# Block Orderings for Tensor-Product Grids in Two and Three Dimensions [1]

Gene H. Golub[2], Chen Greif[3] and James M. Varah[4]

**Abstract.** We consider two-line and two-plane orderings for a convection-diffusion model problem in two and three dimensions, respectively. These strategies are aimed at introducing dense diagonal blocks, at the price of a slight increase of the bandwidth of the matrix, compared to natural lexicographic ordering. Comprehensive convergence analysis is performed for block stationary schemes. We then move to consider a two-step preconditioning technique, and analyze the numerical properties of the underlying linear systems that are solved in each step of the iterative process. For the three-dimensional problem this approach is a viable alternative to the Incomplete LU approach, and may be easier to implement in parallel environments. The analysis is illustrated and validated by numerical examples.

**AMS Subject Classification:** 65F10, 65F50.

**Keywords:** Sparse linear systems, Discretization of PDEs, Orderings, Convergence of iterative solvers.

## 1. Introduction.

Consider the following convection-diffusion model equation with constant coefficients

$$-\Delta u + V^T \nabla u = p \ , \tag{1.1}$$

subject to Dirichlet type boundary conditions. Problem (1.1) is either in two or three dimensions, and the domain is the unit square or the unit cube, respectively. We denote the coefficients of the convective terms by

$$V = (\sigma, \tau, \mu)^T, \tag{1.2}$$

where in two dimensions $V$ consists of $\sigma$ and $\tau$ only. We focus on applying finite difference discretizations: centered differences to the diffusive terms, and centered differences or first order upwind approximations to the convective terms. Let us define $n$ and $h$ so that $h = 1/(n+1)$ is the (uniform) mesh size, and let $L$ denote the corresponding difference operator, after scaling by $h^2$, so that the discrete solution of an interior gridpoint in the three-dimensional case, $u_{i,j,k}$, satisfies:

$$\begin{aligned} L\,u_{i,j,k} = a\,u_{i,j,k} + b\,u_{i,j-1,k} + c\,u_{i-1,j,k} \\ +d\,u_{i+1,j,k} + e\,u_{i,j+1,k} + f\,u_{i,j,k-1} + g\,u_{i,j,k+1} \ . \end{aligned} \tag{1.3}$$

[1] Accepted for publication in Numerical Algorithms, Dec. 20 2001.

[2] SCCM Program, Gates 2B, Stanford University, Stanford CA 94305, USA, golub@sccm.stanford.edu. The work of this author was supported in part by the National Science Foundation, Grant No. DOE DE-FC02-01ER41177.

[3] Parametric Technology Corporation, 2590 North First St., Suite 200, San Jose CA 95131, USA, cgreif@ptc.com. The work of this author was done while he was at Stanford University, and was supported in part by the Natural Sciences and Engineering Research Council of Canada.

[4] Department of Computer Science, University of British Columbia, Vancouver B.C, V6T 1Z4, Canada, varah@cs.ubc.ca. The work of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada, Grant No. 88240.

The computational molecule for the three-dimensional case is depicted in Fig. 1.1. For notational convenience, let the difference equation for the 2D case be the same as (1.3), except $f = g \equiv 0$, and with double subscripts for $u$ rather than triple ones.

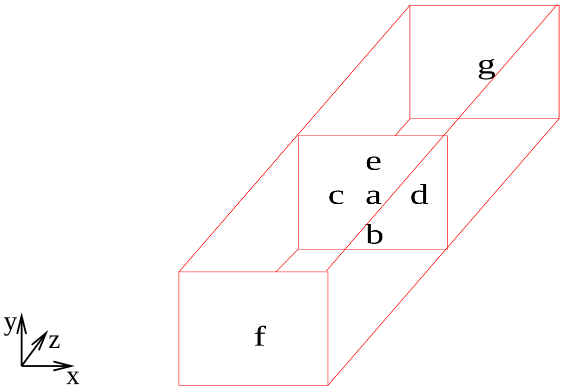


FIG. 1.1. *Computational molecule*

Denoting the mesh Reynolds numbers by

$$\beta = \frac{\sigma h}{2}, \quad \gamma = \frac{\tau h}{2}, \quad \delta = \frac{\mu h}{2}, \tag{1.4}$$

the coefficients $a, b, c, d, e, f$ and $g$ can be expressed in terms of $\beta, \gamma$ and $\delta$ [7].

Our purpose in this report is to introduce a class of orderings which are motivated by attempting to yield matrices with dense diagonal blocks. In other words, we aim at having in the computational molecule as many gridpoints as possible with an associated index in the matrix that is different from the index of the center of the molecule by a quantity that does not depend on $n$. The two-line ordering we discuss was proposed in [11], where it was considered for the classical positive-definite case. Here we consider the general convection-diffusion case and provide analysis for the nonsymmetric constant-coefficient model problem. We then proceed to introduce a two-plane ordering strategy for 3D tensor-product grids.

In the 2D case, gridpoints from two lines are alternately numbered, and in the 3D case, the numbering is done based on blocks of four lines, each consisting of two pairs of lines in two adjacent planes. Hence the names 'two-line' and 'two-plane'.

For stationary methods, it can be shown analytically that if the matrix is an $M$-matrix, strategies that rely on generating dense main diagonal blocks (such as the ones discussed here) lead to fast convergence [13, p. 91]. We provide comprehensive convergence analysis for the block Jacobi scheme. For preconditioned Krylov solvers we present some numerical experiments that illustrate the merits of our approach.

The rest of this paper is organized as follows. In sections 2 and 3 we present the ordering strategies that are examined, for the 2D and 3D problems respectively. In each of these sections a block splitting is proposed, bounds on convergence rates for the block Jacobi scheme are derived, and the amount of computational work involved in solving the linear systems is estimated. In section 4 we discuss a technique of two-step preconditioning, and apply it to the three-dimensional case. In section 5 some numerical results which validate our analysis are presented. In section 6 we draw some conclusions.

