Computational techniques for the numerical solution of parabolic and elliptic partial differential equations

This item was submitted to Loughborough University's Institutional Repository by the/an author.

Additional Information:


Metadata Record: https://dspace.lboro.ac.uk/2134/34344

Publisher: © Christopher Roger Gane

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the published version.
Computational Techniques
for the Numerical Solution of
Parabolic and Elliptic
Partial Differential Equations

by

Christopher Roger Gane, B.Sc.

A Doctoral Thesis
Submitted in partial fulfilment of the requirements
for the award of
Doctor of Philosophy of the Loughborough University of Technology
August 1974

Supervisor: Professor D.J. Evans, Ph.D.
Department of Mathematics

© by Christopher Roger Gane, 1974.
Declaration

I declare that the following thesis is a record of research work carried out by me, and that the thesis is of my own composition. I also certify that neither this thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

C.R. Gane.
Acknowledgements

Firstly, I would like to extend my thanks to Professor D.J. Evans for his help and constant encouragement at every stage of the research work carried out. I am also indebted to him for going out of his way to accommodate me during my time in Sheffield.

I should also like to express my gratitude to Drs. A. Benson and A.R. Courlay for their valuable comments in private communications.

Finally, I would like to thank Miss J. Birds for her expert typing of this thesis.
# CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER 1: BACKGROUND INFORMATION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Partial differential equations - basic concepts</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Finite difference approximations to derivatives</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Irreducibility and diagonal dominance</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Non-negative and positive definite matrices</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Vector and matrix norms</td>
<td>14</td>
</tr>
<tr>
<td>1.6 Eigenvalues of a matrix</td>
<td>15</td>
</tr>
<tr>
<td>1.7 Convergence of sequences of matrices</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 2: PARABOLIC EQUATIONS IN TWO SPACE DIMENSIONS</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Introduction</td>
<td>20</td>
</tr>
<tr>
<td>2.2 Derivation of finite difference formulae</td>
<td>21</td>
</tr>
<tr>
<td>2.3 Explicit methods</td>
<td>22</td>
</tr>
<tr>
<td>2.4 A general implicit formula</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Boundary correction for the Peaceman-Rachford method</td>
<td>24</td>
</tr>
<tr>
<td>2.6 Stability</td>
<td>32</td>
</tr>
<tr>
<td>2.7 Convergence</td>
<td>33</td>
</tr>
<tr>
<td>2.8 Consistency</td>
<td>34</td>
</tr>
<tr>
<td>2.9 Stability of the general formula (2.19)</td>
<td>35</td>
</tr>
<tr>
<td>2.10 Remarks</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 3: ELLIPTIC EQUATIONS IN TWO SPACE DIMENSIONS</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>40</td>
</tr>
<tr>
<td>3.2 The model problem</td>
<td>41</td>
</tr>
<tr>
<td>3.3 Point-iterative methods</td>
<td>43</td>
</tr>
<tr>
<td>3.4 Block S.R.R. iterative methods</td>
<td>44</td>
</tr>
<tr>
<td>3.5 Alternating-Direction Implicit methods</td>
<td>53</td>
</tr>
</tbody>
</table>
CHAPTER 4: ALGORITHMIC SOLUTIONS OF CERTAIN LINEAR SYSTEMS

4.1 Introductory Remarks

Algorithm 4.1

Algorithm 4.2

Algorithm 4.3

Algorithm 4.4

CHAPTER 5: PERIPHERAL ORDERINGS OF GRID POINTS FOR PARABOLIC DIFFERENCE EQUATIONS

5.1 Introduction

5.2 Peripheral hopscotch using cartesian coordinates

5.3 An explicit finite difference formula for solving heat conduction problems in polar coordinates (r-θ)

5.4 The Peaceman-Rachford method in polar coordinates

5.5 Hopscotch methods in polar coordinates

5.6 Conclusions

CHAPTER 6: THE PEACEMAN-RACHFORD (A.D.I.) ITERATIVE METHOD IN r-θ GEOMETRY

6.1 Introduction

6.2 Statement of the problem and formation of the difference equations

6.3 The coefficient matrices of the difference equations

6.4 The iterative process

6.5 The commutative case

6.6 Estimation of eigenvalue bounds

6.7 A non-commutative case

6.8 Computational results

6.9 Conclusions
CHAPTER 1

BACKGROUND INFORMATION
1.1 Partial differential equations - basic concepts

The mathematical formulation of many problems in physics or engineering leads either to a partial differential equation (p.d.e.) or a set of such equations. A p.d.e. is an equation involving rates of change of unknown quantities (such as temperature, pressure, ...) with respect to two or more independent variables usually representing time, length or angle.

The exact solution to a p.d.e. in a region $R$ with boundary $B$ is some function which satisfies the equation at every point in $R$ and matches certain boundary conditions on $B$. For example, suppose $R$ is a bounded, connected, plane region with boundary $B$, and $s(x,y)$ a function defined on $B$. The problem may then be to find a function $U(x,y)$, which is continuous in $R+B$, twice differentiable in $R$, and satisfying in $R$ the general second order partial differential equation

$$a \frac{\partial^2 U}{\partial x^2} + b \frac{\partial^2 U}{\partial x \partial y} + c \frac{\partial^2 U}{\partial y^2} + d \frac{\partial U}{\partial x} + e \frac{\partial U}{\partial y} + fU = g,$$  \hspace{1em} (1.1)

and on $B$ the condition

$$U(x,y) = s(x,y).$$ \hspace{1em} (1.1)'

(Alternatively on $B$, the normal derivative $\partial U/\partial n$ or a linear combination of $U$ and $\partial U/\partial n$ may be specified.)

In (1.1), the coefficients $a,b,c,d,e,f,g$ may be constants (including the value zero) or functions of the independent variables $x$ and $y$, or of the dependent variable $U$. Special cases of (1.1) occur more frequently than any other because they are often the mathematical form of one of the conservation principles of physics.

The above equation is said to be

(i) elliptic if $b^2-4ac < 0$, \hspace{1em}
(ii) parabolic if $b^2 - 4ac = 0$,

and (iii) hyperbolic if $b^2 - 4ac > 0$,

for all $x,y,u$ in the region under consideration. Typical and respective examples are

(i) Poisson's

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2(x,y) ;
\]

(ii) the heat conduction equation, \( \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} ; \)

(iii) the wave equation, \( \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} . \)

The two equations in (i) are generally associated with equilibrium or steady-state problems. For example, the velocity potential for the steady flow of incompressible non-viscous fluid satisfies Laplace's equation, and is the mathematical way of expressing the idea that the rate at which such fluid enters any given region is equal to the rate at which it leaves it. Such a problem is referred to as a boundary value problem, since it is usual to have the function $u$ (say), specified on the boundary of the region, e.g. see equation (1.1)'.

Problems involving time $t$ as one independent variable lead usually to parabolic or hyperbolic equations. For instance, the simplest parabolic equation, given by (ii), governs the flow of heat in a bar or a rod. The boundary conditions consist of either the temperature given at two ends of the bar (two values of $x$), or some measure of diffusion from the ends. It is usual for the temperature distribution in the bar to be known at some instant in time. This is termed the initial condition. Such
problems are therefore called initial-boundary value problems.

Likewise, in the case of the simplest hyperbolic equation given by (iii), the boundary conditions are given on two lines \( x=a,b \) (say), and the initial conditions, usually \( U \) and its time derivative \( \partial U/\partial t \) are given at some instant in time. The majority of hyperbolic equations arise from vibration problems, or those in which discontinuities, such as shock waves, persist in time. This is as far as we shall go with our discussion of hyperbolic equations in this thesis, since as the title suggests, we are solely interested in parabolic and elliptic p.d.e.'s.

For arbitrarily shaped regions and general boundary conditions, it is not usually possible to determine an exact solution to a given p.d.e. In an attempt to solve such problems, approximate methods have been developed. These fall into two classes:

(a) analytical approximate methods where, for example, a truncated series may be obtained for the solution. These methods are usually valid only in certain areas of the region under consideration. They do provide extremely useful information concerning the character of the solution for critical values of the dependent variables, but tend to be more difficult to apply than the numerical methods, and will not be discussed further in this thesis.

(b) numerical approximate methods employing for example, finite difference techniques; these are more frequently used and universally applicable. The concept of a finite difference method is developed in the next section.
1.2 Finite difference approximations to derivatives

Let us consider, without loss of generality, that the problem is to solve an elliptic equation with independent variables \( x \) and \( y \), the region of problem solution being a connected region \( \mathcal{R} \) in the \( x-y \) plane. Let \( \overline{\mathcal{R}} = \mathcal{R} + \mathcal{B} \) denote the closure of the region \( \mathcal{R} \) with boundary \( \mathcal{B} \). We now overlay \( \overline{\mathcal{R}} \) with a system of rectangular meshes formed by two sets of equally spaced lines, one set parallel to \( Ox \) and the other parallel to \( Oy \). An approximate solution to the differential equation is then found at the points of intersection \( P_{i,1}, P_{i,2}, \ldots, P_{i,j}, \ldots \) of the parallel lines, which points are called (unknown) mesh points. (Other terms in common use are pivotal, nodal, grid or lattice points). This solution is obtained by approximating the p.d.e. over the area \( \mathcal{R} \) by \( N \) algebraic equations (linear, if the differential equation is linear) involving approximate values of \( U \) at the \( N \) mesh points internal to \( \mathcal{B} \). Let us assume that the grid is uniform and we have a square mesh size \( h \). This is illustrated below in Figure 1.1, the region \( \mathcal{R} \) being the unit square.

![Diagram](image)

**FIGURE 1.1**
**Taylor series expansion**

This method is probably the best known of all for deriving finite difference approximations. Assuming that $U(x,y)$ is sufficiently differentiable, then by Taylor's theorem

\[
U(x+h,y) = U(x,y) + \frac{h^2}{2} \frac{\partial^2 U}{\partial x^2} + \frac{h^3}{3!} \frac{\partial^3 U}{\partial x^3} + \frac{h^4}{4!} \frac{\partial^4 U}{\partial x^4} + \ldots
\]

\[
U(x,y+h) = U(x,y) + \frac{h^2}{2} \frac{\partial^2 U}{\partial y^2} + \frac{h^3}{3!} \frac{\partial^3 U}{\partial y^3} + \frac{h^4}{4!} \frac{\partial^4 U}{\partial y^4} + \ldots
\]

(1.2)

where the point $(x,y)$ and its four neighbouring points $(x+h,y)$, $(x,y+h)$ are contained in $\Omega$.

Combinations of these formulae yield

\[
\frac{\partial U}{\partial x} = \frac{U(x+h,y) - U(x-h,y)}{2h} + O(h^2) ,
\]

(1.3)

\[
\frac{\partial^2 U}{\partial x^2} = \frac{U(x+h,y) - 2U(x,y) + U(x-h,y)}{h^2} + O(h^2) ,
\]

(1.4)

\[
\frac{\partial U}{\partial y} = \frac{U(x,y+h) - U(x,y-h)}{2h} + O(h^2) ,
\]

(1.5)

and

\[
\frac{\partial^2 U}{\partial y^2} = \frac{U(x,y+h) - 2U(x,y) + U(x,y-h)}{h^2} + O(h^2) ,
\]

(1.6)

where $f(t) = O(g(t))$ as $t \to a$ if there exists a number $M$ such that

\[
\left| \frac{f(t)}{g(t)} \right| < M ,
\]

for all $t$ sufficiently close to $a$.

If for a general mesh point $(x,y) = (ih,jh)$, we denote $U(x,y)$ by $U_{i,j}$, then using equation (1.4) and (1.6), Poisson's equation

\[
\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = g(x,y) ,
\]

(1.7)

can be replaced at the point $(x_i,y_j) = (ih,jh)$ by
\[
\frac{1}{h^2} \left( u_{i+1,j}^* + u_{i-1,j}^* - 2u_{i,j}^* + u_{i+1,j+1}^* + u_{i+1,j-1}^* - 4u_{i,j}^* \right) = \delta_{i,j}
\]

or multiplying through by \(-h^2\),

\[
4u_{i,j}^* - 4u_{i+1,j}^* - 4u_{i-1,j}^* - 4u_{i+1,j+1}^* - 4u_{i+1,j-1}^* + 4u_{i,j}^* = -h^2 \delta_{i,j}
\]

The terms on the right hand side of (1.9), excluding \(-h^2 \delta_{i,j}\), are defined as the local truncation error of formula (1.9). The \(O(h^4)\) term is defined as the principal part of this error. Neglecting the local truncation error in (1.9), and scanning over the mesh points with such a formula gives us a set of simultaneous equations to be solved for the unknown functions \(u_{i,j}\), which denotes the finite difference approximation of the exact solution \(U_{i,j}\) at the point \((ib, jh)\). In matrix notation these equations can be written as

\[
Au = b
\]

Here, \(b\) is a vector whose components consist of the known values \(-h^2 \delta_{i,j}\) plus values of \(u_{i,j}\) given on the boundary \(B\) (N.B. \(u_{i,j}^* = u_{i,j}\) if \((x_i^*, y_j^*) = (ib, jh) \in B\)). If the region under consideration is the unit square, and there are \(h^2\) internal mesh points, then \(u\) and \(b\) are \((h^2 \times 1)\) column vectors, and the matrix \(A\) is of order \(h^2\). In view of (1.99), the smaller we take \(h\), the better accuracy we obtain. But since (for the square) \(h^{-1} = h\), any decrease in \(h\) and increase in accuracy will be offset by the fact that a larger set of
equations will have to be solved. We shall be considering this problem in more detail in Chapters 2 and 3.

Consider next the following initial-boundary value problem defined by

\[ \frac{\partial u}{\partial t} + \frac{x^2}{2} \frac{\partial^2 u}{\partial x^2} = 0 \]  

(1.11)

in the region

\[ 0 \leq x \leq a; \quad t \geq 0, \]

subject to the initial condition

\[ u(x,0) = f(x); \quad 0 \leq x \leq a \]

and the boundary conditions

\[ u(0,t) = g_0(t), \quad u(a,t) = g_a(t); \quad t \geq 0 \]

The region is covered by a rectangular grid with spacing \( h \)
in the \( x \)-direction and \( \Delta t \) in the \( t \) (time)-direction, as shown in Figure 1.2.

![Figure 1.2](image)

Let \( u^k_{i,j} \) denote the exact and finite difference solutions \( U(ih_k,k\Delta t) \), \( u(ih_k,k\Delta t) \) respectively.

Perhaps the simplest replacement of (1.11) using Taylor series expansions would be

\[ \frac{u_{i+1,j}^k - u_{i,j}^k}{\Delta t} \cdot \frac{u_{i-1,j}^k - 2u_{i,j}^k + u_{i+1,j}^k}{h^2} + O(\Delta t + h^2) \]
which is usually written in the form
\[ u^{k+1}_i = (1-2r)u^k_i + r(u^k_{i+1} + u^k_{i-1}) + O(\Delta t^2 + \Delta x^2) \]  
(1.12)
where \( r = \Delta t/\Delta x^2 \) is called the mesh ratio. We notice that
\( u^{k+1}_i \) is obtained solely in terms of the values of \( u \) at the
\((k)\)th time level. Such schemes are termed explicit. An
explicit scheme is one involving more than one point at the
\((k+1)\)th level. Neglecting the local truncation error terms
in (1.12) and scanning over each mesh point (in the interval
\( 0 < x < a \)) in turn, leads to a set of simultaneous equations
expressed in matrix form as
\[ u^{k+1}_i = A u^k_i, \quad k \geq 0 \]  
(1.13)
A corresponding implicit scheme could be likewise written as
\[ A_1 u^{k+1}_i = A_2 u^k_i, \quad k \geq 0 \]  
(1.14)
Both equations (1.13) and (1.14) represent a step by step
procedure for the numerical integration of the differential
equation. This process continues until the final solution
\( u^k \) is obtained, at some time \( T = k \Delta t \). Important concepts
regarding finite difference equations of type (1.13) and
(1.14) will be discussed in Chapter 2.

We have seen that applying finite difference methods
to the numerical solution of p.d.e.'s such as (1.7) and
(1.11), leads to the required solution of a system of
simultaneous equations as represented by (1.10) and (1.14)
respectively. System (1.10) for example has the unique
solution \( u = A^{-1}b \) provided \( A \) is non-singular, i.e. \( \det A \neq 0 \).
The matrix \( A \) derived in this way, is usually sparse (many of
its elements are zero), e.g. in view of formula (1.9)
involving only five mesh points, the matrix \( A \) in (1.10) will
have no more than that number of non-zero elements in each row. The exact structure of coefficient matrices arising from these and other problems will be illustrated in Chapters 2, 3 and 4. Methods of solving such systems of equations, particularly when the order $n$ of the matrix is large, depend very much on this structure, and a variety of techniques for this purpose will be presented throughout this thesis.

The coefficient matrix $A$ derived in these ways, often has particular properties such as diagonal dominance, irreducibility and positive definiteness. These and other properties will be defined and inter-related in the remainder of this Chapter, often without proof (although references for this purpose are given). We shall be principally concerned with real (square) matrices, although some of the results will apply to complex matrices as well. We have presupposed a basic knowledge of matrix and linear algebra theory, which is thoroughly covered in books such as Fox (1964), Faddeeva (1959) and Bellman (1960). Thus the reader is assumed to know for example, what the Jordan canonical form of a square matrix is, and the properties of vector and matrix norms.
2.3 Irreducibility and diagonal dominance

Definition 2.1

A matrix $A$ of order $n$ is irreducible if $n=1$ or if $n>1$ and given any two non-empty disjoint subsets $S$ and $T$ of $W$, the set of the first $n$ positive integers, such that $S \cup T = W$, there exists $i \in S$ and $j \in T$ such that $a_{i,j} \neq 0$.

An alternative definition which is treated and proved as a theorem in Young (1971), p.37 is the following.

Definition 2.2

$A$ is irreducible if and only if there does not exist a permutation matrix $P$ such that $P^{-1}AP$ has the form

$$P^{-1}AP = \begin{pmatrix} C & 0 \\ D & E \end{pmatrix}$$

where $C$ and $E$ are square matrices and where all elements of $O$ vanish.

