NUMERICAL ANALYSIS PROJECT
MANUSCRIPT NA-92-04

MAY 1992

On the Convergence of Line Iterative Methods for Cyclically Reduced Non-Symmetrizable Linear Systems

by

Howard C. Elman
Gene H. Golub
Gerhard Starke

NUMERICAL ANALYSIS PROJECT
COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

ON THE CONVERGENCE OF LINE ITERATIVE METHODS FOR CYCLICALLY REDUCED NON-SYMMETRIZABLE LINEAR SYSTEMS

HOWARD C. ELMAN [*], GENE H. GOLUB[†] AND GERHARD STARKE[‡]

Abstract. We derive analytic bounds on the convergence factors associated with block relaxation methods for solving the discrete two-dimensional convection-diffusion equation. The analysis applies to the reduced systems derived when one step of block Gaussian elimination is performed on red-black ordered two-cyclic discretizations. We consider the case where centered finite difference discretization is used and one cell Reynolds number is less than one in absolute value and the other is greater than one. It is shown that line ordered relaxation exhibits very fast rates of convergence.

Key words. iterative methods, reduced system, convection-diffusion

[*] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, e-mail: elman@cs.umd.edu. The work of this author was supported by the U. S. Army Research Office under grant DAAL-0389-K-0016, and by the National Science Foundation under grants ASC-8958544 and CCR-8818340.

[†] Department of Computer Science, Stanford University, Stanford, CA 94305, e-mail: golub@sccm.stanford.edu. The work of this author was supported in part by the National Science Foundation under grant CCR-8821078.

[‡] Institut für Praktische Mathematik, Universität Karlsruhe, Englerstrasse 2, 7500 Karlsruhe 1, Federal Republic of Germany, e-mail: na.starke@na-net.ornl.gov. The work of this author was supported by the National Science Foundation under grant CCR-8821078 and by the Alexander-von-Humboldt-Stiftung.

1. Introduction. Consider the two-dimensional convection-diffusion equation

$$(1)\qquad \begin{aligned} -[(pu_x)_x + (qu_y)_y] + ru_x + su_y &= f \quad \text{on } \Omega \\ u &= g \quad \text{on } \partial\Omega \end{aligned}$$

where Ω is a smooth domain in $\mathbf{R}^2$ and $p > 0$, $q > 0$ on Ω. Let

$$(2)\qquad Au = f$$

denote the linear system obtained from a five-point finite difference discretization of (1), where the standard scheme [10] is used to discretize the second order term and centered differences are used for the first order term, and u and f now represent vectors in a finite dimensional space. Using a ***red-black** ordering,* this system can be written as

$$(3)\qquad \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} u^{(r)} \\ u^{(b)} \end{bmatrix} = \begin{bmatrix} f^{(r)} \\ f^{(b)} \end{bmatrix},$$

where A_{11} and A_{22} are diagonal matrices. If block elimination is used to decouple the "red" points $u^{(r)}$ ¿from the "black" points $u^{(b)}$, the result is the *reduced system*

$$(4)\qquad (A_{22} - A_{21}A_{11}^{-1}A_{12})u^{(b)} = f^{(b)} - A_{11}^{-1}f^{(r)}.$$

Let $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$, the Schur complement, denote the associated coefficient matrix.

Block (or line) iterative methods for solving (4) in the self-adjoint case have been considered by Parter [8] and Hageman and Varga [7], where it was shown that the rate of convergence was faster than if analogous methods are applied directly to (2). Parter and Steuerwalt [9] and Elman and Golub [3], [4], [5]showed that there are advantages to using the reduced system in the non-self-adjoint case as well. In particular, it was shown in [3], [4], [5] that line iterative methods are rapidly convergent when the reduced operator S can be symmetrized by a real diagonal similarity transformation. For constant coefficient problems, and problems where separation of variable techniques apply, this is the case when the absolute values of the discrete cell Reynolds numbers are either both less than one or both greater than one [3].

In this paper, we extend this analysis, for constant coefficient problems, to the case where the reduced operator is not real symmetrizable. For centered differences this corresponds to the case where one cell Reynolds number is greater than one in absolute value and the other is less than one in absolute value. Although numerical experiments in [3] and [4] showed the method to be effective in this case, there were no analytic bounds on the convergence rate. The key observation is that the reduced matrix can always be transformed into a *complex* symmetric matrix using a diagonal similarity transformation. We will restrict our attention to the two-line ordering considered in [9] and [4], since this is a natural way of ordering the cyclically reduced system in terms of convergence and parallelism.

An outline of the remainder of the paper is as follows. In §2, we describe the centered difference discretization and the reduced matrix corresponding to the two-line ordering, outline the technique used to transform this matrix into a complex symmetric matrix, and define the block Jacobi and Gauß-Seidel iteration strategies. In §3, we state the main results: bounds on the spectral radii of the iteration operators in the cases when one cell Reynolds number is small and the other is large. In addition, we present the results of numerical experiments showing that the analysis is a good indicator of numerical performance. In §4, we present the proofs of the analytic results. Finally, in §5, we make some observations on related methods.

