

# Eigenvalue Analysis of a Block Red-Black Gauss-Seidel Preconditioner Applied to the Hermite Collocation Discretization of Poisson's Equation

Stephen H. Brill

*Department of Mathematics and Computer Science  
Boise State University  
Boise, Idaho, USA*

George F. Pinder

*Department of Civil and Environmental Engineering  
University of Vermont  
Burlington, Vermont, USA*

This is a preprint of an article published in *Numerical Methods for Partial Differential Equations*, Vol. 17, No. 3, pp. 204-228, May 2001. © copyright 2000 John Wiley & Sons, Inc.

*Received June 1, 2000*

This paper is concerned with the numerical solution of Poisson's equation with Dirichlet boundary conditions, defined on the unit square, discretized by Hermite collocation with uniform mesh. In [1], it was demonstrated that the Bi-CGSTAB method of van der Vorst [2] with block Red-Black Gauss-Seidel (RBGS) preconditioner is an efficient method to solve this problem.

In this paper, we derive analytic formulae for the eigenvalues that control the rate at which the Bi-CGSTAB/RBGS method converges. These formulae, which depend upon the location of the collocation points, can be utilized to determine where the collocation points should be placed in order to make the Bi-CGSTAB/RBGS method converge as quickly as possible. Furthermore, using the optimal location of the collocation points can result in significant time savings for fixed accuracy and fixed problem size. © 2001 John Wiley & Sons, Inc.

*Keywords: Hermite collocation, Bi-CGSTAB method, Red-Black, eigenvalue formulae*

## I. INTRODUCTION

The Bi-CGSTAB method of van der Vorst [2], combined with a block Red-Black Gauss-Seidel (RBGS) preconditioner, was shown in [1] to be an efficient method of solving gen-

eral linear partial differential equations in two spatial dimensions with Dirichlet and/or Neumann boundary conditions, discretized by Hermite collocation.

We study herein the specific case of Hermite collocation applied to Poisson's equation with Dirichlet boundary conditions on a uniform mesh, solved by Bi-CGSTAB/RBGS. We derive analytical formulae for the eigenvalues that control the rate at which Bi-CGSTAB/RBGS converges. Because these eigenvalues depend on the location of the collocation points, we are motivated to investigate if the speed of convergence can be enhanced by changing collocation point location. We find that the collocation points can be positioned in an optimal way to maximize the speed of convergence.

This paper is organized as follows. We first provide an overview of preliminary material. We then produce a lengthy analysis that culminates in formulae for the pertinent eigenvalues. This is followed by a discussion of locating the collocation points in an optimal way to accelerate the rate of convergence. Finally, numerical experiments indicate that placing the collocation points optimally results in significant savings in solving time for fixed accuracy and fixed problem size.

## II. PRELIMINARY MATERIAL

Details and derivations of this introductory material can be found in [1].

We wish to solve Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = H(x, y) \quad (2.1)$$

with Dirichlet boundary conditions, on the unit square  $\mathcal{S} = [0, 1] \times [0, 1]$  with a uniform mesh of  $m^2$  square finite elements ( $m$  must be even), discretized by Hermite collocation. Let each of these square finite elements be described in a local coordinate system as

$$\left[ -\frac{1}{2}, \frac{1}{2} \right] \times \left[ -\frac{1}{2}, \frac{1}{2} \right]. \quad (2.2)$$

In order to define a well-posed problem, we require four collocation points per finite element. With reference to (2.2), we set these collocation points to be at coordinates  $(-\xi, -\xi)$ ,  $(-\xi, \xi)$ ,  $(\xi, -\xi)$ ,  $(\xi, \xi)$ , where  $0 < \xi < \frac{1}{2}$ .

It is well known, given certain smoothness conditions, that to minimize discretization error, one chooses the collocation points within each finite element to coincide with the points of Gaussian quadrature [3]. This is equivalent to selecting  $\xi = \frac{1}{\sqrt{12}}$ , from which we obtain  $\mathcal{O}(h^4)$  discretization error, where  $h = \frac{1}{m}$ . If  $\xi \neq \frac{1}{\sqrt{12}}$ , then the discretization error is  $\mathcal{O}(h^2)$ . We initially use the Gaussian value of  $\xi = \frac{1}{\sqrt{12}}$  in the eigenvalue analysis that follows. Later, we will generalize our analysis to allow  $\xi$  to assume any value in the interval  $(0, \frac{1}{2})$ .

We utilize the Red-Black numbering of equations and unknowns described in [1] to discretize (2.1) via Hermite collocation, obtaining the matrix equation

$$A\mathbf{x} = \mathbf{b}, \quad (2.3)$$

whose block structure is

$$\left[ \begin{array}{c|ccc} A_F & & & B_F \\ & A_1 & & C_0 \ B_1 \\ & & A_3 & C_2 \ B_3 \\ & & & \ddots \ \ddots \\ & & & A_{m-3} & C_{m-4} \ B_{m-3} \\ & & & & C_L \ B_{m-3} \\ \hline C_F \ B_0 & & & A_0 \\ & C_1 \ B_2 & & A_2 \\ & & C_3 \ \ddots & \ddots \\ & & & A_{m-4} \\ & & & & A_{m-2} \\ & & & C_{m-3} \ B_L & \end{array} \right] \begin{bmatrix} \mathbf{v}_F \\ \mathbf{v}_1 \\ \mathbf{v}_3 \\ \vdots \\ \mathbf{v}_{m-3} \\ \mathbf{v}_L \\ \mathbf{v}_0 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_{m-4} \\ \mathbf{v}_{m-2} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_F \\ \mathbf{b}_1 \\ \mathbf{b}_3 \\ \vdots \\ \mathbf{b}_{m-3} \\ \mathbf{b}_L \\ \mathbf{b}_0 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_{m-4} \\ \mathbf{b}_{m-2} \end{bmatrix}, \quad (2.4)$$

which we abbreviate

$$\left[ \begin{array}{c|c} R & U \\ \hline L & B \end{array} \right] \begin{bmatrix} \mathbf{v}_R \\ \mathbf{v}_B \end{bmatrix} = \begin{bmatrix} \mathbf{b}_R \\ \mathbf{b}_B \end{bmatrix}. \quad (2.5)$$

Our RBGS preconditioner is

$$P = \left[ \begin{array}{c|c} R & \\ \hline L & B \end{array} \right].$$

### III. EIGENVALUE ANALYSIS

As reported in [4], the rate at which preconditioned conjugate gradient methods (like Bi-CGSTAB) converge “depends on the global eigenvalue distribution [of  $P^{-1}A$ ] more than anything else.” ( $P$  is normally chosen so that  $P^{-1}A \approx I$ , where  $I$  is the identity matrix of appropriate size.) We are thus motivated to find analytical formulae for the eigenvalues of  $P^{-1}A$ , buoyed by the knowledge that analytical formulae for the eigenvalues associated with solving (2.3) via the block Jacobi, Gauss-Seidel, and SOR (successive overrelaxation) methods were determined in [6]. Indeed, we use results reported in [6] in our discussion below as well as using the general approach of [6] as a model for determining the eigenvalues of  $P^{-1}A$ . We will use the term *spectrum* of a matrix to refer to the set whose entries are the eigenvalues of the matrix and denote the spectrum by  $\sigma$ . We thus seek  $\sigma(P^{-1}A)$ .

At times, we will want to consider the vector whose entries are those of the set  $\sigma(P^{-1}A)$ . We will use the same notation, i.e.,  $\sigma(P^{-1}A)$ , for both the vector and the set. We expect that this slight abuse of notation will not be confusing.

#### A. Reformulation of the Problem

To make the problem of finding eigenvalue formulae tractable, we introduce two new matrices. First, we replace matrix  $A$  by the matrix  $AK$ , where  $K$  is a diagonal matrix whose nontrivial entries are non-negative integer powers of  $m$ . Introduction of matrix  $K$  in this regard is equivalent to implementing the scaling procedure introduced in [5] and utilized in [6] and [7].

Secondly, we note that all the blocks of matrix  $A$  with numbered subscripts in (2.4) have the same size, namely  $4m \times 4m$ . However, the blocks with lettered subscripts have

different sizes.  $A_F$  and  $A_L$  are  $2m \times 2m$ ,  $B_F$  and  $C_L$  are  $2m \times 4m$ , and  $B_L$  and  $C_F$  are  $4m \times 2m$ .

In order to obtain a matrix where all the blocks have the same size, a similarity transformation which permutes the rows and columns of the matrix in (2.4) is performed such that the structure of each block is altered but the overall block structure in (2.4) is maintained. The resulting matrix is

$$\begin{array}{c}
 \begin{array}{|c|c|c|c|c|c|c|c|}
 \hline
 Y_2 & & & & & Y_3 & -Y_4 & & \\
 \hline
 Y_1 & -Y_2 & & & & Y_3 & Y_4 & & \\
 Y_1 & Y_2 & & & & Y_3 & -Y_4 & & \\
 \hline
 & & \ddots & & & & Y_3 & Y_4 & \\
 & & & \ddots & & & & \ddots & \ddots \\
 \hline
 & & & Y_1 & -Y_2 & & & & \ddots & \ddots \\
 & & & Y_1 & Y_2 & & & & & Y_3 & -Y_4 \\
 & & & & -Y_2 & & & & & Y_3 & Y_4 \\
 \hline
 Y_4 & & & & & Y_1 & -Y_2 & & & & \\
 Y_3 & -Y_4 & & & & Y_1 & Y_2 & & & & \\
 Y_3 & Y_4 & & & & & Y_1 & -Y_2 & & & \\
 & & \ddots & & & & & Y_1 & Y_2 & & \\
 & & & \ddots & & & & & \ddots & \ddots & \\
 & & & & Y_3 & -Y_4 & & & & & \ddots & \ddots \\
 & & & & Y_3 & Y_4 & & & & & & Y_1 & -Y_2 \\
 & & & & -Y_4 & & & & & & & Y_1 & Y_2 \\
 \hline
 \end{array}
 \end{array} \quad (3.1)$$

where the submatrices  $Y_1, Y_2, Y_3, Y_4$  are all  $2m \times 2m$  and have the structure

$$Y_i = \begin{array}{c}
 \begin{array}{|c|c|c|c|c|c|}
 \hline
 a_{i,2} & a_{i,3} & -a_{i,4} & & & \\
 a_{i,4} & a_{i,1} & -a_{i,2} & & & \\
 \hline
 a_{i,1} & a_{i,2} & a_{i,3} & -a_{i,4} & & \\
 a_{i,3} & a_{i,4} & a_{i,1} & -a_{i,2} & & \\
 \hline
 & a_{i,1} & a_{i,2} & a_{i,3} & -a_{i,4} & \\
 & a_{i,3} & a_{i,4} & a_{i,1} & -a_{i,2} & \\
 \hline
 & & \ddots & \ddots & \ddots & \\
 & & & \ddots & \ddots & \ddots \\
 \hline
 & & & a_{i,1} & a_{i,2} & a_{i,3} & -a_{i,4} \\
 & & & a_{i,3} & a_{i,4} & a_{i,1} & -a_{i,2} \\
 \hline
 & & & & a_{i,1} & a_{i,2} & -a_{i,4} \\
 & & & & a_{i,3} & a_{i,4} & -a_{i,2} \\
 \hline
 \end{array}
 \end{array}$$

