Fall 2013

Methods For Increasing Domains Of Convergence In Iterative Linear System Solvers

David Michael Imberti
Purdue University

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By David Michael Imberti

Entitled
METHODS FOR INCREASING DOMAINS OF CONVERGENCE IN ITERATIVE LINEAR SYSTEM SOLVERS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Ahmed Sameh
Chair

Jianlin Xia

Zhiqiang Cai

Bradley Lucier

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Ahmed Sameh

Jianlin Xia

Approved by: David Goldberg 11/25/2013
Head of the Graduate Program Date

METHODS FOR INCREASING DOMAINS OF CONVERGENCE IN ITERATIVE LINEAR SYSTEM SOLVERS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

David M. Imberti

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2013

Purdue University

West Lafayette, Indiana

To Pascal's tooth

ACKNOWLEDGMENTS

There are many people to pay tribute here. First and foremost, to both Professor Sameh and Xia, I felt like I received quite a balance (the way of the iterative solver and the way of the direct solver). The experience in parallel coding under both situations was in the end worthwhile, and I appreciate having access to the machines necessary to perform such arts.

To Yuanzhe Xi and Shuhao Cao, thanks for letting me lounge in your office and gripe, and thanks for griping back. The world can only go up for both of you. And to Shen Wang, Faisal Saied, Stephen Cauley, and Yao Zhu: each of your codes are a work of art and thanks for dealing with my many pestering questions regarding them.

To Jacob Boswell, Jake Noparstak, Nicholas Stull, and Gabriel Sosa thanks for the evening gaming.

To Jake Noparstak: don't go insane.

To Nicolas Stull: don't go sane.

To Jeffrey Zylinski: may your future be gilded.

To Brad Lucier, Greg Buzzard, Steven Bell, my uncle Ken McGovern, and Chuck thank you for the counsel.

To Joe Kohlhaas, thank you for the long, continuing, years-old discussion.

To my grandmother Elsa, thank you for the many, hours-long discussions.

To my grandfather Mike, thank you the periodic, completely variable tirades.

To Travis Ball, in a way, I wouldn't have been here were it not for you.

To Paul Kepley: peanuts.

To Mark Pentigore: your undergraduate work on tensors was more inspiring than you think.

To Scott Nystrom, who included me in his acknowledgements, I would have not gone through the trouble of writing these acknowledgements had it not have been for the fact that I felt I should at least repay the favor. Thank you, and the letters from Cordell Hull to Argentina were very interesting to hear about. Although, some slack regarding Argentine performance in football would have been nice.

And Kyle Kloster you will go places. You will go to big places.

To James Vogel, Difeng Cai, Xiao Liu, and Zixing Xin: I only wish I had more time to get to know you all.

But of course, it would be odd to not include one's parents. To my father Henry Imberti for his common sense, but also to my mother Marie McGovern for her lack of it.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Imberti, David M. Ph.D., Purdue University, December 2013. Methods for Increasing Domains of Convergence in Iterative Linear System Solvers. Major Professors: Ahmed Sameh and Jianlin Xia.

In this thesis, we introduce and improve various methods for increasing the domains of convergence for iterative linear system solvers. We rely on the following three approaches: making the iteration adaptive, or nesting an inner iteration inside of a previously determined outer iteration; using deflation and projections to manipulate the spectra inherent to the iteration; and/or focusing on reordering schemes. We will analyze a specific combination of these three strategies. In particular, we propose to examine the influence of nesting a Flexible Generalized Minimum Residual algorithm together with an inner Recursive Projection Method using a banded preconditioner resulting from the Fiedler reordering.

1. INTRODUCTION

The performance of classical iterative schemes for solving sparse linear systems is highly dependent on the spectra of the respective iteration matrices. Indeed, most theoretical results for improving convergence, acceleration, or other influences on an iteration necessitates pre-existing bounds of the spectral radius for the iteration matrix to be properly analyzed. Furthermore, methods for improving previously existing iteration methods typically rely on the following three approaches: making the iteration adaptive, or nesting an inner iteration inside of a previously determined outer iteration; using deflation and projections to manipulate the spectra inherent to the iteration; and/or focusing on reordering the coefficient matrix. We will analyze a specific combination of these three strategies. In particular, we propose to examine the influence of nesting a Flexible Generalized Minimum Residual (FGMRES) algorithm [49] together with an inner Recursive Projection Method (RPM) [7] using a banded preconditioner resulting from the Fiedler reordering [17, 35, 38].

The strategy of nesting to improve performance of the outer algorithm is certainly not new [2, 3, 8, 18, 22, 24, 37, 42, 49, 59]. In particular, using the GMRES algorithm as the outer iteration is also a popular choice, due to its superlinear convergence and robustness [21]. Not even combining the GMRES algorithm with a Richardson-like algorithm is new, as it is observed that using a Richardson-like scheme as the inner iteration speeds initial convergence, while it is observed that we still maintain a superlinearity property [3]. What we aim to do is to combine FGMRES and RPM in order to use deflation in the inner step to improve the algorithm overall.

Further, the idea of using some form of deflation together with some form of GMRES is not new [6, 8, 13–15, 43, 49, 50, 52, 59]. These algorithms focus on using deflation directly with GMRES or the Conjugate Gradient method (CG). Thus in the case of GMRES deflation tends to focus on removing "The smallest eigenvalues of A

which are known to slow down the convergence of GMRES" [13]. Here we are looking at an inner-outer iteration which incorporates deflation, more in the line of [59]. In particular, because our deflation occurs in the inner-step by use of RPM, and RPM is essentially a deflated Richardson iteration; we are concerned with deflating out the *largest* eigenvalues of A (as opposed to the proposal in [13]). As will be shown in chapter 8, this improves the degree of positive definiteness of A, and thereby speeds up and can possibly guarantee convergence of the outer FGMRES step.

That we cast our outer step as FGMRES instead of GMRES remains technically accurate. As during each step of the outer iteration in FGMRES causes a different initial vector to be passed to the inner step of RPM. Although theoretically this results in a GMRES-RPM scheme, numerically the calculation of the projectors in RPM depends on the initial vector, and therefore for reasons of stability this requires us to couch our analysis in terms of FGMRES-RPM. However, as we will show in chapter 5, such perturbations will still result in a convergent algorithm. Moreover, and of theoretical importance, the results we obtain for on FGMRES are useful in their own right beyond the results in [8, 49]. We show the relationship between FGMRES and GMRES, and lay the foundation for the relationship between the convergence behavior FGMRES and the geometric mean of convergence behavior of a collection of individual preconditioned GMRES algorithms.

We use RPM in the inner step in order to assume fast initial convergence in the residual norm of a Richardson iteration. More importantly, this necessitates a theoretical analysis of RPM, which we present in chapter 2. Although the majority of these results are somewhat reflected in [7], we improve on these results by deriving an explicit preconditioner expression of RPM. The use of this preconditioner expression allows us to further examine the convergence rate behavior of RPM directly, and permits the analysis of the convergence criteria needed for the inner-outer FGMRES-RPM method presented in chapter 8.

Furthermore, the utility of RPM is directly tied to its underlying preconditioner utilized in RPM (as opposed to the preconditioner expression *of* RPM referred

to in the previous paragraph). For this purpose, we use a banded preconditioner, which, in turn, necessitates a theoretical analysis of banded preconditioners presented in chapter 6. We will show by use of Stein's theorem [34] that convergence using such preconditioners is dependent on the relative size of the extracted central band to the rest of the matrix. As such, the utility of banded preconditioner is inherently tied to the sparse matrix reordering scheme we use.

In order to maximize the size in terms of norm of the banded preconditioner and minimize the norm and the rank of the matrix outside the band, we propose using Fiedler (or spectral) reordering. This, likewise, necessitates a theoretical analysis of Fiedler reordering presented in chapter 7. We review the literature and heuristical results behind using Fiedler reordering in order to concentrate the heaviest elements of the matrix within a central band. This results in introducing the importance of the Hadamard product of A with itself, and the analysis of the Fiedler reordering using the Hadamard product of A with itself as opposed to using the Fiedler reorering using A itself.

In summary, there are a number of new theoretical results generated in this dissertation. For RPM we generate a preconditioner expression and convergence rate results, for FGMRES we produce new convergence results relating FGMRES to certain characteristics of the underlying GMRES algorithm, and in addition, we propose a convergence result for the resultant nested iteration. For the weighted spectral reordering we propose a modified approach involving the Hadamard product in order to improve the effectiveness of the banded preconditioner coincident with a newly developed convergence criteria for banded preconditioning. With all these results established, we can analyze the convergence behavior of the entire algorithm.

The rest of the dissertation is organized as follows. In chapter 2 we review basic facts on stabilization methods, which we will then use in chapter 3 to enhance results pertaining to RPM. In chapter 4 we review basic facts regarding the convergence of GMRES, which we will then use in chapter 5 to enhance results pertaining to the convergence of FGMRES. In chapter 6 we analyze generalized diagonal dominance

criteria necessary for the choice of the banded preconditioner to be used for RPM. In chapter 7 we review and propose a modified Fiedler reordering algorithm in order to improve the diagonal dominance criteria formed in chapter 6. Finally in chapter 8 we discuss the theoretical properties resulting from nesting FGMRES with RPM, propose the overall algorithm in its totality and conclude with numerical experiments before giving a final summary in chapter 9.

2. FIXED POINT ITERATIVE PROCEDURES

Consider a fixed-point iteration of the form

$$u^{(\nu+1)} = F(u^{(\nu)}, \lambda) \tag{2.1}$$

where $F : \Re^N \times \Re \to \Re^N$ is smooth, and $N > 1$.

Fixed-point iterations are often used to approximate solutions for nonlinear problems. Here, we will borrow properties of this approach to aid in solving linear systems. As such, we will first analyze the general abstract convergence properties of fixed point iterative systems, and then succesively apply them to our projection-based algorithm in later chapters.

We say that the above iteration has a solution in a given interval if there is some $\{u^{(\nu)}(\lambda)\} \to u^*(\lambda)$ where

$$u^*(\lambda) = F(u^*(\lambda), \lambda) \tag{2.2}$$

for some $\lambda \in [\lambda_a, \lambda_b]$ [10, 19, 56].

We first note the following

THEOREM 1 (Convergence Criteria for Fixed Point Iteration) *[26,31,33, 36]Equation (2.1) converges locally in a neighborhood of a solution if the spectra of the Jacobian matrix of F ($J := F_u(u^*(\lambda), \lambda)$) lie within the unit disk.*

Proof Let $||\cdot||$ be a norm so that $||J||$ is within a neighborhood of $\rho(J)$ [34].

Since F is smooth, then $\exists r(u)$ with $||r(u)|| < \epsilon$ for $|u - u^*| < \delta$, and $F(u) = F(u^*) + J(u - u^*) + r(u)$ where J is the Jacobian [5] [47].

We then show that $||F^k(u) - F(u^*)|| \leq ||J||^k\delta + \Sigma_{i=0}^k||J||^i\epsilon$ by induction.

The base case is straightforward:

$$||F(u) - F(u^*)|| \leq ||J(u - u^*) + r(u)|| \leq ||J||\delta + \epsilon \tag{2.3}$$

As for the inductive case:

$$\begin{aligned} ||F^k(u) - F(u^*)|| &\leq ||J(F^{k-1}(u)) - u^*) + r(F^{k-1}(u))|| \\ &\leq ||J||||F^{k-1}(u) - F(u^*)|| + ||r(F^{k-1}(u))|| \end{aligned} \tag{2.4}$$

Then by inductive hypothesis $||F^{k-1}(u) - u^*|| \leq ||J||^{k-1}\delta + \Sigma_{i=0}^{k}||J||^i\epsilon$, so if we choose $\epsilon < (\Sigma_{i=0}^{\infty}||J||^i)^{-1}$ (this is well-defined since $\rho(J) < 1$ (since $||J||$ is arbitrarily close to $\rho(J)$ via our choice of norm)) [23] then this gives $||F^{k-1}(u) - u^*|| < \delta$; therefore, $||r(F^{k-1}(u))|| < \epsilon$ by the smoothness of F. Thus using the inductive hypothesis again, then:

$$||F^k(u) - F(u^*)|| \leq ||J||(||J||^{k-1}\delta + \Sigma_{i=0}^{k-1}||J||^i\epsilon) + \epsilon \tag{2.5}$$

Again, due to the choice of ϵ:

$$||F^k(u) - F(u^*)|| \leq ||J||^k\delta + \Sigma_{i=0}^{k}||J||^i\epsilon \tag{2.6}$$

Since $||J||^k\delta \to 0$ as $k \to 0$ and $\sigma_{i=0}^{k}||J||^i\epsilon \to C\epsilon$ given that $||J|| < 1$ [23], then we have convergence within a neighborhood of the solution. ■

This scheme, in general, fails if the spectra of the Jacobian matrix lies outside the unit disk (as a trivial example, consider $F(x) = 2x$).

Therefore, in order to stabilize the procedure, we first decompose the space into $\mathbb{P}$ and $\mathbb{Q}$, where $\mathbb{P}$ is the invariant eigenspace of J corresponding to the eigenvalues of magnitude $> 1 - \delta$, and $\mathbb{Q}$ is orthogonal to $\mathbb{P}$ (it will be theoretically important later to note that $\mathbb{Q}$ is not necessarily an invariant subspace) [23, 56].

Let the corresponding projectors for $\mathbb{P}$ and $\mathbb{Q}$ be P and Q, respectively. Then note that $PQ = 0, Q = I - P$, which we will use implicitly (we will delay a discussion on how these projectors are obtained until the next chapter).

In order to describe the stablization procedure, we define:

$$\begin{aligned} p = f &:= PF \\ q = g &:= QF \end{aligned} \tag{2.7}$$

With this, the central concept is to use the subspace decomposition to improve the general procedure by applying a modified chord method on the system corresponding to the $\mathbb{P}$ eigenspace. This leads to the following scheme:

$$\begin{aligned} (I - f_p^{(0)})(p^{(k+1)} - p^{(k)}) &= f(p^{(k)}, q^{(k)}, \lambda) - p^{(k)} \\ q^{(k+1)} &= g(p^{(k)}, q^{(k)}, \lambda) \end{aligned} \tag{2.8}$$

(where f_p is the derivative of f with respect to the subspace $\mathbb{P}$)

In summary so far then, the stabilized iteration consists of

ALGORITHM 1 (Stabilized Iteration) $p^{(0)} := Pu^{(0)}(\lambda), q^{(0)} := Qu^{(0)}(\lambda)$

Do until convergence:

$$\begin{aligned} p^{(k+1)} &= p^{(k)} + (I - f_p^{(0)})^{-1}(f(p^{(k)}, q^{(k)}, \lambda) - p^{(k)}) \\ q^{(k+1)} &= g(p^{(k)}, q^{(k)}, \lambda) \end{aligned} \tag{2.9}$$

$u^*(\lambda) = p^{(k_{final})} + q^{(k_{final})}$

(here we assume that 1 is not an eigenvalue of the Jacobian of F so that the inversion is well-defined)

Now that this description is complete, we can provide some basic convergence results for this algorithm.

THEOREM 2 (Stabilized Iteration Convergence Theorem) *Let F be smooth and 1 not be an eigenvalue of the Jacobian of F, then algorithm 1 above converges for all initial values $u^{(0)} \in B_\epsilon(u^*)$ for some ϵ*

[56]

Proof We define

$$v^{(k)} := \begin{pmatrix} p^{(k)} \\ q^{(k)} \end{pmatrix} \tag{2.10}$$

Then as in the previous convergence proof, by Taylor's theorem:

$$v^{(k+1)} - v^* = J(v^{(k)} - v^*) + O(||v^{(k)} - v^*||^2) \tag{2.11}$$

Where J is the Jacobian of the stabilized iteration and, again, similar to the previous convergence proof, $||\cdot||$ is a norm for which $||J|| - \rho(J)$ is arbitrarily small [34]).

Thus we need only show that the Jacobian has a spectral radius less than one, from which the rest of the proof follows by application of the previous theorem 1.

By direct calculation:

$$\begin{aligned} J &= \begin{pmatrix} \frac{\partial(p^* + (I - f_p^{(0)})^{-1}(f^* - I))}{\partial p^*} & \frac{\partial(p^* + (I - f_p^{(0)})^{-1}(f^* - I))}{\partial q^*} \\ \frac{g^*}{\partial p^*} & \frac{g^*}{\partial q^*} \end{pmatrix} \\ &= \begin{pmatrix} I + (I - f_p^*)^{-1}(f_p^* - I) & (I - f_p^*)^{-1} f_q^* \\ g_p^* & g_q^* \end{pmatrix} \end{aligned} \tag{2.12}$$

Note that $g_p^* = QJP$, since $\mathbb{P}$ is an invariant space, $JP \in \mathbb{P}$, but $\mathbb{Q}$ and $\mathbb{P}$ are orthogonal, thus $g_p^* = 0$.

It remains to show that $g_q^* = QJQ$ has spectral radius less than one. Using Jordan Canonical Form, there exists a similarity matrix $W = (W_1, W_2)$ so that:

$$J = W \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} W^{-1} \tag{2.13}$$

where J_1 contains all the Jordan blocks associated to eigenvalues of norm < 1 and J_2 contains all Jordan blocks associated to eigenvalues of norm > 1 (there are

no eigenvalues = 1 by hypothesis). If we do this, note that the range of W_1 is $\mathbb{P}$, and the range of W_2 is $\mathbb{Q}$; therefore, $QW_1 = 0$ and:

$$QW_2J_2 = QJW_2 = QJ(PW_2 + QW_2) = QJQW_2 \tag{2.14}$$

Thus, if we let $V = (W_1, QW_2)$, then by a straightforward computation we have:

$$QJQV = V \begin{pmatrix} 0 & 0 \\ 0 & J_2 \end{pmatrix} \tag{2.15}$$

Hence, if we can show that V is nonsingular, then $g_q^* = QJQ$ is similar to a matrix with spectral radius less than one. But if V were singular, then for $w \neq 0$ either $QW_2w = 0$, which can't happen because if the range of W_2 is $\mathbb{Q}$, and Q surjects onto $\mathbb{Q}$, then this would imply that W_2 would be linearly dependent, which would imply that W is singular; or W_2w is in the range of W_1, which also can't happen because W is nonsingular. Thus V is nonsingular.

Therefore, g_q^* has spectral radius less than one, and J itself has spectral radius less than one. And since the Jacobian has spectral radius less than one, the stabilized iteration converges by theorem 1.

■

3. RECURSIVE PROJECTION METHOD

The Recursive Projection Method attempts to find the fixed points of

$$y = F(y, \lambda) : \Re^N \times \Re \to \Re^N \tag{3.1}$$

via deflating appropriate subspaces.