**2. Two-Line Orderings for 2D Problems.** We start with the two dimensional problem. Parter [11] introduced a technique of two-line ordering, which we now analyze for the class of nonsymmetric matrices arising from finite-difference discretization of convection-diffusion equations. The ordering (for a $4 \times 4$ grid) and the sparsity pattern of the matrix (for an $8 \times 8$ grid) are illustrated in Fig. 2.1. The idea is to number the unknowns in groups of two lines. The main diagonal block of this matrix 'captures' four of the five components of the computational molecule for all interior nodes. A two-line ordering strategy was considered for cyclically reduced problems in [2, 3], where grids of a different structure arise.

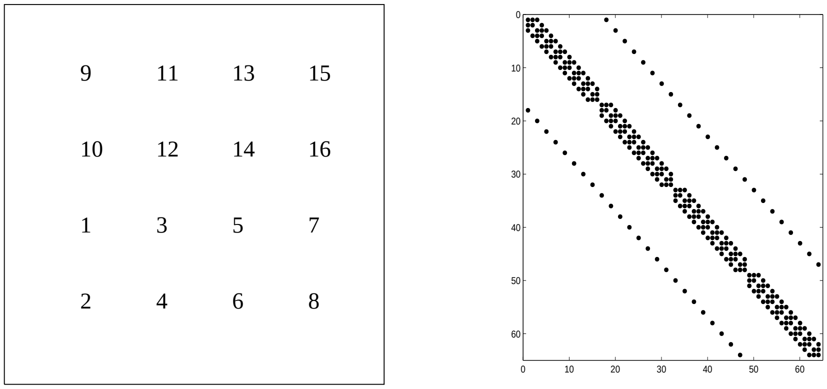


Fig. 2.1. *Two-line ordering*

**2.1. Convergence of the block Jacobi scheme.** Denote by $V_{10}$ a vector with alternating 1s and 0s, and let $V_1$ be a vector of ones. Let the two-line matrix be denoted by $A$, and let $A = M - N$ be a splitting, such that

$$M = \text{penta}[c \cdot V_1, e \cdot V_{10}, a \cdot V_1, b \cdot V_{10}, d \cdot V_1] \ . \tag{2.1}$$

We can think of $M - N$ as a 'two-line Jacobi' splitting. The spectrum of $M$ is given in the following theorem.

Theorem 2.1. *The eigenvalues of the matrix $M$ are given exactly by*

$$a \pm \sqrt{be} - 2\sqrt{cd}\cos(\pi j h) \ , \quad j = 1, \ldots, n \ . \tag{2.2}$$

*Each of these $2n$ eigenvalues has an algebraic multiplicity of $\frac{n}{2}$.*

*Proof.* It is straightforward to show that $M$ can be symmetrized by a diagonal nonsingular matrix, by means analogous to the techniques used in [2]. As a result, each $2n \times 2n$ block of the symmetrized matrix is a block-tridiagonal Toeplitz matrix, relative to $2 \times 2$ blocks, given by

$$\bar{M} = \text{tri}[\sqrt{be} \cdot I_2, G, \sqrt{be} \cdot I_2] \ , \tag{2.3}$$

where

$$G = \begin{pmatrix} a & \sqrt{cd} \\ \sqrt{cd} & a \end{pmatrix} , \tag{2.4}$$

and $I_2$ is the $2 \times 2$ identity matrix. The eigenvalues of $G$ are $a \pm \sqrt{cd}$. Let $\{\lambda, x\}$ be an eigenpair, so that $Gx = \lambda x$, and consider the $2n$-vector:

$$v = \begin{pmatrix} (\sin\alpha)x \\ (\sin 2\alpha)x \\ \vdots \\ (\sin n\alpha)x \end{pmatrix} . \tag{2.5}$$

The $j$th block row of $(\bar{M} - \lambda I)v$ is given by:

(2.6) $[(\bar{M} - \lambda I)v]^{(j)} = \sqrt{cd} \cdot [\sin(j-1)\alpha x + \sin(j+1)\alpha x] + (G - \lambda I)(\sin j\alpha)x$ .

Since $(G - \lambda I))(\sin j\alpha)x = 0$, we get after simplification

$$[(\bar{M} - \lambda I)v]^{(j)} = (2\sqrt{cd}\cos\alpha)v^{(j)} \ . \tag{2.7}$$

For $j = 1, \ldots, n-1$ this holds for *any* $\alpha$, however for $j = n$, it holds only if $\sin(n+1)\alpha = 0$. Thus $\alpha = \frac{k\pi}{n+1}$, $k = 1, \ldots n$, and $2\sqrt{cd}\cos(\frac{k\pi}{n+1})$ is an eigenvalue of $\bar{M} - \lambda I$. From this we get that the eigenvalues of $\bar{M}$ are given by (2.2).

Since the symmetrized matrix $M$ is a block diagonal matrix comprised of $\frac{n}{2}$ copies of $\bar{M}$, the multiplicity stated in the theorem follows. □

Theorem 2.1 leads to an expression for the minimal eigenvalue of $M$, as follows.

COROLLARY 2.2. *For $be, cd > 0$ the minimal eigenvalue of $M$ is given by*

$$a - 2\sqrt{cd} \cdot \cos(\pi h) - \sqrt{be} \ . \tag{2.8}$$

For the matrix $N$ the eigenvalues are given as follows.

PROPOSITION 2.3. *The eigenvalues of the matrix $N$ are either $0$, $\sqrt{b \cdot e}$ or $-\sqrt{b \cdot e}$.*

*Proof.* The matrix $N^2$ is a diagonal matrix whose entries are either $0$ or $b \cdot e$. □

Using the above results, we can establish a convergence result for the two-line Jacobi scheme:

THEOREM 2.4. *For $be, cd > 0$, the spectral radius of the Jacobi scheme associated with the splitting $A = M - N$ is bounded by*

$$\tilde{\rho}_{2L} = \frac{\sqrt{be}}{a - 2\sqrt{cd}\cos(\pi h) - \sqrt{be}} \ . \tag{2.9}$$

*Remark*: We attach the tilde symbol to $\rho$ in order to indicate that this is a bound rather than the actual spectral radius.