2. The discrete problem and symmetrized reduced system. Assume that the discretization is made on a uniform $n \times n$ grid, and let $h = 1/(n+1)$. Let

$$\mathrm{diag}(d_1, d_2, \ldots, d_n)$$

denote a diagonal matrix of order n, and let

$$\mathrm{tridiag}(X_{j,j-1}, X_{jj}, X_{j,j+1}) \tag{5}$$

denote a block tridiagonal matrix. We will use the convention that when upper case characters appear in the representation (5), then the entries are themselves matrices, and if lower case characters are used, then the entries are scalars. If a natural ordering is used for the. discrete grid, then the coefficient matrix has the form

$$A = \mathrm{tridiag}(A_{j,j-1}, A_{jj}, A_{j,j+1}),$$

where

$$\begin{aligned} A_{j,j-1} &= \mathrm{diag}(b_{1j}, \ldots, b_{nj}), \\ A_{jj} &= \mathrm{tridiag}(c_{ij}, a_{ij}, d_{ij}), \\ A_{j,j+1} &= \mathrm{diag}(e_{1j}, \ldots, e_{nj}). \end{aligned}$$

A detailed description of the coefficients of A derived from the centered difference discretization, as well as those of the reduced matrix, is given in [5]; we omit such a description here.

If the operator of (1) is separable, then the discrete coefficients satisfy

$$\begin{aligned} a_{ij} &= a_i^{(x)} + a_j^{(y)}, & & \\ b_{ij} &= b_j\,, & d_{ij} &= d_i, \\ c_{ij} &= c_i\,, & e_{ij} &= e_j\,. \end{aligned}$$

Our convergence analysis depends on the following result concerning a symmetrized form of the reduced matrix.

THEOREM 2.1. *If the* operator of(l) *is separable, then there is a diagonal matrix Q such that $\hat{S} \equiv Q^{-1}SQ$ is a complex symmetric matrix.*
The proof is exactly analogous to the proofs of Theorem 3 of [3], and Theorem 1 of [5], and we omit the details. It makes use of the fact that a general irreducible nonsymmetric tridiagonal matrix can be transformed into a complex symmetric tridiagonal matrix (see, e.g. Cullum and Willoughby [2, Lemma 6.3.2]). If $c_i d_{i-1}$ and $b_j e_{j-1}$ have the same sign for all i and j, then Q and $\hat{S}$ can be chosen to be real.

We perform our convergence analysis for the constant coefficient problem, i.e.,

$$p(x) \equiv q(y) \equiv 1, \quad r(x) \equiv \sigma, \quad s(y) \equiv \tau.$$

In this case, we have

$$\begin{aligned} a_{ij} &\equiv 4, \\ b_{ij} &\equiv -(1+\tau h/2)\,, \quad d_{ij} \equiv -(1-\sigma h/2)\,, \\ c_{ij} &\equiv -(1+\sigma h/2)\,, \quad e_{ij} \equiv -(1-\tau h/2)\,. \end{aligned}$$

The quantities $\gamma = \sigma h/2$ and $\delta = \tau h/2$ are the *cell Reynolds numbers.* We consider the natural two-line ordering, i.e., where grid points are blocked into certain pairs of lines (see Figure 1).

$\times_{13}$	·	$\times_{15}$	·	$\times_{17}$	·
·	$\times_{14}$	·	$\times_{16}$	·	$\times_{18}$
$\times_{7}$	·	$\times_{9}$	·	$\times_{11}$	·
·	$\times_{8}$	·	$\times_{10}$	·	$\times_{12}$
$\times_{1}$	·	$\times_{3}$	·	$\times_{5}$	·
·	$\times_{2}$	·	$\times_{4}$	·	$\times_{6}$

FIG. 1. ***Natural*** two-line ordering for a 6×6 ***grid.***

The points in the k'th block are those with grid indices (i, j) such that $k - 1 < j/2 \leq k$. Since the coefficient matrix is block consistently ordered, the results also apply for block red-black orderings. See [4], [5] for further details. We will restrict our attention to the case where n is even. Modifications for odd n are obvious but make the analysis more complicated. Under these assumptions, the reduced matrix S has block tridiagonal form

$$S = \text{tridiag}(S_{j,j-1}, S_{jj}, S_{j,j+1})$$

where tho diagonal blocks for $j = 2, \ldots, n/2 - 1$ are given by

$$(6) \qquad S_{jj} = \begin{bmatrix} \tilde{p}_0 & p_1 & p_2 & & & \\ p_{-1} & p_0 & & & & \\ p_{-2} & & & \ddots & & \\ & & & & & p_2 \\ & & & & p_0 & p_1 \\ & & & p_{-2} & p_{-1} & \tilde{p}_0 \end{bmatrix} \equiv S_d$$

with $p_{-2} = -c^2, p_{-1} = -2ce, p_1 = -2bd, p_2 = -d^2, p_0 = a^2 - 2be - 2cd$ and $\tilde{p}_0 = a^2 - 2be - cd$. The first and the last diagonal blocks have slightly different form:

$$S_{11} = S_d + be\, E_{01}$$

with $E_{01} = \text{diag}(0, 1, \ldots, 0, 1)$ and

$$S_{n/2,n/2} = S_d + be\, E_{10}$$

with $E_{10} = \text{diag}(1, 0, \ldots, 1, 0)$. The off-diagonal blocks have the irregular tridiagonal form

$$S_{j,j-1} = - \begin{vmatrix} b^2 & 0 & & & & & \\ 2bc & b^2 & 2bd & & & & \\ & 0 & b^2 & 0 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 2bc & b^2 & 2bd & \\ & & & & 0 & b^2 & 0 \\ & & & & & 2bc & b^2 \end{vmatrix}, \quad j = 2, \ldots, n/2,$$

$$S_{j,j+1} = -\begin{bmatrix} e^2 & 2de & & & & & \\ 0 & e^2 & 0 & & & & \\ & 2ce & e^2 & 2de & & & \\ & & & \cdot & & & \\ & & & 0 & e^2 & 0 & \\ & & & & 2ce & e^2 & 2de \\ & & & & & 0 & e^2 \end{bmatrix}, \; j = 1, \ldots, n/2 - 1.$$

(See [4, §4].)

