SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1112–1135

IMPLICITLY RESTARTED GMRES AND ARNOLDI METHODS FOR NONSYMMETRIC SYSTEMS OF EQUATIONS*

RONALD B. MORGAN†

Abstract. The generalized minimum residual method (GMRES) is well known for solving large nonsymmetric systems of linear equations. It generally uses restarting, which slows the convergence. However, some information can be retained at the time of the restart and used in the next cycle. We present algorithms that use implicit restarting in order to retain this information. Approximate eigenvectors determined from the previous subspace are included in the new subspace. This deflates the smallest eigenvalues and thus improves the convergence. The subspace that contains the approximate eigenvectors is itself a Krylov subspace, but not with the usual starting vector. The implicitly restarted FOM algorithm includes standard Ritz vectors in the subspace. The eigenvalue portion of its calculations is equivalent to Sorensen's IRA algorithm. The implicitly restarted GMRES algorithm uses harmonic Ritz vectors. This algorithm also gives a new approach to computing interior eigenvalues.

Key words. GMRES, implicit restarting, iterative methods, nonsymmetric systems, harmonic Ritz values

AMS subject classifications. 65F10, 15A06

PII. S0895479897321362

1. Introduction.

Large systems of linear equations arise in many different scientific applications. Notably, partial differential equations discretized with finite difference or finite element methods yield systems of equations. Large systems can be solved with either sparse factorization techniques or with iterative methods. These two approaches can be combined into a method that uses approximate factorization preconditioning for an iterative method. Here we consider only iterative methods.

For symmetric systems of linear equations, the most popular iterative method is the conjugate gradient method [18, 13, 14, 17]. The conjugate gradient method has several nice properties: it has a Krylov subspace for good convergence, it can be preconditioned, it has a short recurrence for efficiency, and it extracts an optimal solution from the subspace when the matrix is positive definite. Even for indefinite problems, it can be implemented so that the minimum residual solution is found [34].

In the nonsymmetric case, the choice for the iterative method is not so clear. Many methods have been proposed [2, 41, 51], but two of the best known are the generalized minimum residual method (GMRES) [43] and the quasi-minimal residual method (QMR) [9, 12]. These methods are generalizations of the conjugate gradient method. However, neither method retains all of the good properties of the conjugate gradient method. QMR has the Lanczos short recurrence [21, 22, 46, 54] but it has stability problems, is not guaranteed to extract the best solution from the subspace, and uses two matrix-vector products per iteration. Look-ahead [11, 12, 36] can be used to improve stability, the quasi-minimum property is designed to give smoother convergence, and transpose-free versions attempt to take advantage of the extra matrix-vector product. GMRES is stable and has the reassuring minimum

*Received by the editors May 12, 1997; accepted for publication by L. Reichel (in revised form) September 11, 1998; published electronically March 21, 2000. This work was supported by the National Science Foundation under grant NSF-CCR-9522612.

http://www.siam.org/journals/simax/21-4/32136.html

†Department of Mathematics, Baylor University, P.O. Box 97328, Waco, TX 76798-7398 (morganr@baylor.edu).

residual property. However, because the basis vectors for the Krylov subspace are explicitly orthogonalized, GMRES becomes increasingly expensive and requires more storage as the iteration proceeds. It can be restarted, but with the dimension of the subspace limited, the convergence slows down. GMRES often exhibits steady convergence, while QMR convergence curves are characterized by plateaus and sudden drops. In spite of somewhat mysterious behavior and significant complications for stability control, QMR is appealing because its use of the Lanczos recurrence allows for larger subspaces. Restarting is not needed. Large subspaces can be much better for difficult problems.

We propose a method that maintains the good qualities of GMRES but is more competitive with QMR for difficult problems. The new method is mathematically equivalent to a method in [28], but it has a more efficient implementation. The new implementation combines ideas from Sorensen's implicitly restarted Arnoldi algorithm [47] with harmonic Ritz values [27, 33, 45, 30].

We next review GMRES and several related methods that are needed in developing the new version of GMRES. Section 3 discusses how approximate eigenvectors can improve convergence. Section 4 gives a method that uses approximate eigenvectors and implicit restarting with the Arnoldi method for linear equations or full orthogonalization method (FOM) [38]. In section 5, approximate eigenvectors are added to GMRES using implicit restarting. Section 6 has numerical examples, and the related eigenvalue problem is mentioned in section 7.

2. Background on related methods. Here we discuss two methods for solving linear equations and also two eigenvalue methods. The methods are related in that they all use Krylov subspaces and the Arnoldi iteration [1, 37, 39]. With starting vector v_1, the Krylov subspace of dimension $m+1$ is $Span\{v_1, Av_1, A^2v_1, \ldots, A^m v_1\}$. The Arnoldi iteration generates an orthonormal basis for this subspace.

The Arnoldi iteration.

(1) *Start:* Choose a starting vector and normalize for v_1.

(2) *Iterate:* For $j = 1, 2, \ldots, m$ compute:

$h_{ij} = (v_i, Av_j), i = 1, 2, \ldots, j,$

$w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i,$

$h_{j+1,j} = ||w_j||$, and

$v_{j+1} = w_j / h_{j+1,j}$.

Let H_m be the $m \times m$ upper-Hessenberg matrix whose nonzero entries are defined in the Arnoldi iteration. Let $\bar{H}_m$ be the corresponding $m+1$ by m matrix with last row having only the one nonzero element $h_{m+1,m}$. Let V_m be the $n \times m$ matrix whose columns are v_1 through v_m. These relations hold [41, p. 148]:

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T \tag{2.1}$$

$$= V_{m+1} \bar{H}_m, \tag{2.2}$$

$$V_m^T A V_m = H_m. \tag{2.3}$$

All four of the methods that follow have a residual vector usually denoted by r. We add some different accents to distinguish them, except for the GMRES residual which is denoted by plain r.

2.1. Arnoldi for linear equations (FOM). The Arnoldi method for linear equations or FOM uses the Arnoldi iteration to solve nonsymmetric systems of linear equations [16, 38, 43, 41]. We call the Arnoldi residual vector $\acute{r}_0$. Suppose the problem

$Ax = b$ has been recast into form $A(x - x_0) = \acute{r}_0$. The starting vector for the Arnoldi iteration is $\acute{r}_0$. The small problem

$$H_m d = \beta e_1 \tag{2.4}$$

is solved, where $\beta = ||\acute{r}_0||$. Then the approximate solution of the original problem is $x_m = x_0 + V_m d$. It can be shown that the new residual vector satisfies

$$\acute{r} \equiv b - Ax_m \tag{2.5}$$

$$= -d_m h_{m+1,m} v_{m+1}, \tag{2.6}$$

where d_m is the last entry of d. So $\acute{r}$ is a multiple of v_{m+1}, and it becomes the new $\acute{r}_0$ at the restart. Therefore v_1 for the new Krylov subspace is the old v_{m+1}.

Restarted FOM.

(1) *Start:* Choose x_0 and compute $\acute{r}_0 = b - Ax_0$ and $v_1 = \acute{r}_0/||\acute{r}_0||$. Let $\beta = ||\acute{r}_0||$.
(2) *Iterate:* Apply m steps of the Arnoldi iteration.
(3) *Form the approximate solution:* $x_m = x_0 + V_m d$, where d is the solution of $H_m d = \beta e_1$.
(4) *Restart:* Compute the residual norm $||\acute{r}||$; if satisfied then stop, else let $x_0 = x_m$, $\beta = ||\acute{r}||$, $v_1 = \acute{r}/\beta$, $\acute{r}_0 = \acute{r}$, and go to 2.

2.2. GMRES. Most steps of the GMRES method [43] are the same as FOM. Let the recast problem be $A(x - x_0) = r_0$ and the starting vector for the Krylov subspace be r_0. Instead of (2.4), the small least squares problem

$$min||\beta e_1 - \bar{H}_m d||$$

is solved. GMRES finds the approximate solution that minimizes the residual norm.

2.3. Arnoldi for eigenvalue problems. To find eigenvalues of A using the Arnoldi method [1, 37, 39], we solve the small eigenvalue problem

$$H_m g_i = \theta_i g_i.$$

The θ_i's are approximate eigenvalues of A and are called Ritz values. The approximate eigenvectors or Ritz vectors are $y_i = V_m g_i$. The residual vector for each Ritz pair (θ_i, y_i) is defined to be $\dot{r}_i \equiv Ay_i - \theta_i y_i$. It is easy to derive [39, p. 175] that

$$\dot{r}_i = h_{m+1,m} e_m^T g_i v_{m+1}.$$

Defining

$$\beta_{mi} = h_{m+1,m} e_m^T g_i$$

gives

$$\dot{r}_i = Ay_i - \theta_i y_i = \beta_{mi} v_{m+1} \tag{2.7}$$

and

$$||\dot{r}_i|| = |\beta_{mi}|.$$

Note that every residual vector is a scalar multiple of the same vector, namely v_{m+1}. Furthermore, they are all multiples of the residual vector developed by FOM while solving linear equations, assuming the same starting vectors are used. They are not multiples of the GMRES residual vector.

