Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Numerical Mathematics 54 (2005) 222–236
APPLIED NUMERICAL MATHEMATICS
www.elsevier.com/locate/apnum

Restarted block-GMRES with deflation of eigenvalues

Ronald B. Morgan

Department of Mathematics, Baylor University, Waco, TX 76798-7328, USA

Available online 2 November 2004

Abstract

Block-GMRES is an iterative method for solving nonsymmetric systems of linear equations with multiple right-hand sides. Restarting may be needed, due to orthogonalization expense or limited storage. We discuss how restarting affects convergence and the role small eigenvalues play. Then a version of restarted block-GMRES that deflates eigenvalues is presented. It is demonstrated that deflation can be particularly important for block methods.

Keywords: Linear equations; Iterative methods; GMRES; Deflation; Block methods; Eigenvalues

1. Introduction

Block iterative methods are used for large systems of linear equations that have multiple right-hand sides. For nonsymmetric problems, block-GMRES [38] is one such method, but it may need restarting. Here we give a new version of restarted block-GMRES that deflates eigenvalues in order to improve convergence. Note the term deflation has also been used to describe removing some right-hand sides from the process [15]. But here we consider only deflation that involves removing of eigenvalues from the spectrum by including corresponding approximate eigenvectors in the subspace.

Block methods for linear equations [15,30,31,36,38,39,42] were developed for systems with several right-hand sides. There are some alternative approaches for dealing with such problems, especially for the symmetric case [6,11,33,36,41–43,46]. Block methods have also been developed for eigenvalue problems [1,3,8,17,18,35].

GMRES [40] is a popular Krylov method [38] for nonsymmetric linear equations. GMRES builds a fully orthogonal basis. It may need to be restarted due to either storage limits or the orthogonalization

E-mail address: ronald_morgan@baylor.edu (R.B. Morgan).

0168-9274/$30.00
doi:10.1016/j.apnum.2004.09.028

expenses that increase as the subspace grows. But restarted GMRES can suffer from poor performance compared to the unrestarted version. This is because larger Krylov subspaces are more effective [34], particularly for tough problems that have small eigenvalues. It is advantageous for the subspace to be large enough for rough approximations to develop for eigenvectors corresponding to the small eigenvalues. Then the eigenvalues essentially deflate from the spectrum and the convergence rate is accordingly improved [48].

Modifications of restarted GMRES have been developed to try to improve the convergence by deflating eigenvalues. Some approaches modify or precondition the matrix with approximate eigenvectors [2,4,12, 20]. Others save approximate eigenvectors from one cycle to the next and use them to augment the Krylov subspace [2,5,7,23,25,26,39]. Specially chosen vectors are also saved from one cycle of GMRES in [9].

In this paper, we combine the block approach with the idea of deflating eigenvalues. The next two sections give some information about block-GMRES and about deflation for GMRES. Then the new method is given in Section 4.

2. Block methods

We discuss how block methods work, and how they compare to non-block approaches. We concentrate on GMRES, but also quickly mention non-restarted methods such as QMR.

2.1. Block GMRES

Block methods take each right-hand side, build a Krylov subspace, and put the basis vectors all together to form a basis for one large subspace. Solutions for each right-hand side are extracted from this combined subspace. Let the p right-hand sides for a particular cycle of restarted block-GMRES be $r_0^{(1)}, r_0^{(2)}, \ldots, r_0^{(p)}$, and suppose that a subspace of dimension m is generated, with m a multiple of p. Then the subspace is

$$\text{Span}\{r_0^{(1)}, r_0^{(2)}, \ldots, r_0^{(p)}, Ar_0^{(1)}, Ar_0^{(2)}, \ldots, Ar_0^{(p)}, \ldots, A^{l-1}r_0^{(1)}, A^{l-1}r_0^{(2)}, \ldots, A^{l-1}r_0^{(p)}\},$$

where $l = \frac{m}{p}$. Block-GMRES finds minimum residual solutions from this subspace. There is a block Arnoldi recurrence formula [38]

$$AV_m = V_{m+p}\overline{H}_m, \tag{1}$$

where $\overline{H}_m$ is $m + p$ by m and is block or band upper-Hessenberg, and the columns of V_m span the subspace of dimension m.

2.2. Comparing block to non-block

Block methods allow several right-hand sides to be solved at the same time. They are particularly advantageous when the matrix–vector product is difficult to implement, so it is efficient to simultaneously apply the matrix to several vectors. This often happens in large finite element problems for which the matrix is never fully assembled and must be regenerated every time that is it accessed.

Convergence of a block method can be particularly affected by restarting. For example, assume that the maximum dimension subspace that can be used is 30. Then a regular GMRES approach can build a

Krylov subspace of dimension 30, while a block method with block-size of three will have three Krylov subspaces of dimension only 10. As mentioned earlier, Krylov subspaces of high dimension can be more effective than low dimension for difficult problems.

Next, some experiments are given for regular and block-GMRES. Tests with two different size subspaces are done for the block approach, first with subspaces of the same dimension as used for the non-block method, then with larger subspaces corresponding to putting together subspaces of the dimension used for the non-block method for each right-hand side. The experiments will provide support for the following points:

Point 1: For tough problems with small eigenvalues, convergence can be strongly degraded when there is a decrease in the size of the subspace per right-hand side. Then block-GMRES cannot compete well with regular GMRES if storage is quite limited or if the matrix–vector product is inexpensive.

Point 2: Here we assume the right-hand sides are unrelated. Putting Krylov subspaces generated for each right-hand side together, as is done in a block method, does not always improve performance. We conjecture that putting the subspaces together is advantageous only if the small eigenvalues are better determined by the combined subspace.

Point 3: For the case of closely related right-hand sides, putting the subspaces together is again not necessarily helpful.