The symmetrized matrix $\hat{S} = Q^{-1}SQ$ is given by

$$\hat{S} = \mathrm{tridiag}(\hat{S}_{j,j-1}, \hat{S}_{jj}, \hat{S}^T_{j,j-1})$$

with

$$\hat{S}_j = \begin{bmatrix} \tilde{p}_0 & \hat{p}_1 & \hat{p}_2 & & & \\ \hat{p}_1 & p_0 & \ddots & \ddots & & \\ \hat{p}_2 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & \hat{p}_2 \\ & & \ddots & \ddots & p_0 & \hat{p}_1 \\ & & & \hat{p}_2 & \hat{p}_1 & \tilde{p}_0 \end{bmatrix} \equiv \hat{S}_d \,, \; j = 2, \ldots, n/2 - 1 \,, \tag{7}$$

where $\hat{p}_1 = -2\sqrt{bcde}, \hat{p}_2 = -cd$ and $p_0, \tilde{p}_0$ are as in (6). The first and the last diagonal blocks are given by

$$S_{11} = \hat{S}_d + be\, E_{01} \tag{8}$$

and

$$\hat{S}_{n/2,n/2} = \hat{S}_d + be\, E_{10} \,, \tag{9}$$

respectively. The subdiagonal blocks are defined by

$$S_{j,j-1} = -\begin{bmatrix} be & 0 & & & & & \\ 2\sqrt{bcde} & b\ e & 2\sqrt{bcde} & & & & \\ & 0 & be & 0 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 2\sqrt{bcde} & b\ e & 2\sqrt{bcde} & \\ & & & & 0 & be & 0 \\ & & & & & 2\sqrt{bcde} & b\ e \end{bmatrix}, \; j = 2, \ldots, n/2.$$

If $bcde$ is positive, then $\hat{S}$ is a real symmetric matrix; this is the case, for example, if $|\gamma|$ and $|\delta|$ arc either both less than one or both greater than one. These cases were addressed in [3],[4], [5]. Our primary interest here is in the case where $bcde$ is negative, i.e., either $be > 0$ ($|\delta| < 1$) and $cd < 0$ ($|\gamma| > 1$), or $be < 0$ ($|\delta| > 1$) and $cd > 0$ ($|\gamma| < 1$). In these cases $\hat{S}$ is complex symmetric.

Let D denote the block diagonal matrix whose entries are the diagonal blocks $\{S_{jj}\}$, and let C denote the block off-diagonal part $D - S$. The asymptotic rate of convergence of the two-line Jacobi method for solving the reduced system is determined from $\rho(D^{-1}C)$, and because S has block property A and is consistently ordered, the asymptotic rate of convergence for the two-line Gauß-Seidel method is determined from $(\rho(D^{-1}C))^2$ (*see* [10, p. 107], [11, p. 147]). We have $\hat{D} = Q^{-1}DQ$ and $\hat{C} = Q^{-1}CQ$ and, thus, for the two-line Jacobi operator

$$\rho(D^{-1}C) = \rho(Q^{-1}D^{-1}CQ) = \rho(\hat{D}^{-1}\hat{C}) = \rho(\hat{C}\hat{D}^{-1}) . \tag{10}$$

¿From (10), we obtain the bound

$$\rho(D^{-1}C) \leq \|\hat{C}\hat{D}^{-1}\|_2 . \tag{11}$$

The straightforward way to obtain bounds for the convergence of the two-line Jacobi operator would be to use

$$\rho(D^{-1}C) \leq \|\hat{C}\hat{D}^{-1}\|_2 \leq \|\hat{C}\|_2\|\hat{D}^{-1}\|_2 . \tag{12}$$

This approach is equivalent to that used for the cases treated in [3] and [4], when S can be transformed into a real symmetric matrix, and in that case the resulting bounds correspond closely to experimental results. For the case $bcde < 0$, this approach produces pessimistic results, and it is necessary to examine $\|\hat{C}\hat{D}^{-1}\|_2$ more carefully. The details of this analysis will be presented in §4.

3. Statement of main results and numerical examples. Our main result is as follows. The proof is deferred to $4.

THEOREM 3.1. *For even n, we have, for the two-line ordered cyclically reduced system, the following bounds for* $\|\hat{C}\hat{D}^{-1}\|_2$:

$$\frac{2be}{a^2-2be} + 4\sqrt{|bcde|}\frac{[((a^2-4cd)^2-4be(a^2-be))^2-16(be)^3cd]^{1/2}}{[(a^2-4cd)^2-4be(a^2-be)]^{3/2}} \tag{13}$$

for $be > 0$ *and* $\frac{a^2-\sqrt{a^4+3(a^2-2be)^2}}{12} \leq cd < 0$,

$$\frac{2be}{a^2-2be} + \frac{}{((a^2-4cd)^2-4be(a^2-be))^{1/2}}\left[1+\frac{}{a^2(a^2-2be)^2}\right]^{1/2} \tag{14}$$

for $be > 0$ *and* $-\frac{a^2-2be}{4} \leq cd \leq \frac{a^2-\sqrt{a^4+3(a^2-2be)^2}}{12}$,