Arnoldi for eigenvalues.

(1) *Start:* Choose an initial vector for the Krylov subspace and normalize for v_1.
(2) *Iterate:* Apply m steps of the Arnoldi iteration.
(3) *Find approximate eigenpairs:* Find eigenpairs (θ_i, g_i) of H_m as desired. The θ_i's are Ritz values. Compute Ritz vectors, $y_i = V_m g_i$, as desired.
(4) *Restart:* Residual norms can be checked for convergence. If needed, choose a new starting vector for v_1 and go to 2.

The restarting is difficult because one new starting vector must be chosen, while several Ritz vectors may need to be passed on to the new subspace. If not carefully chosen, a combination of Ritz vectors can be a very poor choice for the new starting vector [29]. Sorensen's implicitly restarted Arnoldi (IRA) algorithm [47, 23, 24] elegantly solves this problem. A QR iteration with H_m is used. The right combination of Ritz vectors is formed for the new starting vector [29].

We give a few of the details of the IRA method which will be needed later. Suppose k is the number of Ritz vectors to be saved for the next subspace. Let $p = m - k$. Then p shifts $\tau_1, \ldots, \tau_p$ are selected. Let $H^{(1)} \equiv H_m$. The main steps of the QR method at iteration i are first an orthogonal or unitary factorization of a shifted $H^{(i)}$,

$$H^{(i)} - \tau_i I = Q^{(i)} R^{(i)},$$

and then formation of $H^{(i+1)}$ using

$$R^{(i)} Q^{(i)} = H^{(i+1)} - \tau_i I.$$

In practice, a double shift is used to avoid unitary factorization in the case of a complex τ_i. Define $Q \equiv Q^{(1)} Q^{(2)} \cdots Q^{(p)}$ and $R \equiv R^{(p)} R^{(p-1)} \cdots R^{(1)}$. It is a standard result for the QR iteration [23, p. 24] that

$$QR = \prod_{l=1}^{p} (H_m - \tau_l I). \tag{2.8}$$

At the end of the QR phase, the new Krylov basis V^+ is

$$V^+ = V_m Q, \tag{2.9}$$

and the first k columns are used to start the next Arnoldi iteration. The $(k+1)$st column of the new V can also be calculated. For the case of the shifts being Ritz values, it is the same as the old v_{m+1}.

2.4. Interior or harmonic Arnoldi. The Arnoldi algorithm applies the Rayleigh–Ritz procedure [35, 39] to a Krylov subspace. It has trouble finding interior eigenvalues mainly because the Krylov subspace usually does not contain good approximations to the corresponding eigenvectors, but also due to properties of the Rayleigh–Ritz procedure. For symmetric matrices, Rayleigh–Ritz is optimal at extracting exterior eigenvalues [35, p. 215], but not for interior eigenvalues [27]. With nonsymmetric problems, there is still a tendency for Rayleigh–Ritz to be more reliable at extracting exterior eigenpairs ([30] has some discussion of this). An interior Arnoldi algorithm using a modified Rayleigh–Ritz procedure is given in [30] as a generalization of methods in [27]; see also [33, 45, 55]. Here we look only at the case where eigenvalues near zero are desired. For this case, this method has been mentioned in other recent papers in connection with GMRES [8, 10, 25, 26, 31, 44].

The small eigenvalue problem that needs to be solved is the generalized eigenvalue problem

$$V_m^T A^T V_m \tilde{g}_i = \frac{1}{\tilde{\theta}_i} V_m^T A^T A V_m \tilde{g}_i.$$

Using (2.2) and (2.3), this becomes

$$H_m^T \tilde{g}_i = \frac{1}{\tilde{\theta}_i} \bar{H}_m^T \bar{H}_m \tilde{g}_i. \tag{2.10}$$

This can be simplified with an orthogonal factorization $\bar{H}_m = QR$ to

$$H_m^T \tilde{g}_i = \frac{1}{\tilde{\theta}_i} R^T R \tilde{g}_i.$$

This formula has been stable in experiments, but a possibly better form from [33] is

$$(H_m + h_{m+1,m}^2 f e_m^*)\tilde{g}_i = \tilde{\theta}_i \tilde{g}_i,$$

where $f = H^{-T} e_m$. The $\tilde{\theta}_i$'s are called harmonic Ritz values in [33]. The harmonic Ritz vectors are

$$\tilde{y}_i = V_m \tilde{g}_i.$$

There are eigenvalue approximations available that are more accurate than the harmonic Ritz values. They are the Rayleigh quotients of the $\tilde{y}_i$'s, called ρ values in [27] and [30]. However, in this paper, the harmonic Ritz values are more important. Define the harmonic residual vectors to be

$$\tilde{r}_i \equiv A\tilde{y}_i - \tilde{\theta}_i \tilde{y}_i. \tag{2.11}$$

See [30] for formulas for finding the residual norms in the interior Arnoldi method, but note the residual using the Rayleigh quotient ρ_i is different from the harmonic residual defined in (2.11).

The eigenvalue methods described in this and the previous subsection will be used to improve the convergence of the linear equations methods in the first two subsections. The basic idea is that we want to add approximate eigenvectors to the subspaces for the linear equations methods. This idea is discussed in the next section.

3. Using eigenvectors to improve convergence of GMRES. Restarting generally slows the convergence of GMRES. Useful information is discarded at the restart. In this section we consider saving some vectors from the previous subspace and adding them to the new subspace to deflate eigenvalues. This discussion is a quick recap of [28]; see also [4, 42]. For other approaches to deflating eigenvalues for GMRES, see [3, 7, 20]. Kharchenko and Yeremin [20] modify the matrix A with approximate eigenvectors and thus deflate some eigenvalues. Erhel, Burrage, and Pohl [7] use approximate eigenvectors to build a preconditioner that shifts eigenvalues. Baglama, Calvetti, Golub, and Reichel [3] improve upon this last approach by using implicit restarting. The eigenvectors can be kept in a subspace until they are accurate enough.

We mention some other related work. The natural deflation of unrestarted conjugate gradient type methods has been studied; see, for example, [32, 52]. The GMRESR [53] and FGMRES [40] methods use a double loop to keep a part of the Krylov subspace. Adding an outside loop could perhaps benefit the methods in this paper, but we do not investigate this here. An improvement of GMRESR is given in [50], and another approach to what should be kept after a restart is in [49].

3.1. An augmented subspace. At the time of a restart, let $u_1, u_2, \ldots, u_k$ be vectors chosen from the subspace before it is discarded. They will be used in the next subspace that is developed. Specifically, the augmented subspace

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, u_1, u_2, \ldots, u_k\} \tag{3.1}$$

is generated in the next cycle of GMRES. The approximate solution for the linear equations problem is extracted from this subspace. The subspace has a Krylov portion and an added portion. Many choices are possible for the added vectors. We chose these vectors to be approximate eigenvectors corresponding to the approximate eigenvalues nearest zero. However for some problems, a few large eigenvalues should also be targeted.

3.2. Reasons for using eigenvectors. For symmetric positive definite matrices, slow convergence of the conjugate gradient method is caused by the presence of some relatively small eigenvalues. For nonsymmetric matrices, there can be other factors such as negative and complex eigenvalues. And even with a good eigenvalue distribution, convergence can be poor [15]. Nevertheless, in many difficult linear equations problems, there are small eigenvalues that detract from the convergence. If eigenvectors are added to the subspace, then the corresponding eigenvalues are essentially deflated from the spectrum of the matrix. And if approximate eigenvectors are added, they do not need to be extremely accurate before they are helpful. This is shown in [28, Theorem 2] and has also been observed experimentally.

As an example of how beneficial this deflation of eigenvalues can be, consider a symmetric matrix with eigenvalue distribution $1, 2, 3, \ldots, n-1, n$. If the three smallest eigenvalues are removed from the spectrum, then $\sqrt{\kappa}$, the square root of the condition number, changes from $\sqrt{n}$ to $\frac{\sqrt{n}}{2}$. Using this to estimate convergence gives that convergence should be twice as fast with the eigenvalues removed. If GMRES is unrestarted, removing some eigenvalues may not affect convergence this much, since eigenvalues naturally deflate anyway. But when restarting is used, the convergence will improve about as predicted.

3.3. A few details. The approximate eigenvectors can be chosen to be Ritz vectors from the Arnoldi method or harmonic Ritz vectors from the interior Arnoldi method. In [28], harmonic Ritz vectors are used. For indefinite matrices, they give better approximations for the desired eigenvectors that correspond to eigenvalues near zero. So the subspace is

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}, \tag{3.2}$$

where $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k$ are the harmonic Ritz vectors corresponding to the harmonic Ritz values nearest to the origin.

