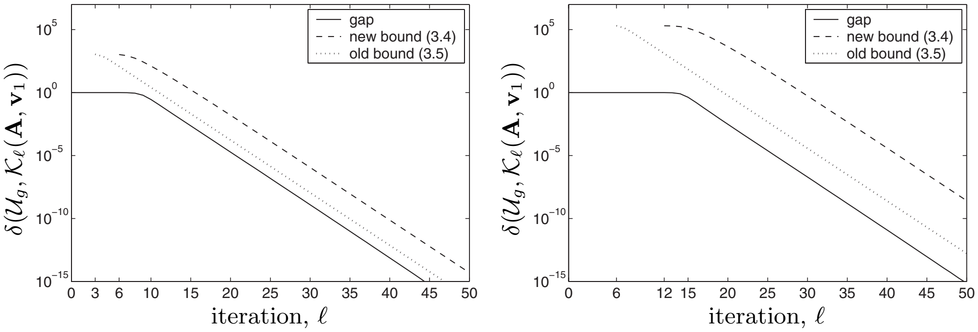

Fig. 4.4 *Comparison of convergence bounds for the Hermitian example again, but now with three good eigenvalues (left) and six good eigenvalues (right). As the number of good eigenvalues increases, the new bound degrades in comparison with* (3.5)*, but the predicted asymptotic rate remains accurate. Note the transient stagnation of the new bound due to the optimization over* $\mathcal{P}_{\ell-2m}$ *rather than* $\mathcal{P}_{\ell-m}$ *for the six-eigenvalue case.*

convergence rate derived from (3.5). The same is true for (3.4): the critical factor determining that convergence rate is the singularity in $1/\alpha_g$ nearest to Ω_b. Adding new eigenvalues to the right of $1/4$ adds more distant singularities to $1/\alpha_g$ without altering the asymptotics.

Though neither convergence rate degrades, the new bound predicts a longer transient phase before the asymptotic rate is realized. This delay, together with the fact that (3.4) gives up *two* polynomial degrees in the approximation problem for every good eigenvalue (the optimization is over $p \in \mathcal{P}_{\ell-2m}$, as opposed to $q \in \mathcal{P}_{\ell-m}$ in (3.5)), causes the new bound to degrade as m grows, though the convergence rate remains descriptive. Figure 4.4 illustrates these properties, first for three good eigenvalues, $\{\frac{1}{4}, \frac{3}{8}, \frac{1}{2}\}$, and then for six good eigenvalues, $\{\frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\}$.

4.3. Defective Examples. Our next examples illustrate the new bound (3.4) for nondiagonalizable matrices, illustrating the use of pseudospectra to compute convergence bounds. We include a situation where (3.4) is superior to (3.5) for a matrix with good eigenvalues that are highly sensitive to perturbations.

First, consider the matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{J}_6(0,1) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{100}(-\frac{5}{2},1) \end{pmatrix},$$

where $\mathbf{J}_k(\lambda,\gamma)$ denotes a k-dimensional Jordan block with eigenvalue λ and off-diagonal entry γ,

$$\mathbf{J}_k(\lambda,\gamma) = \begin{pmatrix} \lambda & \gamma & & \\ & \lambda & \ddots & \\ & & \ddots & \gamma \\ & & & \lambda \end{pmatrix} \in \mathbb{C}^{k\times k};$$

all unspecified entries are zero.

We seek the good eigenvalue $\lambda_1 = 0$, a defective eigenvalue with multiplicity $m = 6$. Since $\mathbf{A}$ is nondiagonalizable, we must take Ω_b and Ω_g to be larger sets than

