BLOCK ITERATIVE METHODS FOR ELLIPTIC AND PARABOLIC DIFFERENCE EQUATIONS

by

Seymour V. Parter
Michael Steuerwalt

Computer Sciences Technical Report #447

September 1981

Block iterative methods for elliptic and parabolic difference equations[1]*

Seymour V. Parter[2] and Michael Steuerwalt[3]

Abstract. Direct iterative methods for solving the linear system AX = Y split A into a difference M-N. By viewing N as a weak multiplication operator, we determine the convergence rates of block direct iterative methods for elliptic and parabolic difference equations. The difference equations may arise from very general partial differential equations on general domains in m space dimensions.

(1) This work was supported by the U.S. Department of Energy under Contract W-7405-Eng-36, and by the Office of Naval Research under Contract N00014-76-C-0341, ID number NR 044-356.

(2) Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

(3) University of California, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

*Will also appear as Los Alamos Scientic Laboratory report LA-UR-81-2576

1. Introduction. The development of computers thirty years ago made it practical to calculate finite difference approximations of elliptic partial differential equations. For these calculations the solution of a linear system $AU = \tilde{F}$, which is the finite difference representation of the differential equation, is fundamental. Hardware characteristics of early computers, particularly memory limitations, spurred the development of direct iterative methods for these linear systems. In direct iterative schemes the matrix A splits into a difference $A = M-N$, and one generates a sequence $\{U^{(\nu)}\}$ according to $MU^{(\nu)} = NU^{(\nu-1)}+\tilde{F}$. Convergence of the method is governed by the spectral radius ρ of $M^{-1}N$: $\{U^{(\nu)}\}$ converges to the solution if $\rho < 1$, and smaller ρ implies faster convergence.

The first iterative methods were point methods -- in any step of the iteration they solved for one component of the unknown solution vector at a time. Intuition suggests that iterative algorithms that solve for several points at once will converge more rapidly than point algorithms. The Gaussian elimination algorithm is seen in this light to converge in one step. Frankel [14], Young [34], Arms, Gates, and Zondek [1], and Varga [32], using the algebraic structure of the linear systems, and Parter [22], [23], by exploiting the nature of the systems as finite difference approximations to elliptic partial differential equations, determined the convergence rates of point and block iterative methods. The results confirmed that iterative methods on blocks comprising several lines of unknowns indeed converged faster than point methods. Much of the work up to 1961 is collected in [33].

The usual finite difference approximations are accurate to second order in the spatial mesh size h. In the middle 1960s attention turned to higher order approximation methods -- finite element and other

projection methods, which are still the subject of intensive study ([36], [7], [2], [30], [5], [10]). Because of their treatment of boundary conditions, these methods are formally easier to obtain than higher order finite difference approximations, and for a given accuracy their corresponding linear system of equations is smaller than the finite difference system. Hence interest in direct factorization methods for linear systems grew, and continues today; see [27], [28], [15], and [16].

At about the same time it was seen that their regular structure made separable finite difference elliptic systems amenable to special fast direct factorization methods ([18], [9], [12], [31]). For a limited class of nice elliptic problems, then, it became practical to compensate for the second order accuracy of the usual finite difference approximation by taking a sufficiently small h and exploiting the regular structure of the linear system.

But not every problem is nice. Moreover, within the past few years a growing desire to solve three-dimensional problems, together with the development of novel computer architectures -- array processors, vector machines, and multiprocessors -- has rekindled interest in block iterative methods for elliptic systems. The effects of special architectures are considered in [29], [17], and [19], while an analysis of the convergence rates of iterative methods for fairly general elliptic problems already appears in [23].

But not every analysis is nice, and that of [23], partly because of its generality, is somewhat opaque. A relatively direct discussion of the basic ideas is given in [3] for the Poisson problem in a square. That presentation uses the strong estimates of Nitsche and Nitsche [21] and of Brandt [8].

Our purpose here is to reexamine the convergence rates of iterative block methods for elliptic difference equations. A feature of the present analysis is that we avoid the estimates of [21] and [8]. For the Poisson problem in two or three dimensions this is of little moment. But the Nitsche estimates have never been extended to general regions, and must fail in dimensions greater than three. In contrast, we will show that our new approach is easily extended to general domains, to any number of dimensions, and to general elliptic difference equations.

In addition, we can deal with certain kinds of singularly perturbed elliptic difference equations. Such equations can arise when solving parabolic problems by discrete time methods. For instance, let $\Delta := \Sigma_{i=1}^{2} \partial^2/\partial x_i^2$ be the two-dimensional Laplacian; the backward Euler method for the parabolic operator $(c_0\partial/\partial t)-\Delta$ leads, at each time slice t_n, to an elliptic operator

$$c/\tau - \Delta, \quad \tau := t_n - t_{n-1}. \tag{1.1}$$

Let Δ_h be a finite difference approximation to Δ on a spatial mesh of size h; we get a matrix A representing the elliptic difference operator $ch^2/\tau - h^2\Delta_h$. If $ch^2/\tau = \bar{c}h^{\alpha}$, then A corresponds to

$$\bar{c}h^{\alpha} - h^2\Delta_h. \tag{1.2}$$

We distinguish four cases. Analysis of the first, in which $\alpha < 0$, is easy: $\rho = O(h^{-\alpha})$, and iterative methods converge very rapidly. In the second, $\alpha = 0$, and (1.2) is a singularly perturbed operator. We have studied this operator in [26], where it arose from plane iterative methods for the Poisson problem in the unit cube; the attack there,

though related to some of the ideas of this report, seems to be particular to the model operator (1.1) and rectangular domains.

In this paper we restrict our attention to the third and fourth cases, wherein $0 < \alpha \leq 2$. If $\alpha = 2$, then (1.2) is a regular elliptic difference operator, to which both the earlier and our new analyses apply. When $0 < \alpha < 2$, (1.2) is again a singularly perturbed operator; but it too can be handled with our present methods, unlike the instance $\alpha = 0$. To justify considering this case, we point out that $\alpha = 1$ for the optimal choice of τ in the Crank-Nicolson method for parabolic problems.

We begin in section 2 with a description of the model elliptic and parabolic problems in the two-dimensional unit square. It is worth remarking that our model problems need not be self-adjoint. Section 3 is devoted to proving the convergence rates of iterative schemes satisfying certain basic assumptions.

In section 4 we describe block structures of particular interest -- k-line and k×k blocks -- and the usual iterative schemes: Jacobi, Gauss-Seidel, and successive overrelaxation. In these schemes A splits into a difference A = M-N. The key to our analysis is that it suffices to consider only the block Jacobi scheme, for which N is essentially a sum of one-dimensional weak multiplication operators $\tilde{N}$. We demonstrate this decomposition of N in section 4, and discuss the action of $\tilde{N}$ in section 5. In section 6 we use the theory of section 3 and properties of $\tilde{N}$ to derive the convergence rates of the block iterative methods of section 4.

Next we take up more general problems: other operators in section 7, and other domains in section 8. We conclude in section 9 with some comments about the general applicability of our method of analysis.

2. The model problems. The basic ideas are clearest in this simple setting. We construct finite difference approximations of the partial differential operators

$$Lu := -\Delta u + du_x + eu_y,$$

$$\ell u := cu_t + Lu \tag{2.1}$$

on the open unit square

$$\Omega := \{(x,y) \in \underset{\sim}{R}^2 : 0 < x, y < 1\}$$

in the usual way. Impose on $\bar{\Omega}$ a mesh with uniform spacing

$$h := 1/(P + 1) \tag{2.2}$$

and let $(x_i,y_j) := (ih,jh)$. Define the set of interior mesh points Ω_h and the discrete boundary $\partial\Omega_h$ by

$$\Omega_h := \{(x_i,y_j) : 1 \leqq i, j \leqq P\},$$

$$\partial\Omega_h := \{(x_i,y_j) : i = 0 \text{ or } = P+1, \text{ or } j = 0 \text{ or } = P+1\}. \tag{2.3}$$

A mesh vector $U = \{U_{i,j} : 0 \leqq i, j \leqq P+1\}$ is a function defined on the entire discrete mesh $\bar{\Omega}_h := \Omega_h \cup \partial\Omega_h$.

The discrete Laplace operator is defined at points in Ω_h by

$$[\Delta_h U]_{i,j} := (U_{i-1,j} - 2U_{i,j} + U_{i+1,j})/h^2 + (U_{i,j-1} - 2U_{i,j} + U_{i,j+1})/h^2. \tag{2.4}$$

We suppose that c, d, and e are smooth functions on $\bar{\Omega}$ and that

$$c(x,y) \geqq c_0 > 0 \text{ on } \bar{\Omega}. \tag{2.5}$$

The discrete operators that arise in approximating (2.1) are then

$$
\begin{aligned}
[L_h U]_{i,j} := [-\Delta_h U]_{i,j} &+ d_{i,j}(U_{i+1,j} - U_{i-1,j})/(2h) \\
&+ e_{i,j}(U_{i,j+1} - U_{i,j-1})/(2h)
\end{aligned} \tag{2.6}
$$

and

$$
[\ell_h U]_{i,j} := (c_{i,j}/\tau)U_{i,j} + [L_h U]_{i,j}, \tag{2.7}
$$

where $\tau > 0$ is given and, for instance, $c_{i,j} := c(x_i, y_j)$.