The implementation for this augmented subspace is more complicated than for standard GMRES [28]. There is added storage, but often a reduction of expense per iteration, because a matrix-vector product can be avoided when an approximate eigenvector is used instead of another Arnoldi vector; see Table 1 in section 5. For the case of inexpensive matrix-vector product, DAXPYs and DDOTs can be reduced to $m^2 + 2km$, instead of saving on the matrix-vector products. This augmented GMRES method from [28] will subsequently be referred to as GMRES-E (GMRES with eigenvectors).

4. Implicitly restarted FOM. We now reconsider the subspace (3.1) discussed in the previous section. With the GMRES-E method, the approximate eigenvectors do not naturally fit into the subspace, but are forced on at the end. It is not even necessary that the extra vectors be approximate eigenvectors. In this and the next section, we look at how the approximate eigenvectors can be selected to correspond to the linear equations method. Then they fit more naturally into the subspace. In fact, the whole subspace is a Krylov subspace, though not with the same starting vector $\acute{r}_0$ or r_0. With FOM, we need the standard Arnoldi Ritz vectors.

4.1. Adding approximate eigenvectors to the subspace. The desired subspace is

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-k-1}\acute{r}_0, y_1, y_2, \ldots, y_k\}. \tag{4.1}$$

Before considering the general case, we look at how just one Ritz vector can be added to the subspace. This turns out to be fairly simple. We want a subspace consisting of the usual Krylov subspace for FOM, but of one less degree and with one Ritz vector as an additional basis vector. So the desired subspace is

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-2}\acute{r}_0, y_1\}. \tag{4.2}$$

We will show that in spite of appearing otherwise, this subspace is itself a Krylov subspace. But it is a Krylov subspace generated with starting vector y_1 instead of $\acute{r}_0$. The key is that the residual vectors for the eigenvalue problem and the linear equations problem are both multiples of the same vector, v_{m+1}.

PROPOSITION 4.1. *Subspace* (4.2) *is the same subspace as*

$$Span\{y_1, Ay_1, A^2y_1, A^3y_1, \ldots, A^{m-1}y_1\}. \tag{4.3}$$

Proof. From (2.7),

$$Ay_1 = \theta_1 y_1 + \beta_{m1} v_{m+1}.$$

Thus Ay_1 is a combination of y_1 and v_{m+1} and

$$Span\{y_1, Ay_1\} = Span\{y_1, v_{m+1}\}.$$

Next, using (2.7) twice,

$$\begin{aligned} A^2y_1 &= \theta_1 Ay_1 + \beta_{m1} Av_{m+1} \\ &= \theta_1^2 y_1 + \theta_1\beta_{m1} v_{m+1} + \beta_{m1} Av_{m+1} \end{aligned}$$

and

$$Span\{y_1, Ay_1, A^2y_1\} = Span\{y_1, v_{m+1}, Av_{m+1}\}.$$

Continuing this, subspace (4.3) is the same as

$$Span\{y_1, v_{m+1}, Av_{m+1}, A^2v_{m+1}, A^3v_{m+1}, \ldots, A^{m-2}v_{m+1}\}. \tag{4.4}$$

Subspace (4.4) is the same as (4.2) because the residual vector $\acute{r}_0$ is a multiple of v_{m+1}. This establishes the claim. ∎

For adding two Ritz vectors to the subspace, the right combination of the two vectors is needed for the new starting vector. Let $s = \beta_{m2}y_1 - \beta_{m1}y_2$. Then it can be shown that

$$Span\{s, As, A^2s, \ldots, A^{m-1}s\} = Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, \ldots, A^{m-3}\acute{r}_0, y_1, y_2\}.$$

See Proposition 5.7 for a similar proof.

For adding k Ritz vectors, there is again a way to combine them into one vector s, so that the Krylov subspace generated with s is the desired subspace (4.1) (see [29, section 3] and also section 7.1). However, Sorensen's implicit restarting [47] in the IRA method provides a better way to implement the formation of subspace (4.1).

4.2. Generating the subspace with implicit restarting. The following theorem establishes the equivalence between augmenting the FOM subspace with Ritz vectors and using implicit restarting. "Exact shifts" [47] should be used for the implicit restarting (the shifts are the unwanted Ritz values).

THEOREM 4.2. *Suppose the IRA method is implemented with the unwanted Ritz values as shifts, and with the initial starting vector the same as the initial $\acute{r}_0$ for FOM. Then IRA generates the subspaces*

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-k-1}\acute{r}_0, y_1, y_2, \ldots, y_k\}, \tag{4.5}$$

where $y_1, \ldots, y_k$ are Ritz vectors and $\acute{r}_0$ is the residual vector for an FOM method using these augmented subspaces.

Proof. The proof has been mostly given in other papers; we will give a sketch. Suppose we have an Arnoldi recurrence $AV_m = V_{m+1}\bar{H}_m$ and apply implicit restarting with the p unwanted Ritz values $\theta_{k+1}, \ldots, \theta_m$ as shifts. Then the first k Arnoldi basis vectors in the next subspace span the subspace $Span\{y_1, \ldots, y_k\}$. This is [29, Theorem 2]; the main step in proving this is that for $j \le k$, $Qe_j = \psi(H_m)t$, where ψ is a polynomial with roots at the undesired Ritz values, and t is some vector (this follows from (2.8), if we let t be the solution of $Rt = e_j$). These unwanted Ritz values are eigenvalues of H_m, so $\psi(H_m)$ purges the unwanted eigencomponents from t and we have that $\psi(H_m)t = Qe_j$ is a combination of only the desired eigenvectors of H_m. Then $V^+e_j = V_mQe_j$ is a combination of the desired Ritz vectors; see [48] for a complete proof.

As observed in the proof of [29, Theorem 3], the v_{k+1} Arnoldi basis vector in the new subspace is the same as the v_{m+1} vector for the previous subspace, so IRA has subspace

$$Span\{y_1, y_2, \ldots y_k, v_{m+1}, Av_{m+1}, A^2v_{m+1}, A^3v_{m+1}, \ldots, A^{m-k-1}v_{m+1}\}.$$

Since the FOM residual vector $\acute{r}_0$ is a multiple of v_{m+1}, this is equivalent to subspace (4.5). ∎

Next, the algorithm is listed for implicitly restarted FOM (FOM-IR). Added to the standard FOM algorithm is the solution of a small eigenvalue problem and the implicit restarting. Also, the reduced problem in step 3 is slightly different. We derive the formula for step 3 using (2.6). Note the superscript *old* denotes a quantity from the previous cycle of FOM. The reduced problem is

$$\begin{aligned} V_m^TAV_md &= V_m^T\acute{r}_0, \\ H_md &= V_m^T(-d_m^{old}h_{m+1,m}^{old}v_{m+1}^{old}), \\ H_md &= -d_m^{old}h_{m+1,m}^{old}e_{k+1}. \end{aligned}$$

The first k entries in the right-hand side are zero because $\acute{r}_0$, being a multiple of v_{m+1}^{old}, is orthogonal to all of the Ritz vectors (which are also from the previous cycle), and as just mentioned in the proof, $v_1, \ldots, v_k$ are combinations of the Ritz vectors. The last $m-k-1$ entries in the right-hand side are zero because $v_{k+1} = v_{m+1}^{old}$ and v_j is orthogonal to v_{k+1} if $j > k+1$.

FOM-IR.

(1) *Start:* Choose m, the maximum size of the subspace, and k, the desired number of approximate eigenvectors. Choose x_0 and compute $\acute{r}_0 = b - Ax_0$. The recast problem is $A(x - x_0) = \acute{r}_0$. Let $v_1 = \acute{r}_0/||\acute{r}_0||$ and $\beta = ||\acute{r}_0||$.
(2) *Iterate:* Apply the Arnoldi iteration out to step m.
(3) *Form the approximate solution:* $x_m = x_0 + V_m d$, where d comes from a reduced problem. In the first cycle, before any restarts, d is the solution of $H_m d = \beta e_1$. In later cycles, d is the solution of $H_m d = -d_m^{old} h_{m+1,m}^{old} e_{k+1}$. Check $||\acute{r}||$ for convergence, and proceed if not satisfied.
(4) *Eigenvalue problem:* Compute the eigenvalues of H_m. These are the Arnoldi–Ritz values, $\theta_1, \ldots, \theta_m$.
(5) *Restart:* Let $x_0 = x_m$. Apply implicit restarting with the unwanted Ritz values as shifts. Both parts of a complex conjugate pair should be included. Go to 2, and resume the Arnoldi iteration from step $k+1$.

FOM-IR solves linear equations and at the same time implements Sorensen's implicitly restarted Arnoldi method for eigenvalues. In spite of the similarity of subspace (4.1) to (3.2), the FOM-IR method is not equivalent to the GMRES-E method discussed in section 3. The residual vectors and approximate eigenvectors are both different.