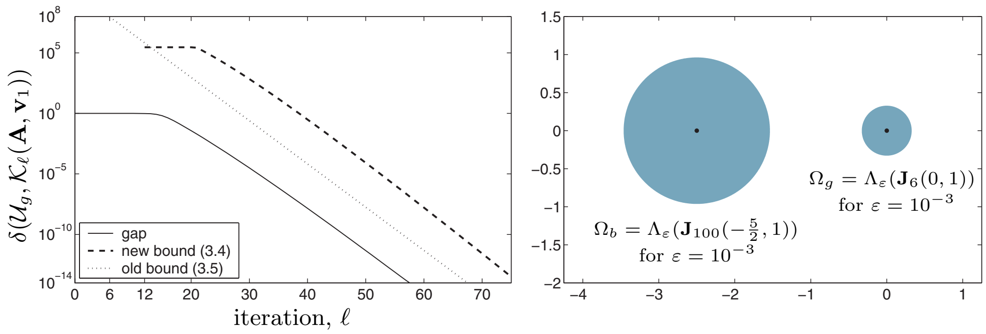

Fig. 4.5 *Comparison of convergence bounds for a nondiagonalizable example. The containing sets are defined as taking $\Omega_g = \Lambda_\varepsilon(\mathbf{J}_6(0,1))$ and $\Omega_b = \Lambda_\varepsilon(\mathbf{J}_{100}(-\frac{5}{2},1))$ for $\varepsilon = 10^{-3}$. The new bound (3.4) shown here is based on an approximation of the optimal polynomial by the interpolant to $1/\alpha_g$ at Fejér points. The right plot shows Ω_g and Ω_b, with the eigenvalues appearing as dots in the center of each circular region.*

the eigenvalues themselves to get finite values for $\kappa(\Omega_b)$ and $\kappa(\Omega_g)$. The pseudospectra of Jordan blocks $\mathbf{J}_k(\lambda,\varepsilon)$ are exactly circular [20] and thus provide convenient choices for Ω_g and Ω_b. We take $\Omega_g = \Lambda_\varepsilon(\mathbf{J}_6(0,1))$ and $\Omega_b = \Lambda_\varepsilon(\mathbf{J}_{100}(-\frac{5}{2},1))$ for $\varepsilon = 10^{-3}$ in both cases. Figure 4.5 illustrates the corresponding convergence bounds. Here, for the bound (3.4) we actually show an upper bound obtained by replacing the optimal polynomial in (3.4) by the polynomial that interpolates $1/\alpha_g$ at Fejér points for Ω_b.

We emphasize that the choice of Ω_g plays no role in the new bound (3.4). It does, however, affect the asymptotic convergence rate of the bound (3.5); taking for Ω_g pseudospectral sets with smaller values of ε will improve the asymptotic convergence rate (better for later iterations), but increase the leading constant (worse for early iterations). The value $\varepsilon = 10^{-3}$ is a good balance for the range of iterations shown here. Regardless of the choice of Ω_g, the asymptotic rate never beats the one derived from the new bound (3.4).

Now suppose the bad eigenvalue remains the same, but we increase the sensitivity of the good eigenvalue, replacing $\mathbf{J}_6(0,1)$ with $\mathbf{J}_6(0,100)$. The only effect this has on the new bound (3.4) is a slight change in the constant C_1 describing bias in the starting vector. (The same change also effects (3.5).) Since the location and multiplicity of the eigenvalue have not changed, α_g remains as before, as does the polynomial approximation problem, and hence the asymptotic convergence rate from (3.4).

The bound (3.5), on the other hand, changes significantly. Enlarging the off-diagonal entry γ in the good Jordan block corresponds to a significant increase in the size of the pseudospectral set $\Omega_g = \Lambda_\varepsilon(\mathbf{J}_6(0,\gamma))$. In particular, replacing $\gamma = 1$ by $\gamma = 100$ increases the radius of $\Omega_g = \Lambda_\varepsilon(\mathbf{J}_6(0,\gamma))$ by a factor of roughly 45. We can't use $\varepsilon = 10^{-3}$ for both Ω_g and Ω_b, as the two sets would intersect, and thus the approximation problem in (3.5) would predict no convergence. Instead, we fix $\Omega_b = \Lambda_\varepsilon(\mathbf{J}_{100}(-\frac{5}{2},1))$ for $\varepsilon = 10^{-3}$, but reduce the value of ε used to define $\Omega_g = \Lambda_\varepsilon(\mathbf{J}_6(0,100))$. In particular, we must take $\varepsilon \approx 10^{-9}$ before Ω_g and Ω_b are disjoint. We find that using $\varepsilon = 10^{-13}$ for Ω_g provides a good bound, and it is this value we use in Figure 4.6. Increasing γ in the good Jordan block dramatically increases the constant term in the bound (3.5). Taking γ ever larger shows that (3.5) can be arbitrarily worse than (3.4) in this special situation.