Note that, although the mesh vector U is defined on $\bar{\Omega}_h$, the vectors $\Delta_h U$, $L_h U$, and $\ell_h U$ are defined only at the interior mesh points. As usual, the forward difference operators are given by

$$
\begin{aligned}
\nabla_x U_{i,j} &:= (U_{i+1,j} - U_{i,j})/h \quad (0 \leqq i \leqq P,\ 1 \leqq j \leqq P), \\
\nabla_y U_{i,j} &:= (U_{i,j+1} - U_{i,j})/h \quad (1 \leqq i \leqq P,\ 0 \leqq j \leqq P).
\end{aligned} \tag{2.8}
$$

Given mesh vectors F and G, the model elliptic problem is to find a mesh vector U satisfying

$$
L_h U = F \text{ in } \Omega_h, \quad U = G \text{ on } \partial\Omega_h \tag{2.9}
$$

and the model parabolic problem requires U to solve

$$
\ell_h U = F \text{ in } \Omega_h, \quad U = G \text{ on } \partial\Omega_h. \tag{2.10}
$$

After choosing an ordering of the mesh points (x_i, y_j) -- or, equivalently, of the components of U -- we let A be the matrix representing $h^2 L_h$ or $h^2 \ell_h$. As indicated in (2.4), Δ_h, L_h, and ℓ_h map vectors with P^2+4P components into vectors with P^2 components. Hence A is a matrix of order P^2; the known boundary values G are put on the

right hand sides of the difference equations (2.9) and (2.10). In either case we arrive at a linear system

$$AU = \tilde{F} \tag{2.11}$$

of order P^2, where $\tilde{F}$ indicates the result of ordering the components of h^2F and of including the G terms.

Every vector U with P^2 components may be viewed as a mesh vector on $\bar{\Omega}_h$ that also satisfies

$$U = 0 \text{ on } \partial\Omega_h. \tag{2.12}$$

Henceforth we assume every mesh vector U satisfies (2.12).

An iterative method for solving (2.11) is determined by a splitting

$$A = M - N. \tag{2.13}$$

Rewrite (2.11) as

$$MU = NU + \tilde{F}.$$

After choosing a first guess $U^{(0)}$, we obtain a sequence $\{U^{(\nu)}\}$ from

$$MU^{(\nu)} = NU^{(\nu-1)} + \tilde{F}. \tag{2.14}$$

It is well known that when A is nonsingular the iterates $\{U^{(\nu)}\}$ converge to the unique solution of (2.11) independently of $U^{(0)}$ if and only if the spectral radius

$$\rho := \max \{|\lambda| : \det(\lambda M - N) = 0\}$$

of $M^{-1}N$ satisfies $\rho < 1$. So the first thing we require of a splitting

is that $\rho < 1$. Evidently the iterates $\{U^{(\nu)}\}$ of (2.14) converge more rapidly for smaller ρ. Hence our task is to determine the asymptotic behavior of ρ as $h \to 0$.

For future reference we note that corresponding to every λ for which $\det(\lambda M-N) = 0$ there is a vector $V \neq 0$ satisfying $\lambda MV = NV$. We also record two lemmas regarding ∇_x, ∇_y, and Δ_h. Let X and Y be mesh vectors; define an inner product and associated norm

$$(X,Y) := \Sigma_{i,j} X_{i,j}\bar{Y}_{i,j}, \quad \|X\|_h := (X,X)^{1/2}.$$

An operator B on mesh vectors is normed in the customary way by

$$\|B\|_h := \sup \{\|BX\|_h : \|X\|_h = 1\}.$$

As usual, $|d|_\infty$ denotes the sup norm of d over $\bar{\Omega}$.

Lemma 2.1. If U is a mesh vector satisfying (2.12), then

$$(\nabla_x U,\nabla_x U) + (\nabla_y U,\nabla_y U) = (-\Delta_h U,U).$$

Proof. Summation by parts; see [11] or [20]. □

Lemma 2.2. If U is a mesh vector satisfying (2.12), then

$$(|\nabla_x U|,|U|) + (|\nabla_y U|,|U|) \leq \|U\|_h [2(-\Delta_h U,U)]^{1/2}.$$

Proof. By the Schwarz inequality,

$$(|\nabla_x U|,|U|) + (|\nabla_y U|,|U|) \leq \|U\|_h [\|\nabla_x U\|_h + \|\nabla_y U\|_h].$$

But the inequality $2ab \leq a^2+b^2$ and Lemma 2.1 show that

$$[\|\nabla_x U\|_h + \|\nabla_y U\|_h]^2 \leq 2[\|\nabla_x U\|_h^2 + \|\nabla_y U\|_h^2] = 2(-\Delta_h U,U). \quad \square$$

3. A general approach. To begin the analysis of the splitting (2.13), we make four assumptions.

A1. $\rho < 1$, so the iterative method (2.14) is convergent.

A2. ρ is an eigenvalue of $M^{-1}N$: there is a mesh vector $U \neq 0$ such that $\rho MU = NU$.

A3. There is a positive constant N_0, independent of h, such that $\|N\|_h \leqq N_0$.

A4. There are a smooth function q and constant q_0 with

$$q(x,y) \geqq q_0 > 0 \text{ on } \bar{\Omega}$$

and a constant $D > 0$, independent of h, so that whenever U and V are mesh vectors satisfying (2.12) we have

$$(NU,V) = (qU,V) + E,$$

where

$$|E| \leqq hD[(|\nabla_x U|+|\nabla_y U|,|V|) + (|U|,|\nabla_x V|+|\nabla_y V|) + (|U|,|V|)]$$

$$+ h^2 D[(-\Delta_h U,U) + (-\Delta_h V,V)].$$

Assumptions A1 - A3 are in effect more or less common; this will become clear in section 6. Our main new concept is A4. As might be expected, verification of A4 and the determination of q are the important technical steps when applying our analysis to any particular splitting. But we shall see that these steps are not difficult.

When a splitting (2.13) satisfies these assumptions, the asymptotic behavior of ρ as a function of h is readily discovered. We begin with the elliptic case.

Theorem 3.1. Let A correspond to h^2L_h. Suppose the splitting (2.13) satisfies A1 - A4. Let Λ_0 be the smallest eigenvalue of the problem

$$Lv = \lambda qv \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega. \tag{3.1}$$

Then

$$\rho = 1 - \Lambda_0 h^2 + o(h^2). \tag{3.2}$$

Proof. Let U be the eigenvector associated with ρ in A2, so that

$$\rho MU = NU.$$

Subtract ρNU from both sides and use (2.13) to see that

$$AU = ((1 - \rho)/\rho)NU. \tag{3.3}$$

By A1,

$$\mu := (1 - \rho)/(\rho h^2) \tag{3.4}$$

is positive. Because A represents h^2L_h, (3.3) corresponds to

$$L_h U = \mu NU \text{ in } \Omega_h, \quad U = 0 \text{ on } \partial\Omega_h. \tag{3.5}$$

Indeed, whenever $\lambda \neq 0$ satisfies

$$\lambda MX = NX \tag{3.6}$$

for some nonzero X, then

$$\mu = \mu(\lambda) := (1 - \lambda)/(\lambda h^2)$$

is an eigenvalue of (3.5). Conversely, if μ is an eigenvalue of (3.5) and $1+\mu h^2 \neq 0$, then

$$\lambda = \lambda(\mu) := 1/(1 + \mu h^2)$$

is an eigenvalue of (3.6).

For fixed h, let $\bar{\mu}$ be an eigenvalue of (3.5) minimal in magnitude. The basic result of [24] shows that $\bar{\mu} \to \Lambda_0$ as $h \to 0$ -- that is, $\bar{\mu} = \Lambda_0+o(1)$. It follows by positivity of Λ_0 that $\mathrm{Re}(1+\bar{\mu}h^2) > 0$ for small h, whence $\bar{\lambda} := 1/(1+\bar{\mu}h^2)$ is a well defined eigenvalue of (3.6). Hence

$$\rho \geq |\bar{\lambda}| = 1/|1 + \bar{\mu}h^2| = 1 - [\Lambda_0 + o(1)]h^2.$$

But μ given by (3.4) is an eigenvalue of (3.5) by construction, and so $(1-\rho)/(\rho h^2) \geq |\bar{\mu}| = \Lambda_0+o(1)$, by the minimality of $\bar{\mu}$. We deduce that

$$\rho \leq 1/(1 + [\Lambda_0 + o(1)]h^2) = 1 - [\Lambda_0 + o(1)]h^2.$$

Comparison of this and the previous inequality proves (3.2). □

Parabolic equations lead to discrete singular perturbation eigenvalue problems, so in the general nonself-adjoint case we can establish only an inequality analogous to (3.2). We arrive as before at (3.3), where A represents $h^2\ell_h$; hence

$$(3.7) \qquad h^2(c/\tau + L_h)U = ((1 - \rho)/\rho)NU.$$

We make a basic assumption about the ratio of the time step τ to the spatial mesh size.

P1. There are constants $c_1 > 0$ and $0 < \alpha < 2$ such that $h^2/\tau = c_1 h^\alpha$.