It is interesting that even though FOM does not use a minimum residual solution, the minimum residual norm can be monitored. It is this minimum residual norm that is given in the examples in section 6. However, when restarting, the approximate solution and residual vector must be found with the FOM approach instead of minimum residual. This is the reason for the jumps in the residual norm at the time of restarts in Figure 1.

5. Implicitly restarted GMRES. Now we include approximate eigenvectors in the subspace for GMRES. For the approximate eigenvectors to fit into a Krylov subspace, we must use the harmonic Ritz vectors mentioned in section 2.4. The subspace is

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}, \tag{5.1}$$

which is the same subspace used in the GMRES-E algorithm. The first subsection looks at why this subspace is a Krylov subspace, and the second proves that implicit restarting can be used.

5.1. The whole subspace is a Krylov subspace. We assume that A is a nonsingular matrix. We will deal mainly with the case of distinct harmonic Ritz values, but we first discuss briefly the special situation of multiple values.

An upper-Hessenberg matrix is called unreduced if there are no zero elements on the subdiagonal. The matrix H_m produced by the Arnoldi iteration is generally unreduced. Otherwise, the Krylov subspace at the time the zero occurs is an invariant subspace of A. We first give a standard result for Arnoldi–Ritz pairs. For a proof, see, for instance, [23].

LEMMA 5.1. *If the Arnoldi iteration produces an unreduced Hessenberg matrix H_m, then multiple eigenvalues of H_m have only one eigenvector associated with them. So multiple Arnoldi–Ritz values have only one corresponding Ritz vector.*

We can give a similar result for harmonic Ritz values.

THEOREM 5.2. *Assume that the Arnoldi–Hessenberg matrix is unreduced. Then multiple harmonic Ritz values have only one harmonic Ritz vector associated with them.*

Proof. We consider a problem equivalent to the harmonic Ritz value problem. The equivalent problem is applying the Rayleigh–Ritz procedure with matrix A^{-1} to the subspace spanned by the columns of the matrix AV_m [27, 30, 33]. The reciprocals of the Ritz values from this problem are the harmonic Ritz values. The Ritz vectors of this equivalent problem are $AV_m\tilde{g}_i$, where the $\tilde{g}_i$ are the same as defined earlier in (2.10) for the harmonic Ritz vectors $V_m\tilde{g}_i$.

This equivalent Rayleigh–Ritz problem can be viewed as being based on an Arnoldi iteration, although different from the one we have been considering. To see this, rewrite the subspace for this equivalent problem, $Span\{Av_1, A^2v_1, \ldots, A^mv_1\}$, as $Span\{A^mv_1, A^{-1}A^mv_1, A^{-2}A^mv_1, \ldots A^{-m+1}A^mv_1\}$. This is a Krylov subspace with starting vector A^mv_1 and with matrix A^{-1}. Thus, there is some Arnoldi iteration that would generate this subspace. The corresponding Hessenberg matrix would also be unreduced (if the first j columns of AV_m spanned an invariant subspace, then the first j columns of V_m would also, and this violates an assumption). Thus, from Lemma 5.1, this different Hessenberg matrix has only one eigenvector associated with each eigenvalue. Going back to the original Arnoldi iteration, there is only one harmonic Ritz vector with each harmonic Ritz value. ∎

In the special case of multiple harmonic Ritz values, there is not a full set of harmonic Ritz vectors. For the rest of the paper, we assume the harmonic Ritz values are distinct and that H_m is unreduced. After two lemmas, we establish a connection between the harmonic residual vectors and the residual vector for GMRES.

LEMMA 5.3. *The GMRES residual vector r is orthogonal to the subspace spanned by the columns of AV_m.*

See [41, chapter 5] for a proof. An orthogonality condition can also be given for the harmonic residual vectors [45, 30].

LEMMA 5.4. *Let $\tilde{y}_j$ be a harmonic Ritz vector with associated harmonic residual vector $\tilde{r}_j = A\tilde{y}_j - \tilde{\theta}_j\tilde{y}_j$. Then $\tilde{r}_j$ is orthogonal to the subspace spanned by the columns of AV_m.*

THEOREM 5.5. *Assuming the same underlying Arnoldi iteration is used, the harmonic residual vectors are multiples of the GMRES residual vector.*

Proof. This follows quickly from the two lemmas. The harmonic residual vectors and the GMRES residual vector all reside in the Krylov subspace of dimension $m+1$ and are orthogonal to the same subspace of dimension m, so they must be multiples of each other. ∎

This relationship between the harmonic and GMRES residual vectors can be used to include harmonic Ritz vectors in a Krylov subspace for GMRES. This is similar to how standard Ritz vectors were included in FOM in section 4.1. At the time of a restart for GMRES, the new r_0 is set equal to the r for the current cycle. Let

$$u = \frac{r_0}{||r_0||}.$$

Then for each i, $\tilde{r}_i = \gamma_i u$, for some scalar γ_i. So

$$A\tilde{y}_i - \tilde{\theta}_i \tilde{y}_i = \gamma_i u. \tag{5.2}$$

Using (5.2), we can see that generating a Krylov subspace with starting vector $\tilde{y}_1$ builds a subspace that contains a Krylov subspace with r_0 as starting vector. The proof of the following proposition is like that for Proposition 4.1, except that it uses (5.2) instead of (2.7).

PROPOSITION 5.6. *For $\tilde{y}_1$ a harmonic Ritz vector and r_0 the GMRES residual vector,*

$$Span\{\tilde{y}_1, A\tilde{y}_1, A^2\tilde{y}_1, \ldots, A^{m-1}\tilde{y}_1\} = Span\{\tilde{y}_1, r_0, Ar_0, A^2r_0, \ldots, A^{m-2}r_0\}.$$

Next, for including two harmonic Ritz vectors in the subspace, we use the starting vector

$$s = \gamma_2 \tilde{y}_1 - \gamma_1 \tilde{y}_2. \tag{5.3}$$

The subspace that is generated contains a Krylov subspace of dimension $m-2$ with r_0 as starting vector, along with the two harmonic Ritz vectors.

PROPOSITION 5.7. *With starting vector s as in* (5.3)*,*

$$Span\{\tilde{s}, A\tilde{s}, A^2\tilde{s}, \ldots, A^{m-1}\tilde{s}\} = Span\{\tilde{y}_1, \tilde{y}_2, r_0, Ar_0, A^2r_0, \ldots, A^{m-3}r_0\}.$$

Proof. After multiplying (5.3) through by A and using (5.2),

$$\begin{aligned} As &= \gamma_2(\tilde{\theta}_1\tilde{y}_1 + \gamma_1 u) - \gamma_1(\tilde{\theta}_2\tilde{y}_2 + \gamma_2 u) \\ &= \gamma_2\tilde{\theta}_1\tilde{y}_1 - \gamma_1\tilde{\theta}_2\tilde{y}_2. \end{aligned}$$

The special choice of s caused the u term to drop out. We have that $Span\{s, As\} = Span\{\tilde{y}_1, \tilde{y}_2\}$.

Multiplying by A and again using (5.2) yields $A^2 s$ as a combination of $\tilde{y}_1$, $\tilde{y}_2$, and u. Continuing this, we get that A^{m-1} is a combination of $\tilde{y}_1$, $\tilde{y}_2$, u, Au, $\ldots$, $A^{m-3}u$, and

$$Span\{s, As, A^2s, \ldots, A^{m-1}s\} = Span\{\tilde{y}_1, \tilde{y}_2, u, Au, A^2u, \ldots, A^{m-3}u\}.$$

We are done, since u is a multiple of r_0. ∎

For adding more harmonic Ritz vectors to the subspace, a careful combination of the harmonic Ritz vectors can be used as starting vector (see section 7.2). However, we will concentrate on using implicit restarting.

5.2. Generating the subspace with implicit restarting. We will develop an implicitly restarted GMRES method, called GMRES-IR, that includes harmonic Ritz vectors in the subspace. The approach is like that for FOM-IR, except the unwanted *harmonic* Ritz values are used as shifts during the QR phase of the algorithm. We need to prove that this approach generates the desired subspace (5.1). This is done in Theorem 5.14, but the proof for GMRES-IR is more difficult than was the proof of Theorem 4.2, the corresponding theorem for FOM. As in Theorem 4.2, we use that $Qe_j = \psi(H_m)t$, but here ψ has roots at harmonic Ritz values instead of standard Ritz values, and we need t to have nonzeros entries only in its first k positions. Then it

can be shown that $\psi(H_m)t$ has components only in the directions of the desired $\tilde{g}_i$'s, but it takes several lemmas. We now start on these.

The GMRES residual vector can be written in terms of a polynomial with roots at the harmonic Ritz values: $r_m = \pi(A)v_1$, where π is a polynomial such that $\pi(\lambda) = \alpha \prod_{l=1}^{m}(\lambda - \tilde{\theta}_l)$, and α is a scalar; see [26] and [8]. We need a similar expression for harmonic Ritz vectors.