In particular, we apply the analysis of the previous chapter to a Richardson scheme:

$$y^{(k+1)} = M^{-1}Ny^{(k)} + M^{-1}b \tag{3.2}$$

In applying deflation to the above iteration, the fixed iteration functional is given by

$$F(y) = M^{-1}Ny + M^{-1}b \tag{3.3}$$

Applying this functional to the previous analysis yields:

$$\begin{gathered} p := Py \in \mathbb{P} \\ q := Qy \in \mathbb{Q} \\ f(p, q, \lambda) = PF(p + q, \lambda) \\ g(p, q, \lambda) = QF(p + q, \lambda) \end{gathered} \tag{3.4}$$

$$(I - PF_y(y^{(0)}, \lambda)P)(p^{(k+1)} - p^{(k)}) = f(p^{(k)}, q^{(k)}, \lambda) - p^{(k)} \tag{3.5}$$

$$q^{(k+1)} = g(p^{(k)}, q^{(k)}, \lambda) \tag{3.6}$$

Equation (3.5) is simplified to:

$$(I - PHP)(p^{(k+1)} - p^{(k)}) = P(Hy^{(k)} - M^{-1}b) - p^{(k)} \tag{3.7}$$

Let Z be an orthogonal basis for $\mathbb{P}$, then $P = ZZ^T, I = Z^TZ$:

$$\begin{aligned}
(I - ZZ^THZZ^T)(p^{(k+1)} - p^{(k)}) &= ZZ^T(Hy^{(k)} - M^{-1}b) - p^{(k)} \\
(I - ZZ^THZZ^T)p^{(k+1)} &= ZZ^T(Hy^{(k)} - M^{-1}b) - ZZ^THPp^{(k)} \\
(I - ZZ^THZZ^T)p^{(k+1)} &= ZZ^T(Hy^{(k)} - Hp^{(k)} - M^{-1}b) \\
(I - ZZ^THZZ^T)p^{(k+1)} &= ZZ^T(Hq^{(k)} - M^{-1}b)
\end{aligned} \tag{3.8}$$

If this equation holds, then if we multiply through by Z^T we obtain:

$$\begin{aligned}
(Z^T - Z^THZZ^T)p^{(k+1)} &= Z^T(Hq^{(k)} - M^{-1}b) \\
(I - Z^THZ)Z^Ty^{(k+1)} &= Z^T(Hq^{(k)} - M^{-1}b)
\end{aligned} \tag{3.9}$$

In order to simplify this expression and later implementation, we define $u =: Z^Ty$.

$$(I - Z^THZ)u^{(k+1)} = Z^T(Hq^{(k)} - M^{-1}b) \tag{3.10}$$

Furthermore, equation (3.6) can also be simplified as follows,

$$\begin{aligned}
q^{(k+1)} &= g(p^{(k)}, q^{(k)}, \lambda) \\
q^{(k+1)} &= Q(Hy^{(k)} + M^{-1}b)
\end{aligned} \tag{3.11}$$

Using our previous notation, we get

$$q^{(k+1)} = Q(Hq^{(k)} + HZu^{(k)} + M^{-1}b) \tag{3.12}$$

In summary, we obtain the following iteration:

$$\begin{aligned}
(I - Z^THZ)u^{(k+1)} &= Z^T(Hq^{(k)} - M^{-1}b) \\
q^{(k+1)} &= Q(Hq^{(k)} + HZu^{(k)} + M^{-1}b)
\end{aligned} \tag{3.13}$$

(where we assume that $(I_r - Z^THZ)$ is nonsingular)

In order to generalize our analysis further, we introduce i, j components into the previous general RPM analysis and allow greater variability over the Jacobians:

ALGORITHM 2 (RPM Iteration)

$$\begin{aligned} (I - Z^T H Z)u^{(k+1)} &= Z^T(Hq^{(i)} - M^{-1}b) \\ q^{(k+1)} &= Q(Hq^{(k)} + HZu^{(j)} + M^{-1}b) \end{aligned} \tag{3.14}$$

This is what we term the coupling factor (as defined in [7]). This is important insofar as its influence on the Jacobian matrices associated upon each iteration.

i	j	coupling
k	k	Jacobi
k	k+1	Gauss-Seidel (GS)
k+1	k	Reverse Gauss-Seidel (RGS)

The only item needed for the description of RPM algorithm above to be complete is a description of how the projectors P, Q are calculated. We do this in particular for the Jacobi coupling.

Note that

$$\begin{aligned} q^{(k+1)} &= Q(M^{-1}b + Hq^{(k)}) \\ q^{(k+1)} - q^{(k)} &= (QHQ)(q^{(k)} - q^{(k-1)}) \end{aligned} \tag{3.15}$$

Therefore, we can use the power method on the successive q vectors to progressively obtain the Q projector ([11, 23, 58, 63]). The above equation can be used to approximate the dominant eigenspace of QHQ by computing a small window of $q^{(k+1)} - q^{(k)}$ for $k = j - wind + 1, \cdots, j$, computing an orthonormal basis S of this space, and then using the Schur vectors T (i.e., the columns of the orthonormal matrix of the Schur decomposition, which are needed to ensure that P is an *invariant* subspace) of the dominant eigenspace $S^T HS$ so that ST approximates the Schur vectors of H [7].

However, in order to utilize parallelism, instead of using $q^{(k+1)} - q^{(k)}$ so that we may apply the power method to extract the necessary corresponding eigenspace, we

suggest using a block of vectors so that we can use a subspace iteration (performed similarly in [11, 43, 52]). This means that instead of needing the *wind* parameter to obtain the corresponding eigenspace, one can use the block of vectors directly. Further, all corresponding u vectors will also be block $n \times m$ matrices. And one can use Hessenberg reduction followed by QR iterations to calculate the corresponding Schur vectors [11, 23, 58].

The only other parameters in the algorithm left to describe as in [7], is the maximal number of deflated eigenvalues (which we denote as *numeig*), and the number of eigenvalues deflated at each iteration (which we denote by def).

In total, then, this yields the following algorithm.

ALGORITHM 3 (Subspace Iteration RPM) *Choose some random $n \times m$ block of initial linearly independent vectors Y [48].*

Let $A = M - N, H = M^{-1}N$, and choose C to be a $n \times m$ matrix with each column $= M^{-1}b$

do k=0:freq-1

$Y^{(k+1)} = C + HY^{(k)}$

$\Delta = Y^{(k+1)} - Y^{(k)}$

enddo

$Z = \varnothing$

$u^{(0)} = 0$

$q^{(0)} = Y^{(0)} - Zu^{(0)}$

$T^{(0)} = C + Hq^{(0)}$

$k = 0$

while not converged

if $size(Z, 2) < numeig$ and $mod(k, freq) = 0$

Orthogonalize Δ

Perform QR iterations on $S^T HS$ to obtain def schur vectors T

$Z1 = ST$

$Z = (Z, Z1)$

Orthogonalize Z

$W = I - Z^T H Z$

endif

$q^{(k+1)} = (I - ZZ^T)(T^{(k)} + (HZ)u^{(k)})$

$T^{(k+1)} = C + Hq^{(k+1)}$

$u^{(k+1)} = W^{-1}(Z^T T^{(k+1)})$

$\Delta = q^{(k+1)} - q^{(k)}$

$Y^{(k+1)} = Zu^{(k+1)} + q^{(k+1)}$

$k = k + 1$

endwhile

Extract the first column of $Y^{(k+1)}$

[7]

Furthermore, we can simplify the expression of the Jacobian for RPM, which will be useful in later convergence analysis.

THEOREM 3 (The Jacobian of RPM) *By denoting the error vector as the appropriate difference in both projection* $\mathbb{P}$ *and* $\mathbb{Q}$*, respectively, as*

$$e^{(k)} = (p^{(k)T} - p, q^{(k)T} - q^T)^T \tag{3.16}$$

Each of the three couplings' iterations can be expressed as $Je^{(k)} = e^{(k+1)}$ *[7] where:*

$$J_J = \begin{pmatrix} 0 & C \\ E & B \end{pmatrix} \tag{3.17}$$

$$J_{GS} = \begin{pmatrix} 0 & C \\ 0 & EC + B \end{pmatrix} \tag{3.18}$$

$$J_{RGS} = \begin{pmatrix} CE & CB \\ E & B \end{pmatrix} \tag{3.19}$$

(note: the Jacobian for RGS is a correction on [7])

Where

$$E := QHP, B := QHQ, C := P(Z(I - Z^T HZ)^{-1}Z^T)PHQ \tag{3.20}$$

Proof We show the derivation of the Jacobi Coupling's Jacobian, which we split into two parts, the first to show that $C(q^{(k)} - q) = p^{(k+1)} - p$:

$$\begin{aligned} C(q^{(k)} - q) \quad = \quad & P(Z(I - Z^T HZ)^{-1}Z^T)PHQ(q^{(k)} - q) \\ & Z(I - Z^T HZ)^{-1}Z^T HQ(q^{(k)} - q) \\ & Z(I - Z^T HZ)^{-1}Z^T Hq^{(k)} - Z(I - Z^T HZ)^{-1}Z^T Hq \end{aligned} \tag{3.21}$$

Using the iteration equation $(I - Z^T HZ)u^{(k+1)} = Z^T(Hq^{(k)} - M^{-1}b)$ (from algorithm 2), then

$$\begin{aligned} C(q^{(k)} - q) \quad = \quad & Z(I - Z^T HZ)^{-1}(I - Z^T HZ)u^{(k+1)} \\ & +Z(I - Z^T HZ)^{-1}Z^T M^{-1}b - Z(I - Z^T HZ)^{-1}Z^T Hq \\ C(q^{(k)} - q) \quad = \quad & Zu^{(k+1)} - Z(I - Z^T HZ)^{-1}Z^T(-M^{-1}b + Hq) \end{aligned} \tag{3.22}$$

We note that the solution satisfies the equation in algorithm 2 exactly, that is $(I - Z^T HZ)u = Z^T(Hq - M^{-1}b)$:

$$\begin{aligned} &= Zu^{(k+1)} - Z(I - Z^T HZ)^{-1}(I - Z^T HZ)u \\ &= Zu^{(k+1)} - Zu \\ &= p^{(k+1)} - p \end{aligned} \tag{3.23}$$

And the second to show that $E(p^{(k)} - p) + B(q^{(k)} - q) = q^{(k+1)} - q$:

$$\begin{aligned} & E(p^{(k)} - p) + B(q^{(k)} - q) \\ = \quad & QHP(p^{(k)} - p) + QHQ(q^{(k)} - q) \\ = \quad & Q(Hp^{(k)} + Hq^{(k)}) - Q(Hp + q) \\ = \quad & Q(HZu^{(k)} + Hq^{(k)} + M^{-1}b) - Q(HZu + q + M^{-1}b) \end{aligned} \tag{3.24}$$

We note that the solution satisfies the equation in algorithm 2 exactly, that is that $q^{(k+1)} = Q(Hq^{(k)} + HZu^{(j)} + M^{-1}b)$ from algorithm 2:

$$\begin{aligned} & Q(HZu^{(k)} + Hq^{(k)} + M^{-1}b) - Q(HZu + q + M^{-1}b) \\ & = q^{(k+1)} - q \end{aligned} \tag{3.25}$$

A similar analysis applies to the other two couplings above.

For the Gauss-Seidel coupling, note that we have already shown that $p^{(k+1)} - p = C(q^{(k)} - q)$ in the above Jacobi Coupling case, since the (i) component is the same for the Jacobi coupling as the Gauss-Seidel.

Therefore, we need only show that $(EC + B)(q^{(k)} - q) = p^{(k+1)} - p$, the proof mimics the algebra for the Jacobi case with the following replacing $E(p^{(k)} + p) + B(q^{(k)} - q)$ in 3.24

$$EC(q^{(k)} - p) + B(q^{(k)} - q) \tag{3.26}$$

We use what we have already shown for the Gauss-Seidel case, namely that $p^{(k+1)} - p = C(q^{(k)} - q)$:

$$\begin{aligned} = \quad & QHP(p^{(k+1)} - p) + QHQ(q^{(k)} - q) \\ = \quad & Q(Hp^{(k+1)} + Hq^{(k)}) - Q(Hp + q) \\ = \quad & Q(HZu^{(k+1)} + Hq^{(k)} + M^{-1}b) - Q(HZu + q + M^{-1}b) \end{aligned} \tag{3.27}$$

We note that the solution satisfies the equation in algorithm 2 exactly because $j = k + 1$ in the Gauss-Seidel case:

$$\begin{aligned} & Q(HZu^{(k)} + Hq^{(k)} + M^{-1}b) - Q(HZu + q + M^{-1}b) \\ = \quad & q^{(k+1)} - q \end{aligned} \tag{3.28}$$

A similar approach works for the reverse Gauss-Seidel case. Note that since the (j) component is the same for the Jacobi coupling as the Reverse Gauss-Seidel that $E(p^{(k)} - p) + B(q^{(k)} - q) = q^{(k+1)} - q$.

All that remains to be shown is that $CE(p^{(k)} - p) + CB(q^{(k)} - q)$ which we already know is equal to $C(q^{(k+1)} - q) = p^{(k+1)} - p$. This follows nearly identical to the $C(q^{(k)} - q) = p^{(k+1)} - p$ case for the Jacobi coupling:

$$\begin{aligned} & P(Z(I - Z^T HZ)^{-1}Z^T)PHQ(q^{(k+1)} - q) \\ = \quad & Z(I - Z^T HZ)^{-1}Z^T HQ(q^{(k+1)} - q) \\ = \quad & Z(I - Z^T HZ)^{-1}Z^T Hq^{(k+1)} - Z(I - Z^T HZ)^{-1}Z^T Hq \end{aligned} \tag{3.29}$$

Using the iteration from algorithm 2 again, $(I - Z^T HZ)u^{(k+1)} = Z^T(Hq^{(k+1)} - M^{-1}b)$:

$$\begin{aligned} & Z(I - Z^T HZ)^{-1}(I - Z^T HZ)u^{(k+1)} + Z(I - Z^T HZ)^{-1}Z^T M^{-1}b \\ & \qquad -Z(I - Z^T HZ)^{-1}Z^T Hq \\ = \quad & Zu^{(k+1)} - Z(I - Z^T HZ)^{-1}Z^T(-M^{-1}b + Hq) \end{aligned} \tag{3.30}$$

We note that the solution satisfies the equation in the iteration exactly:

$$\begin{aligned} & Zu^{(k+1)} - Z(I - Z^T HZ)^{-1}(I - Z^T HZ)u \\ = \quad & Zu^{(k+1)} - Zu \\ = \quad & p^{(k+1)} - p \end{aligned} \tag{3.31}$$

Which shows that the above Jacobians are correct.

■

Note that throughout the rest of the analysis, since $E = QHP = 0$, these Jacobians have the same spectra. Therefore, nothing changes in the following analysis given different couplings.

In particular, this allows us to state and apply the result of theorem 2 from the previous chapter that given 1 not an eigenvalue of PHP and $\rho(QHQ) < 1$, the

above generalized coupling scheme converges. Since by using the Jacobian formulation above, where $E = QHP = 0$ by choice of $\mathbb{P}$, we have $\rho(J) = \rho(B) = \rho(QHQ)$.

Moreover, this allows us to state an important motivating property regarding this method, namely, with $\mathbb{P}, \mathbb{Q}$ properly chosen then either this method can take a divergent splitting and force it to be convergent (by projecting that part of the eigenspace for which $|\lambda(H)| > 1$ onto $\mathbb{P}$) or accelerate convergence (since either all eigenvalues are < 1, in which case the method reduces to the original splitting) [7].

We illustrate the above remarks by the following examples.

Example 3.0.1 *RPM Convergence and Divergence*

First we show that using RPM to deflate larger eigenvalues (in modulus) does indeed help speed up convergence. We do this in figure 1 below showing the residual norm vs. iterations with the typical toy matrix setup, a Poisson matrix of size 100, with the preconditioner being a simple band with bandwidth 21. The maximum number of eigenvalues to be deflated is 4, the subspace size is 4, the frequency of deflation is 1, and the number of eigenvalues deflated at each step is 2. The blue line is RPM deflating 1 eigenvalue, and the green is RPM deflating 4 eigenvalues.

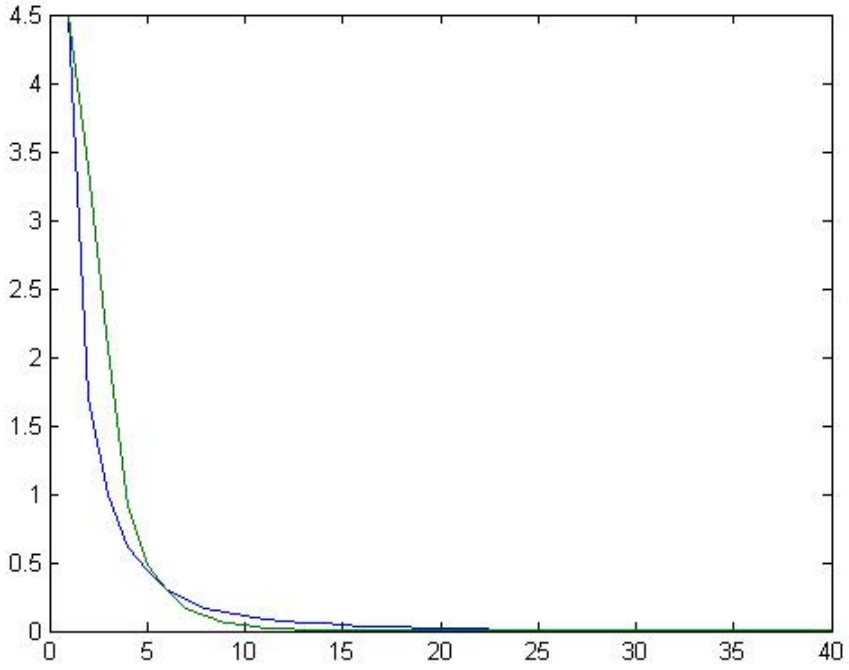

Figure 1. Deflation speeding up convergence

It takes a few iterations for the convergence bound to ensure the speed-up of RPM over traditional Richardson iteration. The following shows a blow-up of convergence upon further iterations:

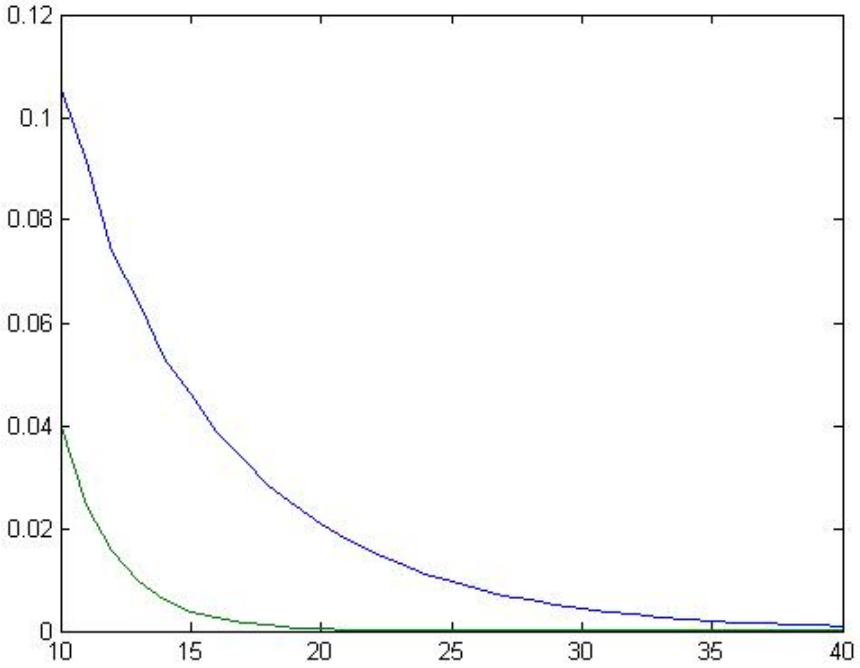

Figure 2. Deflation speeding up convergence blowup

Further, this does not solely speed up convergence, but also can force divergent iterations to become convergent. We show this by keeping all the parameters the same, except we change the problem into a Helmholtz problem by decreasing the value of the diagonal entries down from 4 to 3.6. Here, the blue line is RPM with only 1 eigenvalue deflated, and the green is RPM with the 4 eigenvalues deflated:

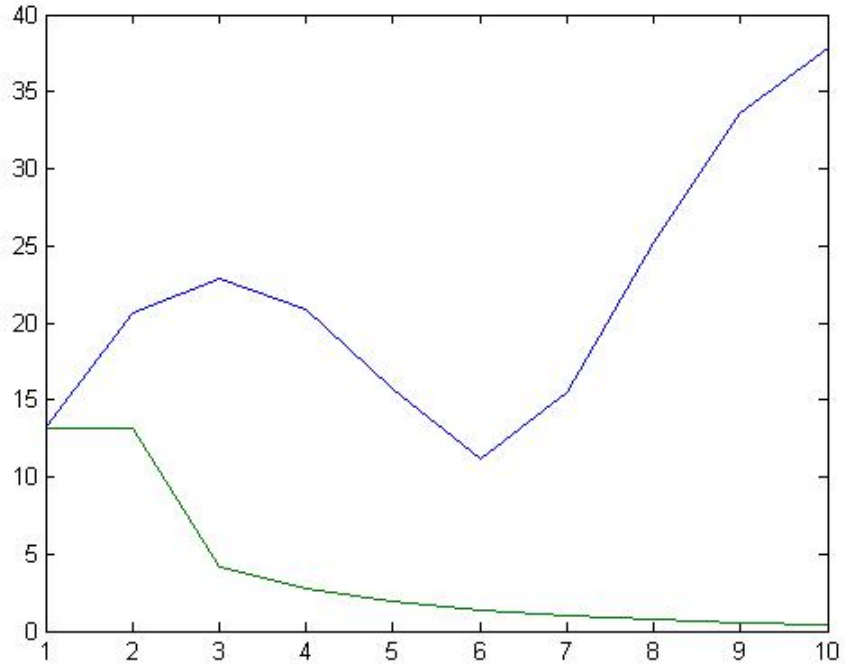

Figure 3. Deflation preventing divergence

Finally, the following not only also illustrates the speed-up upon deflating more eigenvalues, but also demonstrates algorithm 3 and the claim of parallelism underlying the algorithm. The following system has dimension 4194304 with approximately 400 million nonzeros running on 8 cores. Here, the blue line is block Jacobi RPM with 2 eigenvalues deflated, and the green is block Jacobi RPM with 4 eigenvalues deflated.