Now define

$$\mu := (1 - \rho)/(\rho h^\alpha); \tag{3.8}$$

we deduce from (3.7), (3.8), and P1 that in the parabolic case (3.3) corresponds to

$$\bar{c}U + h^{2-\alpha} L_h U = \mu N U \text{ in } \Omega_h, \quad U = 0 \text{ on } \partial\Omega_h, \tag{3.9}$$

where

$$\bar{c}(x,y) := c_1 c(x,y). \tag{3.10}$$

Theorem 3.2. Let A correspond to $h^2 \ell_h$. Suppose P1 holds and the splitting (2.13) satisfies A1 - A4. Let

$$\Lambda_1 := \min \{\bar{c}(x,y)/q(x,y) : (x,y) \in \bar{\Omega}\}.$$

Then

$$\rho \leqq 1 - \Lambda_1 h^\alpha + o(h^\alpha). \tag{3.11}$$

Proof. Because ρ is positive, (3.11) is equivalent to $(1-\rho)/(\rho h^\alpha) = \mu \geqq \Lambda_1 + o(1)$. Suppose this inequality is false. We may then assume

$$0 \leqq \mu \leqq 2\Lambda_1. \tag{3.12}$$

Let U be the eigenvector of (3.9) associated with μ. Normalize $\|U\|_h$ to be 1. By A4,

$$\mu(NU,U) = \mu(qU,U) + E_1, \tag{3.13}$$

where, using Lemma 2.2,

$$|E_1| \leq 2\mu hD[2(-\Delta_h U,U)]^{1/2} + 2\mu h^2 D(-\Delta_h U,U) + \mu hD.$$

Use (3.12) and the inequality $2ab \leq a^2\theta^{-2}+b^2\theta^2$ to get

$$|E_1| \leq 16\Lambda_1^2 D^2 h^\alpha + B/2 + 4\Lambda_1 Dh^\alpha B + 2\Lambda_1 Dh,$$

where we have defined

$$B := h^{2-\alpha}(-\Delta_h U,U).$$

Lemma 2.1 shows that $B > 0$. It follows from (3.9) and (3.13) that

$$(\bar{c}U,U) + B = \mu(qU,U) + E_1 + E_2,$$

with

$$|E_2| \leq h^{2-\alpha}K[2(-\Delta_h U,U)]^{1/2} \leq 2h^{2-\alpha}K^2\theta^{-2} + B\theta^2$$

and $K := |d|_\infty+|e|_\infty$. Choose θ so small that the coefficients of B in the estimates of E_1 and E_2 sum to less than 1 for small h. Then

$$(\bar{c}U,U) \leq \mu(qU,U) + 2h^{2-\alpha}K^2\theta^{-2} + 16\Lambda_1^2 D^2 h^\alpha + 2\Lambda_1 Dh.$$

The theorem follows at once. □

When the splitting is self-adjoint -- a frequent occurrence -- we can use the variational principle to establish equality in (3.11).

Theorem 3.3. Under the assumptions of Theorem 3.2, suppose also that we have

S1. A and M are Hermitian and positive definite.

Then

(3.14) $\rho = 1 - \Lambda_1 h^\alpha + o(h^\alpha).$

Proof. Fix $\varepsilon > 0$ and choose $v_\varepsilon(x,y) \in C^4(\bar{\Omega})$ to vanish on $\partial\Omega$ and to satisfy

$$(3.15) \qquad \frac{\int_\Omega \bar{c} v_\varepsilon^2 \, dx \, dy}{\int_\Omega q v_\varepsilon^2 \, dx \, dy} \leqq \Lambda_1 + \varepsilon.$$

Now A2 and S1 imply that $\rho = \sup \{(NX,X)/(MX,X) : X \neq 0\}$. Choosing X as the mesh vector V_ε determined by point evaluation of v_ε yields

$$(3.16) \qquad \rho \geqq (NV_\varepsilon,V_\varepsilon)/(MV_\varepsilon,V_\varepsilon) = (NV_\varepsilon,V_\varepsilon)/[(AV_\varepsilon,V_\varepsilon) + (NV_\varepsilon,V_\varepsilon)].$$

Observe that

$$(AV_\varepsilon,V_\varepsilon) = h^\alpha[(\bar{c}V_\varepsilon,V_\varepsilon) + h^{2-\alpha}(L_h V_\varepsilon,V_\varepsilon)].$$

It follows from the smoothness of v_ε that

$$h^2(AV_\varepsilon,V_\varepsilon) = h^\alpha[\int_\Omega \bar{c} v_\varepsilon^2 \, dx \, dy + O(h^{2-\alpha})];$$

moreover, by A4

$$h^2(NV_\varepsilon,V_\varepsilon) = \int_\Omega q v_\varepsilon^2 \, dx \, dy + o(1).$$

Combining these equalities with (3.15) and (3.16) yields

$$\rho \geqq 1 - (\Lambda_1 + \varepsilon)h^\alpha + o(h^\alpha),$$

which together with (3.11) establishes the theorem. □

Note that hypothesis S1 requires $d \equiv e \equiv 0$ for the operators (2.1) of the model problems.

4. Some block iterative methods. We take up now a description of specific block iterative methods corresponding to (2.13). The block structure of an iterative scheme for the linear system

$$AX = Y,$$

where A is an $n \times n$ matrix, is completely determined by a block partition of the n-vectors. Suppose every n-vector X is decomposed into subvectors

$$X = (X_1, X_2, \ldots, X_r)^t$$

and each X_j is itself an n_j-vector. This partition of X induces a block partition $A = [A_{i,j}]$ in which each $A_{i,j}$ is an $n_i \times n_j$ matrix. The corresponding block Jacobi iterative scheme is

$$A_{i,i}X_i^{(\nu)} = -\Sigma_{s \neq i} A_{i,s}X_s^{(\nu-1)} + Y_i. \tag{4.1}$$

In terms of (2.14), M is the block diagonal matrix $M = \mathrm{diag}[A_{i,i}]$. The corresponding Gauss-Seidel scheme is

$$A_{i,i}X_i^{(\nu)} = -\Sigma_{s<i} A_{i,s}X_s^{(\nu)} - \Sigma_{s>i} A_{i,s}X_s^{(\nu-1)} + Y_i, \tag{4.2}$$

while the successive overrelaxation (SOR) method with relaxation parameter ω is

$$\begin{aligned} A_{i,i}X_i^{(\nu)} = &-\omega\Sigma_{s<i} A_{i,s}X_s^{(\nu)} - \omega\Sigma_{s>i} A_{i,s}X_s^{(\nu-1)} \\ &+ \omega Y_i + (1-\omega)A_{i,i}X_i^{(\nu-1)}. \end{aligned} \tag{4.3}$$

We are interested in specific block structures that arise in a natural geometric way. Recall that a mesh vector U is defined on the

rectangular set of mesh points Ω_h. We will decompose U into blocks of components corresponding to lines or subsquares of mesh points.

Formally, let k be a fixed integer factor of P, so that

(4.4) $\quad P = kQ$ for some integer Q.

In the k-line block structure (see [22] or [23] for a detailed description), each block of U comprises the unknowns $U_{i,j}$ associated with the points on k consecutive horizontal (or vertical) grid lines. Index the blocks by s; we have

$$\tilde{U}_s := \{U_{i,k(s-1)+j} : 1 \leq i \leq P,\ 1 \leq j \leq k\}. \tag{4.5}$$

The k×k block structure is described in [3], [25], and [26]. Each block comprises the unknowns associated with a k×k square of mesh points. We distinguish these blocks with a double index (r,s):

$$\tilde{U}_{r,s} := \{U_{k(r-1)+i,k(s-1)+j} : 1 \leq i,\ j \leq k\}. \tag{4.6}$$

To write down the matrices A, M, and N of the Jacobi iterative method for each of these block structures is straightforward but tiresome. We shall give a unified analysis of the Jacobi method for these structures. But for illustrative purposes we first sketch a development of the (horizontal) k-line scheme for the elliptic problem (2.9).

For $1 \leq \sigma \leq P$ define the P×P matrices

$$\begin{aligned} D_\sigma &:= [-1-hd_{i,\sigma}/2,\ 4,\ -1+hd_{i,\sigma}/2] \\ S_\sigma &:= \mathrm{diag}[1+he_{i,\sigma}/2] \qquad (1 \leq i \leq P) \\ T_\sigma &:= \mathrm{diag}[1-he_{i,\sigma}/2]. \end{aligned} \tag{4.7}$$

The notation indicates that D_σ is tridiagonal while S_σ and T_σ are diagonal. For example,

$$[D_\sigma]_{i,j} = \begin{cases} 0 & \text{if } |i-j| > 1 \\ -1-hd_{i,\sigma}/2 & \text{if } j = i-1 \\ 4 & \text{if } j = i \\ -1+hd_{i,\sigma}/2 & \text{if } j = i+1. \end{cases}$$

With this ordering of the mesh points into horizontal lines, A is the $P^2 \times P^2$ block tridiagonal matrix

$$A = [-S_\sigma,\ D_\sigma,\ -T_\sigma] \quad (1 \leqq \sigma \leqq P). \tag{4.8}$$

Now collect the lines of unknowns k at a time. For $1 \leqq s \leqq Q$ let M_s be the $kP \times kP$ block tridiagonal matrix

$$M_s := [-S_{k(s-1)+\sigma},\ D_{k(s-1)+\sigma},\ -T_{k(s-1)+\sigma}] \quad (1 \leqq \sigma \leqq k),$$

and define the $kP \times kP$ block matrices

$$R_s := \begin{bmatrix} 0 & 0 \\ T_{ks} & 0 \end{bmatrix}, \qquad W_s := \begin{bmatrix} 0 & S_{k(s-1)+1} \\ 0 & 0 \end{bmatrix}.$$