*Proof.* For positive $be$ and $cd$ the matrix is merely a permutation of the matrix arising from natural lexicographic ordering, which can be symmetrized by a real diagonal nonsingular matrix [2]. Since $M$ and $N$ are symmetric, with $M$ being positive definite, we have (see, for example, [2]) $\rho(M^{-1}N) \leq \dfrac{\rho(N)}{\lambda_{min}(M)}$. □

**2.2. Estimate of computational work.** In [2] it is shown that the spectral radius of the line-Jacobi iteration matrix is given (exactly) by

$$\rho_{NL} = \frac{2\sqrt{be}\cos(\pi h)}{a - 2\sqrt{cd}\cos(\pi h)} \ . \tag{2.10}$$

We have used Maple V to compute symbolically the Taylor expansion of $\tilde{\rho}_{2L}$, and have compared it to the Taylor expansion of $\rho_{NL}$.

The Taylor expansion about $h = 0$ of the Jacobi iteration matrix associated with the lexicographic ordering is

$$1 - \left(\pi^2 + \frac{1}{8}\sigma^2 + \frac{1}{8}\tau^2\right) h^2 + o(h^2).$$

For the bound on the spectral radius of the two-line matrix, the Taylor expansion is

$$1 - \left(\pi^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\tau^2\right) h^2 + o(h^2).$$

A useful quantity to compare here is the asymptotic rate of convergence, which for a given spectral radius $\rho$, is given by $-\ln \rho$. It can roughly indicate the number of iterations it takes to reduce the initial error by a fixed rate. If $\rho = 1 - ch^2$, with $c$ being a constant and $h$ being a small quantity (such as gridsize), then to leading order, we have $-\ln \rho \approx ch^2$.

From a point of view of iteration counts, then, when the convective coefficients are large, it is evident that compared to the scheme associated with the lexicographic ordering, the two-line Jacobi scheme converges faster by a factor of approximately 2.

When $\sigma$ and $\tau$ are very small, it seems by looking at the bound that the natural lexicographic ordering may have approximately the same rate of convergence (in this case the rate of convergence is dominated by the factor $\pi^2 h^2$ in the Taylor expansions given above). However, as is shown in section 5, our numerical experiments indicate that the two-line scheme is faster even for this case.

Table 2.1 provides some insight into the improved spectral radius, and the tightness of the bound.

| $n$ | $\rho_{NL}$ | $\rho_{2L}$ | $\tilde{\rho}_{2L}$ |
|---|---|---|---|
| 8 | 0.686 | 0.524 | 0.575 |
| 16 | 0.741 | 0.589 | 0.605 |
| 24 | 0.753 | 0.604 | 0.612 |
| 32 | 0.758 | 0.610 | 0.614 |

TABLE 2.1

*Spectral radii of the natural lexicographic iteration matrix ($\rho_{NL}$), the two-line iteration matrix ($\rho_{2L}$), and the bound ($\tilde{\rho}_{2L}$), for $\beta = \gamma = 0.5$, using the centered difference discretization for the first and second derivatives.*

## 3. Two-Plane Ordering for 3D Problems.

A strategy of two-plane ordering was proposed and analyzed for a different type of grids: the class of cyclically reduced problems [6, 7, 8], and is now adopted for tensor-product grids. It is based on numbering the nodes by dividing the grid into groups of $4n$ points, gathered from two lines and two planes. The idea is illustrated in Fig. 3.1, where both the ordering (for a $4 \times 4 \times 4$ grid) and the sparsity pattern of the underlying matrix (for a $6 \times 6 \times 6$ grid) are depicted.

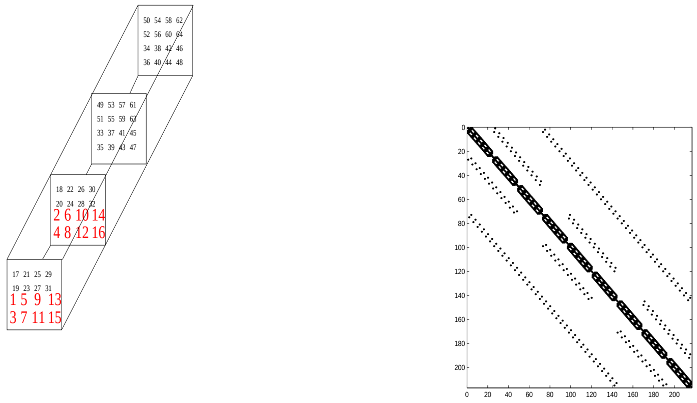
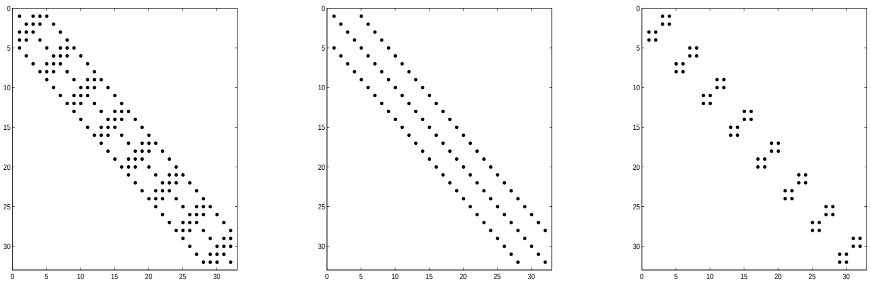


FIG. 3.1. *Two-Plane ordering*

**3.1. Convergence of the block Jacobi scheme.** Consider the splitting $A = M - N$, where $M$ is a block diagonal matrix whose semibandwidth is 4. A single block of $M$ is depicted in the leftmost graph in Fig. 3.2. In order to obtain an upper bound on the spectral radius of the iteration matrix, we use a technique of symmetrization.

PROPOSITION 3.1. *The two-plane matrix can be symmetrized by a real nonsingular diagonal matrix, provided that* $be, cd, fg > 0$.