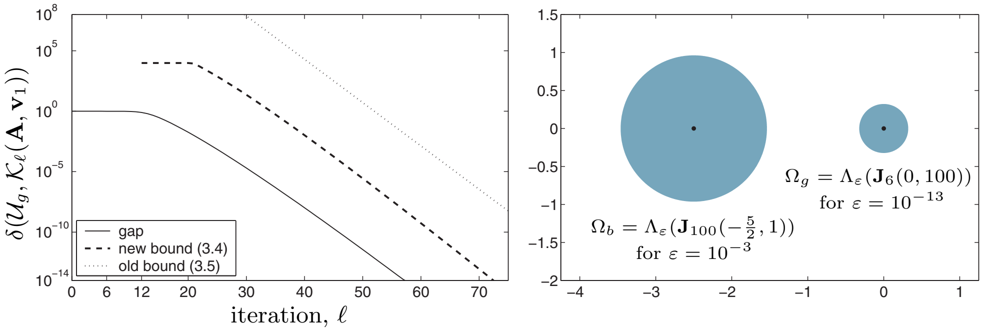

Fig. 4.6 *Analogue of Figure* 4.5*, but with the good Jordan block* $\mathbf{J}_6(0, 1)$ *replaced by* $\mathbf{J}_6(0, 100)$*, thus increasing the sensitivity of the good eigenvalues. Now the new bound* (3.4) *is superior to* (3.5).

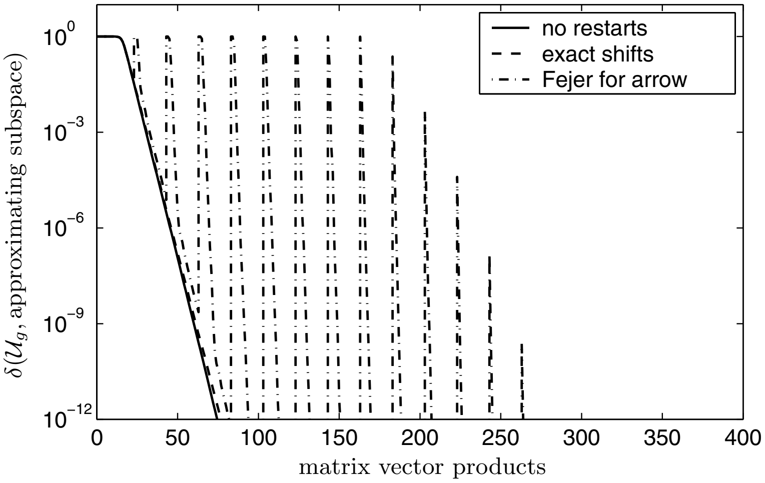

Fig. 4.7 *First restart example: Exact shifts give similar convergence to the optimal iteration with no restarts; restarting with shifts at the degree-*20 *Fejér points for the arrow-shaped* Ω_b *gives slower convergence. The saw-toothed shape indicates that accurate estimates of the good invariant subspace are lost when a restart is performed.*

4.4. Polynomial Restart Examples. Our final examples illustrate the performance of restarted iterations applied to nonnormal matrices. In particular, we modify the example from section 4.1: $\mathbf{A}$ is the direct sum of $\operatorname{diag}(-\frac{1}{4}, 0, \frac{1}{4})$, which contains the perfectly conditioned good eigenvalues, and $\mathbf{A}_{\rm bad} \in \mathbb{C}^{1000\times 1000}$, whose diagonal contains the same bad eigenvalues shown in Figure 4.1, ordered by increasing real part. Unlike the example in section 4.1, we add entries to the first and second superdiagonal of $\mathbf{A}_{\rm bad}$, making the matrix nonnormal. We are interested in the performance of shifting strategies as this nonnormality varies.