$$\frac{2be}{a^2-2be} + \left(\frac{be}{a^2+be}\right)^{1/2}\left[1+\frac{(be)^3}{a^2(a^2-2be)^2}\right]^{1/2} \tag{15}$$

for $be > 0$ *and* $cd \leq -\frac{a^2-2be}{4}$, *and*

$$\frac{2|be|}{[(a^2-4cd)^2+4|be|(a^2+|be|)]^{1/2}} \tag{16}$$

$$+4\sqrt{|bcde|}\frac{[((a^2-4cd)^2+4|be|(a^2+|be|))^2+16|be|^3cd]^{1/2}}{[(a^2-4cd)^2+4|be|(a^2+|be|)]^{3/2}}$$

for $be < 0$ *and* $cd > 0$.

We remark that we only consider the case $bcde < 0$ in Theorem 3.1 since the (real) symmetrizable case $bcde \geq 0$ has been addressed in [4]. However, for be < 0 and cd < 0 ($|\gamma| > 1$ and $|\delta| > 1$) the bounds of [4] can also be improved using the techniques described below.

Note that Theorem 3.1 does not only give us an upper bound for $\rho(D^{-1}C)$, the asymptotic convergence factor of the method, but also a rigorous estimate for the error reduction. since

$$\|(D^{-1}C)^k\|_2 = \|Q\hat{D}^{-1}(\hat{C}\hat{D}^{-1})^{k-1}\hat{C}Q^{-1}\|_2 \leq \text{cond}_2(Q)\frac{\|\hat{D}^{-1}\|_2\|\hat{C}\|_2}{\|\hat{D}^{-1}\hat{C}\|_2}\|\hat{D}^{-1}\hat{C}\|_2^k .$$

COROLLARY 3.2. *For the centered difference scheme, if* $|\gamma| > 1$ *and* $|\delta| < 1$, *then* $\|\hat{C}\hat{D}^{-1}\|_2$ *is bounded by*

$$\frac{1-\delta^2}{7+\delta^2}+2\sqrt{(1-\delta^2)(\gamma^2-1)}\frac{[(4(\gamma^2+3)^2-(1-\delta^2)(15+\delta^2))^2-(1-\delta^2)^3(1-\gamma^2)]^{1/2}}{[4(\gamma^2+3)^2-(1-\delta^2)(15+\delta^2)]^{3/2}} \tag{17}$$

for $\gamma^2 \leq \frac{\sqrt{64+3(7+\delta^2)^2}-2}{6}$,

$$\frac{1-\delta^2}{7+\delta^2}+\frac{2\sqrt{(1-\delta^2)(\gamma^2-1)}}{[4(\gamma^2+3)^2-(1-\delta^2)(15+\delta^2)]^{1/2}}\left[1+\frac{(1-\delta^2)^3}{64(7+\delta^2)^2}\right]^{1/2} \tag{18}$$

for $\frac{\sqrt{64+3(7+\delta^2)^2}-2}{6} \leq \gamma^2 \leq \frac{9+\delta^2}{2}$, *and*

$$\frac{1-\delta^2}{7+\delta^2}+\left(\frac{1-\delta^2}{17-\delta^2}\right)^{1/2}\left[1+\frac{(1-\delta^2)^3}{64(7+\delta^2)^2}\right]^{1/2} \tag{19}$$

for $\gamma^2 \geq \frac{9+\delta^2}{2}$.

Proof. The inequalities for $\|\hat{C}\hat{D}^{-1}\|_2$ result from inserting be $= 1-\delta^2$ and cd $= 1-\gamma^2$ into (13), (14) and (15), resp., of Theorem 3.1. ☐

We remark that for be > 0 and cd < 0 the bound on the right hand side of (12) leads to

$$\rho(P\ C) \leq \frac{4\sqrt{|bcde|}+2be}{a^2-2be} = \frac{4\sqrt{(1-\delta^2)(\gamma^2-1)}+2(1-\delta^2)}{14+\delta^2}$$

which, for the centered difference scheme, if $|\delta| < 1$, obviously tends to ∞ as $\gamma \to \infty$. In contrast, all three bounds in Corollary 3.2 are less than 1, which implies that the method actually converges for any choice of $|\gamma| > 1$ and $|\delta| < 1$.

Table 1 shows different upper bounds for the spectral radius of the Gauß-Seidel operator associated with the two-line ordered cyclically reduced system. The columns indicated by $h = 1/8$, $h = 1/16$ and $h = 1/32$ contain the computed numbers $\|\hat{C}\hat{D}^{-1}\|_2^2$, the last column contains the (asymptotic) bounds from Corollary 3.2. A comparison with [4, Table 6.1] shows a surprising conformity of $\|\hat{C}\hat{D}^{-1}\|_2^2$ and the Gauß-Seidel spectral radius.

As $\gamma \to \infty$, our bound for $\|\hat{C}\hat{D}^{-1}\|_2$ tends to $1/7+\frac{\sqrt{3137}}{56\sqrt{17}} \approx 0.1486$ which implies that the limiting value for the spectral radius of the Gauß-Seidel operator is bounded by 0.1486. For comparison, if SOR with optimal relaxation parameter is applied to the non-reduced system, a convergence factor of 0.1885 is obtained for large values of y (see Chin and Manteuffel [1]).