*Proof.* The two-plane matrix is a symmetric permutation of the matrix associated with the natural lexicographic ordering, which is symmetrizable under the conditions stated above [7]. □



$M$ $M_1$ $M_2$

FIG. 3.2. *Single blocks of the matrix $M$ and its "splitting" into $M_1 + M_2$*

PROPOSITION 3.2. *Suppose $be, cd, fg > 0$. Then the minimal eigenvalue of $M$ is*

$$\lambda_{min}(M) = a - 2\sqrt{cd}\cos(\pi h) - \sqrt{be} - \sqrt{fg}\ . \tag{3.1}$$

*Proof.* By Proposition 3.1, $M$ can be symmetrized. In order to keep notation as simple as possible, we refer below to $M$ as the *symmetrized* matrix rather than the original nonsymmetric matrix. We split $M$ into two matrices, $M = M_1 + M_2$, as depicted in Fig. 3.2. (In the figures a single $4n \times 4n$ block of each of these matrices is given.) $M_1$ is a block diagonal matrix, consisting of $\frac{n^2}{4}$ blocks. Each of the blocks has three nonzero diagonals, located in diagonals -4, 0 and 4. The values in these diagonals are $c, a$ and $d$ respectively.

Let $\otimes$ denotes the Kronecker product operator. Let $Q_1 = \text{penta}[\sqrt{cd}, 0, a, 0, \sqrt{cd}]$ be a $2n \times 2n$ pentadiagonal matrix. $M_1$ can be written as a block pentadiagonal matrix relative to $2 \times 2$ blocks:

$$M_1 = Q_1 \otimes I_2 \ . \tag{3.2}$$

Let $Q_3 = \text{tri}[\sqrt{cd}, a, \sqrt{cd}]$, of size $n \times n$. Then

$$Q_1 = Q_3 \otimes I_2 \ . \tag{3.3}$$

$Q_3$ is a symmetric tridiagonal Toeplitz matrix. Thus its eigenvalues are known explicitly (see [2] or [12, p. 119]), and it follows that the eigenvalues of $Q_1$ are

$$\lambda_j = a + 2\sqrt{cd} \cdot cos(j\pi/(n+1)), \qquad j = 1, \ldots, n, \tag{3.4}$$

each of multiplicity $\frac{n^2}{4}$. Thus the eigenvalues of $M_1$ are (3.4), with algebraic multiplicity $\frac{n^2}{4}$.

Consider now the matrix $M_2$. Let $Q_2$ be a $2n \times 2n$ block diagonal matrix of the form $\text{diag}[Y, \ldots, Y]$, where

$$Y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ . \tag{3.5}$$

Define $T$ by:

$$T = \begin{pmatrix} \sqrt{be} & \sqrt{fg} \\ \sqrt{fg} & \sqrt{be} \end{pmatrix} \ . \tag{3.6}$$

Then $M_2 = Q_2 \otimes T$, and the eigenvalues of $M_2$ are $\pm\sqrt{be} \pm \sqrt{fg}$.

Any four matrices $A, B, C$ and $D$ with the appropriate sizes satisfy [1]

$$(A \otimes B) \cdot (C \otimes D) = (AC) \otimes (BD) \ . \tag{3.7}$$

Thus

$$M_1 \cdot M_2 = (Q_1 \otimes I_2) \cdot (Q_2 \otimes T) = (Q_1 Q_2) \otimes T \ . \tag{3.8}$$

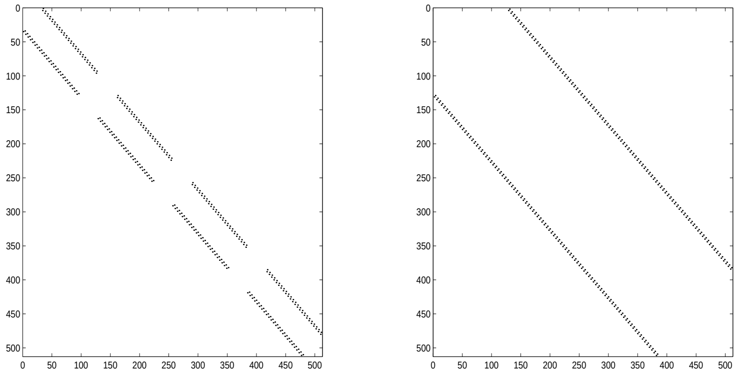
By (3.7), we have

$$Q_1 \cdot Q_2 = (Q_3 \otimes I_2) \cdot (I_n \otimes Y) = Q_3 \otimes Y = Q_2 \cdot Q_1 \ . \tag{3.9}$$

Since $Q_1$ and $Q_2$ commute, it follows from eq. (3.8) that

$$M_1 \cdot M_2 = M_2 \cdot M_1 \ , \tag{3.10}$$

by which it follows that $M_1$ and $M_2$ can be simultaneously diagonalized and have the same eigenvectors. Thus the sum of the exact minimal eigenvalues of $M_1$ and $M_2$ is also the minimal eigenvalue of $M = M_1 + M_2$. □



$N_1$ $N_2$

FIG. 3.3. *The matrices $N_1$ and $N_2$*

Next, we examine the matrix $N$. Here as well we split the matrix into two submatrices, $N_1 + N_2$, in the manner illustrated in Fig. 3.3, and in this case we are able to find the eigenvalues directly, without symmetrizing first.

PROPOSITION 3.3. *The spectral radius of the matrix $N$ is given by $\sqrt{be} + \sqrt{fg}$.*

*Proof.* Relative to $2n^2 \times 2n^2$ blocks, $N_1$ is a block diagonal matrix. The nonzero entries in the superdiagonals indexed $4n \pm 2$ are equal to $e$. On the other hand, the nonzero entries in the subdiagonals $-4n \pm 2$ are equal to $b$. Along each of the above diagonals, any two nonzero entries is separated by three zeros.

Relative to $2n^2 \times 2n^2$ blocks, $N_2$ is a block tridiagonal matrix with zero blocks on its main diagonal. The nonzero entries of the superdiagonals indexed $2n^2 + 1$ and $2n^2 - 3$ are equal to $g$. The entries on the subdiagonals indexed $-2n^2 + 3$ and $-2n^2 - 1$ are equal to either 0 or $f$. Here, as in the diagonals of $N_1$, each two nonzero entries along a diagonal are separated by three zeros.