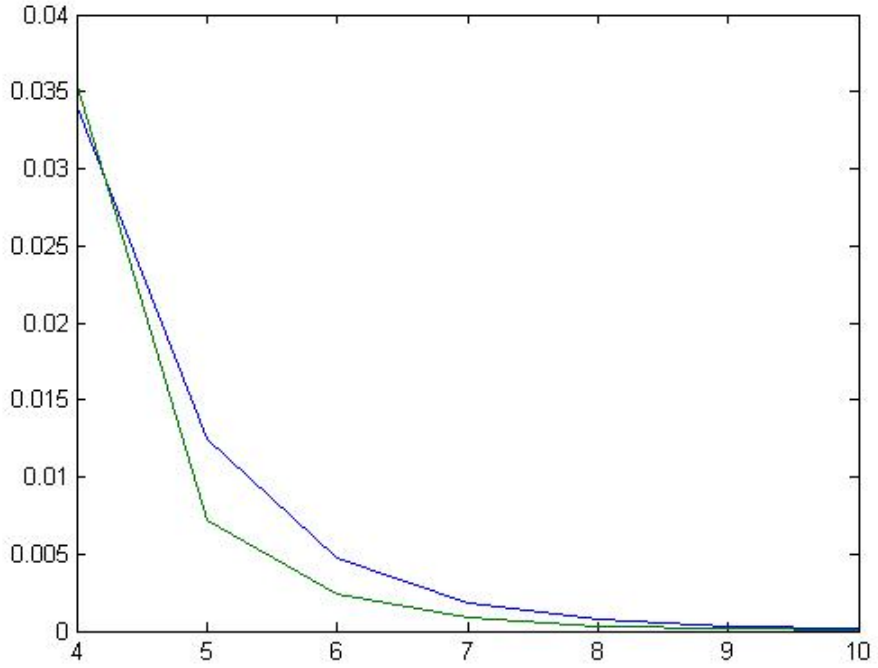

Figure 4. Parallel deflation speeding up convergence

However, the more important consequence of the previous Jacobian expressions, is that it allows us to explicitly state RPM as a preconditioner.

THEOREM 4 (RPM Preconditioner) *k steps of RPM is equivalent to preconditioning the original system by:*

$$M^{-1} = (I, I)J^k \begin{pmatrix} P \\ Q \end{pmatrix} + (I, I)(I - J^k) \begin{pmatrix} P \\ Q \end{pmatrix} A^{-1} \tag{3.32}$$

Assuming that J has no eigenvalues ≥ 1.

Proof The RPM iteration can be expressed via $Je_k = e_{k+1}$, where $e_k = \begin{pmatrix} P(x_k - x) \\ Q(x_k - x) \end{pmatrix}$ as in [7] and discussed above, thus:

$$v := \begin{pmatrix} P \\ Q \end{pmatrix} x, v_k := \begin{pmatrix} P \\ Q \end{pmatrix} x_k, e_k = v_k - v \tag{3.33}$$

$$Jv_k + (v - Jv) = v_{k+1} \tag{3.34}$$

Note that the actual error vector is obtained by premultiplying (3.34) by (I, I) (since $(I, I)\begin{pmatrix} P \\ Q \end{pmatrix} = P+Q = I$). Thus, after premultiplying and telescoping (3.34), k steps of RPM is equivalent to:

$$(I, I)(J^k v_0 + \Sigma_{i=0}^{k-1} J^i (I - J)v) \tag{3.35}$$

Therefore, in solving a system $Ax = b$, with $v_0 = \begin{pmatrix} P \\ Q \end{pmatrix} b$, and $v = \begin{pmatrix} P \\ Q \end{pmatrix} A^{-1}b$, we get that k steps of RPM is equivalent to preconditioning the original system by (Since postmultiplication of (3.36) with b results in (3.35)):

$$M^{-1} = (I, I)J^k \begin{pmatrix} P \\ Q \end{pmatrix} + (I, I)\Sigma_{i=0}^{k-1} J^i (I - J) \begin{pmatrix} P \\ Q \end{pmatrix} A^{-1} \tag{3.36}$$

Since J has no eigenvalues ≥ 1, then $\Sigma_{i=0}^{k-1} J^i = (J^k - I)(J - I)^{-1}$ [44]:

$$M^{-1} = (I, I)J^k \begin{pmatrix} P \\ Q \end{pmatrix} + (I, I)(I - J^k) \begin{pmatrix} P \\ Q \end{pmatrix} A^{-1} \tag{3.37}$$

Which is what we wanted to show.

■

By utilizing this preconditioner expression, we can also properly analyze the convergence rate.

THEOREM 5 (RPM Convergence Criteria) *The convergence rate bound of k steps of RPM is determined by $p(Q(H))$, where $p(x) = x^{k+1}$*

Proof We denote the preconditioner inside RPM by $M_{richardson}$, and the equivalent preconditioner expression of k steps of RPM as denoted in the previous proof by $M_{outside}$.

Without loss of generality, let the preconditioner inside RPM ($M_{richardson}$) be I. We may state that this is without loss of generality because using $M_{richardson}$ in

RPM is equivalent to using this preconditioner initially on A (i.e., applying RPM to the system $M^{-1}_{richardson}Ax = M^{-1}_{richardson}b$) and letting $M_{richardson} = I$ in (3.35).

With this reduction we have $A = I - H_{richardson}$ (where $H_{richardson} := I - M^{-1}_{richardson}A$).

Using the k-step RPM precondioner expression constructed above (and noting that theoretically $E = 0 = C$ in the Jacobian J):

$$\begin{aligned}
H_{outside} &= I - M^{-1}_{outside}(M^{-1}_{richardson}A) \\
&= I - (I,I)\begin{pmatrix} 0 & 0 \\ 0 & (QH_{richardson}Q)^k \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix} M^{-1}_{richardson}A \\
&\quad -(I,I)\begin{pmatrix} P \\ Q \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & (QH_{richardson}Q)^k \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix} \\
&= -(I,I)\begin{pmatrix} 0 & 0 \\ 0 & (QH_{richardson}Q)^k \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix} M^{-1}_{richardson}A \\
&\quad + \begin{pmatrix} 0 & 0 \\ 0 & (QH_{richardson}Q)^k \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix} \\
&= (QH_{richardson}Q)^k(I - M^{-1}_{richardson}A) = (QH^{k+1}_{richardson})
\end{aligned} \tag{3.38}$$

■

With this we can exactly specify how deflating more eigenvalues leads to speeding up convergence.

We will use these convergence results again in later chapters when we combine RPM with FGMRES.

4. GMRES

We first recall some basic facts about Petrov-Galerkin conditions in order to review the minimum residual and GMRES algorithms which we will use and build upon in the subsequent chapter on FGMRES.

In a Petrov-Galerkin algorithm, we wish to find an approximate solution $\hat{x} \in K$ (where K is a subspace with basis $V = [v_1, \cdots, v_m]$) such that the residual $r = f - Ax \perp \mathcal{L}$, where $\mathcal{L}$ is a space with basis $W = [w_1, w_2, \cdots, w_m]$. In short, we wish to make the residual continually orthogonal to a limiting sequence of subspaces (see figure 5 below).

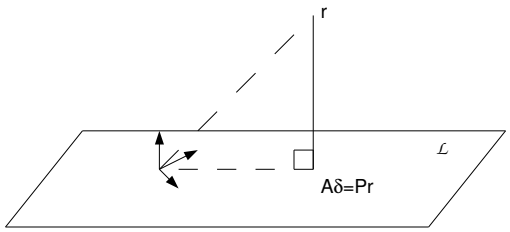

Figure 5. Galerkin Projection

In summary, this algorithm consists of

ALGORITHM 4 (Petrov-Galerkin)

Given kth iterate x_k*, search space* $\mathcal{K}_k, \mathcal{L}_k$

Find $x_{k+1} \in x_k + \mathcal{K}_k$ *so that* $r_0 - A\delta = b - Ax_k - A\delta \perp \mathcal{L}_k, \delta \in \mathcal{K}_k$

Repeat.

In particular, if we let $L = AK$, $K = span\{v = r\}, L = span\{w = Ar\}$ we recover the minimum residual algorithm:

ALGORITHM 5 (Minimum Residual)

$r = f - Ax, \rho = Ar$

Do until convergence:

$\alpha = \frac{r^t \rho}{\rho^T \rho}$

$x = x + \alpha r$

$r = r - \alpha \rho$

$\rho = Ar$

In particular, we will need the following result pertaining to the minimum residual algorithm:

THEOREM 6 (Residual Minimization)

$$\begin{array}{rcl} ||r_{k+1}||_2^2 & \leq & [1 - \frac{\mu^2}{\sigma^2}]||r_k||_2^2 \\ \textit{where} & & \\ \sigma & := & ||A||_2 = \rho^{\frac{1}{2}}(A^T A) \\ \mu & := & \lambda_{min}(A_S) = \lambda_{min}(\frac{1}{2}(A + A^T)) \end{array} \tag{4.1}$$

[36, 51]

Proof Let $A = A_S + A_{SS}$ where $A_S := \frac{1}{2}(A^T + A), A_{SS} := \frac{1}{2}(A - A^T)$, with A_S spd.

Notice that since A_{SS} is skew-symmetric $-u^T A_{SS}^T u = u^T A_{SS} u$, thus $u^T A_{SS} u = 0$.

Therefore,

$$\sigma = ||A||_2 \geq \frac{||Au||_2}{||u||_2}, \frac{u^T A u}{u^T u} = \frac{u^T A_S u + u^T A_{SS} u}{u^T u} \geq \lambda_{min}(A_S) = \mu \tag{4.2}$$

$$x_{k+1} = x_k + \alpha_k r_k \text{ and } r_{k+1} = r_k - \alpha_k A r_k \tag{4.3}$$

So:

$$\begin{aligned} ||r_{k+1}||_2^2 &= ||f - A(x_k + \alpha_k r_k)||_2^2 \\ &= ||r_k - \alpha_k A r_k||_2^2 \\ &= ||(I - \alpha_k A) r_k||_2^2 \\ &= r_{k+1}^T (I - \alpha_k A) r_k \end{aligned} \tag{4.4}$$

Use the Galerkin condition ($r_{k+1} \perp \mathcal{L} = span\{Ar_k\}$):

$$||r_{k+1}||_2^2 = ||r_k||_2^2 [1 - \frac{r_k^T A r_k}{r_k^T r_k}^2 \cdot \frac{1}{||Ar_k||_2^2}] \tag{4.5}$$

By (4.2):

$$||r_{k+1}||_2^2 \geq (1 - \frac{\mu^2}{\sigma^2}) ||r_k||_2^2 \tag{4.6}$$

Which is what we are trying to show.

■

Example 4.0.2 *Minimum Residual Convergence*

We illustrate the previous theorem by applying the minimum residual algorithm to a linear system in which the coefficient matrix is a Toeplitz matrix of order 100 with diagonal elements 2.1 and super and sub diagonal elements -1.

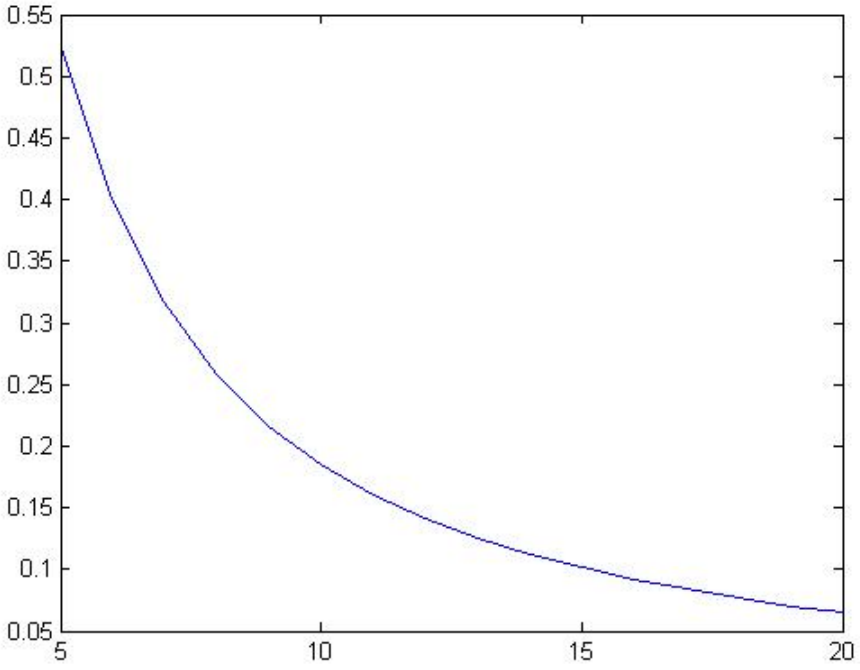

Figure 6. Convergence for the minimum residual algorithm

In the case of the minimum residual algorithm, the positive definiteness condition is quite strong, if we reduce the diagonal down to 1.1:

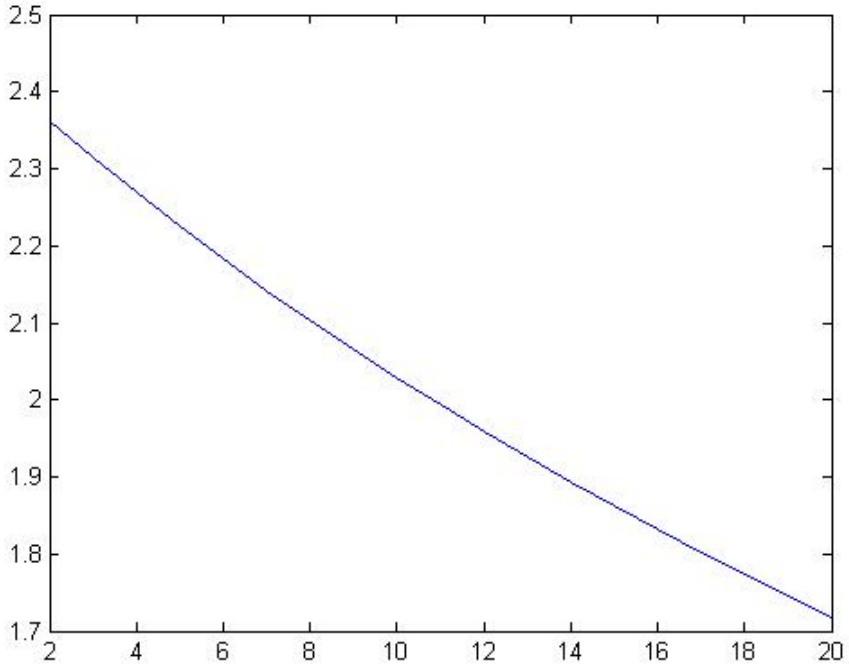

Figure 7. The minimum residual algorithm

applied to a system for which A_S is not symmetric positive definite

Now, if instead we choose $K_k := span\{r_0, Ar_0, A^2r_0, \cdots, A^{k-1}r_0\}$, i.e. a Krylov subspace, we make two observations.

The first, and this is a point which we will bring up later during the convergence proofs of GMRES, is that at step m, $x_m = x_0 + q_m(A)r_0$, where q_m is some mth degree polynomial.

The second, and the important heuristical reason for choosing this as our search space, is that if one notes the characteristic polynomial, $p(\lambda) = \lambda^n - \Sigma_{i=1}^{n-1} p_i \lambda^i - p_0$, it has A as a root, $A^n - \Sigma_{i=1}^{n-1} p_i A^i - p_0 = 0$.

Thus if we premultiply by A^{-1}, we get an expression for A^{-1}:

$$A^{-1} = \frac{1}{p_0}(A^{n-1} - p_{n-1}A^{n-2} - \cdots - p_2 A - p_1 I) \tag{4.7}$$

$x = A^{-1}f$ if $x_0 = 0, r_0 = f$, then this shows that such a choice for the search space eventually converges to the solution. Further, for such a choice, this

shows that there is a polynomial p ($p(0) = 1$) of degree not exceeding n for which $b - Ax = p(A)r_0$ [34, 36, 51].

Now, in an attempt to simplify such a Krylov basis in order to create a practical Petrov-Galerkin algorithm based on the Krylov basis, we utilize a Hessenberg reduction.

I.e., we find an orthogonal V so that $AV = VH$, with H being upper-Hessenberg.

Moreover, if we choose $||v_1||_2 = 1, v_1$ arbitrary (typically $v_1 = \frac{r_0}{||r_0||_2}$), then by analyzing the equality $AV = VH$, we must have that $h_{11} = v_1^T Av_1, h_{21}v_2 = Av_1 - h_{11}v_1, h_{21} = ||Av_1 - h_{11}v_1||_2, v_2 = \frac{(Av_1 - h_{11}v_1)}{h_{21}}, \cdots$

Continuing in the manner, and generalizing this Gram-Schmidt procedure, we obtain the generalized Arnoldi process.

ALGORITHM 6 (Generalized Arnoldi)

Pick $v_1 \in \Re^n$ *with* $||v_1||_2 = 1$.

For $j = 1 : m$

$w_j = Av_j$

For $i = 1 : j$

$h_{ij} =< v_i, w_j >$

$w_j = w_j - h_{ij}v_i$

End

$h_{j+1,j} = ||w_j||_2$

If $h_{j+1,j} = 0$

$m := j$

break

$v_{j+1} = \frac{w_j}{h_{j+1,j}}$

End

Combining this Arnoldi process with the Petrov-Galerkin algorithm on a Krylov subspace above, we obtain the GMRES algorithm.

ALGORITHM 7 (GMRES)

$r_0 = b - Ax_0, \beta := ||r_0||_2, v_1 = \frac{r_0}{\beta}$

For $j = 1, 2, ..., m$ } *Arnoldi*

$w_j := Av_j$

For $i = 1, ..., j$

$h_{ij} := (w_j, v_i); w_j := w_j - h_{ij}v_i$

End

$h_{j+1,j} = ||w_j||_2$

If $h_{j+1,j} = 0$

$m := j$

break

$v_{j+1} = \frac{w_j}{h_{j+1,j}}$

End

$H_m := [h_{ij}]$

Find $y_m = min||\beta e_1 - H_m y||_2$ *via a Givens rotation QR process, keeping in mind that* H *is Hessenberg.*

$x_m := x_0 + V_m y_m$

[36, 51, 53]

In case $h_{j+1,j} = 0$ for both algorithms above, then we happen to hit the minimal polynomial of A with respect to the vector v_1, which is admittedly rare. However, in such a case, this implies that the computed residual is 0, and that we have obtained the exact solution.