For any  $\xi \in (0, \frac{1}{2})$ , the entries  $a_{i,j}$ ,  $i, j = 1, 2, 3, 4$ , are given below. Note the symmetry  $a_{i,j} = a_{j,i}$ . For the Gaussian case  $\xi = \frac{1}{\sqrt{12}}$ , these entries reduce to those given in [6] and

[7].

$$\begin{aligned}
a_{1,1} &= 12m^2\xi(\xi-1)(1+2\xi)^2 \\
a_{2,1} = a_{1,2} &= \frac{1}{2}m^2(1+2\xi)^2(12\xi^2-8\xi-1) \\
a_{3,1} = a_{1,3} &= 12m^2\xi^2(3-4\xi^2) \\
a_{4,1} = a_{1,4} &= -\frac{1}{2}m^2(1+2\xi)(1-2\xi-20\xi^2+24\xi^3) \\
a_{2,2} &= \frac{1}{4}m^2(6\xi+1)(1+2\xi)^2(2\xi-1) \\
a_{3,2} = a_{2,3} &= \frac{1}{2}m^2(1-2\xi)(-1-2\xi+20\xi^2+24\xi^3) \\
a_{4,2} = a_{2,4} &= \frac{1}{4}m^2(2\xi-1)(2\xi+1)(1-12\xi^2) \\
a_{3,3} &= 12m^2\xi(2\xi-1)^2(\xi+1) \\
a_{4,3} = a_{3,4} &= \frac{1}{2}m^2(1-2\xi)^2(-1+8\xi+12\xi^2) \\
a_{4,4} &= \frac{1}{4}m^2(2\xi-1)^2(2\xi+1)(6\xi-1).
\end{aligned} \tag{3.2}$$

Now, let  $W$  be the permutation matrix by which one obtains (3.1) from  $A$  via  $W^{-1}AW$ . Recalling matrix  $K$  above, let  $\bar{A} = W^{-1}AKW$  replace  $A$ . Similarly, replace  $P$  with  $\bar{P} = W^{-1}PKW$ . Then we obtain  $\bar{P}^{-1}\bar{A} = W^{-1}K^{-1}P^{-1}AKW$ . That is,  $\bar{P}^{-1}\bar{A}$  is a similarity transformation of  $P^{-1}A$ , and we may thus perform our eigenvalue analysis on  $\bar{P}^{-1}\bar{A}$ .

Considering the structure of (3.1), we may abbreviate  $\bar{A}$  by

$$\bar{A} = \left[ \begin{array}{c|c} \bar{R} & \bar{U} \\ \hline L & B \end{array} \right]$$

and  $\bar{P}$  by

$$\bar{P} = \left[ \begin{array}{c|c} \bar{R} & \\ \hline L & B \end{array} \right].$$

Because  $\bar{P}$  is a block  $2 \times 2$  matrix, its inverse is computable [8] as

$$\bar{P}^{-1} = \left[ \begin{array}{c|c} \bar{R}^{-1} & \\ \hline -B^{-1}LR^{-1} & B^{-1} \end{array} \right];$$

therefore

$$\bar{P}^{-1}\bar{A} = \left[ \begin{array}{c|c} I & \bar{R}^{-1}\bar{U} \\ \hline I - B^{-1}LR^{-1}U & \end{array} \right],$$

where  $I$  represents the identity matrix of appropriate size. Recall we want  $\bar{P}^{-1}\bar{A}$  to be “close” to  $I$ . This is clearly equivalent to

$$I - \bar{P}^{-1}\bar{A} = \left[ \begin{array}{c|c} & -\bar{R}^{-1}\bar{U} \\ \hline B^{-1}LR^{-1}U & \end{array} \right]$$

being “close” to the “null” matrix (i.e., the matrix whose entries are all zero). The null matrix has all its eigenvalues equal to zero. Since  $I - \bar{P}^{-1}\bar{A}$  may be viewed as a block upper-triangular matrix, its spectrum is given by the union of the spectra of those matrices on its diagonal blocks, namely the null matrix and  $J' = B^{-1}LR^{-1}U$ . We therefore expect the fastest convergence when the eigenvalues of  $J'$  are clustered near the origin of the complex plane.

Finally, we make one more change to the formulation above. We note that the eigenvalues of  $J = LR^{-1}UB^{-1}$  and those of  $J'$  are identical (because  $J$  is obtained from  $J'$  from a similarity transformation), and choose to perform our analysis on  $J$ .

B. The eigenvalues of  $J$  for the case  $\xi = \frac{1}{\sqrt{12}}$

Because  $\bar{R}$  and  $\bar{B}$  are both block diagonal, their inverses are easily computed. Noting that

$$\begin{bmatrix} Y_1 & -Y_2 \\ Y_1 & Y_2 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} Y_1^{-1} & Y_1^{-1} \\ -Y_2^{-1} & Y_2^{-1} \end{bmatrix},$$

we see that

$$\bar{R}^{-1} = \frac{1}{2} \left[ \begin{array}{c|c|c|c|c} 2Y_2^{-1} & & & & \\ \hline & Y_1^{-1} & Y_1^{-1} & & \\ & -Y_2^{-1} & Y_2^{-1} & & \\ \hline & & \ddots & \ddots & \\ & & \ddots & \ddots & \\ \hline & & & Y_1^{-1} & Y_1^{-1} \\ & & & -Y_2^{-1} & Y_2^{-1} \\ \hline & & & & -2Y_2^{-1} \end{array} \right]$$

and

$$\bar{B}^{-1} = \frac{1}{2} \left[ \begin{array}{c|c|c|c} Y_1^{-1} & Y_1^{-1} & & \\ \hline -Y_2^{-1} & Y_2^{-1} & & \\ \hline & Y_1^{-1} & Y_1^{-1} & \\ & -Y_2^{-1} & Y_2^{-1} & \\ \hline & & \ddots & \ddots \\ & & \ddots & \ddots \\ \hline & & & Y_1^{-1} & Y_1^{-1} \\ & & & -Y_2^{-1} & Y_2^{-1} \end{array} \right].$$

$J$  is thus seen to be

$$J = \left[ \begin{array}{c|c|c|c|c} S^2 - QS & SQ - Q^2 & & & \\ \hline SQ & S^2 & QS & Q^2 & \\ \hline Q^2 & QS & S^2 & SQ & \\ \hline & & SQ & S^2 & QS & Q^2 \\ \hline & & \ddots & \ddots & \ddots & \\ \hline & & & Q^2 & QS & S^2 & SQ \\ \hline & & & SQ & S^2 & QS & Q^2 \\ \hline & & & Q^2 & QS & S^2 & SQ \\ \hline & & & & SQ - Q^2 & S^2 - QS \end{array} \right], \quad (3.3)$$

where

$$S = -\frac{1}{2} (Y_3 Y_1^{-1} + Y_4 Y_2^{-1}) \quad (3.4)$$

and

$$Q = -\frac{1}{2} (Y_3 Y_1^{-1} - Y_4 Y_2^{-1}). \quad (3.5)$$



It is clear that the problem of determining the eigenvalues (and eigenvectors) of  $J$  is equivalent to solving the boundary value problem (3.9), (3.10), which we now proceed to do.

With respect to Theorem 8.3 in [9], we form the matrix polynomial  $\mathcal{L}(\lambda)$  that corresponds to (3.9), namely

$$\mathcal{L}(\lambda) = B_2\lambda^2 + (B_1 - \mu I)\lambda + B_0 = \begin{bmatrix} (S^2 - \mu)\lambda + Q^2 & SQ\lambda + SQ \\ SQ\lambda^2 + SQ\lambda & Q^2\lambda^2 + (S^2 - \mu)\lambda \end{bmatrix} \quad (3.11)$$

and compute its determinant

$$\det(\mathcal{L}(\lambda)) = \lambda \left\{ -Q^2\mu\lambda^2 + \left[ (S^2 - Q^2)^2 - 2S^2\mu + \mu^2 \right] \lambda - Q^2\mu \right\}. \quad (3.12)$$

Then (from Theorem 8.3 in [9]), the general solution of (3.9) is given by

$$\mathbf{z}_k = X_F J_F^k \mathbf{w}. \quad (3.13)$$

Here  $(X_F, J_F)$  is a *Jordan pair* (see [9]) of the matrix polynomial  $\mathcal{L}(\lambda)$  in (3.11) and  $\mathbf{w} \in \mathbf{C}^n$ , where  $n$  is the degree (in  $\lambda$ ) of  $\det(\mathcal{L}(\lambda))$ .

We consider two separate cases:  $\mu = 0$  and  $\mu \neq 0$ .

If  $\mu = 0$ , then

$$\mathcal{L}(\lambda) = \begin{bmatrix} S^2\lambda + Q^2 & SQ\lambda + SQ \\ SQ\lambda^2 + SQ\lambda & Q^2\lambda^2 + S^2\lambda \end{bmatrix}$$

and

$$\det(\mathcal{L}(\lambda)) = (S^2 - Q^2)^2 \lambda^2.$$

Thus the only eigenvalue of  $\mathcal{L}(\lambda)$ , i.e., zero of  $\det(\mathcal{L}(\lambda))$ , is  $\lambda = 0$ , which is a double eigenvalue. The *Jordan chain* (see [9]) associated with the eigenvalue  $\lambda = 0$  is seen to be of length two and it thus forms a *canonical set* (see [9]). Its components, using the definition in [9], are easily seen to be the columns of

$$X_F = \begin{bmatrix} 1 & 1 \\ -\frac{Q}{S} & -\frac{S}{Q} \end{bmatrix}$$

while the matrix  $J_F$  is

$$J_F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

and  $\mathbf{w} = [w_0 \ w_1]^T$ . Applying (3.13) with  $k = 0, 1, 2, \dots$  and the definitions of  $X_F$  and  $J_F$  given above, we see that

$$\begin{aligned} \mathbf{z}_0 &= \begin{bmatrix} z_{0,0} \\ z_{0,1} \end{bmatrix} = \begin{bmatrix} w_0 + w_1 \\ -\frac{Q}{S}w_0 - \frac{S}{Q}w_1 \end{bmatrix} \\ \mathbf{z}_1 &= \begin{bmatrix} z_{1,0} \\ z_{1,1} \end{bmatrix} = \begin{bmatrix} w_1 \\ -\frac{Q}{S}w_1 \end{bmatrix} \\ \mathbf{z}_k &= \mathbf{0}, \quad k \geq 2. \end{aligned} \quad (3.14)$$

Now, recall the boundary condition  $b_0 = Q^2(z_{0,0} + z_{1,1}) + QS(z_{0,1} + z_{1,0}) = 0$  from (3.10). Assuming that  $S \neq \pm Q$  (a most reasonable assumption in light of (3.4) and (3.5)) and using the values of  $z_{0,0}$ ,  $z_{1,1}$ ,  $z_{0,1}$ , and  $z_{1,0}$  from (3.14), we conclude that

$w_1 = 0$ . But then  $\mathbf{z}_k = 0$  for all  $k \geq 1$ , which means that  $\mathbf{z}$  in (3.6) is a zero eigenvector, which is impermissible. The case  $\mu = 0$  is therefore eliminated from consideration.