Example 1. We compare regular restarted GMRES with subspaces of dimension 30 to restarted block-GMRES with combined subspaces of dimension 30 and 90. We use five test matrices. The first four are bidiagonal with superdiagonal entries all 1. Matrix 1 has diagonal entries $0.1, 1, 2, 3, \ldots, 999$. Matrix 2 has diagonal entries $1, 2, 3, \ldots, 1000$, and matrix 3 has $11, 12, 13, \ldots, 1010$. Finally, Matrix 4 has diagonal $10.1, 10.2, 10.3, \ldots, 19.9, 20, 21, 22, \ldots, 919, 920$. These matrices are in order of decreased influence of the small eigenvalues. The eigenvalues in the first problem are very important, as their small size makes the problem difficult. The fifth matrix is Sherman4 from the Harwell–Boeing sparse matrix collection [10]. It is nonsymmetric and has dimension 1104. The small eigenvalues are 0.031, 0.085, 0.28, 0.40, 0.43, 0.59, 0.63, 0.89, 0.92, and 1.00, and the largest eigenvalue is 66.5. Sherman4 is a little easier that matrix 1, but the influence of the small eigenvalues is similarly important. For each matrix, there are three right-hand sides, each with entries distributed random Normal$(0, 1)$. We count the total number of matrix–vector products needed to solve the three systems so that the residual norms are below 10^{-8}. Of course the block method solves the systems at the same time, while regular GMRES solves them one at a time. The results are given in Table 1. Bl-GMRES$(30, 3)$ denotes GMRES for three right-hand sides that is restarted whenever the total dimension reaches 30. Table 1 also lists the number (in thousands) of vec-

Table 1
Comparison of GMRES to block-GMRES, three right-hand sides

	mvp's for matrix 1	mvp's for matrix 2	vect. ops matrix 2	mvp's for matrix 3	mvp's for matrix 4	mvp's for Sherman4
GMRES(30)	–	1126	41.0	316	340	1824
Bl-GMRES$(30, 3)$	–	3355	144.1	397	432	3899
Bl-GMRES$(90, 3)$	–	1270	130.0	314	342	946

tor operations of length n for matrix 2 only. We will mainly emphasis the cost in terms of matrix–vector products, because block methods are more useful when matrix–vector products are the main expense.

All three approaches stall for matrix 1. Because of the restarting, they cannot generate an approximation to the smallest eigenvalue, and the problem is too difficult if this approximate eigenvector is not contained in the subspace. For matrix 2, block-GMRES$(30, 3)$ converges much slower than the other methods. For each cycle, it uses three Krylov subspaces of dimension 10, compared to regular GMRES's Krylov subspaces of dimension 30. And for this fairly difficult problem, the larger subspaces are needed. So if due to limited storage or significant orthogonalization costs, the maximum dimension of the subspace is 30, then block-GMRES may not compete well with regular GMRES. Meanwhile, block-GMRES$(90, 3)$ combines subspaces of dimension 30 for each right-hand side and performs better than block-GMRES$(30, 3)$. The comparison of vector operations shows that for such a sparse matrix, block is probably not the way to go.

Now for matrices 3 and 4, the problem is easy enough that small subspaces are fairly effective. GMRES(30) is only a little better than block-GMRES$(30, 3)$. For Sherman4, block-GMRES$(30, 3)$ needs twice as many matrix–vector products as does GMRES(30). As with matrix 2, Krylov subspaces of dimension 10 are not as effective as dimension 30. However, unlike for matrix 2, there is a significant improvement for block-GMRES$(90, 3)$ compared to GMRES(30). So putting together three Krylov subspaces of dimension 30 is more effective than working with them separately. We consider this further in the next example.

Example 2. As just mentioned, Sherman4 gets considerably better results with Krylov subspaces of dimension 30 put together in block-GMRES$(90, 3)$. The associated Ritz values [34,37] show why this happens. The Ritz values and the corresponding Ritz vectors are more accurate when the three subspaces of dimension 30 are put together. After the first cycle, the first five Ritz values for the first right-hand side are 0.12, 0.78, 1.0, 2.2, and 3.6. For the block method, they are more accurate: 0.092, 0.29, 0.40, 0.65, and 0.80 (the true values are 0.031, 0.085, 0.28, 0.40, and 0.43). And it is interesting that the second cycle of block-GMRES$(90, 3)$ generates even more accurate approximations: 0.031, 0.085, 0.29, 0.43, and 0.48. This shows why the block method is more efficient. It develops good enough approximate eigenvectors for deflation of eigenvalues that leads to superlinear convergence [48]. The next example continues this examination of the Ritz values, but for matrix 2.

Example 3. Unlike with Sherman4, the results for matrix 2 did not show improvement when the subspaces of dimension 30 were put together to from a large subspace of dimension 90. Now we try larger subspaces. Subspaces of dimension 60 are put together to create a dimension 180 subspace. The number of matrix–vector products is 1031 for separate solution of the three right-hand sides and 756 for the block approach. So putting together the subspaces of dimension 60 works better than for size 30. The Ritz values are given in Table 2. With the non-block approach, Ritz values are not accurate with either $m = 30$ or $m = 60$. Also, putting the subspaces of dimension 30 together does not help much. However, fairly good approximations are produced when the larger subspaces are put together. The small eigenvalues are then deflated, so the block method gives faster convergence.

Example 4. We now consider the case of the right-hand sides being related to each other. For the matrix Sherman4, let the second and third right-hand sides be equal to the first added to 10^{-4} times other

Table 2
Matrix 2: Ritz values after 1st cycle for GMRES and block-GMRES

mvp's	Ritz val.'s 1st rhs	Ritz val.'s 2nd rhs	Ritz val.'s 3rd rhs	mvp's	Ritz val.'s spaces combined
GMRES(30)				Bl-GMRES(90, 3)	
1126	2.2	2.5	2.6	1270	1.77
	8.9	10.6	8.1		2.69
	20.8	18.2	18.1		5.50
	38.1	38.9	38.2		8.55
	58.8	58.0	58.9		11.4
GMRES(60)				Bl-GMRES(180, 3)	
1031	1.4	0.93	1.1	756	1.00075
	4.1	2.7	3.0		2.00310
	6.3	6.0	5.8		3.00004
	10.4	12.4	9.3		4.01129
	16.2	15.7	15.1		5.12671

Normal$(0, 1)$ random vectors. Bl-GMRES$(90, 3)$ converges in 934 matrix–vector products, only slightly less than the number for the right-hand sides not related. So as mentioned earlier in point 3, having the right-hand sides related does not necessarily help. The next theorem shows why.