Definition 2.3

Let $A$ be an $(n \times m)$ matrix, and consider any $n$ different points $P_1, P_2, \ldots, P_n$ in the plane, which we shall call nodes. If the element $a_{i,j}$ of $A$ is non-zero, connect the node $P_i$ to the node $P_j$ by means of a path $P_i P_j$, directed from $P_i$ to $P_j$ (see Figure 1.3). If $a_{i,i} \neq 0$, then we get a loop from $P_i$ to itself (see Figure 1.4).

When this process has been completed for all the $P_i$ $(i=1, 2, \ldots, n)$ the resulting paths are defined collectively as the finite directed graph, $G(A)$ of $A$. 
Example

If

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

then $G(A)$ is connected.

Figure 1.5

Definition 1.4

A directed graph is strongly connected if, for any ordered pair of nodes $P_i$ and $P_j$, there exists a directed path

$$P_i P_{k_1} P_{k_2} \ldots P_{k_{r-1}} P_{k_r} = j$$

connecting $P_i$ to $P_j$.

It is evident then that $G(A)$ of (1.15) is strongly connected.

The next theorem (which we state without proof) is a useful tool for proving matrices irreducible.

Theorem 1.1

An $n \times n$ matrix $A$ is irreducible if and only if its directed graph $G(A)$ is strongly connected.
With this theorem we can prove a useful Corollary (which is a generalisation of (1.15)) for future reference.

**Corollary**

Let $A$ be an $n \times n$ tri-diagonal matrix with $a_{ij} \neq 0$ for $|i-j| \leq 1$. Then $A$ is irreducible.

**Proof**

This is readily verified by induction on $n$. Let the matrix $A$ be

$$
\begin{bmatrix}
\ b_1 & c_1 \\
\ a_2 & b_2 & c_2 \\
\ & \ddots & \ddots & \ddots \\
\ & & a_{n-1} & b_{n-1} & c_{n-1} \\
\ & & & a_n & b_n
\end{bmatrix}
$$

(1.16)

and let us assume the result true when the order of $A$ is $n-1$.

(We know the result is true for $n=1$, by definition 1.1).

Starting with the node $P_n$, we notice that it is connected to $P_{n-1}$, and vice-versa. Since the nodes $P_1$ (1 is $n-1$) are strongly connected (by the inductive hypothesis), it follows immediately that $P_1P_2\ldots P_n$ must be strongly connected. This completes the proof.

**Definition 2.5**

An $n \times n$ matrix $A$ is **diagonally dominant** if

$$
|a_{1,1}| \geq \sum_{j=1}^{n} |a_{1,j}| \quad \text{for all } 1 \leq i
$$

(1.17)

The matrix $A$ is said to be **strictly diagonally dominant** if strict inequality is satisfied for all $1 \leq i$ in (1.17).

Similarly, $A$ is **irreducibly diagonally dominant** if $A$ is
irreducible and diagonally dominant, with strict inequality in (1.17) for at least one i.

2.6 Non-negative and Positive Definite Matrices

**Definition 2.6**

If $A$ is real and $x$ is complex, then $A$ is **positive definite** if

$$(x, Ax) > 0 \text{ for all } x \neq 0.$$  

(N.B. If $x, z$ are complex, then $(x, z) = \sum_{i=1}^{n} x_i \overline{y}_i$, where $\overline{y}_i$ is the complex conjugate of $y_i$.)

**Definition 2.7**

If $A$ is real and $x$ is complex, then $A$ is **non-negative** (positive semi-) definite if

$$(x, Ax) \geq 0 \text{ for all } x \neq 0,$$

with equality for at least one $x \neq 0$.

Next we state an important theorem (without proof), which is sometimes used as a definition of positive (non-negative) definite.

**Theorem 1.2**

A real matrix is positive (non-negative) definite if and only if it is symmetric and all its eigenvalues are positive (non-negative, with at least one eigenvalue equal to zero).

If $A$ is positive definite therefore, it can be written as $A = GJG^{-1}$ where $J$ is a positive diagonal matrix (i.e. $G^{-1}AG$ is the Jordan canonical form of $A$). Also, by a theorem from Young (1971), p.16, $G$ can be taken to be an orthogonal matrix (i.e. $G^T = G^{-1}$). If $J^1$ denotes the diagonal matrix whose elements are the positive square roots of the elements of $J$. 
then $A^\dagger = GA^{-1}$C is positive definite by Theorem 1.2. It should be noted that $(A^\dagger)^2 = (GA^{-1})^2 = A$.

**Theorem 1.3**

A real symmetric matrix $A$ is positive (non-negative) definite if and only if it can be written in the form $A = F^TP$ where $P$ is some non-singular (singular) matrix.

**Proof**

(i) Assume $A = F^TP$ (det $P \neq 0$). Then for any vector $x \neq 0$,

$$x^T A x = x^T F^T P x = (Fx)^T F x > 0.$$ 

Hence $A$ is positive definite by Definition 1.6.

(ii) Assume $A$ is positive definite (and real). We know $A = A^\dagger A^\dagger$, where $A^\dagger$ is defined above. Hence $A = (A^\dagger)^T A^\dagger$, since $A^\dagger$ is symmetric. Since $A^\dagger$ is also positive definite, $\det A^\dagger \neq 0$.

Thus, putting $P = A^\dagger$ gives the required condition.

The proof for the case of $A$ non-negative definite, follows in a similar manner.

**1.6 Vector and Matrix Norms**

**Definition 1.8**

The following quantities are defined as the 1, 2 and $\infty$ norms of a vector $x$:

$$||x||_1 = |x_1| + |x_2| + \ldots + |x_n|$$

(1.18)

$$||x||_2 = (|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2)^{1/2}$$

(1.19)

$$||x||_\infty = \max_i |x_i|$$

(1.20)

$||x||_2$ is often called the length of $x$.

**Definition 1.9**

A matrix norm is said to be compatible with a vector norm $||x||$ if
\[ ||Ax|| \leq ||A|| \cdot ||x||. \]

We now seek to derive matrix norms compatible with the vector norms of (1.18), (1.19) and (1.20). To do this it seems logical to express

\[ ||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||}, \quad (1.21) \]

which is equivalent to

\[ ||A|| = \sup_{||y||=1} ||Ay||. \quad (1.22) \]

**Definition 1.10**

A matrix norm constructed by means of (1.22) is said to be subordinate to the corresponding vector norm.

**Definition 1.11**

Let \( A \) be an \( n \times n \) matrix with eigenvalues \( \lambda_1, \ldots, \lambda_n \), then

\[ \rho(A) = \max_{\text{signed}} |\lambda_i| \]

is the spectral radius of \( A \).

From Definition 1.9, it is an easy matter to show

\[ ||A|| \geq \rho(A) \quad (1.23) \]

for an arbitrary \( n \times n \) matrix \( A \), and arbitrary norm.

**Proof**

If \( Ax = \lambda x \) is the equation associating the eigenvector \( x \) with the eigenvalue \( \lambda \) of \( A \), then from Definition 1.9,

\[ |\lambda| \cdot ||x|| = ||Ax|| = ||Ax|| \leq ||A|| \cdot ||x||, \]

which proves (1.23).

**Definition 1.12**

The subordinate norms associated with the 1(1.18), 2(1.19) and \(-1.20\) vector norms are

\[ ||A||_1 = \max_{j} \sum_{i=1}^{n} |a_{i,j}| \quad \text{(maximum absolute column sum)}, \quad (1.24) \]
\[ ||A||_2 = \rho(A^TA)^{\frac{1}{2}} \]  \hfill (1.25)
\[ ||A||_\infty = \max \sum_{j=1}^{n} |a_{i,j}| \] (maximum absolute row sum)  \hfill (1.26)

Hence, it follows from (1.25) that if \( A \) is an \( n \times n \) symmetric matrix, then
\[ ||A||_2 = \rho(A) \]  \hfill (1.27)
Moreover, if \( \ell_m(x) \) is any real polynomial of degree \( m \) in \( x \), then
\[ ||\ell_m(A)||_2 = \rho(\ell_m(A)) \]  \hfill (1.28)

**Proof**

Since \( A \) is symmetric,
\[ ||A||_2^2 = \rho(A^TA) = \rho(A^2) = \rho^2(A) \]
proving that \( ||A||_2 = \rho(A) \). Also, it is easily shown that \( \ell_m(A) \) is symmetric, which proves (1.28).

In order to evaluate (1.28) it is worth remembering that if the eigenvalues of \( A \) are \( \lambda \), then
\[ \text{the eigenvalues of } \ell_m(A) \text{ are } \ell_m(\lambda) \]  \hfill (1.29)

For a more detailed analysis of norms and corresponding proofs see for example, Noble (1969).

### 1.6 Eigenvalues of a Matrix

The following theorem is due to Gerschgorin (1931). Its proof may be found in Varga (1962), p.17.

**Theorem 1.4**

Let \( A \) be an arbitrary \( n \times n \) matrix, then all the eigenvalues of \( A \) lie within the union of the discs
\[ |x-a_{i,j}| \leq \sum_{j=1}^{n} |a_{i,j}| \cdot |\text{sign} \]
where
\[ j \neq i \]
An obvious Corollary to this is -

**Corollary**

If $A$ is an arbitrary $n \times n$ matrix, and

$$v = \max_{\text{sign } j=1} \sum_{j=1}^{n} |a_{i,j}|;$$

$$v' = \max_{\text{sign } i=1} \sum_{i=1}^{n} |a_{i,j}|;$$

then $\rho(A) \leq \min(v,v')$.

This follows since $A$ and $A^T$ have the same eigenvalues.

**Theorem 1.5**

Let $A$ be an $n \times n$ strictly or irreducibly diagonally dominant matrix. Then, the matrix $A$ is non-singular. If also $a_{i,j} > 0$ (sign), then the eigenvalues $\lambda_i$ of $A$ satisfy

$$\text{Re}(\lambda_i) > 0, \text{ sign}.$$  

**Proof**

The proof is a simple consequence of Theorem 1.4. For more detail see Varga (1962), p.23.

Since a symmetric matrix has real eigenvalues, we have

**Corollary**

If $A$ is a symmetric $n \times n$ strictly or irreducibly diagonally dominant matrix with positive diagonal entries, then $A$ is positive definite.

**Proof**

This follows from Theorem 1.5 and Theorem 1.2.

A useful theorem of Taussky (1949) is -

**Theorem 1.6**

Let $A$ be a $n \times n$ irreducible matrix which is diagonally dominant (see Definition 1.5). If $a_{i,i} > 0$, $\text{sign}$ and $a_{i,j} \leq 0$ for all $i \neq j$, then $\lambda = 0$ is an eigenvalue of $A$ if and only if

$$\sum_{k=1}^{n} a_{i,k} = 0 \text{ for all sign}$$
In our new developments, such matrices occur quite frequently.

2.7 Convergence of Sequences Of Matrices

Definition 2.13

Let A be an nxn matrix. Then A is convergent to the null matrix 0 if the sequence of matrices $A, A^2, A^3, \ldots$ converges to the null matrix 0, and is divergent otherwise.

Theorem 2.2.

If $A$ is an arbitrary nxn matrix, then $A$ is convergent if and only if $\rho(A) < 1$.

Proof

If

$$J = \begin{bmatrix} \lambda_1 & 1 & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_1 \end{bmatrix}$$

is a Jordan submatrix of $A$, then there exists a non-singular matrix $H$ such that $HAH^{-1}$ is a block diagonal matrix, defined by

$$\text{diag}(HAH^{-1}) = J$$

where $\lambda_1$ (IgIf) are the distinct eigenvalues of $A$. If each Jordan submatrix is raised to the power $m$, then the result tends to the null matrix as $m \to \infty$, if and only if $\rho(A) < 1$.

Two other books (apart from the three mentioned at the end of section 1.2) in which omitted proofs can be found are Varga (1962) and Young (1971).
CHAPTER 2

PARABOLIC EQUATIONS IN TWO SPACE DIMENSIONS
2.2 Introduction

In this Chapter, the aim is to provide background material for the new developments in Chapter 3. We concentrate mainly on two types of finite difference methods here, and also discuss the vital topics of stability, convergence and consistency associated with the numerical solution of time dependent problems.

To be precise, we shall be considering numerical methods of solution of the parabolic equation

\[ \frac{\partial U}{\partial t} = LU + g(x,y,t) , \quad U = U(x,y,t) , \quad (2.1) \]

where $L$ is a second order linear elliptic partial differential operator with respect to $x$ and $y$. The solution will be required in the cylinder $\mathbb{R}^n [0, t_\text{f}]$ where $\mathbb{R}$ is a closed, connected region in the $x$-$y$ plane, with continuous boundary $\partial \mathbb{R}$. Appropriate initial and boundary data are given on $t = 0$ and $\mathbb{R} [0, t_\text{f}]$ respectively. For example, we may be given

\[ U(x,y,0) = e(x,y) , \quad x,y \in \mathbb{R} , \quad (2.2) \]

and

\[ U(x,y,t) = f(x,y,t) , \quad x,y \in \mathbb{R}, \quad 0 \leq t \leq t_\text{f} , \quad (2.3) \]

where $e(x,y), f(x,y,t)$ are given for the prescribed values of $x,y,t$.

The region to be examined in $(x,y,t)$ space is covered by a rectilinear grid with sides parallel to the axes. In this Chapter, we assume a square grid of length $\Delta$ in the $x$-$y$ plane, and a grid spacing (or time step) of $\Delta t$ in the time ($t$) direction. The grid points $(x_i,y_j,t_k)$ are given by $x_i = i \Delta x$, $y_j = j \Delta y$, $t_k = k \Delta t$, where $i,j,k$ are integers. We let $u_{i,j}^k$ denote the approximate solution of (2.1) at the
point \((ih, jh, k At)\). The exact solution of the differential
equation at this point is denoted by \(u_{i,j}^k\).

### 2.3 Derivation of finite difference formulae

The exact difference replacement of equation (2.1) is
obtained from the Taylor expansion

\[
U(x, y, t + At) = (1 + At \frac{\partial}{\partial t} + \frac{(At)^2}{2!} \frac{\partial^2}{\partial t^2} + \ldots)U(x, y, t)
\]

\[
= \exp(At\frac{\partial}{\partial t})U(x, y, t).
\]

Hence, using (2.1) and (2.4) we have \((with g(x, y, t) = 0)\)

\[
u_{i,j}^{k+1} = \exp(At\frac{\partial}{\partial t})u_{i,j}^k.
\]

**Exact formulae connecting** \(\partial^2/\partial x^2\) and \(\delta_x\), the central
difference operator in the x-direction, are

\[
\frac{\partial^2}{\partial x^2} = \frac{2}{h^2} \sinh^{-1}\left(\frac{\delta_x}{2}\right) = \frac{1}{h}\left[\delta_x - \frac{\delta_x^3}{2!} + \frac{\delta_x^5}{4!} - \frac{\delta_x^7}{6!} + \ldots\right]
\]

where

\[
\delta_x u_{i,j}^k = u_{i+1,j}^k - u_{i-1,j}^k,
\]

\[
\delta_x^2 u_{i,j}^k = u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k, \text{ and so on.}
\]

A similar expression to equations (2.6) connects \(\partial^2/\partial y^2\) and
\(\delta_y\), where

\[
\delta_y u_{i,j}^k = u_{i,j+1}^k - u_{i,j-1}^k;
\]

\[
\delta_y^2 u_{i,j}^k = u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k, \text{ etc.}
\]

From (2.6) it readily follows that

\[
\frac{\partial^2}{\partial x^2} = \frac{1}{h^2}(\delta_x^2 - \frac{1}{12} \delta_x^4 + \frac{1}{90} \delta_x^6 - \ldots)
\]
and with "y" replacing "x" a similar expression for \(\frac{\partial^2 u}{\partial y^2}\) is obtained.

### 2.3 Explicit Methods

In this section, we consider the following particular case of equation (2.1), namely

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad (2.10)
\]

In this case, letting \(D_1 = \frac{\partial}{\partial x}\) and \(D_2 = \frac{\partial}{\partial y}\), (2.5) becomes

\[
u^{k+1} = \exp(\Delta t(D_1^2 + D_2^2))u^k
\]

where we have used \(u^k\) instead of \(U_{1,j}^k\). If we substitute the expressions for \(\frac{\partial^2}{\partial x^2}\) and \(\frac{\partial^2}{\partial y^2}\) from (2.9) into equation (2.11) and retain only second order central differences, we obtain the standard explicit replacement

\[
u^{k+1} = [1 + r(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})]u^k \quad (2.12)
\]

involving five points at the time level \(t^k = k\Delta t\). In (2.12), \(r = \Delta t/h^2\) and \(u^k\) replaced \(U_{1,i}^k\). Substituting from (2.7) and (2.8) for \(\frac{\partial^2}{\partial x^2} \frac{\partial^2}{\partial y^2}\) in (2.12) yields

\[
u_{1,i}^{k+1} = (1 - 4r)u_{1,i}^k + r(u_{1,i+1,j}^k + u_{1,i-1,j}^k + u_{1,i+1,j}^k + u_{1,i-1,j}^k + u_{1,i,j+1}^k)
\]

(2.13)

Using Taylor series expansions (see Section 1.2 and equation (2.4)), and defining the difference between the exact solutions of the differential and difference equations at the mesh point \((i\Delta x, j\Delta y, k\Delta t)\) as

\[
\delta_{i,j}^k = u_{i,j}^k - u_{i,j}^k
\]

from (2.13) we have,

\[
\delta_{i,j}^{k+1} = (1 - 4r)\delta_{i,j}^k + r(\delta_{i+1,j}^k + \delta_{i-1,j}^k + \delta_{i,j+1}^k + \delta_{i,j-1}^k)
\]

\[
+ \frac{(\Delta t)^2}{2} \left[ \frac{\partial^2 u}{\partial x^2} \right]_{i,j}^k - \Delta t \frac{h^2}{12} \left[ \frac{\partial^4 u}{\partial x^4} \right]_{i,j}^k + \ldots \ldots \quad (2.15)
\]

Thus showing that the local truncation of formula (2.13) is \(O(\Delta t^2 + \Delta t^2)\), assuming that \((i,j), (i+1,j), (i-1,j), (i,j+1)\)
and \((i,j-1) \in R\). **N.B.** We have assumed that any given partial derivatives of \(U\) are continuous and uniformly bounded for all \(x,y,t \in \mathbb{R}^3[0,T]\).