TABLE 1
Computed values and bounds for $\|\hat{C}\hat{D}^{-1}\|_2$ *for two-line Gauß-Seidel,* 6 = 0

γ	$h = 1/8$	$h = 1/16$	$h = 1/32$	bound
1.0	0.0134	0.0180	0.0197	0.0204
1.2	0.0368	0.0410	0.0426	0.0954
1.4	0.0523	0.0562	0.0575	0.1278
1.6	0.0622	0.0652	0.0662	0.1463
1.8	0.0677	0.0698	0.0705	0.1562
2.0	0.0701	0.0713	0.0717	0.1603
3.0	0.0680	0.0713	0.0715	0.1486

The case $|\gamma| > 1$ and $|\delta| < 1$ corresponds to orienting the Gauß-Seidel sweeps orthogonal to the direction of flow associated with (1). It was observed by Chin and Manteuffel in [1] that, for the non-reduced problem, this orientation gives better convergence rates than if sweeps are oriented in the direction of flow, for large cell Reynolds numbers. We will see below that the same is true for the cyclically reduced system. In particular, if y = 0 and $|\delta| > 1$, our ordering of the unknowns corresponds to sweeps in the direction of flow and the numerical results as well as our analytical bounds give larger convergence factors in this case although both the computed values of $\|\hat{C}\hat{D}^{-1}\|_2$ and analytic bounds are still small for moderate sizes of δ. This can be seen in Table 2. We have also observed that for larger values of 6 than those in the table, the computed values of $\|\hat{C}\hat{D}^{-1}\|_2$ become larger than 1, so that we cannot even guarantee convergence for arbitrary δ in this case.

COROLLARY 3.3. *For the centered difference scheme, if* $|\gamma| < 1$ *and* $|\delta| > 1$, *then we have*

$$(20) \qquad \rho(D^{-1}C) \leq \|\hat{C}\hat{D}^{-1}\|_2 \leq \frac{\delta^2 - 1}{[4(\gamma^2 + 3)' + (\delta^2 - 1)(15 + \delta^2)]^{1/2}}$$

$$+2\sqrt{(\delta^2 - 1)(1 - y")}\,\frac{[(4(\gamma^2 + 3)' + (5" - 1)(15 + \delta^2))^2 + (\delta^2 - 1)^3(1 - \gamma^2)]^{1/2}}{[4(\gamma^2 + 3)" + (52 - 1)(15 + \delta^2)]^{3/2}}.$$

TABLE 2
Computed values and bounds for $\|\hat{C}\hat{D}^{-1}\|_2$ *for two-line* Gauss-Seidel, $\gamma = 0$

δ	$h = 1/8$	$h = 1/16$	$h = 1/32$	bound
1.0	0	0	0	0
1.2	0.0369	0.0426	0.0445	0.1335
1.4	0.0745	0.0858	0.0896	0.2591
1.6	0.1125	0.1289	0.1345	0.3745
1.8	0.1506	0.1716	0.1788	0.4791
2.0	0.1890	0.2139	0.2222	0.5732
3.0	0.3800	0.4159	0.4243	0.9098

4. **Analysis.** We now present the proof of Theorem 3.1. Let

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

so that

$$\hat{C} = 2\sqrt{bcde}\ \text{tridiag}(U_n^T, 0, U_n) + be\ \text{tridiag}(I_n, 0, I_n). \tag{21}$$

We will bound $\|\hat{C}\hat{D}^{-1}\|_2$ using

$$\|\hat{C}\hat{D}^{-1}\|_2 \leq 2\sqrt{|bcde|}\ \|\text{tridiag}(U_n^T, 0, U_n)\hat{D}^{-1}\|_2 + |be|\ \|\text{tridiag}(I_n, 0, I_n)\|_2\|\hat{D}^{-1}\|_2. \tag{22}$$

Note that

$$\|\hat{D}^{-1}\|_2 = \max\{\|\hat{S}_{11}^{-1}\|_2, \|\hat{S}_d^{-1}\|_2, \|\hat{S}_{n/2,n/2}^{-1}\|_2\}, \tag{23}$$

and, using $U_nU_n^T + U_n^TU_n = (U_n + U_n^T)^2 - U_n^2 - (U_n^T)^2 = (U_n + U_n^T)^2 = T_n^2$,

$$\begin{aligned} &\|\text{tridiag}(U_n^T, 0, U_n)\hat{D}^{-1}\|_2 \\ &\leq \max\{\rho(\bar{\hat{S}}_{11}^{-1} U_nU_n^T\hat{S}_{11}^{-1}), \rho(\bar{\hat{S}}_d^{-1} T_n^2\hat{S}_d^{-1}), \rho(\bar{\hat{S}}_{n/2,n/2}^{-1} U_n^TU_n\hat{S}_{n/2,n/2}^{-1})\}^{1/2} \\ &= \max\{\|U_n^T\hat{S}_{11}^{-1}\|_2, \|T_n\hat{S}_d^{-1}\|_2, \|U_n\hat{S}_{n/2,n/2}^{-1}\|_2\}. \end{aligned} \tag{24}$$

The proof' proceeds by first bounding the terms in the center of (23) and (24), i.e.,

$$\|S_d^{-1}\|_2 \text{ and } \|T_n\hat{S}_d^{-1}\|_2,$$

which correspond to the interior blocks of the matrices. After that, we will derive bounds for the first and last blocks in terms of these quantities. It should be noted at this point that the numbers $a^2 - 2be$ and $a^2 - 4cd$ appearing frequently in the following analysis are always positive.