LEMMA 5.8. *The harmonic Arnoldi–Ritz vectors can be written as* $\tilde{y}_i = \phi_i(A)v_1$, *where* ϕ_i *is the polynomial*

$$\phi_i(\lambda) = \alpha_i \prod_{\substack{l=1 \\ l \neq i}}^{m} (\lambda - \tilde{\theta}_l),$$

and α_i *is a scalar.*

For a proof of this lemma; see [30]. The next lemma relates polynomials of A with polynomials of H_m; see [5, 33, 23] for similar results and proofs.

LEMMA 5.9. *For* p *any polynomial of degree less than* m,

$$p(A)v_1 = V_m p(H_m)e_1.$$

For $\tilde{g}_i$ a solution of (2.10), there is an expression in terms of the same polynomials ϕ_i used for $\tilde{y}_i$ in Lemma 5.8.

LEMMA 5.10. *For* $\tilde{g}_i$ *from* (2.10) *and* ϕ_i *as defined in Lemma* 5.8,

$$\tilde{g}_i = \phi_i(H_m)e_1.$$

Proof. We will write $\tilde{g}_i$ in terms of $\tilde{y}_i$, then use Lemmas 5.8 and 5.9.

$$\begin{aligned} \tilde{g}_i &= V_m^T V_m \tilde{g}_i \\ &= V_m^T \tilde{y}_i \\ &= V_m^T \phi_i(A)v_1 \\ &= V_m^T V_m \phi_i(H_m)e_1 \\ &= \phi_i(H_m)e_1. \qquad \square \end{aligned}$$

Lemma 5.9 is also used in proving the next lemma.

LEMMA 5.11. *For* $j \leq m$, *there are some scalars* γ *and* $\beta_1, \beta_2, \ldots, \beta_{j-1}$ *such that*

$$e_j = \gamma \prod_{l=1}^{j-1} (H_m - \beta_l I)e_1.$$

Proof. Because v_j is a member of the Krylov subspace of degree j with v_1 as a starting vector, there are scalars γ and $\beta_1, \beta_2, \ldots, \beta_{j-1}$ such that

$$v_j = \gamma \prod_{l=1}^{j-1} (A - \beta_l I)v_1.$$

Thus,

$$\begin{aligned} e_j &= V_m^T v_j \\ &= \gamma V_m^T \prod_{l=1}^{j-1} (A - \beta_l I) v_1 \\ &= \gamma V_m^T V_m \prod_{l=1}^{j-1} (H_m - \beta_l I) e_1 \\ &= \gamma \prod_{l=1}^{j-1} (H_m - \beta_l I) e_1, \end{aligned}$$

where we used Lemma 5.9 to convert from a polynomial of A to a polynomial of H_m. ∎

In what follows, we define

$$\psi(\lambda) \equiv \prod_{l=k+1}^{m} (\lambda - \tilde{\theta}_l). \tag{5.4}$$

LEMMA 5.12. *For $t \in Span\{e_1, e_2, \ldots, e_k\}$,*

$$\psi(H_m)t \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}.$$

Proof. We will show that for $j \leq k$, $\psi(H_m)e_j \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}$, from which the desired result quickly follows.

Using Lemma 5.11,

$$\begin{aligned} \psi(H_m)e_j &= \gamma \psi(H_m) \prod_{l=1}^{j-1} (H_m - \beta_l I) e_1 \\ &= \gamma \Big(\prod_{l=1}^{j-1} (H_m - \beta_l I) \Big) \psi(H_m) e_1. \end{aligned}$$

We need to show that this vector is a combination of the vectors $\tilde{g}_1, \ldots, \tilde{g}_k$, which from Lemma 5.10 can be written as

$$\begin{aligned} \tilde{g}_i &= \phi_i(H_m) e_1 \\ &= \alpha_i \prod_{\substack{l=1 \\ l \neq i}}^{m} (H_m - \tilde{\theta}_l) e_1 \\ &= \alpha_i \Big(\prod_{\substack{l=1 \\ l \neq i}}^{k} (H_m - \tilde{\theta}_l) \Big) \psi(H_m) e_1. \end{aligned}$$

So expressing $\psi(H_m)e_j$ as a combination of the vectors $\tilde{g}_1, \ldots, \tilde{g}_k$ is equivalent to expressing the polynomial

$$\prod_{l=1}^{j-1} (\lambda - \beta_l),$$

with $j \leq k$, as a combination of the polynomials

$$\alpha_i \prod_{\substack{l=1 \\ l \neq i}}^{k} (\lambda - \tilde{\theta}_l),$$

for i from 1 to k. These last polynomials are Lagrange interpolating polynomials except for possibly different normalization, and they are linearly independent since the $\tilde{\theta}_l$'s are distinct. They can be combined to form any polynomial of degree $k-1$ or less, so we are finished. ∎

We now give the last lemma which deals with the QR iteration [14, 23].

LEMMA 5.13. *Let the matrix Q and shifts τ_i be from the QR iteration. Then for $j \leq k$,*

$$Qe_j = \prod_{l=1}^{p} (H_m - \tau_l I)t,$$

where $t \in Span\{e_1, \ldots, e_k\}$.

Proof. Recall from section 2.3 that $H^{(1)} = H_m$ and $H^{(1)} - \tau_i I = Q^{(1)} R^{(1)}$. We assumed earlier that H_m is unreduced, so $H^{(1)} - \tau_1 I$ is an unreduced Hessenberg matrix. Therefore the first $m-1$ columns of $R^{(1)}$ are linearly independent. And $H^{(i)} - \tau_i I$ has its $m-i+1$ by $m-i+1$ leading principle submatrix unreduced (a zero in the subdiagonal can only move one position during each QR iteration). Therefore $R^{(i)}$ has its first $m-i$ columns linearly independent. Since each $R^{(i)}$ is also upper triangular and $R = R^{(p)} R^{(p-1)} \cdots R^{(1)}$, R has its first $m-p = k$ columns linear independent. With this and the fact that R is upper triangular, we have that for $j \leq k$, there is a solution to

$$Rt = e_j, \tag{5.5}$$

such that

$$t \in Span\{e_1, \ldots, e_k\}. \tag{5.6}$$

Let t be the solution of (5.5) such that (5.6) holds. Then $Qe_j = QRt = \prod_{l=1}^{p} (H_m - \tau_l I)t$, from (2.8). ∎

The GMRES-IR method uses implicit restarting with unwanted harmonic Ritz values as shifts. The theorem establishes that in every cycle, between restarts, it generates subspace (5.1).

THEOREM 5.14. *The GMRES-IR method generates the subspace*

$$Span\{r_0, Ar_0, A^2 r_0, A^3 r_0, \ldots, A^{m-k-1} r_0, \tilde{y}_1, \tilde{y}_2, \ldots \tilde{y}_k\}.$$

Proof. The shifts for the QR iteration are the unwanted harmonic Ritz values, so the $\tau_1, \ldots, \tau_p$ in Lemma 5.13 are $\tilde{\theta}_{k+1}, \tilde{\theta}_{k+2}, \ldots, \tilde{\theta}_m$. Therefore $\prod_{l=1}^{p} (H_m - \tau_l I) = \psi(H_m)$, for ψ as defined in (5.4). Lemma 5.13 becomes

$$Qe_j = \psi(H_m)t,$$

where $t \in Span\{e_1, \ldots, e_k\}$. Applying Lemma 5.12 gives

$$Qe_j \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}. \tag{5.7}$$

The new Arnoldi basis matrix is V^+. Using (2.9),

$$V^+ e_j = V_m Q e_j.$$

With (5.7), if $j \leq k$, then

$$V^+ e_j \in Span\{V_m \tilde{g}_1, V_m \tilde{g}_2, \ldots, V_m \tilde{g}_k\}.$$

And since the harmonic Ritz vectors are defined as $\tilde{y}_i = V_m \tilde{g}_i$,

$$V^+ e_j \in Span\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}.$$

We have that after the implicit restarting, the first k columns of V^+ span the same subspace as $Span\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}$.

So $\tilde{y}_1, \ldots, \tilde{y}_k$ are in the new subspace. We need to show that this subspace also contains the vectors $r_0, Ar_0, \ldots, A^{m-k-1} r_0$. Using (5.2), $A\tilde{y}_i$ is a combination of $\tilde{y}_i$ and r_0. Since v_k is a combination of $\tilde{y}_1, \ldots, \tilde{y}_k$, Av_k is a combination of r_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. Thus v_{k+1} is also a combination of r_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. Next, using a similar argument, v_{k+2} is a combination of r_0, Ar_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. This can be continued until we arrive at the desired result. □

So we have established that implicit restarting can be used with GMRES. The changes in the algorithm from FOM-IR are listed.

GMRES-IR. GMRES-IR is the same as FOM-IR, except replace $\acute{r}$ with r and substitute the following:

(3) d is the solution of $\min||V_{m+1}^T r_0 - \bar{H}_m d||$.