When \(R\) is non-rectangular, internal grid points adjacent to the boundary will normally require special treatment.

![Figure 2.1](image)

For example, the difference equation valid at a grid point such as \(P\) in Figure 2.1 is not given by (2.13), but by

\[
U_{i,j}^{k+1} = (1-2\alpha(\frac{1}{a} + \frac{1}{b})) U_{i,j}^k + 2\alpha \left[ \frac{1}{(a+i \Delta t)} U_{i+1,j}^k + \frac{1}{a(i \Delta t)} U_{i-a,j}^k \right] \\
\frac{1}{b(i \Delta t)} U_{i,j+b}^k + \frac{1}{(i \Delta t)} U_{i,j-1}^k \right] + O(\Delta t^2 + \Delta t \Delta h)
\]

(2.16)

where

\[ PQ = ah, \quad FR = bh \quad (0 < a, b < 1) \]

If \(P = (i,j)\), then (2.16) can be obtained from Taylor series expansions of \(U_{i,j}^{k+1}, U_{i+1,j}^k, U_{i-a,j}^k, U_{i,j+b}^k\) and \(U_{i,j-1}^k\) about \(U_{i,j}^k\).

### 2.4 A General Implicit Formula

Let us write the operator \(L\) in (2.1) as

\[
L = L(1) + L(2)
\]

(2.17)
where \( L^{(1)} \), \( L^{(2)} \) involve only derivatives (up to second order) with respect to \( x,y \) respectively. Hence in the case of (2.10) \( L^{(1)} = \partial^2 / \partial x^2 \) and \( L^{(2)} = \partial^2 / \partial y^2 \). \( L_h^{(1)} \), \( L_h^{(2)} \) are defined as the three point difference replacements of \( L^{(1)}, L^{(2)} \) respectively. Similarly, \( L_h \) will denote the five point difference replacement of \( L \). For equation (2.10), we have

\[
L_h^{(1)} u_{i,j}^k = \frac{u^k_{i-1,j}-2u^k_{i,j}+u^k_{i+1,j}}{h^2} ,
\]

\[
L_h^{(2)} u_{i,j}^k = \frac{u^k_{i,j-1}-2u^k_{i,j}+u^k_{i,j+1}}{h^2} ,
\]

and

\[
L_h u_{i,j}^k = L_h^{(1)} u_{i,j}^k + L_h^{(2)} u_{i,j}^k .
\]  

We now give a formula which includes three hopscotch processes, the Peaceman-Rachford (Alternating Direction Implicit or A.D.I.) and the Crank-Nicolson procedures for the approximate solution of equation (2.1). Consider the formulation

\[
u_{i,j}^{k+1} = \Delta t \left( \theta_{k+1}(i,1)u_{i,j}^{k+1} + \eta_{k+1}(i,2)u_{i,j}^{k+1} \right) - \Delta t (\theta_{k+1}(i,1)u_{i,j}^k + \eta_{k+1}(i,2)u_{i,j}^k) + \Delta t \left( \theta_{k+1}(i,1)u_{i,j}^k + \eta_{k+1}(i,2)u_{i,j}^k \right) ,
\]

with the restrictions

\[
\theta_{k+1}^{i,j} + \theta_{k+1}^{i,j} = 1
\]

and

\[
\eta_{k+1}^{i,j} + \eta_{k+1}^{i,j} = 1 .
\]

The restrictions (2.20) are required for (2.19) to be a consistent approximation (see Definition 2.6) to (2.1). In (2.19) we have used the relation (or splitting)

\[
g = g^{(1)} + g^{(2)} .
\]
We will define \( g^{(1)} \), \( g^{(2)} \) with respect to the above five mentioned methods below.

The value
\[
\theta_{ij}^k = \eta_{ij}^k = 1 \quad \text{for all } k, i, j
\]  
(2.22)
gives the Crank-Nicolson method (Crank and Nicolson, 1947);

\[
\frac{\theta_{ij}^k}{\eta_{ij}^k} = \frac{1}{2(1+(-1)^k)} = \begin{cases} 1 & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases}
\]  
(2.23)
gives the \textit{Peaceman-Rachford} (A.D.I.) procedure with a time step of \( 2\Delta t \) (Peaceman and Rachford, 1955);

\[
\theta_{ij}^k = \eta_{ij}^k = \begin{cases} 1 & \text{if } k+i+j \text{ is even} \\ 0 & \text{if } k+i+j \text{ is odd} \end{cases}
\]  
(2.24)
gives the \textit{odd-even hopscotch} method (Gourlay 1970);

\[
\theta_{ij}^k = \eta_{ij}^k = \theta_{i}^k = \begin{cases} 1 & \text{if } k+i \text{ is even} \\ 0 & \text{if } k+i \text{ is odd} \end{cases}
\]  
(2.25)
gives the \textit{y-line hopscotch} method (where 'y-line' means a line of grid points parallel to the \( y \)-axis) (Gourlay and McGuire, 1971):

\[
\eta_{ij}^{k+1} = \eta_{ij}^k = \eta_{ij}^k = \begin{cases} 1 & \text{if } k+i \text{ is even} \\ 0 & \text{if } k+i \text{ is odd} \end{cases}
\]  
(2.26)
gives the \textit{(y-line) A.D.I. hopscotch} method (Gourlay and McGuire, 1971).

The \( x \)-line hopscotch and \( (x-line) \) A.D.I. hopscotch methods are defined similarly by replacing \( \theta_{ij}^k \) with \( \eta_{ij}^k \) in (2.25) and (2.26) respectively.

We now draw attention to an important point regarding the odd-even and line hopscotch methods. Substituting \( \theta_{ij}^k = \eta_{ij}^k \) in (2.19), we obtain
\[ u_{i,j}^{k+1} - \Delta t \theta_{i,j}^{k+1} \left( L_h u_{i,j}^{k+1} + g_{i,j}^{k+1} \right) \]

\[ = u_{i,j}^k + \Delta t \theta_{i,j}^k \left( L_h u_{i,j}^k + g_{i,j}^k \right) \quad (2.27) \]

which is the odd-even or y-line hopscotch method depending upon whether we use (2.24) or (2.25) respectively. Writing (2.27) with \( k \) replaced by \( k+1 \) and eliminating

\[ \Delta t \theta_{i,j}^{k+1} \left[ L_h u_{i,j}^{k+1} + g_{i,j}^{k+1} \right] \]

from the resulting equation (using (2.27)), we have

\[ u_{i,j}^{k+2} - \Delta t \theta_{i,j}^{k+2} \left[ L_h u_{i,j}^{k+2} + g_{i,j}^{k+2} \right] \]

\[ = 2u_{i,j}^{k+1} - u_{i,j}^k - \Delta t \theta_{i,j}^k \left[ L_h u_{i,j}^k + g_{i,j}^k \right] \quad (2.28) \]

which reduces to

\[ u_{i,j}^{k+2} = 2u_{i,j}^{k+1} - u_{i,j}^k \quad (2.29) \]

when \( k+i+j \) (\( k+1 \)) is odd, since \( \theta_{i,j}^{k+2} \left( \theta_{i,j}^k \right) = \theta_{i,j}^k \left( \theta_{i,j}^k \right) \). This equation is the essence of the so-called fast odd-even (line) hopscotch algorithm; it is given in full in Courlay (1970) p.378, and McGuire (1970), p.6. It should be pointed out that because of the nature of formulae (2.22), (2.23) and (2.26), we cannot obtain an equation of the form (2.29) for the Crank-Nicolson, Peaceman-Rachford and ADI-hopscotch schemes, respectively.

We note at this point that since \( \theta_{i,j}^{k+2} = \theta_{i,j}^k \) and \( n_{i,j}^{k+2} = n_{i,j}^k \), it follows that after two time steps using (2.19) we are back at the same equation. Thus it is reasonable to regard the general method (2.19) as a two step process rather than as a single one.

As yet, we have made no mention of the order in which the finite difference solutions are calculated at the grid points.
Returning then to equation (2.19), we shall see that this is governed first by our choice of \( \delta_{ij}^k \), \( \eta_{ij}^k \) and secondly by our desire to solve the resulting linear systems of equations with a feasible and efficient algorithm. Consider then the region \( R \) defined earlier. If \( L_h \) is a five point difference operator, then ordering (numbering) the mesh points along successive rows or columns of the grid, the Crank-Nicolson scheme (defined by (2.19) and (2.22)) will involve the highly implicit solution (after the finite difference formula has been applied to each grid point \( (x_i, y_j) \in R \) in the order specified) of a five banded sparse matrix system to obtain the solution vector \( u_i^{k+1} \) \((k \geq 0)\). So far, there is no efficient general algorithm for doing this. Hence, in our future discussion, we will ignore this method.

The Peaceman-Rachford method (2.19) and (2.23) is

\[
(1-\Delta t L_h^{(2)})u_{i,j}^{k+1} = (1+\Delta t L_h^{(1)})u_{i,j}^k + \frac{\Delta t}{2} (u_{i,j}^k - u_{i,j}^{k+1})
\] (2.30)

when \( k \) is even, and

\[
(1-\Delta t L_h^{(1)})u_{i,j}^{k+1} = (1+\Delta t L_h^{(2)})u_{i,j}^k + \frac{\Delta t}{2} (u_{i,j}^k - u_{i,j}^{k+1})
\] (2.31)

when \( k \) is odd. Notice that we have taken \( g^{(1)} \), \( g^{(2)} = g/2 \).

If \( L_h^{(2)} \) is a three point finite difference replacement of derivatives involving \( y \) only (up to second order), then ordering the grid points along successive vertical (\( y \)-) lines (columns) would involve the solution of a number of tri-diagonal systems of equations for (2.30), corresponding to the number of columns on the grid. Similarly, a row (or horizontal/\( x \)-line) ordering for (2.31), where \( L_h^{(1)} \) is the three term difference replacement of partial derivatives in \( x \) only (up to second order), would
also involve solution of tri-diagonal systems whose number agreed with the number of rows on the grid. An efficient algorithm for solving such systems will be given in Chapter 4. This procedure of solving equations (2.30) and (2.31) inside the region \( R \), is continued until the desired solution is attained for some time \( t = T \). The name "Alternating Direction Implicit" (A.D.I.) is derived from the alternate solution of equations (2.30) and (2.31), along all the vertical and then along all the horizontal grid lines, respectively.

From (2.19) and (2.24), starting with \( k \) odd, the odd-even hopscotch process involves firstly the solution at points with \((i+j)\) even using the explicit formula

\[
 u_{k+1}^{i,j} = (1 + \Delta t L_h) u_{k}^{i,j} + \Delta t g_{k}^{i,j} \tag{2.32}
\]

and secondly the solution at points with \((i+j)\) odd using the implicit formula

\[
 (1 - \Delta t L_h) u_{k+1}^{i,j} = u_{k}^{i,j} + \Delta t g_{k+1}^{i,j}. \tag{2.33}
\]

For \( k \) even, the roles are reversed, i.e. when \((i+j)\) is odd we use the explicit formula, and after that the implicit formula at points where \((i+j)\) is even. Providing \( L_h \) is an \( E \)-operator (viz. \( L_h u_{k}^{i,j} \) is a finite difference replacement which involves only \( u_{k}^{i,j} \) and \( u_{k+1}^{i,j}, u_{i-1,j}, u_{i+1,j}, u_{i,j-1} \)), then it is evident from the above remarks that the odd-even hopscotch process is essentially explicit. If \( L_h \) were a nine point difference operator however, this would not be the case, and the above procedure would lose its advantages.
The line hopscotch procedure is obtained from (2.19) and (2.25). As indicated previously the term "line" can mean a row or a column. Which direction to choose will be considered in Chapter 5. Let us for now assume a row ordering. In other words, when \( k \) is odd we solve first using (2.32) for the points with \( i \) even. Then the remaining points (\( i \) odd) which consist of alternate "lines of points" are solved for using (2.33). This involves solving tri-diagonal systems using Algorithm 4.2 (Chapter 4) for example. When \( k \) is even, a reverse procedure is adopted. It is now apparent that a nine point formula \( L_h \) can be easily coped with using line hopscotch, since the implicit solution along alternate grid lines using the second formula (2.33), will still be of tri-diagonal systems of equations.

Lastly, we consider the A.D.I. hopscotch method given by (2.19) and (2.26). Again a row or column ordering is possible, and for convenience we assume the former. When \( k \) is odd and \( i \) is even, formula (2.31) is used and this derives tri-diagonal systems of equations to be solved as before. Then when \( i \) is odd, formula (2.30) is used on these points in an explicit nature at the same time level. When \( k \) is even, (2.31) is used when \( i \) is odd, and (2.30) when \( i \) is even at the same time level, and so on.

Having defined the general finite difference formula (2.19) locally and described the respective orderings of the mesh points, we now proceed to discuss the global form of this general method (over the entire mesh). First, we
introduce some notation. Define the solution vector \( u \) as
\[
\mathbf{u} = \{(u_{i,j})_i\}
\]
going over the mesh points ith level (y-line) by level (-line).
We define the vector \( q \) similarly. Also let the diagonal matrices \( \mathbf{I}_1^0, \mathbf{I}_1^n \) be defined by
\[
(\mathbf{I}_1^0)_{i,j} = \delta_{i,j}^0 2k,
(\mathbf{I}_1^n)_{i,j} = \delta_{i,j}^n 2k
\]
and let
\[
\mathbf{I}_2^0 = \mathbf{I} - \mathbf{I}_1^0,
\mathbf{I}_2^n = \mathbf{I} - \mathbf{I}_1^n
\]
a matrix \( \mathbf{A} \) by
\[
(\mathbf{A}u)_{i,j} = -h^2 l_h u_{i,j}
\]
and \( \mathbf{H}, \mathbf{V} \) by
\[
(\mathbf{H}u)_{i,j} = -h^2 (1) u_{i,j},
(\mathbf{V}u)_{i,j} = -h^2 (2) u_{i,j}
\]
and let \( r = \Delta t/h^2 \) be assumed constant. Let also
\[
\mathbf{A}_1 = \mathbf{I}_1^0 \mathbf{H} + \mathbf{I}_1^n \mathbf{V},
\mathbf{A}_2 = \mathbf{I}_2^0 \mathbf{H} + \mathbf{I}_2^n \mathbf{V}
\]
The global form of the general difference equation (2.19) can then be written as
\[
\begin{align*}
(\mathbf{I} + r\mathbf{A}_1)u^k &= (\mathbf{I} - r\mathbf{A}_2)u^{k-1} + \Delta t(\mathbf{I}_2^0 (1) \mathbf{H} + \mathbf{I}_2^n (2) )^k \\
&\quad + \Delta t(\mathbf{I}_2^0 (1) \mathbf{H} + \mathbf{I}_2^n (2) )^{k-1} \\
(\mathbf{I} + r\mathbf{A}_2)u^{k+1} &= (\mathbf{I} - r\mathbf{A}_1)u^k + \Delta t(\mathbf{I}_2^0 (1) \mathbf{H} + \mathbf{I}_2^n (2) )^{k+1} \\
&\quad + \Delta t(\mathbf{I}_2^0 (1) \mathbf{H} + \mathbf{I}_2^n (2) )^k
\end{align*}
\]
with boundary conditions and initial conditions absorbed in the g vectors. The values of \( A_1, A_2 \) which give the five previously mentioned methods are analogous to conditions (2.22)-(2.26). We can now see that the class of methods given by (2.19) is obtained from a splitting of the coefficient matrix \( A \) in the form

\[
A = A_1 + A_2 .
\]

(2.41)

Eliminating \( u^k \) from (2.39) and (2.40) gives

\[
u^{k+1} = Tu^{k-1} + b^k ,
\]

(2.42)

\( k=1,3,5,... \), where \( b^k \) is independent of the \( u \)'s and where

\[
T = [I+TA_1][I+TA_2][I+TA_1][I+TA_2]^{-1}
\]

(2.43)

and

\[
T = [I+TA_2][I+TA_2][I+TA_2][I+TA_2]^{-1} .
\]

(2.44)

**Definition 2.1**

The matrix \( T \) in (2.42) is referred to as the amplification matrix.

### 2.6 Boundary correction for the Peaceman-Rachford (P-R) method

We now investigate the P-R method from a slightly different angle. Let us consider the case when \( L^{(1)} = \partial^2/\partial x^2 \) and \( L^{(2)} = \partial^2/\partial y^2 \) \( (L = L^{(1)} + L^{(2)}) \). Let us assume in (2.5) that we have a time step of \( 2\Delta t \) (i.e., this replaces "\( \Delta t \)"). (2.5) can then be written in the form

\[
\exp[-\Delta t(L^{(1)} + L^{(2)})]u^{k+2} = \exp[\Delta t(L^{(1)} + L^{(2)})]u^k .
\]

(2.45)
Let us use (2.9) as expressions for \( L^{(1)} \) and \( L^{(2)} \). Then correct to second differences equation (2.45) can be written
\[
\exp(-r \delta_x^2) \exp(-r \delta_y^2) u^{k+2} = \exp(r \delta_x^2) \exp(r \delta_y^2) u^k \quad (r=\Delta t/h^2),
\]
and on expansion, the formula
\[
(1-r \delta_x^2)(1-r \delta_y^2) u^{k+2} = (1+r \delta_x^2)(1+r \delta_y^2) u^k + O(\Delta t^3+\Delta t h^2) \tag{2.46}
\]
is obtained. If an intermediate value \( u^{(k+1)a} \) is introduced, formula (2.46) can be split into the two formulae,
\[
\begin{align*}
(1-r \delta_x^2) u^{(k+1)a} &= (1+r \delta_x^2) u^k \\
(1-r \delta_y^2) u^{k+2} &= (1+r \delta_y^2) u^{(k+1)a}
\end{align*}
\tag{2.47}
\]
With \( g=0 \), this is equivalent to equations (2.30) and (2.31) i.e. the Peaceman-Rachford formula.