With this description of GMRES, we can now outline the theoretical results which we will need to compare with FGMRES.

The following lemma will be very similar to theorem 10 below.

LEMMA 1 *[36, 51] Let x_m be the mth step approximate solution obtained by GMRES, and $r_m := b - Ax_m$. Then*

$$x_m = x_0 + q_m(A)r_0 \tag{4.8}$$

and

$$||r_m||_2 = ||(I - Aq_m(A))r_0||_2 = min_{q \in \mathbb{P}_{m-1}}||(I - Aq(A))r_0||_2 \tag{4.9}$$

Where q is a polynomial of degree not exceeding $m - 1$.

Proof Denote the Krylov space by $\mathcal{K}_m, \mathcal{L}_m = A\mathcal{K}_m$.

Then

$$\begin{aligned} min_{y \in \mathcal{L}_m}||b - y|| &= min_{x \in \mathcal{K}_m}||b - Ax|| \\ &= ||b - Ax_m|| \end{aligned} \tag{4.10}$$

which is the case iff $< b - Ax_m, v >= 0$ for $v \in \mathcal{L}_m$, but this is precisely the condition for algorithm 4.

But $\mathcal{K}_m$ is precisely the set of all vectors of the form $x_0 + q(A)r_0$.

■

With this lemma we have only to discuss two important convergence results that can be compared with those of FGMRES.

THEOREM 7 *[51] If $A + A^T$ is spd, then restarted GMRES converges for any choice of k.*

Proof GMRES uses the Krylov subspace K_m at each restart of GMRES.

The minimum residual algorithm is equivalent to GMRES with K_2.

By the previous lemma, therefore, restarted GMRES reduces the residual at least as much as minimum residual.

Since the minimum residual algorithm converges if $A + A^T$ is spd by (4.1), then GMRES converges as well for $A + A^T$ spd.

■

Example 4.0.3 *GMRES Convergence*

Just like with the minimum residual algorithm, we can illustrate the convergence of GMRES for solving a Toeplitz tridiagonal system $Ax = f$ of order 100, with diagonal elements 4 and super and sub diagonal elements -1.

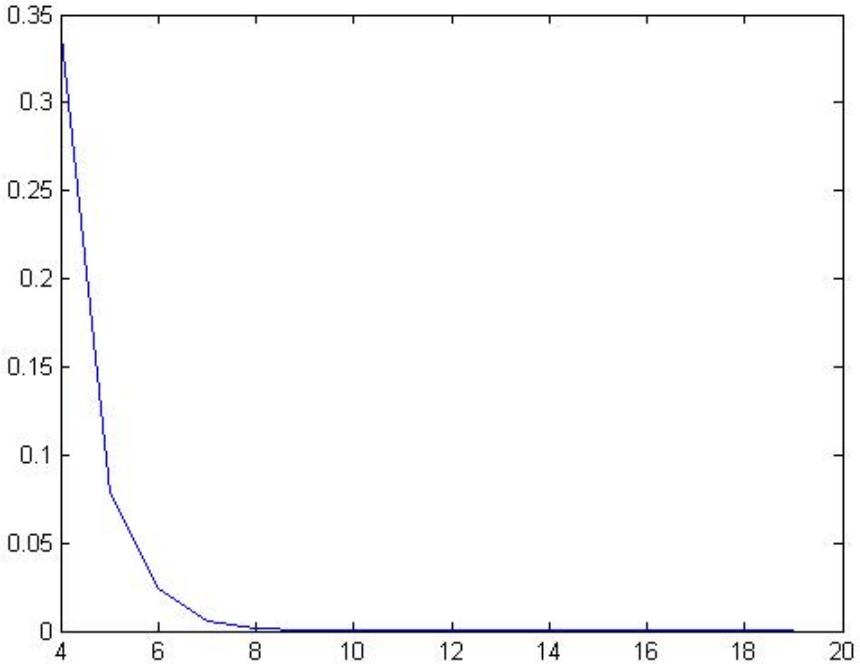

Figure 8. Convergence of GMRES

As with the minimum residual algorithm, the positive definiteness condition is quite strong, if we shift the diagonal of A down to 1.9 to allow the eigenvalues to lie on both sides of the imaginary axis, convergence is not assured, see figure 9.

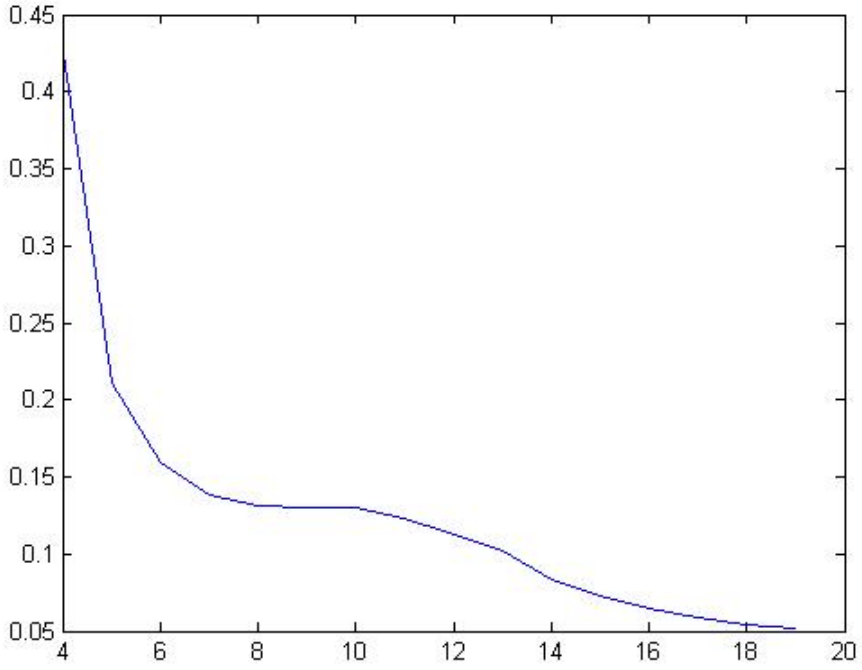

Figure 9. GMRES convergence without a symmetric positive definite part

THEOREM 8 *[51]*

Let A be diagonalizable, i.e. $A = X\Lambda X^{-1}, \Lambda = diag(eigenvalues)$*, then* $||r_m||_2 \leq \kappa_2(X)\epsilon^{(m)}||r_0||_2$.

Where $\epsilon^{(m)} = min_{p(x)\in\mathbb{P}_m, p(0)=1} max_{1\leq i\leq n}|p(\lambda_i)|$

Proof As described in the theory of GMRES above, there is a polynomial p ($p(0) = 1$) of degree not exceeding m for which $b - Ax = p(A)r_0$, thus

$$||b - Ax||_2 = ||Xp(\Lambda)X^{-1}r_0||_2 \leq ||X||_2||X^{-1}||_2||r_0||_2||p(\Lambda)||_2 \tag{4.11}$$

Since Λ is diagonal, then

$$||p(\Lambda)||_2 = max_{i=1,\cdots,n}|p(\lambda_i)| \tag{4.12}$$

Since x_m minimizes the residual norm over $x_0 + \mathcal{K}_m$, then

$$||b - Ax||_2 = ||Xp(\Lambda)X^{-1}r_0||_2 \leq ||X||_2||X^{-1}||_2||r_0||_2(min_{p\in\mathbb{P}_m, p(0)=1}max_{i=1,\cdots,n}|p(\lambda_i)|) \tag{4.13}$$

Which is the desired result.

■

With this appropriate background of GMRES, we can now discuss the connections between GMRES and deflation-based methods discussed in the previous chapters. Like these deflation-based methods, GMRES also implements features that mitigate detrimental influences on convergence via projection, albeit implicitly.

As FGMRES builds on GMRES, and because we will be incorporating FGMRES with an inner-projection based method, it will be important to conclude this section by noting the following theorem which indicates deflation-like properties inherent in GMRES similar to what we previously analyzed for RPM in the preceding chapter.

THEOREM 9 *[21] Let $A = I - B$ be non-singular, with p eigenvalues of B outside the open unit disk, and let Q be the projector onto the invariant subspace of B corresponding to its p largest eigenvalues and P the projector corresponding to the invariant subspace of the $n-p$ smallest eigenvalues (e.g., let the Schur decomposition of $B = URU^*$ with its eigenvalues ordered from largest in modulus to smallest, and let Z be the first p column of U, then $Q = ZZ^T$ and $P = I - ZZ^T$), then for GMRES and $k \geq p$:*

$$||r_k|| \leq K||r_{k-p}^P|| \tag{4.14}$$

Where r_{k-p}^P corresponds to applying GMRES on the projected system according to the projector P, and K a constant.

Proof As expressed in lemma 1 above, there exists a polynomial $p \in \mathcal{P}_k$ so that $deg(p) \leq k, p(0) = 1$, and p minimizes $||p(I - B)r_0||$.

Let $\tau(z) = p(1-z)$, then:

$$||r_k|| \leq ||\tau(B)r_0|| \tag{4.15}$$

Further, if we have such projectors P and Q, then due to the condition that such projectors are constructed so that they are invariant under the appropriate eigenspaces of B, then $B = PBP + QBQ$. Moreover, $p(B) = p(PBP)P + p(QBQ)Q$.

Now, suppose we construct the Lagrangian polynomial p_1 so that it is of the same degree as p, it vanishes at each eigenvalue λ_i not in the unit disk, and so that $p_1(1) = 1$. Then, using the Lagrangian polynomial construction:

$$p_1(z) = \Pi_{\lambda_i \notin \{x \in \mathbb{C} || x| < 1\}} \frac{z - \lambda_i}{1 - \lambda_i} \tag{4.16}$$

Then by construction:

$$p_1(PBP)P = 0 \tag{4.17}$$

Let τ_2 be defined as the GMRES polynomial corresponding to the $k-p$ iteration solving the projected system corresponding to the projector P:

$$\begin{aligned} (P - PBP)x &= Pb \\ Qx &= 0 \end{aligned} \tag{4.18}$$

Thus, $r_{k-p}^P = \tau_2(PBP)Pr_0$.

Let $q := p_1\tau_2$, then (using (4.17)):

$$\begin{aligned} q(B)r_0 &= q(PBP)Pr_0 + q(QBQ)Qr_0 \\ &= p_1(PBP)\tau_2(PBP)Pr_0 + p_1(QBQ)\tau_2(QBQ)Qr_0 \\ &= p_1(QBQ)\tau_2(QBQ)Qr_0 \\ &= p_1(QBQ)r_{k-p}^P \\ &= p_1(QBQ)Qr_{k-p}^P \end{aligned} \tag{4.19}$$

Therefore, using (4.15), we established the proof with $K = ||p_1(QBQ)Q||$.

■

5. FGMRES

As outlined above for GMRES, the Arnoldi loop constructs the following orthogonal basis of a preconditioned Krylov subspace:

$$Span(r_0, AM^{-1}r_0, \cdots, (AM^{-1})^{m-1}r_0) \tag{5.1}$$

In which the new vector is obtained from the previous vector in the process. The last step is a linear combination of the previous vectors $z_i = M^{-1}v_i, i = 1, \cdots, m$. Here, we need only apply M^{-1} to $V_m y_m$. However, if we allow the preconditioner to change at each step, we would have

$$z_j = M_j^{-1}v_j \tag{5.2}$$

If we do this modification, we can modify the above algorithm to create GMRES with flexible preconditioning, or FGMRES:

ALGORITHM 8 (FGMRES)

Let x_0 be an initial vector, m a preset dimension of the Krylov subspace, and define $H_m \in \Re_{(m+1)\times m}$.

Perform Arnoldi

1. *Compute $r_0 = b - Ax_0, \beta = ||r_0||_2, v_1 = \frac{r_0}{\beta}$*
2. *For $j = 1, \cdots, m$ do*
3. *$z_j := M_j^{-1}v_j$*
4. *$w := Az_j$*

5. For $i = 1, \cdots, j$ *do* $h_{i,j} :=< w, v_i >, w := w - h_{i,j} v_i$

6. $h_{j+1,j} = ||w||_2$, *if* $h_{j+1,j} = 0$ *break,* $v_{j+1} = \frac{w}{h_{j+1,j}}$

7. $Z_m := (z_1, \cdots, z_m)$

The approximate solution is then $x_m = x_0 + Z_m y_m$, *where* y_m *is the solution to the linear least squares problem* $H_m y = \beta e_1$.

It is clear that the above algorithm is mathematically equivalent to GMRES when $M_j = M$ for $j = 1, \cdots, m$. [8, 51, 60]

In order to compare FGMRES with GMRES (before extending FGMRES results past the current literature), we note the following basic properties of FGMRES.

First, the following mimics lemma 1 above.

THEOREM 10 $min_{x \in x_0 + span(Z_m)} ||b - Ax||_2 = ||b - Ax_m||$

Proof Since in GMRES we are performing a modified Gram-Schmidt procedure on

$$Span(r_0, AM^{-1} r_0, \cdots, (AM^{-1})^{m-1} r_0) \tag{5.3}$$

We obtain the relation

$$(AM^{-1}) V_m = V_{m+1} H_m \tag{5.4}$$

Similarly, for FGMRES

$$A Z_m = V_{m+1} H_m \tag{5.5}$$

Now let $z = x_0 + Z_m y$ be an arbitrary vector $\in x_0 + span(Z_m)$, then using the above:

$$\begin{aligned} b - Az &= b - A(x_0 + Z_m y) \\ &= r_0 - AZ_m y \\ &= \beta v_1 - V_{m+1} H_m y \\ &= V_{m+1}(\beta e_1 - H_m y) \end{aligned} \tag{5.6}$$

However, the final step of the algorithm minimizes $||\beta e_1 - H_m y||_2 = ||b - A(x_0 + Z_m y)||_2$.

■

Likewise, the following mimics the breakdown case of GMRES.

THEOREM 11 *[49] Assume that $\beta = ||r_0||_2 \neq 0$ and that $k-1$ steps of FGMRES have been successfully performed, thus $h_{i+1,i} \neq 0, i < k$, and that H_k is nonsingular. Then x_k is exact iff $h_{k+1,k} = 0$*

Proof Let $h_{k+1,k} = 0$, then $AZ_k = H_k V_k$, and

$$||\beta v_1 - AZ_j y_j||_2 = ||\beta e_1 - H_k y_k||_2 \tag{5.7}$$

Since H_k is nonsingular, then $y_k = \beta H_k^{-1} e_1$ minimizes the above norm, and in fact yields $x_k = x$ (exact solution). Likewise, if x_k is exact, then:

$$0 = b - Ax_k = V_k(\beta e_1 - H_k y_k) + v_{k+1} e_k^T y_k \tag{5.8}$$

If $e_k^T y_k = 0$, then $H_k y_k = \beta e_1$. But since $h_{i+1,i} \neq 0, i < k$, and $y_k = 0$, then $\beta = 0$, resulting in a contradiction. Thus $e_k^T y_k \neq 0$. Premultiplying the above by V_k^T and v_{k+1}^T and noting orthogonality we conclude that $\beta e_1 = h_k y_k$, and $v_{k+1} = 0$, respectively. Thus $h_{k+1,k} = 0$. ■

Now, in order to extend these GMRES results, and add to the results on FGMRES, we note that FGMRES is equivalent to GMRES on a particular matrix.

THEOREM 12 *FGMRES applied to a linear system $Ax = b$ is equivalent to applying GMRES to a linear system $Yx = b$ for some $n \times n$ matrix Y.*

Proof Notice that if FGMRES uses a sequence of preconditioners $M_i^{-1}A$, then FGMRES minimizes the residual over a polynomial of the vectors $z_0, z_1, \ldots$, and thus $Yz_i = z_{i+1}$ defines a matrix for which performing GMRES on Y is equivalent to applying FGMRES on A with the sequence of preconditioners M_i. Specifically, Y can be found algebraically as:

$$\begin{aligned} Y &= (z_1|z_2|z_3|\ldots)(z_0|z_1|z_2|\cdots)^{-1} \\ &= Z_1 Z_0^{-1} \end{aligned} \tag{5.9}$$

In this way, FGMRES is equivalent to GMRES on Y, and Y describes the convergence behavior of FGMRES. ■

Example 5.0.4

In particular, this shows the strong dependence of FGMRES on the initial test vector.

Take as an example, an FGMRES algorithm which has $M_i^{-1}A$ is a matrix that permutes rows i and $i+1$, $r_0 = e_i$, with $M_{n-1}^{-1}A$ an arbitrary matrix, and $M_n^{-1}A = I$. Thus since $e_i^T e_j = 0$ $(i \neq j)$, the Arnoldi process will trivially produce:

$$\begin{aligned} Y &= \begin{pmatrix} e_1 & e_2 & \cdots & e_{n-1} & \overline{a} \end{pmatrix}^{-1} \begin{pmatrix} e_2 & e_3 & \cdots & \overline{a} & \overline{a} \end{pmatrix} \\ &= P_{1n} \begin{pmatrix} I_{n-1} & C \\ 0 & B \end{pmatrix} \end{aligned} \tag{5.10}$$

Where P_{1n} is the permutation matrix the permutes the first and last rows, and $\overline{a}$ can be made an arbitrary vector via appropriate choice of $M_{n-1}^{-1}A$ and r_0. Thus, in this simple example, the spectrum can depend wholely on r_0 (because P_{1n} is nonsingular, then multiplication by it forms a homeomorphism; therefore, if B is made to vary its eigenvalue in modulus from 0 to ∞, then the same must occur under the homeomorphism [1]).

In order to use this expression for Y carefully, and establish very limited convergence results to compare with the previously exhibited GMRES convergence

[1]The author would like to thank Kyle Kloster and Jake Noparstak for this argument

results, we will need the following lemma. This is essentially a stability result of GMRES applied to FGMRES placing a restriction on the variation of the preconditioners from one iteration to another (similar to some results in [24, 50]). A much stricter bound can be found using [12], but for the purposes of this study we neither need such strict results, and we will use the following result to build a connection between the behavior of the residual norm of FGMRES and the geometric mean of the behavior of the residual norm of the individually preconditioned GMRES iterations.

LEMMA 2 *Assume that* $||M_i^{-1} - M_j^{-1}|| \leq \epsilon$.

Let the initial vector be given as x_0 *and* $r_0 := b - Ax_0$.

Let x_k *be the solution after* k *steps of FGMRES* ($x_k \neq x$) *and* H_k *be nonsingular.*

Let $a_1 = M_1^{-1} r_0$, *and define inductively* $a_k = \Sigma_{j=1}^{k} M_j^{-1} (\Sigma_{i=1}^{k-1} \alpha_{i,j,k} M_i a_i + \gamma_{k-1,j} A a_{k-1})$ *where* $\alpha_{i,j,k}, \gamma_{k-1,j}$ *are given.*

Define the Y*-matrix so that* $Y a_i = a_{i+1}$.

Let y_k *be the solution after* k *steps of GMRES on* Y *with* $M_1^{-1} r_0$ *in place of* r_0.

Then $||x_k - y_k|| \leq C_k \epsilon$ *for some constant* C_k *or* $||b - Ax_k|| \leq ||b - Ay_k||$.

Proof We leave the proof of this in Appendix A.

■

With this, we may now establish some basic results for FGMRES.

The following result parallels theorem 7.

THEOREM 13 (Y is Positive Definite) *If each of the matrices* $M_i^{-1} A$ *has symmetric part positive definite parts and* $||M_i^{-1} - M_j^{-1}|| \leq \epsilon$, *then FGMRES converges.*

Proof In lemma 2, let $\gamma_{2,j} = 1$ for $j = 1, \cdots, m$, else $\alpha, \gamma = 0$, let $z_0 = M_1^{-1} r_0$, then $a_1 = z_0, a_k = M_{k-1}^{-1} z_0$. Consequently, the residual r_m resulting from using GMRES on Y defined by a_i is within a constant times ϵ of the actual residual term using FGMRES or bounds it from above.