(4) and (5) Compute and use harmonic Ritz values.

In step 3 of the algorithm, the vector $V_{m+1}^T r_0$ has zeros except in its first $k+1$ entries. It can also be shown that after Givens rotations are applied during solution of the least squares, the first k entries of this vector become zero.

Having harmonic Ritz vectors in the subspace for GMRES-IR causes the portion of the subspace that is Krylov with starting vector r_0 to have dimension $m-k$ instead of m as for standard GMRES. This slows convergence, but the k approximate eigenvectors generally improve convergence. So there is a tradeoff in choosing k.

We next compare GMRES-IR with some other methods. While GMRES-IR is mathematically equivalent to GMRES-E for a completed cycle, it can give better results at the start and middle of a cycle. This is because eigenvector information is included in the subspace from the beginning instead of being appended at the end of the cycle. For problems using only a moderate number of iterations, this may be of some significance. GMRES-E can be effective for nonlinear problems, while GMRES-IR is not designed for situations with changing equations. The main advantage of GMRES-IR compared to GMRES-E is in expense and storage. We compare these next. Table 1 has a rough count for one cycle of standard GMRES, GMRES-IR, and GMRES-E. Only the major expenses and storage vectors of length n are considered. The solution of the small eigenvalue problem in GMRES-IR and GMRES-E is not included because it is not of order n, however there are cases where it is a significant expense. The cost and storage depends on the implementation. For GMRES-IR, the formation of the new Krylov basis in (2.9) is assumed to be done in place using a

TABLE 1
Expenses and storage for GMRES methods.

	GMRES(m)	GMRES-IR(m,k)	GMRES-E(m,k)
DAXPYs, DDOTs, etc.	m^2	$m^2 + km - \frac{k^2}{2}$	$m^2 + 3km$
matrix-vector products	m	$m - k$	$m - k$
length n vectors	$m + 2$	$m + 2$	$m + k + 2$

Householder factorization of a portion of Q. We see that GMRES-IR can save matrix-vector products without additional storage. Standard GMRES uses fewer length n vector operations (DAXPYs, DDOTs, etc.) per cycle than GMRES-IR, but uses more matrix-vector products. The main advantage of GMRES-IR over GMRES is in reducing the number of iterations.

The next section continues the comparison of methods with some computational experiments.

6. Examples. We compare the new implicitly restarted GMRES and FOM methods with standard GMRES and also with transpose-free QMR (TFQMR) [9]. Some examples also include full, unrestarted GMRES. Full GMRES is not a practical method, but it is interesting to see how closely the other methods can match its performance. FOM-IR(25,6) has $m = 25$ and $k = 6$ (subspaces are dimension 25, including six approximate eigenvectors). Also used are GMRES-IR(25,6) and GMRES-IR(25,10). GMRES(25) refers to standard GMRES with subspaces of size 25. Residual norms are plotted against the number of matrix-vector products.

Example 6.1. The matrix is bidiagonal with entries 0.01, 0.1, 1, 2, 3, 4, ..., 997, 998 on the main diagonal and 1's on the super diagonal. The right-hand side has all 1's. The matrix has small eigenvalues that slow the convergence. No method can give an accurate solution until it develops approximations to the smallest few eigenvalues. GMRES(25) can never develop these approximations because of the restarts. The convergence stagnates; see Figure 1. TFQMR is much more successful because it generates a large subspace. However, the implictly restarted methods are even better. Comparing the number of matrix-vector products required to reduce the residual norm by a factor of 10^{-6} (from 31.6 to below $3.16 * 10^{-5}$), GMRES(25,10) uses 231 and TFQMR needs 378. This shows potential for GMRES-IR, particularly for problems with expensive matrix-vector products.

GMRES-IR(25,10) competes well with full GMRES, even though it uses subspaces of dimension 25. And the portion of these subspaces that is Krylov with starting vector r_0 is of dimension only 15. Meanwhile, full GMRES generates a subspace of size over 200. Yet the two methods have similar convergence. This is surprising, because this problem is fairly difficult. For easy problems, small subspaces can be almost as effective as large subspaces; for tougher problems, the dimension of the Krylov subspace is very important. However, the deflation of eigenvalues in GMRES-IR turns the difficult problem into an easy one. Once approximate eigenvectors have developed for the 10 smallest eigenvalues, the deflated spectrum is $9, 10, 11, \ldots, 998$. We can compare the square roots of the ratios of largest to smallest eigenvalues (similar to comparing the square roots of condition numbers in section 3.2, although the importance is less obvious). The square root of the ratio of largest to smallest

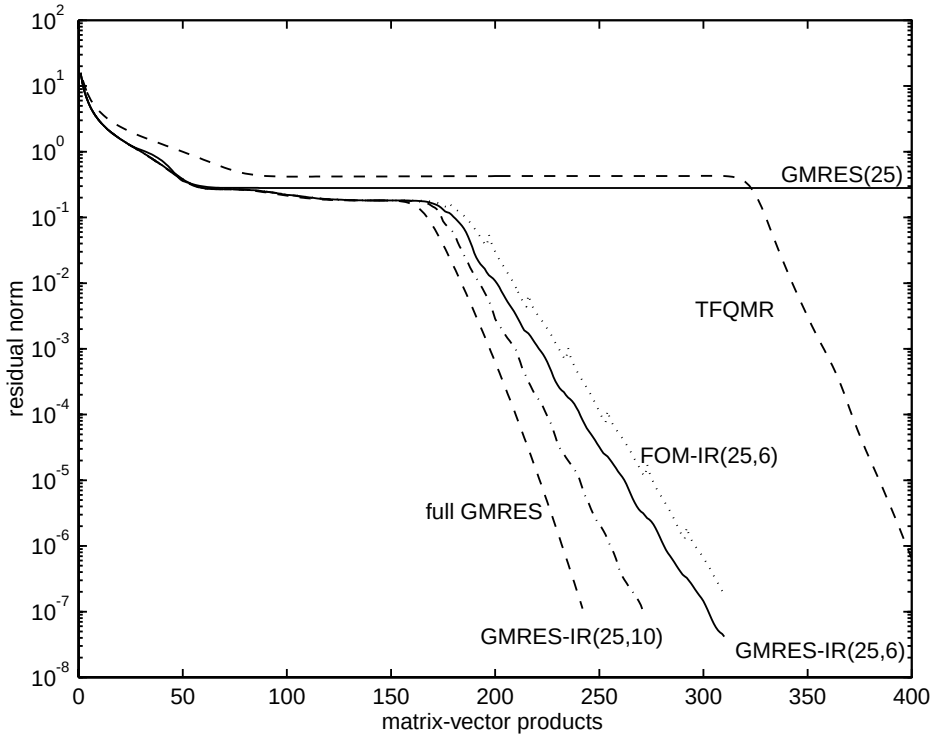

FIG. 1. *The comparison for Example* 6.1.

eigenvalue is 10.5 for the deflated spectrum, compared to the original ratio of 316. This makes the problem considerably easier. Krylov subspaces of dimension 15 are almost as effective for this easier problem as is a single Krylov subspace of dimension 200.

Example 6.2. The matrix is SHERMAN4 from the Harwell–Boeing sparse matrix collection [6]. The right-hand side is the one provided in the collection. This matrix has dimension 1104. All of the eigenvalues are positive. The smallest 11 eigenvalues are 0.031, 0.085, 0.28, 0.40, 0.43, 0.59, 0.63, 0.89, 0.92, 1.00, and 1.13, and the largest is 66.5. See Figure 2 for convergence plots. This problem is not as difficult for standard GMRES as Example 6.1. Convergence of GMRES(25) is slow but steady. TFQMR is erratic, but considerably better than GMRES(25). However, the implicitly restarted GMRES and FOM methods converge both faster and more consistently than TFQMR. Results are very different when comparing length n vector operations. Because the matrix is sparse (3.43 nonzeros per row), GMRES-IR uses about four times as many vector operations per matrix-vector product than does TFQMR, approximately 43 versus 10.4 (7 for DAXPYs and DDOTs, though this depends on the implementation, and 3.4 for the matrix). To improve the residual norm by 10^{-6}, GMRES-IR(25,6) requires about 2.6 times as many vector operations as TFQMR (5.5 versus 2.1 thousands).