In the original Peaceman-Rachford formulation, assuming the region \( R \) to be rectangular,
\[
u^{(k+1)a}_{i,j} = u^{(k+1)}_{i,j} \quad \text{for} \quad (i,j) \in \mathbb{B}_x ,
\tag{2.48}
\]
the part of the boundary parallel to the \( y \)-axis, and hence the intermediate solutions in this case are located at time \( t = (n+1) \Delta t \); Fairweather and Mitchell (1967) show that this formula (2.48) in general leads to a loss of accuracy when the boundary conditions are time-dependent. The local truncation error is only correct to \( O(\Delta t^3+\Delta t h^2) \) provided the following formula is used at the intermediate time level:
\[
u^{(k+1)a}_{i,j} = \frac{1}{2} [1+r \delta_x^2] u^k_{i,j} + \frac{1}{2} [1-r \delta_y^2] u^{k+2}_{i,j} \quad \text{for} \quad (i,j) \in \mathbb{B}_x .
\tag{2.49}
\]

A more detailed investigation of intermediate boundary
values (boundary correction) in the Peaceman-Rachford as well as in other A.D.I. methods is given in Fairweather and Mitchell (1967).

3.6 Stability

Let us firstly assume from now on, that we have a constant time increment, $\Delta t$. If our finite difference equations in global form are

$$u^{k+1} = Tu^k + b^k, \quad k \geq 0,$$

(2.50)

where $b^k$ is independent of the $u$'s, we arrive by induction at

$$u^k = Tu^0 + c^k, \quad k > 0,$$

(2.51)

where $c^k$ is independent of the $u$'s also.

The requirement of stability is that no component of the initial data $u^0$ can be amplified more than a certain amount in the numerical procedure. It is therefore necessary that the sequence of vectors $u^k$ $(k \geq 0)$ be bounded in norm for all $u^0$ as $k \to \infty$, with $\Delta t$ and $h$ remaining fixed. From (2.51) we have

$$||u^k|| \leq ||T|| \cdot ||u^0|| + ||c^k||,$$

(2.52)

so that if $||T||$ is uniformly bounded (and $||u^0||, ||c^k||$) then so is $||u^k||$ for all $k > 0$. Hence a sufficient condition for the stability of (2.50) is

**Condition 2.3**

The finite difference scheme (2.50) is stable if

$$||T|| \leq C, \quad \text{where } C \text{ is some positive constant independent of } k.$$

Since also $T^k > 0$ if and only if $\rho(T) < 1$ (Theorem 1.2) we have
\textbf{Condition 2.3}

Formula (2.50) is stable if $\rho(T) < 1$.

For our final condition, we notice from (2.52) that

\[ ||u^k|| \leq ||\Omega||^k ||u^0|| + ||a^k||. \]

(2.53)

Thus, assuming $||u^0||$ and $||a^k||$ to be bounded for all $k \geq 0$, (2.53) yields

\textbf{Condition 2.4}

Formula (2.50) is stable if $||\Omega|| < 1$.

The concept of stability makes no reference to the differential equations one wishes to solve, but is a property solely of the difference equations.

2.7 \textit{Convergence}

Let $\bar{u}(t_j)$ be the exact solution vector of the differential equation (2.1) say, at all internal grid points at some fixed time $t = t_j$. Let also the vector $u(t_j)$ be the corresponding finite difference solutions of (2.1) at the mesh points, at the same fixed time $t = t_j$. Define the error vector $\varepsilon(t_j)$ as

\[ \varepsilon(t_j) = \bar{u}(t_j) - u(t_j). \]

(2.54)

\textbf{Definition 2.5}

The difference solution $u(t_j)$ is then said to be \textit{convergent} to the exact solution $\bar{u}(t_j)$ (of the differential equation) if and only if

\[ ||\varepsilon(t_j)|| \to 0 \]

(2.55)

when $\Delta t \to 0$ and $h \to 0$ in some manner; such that if $t_k = k \Delta t$,

\[ x_j = jh, y_j = jh, \] 

then $i, j, k \to \infty$ in such a way that $x_j, y_j, t_k$ remain fixed $||\cdot||$ is some suitable norm.
2.8 Consistency

Let \( d_{i,j}^k \) be the local truncation error of a finite difference formula at the grid point \((x_i, y_j, t^k) = (ih, jh, k\Delta t)\). Let also \( (d(t))_{i,j} = d_{i,j}^k \), for all mesh points \((x_i, y_j, t^k)\).

**Definition 2.6**

Then a finite difference approximation to a parabolic equation is consistent if

\[
\left| \frac{d(t)}{\Delta t} \right| \rightarrow 0 \quad (2.56)
\]

when \( \Delta t \to 0, \ h \to 0 \) in some manner, and \( \| \cdot \| \) represents a suitable norm.

2.9 Stability of the general formula (2.19)

Before proving the stability of (2.19), we state two lemmas of Kellog (1964) which we employ below.

**Lemma 2.1**

If \( p > 0 \) and \((B + B^*)\) is non-negative (positive) definite then \((\rho I + B)\) has a bounded inverse and \( \| (\rho I + B)^{-1} \| \leq \rho^{-1} (\rho^{-1}) \).

**Lemma 2.2**

If \( p > 0 \) and \((B + B^*)\) is non-negative (positive) definite, then the operator \((\rho I - B)(\rho I + B)^{-1}\) is bounded and

\[
\| (\rho I - B)(\rho I + B)^{-1} \| \leq 1. \quad (2.57)
\]

Returning to (2.43) and (2.44) we have

\[
T^k = [I + rA_2^{-1}]^{-1} T^k [I + rA_2]. \quad (2.57)
\]

Using the \( L_2 \) matrix norm defined in section 1.5, we have from (2.57),

\[
\| T^k \| \leq \| [I + rA_2^{-1}]^{-1} \| \cdot \| T^k \| \cdot \| [I + rA_2] \|. \quad (2.58)
\]

Hence if \((A_2 + A_2^*)\) is non-negative definite, by Lemma 2.1

\[
\| [I + rA_2^{-1}]^{-1} \| \leq 1,
\]
and further if \((A_1 + A_2)\) is non-negative definite, by

Lemma 2.2,

\[ ||\hat{z}|| \leq 1.\]

Hence from (2.50), \(||x^k|| \leq C\), since \(r\) is fixed, where \(C\)

is some positive constant independent of \(k\). Hence from

Condition 2.2, the method (2.19) will be stable under

the above requirements. Hence we can state

**Theorem 3.1**

The method (2.19) is stable for the solution of (2.1)

if \(A_1\) and \(A_2\) both satisfy

\[ (u, Q u) + (Q u, u) > 0 \quad (Q = A_1, A_2) \]  \hspace{1cm} (3.59)

for all \(u \neq 0\) of appropriate order.

**Corollary**

Algorithm (2.19) is stable for the solution of (2.1)

if

\[ (u, Q u) + (Q u, u) > 0 \quad (Q = L^0, L^v, i=1,2) \]  \hspace{1cm} (2.60)

for all \(u \neq 0\) of appropriate order.

This follows immediately from Theorem 2.1 and (2.38).

The accuracy, convergence and consistency of each of

the five methods contained in the general formula (2.19)

(and defined by (2.22)-(2.26)) will be discussed in

Chapter 5 along with new developments.

So far we have tended to rather ignore the explicit

method of section 2.3. As far as equation (2.10) is

concerned, the explicit formula (2.13) is both convergent

and consistent in the region \(R\) under consideration. As

far as the stability of (2.13) is concerned, it is an easy

matter to show that the method is stable for \(r \leq 1\). (The
analysis for this is outlined in many books on numerical methods, e.g. Mitchell (1969). Hence if \( h = 0.1 \), it would require four hundred time steps to advance one unit in time. It is mainly this drawback, which leads us to prefer methods such as A.D.I. (Peaceman-Rachford) and Hopscotch, since they are stable for all values of \( r > 0 \).

2.10 Remarks

A more detailed study of the general formula (2.19) especially its extension to n-dimensional space, is contained in Gourlay and McGuire (1971) and McGuire (1970). The general formula (2.19) was in fact first presented (in published form) in the former reference.

It seems advantageous at this stage to outline some of the criteria from the above two references regarding a choice of method from the five contained in (2.19). As far as ease of programming is concerned odd-even hopscotch appears to be the easiest and Peaceman-Rachford the most difficult of the four methods (we are excluding Crank-Nicolson for reasons indicated earlier), while the A.D.I. hopscotch is apparently slightly more difficult to program than the line hopscotch method. Speed of solution and storage requirements would seem to indicate the same hierarchy of the four methods. Arithmetic operation counts yielded by the above two authors from their programs, revealed that the odd-even hopscotch method was three to five times faster than any of the other methods. They also pointed out that the A.D.I. hopscotch and Peaceman-Rachford
processes required at least twice the storage space of the odd-even or line hopscotch methods.

The big advantage that Peaceman-Rachford has over the hopscotch methods is accuracy. This stems mainly from the comparative $O(\Delta t^3)$ and $O(\Delta t^3/h^2)$ terms in the local truncation errors of the respective methods, a result which will be proved in subsection 5.2 (vi). It will also be shown there that the line and A.D.I. hopscotch processes tend to be more accurate than the odd-even hopscotch method.

Summarising then, the odd-even hopscotch satisfies the situation where a quick solution with little effort and not too great accuracy is required. The line and A.D.I. hopscotch processes are less efficient (the latter more so than the former), but gain in accuracy. Lastly, the Peaceman-Rachford method gives high accuracy at the expense of development and running time.
CHAPTER 3

ELLIPTIC EQUATIONS IN TWO SPACE DIMENSIONS
3.1 Introduction

In this Chapter, we will be concerned with the application of the method of finite differences to obtain an approximate solution for the self-adjoint problem defined by,

\[ \frac{\partial}{\partial x} \left\{ P(x,y) \frac{\partial u}{\partial x} \right\} + \frac{\partial}{\partial y} \left\{ Q(x,y) \frac{\partial u}{\partial y} \right\} + \sigma(x,y)U(x,y)+\varepsilon(x,y) = 0 \]  

(3.1)

\((x,y) \in \mathbb{R}, \) and

\[ U(x,y) = \alpha(x,y), \quad (x,y) \in \mathbb{B}. \]  

(3.1)'

The region \( \mathbb{R} \) and its boundary \( \mathbb{B} \) are defined at the beginning of section 1.1. The functions \( P(x,y), Q(x,y), \) \( \alpha(x,y) \) satisfy in \( \mathbb{R} \times \mathbb{B} \) the conditions \( P(x,y) > 0, \) \( Q(x,y) > 0, \) \( \alpha(x,y) \geq 0. \) Because of these conditions, it follows from section 1.1, that equation (3.1) is elliptic. The purpose of this chapter, is to provide information for incorporation with or comparison with the new computational techniques (for solving certain linear systems of equations) developed in chapters 6 and 7.

As described in section 1.2, we superimpose a mesh over the region \( \mathbb{R}, \) and as before, unless otherwise noted, we assume a constant mesh size \( h \) in both coordinate directions.

Let \( (x_i,y_j) \) denote the grid point \((ih,jh). \) Then if \( (x_{i+1},y_j), (x_i,y_{j+1}), (x_{i},y_{j-1}) \) are also grid points contained in \( \mathbb{R} \times \mathbb{B}, \) \( (x_i,y_j) \) is called a regular mesh point; otherwise it is called irregular. Now using Taylor series, we can derive, in an analogous manner to equations (1.3)-(1.6), finite difference representations of the form

\[ -h^2 \frac{\partial^2}{\partial x^2} \left\{ P(x,y) \frac{\partial u}{\partial x} \right\} + P(x+h/2,y) \left[ u(x,y) - u(x+h,y) \right] \]

\[ + P(x-h/2,y) \left[ u(x,y) - u(x-h,y) \right]. \]

(3.2)
provided \((x,y)\) is a regular mesh point. With the same provision, a similar expression to (3.2) can be obtained to approximate \(-h^2 \partial^2 U(x,y)/\partial x^2\). If the mesh point \((x,y)\) is irregular, then Taylor series expansions of \(U(x,y)\) about such points may still be used to obtain approximations similar to (1.3)–(1.6) and also to (3.2).

We now consider a fixed labelling of the mesh points of \(R\), which we assume, for convenience, contains only regular mesh points. Scanning over the grid points with the five point difference approximation to (3.1), obtained from (3.2), in the order specified by this labelling, yields, in the manner described in section 1.2, a system of equations represented in matrix notation as

\[
Au = b.
\]

Here \(A\) is a real \(m \times n\) matrix (\(m\) being the number of unknown mesh points), whose diagonal entries are \((-h^2 \sigma(x,y)+P(x+h/2,y)+\nonumber P(x-h/2,y)+Q(x,y+h/2)+Q(x,y-h/2)),\) and whose off-diagonal elements consist of the quantities \(-P(x+h/2,y), -P(x-h/2,y), -Q(x,y+h/2)\) and \(-Q(x,y-h/2),\) which do not correspond to boundary points, or zeros. The vector \(u\) consists of the unknown approximate solutions \(u_{i,j}\), and \(b\) is a vector consisting of the known boundary values plus the quantities \(h^2 f(x,y)\).

The matrix \(A\) can be shown to have the following properties:

\[
\text{(a) } a_{i,i} > 0 \text{ and } a_{i,j} < 0 \text{ for all } i,j \text{ with } |i| \neq |j| \quad \text{(3.4)}
\]

\[
\text{(b) } A \text{ is irreducibly diagonally dominant.}
\]

The latter condition follows from Definition 1.5 and Theorem 1.1. It also follows from Theorem 1.5, that \(A\) is non-singular,
and hence system (3.3) has a unique solution. If all the interior mesh points in $R$ are regular, $A$ will be symmetric, and hence also positive definite (by Theorem 1.5, Corollary). However, if this is not the case, the above procedure will not in general result in a symmetric matrix $A$.

3.3 The model problem

To illustrate our future discussions we shall often consider the following Dirichlet problem. We wish to obtain an approximation to the function $U(x,y)$ which satisfies

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0 \quad \text{Laplace's equation}, \quad 0 < x, y < 1,$$

in the unit square $R$ with boundary $B$, subject to the boundary conditions

$$U(x,y) = \gamma(x,y), \quad x, y \in B.$$  \hspace{1cm} (3.5)

We shall call this the model problem. Note that this is the special case $PH\Omega\frac{1}{2}a_{50}$ in (3.1).

Neglecting the truncation error terms in (1.9) and equating $g_{i,j} = 0$, gives us the finite difference analogue of (3.5), namely

$$4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = 0.$$  \hspace{1cm} (3.6)

If we order (label) the internal mesh points row-wise (see Figure 5.1), we obtain a matrix problem like (3.3).

In general

$$A = \begin{bmatrix}
A_1 & C_1 \\
A_2 & \ddots & C_{m-1} \\
& \ddots & \ddots & C_m \\
& & \ddots & A_m \\
& & & B_m
\end{bmatrix}, \quad (3.8)$$

\hspace{1cm} \hspace{1cm} (\frac{\partial^2 U}{\partial x^2})

\hspace{1cm} \hspace{1cm} (\frac{\partial^2 U}{\partial y^2})

\hspace{1cm} \hspace{1cm} \text{FIGURE 5.1}
The $(m^2 \times 1)$ column vectors $u$ and $b$ are defined as before.

We notice that the matrix $A$ so obtained has properties (3.4).

3.3 Point-iterative methods

We now consider the task of computing the solution of a system such as that represented by the matrix equation (3.3).

As illustrated on page 7, the smaller we take $h$ (the mesh size), the better accuracy we obtain. The use of a small mesh size however, implies a large number of mesh points and consequently a large number of equations. The coefficient matrix $A$ is sparse (most of its elements are zeros), and there are no more than five non-zero elements in any given row for the finite difference approximations which are most often used, e.g. formula (3.7). For large linear systems with sparse matrices, iterative methods of successive approximation are usually preferable to direct methods such as Cramer's rule and Gaussian elimination. In the use of iterative methods, one assumes an arbitrary initial approximation to the solution and then successively modifies the approximation according to some prescribed rule. The approximation can, in general, be made arbitrarily close to the exact solution of the difference equations. We consider two classes of iterative methods for solving systems of equations. In the first, known as point
iterative methods, each component of successive approximations to the solution is modified by an explicit calculation; whereas, in block-iterative methods, several linear systems are solved (each of which is much smaller than the original system) at each stage of the calculation.

We now describe some point-iterative methods which are used to improve an approximate solution of the matrix equation (3.3). We first express the coefficient matrix $A$ in (3.3) as

$$A = D - L - U,$$

where $D$ is a positive diagonal matrix whose elements are the diagonal elements of $A$, and $L$ and $U$ are respectively lower and upper triangular matrices with null diagonals. Equation (3.3) then becomes

$$(D - L - U)u = b .$$

(3.10)

Since $D$ is a positive diagonal matrix, $D^{-1}$ exists, so we let

$$B = D^{-1} (L + U) \quad \text{and} \quad c = D^{-1} b ,$$

(3.11)

then (3.10) can be re-written as

$$u = Bu + c .$$

(3.12)

The Jacobi method is defined by

$$u^{(n+1)} = Bu^{(n)} + c , \quad n \geq 0 ,$$

(3.13)

where $u^{(n)}$ denotes the $n$th approximation to the solution.

Alternatively (3.13) can be written in the form

$$u^{(n+1)} = u^{(n)} + (c - D^{-1}Au^{(n)}) .$$

(3.14)

The method of Simultaneous Displacement is then defined by

$$u^{(n+1)} = u^{(n)} + a (c - D^{-1}Au^{(n)}) ,$$

(3.15)

where $a$ is a positive constant chosen to enhance convergence. In these two methods the order in which one solves for the components $u^{(n+1)}_1$ of $u^{(n+1)}$ is of no
consequence. In methods such as the Gauss-Seidel method, written in matrix notation as

$$u^{(n+1)} = L_u u^{(n+1)} + U_u u^{(n)} + \mathbf{c} \quad (3.16)$$

or as

$$u^{(n+1)} = (I-L) u^{(n)} + (I-L)^{-1} \mathbf{c} \quad (3.17)$$

where

$$L = D^{-1} L, \quad U = D^{-1} U (D+L+U) \quad (3.18)$$

we order the mesh points such that we use the latest estimate \( u_i^{(n+1)} \) of the components of \( u \) where available in determining a new estimate of a component of \( u^{(n+1)} \).