Then by lemma 1:

$$
\begin{aligned}
&||r_m||^m \\
&\quad = \quad (min_{\overline{\alpha}}||z_0 - \alpha_1 \cdot M_1^{-1}Az_0 - \alpha_2 \cdot M_2^{-1}Az_0 - \cdots - \alpha_m \cdot M_m^{-1}Az_0 - \cdots||)^m \\
&\quad \leq \quad \Pi_{i=1}^m min_{\alpha_i}||r_0 - \alpha_i \cdot M_i^{-1}Az_0|| \qquad (5.11)
\end{aligned}
$$

With each item in the product is the minimum residual with respect to $M_i^{-1}A$, and thus by theorem 7:

$$||r_m||^m \leq [\Pi_{i=1}^m(1 - \frac{\mu_i^2}{\sigma_i^2})]||r_0|| \qquad (5.12)$$

where $\mu_i = \lambda_{min}(M^{-1}A + (M^{-1}A)^T)/2, \sigma_i = ||M_i^{-1}A||_2$

■

Important Remark:

This result can be used to give some weak convergence bounds–*even* in the case where not all of $M_i^{-1}A$ have symmetric positive definite parts (since the minimum residual bound still holds).

That is to say, $\mu_i \in \sigma(I) - \sigma(J_i)$, and when we vary the stiff subspace, a weak bound on when convergence still occurs can be thus given by the above result.

Example 5.0.5 *Numerical Test: Convergence rate of FGMRES with components that have symmetric positive definite parts*

The following illustrates the above result and remark. To calculate the bound it uses the minimum residual bound, and then uses the bound formed by the geometric mean of the residual norm bound for each of the individually preconditioned GMRES iterations as in equation (5.12), and compares it with residual norm of FGMRES. Since the minimum residual bound is not tight, the overall bound is not tight.

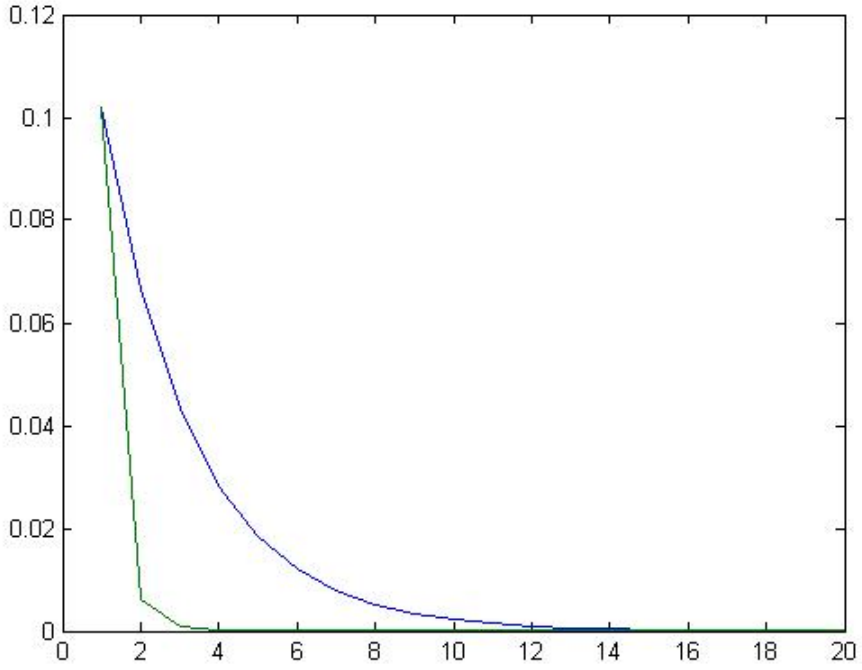

Figure 10. FGMRES bound for components
with symmetric positive definite parts

Of particular interest in theorem 13 is the appearance of the geometric mean in equation (5.12). Following this theme, and recalling theorem 12, we wish to analyze this matrix Y to see if this observation has further merit.

The result we obtain is very limited, but does indeed point to this behavior:

THEOREM 14 (Unit Disk Convergence of Y) *If each of the matrices $I - M_i^{-1}A$ has norm < 1, $||M_i^{-1} - M_j^{-1}|| \leq \epsilon$, and the right-hand side vector $M^{-1}b$ has all nonzero entries under the Jordan basis of the matrix Y described below, then the residual norm of FGMRES is identical to the residual norm of GMRES on a matrix Y whose spectral radius is asymptotically bounded by the geometric mean of the norm of the matrices $I - M_i^{-1}A$.*

Proof Let $b \leftarrow M^{-1}b$, $x_0 \leftarrow 0$, it will be useful to note a similar construction of the FGMRES matrix that if we consider minimizing the residual over the polynomial

of the vectors $b, (I - AM_1^{-1})b, (I - AM_2^{-1})(I - AM_1^{-1})b, \ldots$, then by theorem 12 performing FMGRES is equivalent to performing GMRES on a Y with

$$\begin{aligned} Y &= [((I - AM_1^{-1})b|(I - AM_2^{-1})(I - AM_1^{-1})b|\cdots) \\ &\quad (b|(I - AM_1^{-1})b|(I - AM_2^{-1})(I - AM_1^{-1})b|\cdots)^{-1}] \\ &= Z_1 Z_0^{-1} \end{aligned} \tag{5.13}$$

We will perform the rest of the analysis with this Y. The rest of the result follows by applying the previous lemma 2 with p_i chosen to give the matrix Y above, namely, $\alpha_{i,i,i} = 1$ and $\gamma_{i,i} = -1$ (else $\alpha, \gamma = 0$).

Assume that Y is nonsingular, and that λ_1 is an eigenvalue corresponding to the spectral radius, and if the Jordan canonical form of $Y = XJX^{-1}$ then the right hand side b is such that $e_m^T X^{-1} b \neq 0$ where m is the geometric multiplicity of λ_1. Hence,

$$\left(b|Yb|Y^2b|\cdots\right) c = Y^n b \tag{5.14}$$

Let the Jordan form of $Y = XJX^{-1}$ with J ordered so that the first Jordan block contains λ_1 with geometric multiplicity m. Further, since X is nonsingular, $\exists x_l$ such that $x_l^T X = e_m^T$. Finally, let $d = X^{-1}b$. With these simplifications, multiply the above through by x_l^T:

$$(0, 0, \ldots, 0, \Sigma_{i=1}^n c_i \lambda_1^{i-1}, 0, 0, \ldots, 0)d = x_l^T Y^n b \tag{5.15}$$

Since c_i satisfies the minimal polynomial:

$$\lambda_1^n d_i = x_l^T Y^n b \tag{5.16}$$

By assumption $d_i \neq 0$, so then:

$$|\lambda_1|^n \leq \frac{\frac{|x_l^T Y^n b|}{||Y^n b||}}{\frac{d_i}{||b||}} \frac{||Y^n b||}{||b||} \tag{5.17}$$

Using Bunyakovsky-Cauchy-Schwartz inequality together with the fact that $x_l^T X = e_m^T$, then $||x_l|| \leq ||X^{-1}||$:

$$|\lambda_1|^n \leq \frac{1}{\frac{d_i}{||X^{-1}||||b||}} \frac{||Y^n b||}{||b||} \tag{5.18}$$

Let $C := \frac{1}{\frac{d_i}{||X^{-1}||||b||}}$, then:

$$|\lambda_1| \leq C^{\frac{1}{n}} (\frac{||Y^n b||}{||b||})^{\frac{1}{n}} \tag{5.19}$$

Now using the fact that $Y^n b = \Pi(I - M_i^{-1}A)b$:

$$|\lambda_1| \leq C^{\frac{1}{n}} (||\Pi(I - M_i^{-1}A)||)^{\frac{1}{n}} \tag{5.20}$$

Thus, after noting that $C^{\frac{1}{n}} \to 1$, if the geometric mean of the norm corresponding preconditioners ($||I - M_i^{-1}A||$) are < 1, the spectral radius is also < 1. ■

Although this result does not show that in the unit norm case that the residual norm of FGMRES asymptotically approaches the geometric mean of the residual norm of each individual preconditioned GMRES (it only compares with GMRES performed on the matrix Y), this claim is backed by numerical experiments as shown below.

Example 5.0.6 *Numerical Test: FGMRES with a family of matrices where* $||M_i^{-1}A|| < 1$ *v. a bound which is the geometric mean of the residual norms of the individually preconditioned GMRES iterations*

The following exhibits the tight bound the geometric mean of the residual norm of the individually preconditioned GMRES iterations gives. The places where FGMRES crosses the bound might be due to errors from the constant factor C in the above proof.

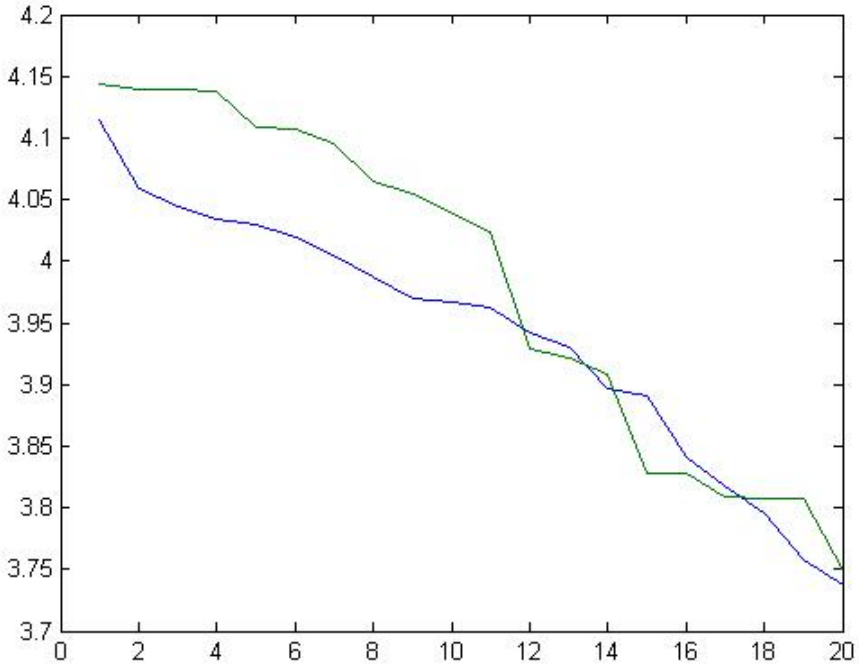

Figure 11. FGMRES residual norm v. geometric mean of GMRES residuals

In summary, the following results pertaining to FGMRES suggest even more general results:

- FGMRES is equivalent to GMRES on a different matrix Y.
- In the case where the symmetric part of the matrices $M_i^{-1}A$ is positive definite, the geometric mean of the positive definite bound obtained in theorem 7 in equation (5.12) forms a bound for FGMRES.
- The bound on the spectral radius of Y follows a similar geometric mean property as in equation (5.20).
- Numerical experiments also suggest an asymptotic bound which is the geometric mean of the residual norms of the individually preconditioned GMRES iterationswhere the preconditioned matrices have norm less than unity.

We would like to offer the following conjecture: that under some suitable restriction on A and b (given example 5 above) that if each of the individual precon-

ditioners of FGMRES converges, then not only will FGMRES converge; but its the residual norm of FGMRES will be the geometric mean of the residual norm of each of the individually preconditioned GMRES iterations.

It should be noted that this result exposes a flaw in the existing literature on FGMRES. Namely, if this conjecture is true, then FGMRES can perform no better than simply choosing the best preconditioner in the adaptive preconditioning in FGMRES. There are two remedies for this. First is that there might be computational advantages for not generating the preconditioner explicitly or fully in the first iteration. Second, given the strong restrictions on the preconditioners placed in lemma 2, the 'suitable restriction' on A and b might be necessarily insignificant.

6. BANDED PRECONDITIONING

If we are utilizing a banded preconditioner inside of RPM, then the preconditioner M inside RPM is a banded matrix. The main advantage of using a banded preconditioner is the use of the SPIKE algorithm [39, 45] as a subroutine in backsolving systems involving M, and thus this can be used as a computationally efficient subroutine inside of RPM.

However, in order to complete our analysis, we need to analyze the spectral properties of the iteration matrix $M^{-1}N$ inherent to this particular choice of a preconditioner. In particular, we require a condition for banded matrices which is similar to the conditions for generalized diagonal dominance. To help in this regard, we first recall the following theorem due to Stein.

THEOREM 15 *[34, 57, 61] For any matrix B, $\rho(B) < 1$ iff there exists a positive definite matrix T s.t. $T - B^H T B$ is also positive definite.*

Proof Suppose that $T, T - B^H T B$ are positive definite.

Then T^{-1} has a Cholesky decomposition and $= PP^H$.

Consider the matrix norm $||A||_P := ||P^{-1}AP||_2$.

Then note that $||B||_P^2 = ||P^{-1}BP||_2^2 = \rho(P^H B^H P^{-H} P^{-1} BP) = \rho(P^H B^H T B P)$.

Thus $||B||_P^2$ is the largest eigenvalue of $P^H B^H T B P$, i.e., $||B||_P^2$ is the largest zero of $det(\lambda I - P^H B^H P^{-H} P^{-1} BP) = det(P^{-H}) det(\lambda T - B^H T B) det(P^{-1}) = det(\lambda T - B^H T B) det(T)$, i.e., the largest zero of $det(\lambda T - B^H T B)$.

Therefore, since this is the largest zero, then for $\lambda > ||B||_P^2$, $\lambda T - B^H T B$ is positive definite.

But since $T - B^H T B$ is positive definite, then λ can at least be 1, thus $1 > ||B||_P^2 > \rho(B)^2$.

Thus $\rho(B) < 1$, finishing the backwards direction.

Now assume that $\rho(B) < 1$.

Let V be the Jordan basis of B, define $\Lambda = diag(1, \epsilon, \epsilon^2, \cdots)$, and then let $P := V^{-1}\Lambda^{-1}$.

Then $||P^{-1}BP||_2 = ||\Lambda J \Lambda^{-1}||_2 < 1$ for the correct choice of ϵ, because $\Lambda J \Lambda^{-1}$ is the matrix of the eigenvalues of B along the diagonal, and ϵ along the subdiagonal (and because $||X||_2$ is a continuous function).

But then $||P^{-1}BP||_2^2 = \rho(P^H B^H P^{-H} P^{-1} BP) < 1$.

But if we let $T = P^{-H}P^{-1}$, then the above implies that $I - P^H B^H T B P$ is positive definite.

But $I - P^H B^H T B P$ is similar to $T - B^H T B$, thus $T - B^H T B$ is positive definite.

And $T = P^{-H}P^{-1}$ is also positive definite because it has a given Cholesky decomposition.

■

Now we may show a condition which ensures that the spectra of the iteration matrix is contained within the unit disk.

THEOREM 16 *If*

$$||M||_2^2 \kappa_2^{-1}(M) > ||N||_2^2 \tag{6.1}$$

Then $\rho(M^{-1}N) < 1$.

Proof Assume that

$$||M||_2^2 \kappa_2^{-1}(M) > ||N||_2^2 \tag{6.2}$$

Then $M^T M - (M^{-1}N)^T M^T M M^{-1} N = M^T M - N^T N$ is p.d., since

$$x M^T M x - x N^T N x > \lambda_{min}(M^T M) - ||N||_2^2 \tag{6.3}$$

$$= \rho(M^T M)\kappa_2^{-1}(M) - ||N||_2^2 \tag{6.4}$$

$$= ||M||_2^2 \kappa_2^{-1}(M) - ||N||_2^2 > 0 \tag{6.5}$$

Now if we apply theorem 6 with $G = M^T M$ and $B = M^{-1}N$, we see that $\rho(M^{-1}N) < 1$. ■

What this theorem immediately implies is that, given that we do not have too ill-conditioned a matrix or preconditioner, the heavier the elements along the band, the better chance we have of ensuring a good spectrum of the iteration matrix. Therefore, if we apply a reordering which brings heavily weighted elements into this band, we can possibly guarantee convergence for a significant class of matrices.

7. FIEDLER

In order to improve the weight of the central in accordance with equation 6.1 above, we suggest to implement Fiedler reordering. However, due to considerations involving the 2-norm in equation 6.1, we will propose a modified Fiedler algorithm after introducing the key heuristics behind its use.

We will assume in this section that A is symmetric and all entries are ≥ 0. To extend to nonsymmetric general matrices, we apply the following section to the symmetric part of $|A|$ (the absolute value of the entries of A). Further, throughout this section $G = G(A)$ will be the weighted graph (with some given orientation) expression given a matrix A (likewise, $A(G)$ is the matrix corresponding to a weighted oriented graph G), λ_i refers to the ith largest eigenvalue in modulus of A where $\lambda_\infty = \lambda_n$ is the largest eigenvalue in modulus of A, $V(G)$ is the collection of the vertices of the graph (with vertices denoted as u or v), $E(G)$ is the collection of the edges of the graph (with individual edges denoted as f or as uv to denote the edge between vertices u and v), ψ is a given labeling of $V(G)$ where $\psi : V(G) \to \{1, 2, \cdots, n\}, n = |V(G)|$, $\overline{1}$ is the vector consisting of all entries equal to 1, and finally $p \in \Re, 0 < p \leq \infty$. We will also assume that all graphs in consideration have only one connected component.

We first define the Laplacian of a graph. In order to motivate this definition, and to aid in proving some essential properties, we first define a weighted incidence matrix with orientation, D, of a given graph G.

DEFINITION 1 (Incidence Matrix) *Given a weighted graph G with some given orientation, define the (u, f) entry of D as the square root of the absolute value of the given weight between vertex u and edge f, where the (u, f) entry is positive if vertex u is the head negative if the tail and 0 otherwise. [28, 29, 41]*

With this, we may now define the weighted Laplacian of a graph.

DEFINITION 2 (Laplacian) *Given a matrix A corresponding to a graph G, let D be the weighted directed incidence matrix corresponding to G, then the Laplacian is $Q = DD^T$ [28, 41].*

We will use this interchangeably with the following equivalent definition of the Laplacian:

LEMMA 3 *[28]*

Let Δ be the vector of row sums of A and let A' be the matrix A with diagonal entries set to 0, then $Q = diag(\Delta) - A'$. [29]

Proof The inner product of any two distinct rows of D is the sum of the square of the weights joining the corresponding vertices. However, since by definition the (u, f) entry of D is the square root of the absolute value, this means that the inner product of any two distinct rows of D is the weight joining the corresponding vertices.

Thus it is 0 or the negative of the absolute value of the weight according to as the vertices are adjacent or not, or the sum of the weights of the edges connected to a given vertex in the case where the two rows are the same.

But given this result for the inner product of the rows of D, we see immediately that $DD^T = diag(\Delta) - A'$, which is what we were trying to show.

■

We use this definition for practical computation of the Laplacian. As a further immediate consequence of this alternative definition, notice that it is clear that $\lambda_1 = 0$. With this observation we define:

DEFINITION 3 (Fiedler Value) *We call λ_2 of $Q(A)$ the Fiedler value of the Laplacian, we also call the corresponding eigenvector, $x^{(2)}$, the Fiedler vector. [16, 17, 29, 35, 41, 46]*

With this terminology, we can define the original Fiedler reordering.

ALGORITHM 9 *Fiedler Reordering*

Let $\lambda_2, x^{(2)}$ be the second smallest eigenvalue and corresponding eigenvector of the Laplacian of a given weighted G associated with a matrix A.
Let $\psi_{fiedler} : V(G) \to \{1, 2, \cdots, n\}$ be the labeling induced by sorting $x^{(2)}$ from smallest to largest. The reordering induced by this labeling $\psi_{fiedler}$ is the Fiedler reordering.