Table 2 gives results for various choices of m and k. First, m is fixed at 25 and k varies from $k = 0$ (standard GMRES) up to $k = 10$. The number of matrix-vector products does not change much for $k \geq 4$. Next, m varies with $k = 4$. With $m = 15$,

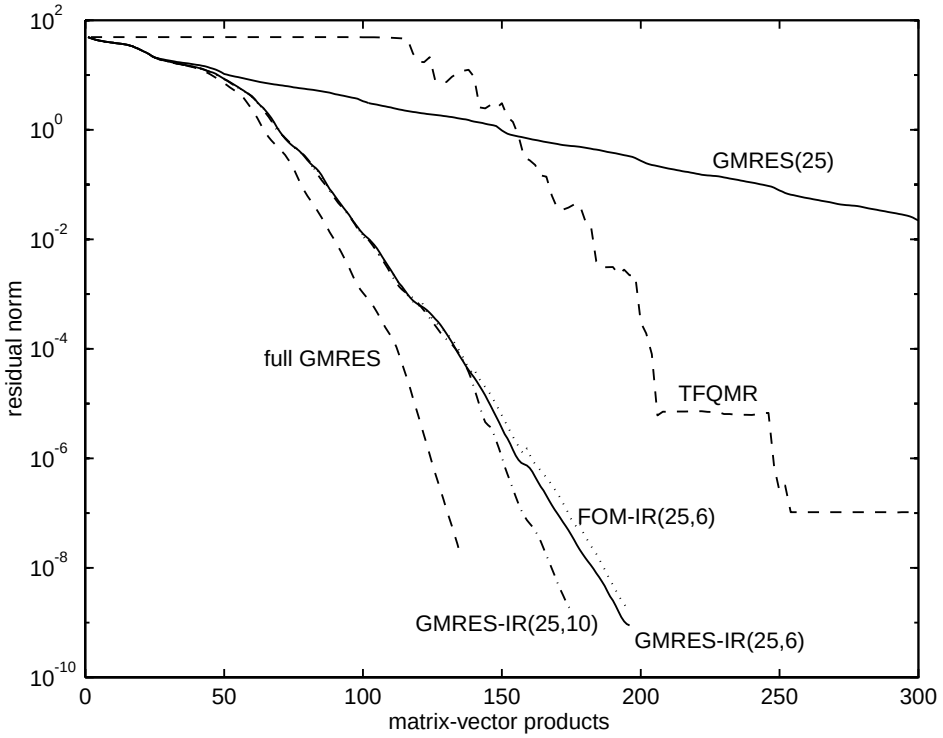

FIG. 2. *SHERMAN4 matrix.*

the number matrix-vector products is 157, only 15 more than with $m = 25$.

The next matrix is difficult for any restarted method.

Example 6.3. The Harwell-Boeing matrix SHERMAN5 is used. The dimension is 3312, and there are eigenvalues with positive and negative real parts. The eight eigenvalues nearest the origin are 0.047, 0.13, 0.40, 0.58, 0.62, 0.85, 0.91, and 1.0. The largest positive and negative eigenvalues are 595 and −189. Without preconditioning this is a difficult problem. GMRES-IR(25,6) stalls out and makes little progress in 350 cycles (over 6000 matrix-vector products). Good approximations to the smallest eigenvalues do not develop. GMRES-IR(30,8) works better; the slightly larger Krylov subspaces make a difference and good eigenvalue approximations are eventually found. Improving the residual norm by a factor of 10^{-6} takes 146 cycles or about 3221 matrix-vector products. Meanwhile, from of [9, Figure 6.2], TFQMR requires about 2700 matrix-vector products. The extremely large subspace generated by TFQMR is needed for this difficult indefinite problem. Finally we note that FOM-IR takes the same number of cycles as GMRES-IR for this indefinite problem. Further comparisons between the two seem warranted.

For more experiments, using the mathematically equivalent GMRES-E approach, see [28, 4, 42].

7. Eigenvalue problems. FOM-IR and GMRES-IR can also be thought of as eigenvalue methods. We briefly look at a few aspects of these methods.

TABLE 2
Comparison with different k and m values.

m	k	mat.-vec. products	thousands of vect.ops
25	0	526	14.9
25	1	298	8.9
25	2	166	5.3
25	3	147	4.9
25	4	142	5.0
25	5	140	5.3
25	6	137	5.5
25	7	138	6.0
25	8	137	6.3
25	9	136	6.7
25	10	137	7.4
7	4	323	8.9
10	4	198	4.9
15	4	157	4.3
20	4	146	4.6
30	4	134	5.4
35	4	135	6.0
40	4	134	6.4

7.1. With FOM-IR. As mentioned in section 5, FOM-IR implements the IRA eigenvalue method. It is interesting how simple it is to add the solution of linear equations on top of IRA. The subspace (4.1) for FOM-IR is not only a Krylov subspace itself, but contains several other Krylov subspaces that are interesting from an eigenvalue viewpoint. The subspaces are $\{y_i, Ay_i, A^2y_i, A^3y_i, \ldots, A^py_i\}$, for each i from 1 to k [29, Theorem 3]. Since these are Krylov subspaces with each Ritz vector as a starting vector, this shows why good eigenvector approximations can develop.

In this paragraph, we give another way of looking at the IRA method, and thus at FOM-IR. It is mentioned in [29] that the starting vector s at the beginning of a new cycle of IRA can be found without the implicit restarting. The vector s is a combination of the k Ritz vectors. The coefficients for this combination are from the solution of the $k-1$ by k homogeneous system of equations with jth row $[\theta_1^{j-1}\beta_{m1}, \theta_2^{j-1}\beta_{m2}, \ldots, \theta_k^{j-1}\beta_{mk}]$. Here we give this solution. Let ω_i be a polynomial with roots at all of the Ritz values except θ_i. Specifically,

$$\omega_i(\lambda) = \prod_{\substack{l=1 \\ l \neq i}}^{k} (\theta_l - \lambda).$$

Then

$$s = \sum_{i=1}^{k} \frac{1}{\beta_{mi}\omega_i(\theta_i)} y_i = V_m \sum_{i=1}^{k} \frac{1}{\beta_{mi}\omega_i(\theta_i)} g_i. \tag{7.1}$$

Formulas can also be given for the vectors $As, A^2s, \ldots, A^{k-1}s$. For $j < k$,

$$A^j s = \sum_{i=1}^{k} \frac{\theta_i^j}{\beta_{mi}\omega_i(\theta_i)} y_i = V_m \sum_{i=1}^{k} \frac{\theta_i^j}{\beta_{mi}\omega_i(\theta_i)} g_i. \tag{7.2}$$

This actually gives a way of implementing an algorithm mathematically equivalent to IRA that does not use the QR iteration (see [29] for yet another mathematically equivalent method). The vectors $s, As, A^2s, \ldots, A^{k-1}s$ need to be orthogonalized in order to form the first k columns of the new V matrix. Just as in IRA, k matrix-vector products are saved. This algorithm is not recommended, because of questionable numerical properties.

7.2. With GMRES-IR. GMRES-IR also solves both linear equations and the associated eigenvalue problem. It finds eigenvalues using the harmonic approach. This is equivalent to the interior Arnoldi with eigenvectors method that is given in [30] (with $\sigma = 0$, but other shifts could be incorporated). The implementation is more efficient. It can also be viewed as an interior eigenvalue version of the IRA method.

One application for the eigenvalue portion of GMRES-IR is with iterative methods for linear equations that require eigenvalue estimates [19, 26, 31]. GMRES-IR can find the eigenvalue estimates and at the same time begin solving the linear equations.

GMRES-IR also develops an interesting polynomial. Since subspace (5.1) is Krylov, there is a polynomial associated with it. And this polynomial has some zeros approximating eigenvalues near the origin. It may be of use for polynomial preconditioning [19].

The subspace for GMRES-IR contains Krylov subspaces with the harmonic Ritz vectors as starting vectors.

THEOREM 7.1. *The GMRES-IR subspace* (5.1) *contains the subspaces* $\{\tilde{y}_i, A\tilde{y}_i, A^2\tilde{y}_i, \ldots, A^p\tilde{y}_i\}$, *for each* i *from* 1 *to* k.

Proof. From (5.2),

$$A\tilde{y}_i = \tilde{\theta}_i\tilde{y}_i + \gamma_i u,$$

where u is a multiple of r_0. So $Span\{\tilde{y}_i, A\tilde{y}_i\} = Span\{\tilde{y}_i, r_0\}$. Next, again using (5.2),

$$A^2\tilde{y}_i = \tilde{\theta}_i^2\tilde{y}_i + \gamma_i\tilde{\theta}_i u + \gamma_i Au.$$

So $Span\{\tilde{y}_i, A\tilde{y}_i, A^2\tilde{y}_i\} = Span\{\tilde{y}_i, r_0, Ar_0\}$. This process can be continued until we reach

$$Span\{\tilde{y}_i, A\tilde{y}_i, \ldots, A^p\tilde{y}_i\} = Span\{\tilde{y}_i, r_0, Ar_0, \ldots, A^{p-1}r_0\}.$$

The right-hand side of this last equation is a subspace of (5.1), so we have the desired result. ∎

Theorem 7.1 shows why good approximations to eigenvectors can be expected in GMRES-IR. We next look at the eigenvalues that are found in Example 6.1 by GMRES-IR and note a correspondence between when the eigenvalues and linear equations converge.