Observe that A is then the block tridiagonal matrix

$$A = [-W_s,\ M_s,\ -R_s] \quad (1 \leqq s \leqq Q).$$

In the k-line Jacobi scheme, A splits into the block matrices

$$M := \text{diag}[M_s], \quad N := [W_s,\ 0,\ R_s].$$

We now seek a simple quantitative description of N for both the k-line and the $k \times k$ block partitions when $k \geqq 2$. If B and C are

matrices, we mean by $B = C+O(h)$ that there is some constant K so that

$$|(BX,Y) - (CX,Y)| \leq Kh|(X,Y)| \text{ for every X and Y.}$$

Because S_σ and T_σ are O(h) perturbations of the P×P identity matrix, let us for the moment ignore the small terms. We define a one-dimensional operator $\tilde{N}$ on vectors $\phi := (\phi_1, \phi_2, \ldots, \phi_P)^t$ as follows:

$$[\tilde{N}\phi]_{ks+\sigma} := \begin{cases} \phi_{ks+1} & 1 \leq s \leq Q-1,\ \sigma = 0 \\ \phi_{ks} & 1 \leq s \leq Q-1,\ \sigma = 1, \end{cases} \tag{4.9}$$

$$[\tilde{N}\phi]_j := 0 \quad \text{for any other subscript j.}$$

$\tilde{N}$ is a weak multiplication operator, as we shall see in the next section. Now let N_x be that operator on mesh vectors U that acts on U only in the x-direction, and in that direction acts as $\tilde{N}$. Define N_y in a similar way. For instance, with $1 \leq i \leq P$ we have

$$[N_y U]_{i,ks+\sigma} := \begin{cases} U_{i,ks+1} & 1 \leq s \leq Q-1,\ \sigma = 0 \\ U_{i,ks} & 1 \leq s \leq Q-1,\ \sigma = 1, \end{cases} \tag{4.10}$$

$$[N_y U]_{i,j} := 0 \quad \text{for any other subscript j.}$$

Observe for each block structure that the Jacobi splitting (4.1) yields the same N for both the elliptic and parabolic operators (2.6) and (2.7). This is so because the matrix representing the operator $\ell_h - L_h$ is a diagonal matrix. A straightforward computation proves the next theorem, which summarizes the essential nature of N.

Theorem 4.1. Let $k \geq 2$. In the k-line Jacobi scheme (4.1)/(4.5),

$$N = N_y + O(h), \tag{4.11}$$

and for the k×k block Jacobi scheme (4.1)/(4.6) N is given by

$$N = N_x + N_y + O(h). \ \square \tag{4.12}$$

5. The operator $\tilde{N}$. We now show that $\tilde{N}U$ converges weakly to $(2/k)U$, so that $\tilde{N}$ is a weak multiplication operator. In this section U and V are real vectors with P components. For each such vector X it is useful to define $X_0 := 0$. It is clear from (4.9) that $\tilde{N}$ samples U twice in each block of k points $\{U_{ks+\sigma} : 0 \leq \sigma \leq k-1\}$, where $0 \leq s \leq Q-1$ -- except in the first and last blocks. Roughly, but perhaps vividly, $\tilde{N}$ sees U about 2/k of the time; precisely, from (4.9) we have

$$(\tilde{N}U,V) = \Sigma_{s=0}^{Q-1} (U_{ks}V_{ks+1} + U_{ks+1}V_{ks}). \tag{5.1}$$

If U and V arise from the evaluation of smooth functions u(x) and v(x) on the points $\{x_i := ih : 0 \leq i \leq P+1\}$, then

$$U_{ks+j} \cong U_{ks} \quad \text{and} \quad V_{ks+j} \cong V_{ks+1} \quad (0 \leq j \leq k-1),$$

whence

$$U_{ks}V_{ks+1} \cong U_{ks+j}V_{ks+j} \quad (0 \leq j \leq k-1).$$

Summing this approximate equality over j and dividing by k gives

$$U_{ks}V_{ks+1} \cong (1/hk)\Sigma_{j=0}^{k-1} U_{ks+j}V_{ks+j}h,$$

which looks like a Riemann sum over the interval $[x_{ks}, x_{ks+k}]$ for $(1/hk)\int u(x)v(x)\,dx$. Consequently,

$$(\tilde{N}U,V) \cong (2/hk) \int_{[0,1]} u(x)v(x)\,dx \cong (2/k)(U,V).$$

Now we make this argument precise. Let ∇ be the forward difference operator, as in (2.8). Fix j for the moment. Obviously

$$U_{ks} = U_{ks+j} - h\Sigma_{\sigma=0}^{j-1} \nabla U_{ks+\sigma}, \quad V_{ks+1} = V_{ks+j} - h\Sigma_{\sigma=1}^{j-1} \nabla V_{ks+\sigma}$$

(as usual, a vacuous sum has value 0). Hence

$$\begin{aligned} U_{ks}V_{ks+1} &= U_{ks+j}V_{ks+j} + h^2(\Sigma_{\sigma=0}^{j-1} \nabla U_{ks+\sigma})(\Sigma_{\sigma=1}^{j-1} \nabla V_{ks+\sigma}) \\ &\quad - h\Sigma_{\sigma=1}^{j-1} U_{ks+j}\nabla V_{ks+\sigma} - h\Sigma_{\sigma=0}^{j-1} V_{ks+j}\nabla U_{ks+\sigma}. \end{aligned} \tag{5.2}$$

Replace U_{ks+j} and V_{ks+j} in the last two terms of (5.2), using the identity

$$X_{ks+\sigma} = X_{ks+j} - h\Sigma_{n=\sigma}^{j-1} \nabla X_{ks+n}.$$

This substitution gives

$$\begin{aligned} U_{ks}V_{ks+1} &= U_{ks+j}V_{ks+j} + h^2 G_{0,j}(U)G_{1,j}(V) \\ &\quad - h\Sigma_{\sigma=1}^{j-1} U_{ks+\sigma}\nabla V_{ks+\sigma} - h\Sigma_{\sigma=0}^{j-1} V_{ks+\sigma}\nabla U_{ks+\sigma} \\ &\quad - h^2\Sigma_{\sigma=1}^{j-1} G_{\sigma,j}(U)\nabla V_{ks+\sigma} - h^2\Sigma_{\sigma=0}^{j-1} G_{\sigma,j}(V)\nabla U_{ks+\sigma}, \end{aligned} \tag{5.3}$$

where for $0 \leq \sigma \leq j-1 \leq k-1$ we define

$$G_{\sigma,j}(X) := \Sigma_{n=\sigma}^{j-1} \nabla X_{ks+n}, \quad G(X,s) := \Sigma_{n=0}^{k-1} |\nabla X_{ks+n}| \geq |G_{\sigma,j}(X)|.$$

Sum (5.3) over $0 \leq j \leq k-1$ and divide by k to get

$$U_{ks}V_{ks+1} = (1/k)\Sigma_{j=0}^{k-1} U_{ks+j}V_{ks+j} + (1/k)E_s, \tag{5.4}$$

with

$$|E_s| \leq hk[\Sigma_{j=0}^{k-1} |U_{ks+j}||\nabla V_{ks+j}| + \Sigma_{j=0}^{k-1} |V_{ks+j}||\nabla U_{ks+j}|]$$

(5.5)

$$+ 3h^2 kG(U,s)G(V,s).$$

By the Schwarz inequality and the inequality $2ab \leq a^2+b^2$,

$$G(U,s)G(V,s) \leq (k/2)[(\Sigma_{j=0}^{k-1} |\nabla U_{ks+j}|^2) + (\Sigma_{j=0}^{k-1} |\nabla V_{ks+j}|^2)].$$

Estimate the last term of (5.5) in this way, and sum (5.4) over s to deduce that

$$\Sigma_{s=0}^{Q-1} U_{ks} V_{ks+1} = (1/k)(U,V) + \tilde{E}/2, \tag{5.6}$$

where

$$|\tilde{E}| \leq 2h[(|U|,|\nabla V|) + (|\nabla U|,|V|)]$$

(5.7)

$$+ 3h^2 k[(\nabla U,\nabla U) + (\nabla V,\nabla V)].$$

Comparison of (5.1) to (5.6) shows that we can exploit the symmetry of this argument in U and V to prove the following theorem, which quantitatively describes $\tilde{N}$.

Theorem 5.1. Let $\tilde{N}$ be given by (4.9). For P-vectors U and V,

$$(\tilde{N}U,V) = (2/k)(U,V) + \tilde{E}, \tag{5.8}$$

and $\tilde{E}$ is estimated by (5.7). □

6. Rates of convergence. In this section we take up the problem of determining the convergence rates of the iterative methods (4.1) - (4.3) when applied to the elliptic and parabolic model problems (2.9)

and (2.10) with the k-line and k×k block structures described in section 4. We limit our discussion to the case where $k \geq 2$; although a similar argument applies when $k = 1$ (and formulas (6.4) - (6.9) are valid for $k = 1$), we have not in that instance described N. We begin by showing that the Jacobi method for these block structures satisfies the assumptions of section 3. After ρ is determined for the Jacobi method it is easy to find the convergence rates of the Gauss-Seidel and SOR methods.