Given the above, it is possible to perform direct computation of $N_1 \cdot N_2$ and $N_2 \cdot N_1$. The resulting matrices are block tridiagonal relative to $2n^2 \times 2n^2$ blocks. Subdividing each of these blocks into $4n \times 4n$ blocks, and taking into account the spacing of nonzero entries along each diagonal, it follows that the nonzero diagonals of the product are all diagonals with identical nonzero values along them, which are either *bf, ef, bg* or *eg*. Thus the matrices $N_1$ and $N_2$ commute, and hence have common eigenvectors.

The matrix $N_1^2$ is a diagonal matrix whose values on the diagonal are either $be$ or 0. Thus the eigenvalues of $N_1$ are given by $0, \pm\sqrt{be}$. Using a similar argument, $N_2^2$ is a diagonal matrix which has either zeros or $fg$ along the main diagonal, thus the eigenvalues of $N_2$ are either zeros or $\pm\sqrt{fg}$. From this it follows that the maximal eigenvalue of $N = N_1 + N_2$ is as stated in the proposition. □

The above results lead to a bound on the spectral radius of the two-plane Jacobi iteration matrix, as follows.

THEOREM 3.4. *An upper bound on the spectral radius of the two-plane Jacobi iteration matrix is given by*

$$\tilde{\rho}_{2P} = \frac{\sqrt{be} + \sqrt{fg}}{a - 2\sqrt{cd}\cos(\pi h) - \sqrt{be} - \sqrt{fg}} \; . \tag{3.11}$$

**3.2. Estimate of computational work.** The spectral radius of the one-line Jacobi iteration matrix associated with the natural lexicographic ordering is [7]

$$\frac{2(\sqrt{be}+\sqrt{fg})\cdot\cos(\pi h)}{a-2\sqrt{cd}\cos(\pi h)}\ . \tag{3.12}$$

It is possible to show that the bound on the two-plane iteration matrix is smaller than the spectral radius of the iteration matrix associated with natural lexicographic ordering. The Taylor expansion about $h=0$ of the Jacobi iteration matrix associated with the lexicographic ordering is

$$1-\left(\frac{3}{4}\pi^2+\frac{1}{16}\sigma^2+\frac{1}{16}\tau^2+\frac{1}{16}\mu^2\right)h^2+o(h^2)\ ,$$

and the Taylor expansion of the bound on the spectral radius of the two-plane Jacobi iteration matrix is

$$1-\left(\frac{1}{2}\pi^2+\frac{1}{8}\sigma^2+\frac{7}{64}\tau^2+\frac{7}{64}\mu^2\right)h^2+o(h^2)\ .$$

However, even though we obtain an impressive saving in iterations, similarly to the 2D case, here there is fill-in when the main block diagonal of the matrix is factorized. This can be observed by examining the leftmost graph of Figure 3.2. Therefore, in the three-dimensional case we should expect a much smaller gain in overall computational work when the block Jacobi scheme is used in conjunction with the two-plane ordering. Nevertheless, this ordering strategy is more effective than the the natural lexicographic ordering.

**4. Two-step preconditioner.** In [10] a class of two-step preconditioners is considered. Suppose we are given a linear system $Ax=b$. The first step is to form a splitting

$$A=A_1+A_2, \tag{4.1}$$

and decompose $A_1$, either exactly or approximately. In the discussion that follows, we shall focus on the LU or incomplete LU decomposition of $A_1$. We note that other decompositions are viable as well. Consider the preconditioner

$$M^{-1}=U^{-1}(I+L^{-1}A_2U^{-1})^{-1}L^{-1}. \tag{4.2}$$

If $A_1=LU$ then $M^{-1}=A^{-1}$, because in this case $M$ is given by:

$$M=L(I+L^{-1}A_2U^{-1})U=A_1+A_2=A. \tag{4.3}$$

Thus, if $A_1=LU+R$ is an incomplete LU factorization with $\|R\|$ small, $M^{-1}$ can be considered an effective approximation to $A^{-1}$. However, computing $(I+L^{-1}A_2U^{-1})^{-1}$ is nearly as costly as computing $A^{-1}$, since $I+L^{-1}A_2U^{-1}=L^{-1}(LU+A_2)U^{-1}\approx L^{-1}AU^{-1}$. We therefore seek an approximation for this quantity rather than its exact value. This is the second step, hence the name *two-step preconditioner*. The approach suggested in [10] is to use a first order truncated Neumann series:

$$(I+L^{-1}A_2U^{-1})^{-1}\approx I-L^{-1}A_2U^{-1}\ . \tag{4.4}$$

This approach is similar in spirit to the technique presented and investigated in [5], where the skew-symmetric part of the matrix forms the basis for the splitting.

If $M$ is computed as specified above, with the approximation as in (4.4), then the algorithm for computing $Mz = r$ is as follows:

1. Find $L$ and $U$ such that $A_1 \approx LU$.
2. Solve $LUy_1 = r$.
3. Compute $y_2 = A_2 y_1$.
4. Solve $LUy_3 = y_2$.
5. $z = y_1 - y_3$.

The splitting $A = A_1 + A_2$ is useful for cases where dealing with $A_1$ is much easier than handling the matrix $A$. For example, if $A$ is positive real, we can define $A_1$ as the symmetric part of $A$, and compute an incomplete Cholesky decomposition [4, p. 535]. Note that for any choice of $A_1$, $||L^{-1}A_2U^{-1}|| < 1$ must be satisfied for the approximation (4.4) to be valid.

It is possible to pick $A_1$ based on sparsity considerations. Having $A_1$ with sparse factors is especially useful for linear systems arising from discretization of three-dimensional elliptic problems, where the loss of sparsity is significant when factorizing $A$. If $A_1$ gives rise to modest fill-in, its full LU decomposition may be computed, and the quality of the preconditioner would depend solely on the quality of the approximation of $(I + L^{-1}A_2U^{-1})^{-1}$.