Example 7.1. We continue Example 6.1 and look at the accuracy of the harmonic Ritz vectors that are developed by GMRES-IR(25,6). Figure 3 plots the norms of the residual vectors $\tilde{r}_i = A\tilde{y}_i - \rho_i\tilde{y}_i$, where ρ_i is the Rayleigh quotient of y_i. Figure 3 also has the residual norm for the linear equations. These linear equations residual norms are calculated at every matrix-vector product, while the eigenvalue residuals are computed only every 19 iterations. The linear equations residual does not begin to improve until after the approximations to the two smallest eigenvalues have made significant progress.

A version of GMRES-IR (equivalently GMRES-E) can be given that uses combinations of harmonic Ritz vectors instead of the QR-iteration, similar to (7.1) and (7.2)

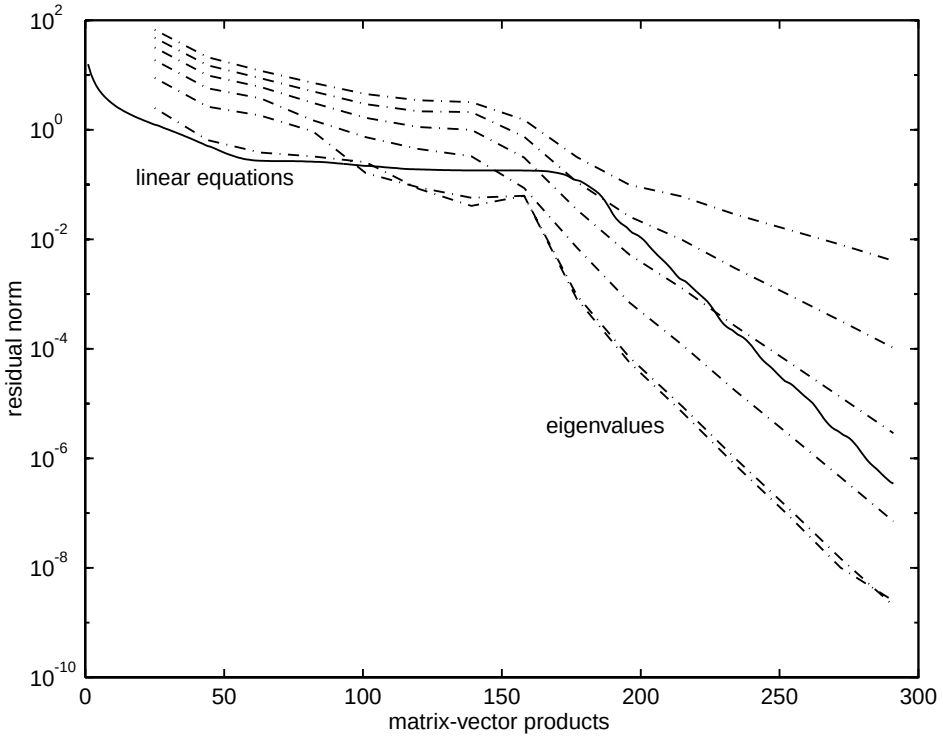

FIG. 3. *Convergence for the linear equations and six eigenvalues.*

for FOM-IR. For s the starting vector of the new GMRES cycle, the γ_i's from (5.2), and

$$\omega_i(\lambda) = \prod_{\substack{l=1 \\ l \neq i}}^{k} (\tilde{\theta}_l - \lambda),$$

if $0 \leq j < k$, then

$$A^j s = \sum_{i=1}^{k} \frac{\tilde{\theta}_i^j}{\gamma_i \omega_i(\tilde{\theta}_i)} \tilde{y}_i = V_m \sum_{i=1}^{k} \frac{\tilde{\theta}_i^j}{\gamma_i \omega_i(\tilde{\theta}_i)} \tilde{h}_i.$$

8. Conclusion. Implicit restarting can be used with the FOM and GMRES methods. A subspace is generated that contains both approximate eigenvectors and the necessary Krylov subspace with the linear equations residual vector as starting vector. Remarkably, the whole subspace is itself a Krylov subspace, though with a different starting vector. With the FOM-IR method, standard Ritz vectors are included in the subspace. With GMRES-IR, harmonic Ritz vectors are used.

Examples show that the implicitly restarted methods can be much better than standard GMRES, and also can be competitive with TFQMR, especially when the matrix-vector product is expensive. These new methods compute eigenvalues at the same time that they solve the linear equations.

We list some topics for future work. Most important is a study of the stability of the implicit restarting. We have seen stability problems when there are eigenvalues that stand out in the spectrum and thus converge rapidly. Such eigenvalues are not uncommon when preconditioning is used. Similar problems have been dealt with for the IRA method [23, 24].

Other possible topics include the possibility that eigenvectors could be released once they have converged to a certain point. An automatic procedure would be desirable for choosing the number of approximate eigenvectors and for selecting exterior Ritz values if they are outstanding; see [48] for related work.

Also, the polynomial preconditioning mentioned in the previous section is currently being investigated and there are possible applications to systems with multiple right-hand sides. The relation

$$AV_k = V_{k+1}\bar{H}_k,$$

that is formed after the implicit restarting, gives both approximate eigenvectors and their matrix-vector products with A in a compact form. With this relation, eigenvalues can be deflated in the solution of subsequent right-hand sides.

Acknowledgments. The author wishes to thank Rich Lehoucq for helpful discussions and the referees for their helpful suggestions.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[2] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.
[3] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1999), pp. 243–269.
[4] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1997), pp. 43–66.
[5] V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of a symmetric matrices*, Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
[6] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.
[7] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES preconditioned by deflation*, J. Comput. Appl. Math., 69 (1996), pp. 303–318.
[8] R. W. Freund, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, J. Comput. Appl. Math., 43 (1992), pp. 135–158.
[9] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
[10] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.
[11] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[12] 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.
[13] G. H. Golub and D. P. O'Leary, *Some history of the conjugate gradient and Lanczos algorithms:* 1948–1976, SIAM Rev., 31 (1989), pp. 50–102.
[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[15] A. Greenbaum, V. Ptak, and Z. Strakos, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
[16] A. Greenbaum and L. N. Trefethen, *GMRESICR and Arnoldi Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[17] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.
[18] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[19] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
[20] S. A. Kharchenko and A. Y. Yeremin, *Eigenvalue translation based preconditioners for the GMRES(k) method*, Numer. Linear Algebra Appl., 2 (1995), pp. 51–77.
[21] C. Lanczos, *An iterative method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282.
[22] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.
[23] R. B. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995.
[24] R. B. Lehoucq and D. C. Sorensen, *Deflation techniques for an implicitly restarted Arnoldi iteration*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 789–821.
[25] T. A. Manteuffel and J. Otto, *On the roots of orthogonal polynomials and residual polynomials associated with a conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 449–475.
[26] T. A. Manteuffel and G. Starke, *On hybrid iterative methods for nonsymmetric systems of linear equations*, Numer. Math., 73 (1996), pp. 489–506.
[27] R. B. Morgan, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154-156 (1991), pp. 289–309.
[28] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[29] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comp., 65 (1996), pp. 1213–1230.
[30] R. B. Morgan and M. Zeng, *Harmonic projection methods for large non-symmetric eigenvalue problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 33–55.
[31] 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.
[32] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.
[33] C. C. Paige, B. N. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.
[34] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[35] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[36] B. N. Parlett, D. R. Taylor, and Z. A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[37] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[38] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.
[39] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, NY, 1992.
[40] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[41] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, MA, 1995.
[42] Y. Saad, *Analysis of augmented Krylov subspace techniques*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 435–449.
[43] Y. Saad and M. H. Schultz, *GMRES: a generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[44] 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.
[45] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[46] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.
[47] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[48] A. Stathopoulos, Y. Saad, and K. Wu, *Dynamic thick restarting of the Davidson, and the implicitly restarted Arnoldi methods*, SIAM J. Sci. Comput., 19 (1998), pp. 227–245.
[49] E. De Sturler, *Truncation strategies for optimal Krylov subspace methods*, Technical report TR-96-38, Swiss Center for Scientific Computing (ETH Zurich), Zurich, Switzerland, 1996.
[50] E. De Sturler and D. R. Fokkema, *Nested Krylov methods and preserving the orthogonality*, in Proceedings Sixth Copper Mountain Conference on Multigrid Methods, NASA Con. Publ. 3224, Part 1, S. F. McCormick, N. D. Melson, and T. A. Manteuffel, eds., NASA Langley Research Center, Hampton, VA, 1993.
[51] C. H. Tong, *A comparative study of preconditioned lanczos methods for nonsymetric linear systems*, Technical report SAND91-8240, Sandia National Laboratories, Albuquerque, NM, 1992.
[52] H. A. van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.
[53] H. A. van der Vorst and C. Vuik, *GMRESR: A family of nested GMRES methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 369–386.
[54] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, New York, 1965.
[55] M. Zeng, *Finding Interior Eigenvalues of Large Nonsymmetric Matrices*, Ph.D. thesis, University of Missouri, Columbia, MO, 1996.