First, place $-1/2$ on the first superdiagonal and $-1/4$ on the second diagonal of any row of $\mathbf{A}_{\rm bad}$ with diagonal entry λ with $\operatorname{Re}\lambda < -3$. This makes the leftmost part of the spectrum highly sensitive to perturbations, with essentially no impact on the good eigenvalues, which are well-separated from these sensitive eigenvalues. Figure 4.7 shows gap convergence for three different iterations: no restarts, restarting

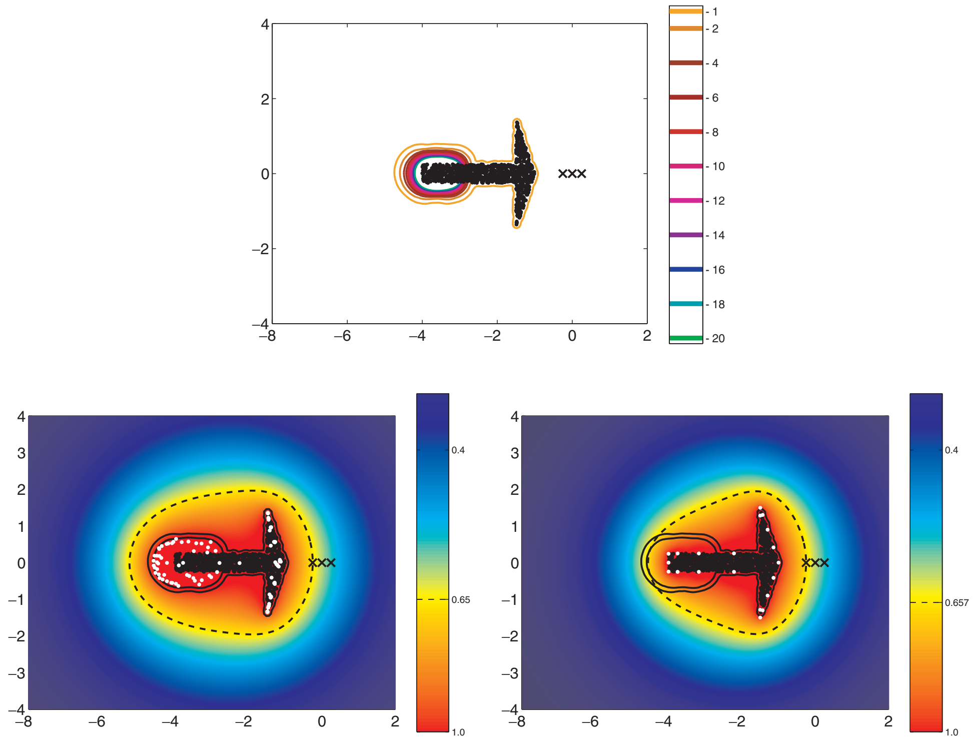

Fig. 4.8 *On the top, ε-pseudospectrum of $\mathbf{A}_{\text{bad}}$ for the first restart example, with $\varepsilon = 10^{-1}, 10^{-2}, 10^{-4}, \ldots, 10^{-20}$. The bad eigenvalues fill an arrow-shaped region; the three good eigenvalues are marked by ×. The bottom plots show the magnitude of the aggregate exact shift polynomial (left) and the degree-20 Fejér polynomial (right). Roots of these polynomials are shown as white dots. Areas with the same color yield the same convergence rate; the broken line marks the lowest level curve of the restart polynomial that passes through a good eigenvalue. The solid lines denote the boundaries of the 10^{-1}- and 10^{-2}-pseudospectra of $\mathbf{A}_{\text{bad}}$.*

with exact shifts, and restarting with Fejér points. These last two methods require some explanation. In both cases, the Krylov subspace is built out to dimension 23, and at the end of each outer iteration, the current starting vector is refined with a polynomial filter, $\mathbf{v}_1 \leftarrow \phi(\mathbf{A})\mathbf{v}_1$, where $\deg(\phi) = 20$. For exact shifts, the roots of ϕ are taken to be the 20 leftmost Arnoldi Ritz values determined from the degree-23 Krylov subspace. For Fejér shifts, the roots of ϕ are the order-20 Fejér points on the boundary of Ω_b,[3] which we take to be the arrow that tightly covers the bad eigenvalues, as shown in Figure 4.1. Exact shifts closely capture the performance of the full iteration, while the Fejér shifts exhibit a sawtooth convergence curve due to the fact that the full 23-dimensional Krylov subspace contains accurate estimates of $\mathcal{U}_g$, but these degrade upon restarting.