Theorem 1. *Let the right-hand sides be* $r_0^{(1)}$ *and* $r_0^{(j)} = r_0^{(1)} + \varepsilon * w^{(j)}$, *for* $j = 2, \ldots, p$. *Then the block Krylov subspace with these related right-hand sides as starting vectors is equivalent to the block Krylov subspace*

$$\text{Span}\{r_0^{(1)}, w^{(2)}, \ldots, w^{(p)}, Ar_0^{(1)}, Aw^{(2)}, \ldots, Aw^{(p)}, \ldots, A^{l-1}r_0^{(1)}, A^{l-1}w^{(2)}, \ldots, A^{l-1}w^{(p)}\},$$

where $l = \frac{m}{p}$.

The proof is trivial, since there is simply a change of basis for the starting subspace. The theorem shows that using a block Krylov method in the situation of related right-hands cannot be expected to be any better than in the non-related case. The solution of the first system with $r_0^{(1)}$ as right-hand side will actually be identical to solving it with the block Krylov subspace with random starting vectors $r_0^{(1)}$ and $w^{(i)}$, $i = 2, \ldots, p$.

On the other hand, Freund and Malhotra [15] give an example with quite a few closely related right-hand sides, for which it is possible to reduce the number of right-hand sides during their block-QMR process. This not only cuts down on the number of systems of equations that must be solved, but also deals with the possible numerical instability of block-QMR for related right-hand sides. With proper reorthogonalization, block-GMRES is not as prone to instability. The next subsection considers further the block-QMR method.

2.3. Non-restarted block methods

Iterative methods based on the nonsymmetric Lanczos algorithm [38] such as QMR [16], TFQMR [14] and Bi-CGSTAB [47] do not need to be restarted. Freund and Malhotra have made software available for their block-QMR [15] method in the BL-QMR package.

Table 3
Comparison of QMR to block-QMR, three right-hand sides

	mvp's for matrix 1	mvp's for matrix 2	vect. ops matrix 2	mvp's for matrix 3	mvp's for matrix 4	mvp's for Sherman4
QMR	1428	1192	12.0	622	678	886
Bl-QMR	786	712	15.7	544	714	540

In the non-block case, QMR generally competes well against restarted GMRES when the matrix–vector product is inexpensive and/or when the problem is difficult. However, GMRES requires only one matrix–vector product per iteration, so it has an advantage over QMR if the matrix–vector product is expensive and if the methods converge in a similar number of iterations. The situation is similar for the block methods. We will give an example that supports the following point.

Point 4: For difficult problems, block-QMR can be better than block-GMRES. However, block-QMR may not be as efficient as regular QMR when the matrix–vector product is inexpensive.

Example 5. Table 3 gives the number of matrix–vector products for solving the systems of linear equations with three right-hand sides using QMR and block-QMR. For matrix 2, QMR takes about the same number of matrix–vector products as GMRES(30) (1192 versus 1126), but it uses less than a third as many length n vector operations. GMRES(30) has greater orthogonalization expenses that are significant for such a sparse matrix. Block-QMR takes less matrix–vector products than block-GMRES$(90, 3)$ and far less than block-GMRES$(30, 3)$. This is because block-QMR puts together large subspaces and is able to deflate eigenvalues. This is similar to what was illustrated in Example 3 for block-GMRES$(180, 3)$, but block-QMR can be even more effective. The three Krylov subspaces that block-QMR is putting together keep getting larger. Block-GMRES is competitive with block-QMR only for the easier problems with matrix 3 and matrix 4.

We note that block-QMR requires less matrix–vector products than regular QMR, except for matrix 4, where deflating eigenvalues is not helpful. However, block-QMR can be more expensive for sparse matrices. For matrix 2, block-QMR takes about one third more vector ops than regular QMR to solve the three systems of equations. But for a more expensive matrix–vector product, the extra overhead of the block approach would not be so significant.

3. Deflated versions of GMRES

Using small subspaces for restarted GMRES can slow convergence for difficult problems (as we saw in the previous section with block-GMRES$(30, 3)$). Deflated versions of restarted GMRES [2,4,5,7,9,12, 20,23,25,26,39] can improve this, when the problem is difficult due to a few small eigenvalues. One of these approaches is related to Sorensen's implicitly restarted Arnoldi method for eigenvalues [21,44] and is called GMRES with implicit restarting [25]. A mathematically equivalent method, called GMRES with deflated restarting (GMRES-DR) [26], is also related to Wu and Simon's restarted Lanczos eigenvalue method [49]. See [24,29,45] for some other related eigenvalue methods.

We will concentrate on GMRES-DR, because it is efficient and relatively simple. Approximate eigenvectors corresponding to the small eigenvalues are computed at the end of each cycle and are put at the

Table 4
Comparison of deflated to regular GMRES

	mvp's for matrix 1	mvp's for matrix 2	vect. ops matrix 2	mvp's for matrix 3	mvp's for matrix 4	mvp's for Sherman4
GMRES(30)	–	403	14.7	106	114	548
GMRES-DR(30, 6)	252	208	9.2	104	114	165

beginning of the next subspace. Letting r_0 be the initial residual for the linear equations at the start of the new cycle and $\tilde{y}_1, \ldots, \tilde{y}_k$ be harmonic Ritz vectors [13,22,28,32], the subspace of dimension m used for the new cycle is

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

This can be viewed as a Krylov subspace generated with starting vector r_0 augmented with approximate eigenvectors. Remarkably the whole subspace turns out to be a Krylov subspace itself (though not with r_0 as starting vector) [25]. Once the approximate eigenvectors are moderately accurate, their inclusion in the subspace for GMRES essentially deflates the corresponding eigenvalues from the linear equations problems.

Example 6. Table 4 shows the improvement over regular restarted GMRES with the test matrices, for the first right-hand side only. GMRES-DR(30, 6) uses subspaces of maximum dimension 30, including six approximate eigenvectors. So even with the same size subspaces (and lesser size Krylov subspaces with r_0 as starting vector), the deflated method is much better for the tougher problems.