Now we may introduce some basic preliminary results. [29, 35, 41, 46]

THEOREM 17 *[28, 29, 35, 41, 46]*

Let x be any vector and Q the Laplacian corresponding to the graph G corresponding to the matrix A, then

$$x^T Q x = \Sigma_{uv \in E(G)} a_{uv}(x_u - x_v)^2 \tag{7.1}$$

[16, 17]

Proof We need only note first that by using the incidence matrix definition of the Laplacian:

$$x^T Q x = x^T D D^T x = (D^T x)^T (D^T x) \tag{7.2}$$

And second that the definition of an incidence matrix is that if $uv \in E(G)$, then the entry of $D^T x$ corresponding to uv equals $\pm\sqrt{|a_{uv}|}(x_u - x_v)$.

■

With this, we can note the following due to Fiedler, which will be essential in demonstrating useful bounds on the Fiedler values:

THEOREM 18 *[17, 28, 35, 40, 41, 46]*

$$min_{x:x\perp\overline{1}}\frac{\Sigma_{uv\in E(G)}a_{uv}(x_u-x_v)^2}{\Sigma_u x_u^2}=\lambda_2 \tag{7.3}$$

$$max_{x:x\perp\overline{1}}\frac{\Sigma_{uv\in E(G)}a_{uv}(x_u-x_v)^2}{\Sigma_u x_u^2}=\lambda_\infty \tag{7.4}$$

Proof If we apply Courant-Fisher [32], and use our a priori knowledge that the eigenvector corresponding to λ_1 of Q is $\overline{1}$:

$$\begin{aligned}\lambda_2 &= min_{x:x\perp\overline{1}}\frac{x^TQx}{x^Tx}\\ \lambda_\infty &= max_{x:x\perp\overline{1}}\frac{x^TQx}{x^Tx}\end{aligned} \tag{7.5}$$

Now we apply equation (7.1):

$$\begin{aligned}\lambda_2 = &\quad min_{x:x\perp\overline{1}}\frac{\Sigma_{uv\in E(G)}a_{uv}(x_u-x_v)^2}{\Sigma_u x_u^2}\\ \lambda_\infty = &\quad max_{x:x\perp\overline{1}}\frac{\Sigma_{uv\in E(G)}a_{uv}(x_u-x_v)^2}{\Sigma_u x_u^2}\end{aligned} \tag{7.6}$$

■

The central heuristical result behind the use of Fiedler is that it minimizes the 2-sum. In order to introduce this notion, we first define the minimum p-sum.

DEFINITION 4 (Minimum p-sum) $m_{p,min}(G) := min_\psi m_p(G,\psi) :=$
$min_\psi(\Sigma_{uv\in E(G)}a_{uv}|\psi(u)-\psi(v)|^p)^{\frac{1}{p}}$ *is the minimum p-sum.*
$m_p(G,\psi)$ *is simply the p-sum. [4, 9, 20, 35, 41]*

We are concerned with the minimum 2-sum problem because if an algorithm minimizes the 2-sum, we feel it should also minimize the norm outside the band in some sense (we will make this more precise later). Unfortunately, the minimum 2-sum problem is not solved. However, we can apply a heuristic to show that in certain cases, Fiedler reordering will 'solve' this. In order to show this, we discuss another important theoretical property of the 2-sum. [4, 9, 35]

THEOREM 19 *[35, 46]*

Let ψ be a given labeling of a graph G

$$\lambda_2(G)\frac{n(n^2-1)}{12} \leq m_2(G,\psi)^2 \leq \lambda_\infty(G)\frac{n(n^2-1)}{12} \tag{7.7}$$

Proof Consider

$$\frac{12m_2^2(G,\psi)}{n(n^2-1)} \tag{7.8}$$

If we define a vector x so that $x_i = \psi(i)$, then as discussed above in the definition of the Laplacian, equation (7.1):

$$\frac{12m_2^2(G,\psi)}{n(n^2-1)} = 2n\frac{< Q(G)x, x >}{\frac{1}{6}(n-1)n^2(n+1)} \tag{7.9}$$

Notice that $\Sigma_{u\in V}\Sigma_{v\in V}(\psi(u)-\psi(v))^2 = \Sigma_{i=1}^n\Sigma_{j=1}^n(i-j)^2 = \frac{(n-1)n^2(n+1)}{6}$, thus

$$\begin{array}{rl} \frac{12m_2^2(G,\psi)}{n(n^2-1)} = & 2n\frac{<Q(G)x,x>}{\Sigma_{u\in V}\Sigma_{v\in V}(\psi(u)-\psi(v))^2} \\ = & 2n\frac{<Q(G)x,x>}{\Sigma_{1\leq i,j\leq n}(x_i-x_j)^2} \end{array} \tag{7.10}$$

In order to simplify the denominator, we first note Lagrange's identity for sequences [25]:

$$\Sigma_{1\leq i<j\leq n}(a_ib_j - a_jb_i)^2 = (\Sigma_{i=1}^n a_i^2)(\Sigma_{i=1}^n b_i^2) - (\Sigma_{i=1}^n a_ib_i)^2 \tag{7.11}$$

If we let $a_i = x_i, b_i = 1$, then:

$$\Sigma_{1\leq i,j\leq n}(x_i-x_j)^2 = 2\Sigma_{1\leq i<j\leq n}(x_i-x_j)^2 = 2n(\Sigma_{i=1}^n x_i^2) - 2(\Sigma_{i=1}^n x_i)^2 \tag{7.12}$$

Applying this to equation (7.10):

$$\frac{12m_2^2(G,\psi)}{n(n^2-1)} = \frac{< Q(G)x, x >}{(\Sigma_{i=1}^n x_i^2) - \frac{1}{n}(\Sigma_{i=1}^n x_i)^2} \tag{7.13}$$

And now we apply Fiedler's bound discussed above (7.3) and note that $(\Sigma_{i=1}^{n} x_i) = (x, \overline{1})$, which since in equation (7.3) $x \perp \overline{1}$, this $= 0$, and thus we may take the minimum over all x such that $||x|| = 1$ and apply (7.3) directly to obtain:

$$\frac{12m_2^2(G,\psi)}{n(n^2-1)} \geq \lambda_2(G) \tag{7.14}$$

If we instead take the maximum over all x such that $||x|| = 1$ and apply Fiedler's other bound (7.4)

$$\frac{12m_2^2(G,\psi)}{n(n^2-1)} \leq \lambda_\infty(G) \tag{7.15}$$

Which solves both sides of the inequality.

■

With this result, we may now state why the Fiedler reordering algorithm may be a good approximation to the optimal minimal 2-sum problem:

Example 7.0.7 *[35]*

Let $x_u^{(2)} = u$ then we attain the lower bound in theorem 7 and $m_2(G) = m_2(G, \psi^2)$

Proof

$$m_2(G, \psi_{fiedler})^2 = \Sigma_{uv \in E} a_{uv} (\psi_{fiedler}(u) - \psi_{fiedler}(v))^2 \tag{7.16}$$

Since $x_u^{(2)} = u$, then upon applying Fiedler reordering, $x_u^{(2)} = \psi_{fiedler}(u)$, so by (7.1)

$$m_2(G, \psi_{fiedler})^2 =< L(G)x^{(2)}, x^{(2)} > \tag{7.17}$$

Note that $L(G) \cdot \overline{1} = 0$

$$m_2(G, \psi_{fiedler})^2 =< L(G)(x^{(2)} - \frac{n+1}{2}\overline{1}), (x^{(2)} - \frac{n+1}{2}\overline{1}) > \tag{7.18}$$

$x^{(2)}$ is an eigenvector

$$m_2(G, \psi_{fiedler})^2 = \lambda_2 ||x^{(2)} - \frac{n+1}{2}\overline{1}||^2 = \lambda_2 \Sigma_{i=1}^{n}(i - \frac{n+1}{2})^2 = \lambda_2 \frac{n(n^2-1)}{12} \quad (7.19)$$

Which, as in theorem 7, is the lower bound for m_2.

■

Although this theorem may describe a specific case in which Fiedler reordering solves the minimum 2-sum problem, we have yet to formalize why solving the minimum 2-sum problem will necessary minimize the weight outside the band, or tighten inequality 6.1.

To formalize this, we use $.*$ to denote the Hadamard product and notice the following:

THEOREM 20

$$m_2^2(A. * A) \geq k^2 ||N||_F^2 \quad (7.20)$$

Where k is the band of M.

Proof

$$\begin{aligned} m_2^2(A. * A) &= \Sigma_{uv \in E} a_{uv}^2 |\psi(u)^2 - \psi(v)|^2 \\ &= \Sigma_{|\psi(u)-\psi(v)|>k, uv \in E} a_{uv}^2 |\psi(u) - \psi(v)|^2 + \Sigma_{|\psi(u)-\psi(v)|\leq k, uv \in E} a_{uv}^2 |\psi(u) - \psi(v)|^2 \\ &\geq k^2 ||N||_F^2 + ||M||_F^2 \\ &\geq k^2 ||N||_F^2 \end{aligned} \quad (7.21)$$

■

Therefore, as a heuristic, if we minimize the $2-sum$ on the Hadamard product of A with itself, we minimize the $||N||_F^2$, making the bound on equation 6.1 tighter.

We have stated the theoretical practicality for prefering Fiedler, and for prefering to perform Fiedler reordering based on $A. * A$ instead of A.

Example 7.0.8 *Fiedler v. Fiedler with Hadamard product*

The following are tables of the ratio of the norms of the banded splitting based on using $A. * A$ over using A respectively. I.e. the ratio $\frac{\frac{||M_1||_F^2}{||N_1||_F^2}}{\frac{||M_2||_F^2}{||N_2||_F^2}}$ where M_1, N_1 is generated after reordering using $A. * A$, and M_2, N_2 is generated after reordering using A. Therefore, entries which are > 1 exhibit an improvement, and < 1 exhibit that Fiedler on A is an improvement.

For random matrices, Fiedler is indifferent. Each row denotes a new test, and the band on M increases with the column index.

1.01367890	0.98327923	0.97597267	0.98108477	0.98187655
1.02382458	1.03980065	1.02336333	1.00773469	1.00816982
0.97623039	0.96404522	0.98467241	0.99594981	0.98809953
0.97476808	0.97995002	0.99117464	0.99666846	0.99288300
1.00313377	1.00202641	1.00285440	1.00517144	1.00535496
1.00376078	0.99546630	0.99399239	0.98974273	0.99586203
1.01553252	0.99817853	0.98772906	0.99065682	1.00082871
0.97640874	0.98936266	0.99380882	1.00518412	1.00594349
0.98423665	0.98268294	0.97379794	0.98052245	0.98024990
1.00669753	0.99506844	0.99373479	1.00566246	1.00323247

The results are more stark for some structured matrices, the following are for H-matrices (a matrix A is an H-matrix iff $|A|$ is an M-matrix, a matrix A is an M-matrix iff $A \geq 0$ and A^{-1} has positive diagonal entries and all other entries ≤ 0 [1, 61]):

9.16060078	9.05846885	8.96060377	8.86664360	8.77634841
9.17742213	9.07535723	8.97782361	8.88463509	8.79481588
9.07408310	8.97183841	8.87318352	8.77913526	8.68847686
9.07716016	8.97404380	8.87542453	8.78058420	8.69003904
9.14021404	9.03808256	8.94013399	8.84629755	8.75712904
9.17217142	9.07002305	8.97212042	8.87874805	8.78972264
9.25558807	9.15445936	9.05870793	8.96601325	8.87660927
9.11227978	9.01019182	8.91239345	8.81831376	8.72821406
9.11292266	9.01060356	8.91305532	8.81977192	8.73045933
9.07118181	8.96795405	8.86929883	8.77523760	8.68447626

This concludes the theoretical discussion of the benefits of using a modified Fiedler reordering. One critical aspect of the above exposition is that all the above results are essentially theoretically heuristical. However, if the goal is to create a test which is of linear time (at least less than the amount of time necessary to solve the original system), then this permits simple a posteriori tests to the utility of using such a method to improve convergence of the overall algorithm. Put differently, because a modified Fiedler reordering may be calculated in linear time, one can simply perform an a posteriori check that the M and N norms satisfy 6.1.

However, this necessitates a discussion of how to achieve such minimal computation time. Because of this, we will now conclude this section with a brief overview on the TRACEMIN-Fiedler algorithm.

We seek to solve

$$Lx = \lambda x(\text{L denotes the weighted Laplacian}) \tag{7.22}$$

For the Fiedler vector x_2. We do this using TRACEMIN.

We know that:

$$min_{Y^TY=I} tr(Y^TAY) = \Sigma_{i=1}^{p}\lambda_i \tag{7.23}$$

Therefore, if we let X_k be an approximation where $X_k^T L X_k = \sigma_k, X_k^T X_k = I, \sigma_k = diag(\rho_1^{(k)}, \rho_2^{(k)}, \cdots, \rho_p^{(k)})$, we can update the approximation by noting that if we find Δ_k so that

$$\begin{aligned} &\Delta_k \text{ minimizes } tr(X_k - \Delta_k)^T A(X_k - \Delta_k) \\ &X_k^T \Delta_k = 0 \end{aligned} \tag{7.24}$$

Then

$$\begin{aligned} &tr(X_k - \Delta_k)^T A(X_k - \Delta_k) < tr X_k^T A X_k \\ &(X_k^T - \Delta_k)^T (X_k - \Delta_k) = I \end{aligned} \tag{7.25}$$

[54]

Finding such a minimum is equivalent to solving the saddle point problem:

$$\begin{pmatrix} L & X_k \\ X_k^T & 0 \end{pmatrix} \begin{pmatrix} \Delta_k \\ N_k \end{pmatrix} = \begin{pmatrix} LX_k \\ 0 \end{pmatrix} \tag{7.26}$$

[55]

This results in the following algorithm:

ALGORITHM 10 (TRACEMIN-Fiedler)

for $k = 1, 2, \cdots, max_i t$ *do*

1. *Orthonormalize* X_k *to* V_k
2. $H_k := V_k^T L V_k$
3. *Find* (Y_k, Σ_k) *the eigenvectors and eigenvalues of* H_k *in ascending order.*
4. $X_k := V_k Y_k$
5. *If* $\frac{||Lx_k - X_k \Sigma_k||_\infty}{||L||_\infty}$ *is less than a predefined tolerance for a vector, move this into* X_{conv}, *set* $n_{conv} = n_{conv} + 1$, *and when* $n_{conv} \geq p$, *stop.*
6. *Deflate, if* $n_{conv} > 1$, $X_k = X_k - X_{conv}(X_{conv}^T X_k)$
7. *if* $k = 1$ *then*

Solve $\hat{L} W_k = X_k$ *via PCG using diagonal preconditioner* $\hat{D}$, *where* $\hat{L}, \hat{D}$ *are* L, D

only perturbed by an additional $+||L||_\infty 10^{-12} I$.

else

Solve $LW_k = X_k$ *via PCG using* D.

end if

8. $S_k = X_k^T W_k$

9. Solve $S_k N_k = X_k^T X_k$ *for* N_k *directly (this is a small system).*

10. $X_{k+1} = X_k - \Delta_k = W_k N_k$.

[38]

With two minor improvements in step 7 to ensure positive semi-definiteness, and in the deflation process in step 6. It should also be noted that the size of the system in step 9 is determined by the chosen number of vectors to be kept in X_k [38]. Further, this has important implications for the parallelism of the overall algorithm, which will be discussed in the next chapter.

8. BANDED FGMRES-RPM WITH SPECTRAL REORDERING

To bring together the diverse topics discussed in previous chapters, we outline our proposed algorithm in total, bring together and summarize our convergence results of the preceding chapters before ending with some numerical experiments.

The algorithm is as follows:

ALGORITHM 11 (Banded FGMRES-RPM with Spectral Reordering)

1. *Perform TRACEMIN-Fiedler as described in algorithm 10 on the symmetric part of the matrix* $|A. * A|$.
2. *If computationally feasible, choose a band large enough to satisfy (6.1). Else choose largest computationally feasible band.*
3. *Use band described in previous step to split* $A = M - N$.
4. *Perform algorithm 8 on original matrix* A *with a given number of outer steps* k.
5. *In step 3 of algorithm 8, use algorithm 3 as a preconditioner.*
6. *For* M, N, H *in algorithm 3, use step 3.*
7. *The other parameters in algorithm 3 should be set as follows: the eigenvalue deflation bound should be set as* $m = \sqrt{k}$*, the number of eigenvalues to deflate at each step should be set as* 2*, frequency should kept at* 1 *or* 2*, the size of the subspace to be iterated should be set as* m*, the number of steps of RPM should be also set as* m.

With this background, we have introduced a number of individual properties of RPM and FGMRES that can now be used in the analysis of the nesting of these procedures as we are doing in the complete outline of the algorithm above. Although we are looking at a nested iteration where-in the stiff subspace contains slight perturbations upon each iteration, it should be noted that a FGMRES-RPM nested iteration will converge (this result is similar to [22, 37, 42], except the outer iteration is not Richardson), as shown as follows:

THEOREM 21 *FGMRES RPM Converges if the Jacobian of RPM has spectra* < 1

Proof The RPM iteration can be expressed via $Je_k = e_{k+1}, e_k = \begin{pmatrix} Px_k \\ Qx_k \end{pmatrix}$ as discussed above.

As discussed previously in theorem 4

k steps of RPM is equivalent to preconditioning by

$$M^{-1} = (I, I)J^k \begin{pmatrix} P \\ Q \end{pmatrix} + (I, I)(I - J^k) \begin{pmatrix} P \\ Q \end{pmatrix} A^{-1} \tag{8.1}$$

We note that the above preconditioned matrix has a symmetric part which is positive definite.

$$\frac{1}{2}(M^{-1}A + (M^{-1}A)^T) \tag{8.2}$$

We apply Rayleigh Ritz [32] as shown below.

$$
\begin{gathered}
\tfrac{1}{2}(x^T M^{-1}Ax + x^T(M^{-1}A)^T x) \\
= \\
\tfrac{1}{2}(x^T(I,I)J^k(P^T,Q^T)^T Ax + x^T A^T(P,Q)(J^T)^k(I,I)^T x \\
+x^T((I,I)(I-J^k)(P^T,Q^T)^T)x + x^T((P,Q)(I-(J^T)^k)(I,I)^T x) \\
= \\
\tfrac{1}{2}(x^T(I,I)J^k(P^T,Q^T)^T Ax + x^T A^T(P,Q)(J^T)^k(I,I)^T x \\
-x^T((I,I)J^k(P^T,Q^T)^T)x - x^T((P,Q)(J^T)^k(I,I)^T x) \\
+2x^T((I,I)(P^T,Q^T)^T)x)
\end{gathered}
\tag{8.3}
$$

The first four components $\to 0$ given that $\rho(J) < 1$ (which is true given that RPM converges), and the last component is identically 2.

So, for sufficiently large k, $(x, \frac{1}{2}(M^{-1}A + (M^{-1}A)^T)x) > 0$, is positive definite.

Therefore, even considering the slight numerical perturbations, by theorem 13 the outer FGMRES step converges.

■

The problem with removing "the *smallest* eigenvalues of A" [13] (emphasis added) are due to the inner nesting of a Richardson iteration (as compared to just applying deflation with GMRES or preconditioned GMRES as in [6, 8, 13]), and the guarantee of positive definiteness.

The central convergence result is theorem 21, which states that if the subspace deflated in algorithm 3 is large enough, that this ensures convergence. However, as discussed in the previous chapter, equation 6.1 and theorems 20 and 7.0.7 imply that the Fiedler step will further improve performance.

Although more importantly, and finally referencing the title, is that each of the key three subroutines of the algorithm (algorithms 10, 8, and 3) all include aspects which improve domains of convergence. We have already stated how algorithm 10 improves the spectra of the iteration matrix inside algorithm 3. However, algorithm 3 reduces the number of eigenvalues outside the unit disk, as seen by theorem 5.

Likewise, algorithm 8 extends the convergence domain as discussed in theorem 9. Again, the central idea behind each algorithm is that they improve convergence by in some way extending the convergence domain in the underlying algorithm.