Thus  $\mu \neq 0$ . In this case,  $\det(\mathcal{L}(\lambda))$  is given by (3.12). Setting  $\det(\mathcal{L}(\lambda)) = 0$  to determine the eigenvalues  $\lambda$  of  $\mathcal{L}(\lambda)$  gives

$$\lambda = \lambda_0 = 0 \text{ or } \lambda = \lambda_1 \text{ or } \lambda = \lambda_2$$

where  $\lambda_1$  and  $\lambda_2$  are obtained from the quadratic formula. Using the well-known formulae for the sum and product of the roots of a quadratic equation, we obtain

$$\lambda_1 + \lambda_2 = \frac{(S^2 - Q^2)^2 - 2S^2\mu + \mu^2}{Q^2\mu} \quad (3.15)$$

and

$$\lambda_1\lambda_2 = 1. \quad (3.16)$$

We note that the Jordan chain corresponding to the eigenvalue  $\lambda_0 = 0$  is  $\begin{bmatrix} 1 \\ -\frac{Q}{S} \end{bmatrix}$  or equivalently,  $\begin{bmatrix} Q \\ -S \end{bmatrix}$ . We now consider two separate cases, namely  $\lambda_1 \neq \lambda_2$  and  $\lambda_1 = \lambda_2$ .

If  $\lambda_1 \neq \lambda_2$ , then the Jordan chain corresponding to  $\lambda_i$  is seen to be  $\begin{bmatrix} \omega_i \\ 1 \end{bmatrix}$ ,  $i = 1, 2$ , where

$$\omega_i = \frac{-SQ\lambda_i - SQ}{(S^2 - \mu)\lambda_i + Q^2}. \quad (3.17)$$

So we obtain, in this case

$$X_F = \begin{bmatrix} Q & \omega_1 & \omega_2 \\ -S & 1 & 1 \end{bmatrix},$$

$$J_F = \begin{bmatrix} 0 & & \\ & \lambda_1 & \\ & & \lambda_2 \end{bmatrix},$$

and  $\mathbf{w} = [w_0 \ w_1 \ w_2]^T$ . We now find  $\mathbf{w}$  to satisfy the boundary conditions (3.10). Using (3.13) for  $k = 0, 1, \frac{m}{2}, \frac{m}{2} + 1$  together with (3.10), we obtain the equation

$$E \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where

$$E = \begin{bmatrix} Q(\omega_1 + \lambda_1) + S(1 + \omega_1\lambda_1) & Q(\omega_2 + \lambda_2) + S(1 + \omega_2\lambda_2) \\ Q\left(\omega_1\lambda_1^{\frac{m}{2}} + \lambda_1^{\frac{m}{2}+1}\right) + S\left(\lambda_1^{\frac{m}{2}} + \omega_1\lambda_1^{\frac{m}{2}+1}\right) & Q\left(\omega_2\lambda_2^{\frac{m}{2}} + \lambda_2^{\frac{m}{2}+1}\right) + S\left(\lambda_2^{\frac{m}{2}} + \omega_2\lambda_2^{\frac{m}{2}+1}\right) \end{bmatrix}.$$

If  $E$  is nonsingular, then  $w_1 = w_2 = 0$ , which implies (from (3.13)) that  $\mathbf{z}_k = 0$  for all  $k \geq 1$ . Thus  $\mathbf{z}$  in (3.6) is a zero eigenvector, which is impermissible. Therefore,  $E$  is singular, so its determinant is zero, which leads to the equation

$$\left(\lambda_2^{\frac{m}{2}} - \lambda_1^{\frac{m}{2}}\right) [Q(\omega_1 + \lambda_1) + S(1 + \omega_1\lambda_1)] [Q(\omega_2 + \lambda_2) + S(1 + \omega_2\lambda_2)] = 0. \quad (3.18)$$

If  $Q(\omega_i + \lambda_i) + S(1 + \omega_i \lambda_i) = 0$ ,  $i = 1, 2$ , then using (3.17) and solving for  $\lambda_i$  yields

$$\lambda_i = \frac{(S + Q)(S - Q)^2 - S\mu}{Q\mu}, \quad (3.19)$$

$i = 1, 2$ . Since  $\lambda_i$  (which is non-zero) is a solution of  $\det(\mathcal{L}(\lambda)) = 0$ , we can conclude from (3.12) that

$$-Q^2\mu\lambda_i^2 + [(S^2 - Q^2)^2 - 2S^2\mu + \mu^2]\lambda_i - Q^2\mu = 0. \quad (3.20)$$

Substitution of (3.19) into (3.20) yields a cubic equation in  $\mu$ , whose solutions are

$$\mu = (S - Q)(S + Q) \text{ or } \mu = (S - Q)^2,$$

where the former solution is of multiplicity two.

If  $\mu = (S - Q)^2$ , then (3.19) reduces to  $\lambda_1 = \lambda_2 = 1$ , which contradicts the assumption  $\lambda_1 \neq \lambda_2$ . If, on the other hand,  $\mu = (S - Q)(S + Q)$ , then (3.19) reduces to  $\lambda_1 = \lambda_2 = -1$ , which also contradicts the assumption  $\lambda_1 \neq \lambda_2$ . Thus, with respect to (3.18), we obtain

$$\left(\lambda_2^{\frac{m}{2}} - \lambda_1^{\frac{m}{2}}\right) = 0.$$

Recalling (3.16) and the assumption  $\lambda_1 \neq \lambda_2$ , we conclude

$$\lambda_1 = e^{i\theta} \quad (3.21)$$

and

$$\lambda_2 = e^{-i\theta} \quad (3.22)$$

where  $\theta = \frac{2k\pi}{m}$ ,  $k = 1, 2, 3, \dots, \frac{m}{2} - 1$ .

If we now add the equation obtained by substituting (3.21) into (3.20) to the equation obtained by substituting (3.22) into (3.20), we obtain a quadratic equation in  $\mu$  whose coefficients are all real. Solving this equation for  $\mu$  yields

$$\mu = (Q^2 \cos \theta + S^2) \pm Q\sqrt{2S^2(1 + \cos \theta) - (Q \sin \theta)^2}, \quad (3.23)$$

$\theta = \frac{2k\pi}{m}$ ,  $k = 1, 2, 3, \dots, \frac{m}{2} - 1$ .

What we have shown so far is this: if  $S$  and  $Q$  are scalars defining the  $m \times m$  matrix  $J$  in (3.3), then  $m - 2$  of the  $m$  eigenvalues of  $J$  are given by (3.23). The remaining two eigenvalues of  $J$  arise from the case  $\lambda_1 = \lambda_2$  and will be determined below.

If  $\lambda_1 = \lambda_2$ , then (3.15) becomes

$$2\lambda_{1,2}Q^2\mu = (Q^2 - S^2)^2 - 2S^2\mu + \mu^2, \quad (3.24)$$

where  $\lambda_{1,2} = \lambda_1 = \lambda_2$ . By (3.16), we must have  $\lambda_1 = \lambda_2 = 1$  or  $\lambda_1 = \lambda_2 = -1$ .

If  $\lambda_1 = \lambda_2 = 1$ , then solving (3.24) for  $\mu$  yields

$$\mu = (S \pm Q)^2.$$

We now consider these two cases separately.

If  $\mu = (S + Q)^2$ , then, using the definition of Jordan pair in [9], we obtain

$$X_F = \begin{bmatrix} S & 1 & -\frac{1}{2} \\ -Q & 1 & \frac{Q}{2S} \end{bmatrix},$$

$$J_F = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

and  $\mathbf{w} = [w_0 \ w_1 \ w_2]^T$ . Using (3.13) with these values and (3.10), we obtain

$$\begin{bmatrix} 4S & Q + 1 \\ 2 & 2m + \frac{1}{2} + \frac{Q}{2S} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Unless  $S = \frac{1}{4m+1}$ , this  $2 \times 2$  matrix is nonsingular, which implies  $w_1 = w_2 = 0$ . Using (3.13), we see that  $\mathbf{z}_k = 0$  for all  $k \geq 1$  which implies that  $\mathbf{z}$  is a zero eigenvector, which is impermissible. We thus eliminate the possibility  $\mu = (S + Q)^2$ .

If  $\mu = (S - Q)^2$ , then, again using the definition of Jordan pair from [9], we obtain

$$X_F = \begin{bmatrix} S & 1 & -\frac{1}{2} \\ -Q & -1 & \frac{Q}{2S} \end{bmatrix},$$

$$J_F = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

and  $\mathbf{w} = [w_0 \ w_1 \ w_2]^T$ . Using (3.13) with these values and (3.10), we find that  $w_2 = 0$  and that  $w_1 \neq 0$  is arbitrary. We conclude that  $\lambda = 1$  provides the eigenvalue  $\mu = (S - Q)^2$  of  $J$  with corresponding eigenvector  $[1 \ -1 \ 1 \ -1 \ \cdots \ 1 \ -1]^T$ .

If  $\lambda_1 = \lambda_2 = -1$ , then solving (3.24) for  $\mu$  yields the solution of multiplicity two

$$\mu = (S - Q)(S + Q).$$

Once more using the definition of Jordan pair in [9], we obtain for this case,

$$X_F = \begin{bmatrix} S & 1 & 0 \\ -Q & 0 & 1 \end{bmatrix},$$

$$J_F = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

and  $\mathbf{w} = [w_0 \ w_1 \ w_2]^T$ . Using (3.13) with these values and (3.10), we find that  $w_1 = w_2$  is arbitrary but non-zero. We thus obtain eigenvalue  $\mu = (S - Q)(S + Q)$  and corresponding eigenvector  $[-1 \ -1 \ 1 \ 1 \ \cdots \ -1 \ -1 \ 1 \ 1]^T$  of  $J$  from  $\lambda = -1$ .

We have therefore proved

**Lemma 3.1.** *Let  $J$  be the  $m \times m$  matrix defined in (3.3) by the real scalars  $S$  and  $Q$ . Then the eigenvalues  $\mu$  of  $J$  are given by*

$$\begin{aligned} \mu &= S^2 - Q^2 \\ \mu &= (S - Q)^2 \\ \mu &= (Q^2 \cos \theta + S^2) \pm Q \sqrt{2S^2(1 + \cos \theta) - (Q \sin \theta)^2}, \end{aligned}$$

where  $\theta = \frac{2k\pi}{m}$ ,  $k = 1, 2, 3, \dots, \frac{m}{2} - 1$ .