Figure 4.8 compares pseudospectra of $\mathbf{A}_{\text{bad}}$ (top) with the relative magnitude of the aggregate restart polynomial Φ for exact shifts and Fejér shifts. The broken line

[3] Strictly speaking, to obtain asymptotically optimal performance we should not repeatedly apply the same degree-20 Fejér polynomial, but rather use Leja points that add distinct new shifts at each outer iteration. We expect our simpler approach to yield qualitatively similar behavior.

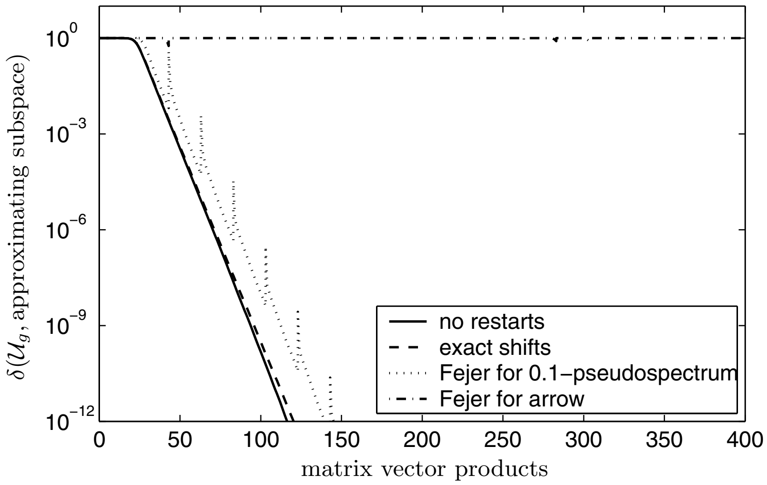

Fig. 4.9 *Second restart example: Fejér points for the arrow-shaped set Ω_b tightly enclosing the eigenvalues yield little convergence, while exact shifts and Fejér points for the $\varepsilon = 10^{-1}$ pseudospectrum both give convergence comparable to the full orthogonalization method. Hence, imprecise spectral information leads to more rapid convergence than precise spectral information.*

on these plots shows the critical curve that determines the convergence rate. The asymptotic convergence rates are very similar for these iterations, so why does the Fejér approach (where nonnormality has no influence on shift selection) fare so much worse in Figure 4.7? Note that there are points in the 10^{-2}-pseudospectrum of $\mathbf{A}_{\text{bad}}$ that are outside the critical level curve that determines the convergence rate (broken line). Krylov methods, at early iterations, are drawn to such false approximate eigenvalues, which appear more prominent than the good eigenvalues. The exact shifts avoid this difficulty: the critical level curve for the potential they generate includes all points in the 10^{-1}-pseudospectrum, and some beyond.

Next, modify the previous example by changing the $-1/2$ and $-1/4$ entries in the superdiagonal of $\mathbf{A}_{\text{bad}}$ to -2 and -1, respectively. We repeat the same restarting experiments as before, with results shown in Figure 4.9. The leftmost eigenvalues remain highly sensitive to perturbations, but now in a fashion rather different geometrically from the previous case, as can be seen in the pseudospectral plot in Figure 4.10. Exact shifts perform nearly as well as the full iteration, but now the Fejér shifts for the arrow enclosing the eigenvalues do not lead to any notable convergence in these iterations. There are points in the 10^{-20}-pseudospectrum of $\mathbf{A}_{\text{bad}}$ that are well outside the critical convergence level curve. Though we predict a superior convergence bound for these Fejér shifts, the constant $\kappa(\Omega_b)$ is enormous. On the other hand, if we take Ω_b to be the 10^{-1}-pseudospectrum of $\mathbf{A}_{\text{bad}}$, then the Fejér points for this set yield convergence similar to that realized by exact shifts. (These Fejér shifts are derived with knowledge of the nonnormality of $\mathbf{A}_{\text{bad}}$.) This is another example where inexact knowledge of the eigenvalues (as derived via the exact shifts, which are Ritz values [24]) leads to markedly better performance than that obtained by exploiting exact spectral information. For similar observations in the context of solving linear systems of equations, see [17].