Lemma 6.1. Assumption A1 holds for both block structures and both problems if h is sufficiently small.

Proof. In all cases, inspection of the submatrices (4.7) of A, as given by (4.8), shows that the diagonal elements of A are positive and the other elements are, for small h, nonpositive. Therefore N is nonnegative and A and M are M-matrices: that is,

$$N \geq 0, \quad M^{-1} \geq 0, \quad \text{and } A^{-1} \geq 0. \tag{6.1}$$

Moreover, A is irreducible for small h. A1 follows from Theorem 3.13 of [33]. □

Lemma 6.2. Assumption A2 holds for both block structures and both problems if h is sufficiently small.

Proof. This follows from (6.1) and the Perron-Frobenius theory; see Theorem 2.1 in [33]. □

We remark that when nonnegativity of M^{-1} or A^{-1} fails, A1 and A2 often can be established by other means. For example, A1 holds when A is positive definite and N is nonnegative. A2 follows from supposing that M is positive definite, N is symmetric, and the splitting satisfies block property A (see [1], [33], [35], [25]), for then eigenvalues of $M^{-1}N$ are real and occur in signed pairs.

Lemma 6.3. Assumption A3 holds for both block structures and both problems if h is sufficiently small.

Proof. In light of Theorems 4.1 and 5.1, $N_0 \leqq 2+O(h) \leqq 3$. □

Lemma 6.4. Assumption A4 holds for both block structures and both problems if h is sufficiently small. For the k-line scheme,

$$q = 2/k, \quad D = \max \{3k+O(h), \ |e|_\infty/k\}, \tag{6.2}$$

while for the k×k block scheme

$$q = 4/k, \quad D = \max \{6k+O(h), \ (|d|_\infty+|e|_\infty)/k\}. \tag{6.3}$$

Proof. These statements essentially follow from Theorems 4.1 and 5.1. We sketch the argument for the k-line block structure (4.5). From Theorem 4.1 and (4.10),

$$(NU,V) = \Sigma_{i=1}^{P} \Sigma_{s=0}^{Q-1} [1 + he_{i,ks+1}/2]U_{i,ks}V_{i,ks+1}$$
$$+ \Sigma_{i=1}^{P} \Sigma_{s=0}^{Q-1} [1 - he_{i,ks+1}/2]U_{i,ks+1}V_{i,ks}.$$

Following the steps from (5.1) to (5.6), we estimate the term

$$T(i,s) := [1 + he_{i,ks+1}/2]U_{i,ks}V_{i,ks+1}$$

to get

$$\Sigma_i \Sigma_s T(i,s) = (1/k)(U,V) + \tilde{E}/2 + hR/2,$$

with $\tilde{E}$ satisfying (5.7) and

$$|R| \leq |e|_\infty[(1/k)(|U|,|V|) + |\tilde{E}|/2].$$

The second term in the expansion of (NU,V) is appraised in the same way, to yield

$$(NU,V) = (2/k)(U,V) + E,$$

$$|E| \leq h(2+h|e|_\infty)[(|\nabla_y U|,|V|)+(|U|,|\nabla_y V|)]$$
$$+ h(|e|_\infty/k)(|U|,|V|) + h^2 3k(1+h|e|_\infty/2)[(-\Delta U,U)+(-\Delta V,V)].$$

But this implies the inequality of A4, with D given by (6.2). □

Our next theorems follow immediately from these lemmas and Theorems 3.1 - 3.3.

Theorem 6.5. Let $\rho(kL)$ and $\rho(kB)$ denote the spectral radii for the k-line and k×k block structures, respectively, of the block Jacobi scheme applied to the elliptic problem (2.9). Let Λ_0 denote the smallest eigenvalue of the problem

$$Lv = \lambda v \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega.$$

Then

$$\rho(kL) = 1 - (k/2)\Lambda_0 h^2 + o(h^2),$$

(6.4)

$$\rho(kB) = 1 - (k/4)\Lambda_0 h^2 + o(h^2). \quad \square$$

Theorem 6.6. Let $\rho(SkL)$ and $\rho(SkB)$ denote the spectral radii for the k-line and k×k block structures, respectively, of the block Jacobi scheme applied to the parabolic problem (2.10), and suppose that P1 holds. Let

$$\Lambda_1 := \min \{\bar{c}(x,y) : (x,y) \in \bar{\Omega}\}.$$

Then

$$\begin{aligned} \rho(SkL) &\leqq 1 - (k/2)\Lambda_1 h^{\alpha} + o(h^{\alpha}), \\ \rho(SkB) &\leqq 1 - (k/4)\Lambda_1 h^{\alpha} + o(h^{\alpha}), \end{aligned} \tag{6.5}$$

and equality holds if $d \equiv e \equiv 0$, so that S1 is satisfied. □

We remark that the character "S" is to remind us of the singular perturbation nature of the parabolic equation.

When a matrix A under a block partition satisfies block property A, then the spectral radii ρ_{GS} of the Gauss-Seidel method (4.2) and ρ_ω of the SOR method (4.3) are determined by the spectral radius ρ of the Jacobi method ([1], [33, chapter 4], [35]):

$$\rho_{GS} = \rho^2, \quad (\rho_\omega + \omega - 1)^2 = \omega^2\rho^2\rho_\omega.$$

Moreover, ρ_ω is minimized for a specific ω:

$$\omega_b = 2/(1 + (1 - \rho^2)^{1/2}), \quad \rho_b = \omega_b - 1.$$

With the block structure imposed by (4.5) or (4.6), A has block property A. This observation proves our next result.

Corollary 6.7. Let A represent $h^2 L_h$. Then

$$\rho_{GS}(kL) = 1 - k\Lambda_0 h^2 + o(h^2),$$
$$\rho_b(kL) = 1 - 2(k\Lambda_0)^{1/2} h + o(h), \tag{6.6}$$

and

$$\rho_{GS}(kB) = 1 - (k/2)\Lambda_0 h^2 + o(h^2),$$
$$\rho_b(kB) = 1 - (2k\Lambda_0)^{1/2} h + o(h). \tag{6.7}$$

Let A represent $h^2 \ell_h$. Then

$$\rho_{GS}(SkL) \leqq 1 - k\Lambda_1 h^\alpha + o(h^\alpha),$$
$$\rho_b(SkL) \leqq 1 - 2(k\Lambda_1)^{1/2} h^{\alpha/2} + o(h^{\alpha/2}), \tag{6.8}$$

and

$$\rho_{GS}(SkB) \leqq 1 - (k/2)\Lambda_1 h^\alpha + o(h^\alpha),$$
$$\rho_b(SkB) \leqq 1 - (2k\Lambda_1)^{1/2} h^{\alpha/2} + o(h^{\alpha/2}). \tag{6.9}$$ □

7. Other operators. In this section we extend our theory to cover the more general operators L and ℓ defined by

$$Lu := -(au_x)_x - (bu_y)_y + du_x + eu_y + fu,$$
$$\ell u := cu_t + Lu. \tag{7.1}$$

For simplicity we have excluded terms in the cross-derivative u_{xy}. Self-adjoint operators L with this term have been discussed in [23]. We assume for convenience that a, b, c, d, e, and f are smooth functions on $\bar{\Omega}$, that c satisfies (2.5), and that

(7.2) $\quad a(x,y) \geq a_0 > 0, \quad b(x,y) \geq b_0 > 0, \quad f(x,y) \geq 0 \quad \text{on } \bar{\Omega}.$

L is uniformly elliptic by the strict positivity of a and b, and satisfies a maximum principle by virtue of the nonnegativity of f.

As in section 2, we let U be a mesh vector on the mesh points Ω_h defined by (2.3). At points (x_i,y_j) of Ω_h we define

$$a_{i+\frac{1}{2},j} := (a_{i,j} + a_{i+1,j})/2, \quad b_{i,j+\frac{1}{2}} := (b_{i,j} + b_{i,j+1})/2.$$

The discrete approximations to (7.1) are then

$$\begin{aligned}[L_hU]_{i,j} := &- [a_{i+\frac{1}{2},j}(U_{i+1,j} - U_{i,j}) - a_{i-\frac{1}{2},j}(U_{i,j} - U_{i-1,j})]/h^2 \\ &- [b_{i,j+\frac{1}{2}}(U_{i,j+1} - U_{i,j}) - b_{i,j-\frac{1}{2}}(U_{i,j} - U_{i,j-1})]/h^2 \\ &+ d_{i,j}(U_{i+1,j} - U_{i-1,j})/(2h) + e_{i,j}(U_{i,j+1} - U_{i,j-1})/(2h) \\ &+ f_{i,j}U_{i,j}\end{aligned} \tag{7.3}$$

and

$$[\ell_hU]_{i,j} := (c_{i,j}/\tau)U_{i,j} + [L_hU]_{i,j}. \tag{7.4}$$

It is not difficult to see that the machinery of section 3 still works. The main theorem of [24], which relates the minimal eigenvalues of (3.1) and (3.5), is easy to establish with L and L_h given by (7.1) and (7.3), respectively. Consequently Theorems 3.1, 3.2, and 3.3 apply, mutatis mutandis, to splittings of the matrix A arising from (7.3) or (7.4).

Now we must determine q for the Jacobi scheme, using either of the block structures (4.5) or (4.6).