The next method which is derived from (3.16) is called the (point) successive-overrelaxation (S.O.R.) method. In matrix notation it is

$$u^{(n+1)} = (I-\omega) u^{(n)} + \omega (L_u u^{(n+1)} + U_u u^{(n)} + \mathbf{c}) \quad (3.19)$$

or

$$u^{(n+1)} = M_u u^{(n)} + \omega (I-M_u)^{-1} \mathbf{c} \quad (3.20)$$

where

$$M_u = (I-M_u)^{-1} ((I-\omega) I + \omega U) \quad (3.21)$$

Here \( \omega \) is a parameter known as a relaxation factor, the choice of which determines how rapidly the method converges.

For \( \omega = 1 \), the S.O.R. method reduces to the Gauss-Seidel method.

**Convergence of point-iterative methods**

All of the iterative methods described above, can be written in the form

$$u^{(n+1)} = M u^{(n)} + \mathbf{d} \quad (3.22)$$

where \( M \) is known as the iteration matrix (of the particular method), and \( \mathbf{d} \) is a column vector of constants.

**Definition 3.1**

The consistency condition (for an iterative process) is satisfied if, when the solution \( u \) of \( Au = b \) is substituted
for $u^{(n)}$ in (3.22), $u^{(n+1)}$ is also the solution $u$ of the system, i.e. once the solution is obtained, the iterative procedure makes no further modification of successive iterates.

**Definition 3.2**

An iterative method is said to converge if, for any given $b$, each component of the successive iterates $u^{(n)}$ tends toward the corresponding component of the solution vector $u$ for all initial vectors $u^{(o)}$.

**Theorem 3.1**

An iterative method which can be expressed in the form of equation (3.22) converges if and only if $\rho(A)<1$.

**Proof**

Defining $e^{(n)} = u^{(n)} - u$ (the error vector after $n$ iterations) and assuming the method consistent, i.e.

$$ u = H u + d , $$

(3.23)

we have from (3.22) and (3.23),

$$ e^{(n+1)} = H e^{(n)} . $$

(3.24)

Therefore, we have

$$ e^{(n)} = H^n e^{(o)} , $$

(3.25)

where $e^{(o)}$ is the error vector associated with the initial vector $u^{(o)}$. We therefore require the condition(s) under which $e^{(n)} \to 0$ as $n \to \infty$, i.e. $e_i^{(n)} \to 0$ for $i=1,2,\ldots,N$.

If $e^{(o)}$ is bounded, i.e. $|e_i^{(o)}| < k, (i=1,2,\ldots,N)$, for some constant $k$, then from (3.25) it follows that $e^{(n)} \to 0$ (as $n \to \infty$) if and only if $H^2 \to 0$ (the null matrix) as $n \to \infty$.

By Theorem 1.7, this will be true if and only if $\rho(A)<1$. This completes the proof.

By assuming the conditions (3.4) on the matrix $A$, it can be shown (e.g., see Young and Frank (1963) pp.13-14)
that the Jacobi and Gauss-Seidel methods ((3.13) and (3.16)) converge. A convergence theorem for the point S.O.R. method (3.19) whose proof can be found in Varga (1962) is as follows.

**Theorem 3.2**

If A is symmetric and $a_{1,i}>0$, $i=1,2,...,N$, then $\sigma(M)<1$ if and only if A is positive definite and $0<\omega<2$.

**Rate of Convergence**

In order to evaluate the effectiveness of an iterative method, one must consider both the work required per iteration and the number of iterations required for convergence. For practical purposes, we shall say that an iterative method has converged when the norm of the error vector $\varepsilon^{(n)}$ is less than some predetermined factor $\epsilon$ (say $10^{-6}$) of the norm of the original error vector $\varepsilon^{(0)}$.

From (3.25) we have

$$||\varepsilon^{(n)}|| = ||x^n - e^{(0)}|| \leq ||x^n|| - ||e^{(0)}||.$$  \hspace{1cm} (3.26)

Then if $e^{(0)}$ is not the null vector,

$$||\varepsilon^{(n)}||/||e^{(0)}|| \leq ||x^n||,$$  \hspace{1cm} (3.27)

and we require

$$||\varepsilon^{(n)}|| \leq \epsilon||e^{(0)}||.$$  \hspace{1cm} (3.28)

where $||.||$ denotes the spectral (or $L_2$) norm. By Theorem 1.8, we know that $||x^n||<0$ as $n \to \infty$ if $\sigma(M)<1$. Hence we can satisfy (3.28) if we choose n large enough so that

$$||x^n|| \leq \epsilon$$  \hspace{1cm} (3.29)

For all $n$ sufficiently large that $||x^n||<1$, is equivalent to

$$n \geq -log\epsilon/(\frac{1}{\eta} \log||x^n||).$$  \hspace{1cm} (3.30)
**Definition 3.3**

We now define

\[ R_n(h) = -(1/n) \log ||h^n||, \]  

(3.31)

as the average rate of convergence of the iterative method associated by (3.22) with the matrix \( M \).

**Definition 3.4**

We also define the asymptotic average rate of convergence, or asymptotic rate of convergence by

\[ R(h) = \lim_{n \to \infty} R_n(h) = -\log \rho, \]  

(3.32)

the latter equality holding since \( \rho(h) = \lim_{n \to \infty} (||h^n||)^{1/n}, \) a result proved by Young (1971) p. 87. We shall usually refer to \( R(h) \) as the rate of convergence.

**Comparison of rates of convergence**

Direct comparisons of the Jacobi, Gauss-Seidel, and S.O.R. methods can be made when the matrix \( A \) has certain properties. D.H. Young has developed the theory of the S.O.R. method subject to the conditions that the matrix \( A \) has Property (A) and is consistently ordered. We shall not discuss these concepts here, since they are covered in many books, particularly Young (1971). We will assume that these two properties hold in any future discussion on the S.O.R. method (unless otherwise stated).

We now summarise the argument, and for more detail the reader is referred to the book just mentioned [or Varga (1962)]. If \( \mu \) is an eigenvalue of the Jacobi matrix \( B \) [see (3.13)], \( \omega(\text{real}) > 0 \) and \( \lambda \) is an eigenvalue of \( M_0 \) [see (3.21)] then

\[ \frac{\lambda \omega - 1}{\omega^1} = \mu. \]  

(3.33)
It can be shown that

$$\overline{\omega} = \frac{2}{1+\sqrt{1-\frac{4}{\omega^2}}} \left[ \overline{\rho}(\omega) \right]$$  \hspace{1cm} (3.34)$$

is the optimum value of $\omega$ in the sense that $\rho(\overline{\omega})$ is a minimum. Using this value of $\omega$ it can also be shown that

$$\rho(\overline{\omega}) = \frac{1-\sqrt{1-\frac{4}{\omega^2}}}{1+\sqrt{1-\frac{4}{\omega^2}}} = \overline{\omega} - 1.$$  \hspace{1cm} (3.35)$$

Remembering that $\omega=1$ gives the Gauss-Seidel method, from

$$\rho(1) = [\overline{\rho}(1)]^2$$  \hspace{1cm} (3.36)$$

and so from (3.32),

$$R(1) = 2R(1).$$  \hspace{1cm} (3.37)$$

Thus the rate of convergence of the Gauss-Seidel method is just twice that of the Jacobi method.

By the use of (3.35) it can be verified that, asymptotically as $\rho(1)$ tends to one,

$$R(1) \approx 2\sqrt{R(1)}.$$  \hspace{1cm} (3.38)$$

We now investigate the rate of convergence of the method of Simultaneous Displacements given by (3.15). The optimum value $\overline{\alpha}$ of $\alpha$ in the sense that $\rho((1-\alpha D^{-1})A)$ is minimised, is given by

$$\overline{\alpha} = \frac{2}{\omega + b};$$  \hspace{1cm} (3.39)$$

where we have assumed

$$0 < a \lambda_s < b < \infty \quad (\lambda_s \text{ is an eigenvalue of } D^{-1}A).$$  \hspace{1cm} (3.40)$$

(see Forsythe and Wasow (1960), p.225). With this choice of $\alpha$, the spectral radius of $(1-\alpha D^{-1})A$ is

$$\max \left|1-\alpha \lambda \right| \leq \frac{b-a}{b+a} = \frac{P-1}{P+1},$$  \hspace{1cm} (3.41)$$

where $P = \max \left|\lambda \right|$. 

Definition 3.5

\( P = \frac{b}{a} \) is the \textit{P-condition number} of \( D^{-1}A \) and is defined as the ratio of the maximum eigenvalue to the minimum eigenvalue of a positive definite matrix.

From (3.41), the (asymptotic) rate of convergence is given by

\[
\hat{\rho}(\hat{H}) = \frac{2}{P}, \quad P \to \infty.
\]  

(3.42)

The iterative methods we have discussed so far of the form (3.22) are \textit{linear stationary iterative methods of first degree}. A general linear stationary \textit{second degree} method [see Young (1971), Chapter 16] is

\[
\begin{pmatrix}
\hat{u}^{(n+1)} \\
\hat{u}^{(n)}
\end{pmatrix} = \begin{pmatrix}
0 & 1 \\
H & G
\end{pmatrix} \begin{pmatrix}
\hat{u}^{(n-1)} \\
\hat{u}^{(n)}
\end{pmatrix} + \begin{pmatrix}
0 \\
G_k
\end{pmatrix}.
\]  

(3.43)

(This is equivalent to splitting the matrix \( D^{-1}A \) as

\( D^{-1}A = I - H - G \). Now (3.43) can be expressed as

\[
\begin{pmatrix}
\hat{u}^{(n+1)} \\
\hat{u}^{(n)}
\end{pmatrix} = \begin{pmatrix}
0 & 1 \\
H & G
\end{pmatrix} \begin{pmatrix}
\hat{u}^{(n-1)} \\
\hat{u}^{(n)}
\end{pmatrix} + \begin{pmatrix}
0 \\
0
\end{pmatrix}.
\]  

(3.44)

Hence a necessary and sufficient condition for the convergence of (3.43), for all \( \hat{u}^{(0)} \) and \( \hat{u}^{(1)} \), is

\[
\hat{\rho}(\hat{H}) < 1,
\]  

(3.45)

where

\[
\hat{H} = \begin{pmatrix}
0 & 1 \\
H & G
\end{pmatrix}.
\]  

(3.45)'

A special case of (3.46) for the solution of (3.3) is \textit{Richardson's second order (degree) method},

\[
\hat{u}^{(n+1)} = \hat{u}^{(n)} + a \hat{u}^{(n)} - \frac{1}{\hat{u}^{(n)} - b \hat{u}^{(n-1)}}.
\]  

(3.46)

The theoretical justification for the choice of \( a \) and \( b \) has been given by Frankel (1950). The optimum values are

\[
a = -\left(\frac{2}{a+b}\right)^2 \quad \text{and} \quad b = -\left(\frac{2}{a+b}\right)^2.
\]  

(3.47)

[where \( a, b \) are defined by (3.40)].
\[ \rho(\lambda) = \sqrt{\lambda} , \]  
(3.48)

and the rate of convergence is

\[ R = \frac{2}{\sqrt{F}} \quad \text{as } F \to \infty . \]  
(3.49)

For the model problem it is easily verified (e.g. Forsythe and Wasow (1960), p.226) that the method of simultaneous

displacements (3.15) is equivalent to the Jacobi method of (3.13).

For the model problem of section 3.2, we have from (3.3) and (3.40)

\[ \alpha = 2\sin^2 \left( \frac{\pi}{2(n+1)} \right) \cos^2 \left( \frac{\pi}{2(n+1)} \right) + \sin^2 \left( \frac{\pi}{2(n+1)} \right) \]  
(3.50)

where \( n \) is the order of the matrix \( A \). This gives us from (3.13),

\[ \overline{w} = \rho(0) = \cos \left( \frac{\pi}{n+1} \right) . \]  
(3.51)

Hence from (3.34),

\[ \overline{w} = \frac{2}{1 + \sin \left( \frac{\pi}{n+1} \right)} = \frac{2}{1 + \sin (\pi n)} \]  
(3.52)

Thus from (3.36), (3.37), (3.38) and (3.51) we obtain

\[
\begin{array}{l}
(i) \quad R(B) = \frac{x^2}{2(n+1)^2} \\
(ii) \quad R(C) = \frac{x^2}{(m+1)^2} \\
(iii) \quad R(D) = \frac{2x}{(m+1)}
\end{array}
\]  
(3.53)

for the model problem. [See also, for example, Varga (1962), p.201-205].
3.6 **Block S.O.R. iterative methods**

In point iterative methods, each component of \( u^{(n)} \) is determined *explicitly*, i.e. can be determined by itself using already computed approximate values of the other unknowns. With \( k \)-line iterative methods one improves the values of the approximate solution simultaneously on \( k \) horizontal or vertical lines of points in \( R \). Such methods are called *implicit methods* since we solve a linear system for a whole subset of components at once. The advantage of implicit over explicit methods is that the convergence rate of the former may be appreciably greater, at the cost of some complication in the method. The system of linear equations will still be of the form (3.3) with (3.8), where for \( k \)-line methods the "largest" \( B_i \) (\( i=1,2,\ldots,n \)) will have a band-width of \( 2k+1 \). Parter (1961a) tells us, that unless one finds a really good way to invert such matrices, one should use as small a "\( k \)" as possible, i.e. \( k=1 \) or 2 for second order equations. We consider below the case when \( k=1 \).

In an analogous manner to point S.O.R. theoretical results for 1-line S.O.R. (S.L.O.R.) have been achieved by assuming that the coefficient matrix \( A \) is *block consistently ordered* and possesses property \( A^n \) (or *block property A*). Both these properties are discussed in Arns, Gates and Zondek, (1956). For the S.L.O.R. method with *Dirichlet boundary conditions* given on \( B \), the \( B_i \) in (3.8)' will be tri-diagonal, and it will be necessary to solve subsystems of equations of the form

\[
A_i^* u_i^{(n+1)} + B_i u_i^{(n+1)} + C_i u_{i+1}^{(n)} = d_i, \quad i=1,2,\ldots,n \quad (A_i^* C_n = 0) \quad (3.54)
\]
or
\[ \hat{u}_{i}^{(n+1)} = k_{i} \cdot \hat{u}_{i}, \quad i = 1, 2, \ldots, n \]  
(3.55)
where \( \hat{u}_{i} \) is the column vector of values along the \( i \)th line
and \( k_{i} \) (lasts) are known since \( \hat{u}_{i}^{(n+1)} \) are calculated before
\( \hat{u}_{i}^{(n+1)} \). A well-known algorithm for solving system (3.55) is
given in Chapter 4.

The entire theory of point overrelaxation is also
applicable to line overrelaxation (e.g. see Young (1971)).

If \( \bar{u} \) is the spectral radius of the Jacobi line method, then
the optimum value of \( \omega \) for S.L.O.R. is given by

\[ \bar{u}_{L} = 1 + \left[ \frac{\bar{u}_{L}}{1+\sqrt{1-\bar{u}_{L}}} \right]^{2} \]  
(3.56)

From Varga (1962) pp. 204-205, we have for the model
problem,

\[ R(H_{m}^{(\text{line})}) = \frac{2\sqrt{2}}{m+1}, \quad m \to \infty \]  
(3.57)

where \( H_{m}^{(\text{line})} \) denotes the iteration matrix of the S.L.O.R.
method, so that from (3.53) (iii),

\[ \frac{R(H_{m}^{(\text{line})})}{R(H_{m}^{(\text{SOR})})} = \sqrt{2}, \quad m \to \infty \]  
(3.58)

Parter (1961a) has shown for the model problem (as well as
a similar expression for more general problems that

\[ R(H_{m}^{(k \text{ line})}) = 2\sqrt{2k}, \quad m \to \infty \]  
(3.59)

This shows how the ratio of the asymptotic convergence
rates of different line S.O.R. iterative methods behaves as
a function of \( k \). It is then evident that practically,
little is gained for "large" values of \( k \) (e.g. \( k = 10 \)) since
arithmetic requirements now also increase linearly with \( k \),
as pointed out by Parter.
As shown in Cuthill and Varga (1959) and Varga (1960), it is possible in many cases to perform both single and double-line overrelaxation in approximately the same number of arithmetic operations per mesh point as required by the point S.O.R. method. Therefore, the increase in rates of convergence of line-iterative methods over point-iterative methods result in corresponding decreases in total computational effort.

3.6 Alternating-Direction Implicit methods

Alternating-Direction Implicit (A.D.I.) methods are somewhat similar to single-line iterative methods with alternating directions. There are two basic A.D.I. methods, one presented by Peaceman and Rachford (1955) and the other by Douglas and Rachford (1956). In this Chapter we will consider only the former, which is known as the Peaceman-Rachford (P-R) method.

Consider the self-adjoint partial differential equation of section 3.1,

\[
\frac{\partial}{\partial x}(P(x,y) \frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(Q(x,y) \frac{\partial u}{\partial y}) + a(x,y)u(x,y) + f(x,y) = 0, \tag{3.60}
\]

where \( P, Q > 0 \) and \( a \leq 0 \) for all \((x,y) \in \mathbb{R}\). Let

\[
lu(x,y) = -a(x,y)u(x+h,y) + 2b(x,y)u(x,y) - c(x,y)u(x-h,y), \tag{3.61}
\]

\[
vu(x,y) = -a(x,y)u(x,y+h) + 2b(x,y)u(x,y) - \gamma(x,y)u(x,y-h), \tag{3.62}
\]

where

\[
a = P(x+h/2,y), \quad c = P(x-h/2,y), \quad 2b = a+c \bigg\}
\]

\[
a = Q(x,y+h/2), \quad \gamma = Q(x,y-h/2), \quad 2b = a+\gamma \bigg\}
\]

We see from equation (3.2), that \( lu(x,y), vu(x,y) \) are the central difference approximations of \(-h^2 \partial^2 u/\partial x^2(P(x,y)\partial u/\partial x)\), \(-h^2 \partial^2 u/\partial y^2(Q(x,y)\partial u/\partial y)\) for a square mesh of mesh size \( h \).