4. The new approach

We wish to give a block version of GMRES with deflation of eigenvalues. Saad has unpublished work on this. Also Gu and Cao [19] have a version of block-GMRES augmented by eigenvectors. It is generalization of the method from [23] which puts the eigenvectors after the Krylov vectors and is not as efficient as the GMRES-DR approach. We now give a block generalization of GMRES-DR.

4.1. Block-GMRES with deflated restarting

Eigenvalues can be deflated from block-GMRES in a fashion similar to the GMRES-DR method. Harmonic Ritz vectors are added to the subspace. This gives us the block-GMRES-DR method, which uses the subspace

$$\text{Span}\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k, r_0^{(1)}, \ldots, r_0^{(p)}, Ar_0^{(1)}, \ldots, Ar_0^{(p)}, \ldots, A^{l-1}r_0^{(1)}, \ldots, A^{l-1}r_0^{(p)}\}, \tag{2}$$

where $l = \frac{m-k}{p}$. It will be shown that if m and k are divisible by p, then this subspace contains p Krylov subspaces of dimension m/p. First, we need a lemma establishing that eigenvalue residuals are related to linear equations residuals. In the lemma and the following theorems, we assume that there are not any special cases with sets of vectors becoming linearly dependent.

Lemma 2. *If a block Krylov subspace is used for solving both linear equations with multiple right-hand sides and for computing approximate eigenpairs, the harmonic residual vectors are all linear combinations of the residual vectors from the minimum residual solution of the linear equations problems.*

Proof. Let V be the orthonormal matrix with m columns spanning a block Krylov subspace

$$\text{Span}\{s^{(1)}, \ldots, s^{(p)}, As^{(1)}, \ldots, As^{(p)}, \ldots, A^{m/p-1}s^{(1)}, \ldots, A^{m/p-1}s^{(p)}\}.$$

The minimum residual solution of the linear equations with the jth right-hand side can come from solving the equations

$$V^{\mathrm{T}}A^{\mathrm{T}}AVd^{(j)} = V^{\mathrm{T}}A^{\mathrm{T}}r_0^{(j)}. \tag{3}$$

The new residual from the minimum residual solution is then $r^{(j)} = r_0^{(j)} - AVd^{(j)}$, so Eq. (3) is equivalent to

$$(AV)^{\mathrm{T}}r^{(j)} = 0.$$

We next show that the harmonic residual vectors satisfy the same orthogonality condition. For harmonic Rayleigh–Ritz, we solve

$$V^{\mathrm{T}}A^{\mathrm{T}}AV\tilde{g} = \tilde{\theta}V^{\mathrm{T}}A^{\mathrm{T}}V\tilde{g}, \tag{4}$$

with the harmonic Ritz values being the $\tilde{\theta}_i$'s. The harmonic Ritz vectors are then the vectors $\tilde{y}_i = V\tilde{g}_i$. The harmonic residual vectors are $A\tilde{y}_i - \tilde{\theta}_i\tilde{y}_i$. To see the orthogonality to the columns of AV, write Eq. (4) as

$$(AV)^{\mathrm{T}}\big(A\tilde{y}_i - \tilde{\theta}_i\tilde{y}_i\big) = 0.$$

So the p-dimensional subspaces of linear equations residuals and harmonic residuals are both orthogonal to a subspace of dimension m. They are also both contained in a subspace of dimension $m + p$, so they are the same subspace. □

Theorem 3. *We consider a cycle of block-GMRES-DR. Assume both m and k are divisible by p. Let $r_0^{(1)}, \ldots, r_0^{(p)}$ be the residual vectors and $\tilde{y}_1, \ldots, \tilde{y}_k$ be the harmonic Ritz vectors calculated during the previous cycle. Then*

$$\begin{aligned}&\text{Span}\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k, r_0^{(1)}, \ldots, r_0^{(p)}, Ar_0^{(1)}, \ldots, Ar_0^{(p)}, \ldots, A^{l-1}r_0^{(1)}, \ldots, A^{l-1}r_0^{(p)}\}\\ &\quad = \text{Span}\{s^{(1)}, s^{(2)}, \ldots, s^{(p)}, As^{(1)}, As^{(2)}, \ldots, As^{(p)}, \ldots,\\ &\qquad A^{m/p-1}s^{(1)}, A^{m/p-1}s^{(2)}, \ldots, A^{m/p-1}s^{(p)}\},\end{aligned}$$

for some vectors $s^{(1)}, s^{(2)}, \ldots, s^{(p)}$. Here $l = \frac{m-k}{p}$. So the subspace for block-GMRES-DR is a block Krylov subspace (*is a combination of Krylov subspaces*).

Proof. From the lemma,

$$A\tilde{y}_i - \tilde{\theta}_i\tilde{y}_i = \sum_{j=1}^{p}\alpha_i^{(j)}r_0^{(j)}. \tag{5}$$

We first claim that

$$\text{Span}\{\tilde{y}_1, \ldots, \tilde{y}_k\} = \text{Span}\left\{s^{(1)}, \ldots, s^{(p)}, As^{(1)}, \ldots, As^{(p)}, \ldots, A^{k/p-1}s^{(1)}, \ldots, A^{k/p-1}s^{(p)}\right\}, \tag{6}$$

for some vectors $s^{(1)}, s^{(2)}, \ldots, s^{(p)}$. Let

$$s = \sum_{i=1}^{k} \beta_i \tilde{y}_i. \tag{7}$$

Multiplying by A and using Eq. (5),

$$As = \sum_{i=1}^{k} \beta_i \tilde{\theta}_i \tilde{y}_i + \sum_{i=1}^{k} \beta_i \left(\sum_{j=1}^{p} \alpha_i^{(j)} r_0^{(j)} \right) = \sum_{i=1}^{k} \beta_i \tilde{\theta}_i \tilde{y}_i + \sum_{j=1}^{p} \left(\sum_{i=1}^{k} \alpha_i^{(j)} \beta_i \right) r_0^{(j)}.$$