For the k-line structure, a direct computation yields

$$\begin{aligned}(NU,V) = &\sum [b_{i,ks+\frac{1}{2}} + he_{i,ks+1}/2]U_{i,ks}V_{i,ks+1} \\ &+ \sum [b_{i,ks+\frac{1}{2}} - he_{i,ks}/2]U_{i,ks+1}V_{i,ks},\end{aligned}$$

where the sum is over $1 \leq i \leq P$, $0 \leq s \leq Q-1$. Consider a term

$$T(i,s) := [b_{i,ks+\frac{1}{2}} + he_{i,ks+1}/2]U_{i,ks}V_{i,ks+1}$$
$$= b_{i,ks+j}U_{i,ks}V_{i,ks+1} + h(e_{i,ks+1}/2)U_{i,ks}V_{i,ks+1}$$
$$+ [b_{i,ks+\frac{1}{2}} - b_{i,ks+j}]U_{i,ks}V_{i,ks+1}.$$

The factor in square brackets in the last term above is bounded by $hk|\nabla_y b|_\infty$, because b is smooth. Proceeding as in the proof of Lemma 6.4, we establish the validity of A4 with

$$(7.5) \qquad q = 2b/k, \quad D = \max \{3k|b|_\infty + O(h),\ 2|\nabla_y b|_\infty + |e|_\infty/k\}.$$

Observe that the variable coefficient b has led to a variable q.

In the same way, for the k×k block scheme we obtain

$$(7.6) \qquad \begin{aligned} &q = (2a + 2b)/k, \\ &D = \max \{3k|a|_\infty + 3k|b|_\infty + O(h),\ 2|\nabla_x a|_\infty + 2|\nabla_y b|_\infty + (|d|_\infty + |e|_\infty)/k\}. \end{aligned}$$

We collect our results in the following two theorems.

Theorem 7.1. Let $\rho(kL)$ and $\rho(kB)$ denote the spectral radii for the horizontal k-line and k×k block structures, respectively, of the block Jacobi scheme applied to the elliptic problem (2.9) with L_h given by (7.3). Let $\bar{\Gamma}_0$ denote the smallest eigenvalue of the problem

$$Lv = \gamma bv \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega,$$

and let Γ_0 denote the smallest eigenvalue of the problem

$$Lv = \gamma(a+b)v \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega.$$

Then

$$\rho(kL) = 1 - (k/2)\bar{\Gamma}_0 h^2 + o(h^2),$$

(7.7)

$$\rho(kB) = 1 - (k/2)\Gamma_0 h^2 + o(h^2). \quad \square$$

Theorem 7.2. Let $\rho(SkL)$ and $\rho(SkB)$ denote the spectral radii for the horizontal k-line and k×k block structures, respectively, of the block Jacobi scheme applied to the parabolic problem (2.10) with ℓ_h given by (7.4), and suppose that P1 holds. Let

$$\bar{\Gamma}_1 := \min \{\bar{c}(x,y)/b(x,y) : (x,y) \in \bar{\Omega}\},$$
$$\Gamma_1 := \min \{\bar{c}(x,y)/(a(x,y) + b(x,y)) : (x,y) \in \bar{\Omega}\}.$$

Then

$$\rho(SkL) \leqq 1 - (k/2)\bar{\Gamma}_1 h^{\alpha} + o(h^{\alpha}),$$

(7.8)

$$\rho(SkB) \leqq 1 - (k/2)\Gamma_1 h^{\alpha} + o(h^{\alpha}),$$

and equality holds if $d \equiv e \equiv 0$, so that S1 is satisfied. $\square$

The nonzero pattern of A is the same whether A arises from (2.6) or (7.3). In the more general case, then, A retains block property A for both the k-line and k×k block partitions. Consequently the analogue of Corollary 6.7 is valid. We leave a statement of this Corollary 7.3 to the reader.

8. Other domains. Extension of our results from two to m space dimensions is straightforward. We sketch this for the model problems set in the m-dimensional unit cube, and then return to the two-dimensional setting to discuss domains other than the unit square. Treatment of general domains in higher dimensions is similar, while more general operators on these domains can be handled as outlined in section 7.

Impose a uniform mesh of size $h = 1/(P+1)$ on the unit cube

$$\Omega(m) := \{x = (x_1, \ldots, x_m) \in \underset{\sim}{R}^m : 0 < x_i < 1\}.$$

Let β be a multi-index $\beta = (\beta_1, \ldots, \beta_m) \in \underset{\sim}{Z}^m$, and let $\gamma(i)$ be the multi-index whose ith component is 1 and whose other components are 0. By x_β we mean the point $x = (\beta_1 h, \ldots, \beta_m h)$ in $\underset{\sim}{R}^m$. Hence putting

$$B := \{\beta \in \underset{\sim}{Z}^m : 1 \leqq \beta_i \leqq P\}, \quad \bar{B} := \{\beta \in \underset{\sim}{Z}^m : 0 \leqq \beta_i \leqq P+1\}$$

allows us to write

$$\Omega_h(m) = \{x_\beta : \beta \in B\},$$

$$\partial\Omega_h(m) = \{x_\beta : \beta \in \bar{B} \text{ and at least one } \beta_i = 0 \text{ or } = P+1\}.$$

With the m-dimensional Laplacian $\Delta(m) := \Sigma_i \, \partial^2/\partial x_i^2$, our model operators are

$$Lu := -\Delta(m)u + \Sigma_{i=1}^m \, d_i \partial u/\partial x_i,$$

$$\ell u := c\partial u/\partial t + Lu; \tag{8.1}$$

we suppose that c and all d_i are smooth and that c satisfies (2.5). Discretization of these operators is done as in (2.6) and (2.7). Let $U = (U_\beta)$ be a mesh vector. The approximation of $\Delta(m)$ is given by

$$(8.2)\qquad [\Delta_h(m)U]_\beta := \Sigma_i\ (U_{\beta-\gamma(i)} - 2U_\beta + U_{\beta+\gamma(i)})/h^2,$$

and the discrete operators corresponding to (8.1) are

$$(8.3)\qquad [L_hU]_\beta := [-\Delta_h(m)U]_\beta + \Sigma_i\ d_{i,\beta}(U_{\beta+\gamma(i)} - U_{\beta-\gamma(i)})/(2h),$$

$$(8.4)\qquad [\ell_hU]_\beta := (c_\beta/\tau)U_\beta + [L_hU]_\beta.$$

In m dimensions there are m obvious block partitions. Suppose for example that m = 3. In the k-plane block structure, each block of U comprises the unknowns U_β associated with the points x_β on k consecutive planes. Indexing the blocks by s, we have

$$\tilde{U}_s := \{U_\beta : \beta \in B,\ k(s-1) < \beta_3 \leqq ks\}.$$

In like fashion, for blocks of k×k lines the basic subblock is

$$\tilde{U}_{r,s} := \{U_\beta : \beta \in B,\ k(r-1) < \beta_2 \leqq kr,\ k(s-1) < \beta_3 \leqq ks\},$$

and k×k×k blocks are given by

$$\tilde{U}_{r,s,t} := \{U_\beta : \beta \in B,\ k(r-1) < \beta_1 \leqq kr,\ k(s-1) < \beta_2 \leqq ks,\ k(t-1) < \beta_3 \leqq kt\}.$$

Let us agree to call a basic block an s-slice of U if we partition U by imposing restrictions of the form $k(s_i-1) < \beta_i \leqq ks_i$ on the indices $\beta_{m-s+1}, \ldots, \beta_m$. With this notation, it is easy to state and prove the m-dimensional version of Theorem 4.1.

Theorem 8.1. Let $k \geqq 2$. Denote by N_j the operator acting as $\tilde{N}$ in the x_j-direction. Let A represent h^2L_h or $h^2\ell_h$. For the block Jacobi scheme (4.1) based on s-slice blocks,

$$N = \Sigma_{j=m-s+1}^{m}\ N_j + O(h). \quad \square$$

Observe that s-slice decompositions preserve block property A. Assertions similar to Lemmas 6.1 - 6.4 are readily demonstrated; the following lemma collects the results.

Lemma 8.2. Let $k \geq 2$ and let A correspond to h^2L_h or $h^2\ell_h$. Assumptions A1 - A4 are satisfied by the block Jacobi method (4.1) based on s-slices. Specifically, for $1 \leq s \leq m$,

$$\|N\|_h \leq N_0 = 2s/k + O(h) \leq s + 1, \quad q = 2s/k,$$

$$D = \max \{3sk+O(h), \Sigma_{j=m-s+1}^{m} |d_j|_\infty/k\},$$

(8.5)

$$|E| \leq hD[\Sigma_{j=1}^{m} (|\nabla_j U|,|V|) + \Sigma_{j=1}^{m} (|U|,|\nabla_j V|) + (|U|,|V|)]$$

$$+ h^2D[(-\Delta_h(m)U,U) + (-\Delta_h(m)V,V)]. \quad \square$$

Now the machinery of section 3 grinds out theorems like those of section 6. Rather than turn the crank, we choose to consider problems with the model operators (2.1) set in more general domains Ω of $\underset{\sim}{R}^2$.