Using (3.61) and (3.62), we define a system of linear algebraic equations analogous to equation (3.3), which may be
represented in matrix notation as

\[(H + V + \Sigma)\mathbf{u} = \mathbf{b}\]  \hspace{1cm} (3.64)

where \(\Sigma\) is a non-negative diagonal matrix whose \(i\)th diagonal entry associated with the mesh point \((x_i, y_i)\) is \(-h^2\sigma(x_i, y_i)\). The column matrix \(\mathbf{b}\) consists of the terms \(h^2 f(x_i, y_i)\) plus the terms associated with points on the boundary, and \(\mathbf{u}\) is an approximate solution to (3.60).

Providing the previously defined region \(R\) contains only regular mesh points, we can make the following assertions, which can be verified using the theorems of sections 1.3, 1.4 and 1.6.

(a) \(A = H + V + \Sigma\) is an irreducible Stieltjes matrix, i.e., \(A\) is a real, symmetric, and positive definite irreducible matrix with non-positive off-diagonal entries.

(b) \(H\) and \(V\) are real, symmetric, diagonally dominant matrices with positive diagonal entries and non-positive off-diagonal entries.

(c) \(\Sigma\) is a non-negative diagonal matrix.

By ordering the mesh points by rows, \(H\) can be made tri-diagonal, and by ordering by columns, \(V\) can be made tri-diagonal. Hence using this, conditions (3.65) and Dirichlet boundary conditions, we can state that there exist \(n \times n\) permutation matrices \(P_1\) and \(P_2\) such that \(P_1 H P_1^T\) and \(P_2 V P_2^T\) are the direct sum of tri-diagonal irreducible Stieltjes matrices.

Equation (3.64) is equivalent to each of the two matrix equations

\[\begin{align*}
(rI + H + \Sigma/2)\mathbf{u} &= (rI - V - \Sigma/2)\mathbf{u} + \mathbf{b} \\
(rI + V + \Sigma/2)\mathbf{u} &= (rI - H - \Sigma/2)\mathbf{u} + \mathbf{b}
\end{align*}\] \hspace{1cm} (3.66)
for any positive scalar \( r \). If we let

\[
\mathbf{n}_1 = \mathbf{n} \cdot \frac{1}{\sqrt{2}}, \quad \mathbf{v}_1 = \mathbf{v} \cdot \frac{1}{\sqrt{2}},
\]

(3.67)

the Peaceman-Rachford alternating direction implicit iterative method is defined by

\[
\begin{align*}
(r_k \mathbf{I} + \mathbf{h}_1) \mathbf{u}^{(k+1)} &= (r_k \mathbf{I} - \mathbf{v}_1) \mathbf{u}^{(k)} + \mathbf{b}, \\
(r_k \mathbf{I} + \mathbf{v}_1) \mathbf{u}^{(k+1)} &= (r_k \mathbf{I} - \mathbf{h}_1) \mathbf{u}^{(k)} + \mathbf{b}.
\end{align*}
\]

(3.68)

The \( r_k \)'s are positive acceleration parameters chosen to make the process converge rapidly, and \( \mathbf{u}^{(0)} \) is an arbitrary initial vector approximation of the unique solution of (3.64).

Since both \( \mathbf{h}_1 \) and \( \mathbf{v}_1 \) are, by suitable rearrangement of their rows and corresponding columns, tri-diagonal matrices, the iterative method (3.68) can be carried out directly using Algorithm 4.2, Chapter 4. The vector \( \mathbf{u}^{(k+1)} \) is treated as an auxiliary vector which is discarded as soon as it has been used in the calculation of \( \mathbf{u}^{(k+1)} \).

We combine the two equations of (3.68) into the form

\[
\mathbf{u}^{(k+1)} = T_{k+1} \cdot \mathbf{u}^{(k)} + H_{k+1} \mathbf{b}, \quad n > 0
\]

(3.69)

where

\[
T_k = (r_k \mathbf{I} + \mathbf{v}_1)^{-1} (r_k \mathbf{I} - \mathbf{h}_1) (r_k \mathbf{I} + \mathbf{v}_1)^{-1} (r_k \mathbf{I} - \mathbf{v}_1),
\]

(3.70)

and

\[
H_k \mathbf{b} = (r_k \mathbf{I} + \mathbf{v}_1)^{-1} ((r_k \mathbf{I} - \mathbf{h}_1) (r_k \mathbf{I} + \mathbf{v}_1)^{-1} + \mathbf{b}).
\]

(3.71)

It will be noticed that (3.69) is of the same form as (3.22).

We first consider the case, where all the constants \( r_j \)

are equal to the fixed constant \( r > 0 \). The following convergence theorems will be stated here and proved in section 6.5.

**Theorem 3.3**

Let \( \mathbf{h}_1 \) and \( \mathbf{v}_1 \) be full symmetric non-negative definite matrices, where at least one of the matrices \( \mathbf{h}_1 \) and \( \mathbf{v}_1 \) is
positive definite. Then, for any \( r > 0 \), the Peaceman-Rachford matrix \( T_x \) of (3.70) is convergent.

Returning our attention to the model problem (section 3.2), if the vector \( a^{(k,t)} \) is defined so that its component for the \( i \)th column and \( j \)th row of the mesh is \([\text{for } b=1/(m+1)]\):

\[
a_{i,j}^{(k,t)} = a_{k,t} \sin \left( \frac{k \pi}{m+1} \right) \sin \left( \frac{j \pi}{m+1} \right) l_{i,j}, k \in \mathbb{Z}, t, \quad \text{for all } 1 \leq k, \leq m.
\]  

(3.72)

then, using the definitions of the matrices \( R \) and \( V \) of (3.61) and (3.62) applied to Laplace's equation (vis. \( a = b = c = \beta = 0 = 1 \)), it follows that:

\[
R_{a{k,t}}^{(k,t)} = 4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right) \frac{a(k,t)}{a_{k,t}}, \quad \text{for all } 1 \leq k \leq m.
\]

(3.73)

For the model problem \( \sum = 0 \), hence from (3.73) and (3.70) we have:

\[
T_x a(k,t) = \begin{pmatrix}
4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right) & 4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right) \\
4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right) & 4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right)
\end{pmatrix}
\]

\[
a(k,t), \quad 1 \leq k \leq m.
\]

(3.74)

We therefore conclude that:

\[
\rho(T_x) = \left\{ \max_{1 \leq k \leq m} \frac{4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right)}{4 \sin^2 \left( \frac{\pi k}{2(m+1)} \right)} \right\}^2.
\]

(3.75)

Consider the simple function:

\[
g(x;r) = \frac{r-x}{r+x}, \quad r > 0.
\]

(3.76)

where \( 0 < x_1 \leq x_2 \). The derivative of \( g(x;r) \) with respect to \( x \) is negative for all \( x > 0 \), so that
\[
\max_{x_1, x_2} |g(x; r)| = \max \left\{ \frac{r-x_1}{r+x_1} \cdot \frac{r-x_2}{r+x_2} \right\}. \quad (3.77)
\]

From this, it is easily shown that

\[
\max_{x_1, x_2} |g(x; r)| = \begin{cases} \frac{x_2-r}{x_2+r} & 0 < x_2 \sqrt{x_1 x_2} \\ \frac{r-x_1}{r+x_1} & x_2 \sqrt{x_1 x_2} \end{cases}. \quad (3.78)
\]

from which it follows that

\[
\min \{ \max_{r>0} g(x; r) \} = g(x_1; x_2; x) = \frac{1-(x_1/x_2)^{1}}{1+(x_1/x_2)^{1}}.
\]

(3.79)

Setting \( x_1 = 4 \sin^2(\pi/2(m+1)) \) and \( x_2 = 4 \cos^2(\pi/2(m+1)) \), we have \( \hat{r} = x_2 = 2 \sin(\pi/(m+1)) \) and from (3.75)

\[
\min_{r>0} \rho(\hat{T}_r) = \rho(\hat{T}_T) = \frac{1-\sin(\pi/(m+1))}{1+\sin(\pi/(m+1))}.
\]

(3.80)

Substituting \( \hat{u} = \cos(\pi/(m+1)) \) = the spectral radius of the point Jacobi matrix, we see that

\[
\min_{u} \rho(\hat{T}_u) = \rho(\hat{T}_T) = \frac{1-\sin(\pi/(m+1))}{1+\sin(\pi/(m+1))} = \min_{r>0} \rho(\hat{T}_r) = \rho(\hat{T}_T).
\]

(3.81)

Hence, as optimised one-parameter iterative methods, the

Peaceman-Rachford iterative method and the point S.O.R.
method have identical asymptotic rates of convergence for all \( b>0 \) for the model problem.

The Commutative Case

The rate of convergence of the A.D.I. methods can be
appreciably increased by the use of several iteration parameters.
In this case the parameters are used successively in a cyclic order. However, the theory of convergence and of the selection of good iteration parameters when more than one cycle of alternating directions is used has not been developed for arbitrary self-adjoint elliptic partial differential equations. For those cases which give rise to matrices $H$, $V$ and $J$ which are pairwise commutative, a satisfactory theory does exist. This occurs when $J = eta I$ ($\beta$ a non-negative constant) and when the conditions of the following theorem of Frobenius (see, for instance, Varga (1962) pp. 220–221) are satisfied.

**Theorem 3.4**

The matrices $H$ and $V$ have a common set of eigenvectors if and only if $HV = VH$ and $H$ and $V$ are similar to diagonal matrices.

(Further, if $H$ and $V$ are symmetric then there exists a common set of orthonormal eigenvectors).

The iteration matrix of the multi-parameter Peaceman-Rachford method is given by

$$
\prod_{j=1}^{N} T_j = \prod_{j=1}^{N} (r_j I + V_j)^{-1} (r_j I + H_j) (r_j I + H_j)^{-1} (r_j I - V_j). 
$$

(3.82)

If Theorem 3.4 is satisfied, then the eigenvalues of $\prod_{j=1}^{N} T_j$ are all of the form

$$
\lambda(\mu, \nu) = \prod_{j=1}^{N} \frac{(\mu - r_j)(\nu - r_j)}{(\mu + r_j)(\nu + r_j)}. 
$$

(3.83)

where $\mu, \nu$ satisfy

$$
H_1 \mu = \nu, \quad V_1 \nu = \nu. 
$$

(3.84)

Let us assume we can estimate the spectral bounds $a$ and $b$ for the eigenvalues $\mu_1$ and $\nu_1$ of the positive definite matrices $H_1$ and $V_1$, i.e.

$$
a \leq \mu_1 \leq b, \quad \nu_1 \leq b \implies \mu_1 \leq b. 
$$

(3.85)
Then

\[
\max_{1\leq n} \left\{ \frac{\| T^{-1} x_j \|}{\| T_j \|} \cdot \frac{\| T^{-1} y \|}{\| T_j \|} \right\}
\]

is exactly equal to \( \left( \max_{1\leq n} \left\{ \frac{\| T^{-1} x_j \|}{\| T_j \|} \right\} \right)^2 \times \left( \max_{1\leq n} \| T_j \| \right)^2 \times \left( \max_{1\leq n} \| T_j \| \right)^2 \times \left( \max_{1\leq n} \left\{ \frac{\| T^{-1} y \|}{\| T_j \|} \right\} \right)^2 \) (3.86)

where

\[
g_M(x_j, y) = \prod_{j=1}^{m} \left( \frac{T_i \cdot x}{T_j \cdot y} \right)
\]

If \( S_M \) is the set of all functions \( g_M(x_j, y) \), where \( r_1, r_2, \ldots, r_M \)
are positive or non-negative real numbers, let

\[
d_M[\alpha, \beta] = \min_{g_M \in S_M} \{ \max_{0 < x < \beta} g_M(x, y) \} \quad (3.87)
\]

Wachpress's method (1.92) gives an exact determination of the

\( \bar{r}_j \) which solve the min-max problem (3.87) for \( n=2^k \), \( k \geq 0 \); also
there are two approximate methods given by Varga (1962) pp. 226-229.

With Wachpress's method for \( n=2^k \), (3.87) is satisfied when

\[
d_{2^k}[\alpha, \beta] = \frac{\sqrt{\alpha - \alpha_k}}{\sqrt{\alpha_k + \beta}} \quad (3.88)
\]

where \( \alpha_k, \beta_k \) are given by the sequence

\[
\begin{align*}
\alpha_0 &= \alpha, & \beta_0 &= \beta, \\
\alpha_{i+1} &= \sqrt{\alpha_i \beta_i}, & \beta_{i+1} &= \frac{\alpha_i + \beta_i}{2}, & i \geq 0.
\end{align*}
\]

(3.89)

It then follows from equations (3.86) and (3.87) that

\[
\rho \left\{ \prod_{j=1}^{M} T_j \right\} \leq \left( d_M[\alpha, \beta] \right)^2 \quad (3.90)
\]

Hence, the average spectral radius \( \rho \left( \prod_{j=1}^{M} T_j \right) \) is

\[
\geq \left( d_M[\alpha, \beta] \right)^{2/M} \quad (3.91)
\]
For the model problem, from (3.73) and (3.85)

\[ a = 4\sin^2 \frac{y}{2(m+1)}, \quad b = 4\cos^2 \frac{y}{2(m+1)} \quad (3.92) \]

Hence if we take two parameters alternately i.e. \( k = 1 \) in (3.88) we obtain

\[ d_2[a, b] = \frac{1 - \sqrt{\sin[\pi/(m+1)]}}{1 + \sqrt{\sin[\pi/(m+1)]}} = 1 - 2\sqrt{\frac{\pi}{m+1}} \quad (3.93) \]

which from (3.91) is an upper bound for the average spectral radius. Hence the average rate of convergence satisfies

\[ R \left[ \sum_{j=1}^{H} \bar{T}_{\pi} \right] = 2\sqrt{\frac{\pi}{m+1}} \quad (3.94) \]

Comparing this with (3.53), we see that for the model problem, the A.D.I. (F-R) method with only two parameters, is considerably faster (asymptotically) than the point S.O.R. method.

For general \( k \) we have an average rate of convergence:

\[ R \left[ \sum_{j=1}^{H} \bar{T}_{\pi} \right] = \frac{-2\log(d_2[a, b])}{H} \quad (3.95) \]

where \( H = 2^k \). This follows from (3.91).

The non-commutative case

In order to develop the theoretical results above, it was essential in view of Theorem 3.4, that \( H_1 \) and \( V_1 \) commuted. Birkhoff, et.al. (1962), showed that for a rectangular grid, a necessary and sufficient condition for commutativity is that the region \( R \) be rectangular, and that the differential equation have the form

\[ KE_2(u)Y_1(y)U(x, y) - F_1(y) \frac{\partial}{\partial x} (E_1(x) \frac{\partial U}{\partial x}) - E_2(x) \frac{\partial}{\partial y} (F_2(y) \frac{\partial U}{\partial y}) = S(x, y) \]

(3.96)
where the functions $E_1(x), F_1(y), E_2(x)$ are positive in $E$, and $\kappa$ is a non-negative constant.

Hence, for non-rectangular regions of problem solution and/or p.d.e.'s not satisfying (3.96), we have $H_1 V_1^* V_1 H_1$, viz. the "non-commutative case". In Section 6.7, we state two theorems which with certain restrictions on $H_1$ and $V_1$ on the number of acceleration parameters and on the order in which these parameters are used in each cycle, guarantee convergence in the non-commutative case. However, as yet there is no useful technique in the non-commutative case of explaining fairly rapid convergence in many situations, an example of which is described in Section 6.7.

The Huchprase and Habetler (1980) variant of the F-R iterative method

The basic idea of this technique, is to create a "commutative case" out of a "non-commutative" one. Their variant of the F-R (A.D.I.) iterative method is defined by

\[
(r_{n+1} D + H_1) u^{(n+1)} = (r_{n+1} D - V_1) u^{(n)} + b, \quad n \geq 0, \tag{3.97}
\]

and

\[
(r_{n+1} D + V_1) u^{(n+1)} = (r_{n+1} D - H_1) u^{(n)} + b, \quad n \geq 0, \tag{3.98}
\]

where $D$ is a positive diagonal matrix replacing $I$ in (3.68).

Defining \( D \hat{H}_1 \hat{V}_1 \), this variant takes the more familiar form

\[
(r_{n+1} I + \hat{H}_1) v^{(n+1)} = (r_{n+1} I - \hat{V}_1) v^{(n)} + \hat{b}, \quad n \geq 0, \tag{3.99}
\]

and

\[
(r_{n+1} I + \hat{V}_1) v^{(n+1)} = (r_{n+1} I - \hat{H}_1) v^{(n)} + \hat{b}, \quad n \geq 0, \tag{3.100}
\]

where

\[
\hat{H}_1 = D^{-1} H_1 D^{-1}, \quad \hat{V}_1 = D^{-1} V_1 D^{-1}. \tag{3.101}
\]
In order to apply the theory for the commutative case to (3.99) and (3.100), we now ask if

$$\tilde{H}_1 \tilde{V}_1 = \tilde{V}_1 \tilde{H}_1,$$

or equivalently, using (3.101), if

$$H_1 D^{-1} V_1 = V_1 D^{-1} H_1.$$  \hspace{1cm} (3.102) \hspace{1cm} (3.103)

In effect, as the diagonal entries of the matrix $D$ are parameters that are to be chosen, it is clear that there are cases when $H_1$ and $V_1$ do not commute, but there exists a positive diagonal matrix $D$ for which (3.103) is satisfied. Such a case is described in Section 6.5. It follows from Varga (1962), Lemma 7.2, however, that such extensions of the commutative theory are still restricted to rectangular domains.
CHAPTER 4

ALGORITHMIC SOLUTIONS

OF CERTAIN LINEAR SYSTEMS
4.1 Introductory Remarks

In Chapters 5 to 7, it will be necessary to solve certain special linear systems of equations. It has already come to our attention in Chapters 2 and 3, that tri-diagonal systems of equations occur very frequently in connection with the finite difference methods outlined in these Chapters.
Examples of such methods are the Peaceman-Rachford (A.D.I.) and line hopscotch procedures in Chapter 2, and the F-R (A.D.I.) and S.L.O.R. iterative processes in Chapter 3. Our second well known algorithm is an efficient procedure for solving such systems. We shall also describe other important algorithms for solving periodic tri-diagonal and quin-diagonal systems, which crop up in connection with new developments (and old) in Chapters 5 to 7.