To have As in the span of the $\tilde{y}_i$ vectors, we will make

$$\sum_{i=1}^{k} \alpha_i^{(j)} \beta_i = 0,$$

for j from 1 to p. We set these equations up as part of a homogeneous system of linear equations with the β_i's as unknowns. So the first p equations of this system have the $\alpha_i^{(j)}$'s as coefficients, and more equations will be added. With the $r_0^{(j)}$ terms eliminated, we have

$$As = \sum_{i=1}^{k} \beta_i \tilde{\theta}_i \tilde{y}_i.$$

Multiplying again by A and using Eq. (5) gives

$$A^2 s = \sum_{i=1}^{k} \beta_i \tilde{\theta}_i^2 \tilde{y}_i + \sum_{j=1}^{p} \left(\sum_{i=1}^{k} \alpha_i^{(j)} \theta_i \beta_i \right) r_0^{(j)}.$$

To again eliminate the $r_0^{(j)}$ terms, we let the next p equations for our homogeneous system have coefficients $\alpha_i^{(j)} \theta_j$. Similarly, the next p equations will have $\alpha_i^{(j)} \theta_j^2$'s as coefficients. We continue this until we have $A^{k/p-1}s$ is a linear combination of the y_i's. The homogeneous system then has $k - p$ equations. Since the system has k unknowns, there are p linearly independent solutions for the β_i's. Putting these in (7) gives the desired vectors $s^{(1)}, \ldots, s^{(p)}$ that satisfy Eq. (6).

We now multiply again by A and use Eq. (5):

$$A^{k/p} s = \sum_{i=1}^{k} \beta_i \tilde{\theta}_i^{k/p} \tilde{y}_i + \sum_{j=1}^{p} \left(\sum_{i=1}^{k} \alpha_i^{(j)} \theta_j^{k/p-1} \beta_i \right) r_0^{(j)}.$$

So $A^{k/p}s$ is a linear combination of the $\tilde{y}_i$'s and the $r_0^{(j)}$'s. $A^{k/p+1}s$ will be a linear combination of y_i's, $r_0^{(j)}$'s, and $Ar_0^{(j)}$'s. We continue until we have $A^{m/p-1}s$ is a linear combination of the y_i's and $r_0^{(j)}$'s multiplied by powers of A up to $\frac{m-k}{p} - 1$. This establishes the result in the theorem. □

Theorem 4.2 is important, because it tells us there can be a fairly simple and efficient implementation. There is a block Arnoldi-type recurrence formula similar to (1):

$$AV_m = V_{m+p} \overline{H}_m, \tag{8}$$

where $\overline{H}_m$ is block or band upper-Hessenberg, *except* for a full leading $k+p$ by $k+p$ portion. With (8), the vectors $A\tilde{y}_i$ can be used without actually being computed and stored. This saves expense and saves storage of k vectors of length n. The theorem tells us that we still have Krylov subspaces even though the approximate eigenvectors are put first in the subspace. Note that Theorem 4.2 works only if the approximate eigenvectors are chosen to be harmonic Ritz vectors.

For the case of k not being divisible by p, subspace (2) is no longer strictly a block Krylov subspace. However, the algorithm will still work. Using (5), it can be shown that there still is a recurrence (8) even when there are extra y_i's in the subspace.

The next theorem tells us that subspace (2) contains Krylov subspaces with every harmonic Ritz vector as a starting vector. This property gives reason to expect efficient computation of the eigenvalues. See [26] for a discussion of the advantages of this for regular GMRES-DR.

Theorem 4. *Assume both m and k are divisible by p. The block-GMRES-DR subspace* (2) *contains the subspaces* $\{\tilde{y}_i, A\tilde{y}_i, A^2\tilde{y}_i, \ldots, A^{(m-k)/p}\tilde{y}_i\}$, *for each i from* 1 *to* k.

We omit the proof, but it just involves taking $\tilde{y}_i$, multiplying by A $\frac{m-k}{p}$ times, and using Eq. (5) after each multiplication. Then one can observe that all of the vectors thus generated are combinations of $\tilde{y}_i$, the $r_0^{(j)}$'s, and $r_0^{(j)}$'s multiplied by powers of A.

4.2. The algorithm

We use Ruhe's variant of block Arnoldi, and do not need m and k to be multiples of p. Superscripts denote different right-hand sides and the corresponding different solution vectors. Some MATLAB notation is used; for instance, $\overline{H}(m+1:m+p, 1:m)$ denotes the portion of $\overline{H}$ with rows from $m+1$ to $m+p$ and columns from 1 to m. Meanwhile, H_m is the m by m portion of the $m+p$ by m matrix $\overline{H}_m$.

Block-GMRES-DR

(1) *Start*: Let the p right-hand sides be $b^{(1)}, b^{(2)}, \ldots, b^{(p)}$. Choose m, the maximum size of the subspace, and k, the desired number of approximate eigenvectors. Choose initial guesses $x_0^{(1)}, x_0^{(2)}, \ldots, x_0^{(p)}$, and compute initial residual vectors $r_0^{(1)} = b - Ax_0^{(1)}, \ldots$. The recast problems are $A(x^{(1)} - x_0^{(1)}) = r_0^{(1)}, \ldots$. Form V_p by orthonormalizing $r_0^{(1)}, r_0^{(2)}, \ldots, r_0^{(p)}$.

(2) *Find approximate solutions*: Generate V_{m+p} and $\overline{H}_m$ with the block-Arnoldi iteration. Solve $\min\|C - \overline{H}_m D\|$ for D, where the ith column of C is $V_{m+p}^{\mathrm{T}} r_0^{(i)}$. Form the new approximate solutions $x_m^{(i)} = x_0^{(i)} + V_m d_i$, with d_i the ith column of D. Compute the residual vectors $r^{(i)} = b - Ax_m^{(i)} = V_{m+1}(c_i - \overline{H}_m d_i)$. Check residual norms $\|r^{(i)}\| = \|c_i - \overline{H}_m d_i\|$ for convergence, and proceed if not satisfied.