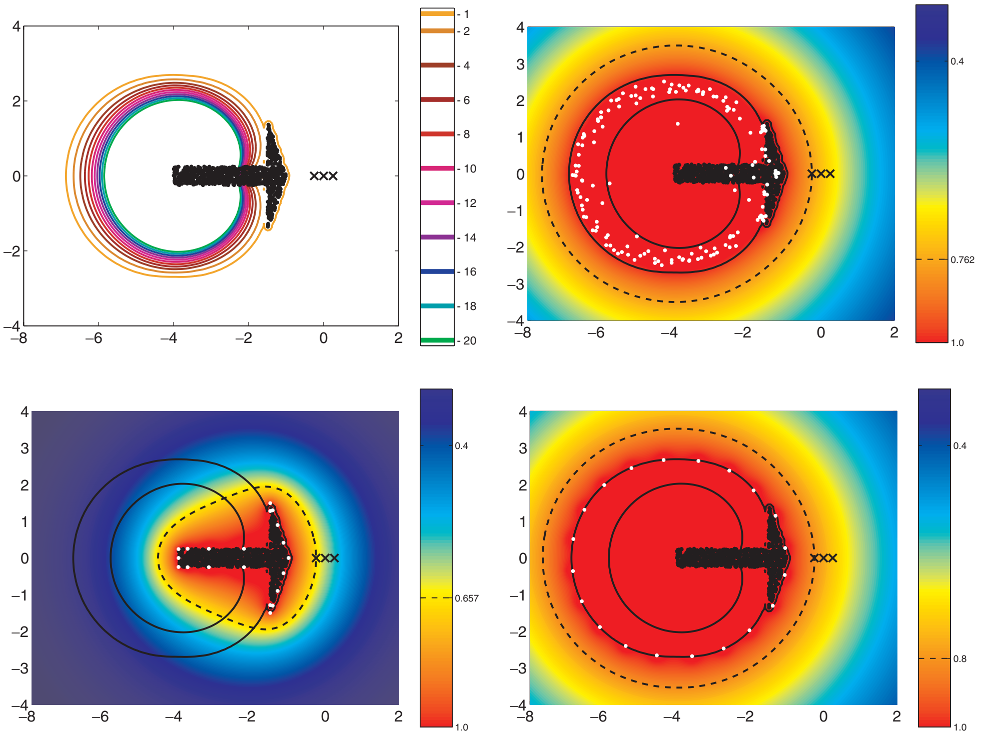

Fig. 4.10 *On the top left, ε-pseudospectrum of $\mathbf{A}_{\rm bad}$ for the second restart example, with $\varepsilon = 10^{-1}, 10^{-2}, 10^{-4}, \ldots, 10^{-20}$. The bad eigenvalues fill an arrow-shaped region; the three good eigenvalues are denoted by $\times$. The top right and bottom plots show the magnitude of the aggregate exact shift polynomial (top right), the degree-20 Fejér polynomial for the arrow (bottom left), and the degree-20 Fejér polynomial for the 10^{-1}-pseudospectrum (bottom right). The solid lines now denote the boundaries of the 10^{-1}- and 10^{-20}-pseudospectra of $\mathbf{A}_{\rm bad}$. The Fejér shifts for the arrow give essentially no convergence for these iterations: though the predicted asymptotic convergence is better than the others, the nonnormality constant $\kappa(\Omega_b)$ will be enormous.*

Conclusions. Our bounds for Krylov subspace eigenvalue algorithms describe convergence in terms of three primary components:

- a constant influenced by the starting vector, the location of the eigenvalues, and nonnormality;
- a constant that describes the conditioning of the undesired eigenvalues;
- a linear convergence rate that depends on the separation of the desired eigenvalues from the remainder of the spectrum.

More complicated behavior is often observed in practice: convergence frequently accelerates as the iteration proceeds, eventually yielding convergence that is better than any fixed linear rate. (For analysis of such "superlinear" convergence, see [3].) For nonnormal problems, this behavior can be preceded by an early "sublinear" phase of apparent stagnation.