We cannot of course, mention all the algorithms in use at present for solving the above systems, but will endeavour to present those which appear to be among the most efficient for solving the particular system in question. Arithmetic operation counts will be given with the last three algorithms, since they occur in conjunction with numerical methods whose efficiency we wish to compare.

We will assume all the linear systems of equations are written in matrix notation as

$$Au = k,$$  \hspace{1cm} (4.1)

where \( k \) will be some given \( n \times 1 \) column vector, \( A \) is an \( n \times n \) non-singular matrix which will be specified in each case, and \( u \) is the required solution vector.

Our first algorithm is due to Busbee et al. (1970).
Algorithm 4.1 (Matrix decomposition)

Consider system (4.1) where

\[
A = \begin{bmatrix}
B & C \\
C & B & C \\
& C & B \\
\end{bmatrix}
\]  

where \( B, C \) are \((p \times p)\) symmetric matrices, and we assume that

\[ BC = CB. \]  

From Theorem 3.4, it follows that there exists an orthogonal matrix \( Q \) (i.e. \( Q^T = Q^{-1} \)) such that

\[
Q^T EQ = A, \quad Q^T CQ = \Omega
\]

and \( A \) and \( \Omega \) are real diagonal matrices. The matrix \( Q \) is the fundamental matrix (set of eigenvectors) of \( B \) and \( C \) and \( A \) and \( \Omega \) are the diagonal matrices of eigenvalues of \( B \) and \( C \) respectively.

If we write

\[
u_j = \begin{bmatrix} u_{1j} \\ u_{2j} \\ \vdots \\ u_{pj} \end{bmatrix}, \quad k_j = \begin{bmatrix} k_{1j} \\ k_{2j} \\ \vdots \\ k_{pj} \end{bmatrix}
\]

\[
u_j = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix}, \quad k = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_q \end{bmatrix}
\]

then

\[ (4.5) \]

\[ (4.6) \]
System (4.1) with (4.2), may be written

\[ Bu_1 + Cu_2 = k_1, \]  
\[ Cu_{j-1} + Bu_j + Cu_{j+1} = k_j, \quad j=2,3,\ldots,q-1, \]  
\[ Cu_{q-1} + Bu_q = k_q. \]

Using (4.4), this becomes

\[ A\tilde{u}_1 + \tilde{u}_2 = \tilde{k}_1, \]  
\[ \tilde{u}_{j-1} + A\tilde{u}_j + \tilde{u}_{j+1} = \tilde{k}_j, \quad j=2,3,\ldots,q-1 \]  
\[ \tilde{u}_q + A\tilde{u}_q = \tilde{k}_q, \]

where

\[ \tilde{u}_j = q^T u_j, \quad \tilde{k}_j = q^T k_j, \quad j=1,2,\ldots,q. \]

The components of \( \tilde{u}_j \) and \( \tilde{k}_j \) are labelled as in (4.5). Then (4.8) can be rewritten for \( i=1,2,\ldots,p, \)

\[ \lambda_i \tilde{u}_{11} + w_i \tilde{u}_{12} = \tilde{k}_{11}, \]  
\[ w_{i-1} \tilde{u}_{jj-1} + \lambda_i \tilde{u}_{jj} + w_i \tilde{u}_{jj+1} = \tilde{k}_{jj}, \quad j=2,\ldots,q-1 \]  
\[ w_{i-1} \tilde{u}_{iq-1} + \lambda_i \tilde{u}_{iq} = \tilde{k}_{iq}. \]

Now let us write

\[ \Gamma_i = \begin{bmatrix} \lambda_i & w_i \\ w_i & \lambda_i & w_i \\ & & \ddots & \ddots \\ & & & \lambda_1 & w_1 \end{bmatrix} \]  

so that (4.10) is equivalent to the system of equations
\[ T_i \hat{u}_i = \hat{k}_i \quad (4.12) \]

Thus the algorithm proceeds as follows:

(i) Compute or determine the eigenvectors of \( B \) and the
eigenvalues of \( B \) and \( C \).

(ii) Compute \( \overline{k}_j = Q^T k_j \), \( j=1,2,\ldots,q \).

(iii) Solve \( T_i \hat{u}_i = \hat{k}_i \), \( i=1,2,\ldots,p \).

(iv) Compute \( u_j = Q\hat{u}_j \), \( j=1,2,\ldots,q \).

Hockney (1965) has taken advantage of the fact, that
in the case of solving Poisson's equation in a square, the
matrix \( Q \) is known and that one can use the fast Fourier
transform (Cooley and Tukey, 1965) to perform steps (ii) and
(iv).

The next three algorithms are all based upon the
Gaussian elimination method without pivoting for size.
Wilkinson (1961) has shown that such algorithms are stable
with respect to the growth of rounding errors if the
coefficient matrix \( A \) is diagonally dominant and/or positive
definite. Hence in any future application of these
algorithms, it will be assumed that one or both of these
properties hold.

**Algorithm 4.2**

Consider system (4.1) where

\[
A = \begin{bmatrix}
  b_1 & c_1 \\
  a_2 & b_2 & c_2 \\
  & & \ddots \\
  & & & c_{n-1} \\
  & & & a_n & b_n
\end{bmatrix} \quad (4.14)
\]
Defining
\[ u_1 = \frac{c_1}{b_1} ; \quad u_1 = \frac{c_1}{b_1 - a_1 u_{i-1}} , \quad 2\text{sign}, \]
\[ g_1 = \frac{k_1}{b_1} ; \quad g_1 = \frac{k_1 - a_1 g_{i-1}}{b_1 - a_1 u_{i-1}} , \quad 2\text{sign}, \]
the components \( u_i \) of the solution vector \( u \) are then given recursively by
\[ u_n = g_n ; \quad u_i = g_i - a_i u_i + 1 ; \quad 1\text{sign} - 1. \]

This algorithm was first presented by Thomas (1949).

In terms of arithmetic calculations, the above method requires at most 3 multiplications, 2 divisions and 3 additions per unknown \( u_i \). If system (4.1) with (4.14) needs to be solved for more than once for different \( u \) and \( k \), e.g. and in an iterative method, the coefficient matrix \( A \) remains unchanged, then the amount of work reduces to at most 2 multiplications, 1 division and 2 additions per unknown \( u_i \). This is because the vector \( u \) need only be calculated once in a situation like this.

**Algorithm 4.3**

Consider system (4.1) where
\[
A = \begin{bmatrix}
     b_1 & c_1 & a_1 \\
     a_2 & b_2 & c_2 \\
     \vdots & \vdots & \vdots \\
     c_{n-1} & c_n & a_n \\

e_n & a_n & b_n
\end{bmatrix}.
\]  
\quad (4.15)

Define
\[ s_1 = \frac{c_1}{b_1} ; \quad s_1 = \frac{c_1}{b_1 - a_1 s_{i-1}} , \quad 2\text{sign} - 1 ; \]
The components $u_i$ of $u$ are then given recursively by

$$ u_n = \frac{f_n}{H_n}; \quad u_i = f_i + s_i u_{i+1} + h_i u_{i-1}, \quad i = n-1(-1)1. $$

This algorithm has been published by Evans and Atkinson (1971). It requires at most 8 multiplications, 3 divisions and 6 additions per unknown $u_i$ at the first application. If the matrix elements remain unchanged for subsequent applications, this reduces to at most 4 multiplications, 1 division and 4 additions per unknown $u_i$. In other words, Algorithm 4.3 requires approximately twice the amount of work to obtain $u$ as the previous algorithm, 4.2, for both comparative cases.
Algorithm 4.1

Consider system (4.1) where

\[ A = \begin{bmatrix} c_1 & d_1 & e_1 \\ b_2 & c_2 & d_2 & e_2 \\ a_3 & b_3 & c_3 & d_3 & e_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ a_{n-2} & \ddots & \ddots & \ddots & \ddots & e_{n-1} \\ a_n & b_n & c_n & \ddots & \ddots & \ddots & e_{n-1} & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & \ddots & \ddots \\ & & & & & & & & \ddots \\ & & & & & & & & & \end{bmatrix} \]

(4.16)

The system may be solved by the following (gaussian) elimination scheme. Let

\[ u_1 = c_1 ; \ b_1 = d_1/u_1 ; \ \beta_0 = 0, \ \beta_n = 0 ; \]

\[ v_1 = e_1/u_1 ; \ \gamma_0 = 0, \ \gamma_n = \gamma_{n-1} = 0 ; \]

and define recursively:

\[ \alpha_j = b_j - a_j \beta_{j-2} ; \]

\[ u_j = c_j - a_j \gamma_{j-2} - \gamma_j \beta_{j-1} ; \quad \text{for } 2 \leq j \leq n \]

\[ \beta_j = (d_j - \gamma_j \gamma_{j-1})/u_j ; \]

\[ \gamma_j = e_j/u_j ; \]

Next form

\[ h_0 = 0, \ h_1 = b_1 / u_1, \ h_j = (k_j - a_j h_{j-2} - c_j h_{j-1}) / u_j, \quad 2 \leq j \leq n. \]

The values of \( u \) are then obtained recursively from the formulae

\[ u_n = h_n, \ u_j = h_j - \beta_j u_{j+1} - \gamma_j u_{j+2}, \quad j = n-1, n-2, \ldots, 1. \]

This algorithm was first presented by Conte and Darmaz (1958).

It requires not more than 8 multiplications, 3 divisions and 8 additions per unknown \( u_j \), and not more than 4 multiplications, 1 division and 4 additions per unknown for subsequent applications if the matrix \( A \) remains unchanged.
CHAPTER 5

PERIPHERAL ORDERINGS OF GRID POINTS

FOR PARABOLIC DIFFERENCE EQUATIONS
5.1 Introduction

In this chapter we will be considering finite difference methods of solution of time dependent problems in two space variables over a plane, connected region R. We look at such problems in two different coordinate systems, namely cartesian (x-y) and polar (r-θ). The regions we will be particularly interested in, which correspond to these respective geometries, are the square and the unit circle and ring (\(a \leq r \leq 1\)).

On the first region we superimpose a square grid of size \(h\), and on the second, a polar grid with mesh lengths \(\Delta r, \Delta \theta\). Their common factor will be a peripheral ordering of the internal grid points. This is defined as a serial labelling of the mesh points along successive circuits parallel to the boundary \(B\), of the region \(R\). We shall often refer to these circuits as peripherals. This ordering has been investigated by Benson (1969) and Benson and Evans (1973), in connection with the iterative solution (by S.O.R. techniques) of elliptic difference equations on regions such as the hollow square bar and the ring (\(a \leq r \leq 1\)).

On each of the first three regions mentioned, we will be especially interested in the solution of \(\frac{\partial u}{\partial t} = \nabla^2 u\), where \(\nabla^2\) denotes the well known Laplacian operator (in two space dimensions). In section 5.2, we investigate the use of a peripheral hopscotch method for solving such a problem when \(\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\), and at the end of that section consider applying the process to time dependent problems on open regions in the x-y plane. In sections 5.3-5.5, we examine three types of difference methods for the solution of the above problem when \(\nabla^2 = \frac{\partial^2}{\partial x^2} + (1/x)\frac{\partial}{\partial x} + (1/x^2)\frac{\partial^2}{\partial \theta^2}\). These are respectively explicit, alternating-direction implicit (Peaceman-Rachford) and hopscotch methods applied to a polar grid.

5.2 Peripheral Hopscotch using cartesian coordinates

(1) Problem definition

We are required to solve

\[
\frac{\partial u}{\partial t} = \nu + g(x,y,t), \quad u = U(x,y,t)
\]  

(5.1)
where

$$LU = \frac{\partial}{\partial x} \{ P(x,y) \frac{\partial U(x,y)}{\partial x} \} + \frac{\partial}{\partial y} \{ Q(x,y) \frac{\partial U(x,y)}{\partial y} \} - \sigma(x,y)U(x,y,t) \quad (5.2)$$

and $P, Q$ and $\sigma$ are piecewise continuous in $\bar{R}$, the closure of $R$, and satisfy

$$P(x,y) > 0, Q(x,y) > 0, \sigma(x,y) > 0, \text{ for all } (x,y) \in \bar{R}. \quad (5.3)$$

Let us then consider the case when the plane region $R$ is a SQUARE. The solution $U(x,y,t)$ is then required in the cylinder $R \times [0, t_f]$ subject to the following initial and boundary data:

$$U(x,y,0) = f_1(x,y), \quad (x,y) \in \bar{R};$$
$$U(x,y,t) = f_2(x,y,t), \quad \text{ for all } (x,y,t) \in \bar{R} \times [0, t_f], \quad (5.4)$$

where $f_1(x,y)$ is prescribed for all $(x,y) \in \bar{R}$, and $f_2(x,y,t)$ is given for all $(x,y,t) \in \bar{R} \times [0, t_f]$, where $B$ is the continuous boundary of the region $R$.

(ii) Peripheral ordering

We will now superimpose a square grid of size $h$ on $R$, and order the mesh points as in Figure 5.1, i.e., a peripheral ordering.

Let the matrix $A$ (as in Chapter 2), be defined by

$$(Au)_{i,j} = -h^2 L_h u_{i,j}, \quad (5.5)$$

where $L_h$ is defined as in section 2.4.

Hence, using finite difference approximations of the type (3.61)
and \((3.62)\), we have
\[
(Au)_{i,j} = D_{i,j} u_{i+1,j} - P_{i-1,j} u_{i-1,j} - P_{i+1,j} u_{i+1,j} - Q_{i,j+1} u_{i,j+1} - Q_{i,j-1} u_{i,j-1} \quad (5.6)
\]
where
\[
D_{i,j} = P_{i-1,j} + P_{i+1,j} + Q_{i,j+1} + Q_{i,j-1} \quad (5.7)
\]
In general, if the length of each side of the square is \(h\), then the matrix \(A\) has the form
\[
A = \begin{bmatrix}
A_{1,1} & A_{1,2} \\
A_{2,1} & A_{2,2} & A_{2,3} \\
& & \ddots & \ddots \\
& & & A_{k-1,k} \\
& & & & A_{k,k}
\end{bmatrix}
\quad (5.8)
\]
where \(k\), the number of peripherals, is the greatest number not exceeding \(N/2\). The matrices \(A_{i,i}\) are square and of order
\[
\begin{align*}
4(N-2i) & \quad \text{for } 1 \leq i \leq k-1 \\
4(N-2i) & \quad \text{for } N \text{ odd and } i=k \\
1 & \quad \text{for } N \text{ even and } i=k,
\end{align*}
\]
i.e. if \(N\) is even, the last diagonal matrix \(A_{k,k}\) is just a single element. The rectangular matrices \(A_{i,i-1}\) (\(2 \leq i \leq k-1\)) are of order \(4(N-2i) \times 4[N-2(i-1)]\) and
\[
A_{i,i-1} = \begin{cases} 
4(N-2i) \times 4[N-2(i-1)] & \text{if } N \text{ is odd,} \\
1 \times 4[N-2(i-1)] & \text{if } N \text{ is even.}
\end{cases}
\quad (5.10)
\]
In particular, applying equation \((3.6)\) at each internal grid point in Figure 5.1 (in the specified order) gives
where we have excluded boundary values. (5.11) is then a special case of (5.8). We notice immediately that $A$ has the following properties:

(a) it is symmetric with positive diagonal entries,

and (b) is irreducibly diagonally dominant.

That $A$ is irreducible, follows from the proof of Theorem 1.1, Corollary, since $A$ has a tri-diagonal band about its main diagonal consisting of all non-zero elements.

It now follows from Theorem 1.5, Corollary that $A$ is positive definite.

**(iii) The peripheral hopscotch process**

Let us return to the general formula given by (2.19). Let us
specify
\[ \theta^k_{ij} = w^k_{ij}, \]
and
\[ \theta^k_{ij} = c^k_p = \begin{cases} 1 & \text{if } k+p \text{ is even} \\ 0 & \text{if } k+p \text{ is odd} \end{cases} \]  
(5.13)

where the mesh point \((ih,jb)\) lies on the \((p)\)th peripheral, for integer \(p\).

We shall assume that the first peripheral is the largest one, i.e. the \((p)\)th peripheral contains fewer mesh points than the \((p-1)\)th peripheral (eight less in fact - see (5.9)). Substituting (5.13) into (2.19) gives the scheme

\[ u^{k+1}_{i,j} - \Delta t \theta^k_p \left( \frac{L^k_{h} u^k_{i,j} + \theta^k_{i,j}}{p} \right) = u^k_{i,j} + \Delta t \theta^k \left( \frac{L^k_{h} u^k_{i,j} + \theta^k_{i,j}}{p} \right), \]  
(5.14)

which we shall refer to as the peripheral hopsotch method.

In an analogous manner to the derivation of equation (2.29), it can be shown from (5.14) that

\[ u^{k+2}_{i,j} = 2u^{k+1}_{i,j} - u^k_{i,j}, \]  
(5.15)

when \(k+p\) is odd, since \(\theta^{k+2}_{i,j} = \theta^k_{i,j} = \theta^k_{i,j}\), assuming \((ih,jb)\) lies on the \((p)\)th peripheral.

**Computational procedure**

The computation for this method now proceeds as follows. For the first step \((k=1)\) when \(k+p\) is even, calculate

\[ u^{k+1}_{i,j} = u^k_{i,j} + \Delta t \left[ \frac{L^k_{h} u^k_{i,j} + \theta^k_{i,j}}{p} \right], \]  
(5.16)

where \((ih,jb)\) is \((p)\)th peripheral. Then for the remaining mesh points, with \(k+p\) odd calculate

\[ u^{k+1}_{i,j} = u^k_{i,j} + \Delta t \frac{L^k_{h} u^k_{i,j} + \theta^k_{i,j}}{p} \]  
or

\[ u^{k+1}_{i,j} = u^k_{i,j} + \Delta t \theta^k_{i,j}. \]  
(5.17)

For each value of \(p\), such that \(k+p\) is odd, (5.17) involves the implicit solution of a periodic tri-diagonal system of equations (of the form (4.14)) by Algorithm 4.3, whose coefficient matrix is of the order given
by \( i=p \) in (5.9).