Further, each algorithm chosen is highly parallelizable. Algorithm 10 [38] and the SPIKE algorithm used in backsolving the underlying preconditioner [39,45] inside algorithm 3 are both highly parallelizable algorithms. Algorithm 8 is also already a highly parallelizable algorithm due to its critical use of matrix-vector products. Finally, the minor addition of making algorithm 3 use subspace iteration was for this very purpose of parallelizability in the overall algorithm as well.

The only item that remains to be discussed that has not been described previously is to explain the rationale for the parameter choices in expressed in step 7.

The eigenvalue deflation bound is based on a consideration of cost, since ideally the eigenvalue deflation bound should be set as large as possible. However, since the most feasible application is to use this algorithm because of a failure from simply increasing the number of steps of GMRES, then this suggests a bound for the eigenvalue deflation bound. In particular, since k steps of GMRES is roughly $\mathcal{O}(kn)$, the reorthogonalization step in the subspace iteration would cost $\mathcal{O}(m^2n)$, and $m << n$ so that the cost of the eigenvalue problem inside RPM is marginal; then in order to force the cost of RPM to not exceed the order GMRES step we set $m = \sqrt{k}$.

The number of eigenvalues to deflate at each step is set as 2 because if it were larger, then it would necessitate counterbalancing higher frequency, more steps of RPM in the inner iteration, or an even larger subspace size. If it were smaller, then it introduces numerical issues from the necessity of deflating out complex eigenvalues pairs.

The frequency is determined as part of a balance between frequency and the subspace size. Part of this balance needs to take into account the number of nodes (if the algorithm is implemented in parallel) and the distribution of the spectra of the iteration matrix. However, one of the advantages to utilizing a subspace size that is larger than the number of eigenvalues to be iterated is that further iterations smooth

out errors apparent earlier in the iterative deflation process. Therefore, subspace size is to be preferred over frequency (this assumes that the spectra of the matrix is widely distributed enough that subspace iteration is a significant advantage over the power method). Thus, frequency should be kept at 1 or 2. The subspace size is then to be bounded by a consideration of a cost comparison to the outer GMRES steps, as was done with the bound on the number of eigenvalues to be deflated, which is $m = \sqrt{k}$.

This leaves an analysis of the description for the number of RPM steps. Ideally, this would depend on the particular spectral distribution of the iteration, like we mentioned with respect to the balance between the frequency and subspace size. We also set the number of steps of RPM to be the same m as the subspace size and bound of the number of eigenvalues to be deflated in order that there are enough successive iterations to accurately compute up to the bound of the number of eigenvalues. However, as per theorem 5, ideally either the number of RPM steps should be set at 0 depending on whether m is large enough to deflate out all eigenvalues larger than 1 in norm. However, upon assumption that a full calculation or precise a priori knowledge of the spectra of the iteration matrix would be computationally infeasible, and that projection methods are computationally preferrable to Richardson-like methods; then the ideal practical application would be linear systems in which GMRES is failing (as noted above), but for which a modest amount of deflation may ensure convergence. E.g., solving a set of linear systems dependent on some parameter l, $A(l)x = b$, under which l undergoes a small perturbation at each step. In this way, we can be assured that if, for example, vanilla GMRES converges for $l < a$, then at $l = a + \epsilon$, the spectra of the iteration matrix associated to $A(l + \epsilon)$ will have only a modest number of eigenvalues necessary to deflate.

Such an example is with the frequency parameter in solving Helmholtz problems, which we exhibit below by comparison with the standard comparison with Poisson in examples 8.0.9 and 8.0.10 below. We will conclude this section by further exhibiting some other examples along with other numerical experiments.

In all the following examples we compare restarted GMRES (in blue) with restarted GMRES-RPM (in green) as previously described.

Example 8.0.9 *Poisson*

In this example, we look at the canonical toy example for numerical experiments: the Poisson matrix. The following is a size $n = 1000000$ Poisson matrix where the preconditioner band was chosen as 10, the number of GMRES steps was 40, and the number of restarts was also 40.

The overall convergence pattern matches that of just restarted GMRES.

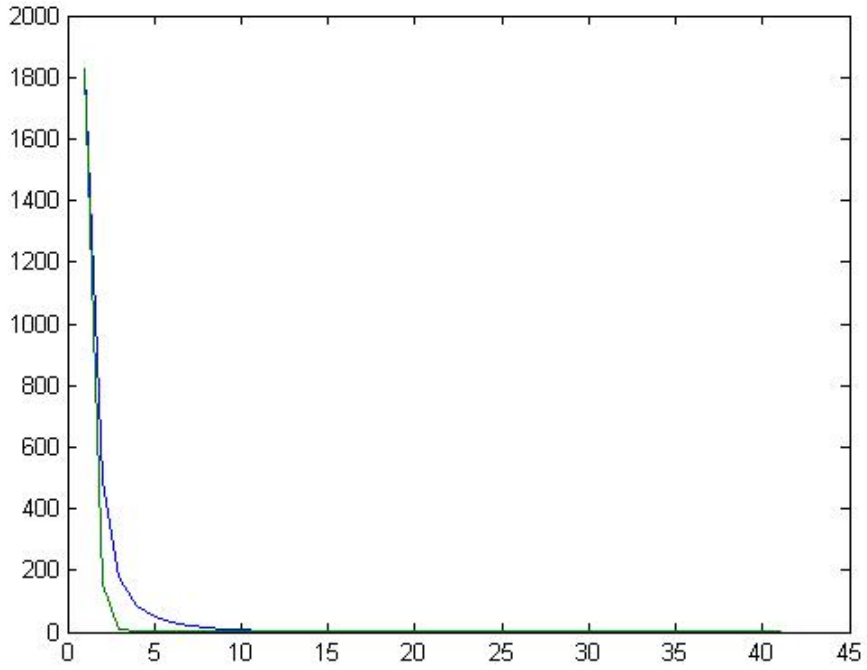

Figure 12. GMRES-RPM on Poisson

However, the actual gain in accuracy was roughly two order of magnitudes.

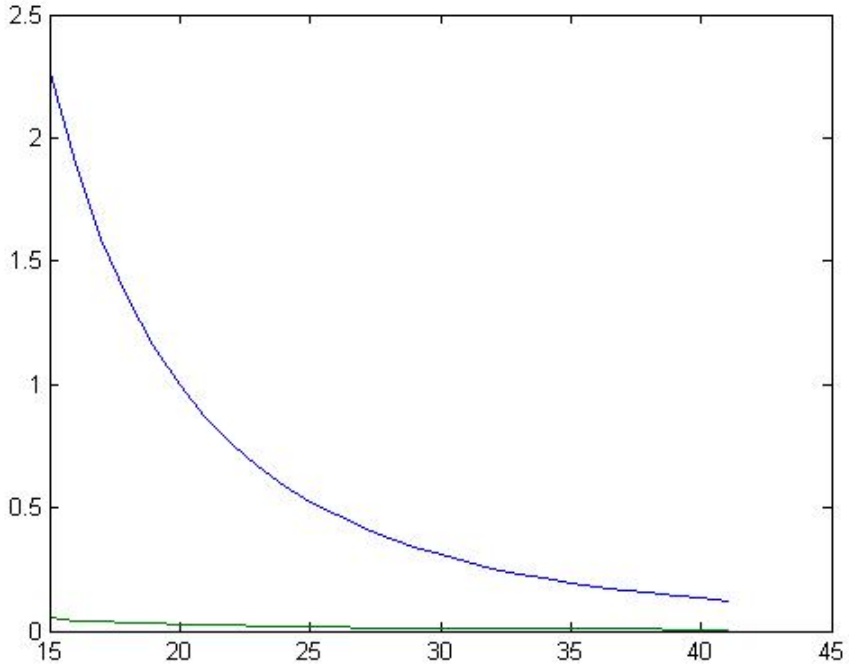

Figure 13. GMRES-RPM on Poisson blowup

Example 8.0.10 *Helmholtz*

Here we adjust the previous Poisson matrix into a Helmholtz matrix with the frequency set to a low $\sqrt{5}$. The size is still $n = 1000000$, the preconditioner band was chosen as 10, the number of GMRES steps was 40, and the number of restarts was also 40.

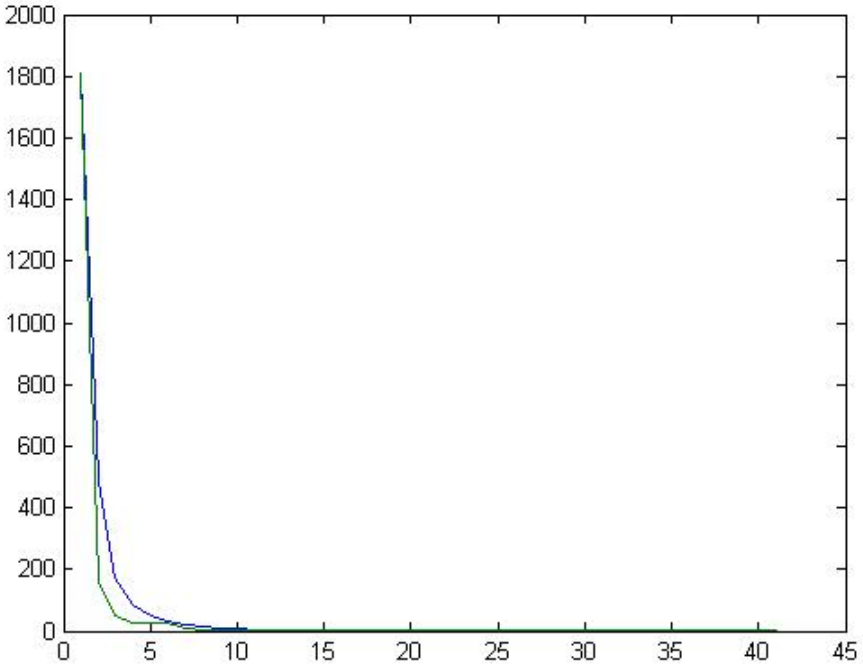

Figure 14. GMRES-RPM on Helmholtz

Likewise, there was a similar gain in accuracy, although not as significant as in Poisson. However, this behavior does exhibit that such a procedure can extend the domain of convergence over ordinary restarted GMRES.

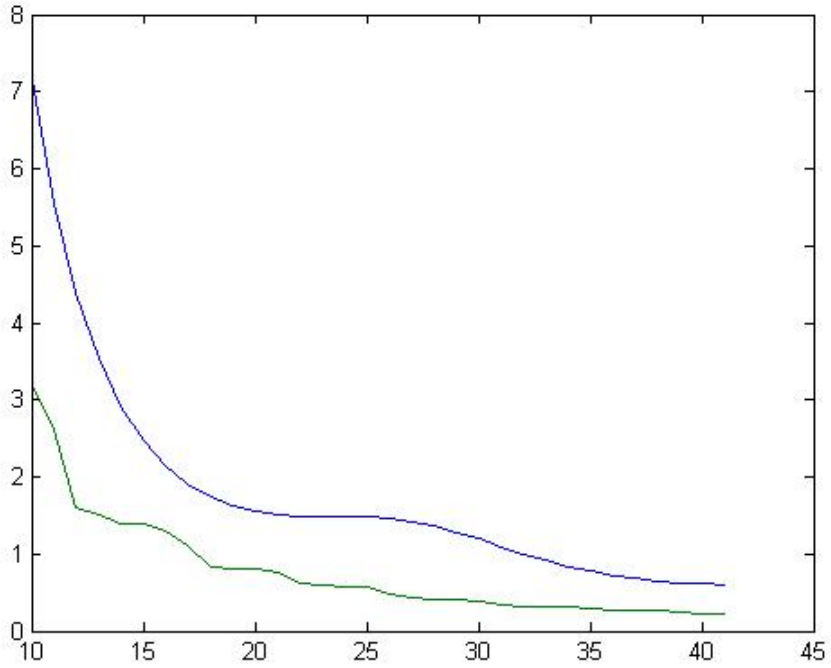

Figure 15. GMRES-RPM on Helmholtz blowup

Example 8.0.11 *nd3k*

Our next two examples come from the Tim Davis matrix collection.

The following is a 3D sparse problem 'nd3k', size 9000 spd matrix, with roughly $3E6$ nonzero entries. The preconditioner band was chosen at 200, the number of GMRES steps at 30, and the number of restarts at 30.

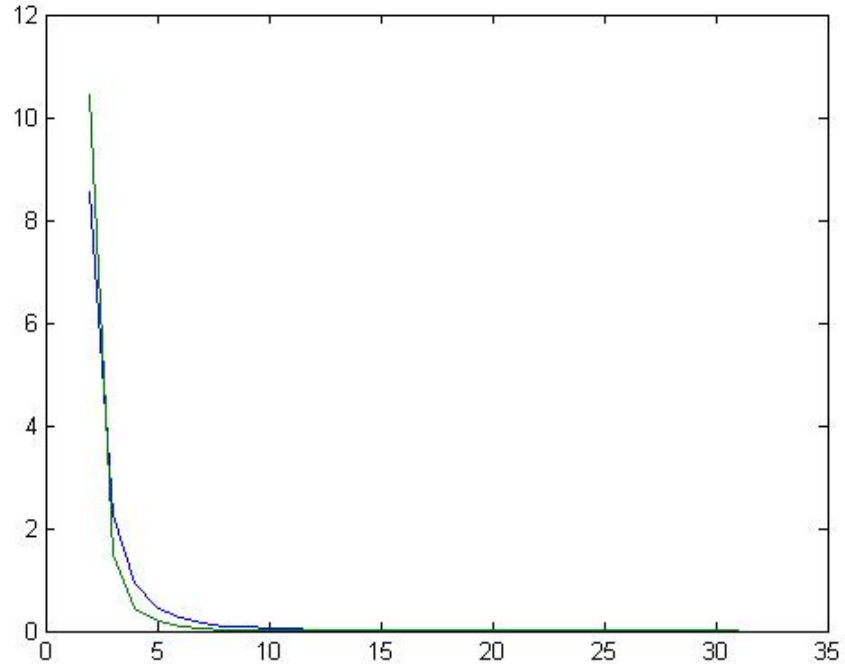

Figure 16. GMRES-RPM on nd3k

A similar pattern convergence pattern continues throughout:

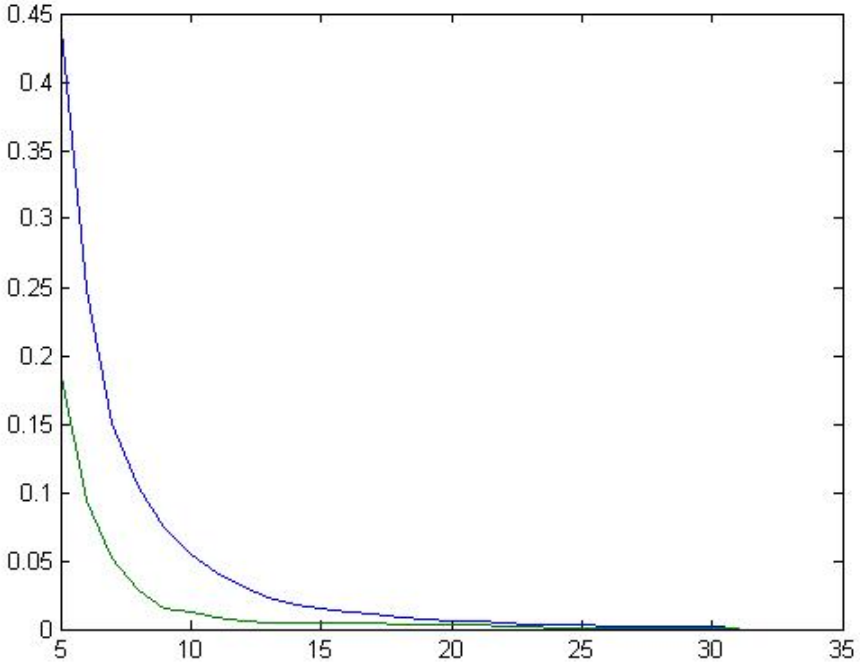

Figure 17. GMRES-RPM on nd3k blowup

Example 8.0.12 *Thermal*

The following is a steady-state thermal problem 'thermal2', size $1,228,045$, with roughly $8.5E6$ nonzero entries. The preconditioner band was chosen at 10, the number of GMRES steps at 10, and the number of restarts at 30.

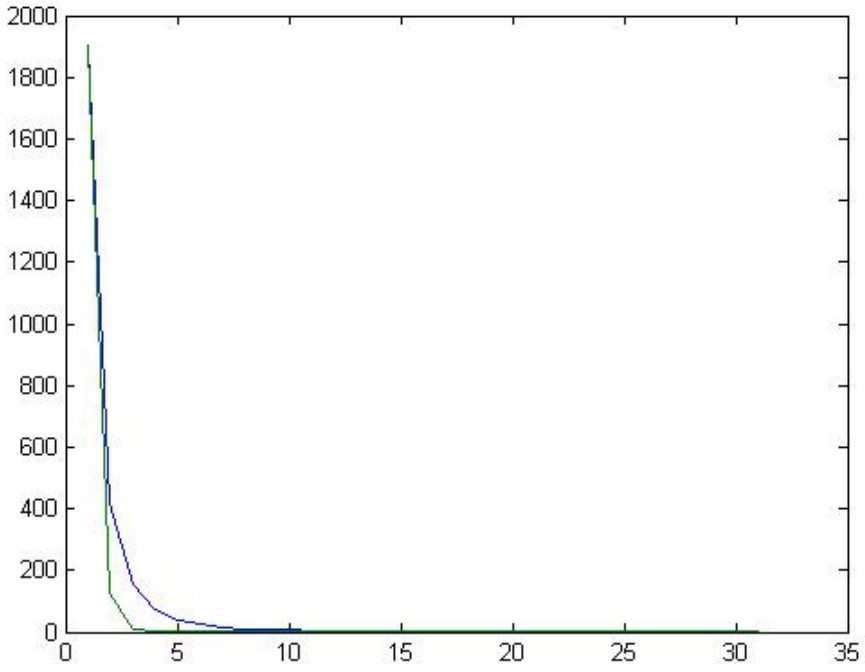

Figure 18. GMRES-RPM on thermal2

Not only does a similar convergence patten continue throughout, but the gain in accuracy is similar as with Poisson.

9. SUMMARY

We have analyzed various methods that can expand domains of convergence for iterative linear system solvers, and combined these methods to create a parallel algorithm that utilizes deflation, adaptive GMRES, multiple nesting layers, spectral reordering, and banded preconditioning. We made use of spectral reordering to improve the banded preconditioner. We used this banded preconditioner to decrease the amount of deflation necessary. The deflation was in turn nested inside of an adaptive GMRES scheme to guarantee convergence.

In our analysis of RPM, we traced the theoretical dependence of this algorithm from nonlinear stabilization procedures to the specific application of Richardson methods. This in turn created a simplified expression for the Jacobian of the algorithm. From this, we developed a new preconditioner expression and convergence results for RPM. With this background, we introduced a new block version of RPM, and illustrated a parallel implementation of this algorithm.

Our discussion of adaptive GMRES, and FGMRES in particular, outlined the dependence of FGMRES convergence results from underlying GMRES convergence results. This included showing that FGMRES is in fact identical to GMRES on certain systems. We illustrated the robustness of GMRES algorithms, particularly on systems with symmetric part positive definite. Moreover, we were able to develop new results which suggested a relation between the residual norm of FGMRES and the residual norms of the individually preconditioned GMRES iterations.

We noted the relation between spectral reordering and banded preconditioning. We illustrated how spectral reordering concentrates the heaviest elements of a matrix along the central band, and introduced a new procedure to improve this reordering in theory and practice. Using this result and Stein's theorem, we were able to show how this improves our choice for a banded preconditioner.

Our analysis from RPM and FGMRES were consequentially used to show that upon nesting convergence can be guaranteed. Our analysis from spectral reordering and RPM were used to show that upon banded preconditioning, the performance of RPM is enhanced. This resulted in a robust algorithm with the ability to be applied to a wide range of linear systems, or to improve currently existing linear system solvers.