Suppose indeed that

$$M^{-1} = U^{-1}(I - L^{-1}A_2U^{-1})L^{-1} \ ; \quad A_1 = LU. \tag{4.5}$$

For any splitting of the matrix $A$, we have the following useful result:

Proposition 4.1. *Let $A = A_1 + A_2$ and let $M^{-1}$ be the matrix defined in (4.5). Then*

$$M^{-1}A = I - (A_1^{-1}A_2)^2 \ . \tag{4.6}$$

*Proof.*

$$M^{-1}A = U^{-1}(I - L^{-1}A_2U^{-1})L^{-1}A = (A_1^{-1} - A_1^{-1}A_2A_1^{-1})(A_1 + A_2) = I - (A_1^{-1}A_2)^2.$$

□

Thus the convergence rate of a scheme of the form

$$Mx^{(k+1)} = Nx^{(k)} + b$$

depends on the spectral radius of

$$M^{-1}N = I - M^{-1}A = \left(A_1^{-1}A_2\right)^2 .$$

Below we provide analytical observations specifically for the two-plane ordering. We define $A_1$ to be the matrix $A$, with its fourth superdiagonal and fourth subdiagonal eliminated. This actually corresponds to setting $c = d = 0$. It turns out that the amount of fill-in for $A_1$ is significantly lower than the fill-in for the whole matrix $A$: see Fig. 4.1. Since only two diagonals of the original matrix have been zeroed out, we can expect good convergence rates.

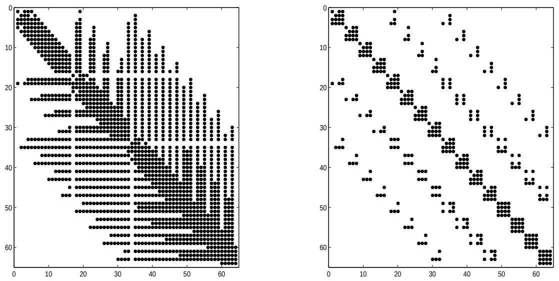


FIG. 4.1. *Fill-in in $A$ vs. fill-in in $A_1$ for a 64-point grid. Shown are the sparsity patterns of $L+U$, where $L$ and $U$ are the matrices associated with the full $LU$ decompositions (without pivoting) of $A$ (left) and $A_1$ (right).*

PROPOSITION 4.2. *The eigenvalues of $A_1$ are given by*

$$\lambda_{j,k}(A_1) = a + 2\sqrt{be}\cos(\pi jh) + 2\sqrt{fg}\cos(\pi kh) \;, \qquad j,k = 1,\dots,n. \tag{4.7}$$

*Proof.* $A_1$ is a permuted version of the lexicographically ordered matrix with the convection term in the $x$-direction set to 0. □

COROLLARY 4.3. *For $be, fg > 0$ the minimal eigenvalue of $A_1$ is*

$$\lambda_{min}(A_1) = a - 2(\sqrt{be} + \sqrt{fg})\cos(\pi h). \tag{4.8}$$

PROPOSITION 4.4. *The eigenvalues of $A_2$ are given by*

$$\lambda_j = 2\sqrt{cd}\cos(\pi jh) \;, \qquad j = 1,\dots,n. \tag{4.9}$$

*Proof.* The matrix $A_2$ is a permuted version of the lexicographically ordered matrix with zero convection in the $y$ and $z$ directions, modified so that its main diagonal is zero. □

The above propositions lead us to the following result.

THEOREM 4.5. *Let*

$$\lambda_{i,j,k} = \frac{2\sqrt{cd}\cos(\pi ih)}{a + 2\sqrt{be}\cos(\pi jh) + 2\sqrt{fg}\cos(\pi kh)} \;, \qquad i,j,k = 1,\dots,n. \tag{4.10}$$

*The eigenvalues of $M^{-1}A$ are given by $1 - \lambda_{i,j,k}^2$, $i,j,k = 1,\dots,n$.*

We can make a few further observations on the eigenvalues of the preconditioned matrix. For example:

COROLLARY 4.6. *Suppose centered difference discretization is used for the convective terms. For the region of numerical stability, namely $|\beta|, |\gamma|, |\delta| < 1$, the eigenvalues of $M^{-1}A$ are all real, positive and smaller than 1.*

*Proof.* For the centered difference scheme we have $cd = 1 - \beta^2$, hence if $|\beta| < 1$ we have $0 < cd < 1$. Similarly, $0 < be, fg < 1$. Since $a = 6$, it readily follows by (4.10) that $0 < 1 - \lambda_{i,j,k}^2 < 1$ for all $i,j,k$. □

Suppose we wish to use the splitting $A = M - N$ for fixed point iterations. We can estimate the spectral radius of the iteration matrix $I - M^{-1}A$.

PROPOSITION 4.7. *Let $A_1 = LU$, and let $be, cd, fg > 0$. Denote*

$$\rho_1 = \frac{2\sqrt{cd}\cos(\pi h)}{a - 2(\sqrt{be} + \sqrt{fg})\cos(\pi h)} \ . \tag{4.11}$$

*Then the spectral radius of $I - M^{-1}A$ is given by $\rho_1^2$.*

*Proof.* $A_1$ and $A_2$ commute. By Proposition 4.1 we have $I - M^{-1}A = (A_1^{-1}A_2)^2$, thus the spectral radius is given by $\left(\frac{\rho(A_2)}{\lambda_{min}(A_1)}\right)^2$. □

From Proposition 4.7 (or Cor. 4.6) we can deduce the following on the convergence of the scheme.