Before we consider the general case where  $S$  and  $Q$  are matrices, we require another lemma:

**Lemma 3.2.** *Let  $S$  and  $Q$  be as defined in (3.4) and (3.5), respectively. Then  $SQ = QS$ .*

**Proof.** Using Lemma 5.2 in [6], we are given the existence of a nonsingular matrix  $X$  and explicit diagonal matrices  $D$  and  $\overline{D}$  such that

$$X^T (Y_4 Y_2^{-1}) X^{-T} = D \quad (3.25)$$

and

$$X^T (Y_3 Y_1^{-1}) X^{-T} = \overline{D}. \quad (3.26)$$

Thus  $Y_4 Y_2^{-1}$  and  $Y_3 Y_1^{-1}$  have the same complete set of eigenvectors and must therefore commute [10], i.e.,

$$Y_4 Y_2^{-1} Y_3 Y_1^{-1} = Y_3 Y_1^{-1} Y_4 Y_2^{-1}. \quad (3.27)$$

But this is precisely the condition that is required to show that  $SQ = QS$ .  $\blacksquare$

Now that we have established that  $S$  and  $Q$  commute, we see that we can use the analysis culminating in Lemma 3.1 to determine the eigenvalues of  $J$  for the case where  $S$  and  $Q$  are defined as in (3.4) and (3.5). Let us begin by recalling (3.11):

$$\mathcal{L}(\lambda) = \begin{bmatrix} (S^2 - \mu I) \lambda + Q^2 & SQ\lambda + SQ \\ SQ\lambda^2 + SQ\lambda & Q^2\lambda^2 + (S^2 - \mu I) \lambda \end{bmatrix},$$

where we are making use of the fact that  $S$  and  $Q$  are commuting matrices. As above, let us now consider

$$0 = \det(\mathcal{L}(\lambda)) = \det \begin{bmatrix} (S^2 - \mu I) \lambda + Q^2 & SQ\lambda + SQ \\ SQ\lambda^2 + SQ\lambda & Q^2\lambda^2 + (S^2 - \mu I) \lambda \end{bmatrix}. \quad (3.28)$$

From our work above, we know the solutions  $\lambda$  of (3.28), namely  $\lambda = 0$ ,  $\lambda = \pm 1$ , and  $\lambda = e^{\pm i\theta_k}$ ,  $\theta_k = 1, 2, \dots, \frac{m}{2} - 1$ . We now exploit this knowledge to compute the eigenvalues  $\mu$  of  $J$ .

If we use the values of  $\lambda = e^{i\theta_k}$ ,  $\theta_k = \frac{2k\pi}{m}$ ,  $k = 1, 2, \dots, \frac{m}{2} - 1$  in (3.28), we can easily show that

$$0 = \det \begin{bmatrix} S^2 + \frac{1}{\lambda} Q^2 - \mu I & SQ + \frac{1}{\lambda} SQ \\ SQ + \lambda SQ & S^2 + \lambda Q^2 - \mu I \end{bmatrix}. \quad (3.29)$$

Finding the values  $\mu$  that satisfy (3.29) is equivalent to finding the eigenvalues  $\mu$  of the matrix

$$M_k = \begin{bmatrix} S^2 + \frac{1}{\lambda} Q^2 & SQ + \frac{1}{\lambda} SQ \\ SQ + \lambda SQ & S^2 + \lambda Q^2 \end{bmatrix}.$$

To eliminate the complex numbers (i.e., the  $\lambda$ 's), we perform the similarity transformation  $T_k = R_k M_k R_k^{-1}$ , where

$$R_k = \begin{bmatrix} -i(Q^2 - SQ) \sin \theta_k & i(Q^2 - SQ) \sin \theta_k \\ (Q^2 + SQ) \sin \theta_k & (Q^2 + SQ) \sin \theta_k \end{bmatrix}.$$

We thus seek the eigenvalues  $\mu$  of

$$T_k = \begin{bmatrix} (S - Q)(S - Q \cos \theta_k) & (S - Q)Q \sin \theta_k \\ (S + Q)Q \sin \theta_k & (S + Q)(S + Q \cos \theta_k) \end{bmatrix},$$

$$\theta_k = \frac{2k\pi}{m}, k = 1, 2, \dots, \frac{m}{2} - 1.$$

The characterization of all the eigenvalues  $\mu$  of  $J$  is found in the following lemma, the proof of which is, except for the obvious notational differences, analogous to that given in Lemma 4.2 in [6]:

**Lemma 3.3.** *Let  $J$  be the  $2m^2 \times 2m^2$  matrix defined in (3.3) by the matrices  $S$  and  $Q$  defined in (3.4) and (3.5). Then*

$$\sigma(J) = \bigcup_{k=1}^{\frac{m}{2}-1} \sigma(T_k) \cup \sigma(S^2 - Q^2) \cup \sigma((S - Q)^2). \quad (3.30)$$

Now that we have characterized the spectrum of  $J$  as the union of spectra of other matrices, we now determine these latter spectra, namely for  $T_k$ ,  $k = 1, 2, \dots, \frac{m}{2} - 1$ ;  $S^2 - Q^2$ ; and  $(S - Q)^2$ . We first make some observations and definitions. Let  $Y_{31} = Y_3 Y_1^{-1}$  and let  $Y_{42} = Y_4 Y_2^{-1}$ . With respect to (3.4) and (3.5), we see that

$$S + Q = -Y_{31}$$

and

$$S - Q = -Y_{42}.$$

Using these definitions, trigonometric identities, and (3.27), we can show

$$T_k = \begin{bmatrix} \sin^2 \frac{\theta_k}{2} Y_{42} Y_{31} + \cos^2 \frac{\theta_k}{2} Y_{42}^2 & \frac{\sin \theta_k}{2} (Y_{42} Y_{31} - Y_{42}^2) \\ \frac{\sin \theta_k}{2} (Y_{31}^2 - Y_{42} Y_{31}) & \sin^2 \frac{\theta_k}{2} Y_{42} Y_{31} + \cos^2 \frac{\theta_k}{2} Y_{31}^2 \end{bmatrix}.$$

We now use (3.25) and (3.26) to compute the similarity transformation

$$\begin{bmatrix} X^T & \\ & X^T \end{bmatrix} T_k \begin{bmatrix} X^{-T} & \\ & X^{-T} \end{bmatrix} = \begin{bmatrix} \sin^2 \frac{\theta_k}{2} D \bar{D} + \cos^2 \frac{\theta_k}{2} D^2 & \frac{\sin \theta_k}{2} (D \bar{D} - D^2) \\ \frac{\sin \theta_k}{2} (\bar{D}^2 - D \bar{D}) & \sin^2 \frac{\theta_k}{2} D \bar{D} + \cos^2 \frac{\theta_k}{2} \bar{D}^2 \end{bmatrix}, \quad (3.31)$$

where  $D$  and  $\bar{D}$  are diagonal matrices given explicitly in Lemma 5.2 of [6]. We write

$$D = \text{diag}(d_0, d_1, \dots, d_{2m-1})$$

and

$$\bar{D} = \text{diag}(\bar{d}_0, \bar{d}_1, \dots, \bar{d}_{2m-1}).$$

If we then permute the rows and columns of (3.31) in an obvious way, we find that  $T_k$  is similar to  $\text{diag}(D_{k,0}, D_{k,1}, \dots, D_{k,2m-1})$ , where

$$D_{k,i} = \begin{bmatrix} d_i \left( \bar{d}_i \sin^2 \frac{\theta_k}{2} + d_i \cos^2 \frac{\theta_k}{2} \right) & d_i \frac{\sin \theta_k}{2} (\bar{d}_i - d_i) \\ \bar{d}_i \frac{\sin \theta_k}{2} (\bar{d}_i - d_i) & \bar{d}_i \left( d_i \sin^2 \frac{\theta_k}{2} + \bar{d}_i \cos^2 \frac{\theta_k}{2} \right) \end{bmatrix}.$$

Since each  $D_{k,i}$  is a  $2 \times 2$  matrix, the eigenvalues of each are easily determined:

$$\sigma(D_{k,i}) = \left\{ \mu : \mu = \frac{\zeta_{k,i}^2 + 2\bar{d}_i d_i \pm \sqrt{\zeta_{k,i}^2 (\zeta_{k,i}^2 + 4\bar{d}_i d_i)}}{2} \right\},$$

where  $\zeta_{k,i} = (\bar{d}_i - d_i) \cos \frac{\theta_k}{2}$ ,  $\theta_k = \frac{2k\pi}{m}$ ,  $k = 1, 2, \dots, \frac{m}{2} - 1$ ,  $i = 0, 1, \dots, 2m - 1$ .

To find  $\sigma(S^2 - Q^2)$ , we note that  $S^2 - Q^2 = Y_{42} Y_{31}$  and  $X^T Y_{42} Y_{31} X^{-T} = \bar{D} D$ , which is a diagonal matrix. Therefore,  $\sigma(S^2 - Q^2) = \{\bar{d}_i d_i\}_{i=0}^{2m-1}$ .

To find  $\sigma((S - Q)^2)$ , we note that  $(S - Q)^2 = Y_{42}^2$ , which is similar to  $D^2$ , which is a diagonal matrix. Therefore,  $\sigma((S - Q)^2) = \{d_i^2\}_{i=0}^{2m-1}$ .

We may now state:

**Theorem 3.4.** *Let  $J$  be the  $2m^2 \times 2m^2$  matrix defined in (3.3) with  $S$  and  $Q$  defined as in (3.4) and (3.5), respectively. Let  $\xi = \frac{1}{\sqrt{42}}$  in (3.2). Then*

$$\sigma(J) = \{\mu : \mu = d_i^2\} \cup \{\mu : \mu = \bar{d}_i d_i\} \cup \left\{ \mu : \mu = \frac{\zeta_{k,i}^2 + 2\bar{d}_i d_i \pm \sqrt{\zeta_{k,i}^2 (\zeta_{k,i}^2 + 4\bar{d}_i d_i)}}{2} \right\},$$

where  $\zeta_{k,i} = (\bar{d}_i - d_i) \cos \theta_k$ ,  $\theta_k = \frac{k\pi}{m}$ ,  $k = 1, 2, \dots, \frac{m}{2} - 1$ ,  $i = 0, 1, \dots, 2m - 1$ , and where [6]

$$\{d_i\}_{i=0}^{2m-1} = \{\alpha_0^-, \alpha_1^+, \alpha_1^-, \dots, \alpha_{m-1}^+, \alpha_{m-1}^-, \alpha_m^-\},$$

$$\{\bar{d}_i\}_{i=0}^{2m-1} = \{\beta_0^-, \beta_1^+, \beta_1^-, \dots, \beta_{m-1}^+, \beta_{m-1}^-, \beta_m^-\},$$

$$\alpha_j^\pm = \frac{3\sqrt{3} \pm q_j}{-28 - 16\sqrt{3} + (\sqrt{3} + 1) \cos \varphi_j},$$

$$\beta_j^\pm = \frac{(37 + 8 \cos \varphi_j) \pm 3\sqrt{3} q_j}{-64 - 36\sqrt{3} + (19 + 9\sqrt{3} \cos \varphi_j)},$$

$$q_j = \sqrt{43 + 40 \cos \varphi_j - 2 \cos^2 \varphi_j},$$

$$\varphi_j = \frac{j\pi}{m}, j = 0, 1, \dots, m.$$

C. The eigenvalues of  $J$  for the case of general  $\xi \in (0, \frac{1}{2})$

The result in Theorem 3.4 is for the case  $\xi = \frac{1}{\sqrt{12}}$ . If we seek an analogous result for arbitrary  $\xi \in (0, \frac{1}{2})$ , we note that we must consider only three things. The first (and obvious) one is that the entries of matrix  $A$  are now given by (3.2) for arbitrary  $\xi$  (as opposed to the specific value  $\xi = \frac{1}{\sqrt{12}}$ ). The second is to check whether we obtain the result  $w_1 = \bar{w}_1$ , where  $w_1$  and  $\bar{w}_1$  are given in Lemmas 5.1 and 5.2 in [6], which a long and tedious calculation does indeed confirm. The third is to determine how  $\alpha_j^\pm$  associates with  $\beta_j^\pm$  (see Lemma 5.2 in [6]). In this case, another long and tedious calculation provides that  $\alpha_j^+$  associates with  $\beta_j^-$  and that  $\alpha_j^-$  associates with  $\beta_j^+$  for all values of  $\xi \in (0, \frac{1}{2})$ . We can therefore state, for arbitrary  $\xi \in (0, \frac{1}{2})$ , the result analogous to Theorem 3.4:

**Theorem 3.5.** *Let  $J$  be the  $2m^2 \times 2m^2$  matrix defined in (3.3) with  $S$  and  $Q$  defined as in (3.4) and (3.5), respectively. Let  $\xi$  in (3.2) be an arbitrary number in the interval  $(0, \frac{1}{2})$ . Then*

$$\sigma(J) = \{\mu : \mu = d_i^2\} \cup \{\mu : \mu = \bar{d}_i d_i\} \cup \left\{ \mu : \mu = \frac{\zeta_{k,i}^2 + 2\bar{d}_i d_i \pm \sqrt{\zeta_{k,i}^2 (\zeta_{k,i}^2 + 4\bar{d}_i d_i)}}{2} \right\};$$

where  $\zeta_{k,i} = (\bar{d}_i - d_i) \cos \theta_k$ ;  $\theta_k = \frac{k\pi}{m}$ ;  $k = 1, 2, \dots, \frac{m}{2} - 1$ ;  $i = 0, 1, \dots, 2m-1$ ;  $c_j = \cos \theta_j$ ;  $j = 0, 1, \dots, m$ ;

$$\{d_i\}_{i=0}^{2m-1} = \{\alpha_0^+, \alpha_1^+, \alpha_1^-, \dots, \alpha_{m-1}^+, \alpha_{m-1}^-, \alpha_m^+\};$$

$$\{\bar{d}_i\}_{i=0}^{2m-1} = \{\beta_0^-, \beta_1^-, \beta_1^+, \dots, \beta_{m-1}^-, \beta_{m-1}^+, \beta_m^-\};$$

$$q_j = 2\xi \sqrt{(16\xi^4 - 24\xi^2 + 21) + c_j (-32\xi^4 + 18) + c_j^2 (16\xi^4 + 24\xi^2 - 3)};$$

$$\alpha_j^\pm = \frac{96\xi^4 - 92\xi^2 + 5 + c_j (1 - 12\xi^2) (-1 + 8\xi^2) \pm q_j}{-96\xi^4 - 56\xi^3 + 76\xi^2 + 42\xi + 5 + c_j (-1 - 2\xi) (1 + 4\xi - 4\xi^2 - 48\xi^3)};$$

$$\beta_j^\pm = \frac{128\xi^6 - 192\xi^4 + 60\xi^2 - 1 + c_j (-128\xi^6 + 96\xi^4 - 12\xi^2 + 1) \pm q_j}{(1 + 2\xi) (-1 - 16\xi - 28\xi^2 + 80\xi^3 + 32\xi^4 - 64\xi^5) + c_j (1 - 2\xi) (1 + 2\xi)^2 (1 + 4\xi + 8\xi^2 - 16\xi^3)}.$$

#### IV. SHIFTING THE COLLOCATION POINTS TO MINIMIZE $\|\sigma(\mathbf{I} - \mathbf{P}^{-1}\mathbf{A})\|_2$

In this section we investigate the possibility of increasing the speed at which the Bi-CGSTAB/RBGS algorithm converges by using values for  $\xi$  other than  $\frac{1}{\sqrt{12}}$ , acknowledging that any increase in convergence rate will come at the expense of losing the  $\mathcal{O}(h^4)$  accuracy provided by the Gaussian value  $\xi = \frac{1}{\sqrt{12}}$ . In brief, we show that the collocation points can be located in an optimal fashion that results in significant time savings for fixed accuracy and fixed problem size.

Recall the discussion above where we stated that we expect the fastest convergence of the preconditioned Bi-CGSTAB algorithm to occur when the eigenvalues are clustered

about the origin of the complex plane. In order to measure the clustering, we consider the vector norm  $\|\sigma(I - P^{-1}A)\|_2 = \|\sigma(J)\|_2$ . In light of the geometric interpretation of the 2-norm, we may say that optimal clustering about the origin of the complex plane is equivalent to minimizing  $\|\sigma(J)\|_2$  as a function of  $\xi$ , the parameter which controls collocation point location.

In the literature (e.g. [4] and [11]), the *condition number*  $\kappa_\gamma(P^{-1}A)$  is often used as a measure of how close  $P^{-1}A$  is to the identity matrix  $I$ , with rapid convergence occurring when  $P^{-1}A \approx I$ . The condition number is defined

$$\kappa_\gamma(M) = \|M\|_\gamma \|M^{-1}\|_\gamma,$$

where  $\gamma$  is a given matrix norm. The four most common matrix norms correspond to  $\gamma = 1, 2, \infty, F$  (see [11] for the definitions of these norms). Because these matrix norms satisfy the *consistency* [12] or *submultiplicative* [11] property, the minimum value that  $\kappa_\gamma(M)$  can attain is unity (which occurs when  $M = I$ ). Thus if we were able to achieve  $P = A$ , then we would have  $\|\sigma(J)\|_2 = 0$  (the minimum value a vector norm can attain) and  $\kappa_\gamma(P^{-1}A) = 1$  (the minimum value a condition number can attain). We are therefore motivated to examine the relationships between the value of  $\xi$  that produces the fastest convergence of Bi-CGSTAB/RBGS, the value of  $\xi$  that minimizes  $\|\sigma(J)\|_2$  and the value of  $\xi$  that minimizes each of the four condition numbers.

In Figure 1, we summarize results obtained for Poisson's equation (2.1) for the case  $m = 10$ . Convergence is defined by the 2-norm of the residual vector being less than  $10^{-8}$ . For  $\xi = 0.1, 0.2, \dots, 0.49$ , we solved Poisson's equation using Bi-CGSTAB/RBGS, computed  $\|\sigma(J)\|_2 = \|\sigma(I - P^{-1}A)\|_2$  using the eigenvalue formulae derived above, and computed the condition numbers using Matlab. With reference to Figure 1, we see that the general behavior in all six graphs is similar: i.e., as  $\xi$  increases, the curves rapidly decrease to a minimum, then increase gradually. Upon more careful examination, we see that the value of  $\xi$  that minimizes each of  $\|\sigma(I - P^{-1}A)\|_2$ ,  $\kappa_F(P^{-1}A)$ , and  $\kappa_\infty(P^{-1}A)$  corresponds well to the value of  $\xi$  that minimizes the number of iterations required for convergence of Bi-CGSTAB/RBGS. The condition numbers using the 1- and 2-norms are less effective in predicting the value of  $\xi$  that produces the fastest convergence of Bi-CGSTAB/RBGS.

In Figure 2, we repeat our study for  $m = 20$ . The results are similar, except that now the condition number using the  $\infty$ -norm also is a poor predictor of optimal convergence, leaving only  $\|\sigma(I - P^{-1}A)\|_2$  and  $\kappa_F(P^{-1}A)$  as effective predictors for the value of  $\xi$  that produces optimal convergence of Bi-CGSTAB/RBGS.

It is worth noting that computing  $\|\sigma(I - P^{-1}A)\|_2$  is easy, since we have formulas for the eigenvalues of  $J$ . Furthermore, finding the value of  $\xi$  that minimizes  $\|\sigma(I - P^{-1}A)\|_2$  for any given value of  $m$  is also easy if we use the MINOS optimization software (see [13] for documentation). Using MINOS, we determined the value of  $\xi$  that minimizes  $\|\sigma(I - P^{-1}A)\|_2$  for Poisson's equation for all even  $m$  between 2 and 100, inclusive. The results are given in Figure 3.

In examining Figure 3, we see that the optimal value of  $\xi$  is relatively insensitive to problem size  $m$  for sufficiently large  $m$ . This is an important result, for it means that a separate optimization for each value of  $m$  is not necessary each time we wish to solve Poisson's equation as quickly as possible. A qualitative explanation for the insensitivity of optimal  $\xi$  with respect to  $m$  follows.

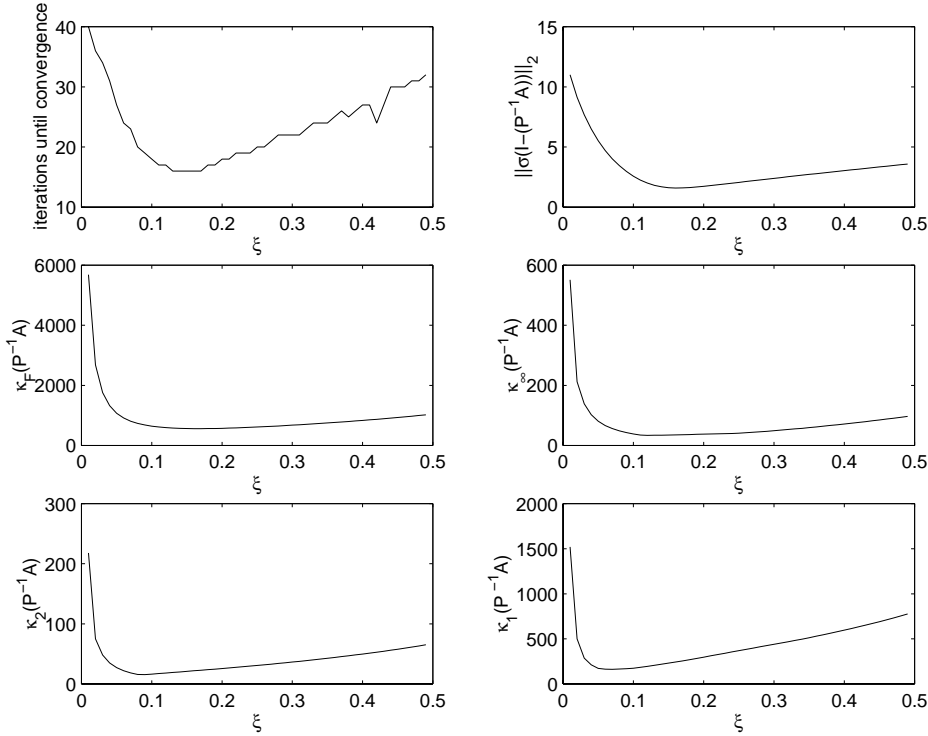


FIG. 1. Comparison of numerical results, condition numbers and  $\|\sigma(I - P^{-1}A)\|_2$  for Poisson's equation for  $m = 10$

With respect to Theorem 3.5, we see that there are eight flavors of eigenvalues. Four of them are  $(\alpha^+)^2$ ,  $(\alpha^-)^2$ ,  $\alpha^+\beta^-$ , and  $\alpha^-\beta^+$ . The remaining four are of the form

$$\frac{1}{2} \left( \zeta^2 + 2\bar{d}d \pm \sqrt{\zeta^2 (\zeta^2 + 4\bar{d}d)} \right).$$