We begin by describing Ω and Ω_h. Assume that Ω is a bounded domain in $\underset{\sim}{R}^2$ with Lipschitz boundary $\partial\Omega$, and that locally Ω lies always on one side of $\partial\Omega$. This last condition ensures that Ω has no internal cusps. Boundedness implies that $\bar{\Omega}$ has "leftmost" and "bottommost" tangents $x = x_0 := \min \{x : (x,y) \in \bar{\Omega}\}$, $y = y_0 := \min \{y : (x,y) \in \bar{\Omega}\}$. Choose $h > 0$ and impose on $\underset{\sim}{R}^2$ a grid of lines

$$x = x_i := x_0 + ih, \quad y = y_j := y_0 + jh; \tag{8.6}$$

the intersections (x_i,y_j) are called grid points. Define $\bar{\Omega}_h$ to be the collection of all the grid points $(x_i,y_j) \in \bar{\Omega}$ together with all the

points of intersection of the grid lines (8.6) with $\partial\Omega$. Then $\partial\Omega_h :=$ $\bar{\Omega}_h \cap \partial\Omega$, and Ω_h consists of those points of $\bar{\Omega}_h$ that lie in Ω. The points of $\bar{\Omega}_h$ are called mesh points.

The points of Ω_h are conveniently viewed as being of two types: the four nearest neighbors of a regular point are themselves grid points (x_i, y_j), and the other points of Ω_h are called irregular. We consider finite difference approximations of the operators (2.1) that differ from the earlier constructions (2.6) and (2.7) only at irregular points. Any of a number of approximations at an irregular point will suffice for our purposes; it is only necessary that the approximation be at least an interpolation of degree 0 (see [13, pp. 199 ff.], [6]). Proceeding as before brings us to a linear system (2.11), to which the block iterative methods of section 4 can be applied.

To use the theory of section 3, we need to establish the relations

$$(NU,V) = (qU,V) + E,$$

$$|E| \leq hD[(|\nabla_x U|+|\nabla_y U|,|V|) + (|U|,|\nabla_x V|+|\nabla_y V|) + (|U|,|V|)] \tag{8.7}$$

$$+ h^2 D[(-\Delta_h U,U) + (-\Delta_h V,V)]$$

of A4 for mesh vectors U, V that vanish on $\partial\Omega_h$. It is clear from the argument of Lemma 6.4 that (8.7) remains true if $\bar{\Omega}$ is composed of "grid rectangles," whose sides are grid lines (8.6). All that remains is to prove (8.7) for more general domains. But this will follow if we can show that $\bar{\Omega}$ is almost a union of grid rectangles. To this end, note that $\bar{\Omega}$, being bounded, is contained in some rectangle

$$\tilde{R} := \{(x,y) \in \underset{\sim}{R}^2 : x_0 \leq x \leq x_{P+1},\ y_0 \leq y \leq y_{P+1}\}.$$

Pick an integer $k \geq 2$ and subdivide $\tilde{R}$ into closed subrectangles

$$R_{r,s} := \{(x,y) : x_{k(r-1)} \leq x \leq x_{kr},\ y_{k(s-1)} \leq y \leq y_{ks}\}.$$

$\bar{\Omega}$ is the union of rectangles $R_{r,s}$ that lie entirely within $\bar{\Omega}$, and fragments of such rectangles. Let R be the interior of the set

$$\bar{R} := \{R_{r,s} : R_{r,s} \subset \bar{\Omega}\},$$

and for any subset G of $\tilde{R}$ let $(U,V)_G$ denote the inner product of mesh vectors U and V over the mesh points in G. Because N is bounded, (8.7) is an easy consequence of the following lemma, whose proof is similar to the proof of inequality (13) in [11].

Lemma 8.3. Let U and V be mesh vectors that vanish on $\partial\Omega_h$. Then

$$(8.8) \qquad |(U,V)_\Omega - (U,V)_R| \leq 2h^2w^2[(-\Delta_h U,U) + (-\Delta_h V,V)]$$

whenever $h \leq h_0$, for some constants w and h_0 that depend only on Ω.

Proof. Divide $\partial\Omega$ into a finite number of pieces for which the angle of the tangent with either the x- or y-axis exceeds some positive value (say 30°). For instance, let B be a piece of the boundary that is this steep with respect to the x-axis. Let S be the horizontal strip of $\bar{\Omega}$ that abuts B and is k grid lines high, so that S is for some fixed s the union of rectangles $R_{r,s}$ and at most two pieces composed of fragments of such rectangles. Denote by F the piece touching B -- say the leftmost piece. The smoothness of $\partial\Omega$ ensures that there are positive constants h_0 and w, depending only on Ω, so that the x-width of F is at most wh when $h \leq h_0$. See Figure 1.

Now consider a horizontal mesh line $y = y_j$ in this strip S; let x_a be the leftmost and x_b the rightmost mesh points on the line, and let x_f be the first point in the leftmost subrectangle in S. Observe that $|x_f - x_a| \leq wh$. Because U vanishes on B, at each point x_i between x_a and x_f we have

$$U_{i,j} = \Sigma_{\sigma=a}^{i-1} (U_{\sigma+1,j} - U_{\sigma,j}),$$

and similarly for V. Hence $|U_{i,j}| \leq h\Sigma_{\sigma=a}^{i-1} |\nabla_x U_{\sigma,j}|$, so

$$\begin{aligned} |U_{i,j}V_{i,j}| &\leq h^2(\Sigma_{\sigma=a}^{f-1} |\nabla_x U_{\sigma,j}|)(\Sigma_{\sigma=a}^{f-1} |\nabla_x V_{\sigma,j}|) \\ &\leq h^2 w(\Sigma_{\sigma=a}^{f-1} |\nabla_x U_{\sigma,j}|^2)^{1/2}(\Sigma_{\sigma=a}^{f-1} |\nabla_x V_{\sigma,j}|^2)^{1/2} \\ &\leq h^2 w\Sigma_{\sigma=a}^{f-1} [|\nabla_x U_{\sigma,j}|^2 + |\nabla_x V_{\sigma,j}|^2]/2 \\ &\leq h^2 w\Sigma_{\sigma=a}^{b-1} [|\nabla_x U_{\sigma,j}|^2 + |\nabla_x V_{\sigma,j}|^2]/2. \end{aligned}$$

Now sum over x_i between x_a and x_f to get

$$\begin{aligned} |\Sigma_i U_{i,j}V_{i,j}| &\leq \Sigma_i |U_{i,j}V_{i,j}| \\ &\leq h^2 w^2 \Sigma_{\sigma=a}^{b-1} [|\nabla_x U_{\sigma,j}|^2 + |\nabla_x V_{\sigma,j}|^2]/2, \end{aligned}$$

and sum this last inequality over j to see that

$$|(U,V)_F| \leq h^2 w^2 [\|\nabla_x U\|_{h,S}^2 + \|\nabla_x V\|_{h,S}^2]/2.$$

Repeat this over every fragment F and boundary piece B. Then each subrectangle $R_{r,s}$ within $\bar{\Omega}$ is covered at most four times, so

$$|(U,V)_\Omega - (U,V)_R| \leq 2h^2 w^2 [\|\nabla_x U\|_h^2 + \|\nabla_y U\|_h^2 + \|\nabla_x V\|_h^2 + \|\nabla_y V\|_h^2];$$

(8.8) follows from Lemma 2.1. □

9. How general is the method? We expect that our techniques will permit us to analyze any block iterative scheme in which the blocks have a regular pattern, so that NU constitutes an orderly, weighted sampling of any mesh vector U. Evidently most finite difference approximations on nice meshes give rise to such operators N.

For instance, on uniform meshes in $\underset{\sim}{R}^m$ it will be true that

$$h^2[L_h U]_\beta = \Sigma_\xi \, A_\beta(x_\xi)U_\xi,$$

where ξ runs over the set of indices $\Xi(\beta)$ of "neighboring" mesh points of the point x_β. Each smooth enough coefficient $A_\beta(x_\xi)$ will satisfy

$$A_\beta(x_\xi) = \hat{A}_\xi(x_\beta) + O(h)$$

for some smooth function $\hat{A}_\xi$, and will be a linear combination of the coefficients of the differential operator L to which L_h is an approximation. Because N is derived from the matrix A representing h^2L_h, we will have

$$[NU]_\beta = \Sigma_\xi \, B_\beta(x_\xi)U_\xi. \tag{9.1}$$

If N is sufficiently regular, then there will be some regularly distributed subset S of points of Ω_h so that

$$B_\beta(x_\xi) = \begin{cases} 0 & x_\beta \notin S \\ \hat{B}_\xi(x_\beta) + O(h) & \xi \in \Xi(\beta),\ x_\beta \in S. \end{cases} \tag{9.2}$$

Consider now a typical term in (NU,V). At a point x_β of S,

$$[NU]_\beta V_\beta = \Sigma_\xi \, B_\beta(x_\xi)U_\xi V_\beta.$$

It follows from (9.1) and (9.2) that if U and V are smooth then

$$[NU]_\beta V_\beta \cong \Sigma_\xi \hat{B}_\xi(x_\beta) U_\beta V_\beta;$$

hence

$$(9.3) \qquad (NU,V) \cong (qU,V),$$

where q accounts for the terms $\hat{B}_\xi$ and the relative cardinalities of $\Xi(\beta)$, S, and Ω_h. The approximate inequality of (9.3) indicates that N is a weak multiplication operator. Development of an estimate of the error in (9.3) proceeds as in section 5 and in the proof of Lemma 6.4.