LEMMA **4.1.** *The matrix norm $\|\hat{S}_d^{-1}\|_2$ satisfies the following bounds:*

$$\|\hat{S}_d^{-1}\|_2 \leq \frac{1}{a^2 - 2be} \quad \text{for } be > 0 \text{ and } cd < 0, \tag{25}$$

$$\|\hat{S}_d^{-1}\|_2 \leq \frac{1}{((a^2 - 4cd)^2 + 4|be|(a^2 + |be|))^{1/2}} \quad \text{for } be < 0 \text{ and } cd > 0. \tag{26}$$

Moreover, for the matrix norm $\|T_n\hat{S}_d^{-1}\|_2$, we have

$$\|T_nS_d^{-1}\|_2 \leq 2(|cd|(a^2 + be))^{1/2} \quad \text{for } be > 0 \text{ and } cd \leq -\frac{a^2 - 2be}{4}, \tag{27}$$

$$\|T_n\hat{S}_d^{-1}\|_2 \leq \frac{2}{((a^2 - 4cd)^2 + 4|be|(a^2 + |be|))^{1/2}} \tag{28}$$

for $be > 0$ and $-\frac{a^2-2be}{4} \le cd < 0$ and for $be < 0$ and $cd \ge 0$.

Proof. From Lemmas 2 and 3 in [4], we have $\hat{S}_d = (a^2-2be)I_n - 2\sqrt{bcde}T_n - cdT_n^2$. It follows that

$$(29) \qquad \|\hat{S}_d^{-1}\|_2 = \rho(\hat{S}_d^{-1}) \le \max_{x\in[-2,2]} \frac{1}{|a^2 - 2be - 2\sqrt{bcde}x - cdx^2|}$$

$$= \max_{x\in[-2,2]} \frac{1}{[(a^2-cdx^2)^2 - 4be(a^2-be)]^{1/2}}.$$

The maximum is attained at x = 0 for $cd < 0$ and at $x = \pm 2$ for $cd > 0$ which leads to (25) and (26). Analogously, we have

$$(30) \qquad \|T_n\hat{S}_d^{-1}\|_2 = \rho(T_n\hat{S}_d^{-1}) \le \max_{x\in[-2,2]} \left|\frac{x}{a^2 - 2be - 2\sqrt{bcde}x - cdx^2}\right|$$

$$= \max_{x\in[-2,2]} \frac{|x|}{[(a^2-cdx^2)^2 - 4be(a^2-be)]^{1/2}}$$

which, after an exercise in elementary calculus, leads to (27) and (28). □

We will need the following result to handle the other terms of (24).

LEMMA 4.2. *For the tridiagonal Toeplitz matrix T_n = tridiag(1, 0, 1),*

$$(31) \qquad T_nE_{10} = E_{01}T_n, \quad T_nE_{01} = E_{10}T_n.$$

Moreover, there exist matrices V_n, W_n that satisfy the following relations:

$$(32) \qquad V_nE_{10} = E_{10}V_n, \quad V_nE_{01} = E_{01}V_n,$$

$$(33) \qquad W_nE_{10} = E_{01}W_n, \quad W_nE_{01} = E_{10}W_n,$$

$$(34) \qquad \hat{S}_d^{-1}E_{01} = E_{01}V_n + E_{10}W_n, \ E_{01}\hat{S}_d^{-1} = V_nE_{01} + W_nE_{10}$$

and

$$(35) \qquad \hat{S}_d^{-1}E_{10} = E_{10}V_n + E_{01}W_n, \ E_{10}\hat{S}_d^{-1} = V_nE_{10} + W_nE_{01}.$$

In addition, V_n is positive definite and the norm of W_n is bounded by

$$(36) \quad \|W_n\|_2 \le \frac{4\sqrt{|bcde|}}{(a^2+4|cd|)^2 - 4be(a^2-be)} \text{ for } cd \ge \frac{a^2 - \sqrt{a^4 + 3(a^2-2be)^2}}{12},$$

$$(37) \quad \|W_n\|_2 \le \frac{\sqrt{be}}{a(a^2-2be)} \text{ for } be > 0 \text{ and } cd \le \frac{a^2 - \sqrt{a^4 + 3(a^2-2be)^2}}{12}.$$

Proof. The verification of (31) is straightforward. We will only prove the first part of (34) since the second part and (35) are completely analogous. We wish to find matrices V_n and W_n such that

$$\hat{S}_d^{-1}E_{01} = E_{01}V_n + E_{10}W_n$$

holds. This is true if and only if

$$\begin{aligned} E_{01} &= \hat{S}_d E_{01} V_n + \hat{S}_d E_{10} W_n \\ &= E_{01}[((a^2 - 2be)I_n - cdT_n^2)V_n - 2\sqrt{bcde}T_n W_n] \\ &\quad + E_{10}[((a^2 - 2be)I_n - cdT_n^2)W_n - 2\sqrt{bcde}T_n V_n] \,. \end{aligned}$$

It is sufficient to find V_n and W_n that satisfy

$$\begin{aligned} 2\sqrt{bcde}T_n V_n &= ((a^2 - 2be)I_n - cdT_n^2)W_n \,, \\ ((a^2 - 2be)I_n - cdT_n^2)V_n - 2\sqrt{bcde}T_n W_n &= I_n \,. \end{aligned}$$