This represents four additional flavors since  $\zeta$  may be formed by combining  $\alpha^+$  and  $\beta^-$  or by combining  $\alpha^-$  and  $\beta^+$  and, for each form of  $\zeta$ , we either add or subtract the radical. To summarize:

flavor number	characterization	number of eigenvalues per flavor
1	$(\alpha^+)^2$	$m + 1$
2	$(\alpha^-)^2$	$m - 1$
3	$\alpha^+\beta^-$	$m + 1$
4	$\alpha^-\beta^+$	$m - 1$
5	$\zeta$ with $\alpha^+$ , $\beta^-$ and $+\sqrt{\quad}$	$(m + 1) \left(\frac{m}{2} - 1\right)$
6	$\zeta$ with $\alpha^+$ , $\beta^-$ and $-\sqrt{\quad}$	$(m + 1) \left(\frac{m}{2} - 1\right)$
7	$\zeta$ with $\alpha^-$ , $\beta^+$ and $+\sqrt{\quad}$	$(m - 1) \left(\frac{m}{2} - 1\right)$
8	$\zeta$ with $\alpha^-$ , $\beta^+$ and $-\sqrt{\quad}$	$(m - 1) \left(\frac{m}{2} - 1\right)$

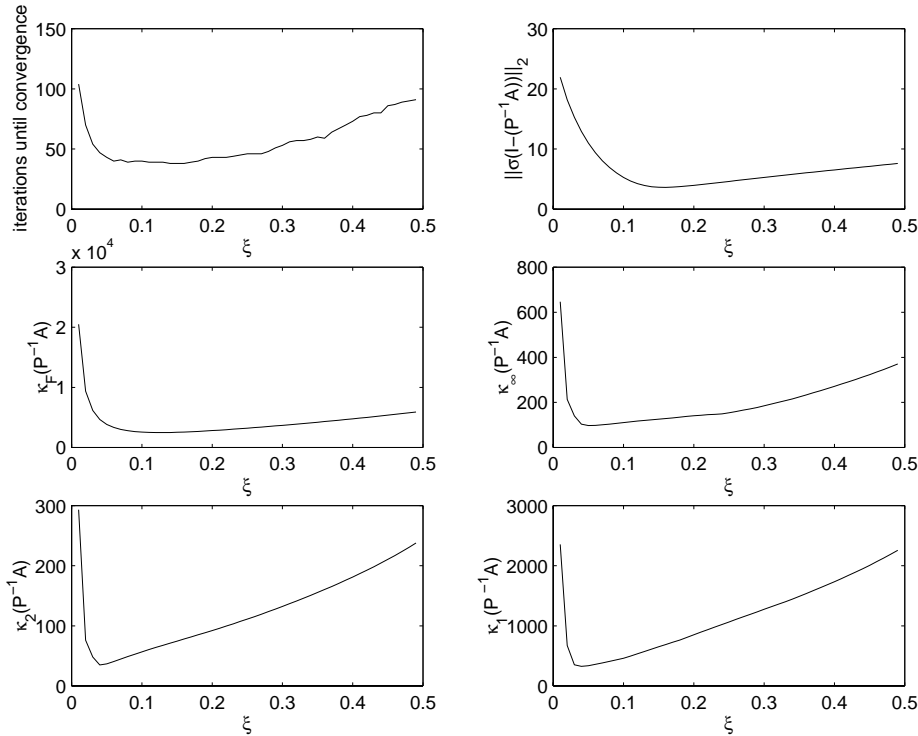


FIG. 2. Comparison of numerical results, condition numbers and  $\|\sigma(I - P^{-1}A)\|_2$  for Poisson's equation for  $m = 20$

Figures 4 through 11 show the eigenvalues associated with each of the eight flavors for  $\xi = 0.01, 0.05, 0.10, 0.15, 0.20,$  and  $0.40$  and  $m = 10$ . The fact that we use the specific value  $m = 10$  should not obfuscate what occurs in the general case, for the eigenvalues are merely values of continuous functions (i.e., the eight flavors) evaluated at the discrete points defined by  $\theta_j$  ( $j = 0, 1, \dots, m$ ) and  $\theta_k$  ( $k = 1, 2, \dots, \frac{m}{2} - 1$ ) where  $\theta_i = \frac{i\pi}{m}$ ,  $i = j$  or  $k$  (see Theorem 3.5).

For very small  $\xi$ , all flavors have a large real part. As  $\xi$  increases to 0.15, all eigenvalues associated with all flavors cluster tightly around the origin, with the exception of those of flavor 7, which has a few large real eigenvalues. As  $\xi$  increases further, we see that the eigenvalues of flavors 1, 2, 3, 5, and 6 remain tightly clustered about the origin, while eigenvalues of flavor 7 gain more outliers and eigenvalues of flavors 4 and 8 disperse significantly. This behavior is clearly not a function of  $m$ , but rather of the eight flavors. This explains why the optimal value of  $\xi$  (i.e., the value corresponding to greatest clustering) is approximately 0.154, irrespective of problem size  $m$ .

We also compare the ease of finding the value of  $\xi$  that minimizes condition numbers as compared to minimizing  $\|\sigma(I - P^{-1}A)\|_2$ . Assuming sufficiently large  $m$ , the value of  $\xi$  that minimizes  $\|\sigma(I - P^{-1}A)\|_2$  is seen *a priori* to be about 0.154. On the other hand, computing condition numbers, let alone minimizing them with respect to  $\xi$ , is an exceedingly difficult task. Indeed, both [11] and [12] report on methods which estimate

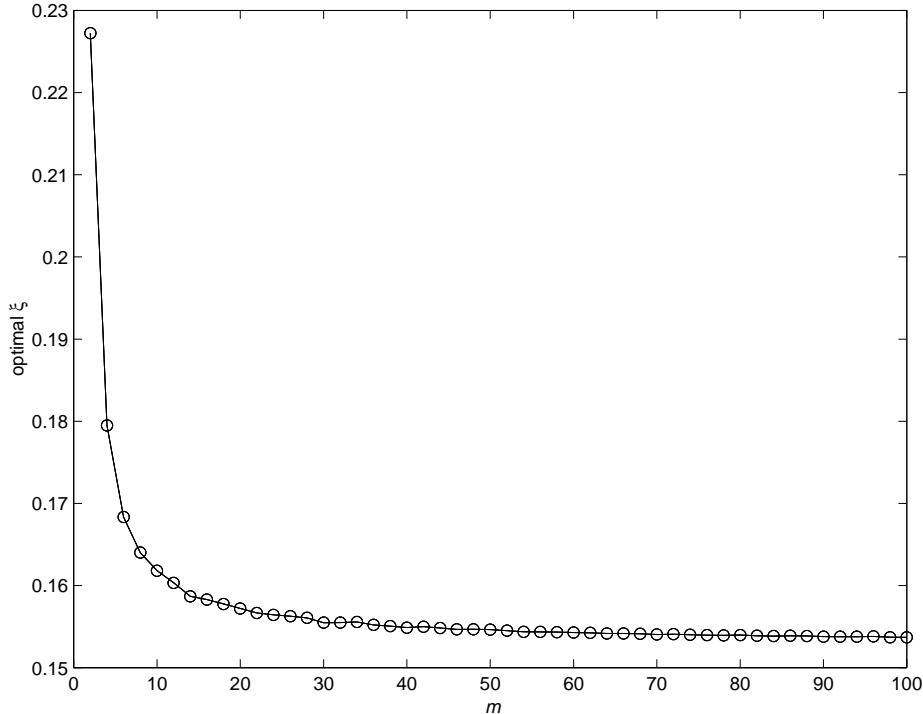


FIG. 3. Optimal value of  $\xi$  for various values of  $m$  for Poisson's equation

only the *order of magnitude* of condition numbers. However, it is evident that if we wish to use condition numbers to determine the optimal value of  $\xi$ , we require more precise information than mere orders of magnitude. And, as is readily seen from Figures 1 and 2, the insensitivity of optimal  $\xi$  with respect to  $m$  that we find for  $\|\sigma(I - P^{-1}A)\|_2$  does not hold for the condition numbers, negating the possibility of *a priori* optimality.

Finally, we examine the issue of the tradeoff between increasing the speed of convergence of Bi-CGSTAB/RBGS using the optimal value of  $\xi$  and the commensurate loss of accuracy. To address this question, we solved Poisson's equation with Dirichlet boundary conditions for two cases in which the analytical solution is known. Let  $\mathbf{v}$  be the vector whose entries are the values of  $u$ , the solution of Poisson's equation, at the interior mesh points of our finite element domain. Since the analytical solution is known, we had the Bi-CGSTAB/RBGS iterations stop when

$$\|\mathbf{v}_{\text{analytical}} - \mathbf{v}_{\text{numerical}}\|_{\infty} < \epsilon,$$

where  $\epsilon$  is a given tolerance. For all reported run times, we compiled an average of five individual runs.

We solved both of these PDEs with problem sizes  $m = 10, 20, 30, 40$ . For each of these values of  $m$ , we ran our code using both the Gaussian value of  $\xi = \frac{1}{\sqrt{12}}$  and the optimal value of  $\xi$  as determined from the MINOS optimization software. For  $\epsilon$ , we used values

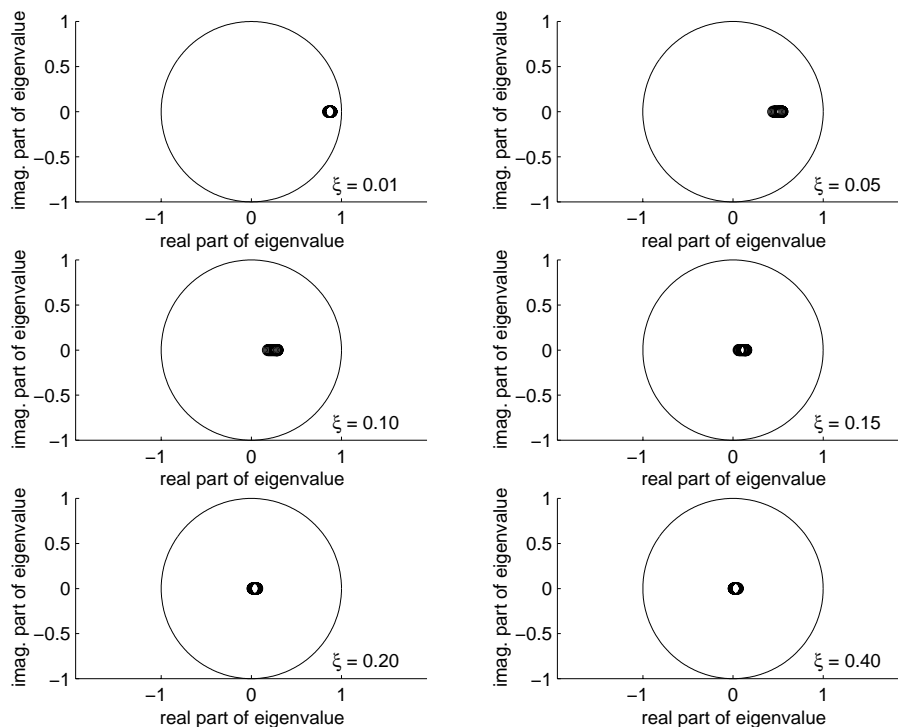


FIG. 4. Eigenvalues of flavor 1 for Poisson's equation for various values of  $\xi$  for  $m = 10$

of the form

$$\epsilon = 10^{-j}, \quad (4.1)$$

$j = 2, 3, 4, 5, 6$ . Because the Gaussian value of  $\xi = \frac{1}{\sqrt{12}}$  provides  $\mathcal{O}(h^4)$  accuracy while the optimal value of  $\xi$  provides only  $\mathcal{O}(h^2)$  accuracy, there are sufficiently large values of  $j$  in (4.1) for which Bi-CGSTAB/RBGS converges when using the Gaussian value of  $\xi$  but for which it fails to converge when using the optimal value of  $\xi$ .