In effect, our view of N as a weak multiplication operator has already been used in [23, section 7], where a splitting somewhat different from the usual k-line block Jacobi scheme is treated. Rectangular meshes, which have uniform spacing h_i in the x_i-direction, can also be handled, as can other boundary conditions. We see then that this viewpoint unifies the derivation of the convergence rates of block iterative methods for elliptic and parabolic finite difference equations. In fact, the technique will also yield estimates of the rates of convergence of block relaxation methods applied to the matrices arising from finite element approximations. But finite elements are powerful in part because they admit irregular partitions of Ω. For such partitions it is not apparent how to group the unknowns so that a systematic block iterative scheme is easy to implement. We direct the reader to [4] for an example of a successful application of these ideas in a simple case, where the iterative method was easy to program.

Acknowledgements. We are grateful to Bill Buzbee for support and encouragement during the evolution of this work.

References

[1] R. J. Arms, L. D. Gates, and B. Zondek, A method of block iteration, J. Soc. Ind. Appl. Math., 4 (1956), pp. 220-229.

[2] K. Aziz and I. Babuska, eds., The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1973.

[3] D. L. Boley, B. L. Buzbee, and S. V. Parter, On block relaxation techniques, University of Wisconsin M.R.C. report no. 1860 (1978); also in Proceedings of the 1978 Army Numerical Analysis and Computer Conference, ARO Report 78-3, pp. 277-295.

[4] D. L. Boley and S. V. Parter, Block relaxation techniques for finite element elliptic equations: an example, Los Alamos Scientific Laboratory report LA-7870-MS (1979).

[5] C. de Boor, ed., Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974.

[6] J. H. Bramble and B. E. Hubbard, New monotone type approximations for elliptic problems, Math. Comp., 18 (1964), pp. 349-367.

[7] J. H. Bramble and A. H. Schatz, Rayleigh-Ritz-Galerkin Methods for Dirichlet's problem using subspaces without boundary conditions, Comm. Pure Appl. Math., 23 (1970), pp. 653-675.

[8] A. Brandt, Estimates for difference quotients of solutions of Poisson type difference equations, Math. Comp., 20 (1966), pp. 473-499.

[9] B. L. Buzbee, G. H. Golub, and C. W. Neilson, On direct methods for solving Poisson's equation, SIAM J. Numer. Anal., 7 (1970), pp. 627-656.

[10] P. G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, The Netherlands, 1978.

[11] R. Courant, K. Friedrichs, and H. Lewy, Über die Partiellen Differenzengleichungen der Mathematischen Physik, Math. Ann., 100 (1928), pp. 32-74 = On the partial difference equations of mathematical physics, IBM J. Res. Develop., 11 (1967), pp. 215-234.

[12] F. W. Dorr, The direct solution of the discrete Poisson equation on a rectangle, SIAM Rev., 12 (1970), pp. 248-263.

[13] G. E. Forsythe and W. R. Wasow, Finite Difference Methods for Partial Differential Equations, Wiley, New York, 1960.

[14] S. P. Frankel, Convergence rates of iterative treatments of partial differential equations, M. T. A. C., 4 (1950), pp. 65-76.

[15] A. George, Nested dissection of a regular finite element mesh, SIAM J. Numer. Anal., 10 (1973), pp. 345-363.

[16] _____, Numerical experiments using dissection to solve n by n grid problems, SIAM J. Numer. Anal., 14 (1977), pp. 161-179.

[17] C. E. Grosch, Poisson solvers on a large array computer, in Proceedings of the 1978 LASL Workshop on Vector and Parallel Processors, Los Alamos Scientific Laboratory, Los Alamos, NM, 1978, pp. 93-132.

[18] R. W. Hockney, A fast direct solution of Poisson's equation using Fourier analysis, J. Assoc. Comput. Mach., 12 (1965), pp. 341-361.

[19] M. J. Kascic, Jr., A direct Poisson solver on Star, in Proceedings of the 1978 LASL Workshop on Vector and Parallel Processors, Los Alamos Scientific Laboratory, Los Alamos, NM, 1978, pp. 137-164.

[20] M. Lees, Discrete methods for nonlinear two-point boundary value problems, in Numerical Solution of Partial Differential Equations, James H. Bramble, ed., Academic Press, New York (1966), pp. 59-72.

[21] J. Nitsche and J. C. C. Nitsche, Error estimates for the numerical solution of elliptic differential equations, Arch. Rational Mech. Anal., 5 (1960), pp. 293-306.

[22] S. V. Parter, Multi-line iterative methods for elliptic difference equations and fundamental frequencies, Numerische Math., 3 (1961), pp. 305-319.

[23] _____, On estimating the "rates of convergence" of iterative methods for elliptic difference equations, Trans. Amer. Math. Soc., 114 (1965), pp. 320-354.

[24] _____, On the eigenvalues of second order elliptic difference operators, University of Wisconsin Computer Sciences Department technical report 405 (1980).

[25] S. V. Parter and M. Steuerwalt, Another look at iterative methods for elliptic difference equations, University of Wisconsin Computer Sciences Department technical report 358 (1979).

[26] _____, On k-line and k×k block iterative schemes for a problem arising in three-dimensional elliptic difference equations, SIAM J. Numer. Anal., 17 (1980), pp. 823-839.

[27] D. J. Rose, A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972.

[28] D. J. Rose and R. A. Willoughby, eds., Sparse Matrices and Their Applications, Plenum Press, New York, 1972.

[29] A. H. Sameh, S. C. Chen, and D. J. Kuck, Parallel Poisson and biharmonic solvers, Computing, 17 (1976), pp. 219-230.

[30] G. Strang and G. J. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[31] P. N. Swarztrauber and R. A. Sweet, Efficient Fortran subprograms for the solution of separable elliptic partial differential equations, ACM Trans. Math. Software, 5 (1979), pp. 352-364.

[32] R. S. Varga, Factorization and normalized iterative methods, in Boundary Problems in Differential Equations, R. E. Langer, ed., University of Wisconsin Press, Madison, WI, 1960, pp. 121-142.

[33] _____, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[34] D. M. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc., 76 (1954), pp. 92-111.

[35] _____, Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

[36] O. C. Zienkiewicz, The finite element method: from intuition to generality, Appl. Mech. Rev., 23 (1970), pp. 249-256.

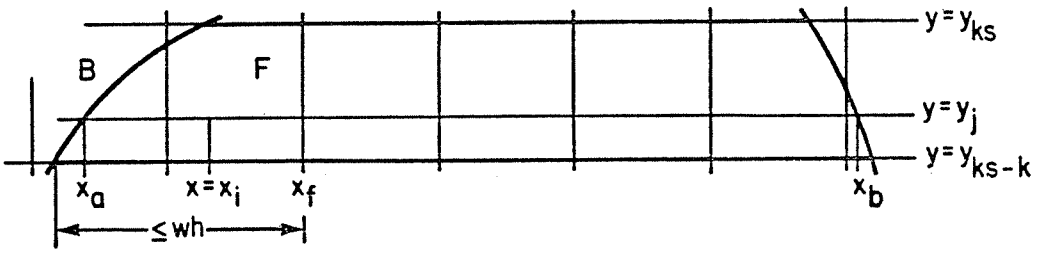

Figure 1. The strip S of $\bar{\Omega}$.

F is the union of the two leftmost fragments.

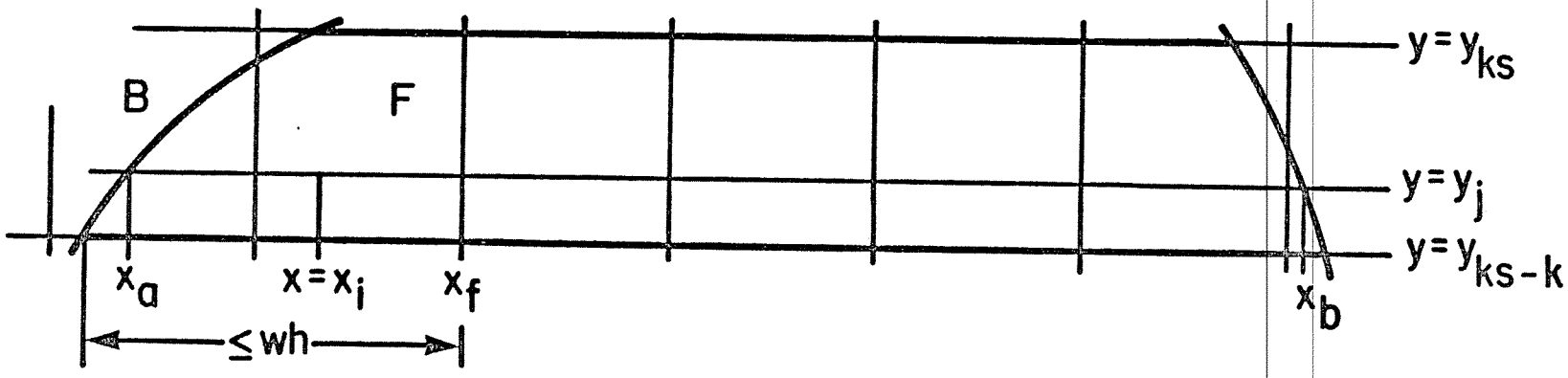

Figure 1. The strip S of $\bar{\Omega}$.

F is the union of the two leftmost fragments.