Obviously, these equations are solved by

$$\begin{aligned} V_n &= ((a^2 - 2be)I_n - cdT_n^2)((a^2 I_n - cdT_n^2)^2 - 4be(a^2 - be)I_n)^{-1} \,, \\ W_n &= 2\sqrt{bcde}T_n((a^2 I_n - cdT_n^2)^2 - 4be(a^2 - be)I_n)^{-1} \,, \end{aligned}$$

the relations (32) and (33) can now be obtained from (31). Since the matrix V_n is a rational function in T_n,

$$\sigma(V_n) = \{\frac{a^2 - 2be - cd\lambda^2}{(a^2 - cd\lambda^2)^2 - 4be(a^2 - be)} : \lambda \in \sigma(T_n)\} \,.$$

This implies that all the eigenvalues of the symmetric matrix V_n are bounded below by

$$\min_{x\in[-2,2]} \frac{a^2 - 2be - cdx^2}{(a^2 - cdx^2)^2 - 4be(a^2 - be)} > 0 \,,$$

and therefore V_n is positive definite. The norm bounds (36) and (37) follow from

$$||W_n||_2 = \rho(W_n) \le 2\sqrt{|bcde|} \max_{x\in[-2,2]} \left| \frac{x}{(a^2 - cdx^2)^2 - 4be(a^2 - be)} \right| \,.$$

□

LEMMA **4.3.** *For the first and last blocks of* $\hat{D}^{-1}$, *we have*

(38) $$||\hat{S}_{11}^{-1}||_2 = ||\hat{S}_{n/2,n/2}^{-1}||_2 \le \begin{cases} ||\hat{S}_d^{-1}||_2 & \text{if } be > 0, \\ (1 + |be|\,||\hat{S}_d^{-1}||_2)||\hat{S}_d^{-1}||_2 & \text{if } be < 0 \,. \end{cases}$$

Analogously, for the norm of the first and last blocks of tridiag$(U_n^T, 0, U_n)\hat{D}^{-1}$, *we have*

(39) $$||U_n^T \hat{S}_{11}^{-1}||_2 = ||U_n \hat{S}_{n/2,n/2}^{-1}||_2 \le (1 + (be)^2||W_n||_2^2)^{1/2}||T_n \hat{S}_d^{-1}||_2$$

where W_n is the matrix defined by (4) in the proof of Lemma 4.2.

Proof. Consider the case $be > 0$. From (8) and the fact that $x^H \hat{S}_d x$ is located in the right half plane for any $x \in \mathbb{C}^n$, $x \neq 0$, we have

$$|x^H \hat{S}_{11} x| = |x^H \hat{S}_d x + be(|x_2|^2 + \cdots + |x_n|^2)| \le |x^H \hat{S}_d x| \,.$$

Moreover, since S_d is normal, this leads us to

$$\min_{||y||_2=1} ||\hat{S}_d y||_2 = \min_{||y||_2=1} |y^H \hat{S}_d y| \le \min_{||y||_2=1} |y^H \hat{S}_{11} y| \le \min_{||y||_2=1} ||\hat{S}_{11} y||_2 \,,$$

which implies $||\hat{S}_{11}^{-1}||_2 \leq ||\hat{S}_d^{-1}||_2$.

For the case be < 0, we apply the Sherman-Morrison-Woodbury formula (see [6, p. 51]) to (I), giving

$$\hat{S}_{11}^{-1} = \hat{S}_d^{-1} - be\hat{S}_d^{-1}E_{01}(I_n + be\hat{S}_d^{-1}E_{01})^{-1}\hat{S}_d^{-1} . \tag{40}$$

Moreover, by (34), we have

$$E_{01}\hat{S}_d^{-1}E_{01} = E_{01}(E_{01}V_n + E_{10}W_n) = E_{01}V_n = E_{01}V_nE_{01}$$

which leads to

$$(I_n + beE_{01}V_n)E_{01} = E_{01}(I_n + be\hat{S}_d^{-1}E_{01}) . \tag{41}$$

Since $|be|||\hat{S}_d^{-1}||_2 < 1$ (see (25) and (26)) and since V_n is positive definite, the existence of $(I_n + beE_{01}\hat{S}_d^{-1})^{-1}$ and $(I_n + beE_{01}V_n)^{-1}$ is guaranteed. With this, we obtain from (41)

$$E_{01}(I_n + be\hat{S}_d^{-1}E_{01})^{-1} = E_{01}(I_n + beE_{01}V_n)^{-1} . \tag{42}$$

Inserting (42) into (40) leads to

$$\hat{S}_{11}^{-1} = \hat{S}_d^{-1} - be\hat{S}_d^{-1}E_{01}(I_n + beE_{01}V_n)^{-1}\hat{S}_d^{-1} . \tag{43}$$

But V_n is positive definite (see Lemma 4.2), so that by (32)

$$E_{01}V_n = E_{01}V_nE_{01}$$

is positive semi-definite and we have

$$||(I_n + beE_{01}V_n)^{-1}||_2 \leq 1 . \tag{44}$$

The second bound of (38) then follows from (43) and (44).