In the numerical experiments, the analytical solutions used are

$$u(x, y) = \sin x \cos y \quad (4.2)$$

and

$$u(x, y) = \exp(2x + y). \quad (4.3)$$

The results are given in Tables I and II, respectively.

In Tables I and II, the first column is problem size  $m$  while the second column is the corresponding value of the optimal value of  $\xi$ , as determined from MINOS. Column three gives the tolerance  $\epsilon$  from (4.1). Columns four and five (respectively, six and seven) give the average run time (in seconds) and number of iterations required to reach convergence using the optimal value of  $\xi$  (respectively, the Gaussian value of  $\xi$ ). The eighth and final column gives the “improvement” one achieves using the optimal value of  $\xi$  instead of its

TABLE I. Results for Poisson's equation with analytical solution (4.2)

$m$	$\xi_{\text{opt}}$	$\log_{10} \epsilon$	$\xi_{\text{opt}}$		$\xi = \frac{1}{\sqrt{12}}$		% impr
			time	its	time	its	
10	0.16182	-2	0.013781	4	0.017010	5	18.98
		-3	0.016990	5	0.020255	6	16.12
		-4	0.017031	5	0.023534	7	27.63
		-5			0.026753	8	
20	0.15721	-2	0.118716	8	0.149149	10	20.40
		-3	0.160718	11	0.180121	12	10.77
		-4	0.162537	11	0.204349	14	20.46
		-5	0.175132	12	0.231695	16	24.41
		-6			0.263750	18	
30	0.15547	-2	0.479577	12	0.594175	15	19.29
		-3	0.565698	14	0.753442	19	24.92
		-4	0.632023	16	0.932896	23	32.25
		-5	0.713848	18	0.986615	25	27.65
		-6			1.063438	27	
40	0.15488	-2	1.255842	17	1.555441	21	19.26
		-3	1.481409	20	1.841658	25	19.56
		-4	1.695135	23	2.199651	30	22.94
		-5	1.840153	25	2.497365	34	26.32
		-6			2.571315	35	

TABLE II. Results for Poisson's equation with analytical solution (4.3)

$m$	$\xi_{\text{opt}}$	$\log_{10} \epsilon$	$\xi_{\text{opt}}$		$\xi = \frac{1}{\sqrt{12}}$		% impr
			time	its	time	its	
10	0.16182	-2	0.017062	5	0.024341	7	29.91
		-3			0.026751	8	
20	0.15721	-2	0.148577	10	0.204783	14	27.45
		-3	0.162621	11	0.238577	16	31.84
		-4			0.252492	17	
30	0.15547	-2	0.595449	15	0.788397	20	24.47
		-3	0.669760	17	0.909203	23	26.34
		-4			0.949999	24	
40	0.15488	-2	1.432224	20	1.782991	25	19.67
		-3	1.638079	23	2.209249	31	25.85
		-4			2.355495	33	

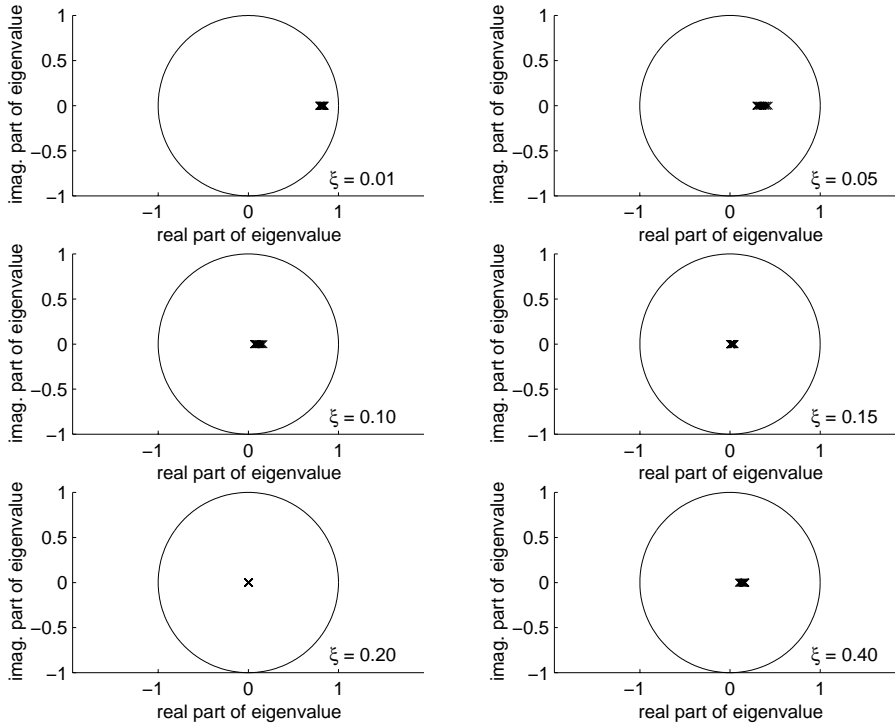


FIG. 5. Eigenvalues of flavor 2 for Poisson's equation for various values of  $\xi$  for  $m = 10$

Gaussian value. Improvement is defined

$$\frac{\text{time}(\text{Gaussian}) - \text{time}(\text{optimal})}{\text{time}(\text{Gaussian})} \times 100\%.$$

In examining Tables I and II, we see, for fixed accuracy and fixed problem size, that the convergence using the optimal value of  $\xi$  is always faster than that using the Gaussian value of  $\xi$ . The value of the “improvement” is never less than 10% and as high as more than 32%. The only times when the optimal value of  $\xi$  fails to do better than the Gaussian value is when  $\epsilon$  is so small that convergence with the optimal value of  $\xi$  is unattainable.

## V. SUMMARY

We derived analytical formulae for the eigenvalues that govern the rate at which the Bi-CGSTAB/RBGS method converges to the solution of the matrix equation arising from the Hermite collocation discretization of Poisson's equation. The eigenvalue formulae depend upon collocation point location, which can be chosen optimally to minimize the number of iterations required to converge to a predetermined tolerance. The optimal location of the collocation points is insensitive to problem size  $m$  for sufficiently large  $m$ . Results of numerical experiments indicate significant speedup when we use the optimal collocation point location as compared to the Gaussian location. Additionally, although

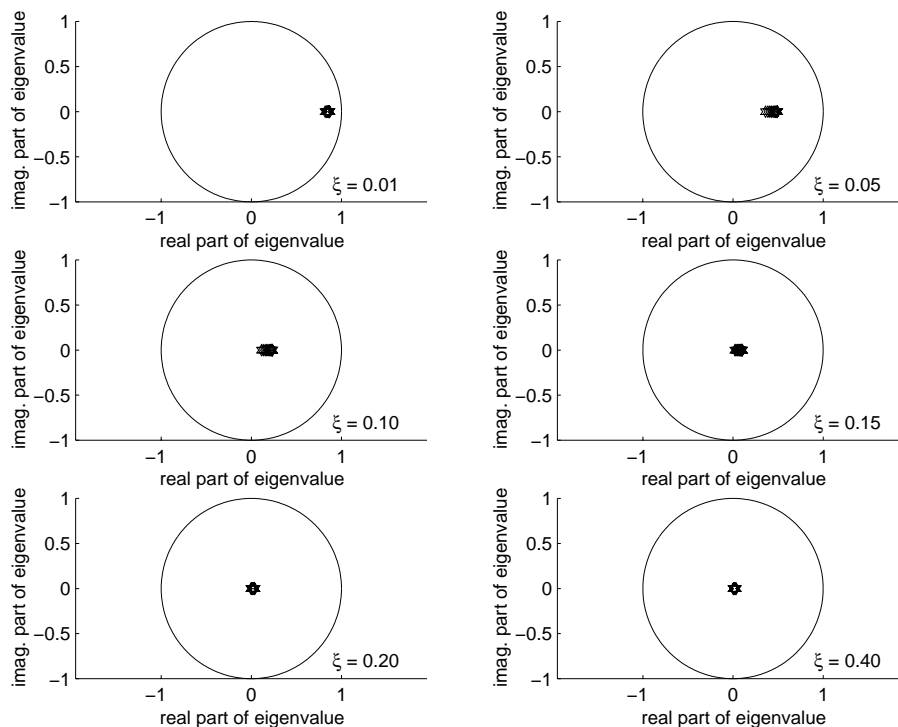


FIG. 6. Eigenvalues of flavor 3 for Poisson's equation for various values of  $\xi$  for  $m = 10$

the issue was not addressed in this work, our preconditioner makes our method of solution amenable to parallel processing.

## REFERENCES

1. Brill, S. H. Hermite Collocation Solution of Partial Differential Equations via Preconditioned Krylov Methods. *Numerical Methods for Partial Differential Equations*, in press.
2. van der Vorst, H. A. (1992). Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.*, 13:631-644.
3. Prenter, P. M. and Russell, R. D. (1976). Orthogonal Collocation for Elliptic Partial Differential Equations. *SIAM J. Numer. Anal.*, 13:923-939.
4. Saad, Y. and Schultz, M. H. (1985). Parallel Implementation of Preconditioned Conjugate Gradient Methods. Research Report YALEU/DCS/RR-425, Department of Computer Science, Yale University, New Haven, Connecticut.
5. Dyksen, W. R. and Rice, J. R. (1986) The Importance of Scaling for the Hermite Bicubic Collocation Equations. *SIAM J. Sci. Stat. Comput.*, 7:707-719.
6. Lai, Y.-L., Hadjidimos, A., Houstis, E. N., and Rice, J. R. (1995). On the Iterative Solution of Hermite Collocation Equations. *SIAM J. Matrix Anal. Appl.*, 16:254-277.