APPENDICES

A. FGMRES DECOMPOSITION LEMMA

In this appendix, we restate and prove the lemma required to analyze the preconditioner dependence of FGMRES.

LEMMA 4 *Assume that* $||M_i^{-1} - M_j^{-1}|| \leq \epsilon$.

Let the initial vector be given as x_0 *and* $r_0 := b - Ax_0$.

Let x_k *be the solution after* k *steps of FGMRES* ($x_k \neq x$) *and* H_k *be nonsingular.*

Let $a_1 = M_1^{-1} r_0$, *and define inductively* $a_k = \Sigma_{j=1}^{k} M_j^{-1} (\Sigma_{i=1}^{k-1} \alpha_{i,j,k} M_i a_i + \gamma_{k-1,j} A a_{k-1})$ *where* $\alpha_{i,j,k}, \gamma_{k-1,j}$ *are given.*

Define the Y*-matrix so that* $Y a_i = a_{i+1}$.

Let y_k *be the solution after* k *steps of GMRES on* Y *with* $M_1^{-1} r_0$ *in place of* r_0.

Then $||x_k - y_k|| \leq C_k \epsilon$ *for some constant* C_k *or* $||b - Ax_k|| \leq ||b - Ay_k||$.

Proof We show this inductively.

$k = 1$:

Note that x_1 minimizes the residual over $x_0 + Span(M_1^{-1} r_0)$ by theorem 10 and y_1 minimizes the residual over $x_0 + Span(M_1^{-1} p(r_0)) = x_0 + Span(\alpha_1 M_1^{-1} r_0) = x_0 + Span(M_1^{-1} r_0)$ by lemma 1.

Thus $x_1 = y_1$.

$k = N$:

Assume that $||x_i - y_i|| \leq C_i \epsilon, \forall 1 < i < N$.

Note that $x_i - x_0$ is a linear combination of $z_1, z_2, \cdots$ up to z_i (where z_i are the same z_i vectors in the FGMRES algorithm).

Thus x_k minimizes the residual over $x_0 + Span(Z_k) = x_0 + Span(\{x_i - x_0 | 1 < i < N\} \cup z_k) = x_0 + Span(X_1)$ (by theorem 10 above).

Note that $y_i - x_0$ is a linear combination of $a_1, a_2, \cdots$ up to a_i.

Thus y_k minimizes the residual over $x_0 + Span(z_0, Yz_0, \cdots, Y^{i-1}z_0) = x_0 + Span(\{y_i - x_0 | 1 < i < N\} \cup a_k) = x_0 + span(X_2)$ (by lemma 1).

By algorithm 8 above, we have

$$z_k = M_k^{-1}(\Sigma_{i=1}^{k-1}\beta_i M_i z_i + \zeta_{k-1} A z_{k-1}). \tag{A.1}$$

And because the spans are the same as shown above, then under a different linear combination

$$z_k = M_k^{-1}(\Sigma_{i=1}^{k-1}\beta_i' M_i (x_i - x_0) + \zeta_{k-1}' A(x_{k-1} - x_0)) \tag{A.2}$$

Furthermore, by hypothesis of a_k as a polynomial expression of the prior a_i:

$$a_k = \Sigma_{j=1}^{k} M_j^{-1}(\Sigma_{i=1}^{k-1}\alpha_{i,j,k} M_i a_i + \gamma_{k-1,j} A a_{k-1}) \tag{A.3}$$

And because the spans are the same, then under a different linear combination

$$a_k = \Sigma_{j=1}^{k} M_j^{-1}(\Sigma_{i=1}^{k-1}\alpha_{i,j,k}' M_i (y_i - x_0) + \gamma_{k-1,j}' A(y_{k-1} - x_0)) \tag{A.4}$$

Note that $\gamma_{k-1,j}' \neq 0$ as when this $= 0$ this corresponds to the breakdown case of FGMRES, which can not happen since H_k is nonsingular. Now define z_k' as:

$$z_k' = \Sigma_{j=1}^{k}(\frac{\gamma_{k-1,j}'}{\zeta_{k-1}'} z_k - \Sigma_{i=1}^{k-1}\frac{\gamma_{k-1,j}'\beta_i'}{\zeta_{k-1}'}(x_i - x_0) + \Sigma_{i=1}^{k-1}\alpha_{i,j,k}'(x_i - x_0)) \tag{A.5}$$

Further, if $||M_i^{-1} - M_j^{-1}||_2 \leq \epsilon$ then for any vectors $\overline{p}, \overline{q}$

$$\begin{aligned} ||\overline{p} + M_j^{-1} M_i \overline{q}|| &= ||\overline{y} + (M_j^{-1} - M_i^{-1} + M_i^{-1})\overline{q}|| \\ &\leq ||\overline{p} + \overline{q}|| + \epsilon||\overline{q}|| \\ &= ||\overline{p} + \overline{q}|| + C||\epsilon|| \end{aligned} \tag{A.6}$$

We will use A.6 repeatedly in what follows.

Then, also recalling that $||x_i - y_i|| \leq C_i\epsilon$:

$$||z'_k - a_k||_2 \;=\; ||\Sigma_{j=1}^{k}(\tfrac{\gamma'_{k-1,j}}{\zeta'_{k-1}} z_k - \Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}}(x_i - x_0) + \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0)) - a_k||_2 \tag{A.7}$$

By equation A.2:

$$\begin{aligned} &= ||\Sigma_{j=1}^{k}(\tfrac{\gamma'_{k-1,j}}{\zeta'_{k-1}} M_k^{-1}(\Sigma_{i=1}^{k-1}\beta'_i M_i(x_i - x_0) + \zeta'_{k-1}A(x_{k-1} - x_0)) \\ &\quad -\Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}}(x_i - x_0) + \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0)) - a_k||_2 \end{aligned} \tag{A.8}$$

By equation A.4:

$$\begin{aligned} = \quad & ||\Sigma_{j=1}^{k}((M_k^{-1}(\Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}} M_i(x_i - x_0) \\ &+\gamma'_{k-1,j}A(x_{k-1} - x_0)) - \Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}}(x_i - x_0) + \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0)) \\ &-M_j^{-1}(\Sigma_{i=1}^{k-1}\alpha'_{i,j,k}M_i(y_i - x_0) + \gamma'_{k-1,j}A(y_{k-1} - x_0)))|| \end{aligned} \tag{A.9}$$

Using A.6:

$$\begin{aligned} \leq \quad & ||\Sigma_{j=1}^{k}(((\Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}}(x_i - x_0) \\ &+\gamma'_{k-1,j}M_k^{-1}A(x_{k-1} - x_0)) - \Sigma_{i=1}^{k-1}\tfrac{\gamma'_{k-1,j}\beta'_i}{\zeta'_{k-1}}(x_i - x_0) + \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0)) \\ &-(\Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(y_i - x_0) + \gamma'_{k-1,j}M_j^{-1}A(y_{k-1} - x_0)))|| + C\epsilon \end{aligned} \tag{A.10}$$

Algebraic simplification:

$$\begin{aligned} &= ||\Sigma_{j=1}^{k}(((\gamma'_{k-1,j}M_k^{-1}A(x_{k-1} - x_0)) + \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0)) \\ &\quad -(\Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(y_i - x_0) + \gamma'_{k-1,j}M_j^{-1}A(y_{k-1} - x_0)))|| + C\epsilon \end{aligned} \tag{A.11}$$

Triangle inequality:

$$\begin{aligned} &\leq \Sigma_{j=1}^{k}(||\gamma'_{k-1,j}M_k^{-1}A(x_{k-1} - x_0) - \gamma'_{k-1,j}M_j^{-1}A(y_{k-1} - x_0)|| \\ &\qquad +||\Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(x_i - x_0) - \Sigma_{i=1}^{k-1}\alpha'_{i,j,k}(y_i - x_0)||) + C\epsilon \\ &\leq \Sigma_{j=1}^{k}(||\gamma'_{k-1,j}M_k^{-1}A(x_{k-1} - x_0) - \gamma'_{k-1,j}M_j^{-1}A(y_{k-1} - x_0)|| \\ &\qquad +\Sigma_{i=1}^{k-1}|\alpha'_{i,j,k}|||x_i - y_i||) + C\epsilon \end{aligned} \tag{A.12}$$

Induction hypothesis:

$$\leq \quad \Sigma_{j=1}^{k}(||\gamma'_{k-1,j}M_k^{-1}A(x_{k-1}-x_0)-\gamma'_{k-1,j}M_j^{-1}A(y_{k-1}-x_0)||)+C\epsilon \tag{A.13}$$

Matrix norm:

$$\leq \quad \Sigma_{j=1}^{k}(||M_k^{-1}||||\gamma'_{k-1,j}A(x_{k-1}-x_0)-\alpha'_{k-1,j,k}M_kM_j^{-1}A(y_{k-1}-x_0)||)+C\epsilon \tag{A.14}$$

A.6 again:

$$\leq \quad \Sigma_{j=1}^{k}(||M_k^{-1}||||\gamma'_{k-1,j}A(x_{k-1}-y_{k-1})||)+C\epsilon \tag{A.15}$$

Induction hypothesis:

$$\leq \quad C\epsilon \tag{A.16}$$

Therefore, under the assumption that $\gamma'_{k-1,j} \neq 0$ for some j, then $span(X_1) = span(X'_1)$ where the last column of X'_1 is z'_k and $||z'_k - a_k||_2 \leq C\epsilon$.

But since $||(y_i - x_0) - (x_i - x_0)|| \leq C_i\epsilon$, then $||X_1 - X_2||_2 \leq ||X_1 - X_2||_F \leq nC_X\epsilon$

Then by Wedin [11, 30, 62], we know that the forward error for a linear least square problem is $||x_k - y_k||_2 \leq (1 + 2\kappa_2(X_1))nC_X\epsilon = C_k\epsilon$.

Should $\gamma'_{k-1,j} = 0 \forall j$, then this corresponds to $span(X_1) \supsetneq span(X'_1)$ where the last column of X'_1 is z'_k and $||z'_k - a_k|| \leq C\epsilon$. Therefore, if we let x'_k correspond to the solution using X'_1, then similar to the previous line $||x'_k - y_k||_2 \leq C_k\epsilon$, and $||b - Ax_k|| \leq ||b - Ax'_k|| = ||b - Ay_k + A(y_k - x'_k)|| \leq ||b - Ay_k|| + C_k\epsilon$.

■

LIST OF REFERENCES

LIST OF REFERENCES

[1] Axelsson, O. (1996). *Iterative solution methods*. Cambridge University Press.

[2] Baglama, J., Calvetti, D., Golub, G. H., & Reichel, L. (1998). Adaptively preconditioned GMRES algorithms. *SIAM Journal on Scientific Computing*, 20(1), 243-269.

[3] Baggag, A. (2012). A Preconditioned Scheme for Nonsymmetric Saddle-Point Problems. In *High-Performance Scientific Computing* (pp. 219-250). Springer London.

[4] Barnard, S. T., Pothen, A., & Simon, H. D. (1993, December). A spectral algorithm for envelope reduction of sparse matrices. In *Proceedings of the 1993 ACM/IEEE conference on Supercomputing* (pp. 493-502). ACM.

[5] Buck, C. R. (1956). *Advanced Calculus*. McGraw-Hill Book Company.

[6] Burrage, K., & Erhel, J. (1998). On the performance of various adaptive preconditioned GMRES strategies. *Numerical linear algebra with applications*, 5(2), 101-121.

[7] Burrage, K., Erhel, J., Pohl, B., & Williams, A. (1998). A deflation technique for linear systems of equations. *SIAM Journal on Scientific Computing*, 19(4), 1245-1260.

[8] Chapman, A., & Saad, Y. (1997). Deflated and augmented Krylov subspace techniques. *Numerical linear algebra with applications*, 4(1), 43-66.

[9] Chung, F.R.K. (1988). *Graph Theory*. Academic Press Limited.

[10] Dahlquist, G., & Bjorck, A. (2008). *Numerical Methods in Scientific Computing*. SIAM.

[11] Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. SIAM.

[12] Drkoov, J., Greenbaum, A., Rozlonk, M., & Strako, Z. (1995). Numerical stability of GMRES. *BIT Numerical Mathematics*, 35(3), 309-330.

[13] Erhel, J., Burrage, K., & Pohl, B. (1996). Restarted GMRES preconditioned by deflation. *Journal of computational and applied mathematics*, 69(2), 303-318.

[14] Erhel, J., & Frdric, G. H. (1997). An augmented subspace conjugate gradient.

[15] Nuentsa Wakam, D., & Erhel, J. (2011). Parallelism and robustness in GMRES with the Newton basis and the deflated restarting.

[16] Fiedler, M. (1974). An Algebraic Approach to the Connectivity of Graphs. In *Recent Advances in Graph Theory*. Paper presented at the Proceedings of the Composium Held in Prague.

[17] Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4), 619-633.

[18] Frommer, A., & Szyld, D. B. (1992). H-splittings and two-stage iterative methods. *Numerische Mathematik*, 63(1), 345-356.

[19] Gautschi, W. (1997). *Numerical Analysis: an Introduction*. Birkhauser.

[20] George, A., & Pothen, A. (1997). An analysis of spectral envelope reduction via quadratic assignment problems. *SIAM Journal on Matrix Analysis and Applications*, 18(3), 706-732.

[21] Gmati, N., & Philippe, B. (2008). Comments on the GMRES convergence for preconditioned systems. In *Large-Scale Scientific Computing* (pp. 40-51). Springer Berlin Heidelberg.

[22] Golub, G. H., & Overton, M. L. (1982). *Convergence of a two-stage Richardson iterative procedure for solving systems of linear equations* (pp. 125-139). Springer Berlin Heidelberg.

[23] Golub, G. H., & Van Loan, C. F. (2012). *Matrix computations* (Vol. 3). JHUP.

[24] Golub, G. H., & Ye, Q. (1999). Inexact preconditioned conjugate gradient method with inner-outer iteration. SIAM Journal on Scientific Computing, 21(4), 1305-1320.

[25] Graham, R. L., Knuth, D. E., & Patashnik, O. (1989). *Concrete mathematics: a foundation for computer science* (Vol. 2). Reading: Addison-Wesley.

[26] Graves, L. M.. (1946). *The Theory of Functions of Real Variables*. McGraw-Hill Book Company.

[27] Greenbaum, A. (1997). *Iterative Methods for Solving Linear Systems*. SIAM.

[28] Godsil, C. and G. R. (2001). *Algebraic Graph Theory*. Springer-Verlag New York, Inc..

[29] Guattery, S., & Miller, G. L. (1995, January). On the performance of spectral graph partitioning methods. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms* (pp. 233-242). Society for Industrial and Applied Mathematics.

[30] Higham, Nicholas J. (2002). *Accuracy and Stability of Numerical Algorithms*. SIAM.

[31] Hildebrandt, T. H., & Graves, L. M. (1927). Implicit functions and their differentials in general analysis. *Transactions of the American Mathematical Society*, 29(1), 127-153.

[32] Horn, R. A., and Charles R. J. (1985). *Matrix Analysis*. Cambridge University Press.

[33] Householder, A. S. (1953). *Principles of Numerical Analysis*. McGraw-Hill Book Company.

[34] Householder, A. S. (1964). *The Theory of Matrices in Numerical Analysis*. Dover Publications.

[35] Juvan, M., & Mohar, B. (1992). Optimal linear labelings and eigenvalues of graphs. *Discrete Applied Mathematics*, 36(2), 153-168.

[36] Kelley, C.T. (1995). *Iterative Methods for Linear and Nonlinear Equations*. SIAM.

[37] Lanzkron, P. J., Rose, D. J., & Szyld, D. B. (1990). Convergence of nested classical iterative methods for linear systems. *Numerische Mathematik*, 58(1), 685-702.

[38] Manguoglu, M., Cox, E., Saied, F., & Sameh, A. (2011). TRACEMIN-fiedler: A parallel algorithm for computing the Fiedler vector. In *High Performance Computing for Computational ScienceVECPAR 2010* (pp. 449-455). Springer Berlin Heidelberg.

[39] Mikkelsen, C. C. K., & Manguoglu, M. (2008). Analysis of the truncated SPIKE algorithm. *SIAM Journal on Matrix Analysis and Applications*, 30(4), 1500-1519.

[40] Mohar, B. (1991). Eigenvalues, diameter, and mean distance in graphs. *Graphs and combinatorics*, 7(1), 53-64.

[41] Mohar, B., & Alavi, Y. (1991). The Laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2, 871-898.

[42] Nichols, N. K. (1973). On the convergence of two-stage iterative processes for solving linear equations. *SIAM Journal on Numerical Analysis*, 10(3), 460-469.

[43] Nicolaides, R. A. (1987). Deflation of conjugate gradients with applications to boundary value problems. *SIAM Journal on Numerical Analysis*, 24(2), 355-365.

[44] Pedersen, Gert K. (1995). *Analysis Now*. Springer-Verlag New York Inc..

[45] Polizzi, E., & Sameh, A. H. (2006). A parallel hybrid banded system solver: the SPIKE algorithm. *Parallel computing*, 32(2), 177-194.

[46] Pothen, A., Simon, H. D., & Liou, K. P. (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3), 430-452.

[47] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill Book Company.

[48] Rutishauser, H. (1970). Simultaneous iteration method for symmetric matrices. *Numerische Mathematik*, 16(3), 205-223.

[49] Saad, Y. (1993). A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal on Scientific Computing*, 14(2), 461-469.

[50] Saad, Y. (1997). Analysis of augmented Krylov subspace methods. *SIAM Journal on Matrix Analysis and Applications*, 18(2), 435-449.

[51] Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*. SIAM.

[52] Saad, Y., Yeung, M., Erhel, J., & Guyomarc'h, F. (2000). A deflated version of the conjugate gradient algorithm. *SIAM Journal on Scientific Computing*, 21(5), 1909-1926.

[53] Saad, Y., & Schultz, M. H. (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM Journal on scientific and statistical computing, 7(3), 856-869.

[54] Sameh, A. H., & Wisniewski, J. A. (1982). A trace minimization algorithm for the generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 19(6), 1243-1259.

[55] Sameh, A., & Tong, Z. (2000). The trace minimization method for the symmetric generalized eigenvalue problem. *Journal of Computational and Applied Mathematics*, 123(1), 155-175.

[56] Shroff, G. M., & Keller, H. B. (1993). Stabilization of unstable procedures: the recursive projection method. *SIAM Journal on Numerical Analysis*, 30(4), 1099-1120.

[57] Stein, P., & Rosenberg, R.L. (1948). On the Solution of Linear Simultaneous Equations by Iteration, *Journal of London Mathematical Society*, 23, 111-118.

[58] Stewart, G.W. (1998). *Matrix Algorithm Volume II: Eigensystem*. SIAM.

[59] Sturler, E. de (1996). Inner-Outer Methods with Deflation for Linear Systems with Multiple Right-Hand Sides. Presentation in Householder Symposium XIII.

[60] Van der Vorst, H. A., & Vuik, C. (1994). GMRESR: A family of nested GMRES methods. Numerical linear algebra with applications, 1(4), 369-386.

[61] Varga, R. S. (1962). *Matrix Iterative Analysis*. Prentice-Hall, Inc..

[62] Wedin, P. . (1973). Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2), 217-232.

[63] Wilkinson, J.H. (1965). *The Algebraic Eigenvalue Problem*. Oxford Science Publications.

VITA

VITA

David Imberti was born in Boone, Iowa, in 1987. During High School, he spent two years enrolled at DMACC before he went to Iowa State University in 2005. In the summer of 2005, he participated in tachyon research for Iowa State University, and later as a teaching assistant in Trigonometry. David received his bachelor of science in mathematics and a minor in physics Summa Cum Laude in May 2008. He then went to Purdue University in 2008 entering in the graduate school of Mathematics in 2008.