(3) *Begin restart*: Let $x_0^{(i)} = x_m^{(i)}$ and $r_0^{(i)} = r^{(i)}$. Compute the k smallest (or others, if desired) eigenpairs $(\tilde{\theta}_i, \tilde{g}_i)$ of $H_m + H_m^{-\mathrm{T}} \overline{H}_m(m+1:m+p, 1:m)^{\mathrm{T}} \overline{H}_m(m+1:m+p, m-p+1:m)$. The $\tilde{\theta}_i$ are harmonic Ritz values.

(4) *Orthonormalization of first k vectors*: Orthonormalize the $\tilde{g}_i$'s, first separating into real and imaginary parts if complex, in order to form an m by k matrix P_k. (It may be necessary to adjust k in order to make sure both parts of complex vectors are included.)

(5) *Orthonormalization of $k+1,\ldots,k+p$ vectors*: First extend with zero entries the vectors $p_1,\ldots,p_k$ to length $m+p$, then orthonormalize the columns of $C-\overline{H}_m D$ against $p_1,\ldots,p_k$ to form $p_{k+1},\ldots,p_{k+p}$. P_{k+p} is $m+p$ by $k+p$.
(6) *Form portions of new H and V using the old H and V*: Let $\overline{H}_k^{\text{new}} = P_{k+p}^{\mathrm{T}}\overline{H}_m P_k$ and $V_{k+p}^{\text{new}} = V_{m+p}P_{k+p}$. Then let $\overline{H}_k = \overline{H}_k^{\text{new}}$ and $V_{k+p} = V_{k+p}^{\text{new}}$.
(7) *Reorthogonalization of $k+1,\ldots,k+p$ vectors*: Orthogonalize $v_{k+1},\ldots,v_{k+p}$ against the earlier columns of the new V_{k+p}. Go to step 2.

We now derive the eigenvalue portion of Step 3 of the algorithm. Harmonic Rayleigh–Ritz applied to the subspace spanned by the columns of V_m gives the small generalized eigenvalue problem

$$V_m^{\mathrm{T}}A^{\mathrm{T}}AV_m g = \theta V_m^{\mathrm{T}}A^{\mathrm{T}}V_m g.$$

The block Arnoldi recurrence (8) converts this to

$$\overline{H}_m^{\mathrm{T}}\overline{H}_m g = \theta H_m^{\mathrm{T}} g. \tag{9}$$

Breaking $\overline{H}_m$ into two pieces $\overline{H}_m = H_m + \overline{H}_m(m+1:m+p,1:m)$, (9) becomes

$$\left(H_m^{\mathrm{T}}H_m + \overline{H}_m(m+1:m+p,1:m)^{\mathrm{T}}\overline{H}_m(m+1:m+p,1:m)\right)g = \theta H_m^{\mathrm{T}} g. \tag{10}$$

Next we multiply through by $H_m^{-\mathrm{T}}$ and use that $\overline{H}_m(m+1:m+p,1:m)$ is zero except in the last p columns to get the standard eigenvalue problem in Step 3 with matrix

$$\left(H_m + H_m^{-\mathrm{T}}\overline{H}_m(m+1:m+p,1:m)^{\mathrm{T}}\overline{H}_m(m+1:m+p,m-p+1:m)\right). \tag{11}$$

Solving the generalized eigenvalue problem in Eq. (10) might be more stable than (11) if H_m is nearly singular, and more study is needed of this.

4.3. Eigenvalue computations

Associated with block-GMRES-DR is an eigenvalue computation method. It can be called a block harmonic restarted Arnoldi method. A block restarted Arnoldi method can also be devised with standard Rayleigh–Ritz instead of harmonic. Unlike in the non-block case, this approach is not equivalent to block implicitly restarted Arnoldi. Comparisons would be interesting.

4.4. Experiments

We now test the new deflated version of block-GMRES against other GMRES methods. The following point will be supported.

Point 5: Deflating eigenvalues can be even more important for a block method.

Example 7. The matrices are the same as in Example 1. Block-GMRES-DR(m,p,k) uses a subspace of maximum dimension m, including k approximate eigenvectors, to solve systems with p right-hand sides. Table 5 has first three rows the same as in Table 1 and has results for deflated methods on the next four rows. Again the number of matrix–vector products is given for solving systems with three right-hand sides. Deflating eigenvalues makes the block methods much more effective for the tougher problems. For example with matrix 2, block-GMRES-DR$(30,3,6)$ uses only one fifth of the number of matrix–vector

Table 5
Comparison of deflated to regular versions of GMRES and block-GMRES, 3 rhs's

	mvp's for matrix 1	mvp's for matrix 2	vect. ops matrix 2	mvp's for matrix 3	mvp's for matrix 4	mvp's for Sherman4
GMRES(30)	–	1126	41.0	316	340	1824
Bl-GMRES(30, 3)	–	3355	144.1	397	432	3899
Bl-GMRES(90, 3)	–	1270	130.0	314	342	946
GMRES-DR(30, 6)	737	609	26.8	306	340	501
Bl-G.-DR(30, 3, 6)	836	671	36.4	328	426	559
Bl-G.-DR(90, 3, 6)	541	460	48.5	272	339	352
Bl-G.-DR(90, 3, 18)	412	371	45.0	263	336	295