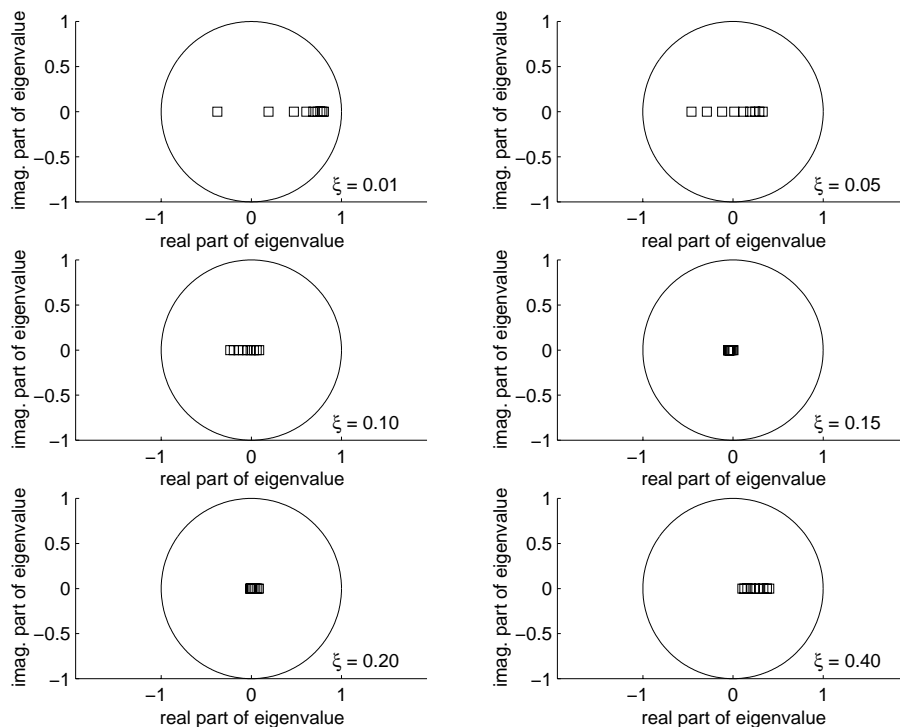


FIG. 7. Eigenvalues of flavor 4 for Poisson's equation for various values of  $\xi$  for  $m = 10$

7. Papatheodorou, T. S. (1983). Block AOR Iteration for Nonsymmetric Matrices. *Mathematics of Computation*, 41:511–525.
8. Cottle, R. W. (1974). Manifestations of the Schur Complement. *Linear Algebra and its Applications*, 8: 189–211.
9. Gohberg, I., Lancaster, P., and Rodman, L. (1982). *Matrix Polynomials*, Academic Press, New York.
10. Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*. Oxford University Press, London.
11. Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*, Third ed. The Johns Hopkins University Press, Baltimore.
12. Watkins, D. S. (1991). *Fundamentals of Matrix Computations*. John Wiley & Sons, Inc., New York.
13. Murtaugh, B. A. and Saunders, M. A. (1987). MINOS 5.1 User's Guide. Technical Report SOL 83-20R, Department of Operations Research, Stanford University, Stanford, California.

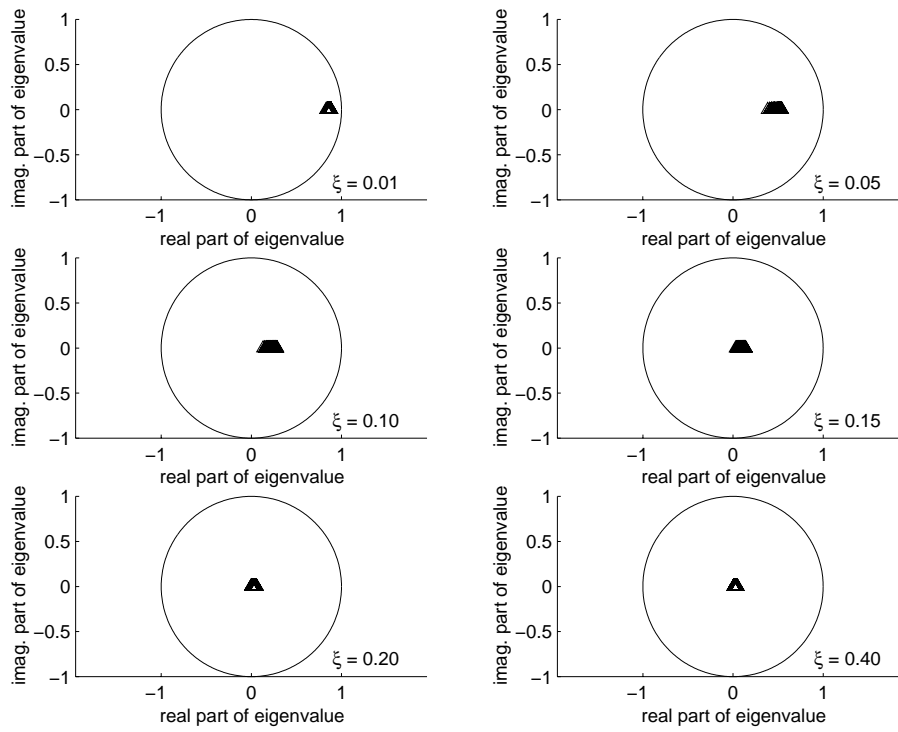


FIG. 8. Eigenvalues of flavor 5 for Poisson's equation for various values of  $\xi$  for  $m = 10$

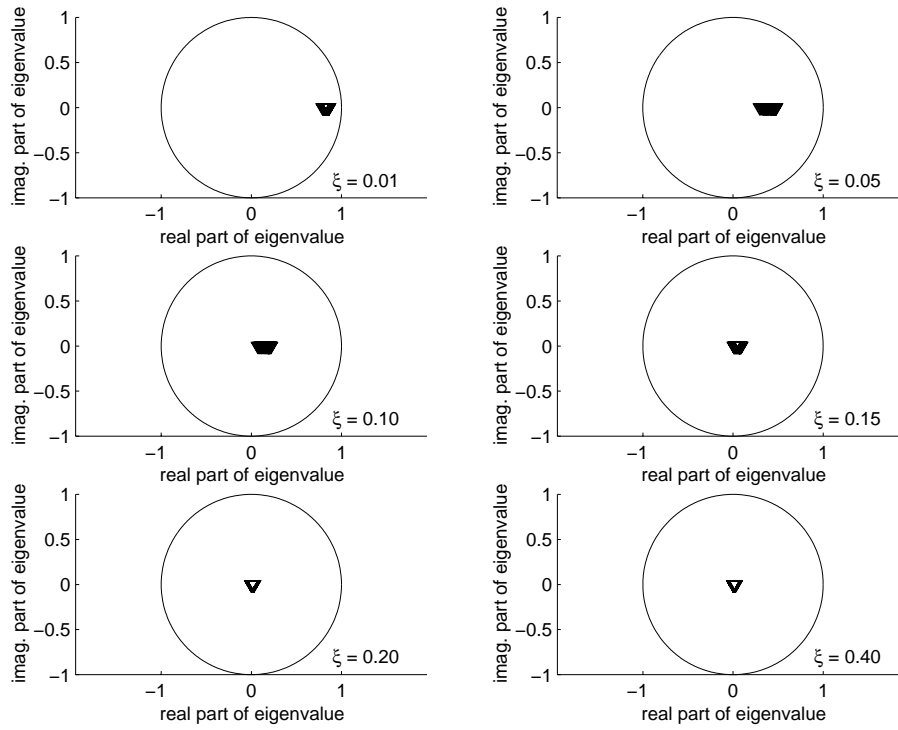


FIG. 9. Eigenvalues of flavor 6 for Poisson's equation for various values of  $\xi$  for  $m = 10$

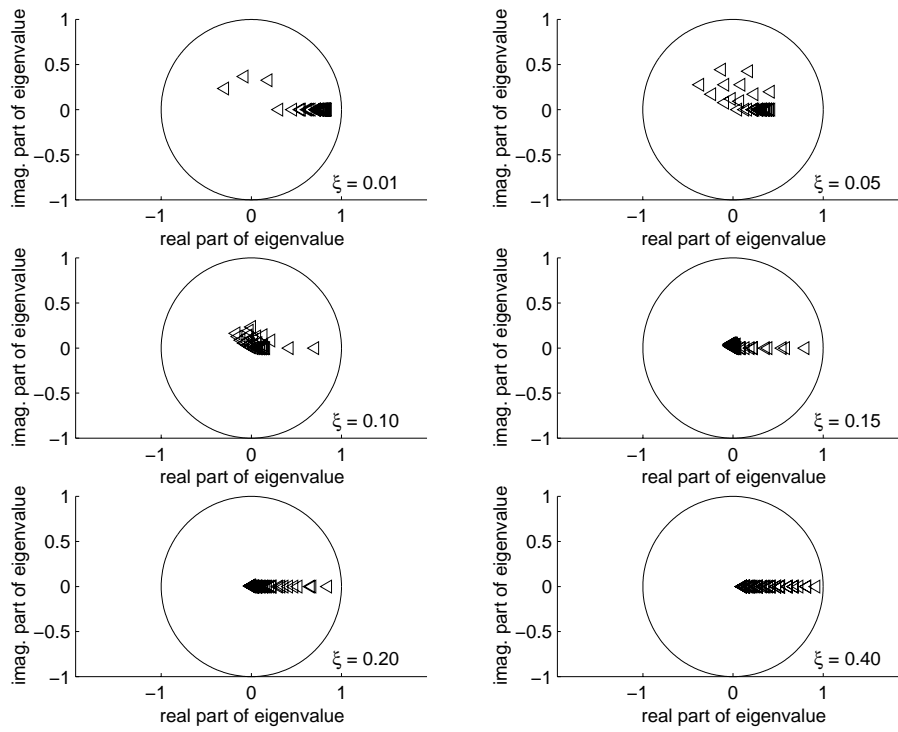


FIG. 10. Eigenvalues of flavor 7 for Poisson's equation for various values of  $\xi$  for  $m = 10$

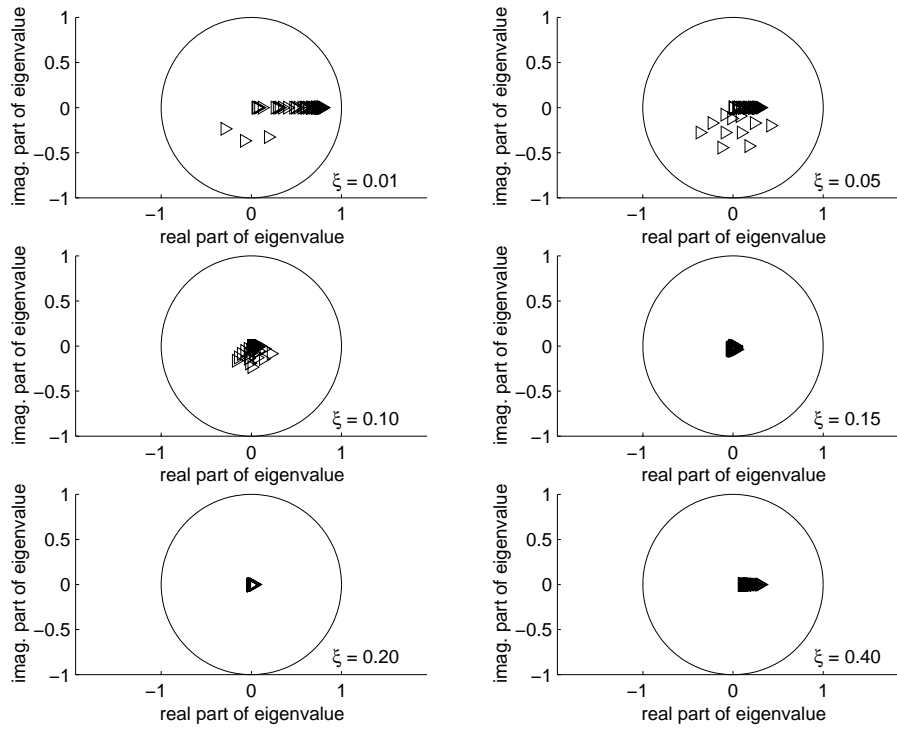


FIG. 11. Eigenvalues of flavor 8 for Poisson's equation for various values of  $\xi$  for  $m = 10$