For (39), we have $U_n^T = T_n E_{10}$, so that

$$||U_n^T\hat{S}_{11}^{-1}||_2 = ||T_nE_{10}\hat{S}_{11}^{-1}||_2 .$$

Then, from (43) and Lemma 4.2, we have

$$\begin{aligned}
T_nE_{10}\hat{S}_{11}^{-1} &= T_nE_{10}\hat{S}_d^{-1} - beT_nE_{10}\hat{S}_d^{-1}E_{01}(I_n + beE_{01}V_n)^{-1}\hat{S}_d^{-1} \\
&= T_nE_{10}\hat{S}_d^{-1} - beT_nE_{10}W_n(I_n + beE_{01}V_n)^{-1}\hat{S}_d^{-1} && \text{by (34)} \\
&= T_nE_{10}\hat{S}_d^{-1} - beT_nW_nE_{01}(I_n + beE_{01}V_n)^{-1}\hat{S}_d^{-1} && \text{by (33)} \\
&= E_{01}\left[I_n - be(I_n + beE_{01}V_n)^{-1}W_n\right]T_n\hat{S}_d^{-1} && \text{by (31) and (33)}
\end{aligned}$$

where we have used the facts that $E_{10}E_{01} = E_{01}E_{10} = 0$, and that T_n commutes with V_n and W_n . Consequently, (44) implies

$$||T_nE_{10}\hat{S}_{11}^{-1}||_2 \leq \left(1 + (be)^2||W_n||_2^2\right)^{1/2} ||T_n\hat{S}_d^{-1}||_2 .$$

□

Proof of Theorem 3.1. ¿From (22) we get, using (23), (24), (38), (39) and

$$||\text{tridiag}(I_n, 0, I_n)||_2 \leq 2\cos(\frac{2\pi h}{1+h}) < 2 ,$$

$$||\hat{C}\hat{D}^{-1}||_2 \le 2\sqrt{|bcde|}(1+(be)^2||W_n||_2^2)^{1/2}||T_n\hat{S}_d^{-1}||_2 + 2|be|||\hat{S}_d^{-1}||_2$$

for be > 0 and $cd < 0$, and

$$||\hat{C}\hat{D}^{-1}||_2 \le 2\sqrt{|bcde|}(1+(be)^2||W_n||_2^2)^{1/2}||T_n\hat{S}_d^{-1}||_2 + 2|be|(1+|be|||\hat{S}_d^{-1}||_2)||\hat{S}_d^{-1}||_2$$

for $be < 0$ and cd > 0. All that is left to do now, is to insert the bounds (25) and (26) for $||\hat{S}_d^{-1}||_2$, (27) and (28) for $||T_n\hat{S}_d^{-1}||_2$, (36) and (37) for $||W_n||_2$. This leads to (13), (14), (15) and (16) of Theorem 3.1. □

5. **Concluding remarks.** We conclude with some observations relating these results lo t hose of [3], [4], and [5]. There are three essential differences in the analysis:

1. the use of a complex diagonal similarity similarity transformation to produce the complex symmetric matrix $\hat{S}$;
2. the use of (22) rather than (12) to derive bounds on $\rho(D^{-1}C)$;
3. the use of the Sherman-Morrison-Woodbury formula in Lemma 4.3, to bound the norms of first and last blocks of the block diagonal matrix $\hat{D}$.

Item (1) can be thought of as a generalization of the real diagonal similarity transformation used in [3] et. al. The more careful analysis determined by items (2) and (3) were not, necessary for the cases of small cell Reynolds or upwind difference schemes considered in [R] et. al. In those cases, the nonzero entries of $\hat{C}$ are real numbers of the same sign, and the analysis based on (12) is essentially tight; similarly, the block diagonals satisfy $||\hat{S}_{11}^{-1}||_2 \le ||\hat{S}_d^{-1}||_2$ (essentially as in the proof of the first part of Lemma 4.3), and there is no need for a careful perturbation analysis. We remark, however, that the techniques of items (2) and (3) can be used to improve the previous results in the case where both cell Reynolds numbers are greater than one in absolute value.

Finally, although we have restricted our attention to the two-line ordering considered in [4], it is also possible to derive analogous bounds for the one-line ordering discussed in [3]. For this case, however, the analysis is considerably less clean and the results not as strong, so we do not consider them here.

REFERENCES

[1] R. C. Y. Chin and T. A. Manteuffel. An analysis of block successive overrelaxation for a class of matrices with complex spectra. *SIAM J. Numer. Anal.*, 25:564–585, 1988.

[2] J. K. Cullum and R. A. Willoughby. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations.* Birkhäuser, Boston, 1985.

[3] H. C. Elman and G. H. Golub. Iterative methods for cyclically reduced non-self-adjoint linear systems. *Math. Comp.*, 54:671–700, 1990.

[4] H. C. Elman and G. H. Golub. Iterative methods for cyclically reduced non-self-adjoint linear systems II. *Math. Comp.*, 56:215–242, 1991.

[5] H. C. Elman and G. H. Golub. Line iterative methods for cyclically reduceddiscrete convection-diffusion problems. *SIAM J. Sci. Statist. Comput.*, 13:339–363, 1992.

[6] G. H. Golub and C. F. V. Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore - London, 2nd edition, 1989.

[7] L. A. Hageman and R. S. Varga. Block iterative methods for cyclically reduced matrix equations. *Numer. Math.*, 6:106–119, 1964.

[8] S. V. Parter. On estimating the "rates of convergence" of iterative methods for elliptic difference equations. *Trans. Amer. Math. Soc.*, 114:320–354, 1965.

[9] S. V. Parter and M. Steuerwalt. Block iterative methods for elliptic and parabolic difference equations. *SIAM J. Numer. Anal.*, 19:1173–1195, 1982.

[10] R. S. Varga. *Matrix Iterative Analysis.* Prentice Hall, New York, 1962.

[11] D. M. Young. *Iterative Solution of Large Linear Systems.* Academic Press, New York - London, 1971.