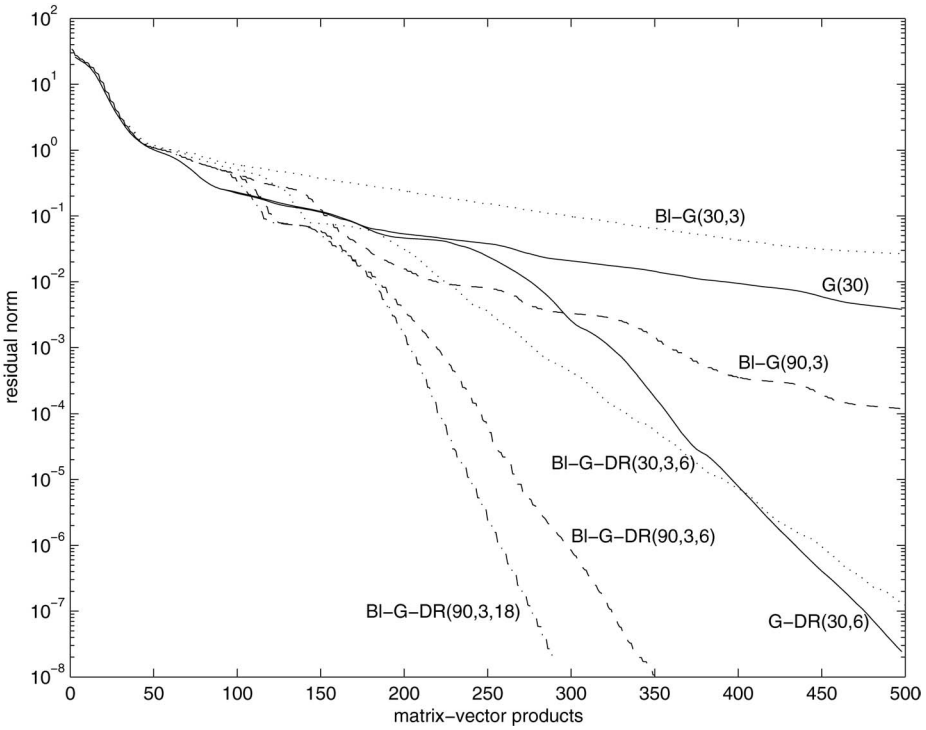

Fig. 1. Comparison of convergence for Sherman4 matrix.

products as does block-GMRES(30, 3) (671 compared to 3355). Also, block-GMRES-DR(30, 3, 6) is almost as effective as GMRES-DR(30, 6) in spite of the fact that it uses much smaller Krylov subspaces. So, deflating eigenvalues appears especially important for block methods. The deflation can sometimes turn the problem into an easier one that can be handled with small subspaces.

Fig. 1 has all the methods listed in Table 5 for the matrix Sherman4. Both non-block methods have solid lines. Both block-GMRES methods (with and without deflated restarting) with maximum dimension of 30 have dotted lines. Bl-GMRES(90, 3) and Bl-GMRES-DR(90, 3, 6) have dashed lines and

Bl-GMRES-DR(90, 3, 18) is dot-dashed. All the deflated methods give significantly improved convergence over the non-deflated methods with corresponding size subspaces.

We now compare the deflated block-GMRES approach with block-QMR (see also Table 3). For all matrices, block-GMRES(90, 3, 6) uses less matrix–vector products than block-QMR. For instance, 460 versus 712 for matrix 2. So for the case of expensive matrix–vector product, GMRES can be better. For particularly sparse matrices, it may be considerably more costly.

5. Conclusion

We have presented a block-GMRES method with deflation of eigenvalues. For tough problems with small eigenvalues, it is much better than standard block-GMRES. It can compete with block-QMR, at least for some problems with fairly expensive matrix–vector products and preconditioners. For problems with inexpensive matrix–vector products, block methods may not always be the best approach.

The rest of this section gives some possible extensions of this project. We plan to develop a method for the case where there are many right-hand sides. This would involve applying the block-GMRES-DR method to only some of the right-hand sides in order to keep down storage needs and orthogonalization expense. The eigenvector information generated while solving this first group of right-hand sides can then be used to deflate the eigenvalues for the other right-hand sides in an approach that alternates block-GMRES with projection over the eigenvectors (see [26,27] for some discussion of a non-block version). Proper choice of the block-size could then make the method more efficient. We also plan to look at solving the other right-hand sides with projection over the eigenvectors followed by block-QMR.

Another project would involve removal of some right-hand sides from the problem as block-GMRES-DR proceeds. This is usually called deflation of right-hand sides [1,8,15,30,35].

We wish to investigate implicitly restarted block methods for both computing eigenvalues and for solving linear equations and compare these methods with the block, deflated approaches given here. As mentioned earlier, implicit restarting is not equivalent to the deflated restarting approach in the block case.

Acknowledgements

The author wishes to thank Oliver Ernst for suggesting that Eq. (10) may be more stable. Also thanks to the referees for their suggestions for improving the paper.

References

[1] J.I. Aliaga, D.L. Boley, R.W. Freund, V. Hernández, A Lanczos-type method for multiple starting vectors, Math. Comp. 69 (2000) 1577–1601.

[2] J. Baglama, D. Calvetti, G.H. Golub, L. Reichel, Adaptively preconditioned GMRES algorithms, SIAM J. Sci. Comput. 20 (1998) 243–269.

[3] Z. Bai, D. Day, Q. Ye, Able: an adaptive block Lanczos method for non-Hermitian eigenvalue problems, SIAM J. Matrix Anal. Appl. 20 (1999) 1060–1082.