COROLLARY 4.8. *Given $|\beta|, |\gamma|, |\delta| < 1$, the iteration*

$$Mx^{(k+1)} = (M - A)x^{(k)} + b \tag{4.12}$$

*with $M$ as defined in (4.5), converges for any initial guess.*

It is clear that if the matrices $L$ and $U$ defining $M$ satisfy $A_1 = LU$, improving the rate of convergence lies solely on obtaining a better approximation to the matrix $(I + L^{-1}A_2U^{-1})^{-1}$. An alternative to approximating it by the leading two terms of the Neumann series is to perform a small number of, say, Jacobi iterations of the form

$$x^{(k)} = -L^{-1}A_2U^{-1}x^{(k-1)} + p, \quad k = 1, 2, \ldots \tag{4.13}$$

We have (see [9, p. 53] for justification)

$$\rho(L^{-1}A_2U^{-1}) = \rho(A_1^{-1}A_2) \ , \tag{4.14}$$

and the spectral radius of the iteration matrix is known by the previous propositions. We can thus estimate the number of iterations required for obtaining a solution to a certain level of accuracy. Furthermore, by induction it is straightforward to show:

PROPOSITION 4.9. *Suppose $m$ steps of (4.13) are performed, and suppose that $A_1 = LU$. Then if the initial guess is zero, the corresponding preconditioned matrix is*

$$M^{-1}A = I - \left(A_1^{-1}A_2\right)^m . \tag{4.15}$$

**4.1. Evaluation of the bounds for small mesh size.** Taylor expansions of the quantities discussed in the previous sections can help gain insight into asymptotic rates of convergence.

PROPOSITION 4.10. *For $h$ sufficiently small the spectral radius of $I - M^{-1}A$, where $M$ is the matrix defined in (4.5), can be written as:*

$$\rho(I - M^{-1}A) = 1 - (3\pi^2 + \frac{\sigma^2}{4} + \frac{\tau^2}{4} + \frac{\mu^2}{4})h^2 + o(h^2). \tag{4.16}$$

*Proof.* Expanding $\sqrt{cd} = \sqrt{1 - \beta^2} = 1 - \frac{\sigma^2 h^2}{8} + o(h^2)$, and similarly for $\sqrt{be}$ and $\sqrt{fg}$, using $\cos(\pi h) = 1 - \frac{\pi^2 h^2}{2} + o(h^2)$, and using $\frac{1}{1-x^2} = 1 + x^2 + o(x^2)$ for $x$ sufficiently small, eq. (4.16) readily follows. This result has been verified using Maple V. □

From the point of view of iteration counts, the two-step preconditioning solve, if used as a fixed point iteration scheme, is faster than the Jacobi scheme associated with the two-plane matrix by a factor of 2, and is faster than the Jacobi scheme associated with the lexicographic ordering by a factor of 4; but the computational work involved in each iteration of the two-step preconditioner is larger: it is equivalent to solving *two* linear systems, and requires that the fill-in in the factorization of $A_1$ be minimal.

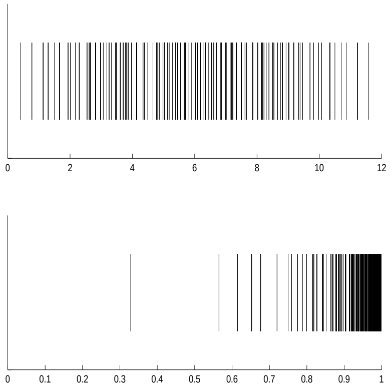


FIG. 4.2. *Eigenvalues of $A$ and of $M^{-1}A$ (upper and lower plots respectively). $M$ is the two-step preconditioner. Note that the scales are different. In this example we have a 512-point grid and the problem is described in section 5 (Test problem 2, with $\varepsilon = 1/10$).*

In Figure 4.2 we observe the structure of the eigenvalues of the preconditioned matrix. The eigenvalues of the matrix associated with the two-step preconditioner are all positive and smaller than 1 (for this example Cor. 4.6 applies), and are clustered near 1.

## 5. Numerical experiments.

The experiments were performed on an SGI machine, using MATLAB.

### 5.1. Test Problem 1.

Consider Problem (1.1). For the 2D problem we tested with $n = 32$, for the 3D problem we tested with $n = 16$. The tables validate our analysis, both in terms of iteration counts and solve time.

In Table 5.1 we present iteration counts and solution times for the block Jacobi scheme in the two-dimensional case. As is predicted by our bounds and estimate of computational costs, the iteration count is faster by a factor of approximately 2, and so is the solution time. The analysis predicts an improvement by a factor of 2 for iterations, and at least 80% for solution time. The same rates of improvement occur for the Gauss-Seidel and the SOR schemes — Tables 5.2 and 5.3. (In these tables only iteration counts are given.)

For the SOR scheme applied to the matrix associated with the two-line ordering, the relaxation parameter was determined by using the bound for the block Jacobi scheme, and substituting it in Young's formula [15]. (The two-line matrix is consistently ordered relative to $2n \times 2n$ blocks, and hence the Young analysis can be applied.) Even though the chosen relaxation parameter is not optimal, since a bound rather than the actual spectral radius of the Jacobi iteration matrix is used, numerical experiments verify that it is very close in value to the optimal relaxation parameter. For example, for $\beta = \gamma = 0.1$, we have obtained 1.62 as the approximate optimal relaxation and the scheme converged within 42 iterations (see Table 5.3) ; the optimal

| $\beta,\ \gamma$ | Two-Line | | Lexicographic | |
|---|---|---|---|---|
| —— | Iterations | time | Iterations | time |
| 0.1 | 526 | 14.0 | 1042 | 28.6 |
| 0.3 | 141 | 3.8 | 269 | 7.4 |
| 0.5 | 74 | 2.0 | 135 | 3.7 |
| 0.7 | 49 | 1.3 | 84 | 2.3 |
| 0.9 | 34 | 0.9 | 56 | 1.5 |

TABLE 5.1

*2D Block Jacobi: comparison between the two ordering strategies for different values of the mesh Reynolds numbers.*

| $\beta,\ \gamma$ | Two-line | Lexicographic |
|---|---|---|
| 0.1 | 260 | 516 |
| 0.3 | 65 | 124 |
| 0.5 | 31 | 55 |
| 0.7 | 18 | 29 |
| 0.9 | 10 | 14 |

TABLE 5.2

*2D Block Gauss-Seidel: comparison of iteration counts for different values of the mesh Reynolds numbers.*

relaxation parameter in this case is 1.58 and the scheme converges within 40 iteration when the value is used, i.e. an insignificant difference.