The bound (3.10) suggests that a restarted Krylov iteration can recover the linear convergence rate of the full iteration with no restarts, provided the polynomial filters are designed according to potential-theoretic principles. In practice, one does not have access to sufficient spectral information to determine such restart polynomials a priori;

instead, one restarts using "exact shifts" [16, 24]. While our analysis does not explicitly address this restarting scheme, the computational examples in section 4 illustrate its success. A rigorous convergence theory for exact shifts currently exists only in the Hermitian setting [24]. The development of such a theory for non-Hermitian matrices remains an important open problem.

Acknowledgments. We thank Rich Lehoucq and Jörg Liesen for their constructive comments. The pseudospectra in section 4 were computed using software developed by Trefethen and Wright [28, 30].

REFERENCES

[1] J. Baglama, D. Calvetti, and L. Reichel, *Iterative methods for the computation of a few eigenvalues of a large symmetric matrix*, BIT, 36 (1996), pp. 400–421.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[3] C. Beattie, M. Embree, and J. Rossi, *Convergence of restarted Krylov subspaces to invariant subspaces*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1074–1109.

[4] E. A. Burroughs, L. A. Romero, R. B. Lehoucq, and A. G. Salinger, *Linear stability of flow in a differentially heated cavity via large-scale eigenvalue calculations*, Internat. J. Numer. Methods Heat Fluid Flow, 14 (2004), pp. 803–822.

[5] D. Calvetti, L. Reichel, and D. C. Sorensen, *An implicitly restarted Lanczos method for large symmetric eigenvalue problems*, Electron. Trans. Numer. Anal., 2 (1994), pp. 1–21.

[6] T. A. Driscoll, *A* Matlab *toolbox for Schwarz–Christoffel mapping*, ACM Trans. Math. Software, 22 (1996), pp. 168–186. Associated software available from http://www.math.udel.edu/~Driscoll/SC/.

[7] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.

[8] B. Fischer and J. Modersitzki, *An algorithm for complex linear approximation based on semi-infinite programming*, Numer. Algorithms, 5 (1993), pp. 287–297. Associated software available from http://www.math.mu-luebeck.de/workers/modersitzki/COCA/coca5.html.

[9] B. Fischer and L. Reichel, *Newton interpolation in Fejér and Chebyshev points*, Math. Comp., 53 (1989), pp. 265–278.

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

[11] F. R. Gantmacher, *Theory of Matrices*, Vol. 1, 2nd ed., Chelsea, New York, 1959.

[12] V. Heuveline and M. Sadkane, *Arnoldi–Faber method for large non-Hermitian eigenvalue problems*, Electron. Trans. Numer. Anal., 5 (1997), pp. 62–76.

[13] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 843–862.

[14] T. Kato, *Perturbation Theory for Linear Operators*, corrected 2nd ed., Springer-Verlag, Berlin, 1980.

[15] R. B. Lehoucq, *Implicitly restarted Arnoldi methods and subspace iteration*, SIAM J. Matrix. Anal. Appl., 23 (2001), pp. 551–562.

[16] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998. Associated software available from http://www.caam.rice.edu/software/ARPACK.

[17] 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.

[18] V. Paulsen, *Completely Bounded Maps and Operator Algebras*, Cambridge University Press, Cambridge, UK, 2002.

[19] L. Reichel, *Newton interpolation at Leja points*, BIT, 30 (1990), pp. 332–346.

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

[21] M. Robbé and M. Sadkane, *A priori error bounds on invariant subspace approximations by block Krylov subspaces*, Linear Algebra Appl., 350 (2002), pp. 89–103.

[22] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[23] V. Simoncini, *Ritz and pseudo-Ritz values using matrix polynomials*, Linear Algebra Appl., 241–243 (1996), pp. 787–801.

[24] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix. Anal. Appl., 13 (1992), pp. 357–385.

[25] D. C. Sorensen, *Numerical methods for large eigenvalue problems*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 2002, pp. 519–584.

[26] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.

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

[28] L. N. Trefethen, *Computation of pseudospectra*, in Acta Numerica 8, Cambridge University Press, Cambridge, UK, 1999, pp. 247–295.

[29] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 5th ed., AMS, Providence, RI, 1969.

[30] T. G. Wright, *EigTool*, 2002. Software available from http://www.comlab.ox.ac.uk/pseudospectra/eigtool.