[4] K. Burrage, J. Erhel, On the performance of various adaptive preconditioned GMRES strategies, Numer. Linear Algebra Appl. 5 (1998) 101–121.
[5] C.L. Calvez, B. Molina, Implicitly restarted and deflated GMRES, Numer. Algorithms 21 (1999) 261–285.
[6] T.F. Chan, W. Wan, Analysis of projection methods for solving linear systems with multiple right-hand sides, SIAM J. Sci. Comput. 18 (1997) 1698–1721.
[7] A. Chapman, Y. Saad, Deflated and augmented Krylov subspace techniques, Numer. Linear Algebra Appl. 4 (1997) 43–66.
[8] J. Cullum, W.E. Donath, A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace of large sparse real symmetric matrices, in: Proc. of the 1974 IEEE Conf. on Decision and Control, Phoenix, Arizona, 1974, pp. 505–509.
[9] E. De Sturler, Truncation strategies for optimal Krylov subspace methods, SIAM J. Numer. Anal. 36 (1999) 864–889.
[10] I.S. Duff, R.G. Grimes, J.G. Lewis, Sparse matrix test problems, ACM Trans. Math. Software 15 (1989) 1–14.
[11] J. Erhel, F. Guyomarc'h, An augmented conjugate gradient method for solving consecutive symmetric positive definite linear systems, SIAM J. Matrix Anal. Appl. 21 (2000) 1279–1299.
[12] J. Erhel, K. Burrage, B. Pohl, Restarted GMRES preconditioned by deflation, J. Comput. Appl. Math. 69 (1996) 303–318.
[13] R.W. Freund, Quasi-kernel polynomials and their use in non-Hermitian matrix iterations, J. Comput. Appl. Math. 43 (1992) 135–158.
[14] R.W. Freund, A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems, SIAM J. Sci. Comput. 14 (1993) 470–482.
[15] R.W. Freund, M. Malhotra, A block QMR algorithm for non-Hermitian linear systems with multiple right-hand sides, Linear Algebra Appl. 254 (1997) 119–157.
[16] R.W. Freund, N.M. Nachtigal, QMR: A quasi-minimal residual method for non-Hermitian linear systems, Numer. Math. 60 (1991) 315–339.
[17] G.H. Golub, R. Underwood, The block Lanczos method for computing eigenvalues, in: J. Rice (Ed.), Mathematical Software III, Academic Press, New York, 1977, pp. 364–377.
[18] G.H. Golub, R. Underwood, J.H. Wilkinson, The Lanczos algorithm for the symmetric $Ax = \lambda Bx$ problem, Tech. Rep., Stanford University Computer Science Department, Stanford, CA, 1972.
[19] G. Gu, Z. Cao, A block GMRES method augmented with eigenvectors, Appl. Math. Comput. 121 (2001) 271–289.
[20] S.A. Kharchenko, A.Y. Yeremin, Eigenvalue translation based preconditioners for the GMRES(k) method, Numer. Linear Algebra Appl. 2 (1995) 51–77.
[21] R.B. Lehoucq, D.C. Sorensen, Deflation techniques for an implicitly restarted Arnoldi iteration, SIAM J. Matrix Anal. Appl. 17 (1996) 789–821.
[22] R.B. Morgan, Computing interior eigenvalues of large matrices, Linear Algebra Appl. 154–156 (1991) 289–309.
[23] R.B. Morgan, A restarted GMRES method augmented with eigenvectors, SIAM J. Matrix Anal. Appl. 16 (1995) 1154–1171.
[24] R.B. Morgan, On restarting the Arnoldi method for large nonsymmetric eigenvalue problems, Math. Comp. 65 (1996) 1213–1230.
[25] R.B. Morgan, Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations, SIAM J. Matrix Anal. Appl. 21 (2000) 1112–1135.
[26] R.B. Morgan, GMRES with deflated restarting, SIAM J. Sci. Comput. 24 (2002) 20–37.
[27] R.B. Morgan, W. Wilcox, Deflation of eigenvalues for GMRES in lattice QCD, Nucl. Phys. B (Proc. Suppl.) 106 (2002) 1067–1069.
[28] R.B. Morgan, M. Zeng, Harmonic projection methods for large non-symmetric eigenvalue problems, Numer. Linear Algebra Appl. 5 (1998) 33–55.
[29] R.B. Morgan, M. Zeng, Harmonic restarted Arnoldi for calculating eigenvalues and determining multiplicity, Preprint, 2003.
[30] A.A. Nikishin, Y.Y. Yeremin, Variable block CG algorithms for solving large sparse symmetric positive definite linear systems on parallel computers, I: General iterative scheme, SIAM J. Matrix Anal. Appl. 16 (1995) 1135–1153.
[31] D.P. O'Leary, The block conjugate gradient algorithm and related methods, Linear Algebra Appl. 29 (1980) 293–322.
[32] C.C. Paige, B.N. Parlett, H.A. van der Vorst, Approximate solutions and eigenvalue bounds from Krylov subspaces, Numer. Linear Algebra Appl. 2 (1995) 115–133.
[33] B.N. Parlett, A new look at the Lanczos algorithm for solving symmetric systems of linear equations, Linear Algebra Appl. 29 (1980) 323–346.

[34] B.N. Parlett, The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[35] A. Ruhe, The block Lanczos method for computing eigenvalues, Math. Comp. 33 (1979) 680–687.
[36] Y. Saad, On the Lanczos method for solving symmetric linear systems with several right-hand sides, Math. Comp. 48 (1987) 651–662.
[37] Y. Saad, Numerical Methods for Large Eigenvalue Problems, Halsted Press, New York, 1992.
[38] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS Publishing, Boston, MA, 1996.
[39] Y. Saad, Analysis of augmented Krylov subspace techniques, SIAM J. Matrix Anal. Appl. 18 (1997) 435–449.
[40] Y. Saad, M.H. Schultz, GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 7 (1986) 856–869.
[41] Y. Saad, M.C. Yeung, J. Erhel, F. Guyomarc'h, A deflated version of the conjugate gradient algorithm, SIAM J. Sci. Comput. 21 (2000) 1909–1926.
[42] V. Simoncini, E. Gallopoulos, An iterative method for nonsymmetric systems with multiple right-hand sides, SIAM J. Sci. Comput. 16 (1995) 917–933.
[43] C. Smith, A. Peterson, R. Mittra, A conjugate gradient algorithm for the treatment of multiple incident electromagnetic fields, IEEE Trans. Antennas Propag. 37 (1989) 1490–1493.
[44] D.C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, SIAM J. Matrix Anal. Appl. 13 (1992) 357–385.
[45] G.W. Stewart, A Krylov–Schur algorithm for large eigenproblems, SIAM J. Matrix Anal. Appl. 23 (2001) 601–614.
[46] H.A. van der Vorst, An iterative method for solving $f(A)x = b$ using Krylov subspace information obtained for the symmetric positive definite matrix A, J. Comput. Appl. Math. 18 (1987) 249–263.
[47] H.A. van der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems, SIAM J. Sci. Statist. Comput. 12 (1992) 631–644.
[48] H.A. van der Vorst, C. Vuik, The superlinear convergence behaviour of GMRES, J. Comput. Appl. Math. 48 (1993) 327–341.
[49] K. Wu, H. Simon, Thick-restart Lanczos method for symmetric eigenvalue problems, SIAM J. Matrix Anal. Appl. 22 (2000) 602–616.