In Table 5.4 we present results for the block Jacobi scheme in the 3D case. Here as well, the analysis predicts the behavior. In the 3D case the factor of 2 in iteration count does not carry over to a similar saving in running time, due to the fill-in in the main diagonal block of the matrix. Nevertheless, the two-plane ordering is superior to the natural lexicographic ordering when the block Jacobi scheme is used, and an optimized block solver could possibly increase the savings.

**5.2. Test Problem 2.** Consider the following PDE with variable coefficients:

$$-\varepsilon \Delta u + (x, y, z) \cdot \nabla u = w(x, y, z) \tag{5.1}$$

on $\Omega = (0,1)\times(0,1)\times(0,1)$ , with zero Dirichlet boundary conditions, where $w(x,y,z)$ is constructed so that the solution is

$$u(x,y,z) = x\ y\ z\ (1-x)\ (1-y)\ (1-z)\ \exp(x+y+z)\ . \tag{5.2}$$

We experimented with the standard centered second order accurate finite difference scheme on a uniform grid, with $\varepsilon = \frac{1}{10}$. Bi-CGSTAB [14] was used, and the stopping criterion was: relative residual smaller than $10^{-10}$. The grid size was $32 \times 32 \times 32$ (32,768 gridpoints). Note that for this grid the mesh Reynolds numbers do not exceed the value 1, hence the centered difference scheme is adequate.

Our results are presented in Table 5.2. The solver preconditioned with ILU and drop tolerance 0.04 is the overall winner. (We note that larger values of the drop tolerance did not reduce the amount of computational work.) However, while the variation in performance for the ILU preconditioned solvers is high (this is true for both the construction time and iteration counts), the two-step preconditioner demonstrates robustness in that regardless of the numerical drop tolerance that is used, the overall

| $\beta,\ \gamma$ | Two-Line | | Lexicographic | |
|---|---|---|---|---|
| —— | $\omega$ | Iterations | $\omega$ | Iterations |
| 0.1 | 1.62 | 42 | 1.68 | 59 |
| 0.3 | 1.40 | 21 | 1.49 | 31 |
| 0.5 | 1.12 | 16 | 1.21 | 21 |
| 0.7 | 1.04 | 13 | 1.09 | 17 |
| 0.9 | 1.01 | 9 | 1.01 | 11 |

TABLE 5.3

*2D Block SOR: comparison between iteration counts and relaxation parameters for different values of the mesh Reynolds numbers.*

| $\beta,\ \gamma,\ \delta$ | Two-Plane | | Lexicographic | |
|---|---|---|---|---|
| —— | Iterations | time | Iterations | time |
| 0.1 | 304 | 60.9 | 600 | 66.4 |
| 0.3 | 129 | 25.9 | 248 | 27.1 |
| 0.5 | 71 | 14.3 | 130 | 14.4 |
| 0.7 | 47 | 9.5 | 81 | 9.0 |
| 0.9 | 31 | 6.2 | 53 | 5.8 |

TABLE 5.4

*3D Block Jacobi: comparison between the two ordering strategies for different values of the mesh Reynolds numbers.*

computation time does not significantly vary. The two-step preconditioner is parameter free; for ILU it may be difficult to find the optimal value without performing repeated experiments.

| Method | ILU p/c | two-step p/c |
|---|---|---|
| ILU($4 \cdot 10^{-2}$) | 36.60 | 38.75 |
| ILU($10^{-2}$) | 42.04 | 38.40 |
| ILU($7 \cdot 10^{-3}$) | 45.59 | 38.52 |
| ILU($4 \cdot 10^{-3}$) | 52.61 | 40.43 |
| ILU($10^{-3}$) | 85.30 | 46.55 |

## 6. Conclusions.

Two ordering strategies have been analyzed: the two-line ordering for 2D problems, and the two-plane ordering for 3D problems. Bounds on convergence rates of the block Jacobi scheme have been derived and have been shown to be effective and reliable in predicting the convergence behavior of this scheme. Analysis and numerical experiments demonstrate improvement in comparison to the natural lexicographic ordering.

A two-step preconditioning technique based on ideas in [5, 10] has been presented. For a certain choice of the splitting with the ordering strategy we have considered, we obtain a rapid construction of the preconditioner. While convergence is slower compared to the optimal ILU preconditioned schemes that are based on drop tolerance, the two-step preconditioning technique is not parameter dependent, which is an important advantage. Preserving the sparsity throughout the computation may make it possible to gain savings in a parallel environment.

## REFERENCES

[1] S. Barnett. *Matrices - Methods and Applications.* Clarendon Press, Oxford, 1990.

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

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

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Third Edition, Johns Hopkins University Press, Baltimore, MD, 1996.

[5] G. H. Golub and D. Vanderstraeten. On the preconditioning of matrices with skew-symmetric splittings. *Numerical Algorithms*, 25:223–239, 2000.

[6] C. Greif. Cyclic reduction for three-dimensional elliptic equations with variable coefficients. *SIAM J. Matrix Anal. Appl.*, 21(1):29–44, 1999.

[7] C. Greif and J. M. Varah. Iterative solution of cyclically reduced systems arising from discretization of the three-dimensional convection-diffusion equation. *SIAM J. Sci. Comput.*, 19(6):1018–1040, 1998.

[8] C. Greif and J. M. Varah. Block stationary methods for nonsymmetric cyclically reduced systems arising from three-dimensional elliptic equations. *SIAM J. Matrix Anal. Appl.*, 20(4):1038–1059, 1999.

[9] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[10] H. Melbø. *Preconditioning and Error Estimates for Iterative Linear Solvers.* Ph.D. thesis, Studieretning for Industriell matematikk, NTNU, 1998.

[11] S. V. Parter. On 'two-line' iterative methods for the Laplace and biharmonic difference equations. *Numer. Math.*, 11:240–252, 1959.

[12] Y. Saad. *Iterative Methods for Sparse Linear Systems.* PWS Publishing Company, Boston, 1996.

[13] R. S. Varga. *Matrix Iterative Analysis.* Prentice-Hall, 1962.

[14] H. A. Van Der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 13:631–644, 1992.

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