For subsequent steps, assuming we are on the \((k+2)\)th time level, we use equation (5.15) at those points with \(k+p\) odd. Then we use (5.17) again on the remainder of the mesh points along alternate peripherals. This process continues until the final solution \(u(T)\) at some desired time \(t=T\) is reached. It will be noticed, that it is a similar type of procedure to the line hopscotch method, described in section 2.4.

Next let us define the diagonal matrix \(I_1^{(p)}\) such that if \(\left[I_1^{(p)}u^{2k}\right]_{i,j}\) denotes the component of the vector \(I_1^{(p)}u^{2k}\) corresponding to the spatial mesh point \((ih,jh)\), then with a peripheral ordering of these unknown grid points (see Figure 5.1),

\[
\left[I_1^{(p)}u^{2k}\right]_{i,j} = \theta_{p}^{2k} u_{ij}^{2k}, \text{ for all } (ih,jh) \in \mathbb{R}, \tag{5.18}
\]

where \((ih,jh) \in (p)\)th peripheral, and \(\theta_{p}^{2k}\) is defined by (5.13). Hence \(I_1^{(p)}\) is a block diagonal matrix, whose block diagonal elements are alternately unit and null matrices, whose orders are given by (5.9).

Let also

\[
I_2^{(p)} = I - I_1^{(p)}. \tag{5.19}
\]

We may now define the two step peripheral hopscotch process globally in the form

\[
\begin{equation}
\left[ I + rI_2^{(p)} \right]u^{2k+1} = \left[ I - rI_2^{(p)} \right]u^{2k} + \Delta t \left[ I_1^{(p)}u^{2k+1} + I_1^{(p)}u^{2k} \right]
\end{equation}
\]

and

\[
\begin{equation}
\left[ I + rI_2^{(p)} \right]u^{2k+2} = \left[ I - rI_2^{(p)} \right]u^{2k+1} + \Delta t \left[ I_1^{(p)}u^{2k+2} + I_1^{(p)}u^{2k+1} \right]
\end{equation}
\]

(5.20)

where \(r=h^2/\Delta t\), and the data from the boundary of the square has been absorbed in the \(g\) vectors. It is seen that (5.20) is a Peaceman-Rachford type procedure with the coefficient matrix \(A\) defined by (5.5) split as

\[
A = I_1^{(p)} + I_2^{(p)}A. \tag{5.21}
\]

(1v) Stability

In this subsection we treat not only the stability of the peripheral hopscotch process, but that of the odd-even and line hopscotch methods as
well. In each case, we assume that the matrix $A$ defined by (5.5) with the respective orderings of the grid points, is positive definite.

For the odd-even and line hopscotch processes we define respective diagonal matrices $I_1^{(OE)}, I_2^{(OE)}$ and $I_1^{(L)}, I_2^{(L)}$ in an analogous manner to equations (5.18) and (5.19). The difference being that in these two cases, we adopt a line ordering of the spatial mesh points, and $6_{ij}$ are defined by (2.24) and (2.25) respectively. Let $(I_1, I_2)$ denote any one of the pairs $(I_1^{(C)}, I_2^{(C)}), (I_1^{(OE)}, I_2^{(OE)})$ and $(I_1^{(L)}, I_2^{(L)})$, then

$$
\begin{align*}
I_1 I_2 &= I_2 I_1 = 0, \\
I_1^2 &= I_1, \\
I_2^2 &= I_2.
\end{align*}
$$

(5.22)

Eliminating $u^{2k+1}$ from (5.20), gives

$$
u^{2k+2} = T u^{2k} + b^{2k}, \quad k=0,1,2,\ldots,$$

(5.23)

where $b^{2k}$ is independent of the $u$'s and where

$$
T = \begin{bmatrix} \tau & 0 \\ 0 & \tau \end{bmatrix},
$$

(5.24)

Then

$$
T^k = \begin{bmatrix} \tau^k & 0 \\ 0 & \tau^k \end{bmatrix}.
$$

(5.25)

Since the three hopscotch processes we are considering are all special cases of the general formula (2.19), a sufficient condition for stability is given by Theorem 2.1 (p. 37). A simple $2 \times 2$ counter-example however, shows that the theorem is not generally applicable in this particular context. Namely, let the positive definite matrix $A$ and the diagonal matrix $I_1$ be defined by

$$
A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad I_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.
$$

(5.26)

Then $Q = I_1 A$. It is then easily verified that the eigenvalues of $Q+Q^T$ are $2+\sqrt{5}$ and $2-\sqrt{5}$, showing that $Q+Q^T$ is not non-negative definite.

Hence condition (2.59) [Theorem 2.1] is not satisfied.
To remedy this situation, we introduce another norm and prove stability with respect to it. Firstly, define

$$||u||_{H_2} = \left(\frac{1}{N} \sum_{i=1}^{N} |u_i|^2 \right)^{1/2} = \frac{1}{n} ||u||_2.$$  

(5.27)

This is known as the $H_2$ or mean 2nd power norm (Rosen, 1966) of a $\mathbb{R}^N$ vector $u$. Further, define the $A_i$-vector norm as

$$||u||_{A_i} = ||A_i u||_{H_2},$$  

(5.28)

for positive definite $A$ ($A_i$ is defined in section 1.4). It is then easily shown that the $A_i$-matrix norm subordinate to (5.28) is

$$||c||_{A_i} = ||A_i c A_i^{-1}||_2.$$  

(5.29)

Until stated otherwise, it will be assumed from now on that $||\cdot||$ denotes $||\cdot||_{A_i}$.

Returning our attention to (5.24), we have

$$\tilde{z} = A^{-1} D_2 P_A \tilde{z},$$  

(5.30)

where

$$D_j = [A^i A_j A^i] [A^i A_j A^i]^{-1}, \quad j=1,2.$$  

(5.31)

Since $A$ is positive definite, we know from section 1.4 that $A_i$ is positive definite also. Because we can write

$$A_i^i A_j A^i = \langle A^i A_j A^i, A_j A^i \rangle, \quad j=1,2,$$  

using (5.22)

it follows from Theorem 1.3 that $A_i^i A_j A^i$ ($j=1,2$) are non-negative definite.

Thus from (5.30) and (5.31)

$$||\tilde{z}|| = ||D_2 P_A \tilde{z}||_2 \leq ||D_2||_2 \cdot ||P_A \tilde{z}||_2 \leq 1,$$  

(5.32)

since $||D_j||_2 \leq 1$ for $j=1,2$ by Lemma 2.2 (p.36). Also we have

$$||[A^i A_j A^i]^{-1}|| = ||[A^i A_j A^i]^{-1}||_2 \leq 1,$$  

(5.33)

by Lemma 2.1 (p.36). Next

$$||[A^i A_j A^i]|| = ||[A^i A_j A^i]||_2 = \rho(A^i A_j A^i),$$  

(5.34)

since $(A^i A_j A^i)$ is symmetric [see (1.25)] and $A^i A_j A^i A_j A^i = A^i A_j A^i A_j A^i A^i$, i.e. $A_j A^i$ and $A^i A_j A^i$ are similar. Now, because $A_j A^i$ is similarly composed of rows from $A$ and null rows, it follows from Theorem 1.4 Corollary and equation (5.6) that
\begin{equation}
s(L_n) \leq \max_{(x,y) \in \mathbb{R}} \left( 4(f(x,y) + g(x,y)) + h^2 \sigma(x,y) \right) = M, \text{ say}
\end{equation}

where we assume that \( M \) is bounded. (For example, if \( I_n \) is the usual central difference approximation of \( L = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \), then \( f(x,y) = g(x,y) = 1, \sigma(x,y) = 0 \), for all \( (x,y) \in \mathbb{R} \), the square (say). In this case \( M = 8 \). It then follows from (5.34) and (5.35) that

\begin{equation}
|||I + L_n||| \leq 1 + \beta M.
\end{equation}

Therefore, from (5.36) and equations (5.35), we have

\begin{equation}
|||T^k||| \leq 1 + \alpha M, \text{ (for all integer } k > 0),
\end{equation}

a constant independent of \( k \), assuming \( r = \Delta t / h^2 \) remains fixed.

Hence, under the above requirements, it follows from Condition 2.2 that the odd-even, line and peripheral hopscotch methods are stable.

We can therefore state

**Theorem 5.1**

The odd-even, line and peripheral hopscotch processes are stable (in the \( A^1 \)-norm) for the solution of equation (5.1) if the matrix \( A \) [defined by (5.5)] is positive definite.

From the remarks made in section 3.1, the above hypothesis will be satisfied if the region \( \mathbb{R} \) contains only regular mesh points.

**Convergence**

In this subsection, we look at the convergence of the three afore-mentioned hopscotch methods in the \( A^1 \)-norm. The general two step process (2.19) is

\begin{equation}
\begin{aligned}
[1 - \Delta t L_1] u_{k+1,j}^* &= [1 + \Delta t L_2] u_{k,j}^* + \Delta t b_{k,j}^* \\
[1 - \Delta t L_2] u_{k,j}^* &= [1 + \Delta t L_1] u_{k,j}^* + \Delta t b_{k,j}^*.
\end{aligned}
\end{equation}

where

\begin{equation}
\begin{aligned}
L_1 &= \sigma_{k+1,1}^1 (1) + \sigma_{k+1,2}^1 (2), \\
L_2 &= \sigma_{k,1}^1 (1) + \sigma_{k,2}^2 (2),
\end{aligned}
\end{equation}

\begin{equation}
\begin{aligned}
b_{k,j}^* &= (\sigma_{k,1}^1 (1) + \sigma_{k,2}^2 (2)) + (\sigma_{k+1,1}^1 (1) + \sigma_{k+1,2}^1 (2)) + (\sigma_{k,1}^1 (1) + \sigma_{k,2}^2 (2)),
\end{aligned}
\end{equation}

(5.39)
In the case of the odd-even, line and peripheral hopscotch methods, \( e_{i,j}^k \) are defined by (2.24), (2.25) and (5.13), and \( L_h^{(1)} \), \( L_h^{(2)} \), \( g^{(1)} \) and \( g^{(2)} \) are defined in section 2.4. We assume that \( U(x,y,t), g(x,y,t), g^{(1)}(x,y,t) \) and \( g^{(2)}(x,y,t) \) are sufficiently smooth functions so that they have bounded derivatives of any order required.

If we define
\[
e_{i,j}^k = u_{i,j}^k - u_{i,j}^k,
\]
as the difference between the true and computed solution (respectively) at the mesh point \((ih,jh,kAt)\), this leads to the error equation (on eliminating \( u_{i,j}^k \) from (5.38)):
\[
\left[ I + \Delta t L_1 \right] \left[ I + \Delta t L_2 \right] e_{i,j}^{k+1} = \left[ I + \Delta t L_1 \right] \left[ I + \Delta t L_2 \right] e_{i,j}^{k-1} + d_{i,j}^k.
\]

McGuire (1970) pp. 21-23, has shown that
\[
d_{i,j}^k = 2At^3 L_2 \frac{\partial U}{\partial t} e_{i,j}^k + 2At^3 L_1 e_{i,j}^k \frac{\partial g^{(1)}}{\partial t} + \frac{\partial g^{(2)}}{\partial t} e_{i,j}^k
\]
\[+ O(At^3 + At^a) + O(At^5/h^2) + \frac{\partial g^{(1)}}{\partial t} e_{i,j}^k + \frac{\partial g^{(2)}}{\partial t} e_{i,j}^k
\]
\[= O(At^a + At^3/h^2 + At^3)
\]
is the local truncation error for the general formula (2.19) where \( a = 1 \) or \( 2 \) depending upon whether \((ih,jh)\) is an irregular or regular grid point, respectively [see, for instance, equation (2.16)]. Let us assume in this case, that we have only regular mesh points in \( R \), so that \( a = 2 \) for all \((ih,jh) \in R \).

From (5.41) the global error equation for our hopscotch methods is
\[
\left[ I + \tau L_2 \right] \left[ I + \tau L_1 \right] e_{k+1} = \left[ I - \tau L_2 \right] \left[ I - \tau L_1 \right] e_{k-1} + d_k
\]
where \( k \) is an odd integer (> 0) and \( L_1, L_2, A \) may be defined for the odd-even, line or peripheral hopscotch. Replacing \( k \) by \( 2k+1 \), (5.43) may be written as
\[
e^{2k+2} = \tau e^{2k} + \frac{\partial}{\partial t} e^{2k+1}, \quad k=0,1,2,...
\]
where $T$ is defined by $(5.24)$ and
\[
\xi^{2k+1} = [I + rT_{k}A]^{-1}[I + rT_{k}A]^{-1} d^{2k+1}.
\]
(5.45)

Thus, from $(5.44)$ we have
\[
\xi^{2k} = T_{k}^{2} + T_{k-1}^{2} + T_{k-2}^{2} + \ldots + T_{2}^{2} + T_{1}^{2} + T_{0}^{2} + 1.
\]
(5.46)

Now, if the processes are stable, we have $||T^{k}||_{SC}$ (a constant) for all integers $k > 0$. Also from $(5.45)$,
\[
||\xi^{2k+1}|| \leq ||(I + rT_{k}A)_{2}^{-1}||(I + rT_{k}A)_{1}^{-1}||d^{2k+1}||
\]
\[
\leq ||d^{2k+1}||_{2}, \quad \text{using Lemma 2.1 (p. 36)}.
\]

Now, from $(5.46)$, we obtain
\[
||\xi^{2k}|| \leq ||T^{k}||_{2} ||d^{0}|| + ||\xi_{k}||.
\]
where
\[
||\xi_{k}|| = ||T_{k}^{2} + \ldots + T_{1}^{2} + T_{0}^{2} + 1||.
\]

Thus
\[
||\xi_{k}|| \leq ||T^{k}||_{2} ||d^{2k-1}||, \quad \max 1 \leq 4 \leq 2k
\]

Hence, since $||\xi^{0}|| = ||\xi^{0} - \xi^{0}|| = 0$ (same initial data for exact finite difference solutions), we obtain
\[
||\xi^{2k}|| \leq kC \max 1 \leq 4 \leq 2k.
\]
(5.47)

Now from $(5.28)$ we have
\[
||d^{2k-1}|| = ||d^{2k-1}||_{M} = ||A_{1}d^{2k-1}||_{H_{2}} + ||A_{1}||_{H_{2}} ||d^{2k-1}||_{H_{2}}
\]
\[
= ||A_{1}||_{2} ||d^{2k-1}||_{H_{2}}.
\]
(5.48)

It follows from $(5.27)$ that since $d_{k}^{j} = 0(\Delta t^{2} + \Delta t^{3}/h^{2} + \Delta t^{3})$, for all $i, j, k, \text{ then}$
\[
||d^{2k-1}||_{H_{2}} = 0(\Delta t^{2} + \Delta t^{3}/h^{2} + \Delta t^{3}) \quad \text{for all $A_{1}$}. \quad (5.49)
\]

Since the eigenvalues of the positive definite matrix $A_{1}$ are the positive square roots of the eigenvalues of $A$, we have in an analogous manner to equation $(5.35)$
\[
||A_{1}||_{2} = \rho(A_{1}) \leq \max_{(x, y) \in \mathbb{R}} (A_{1}(x, y) + o(x, y) + h^{2}o(x, y))^{1/2} = \mathcal{M}^{1}, \quad \text{say}, \quad (5.50)
\]

where $\mathcal{M}$ is bounded, since for stability we assumed $||w|| = (\mathcal{M})^{2}$ was bounded. Therefore, equations $(5.48) - (5.50)$ yield
\[
||d^{2k-1}||_{M} = 0(\Delta t^{2} + \Delta t^{3}/h^{2} + \Delta t^{3}). \quad (5.51)
\]
It now follows from (5.47) and (5.51) that
\[ ||x^{2k}|| \leq t_{2k} C M' h^2 + (h^4 + \Delta t^2/h^2 + \Delta t^2), \tag{5.52} \]
where \( t_{2k} = 2k \Delta t \). Hence the three afore-mentioned hopscotch methods will converge in the \( A^1 \)-norm (in the sense of Definition 2.5) as \( \Delta t \to 0, \ h \to 0 \) if \( \Delta t/h \to 0 \) also. This is achieved if we keep \( r = \Delta t/h^2 \) constant.

Hence we can state the following convergence theorem.

**Theorem 5.2**

The odd-even, line and peripheral hopscotch procedures for the solution of equation (5.1) have a convergence rate of \( O(\Delta t + h^2) \) (in the \( A^1 \)-norm), if \( r = \Delta t/h^2 \) is constant, if the processes are stable, and if the region \( R \) under consideration contains only regular mesh points.

(As mentioned before, the last condition guarantees the positive definiteness of the matrix \( A \).)

We also notice from (5.51) that the above three methods will be consistent (in the sense of Definition 2.6) if \( M' \) is bounded, and more especially if \( \Delta t/h \to 0 \) as \( \Delta t \to 0, \ h \to 0 \). This latter condition is realised if \( r \) is kept constant.

(vi) **Comparative truncation error analysis for the peripheral and other hopscotch methods**

In this section we look at the truncation error of the general two step process (2.19), in particular for the peripheral hopscotch process. Let us assume that

\[ L_h u_{i, j} = \left( A_{i, j} u_{i+1, j} + B_{i, j} u_{i-1, j} + C_{i, j} u_{i, j} + D_{i, j} u_{i, j+1} + E_{i, j} u_{i, j+1} \right)/h^2, \tag{5.53} \]

also that

\[ L_h^{(1)} u_{i, j} = \left( A_{i, j} u_{i+1, j} + B_{i, j} u_{i-1, j} + C_{i, j}^{(1)} u_{i, j} + D_{i, j} u_{i, j+1} \right)/h^2, \tag{5.54} \]

and

\[ L_h^{(2)} u_{i, j} = \left( A_{i, j} u_{i+1, j} + B_{i, j} u_{i-1, j} + C_{i, j}^{(2)} u_{i, j} + D_{i, j} u_{i, j+1} \right)/h^2, \tag{5.55} \]

where

\[ C_{i, j} = C_{i, j}^{(1)} + C_{i, j}^{(2)}. \]
Let us specify for convenience that the region $R$ is a square, and that the peripherals and the sides of the peripherals are numbered as in Figure 5.2. (The boundary is the 0th peripheral). In this case then, $L_n^k, L_n^{(1)}$ and $L_n^{(2)}$ will be $O(h^2)$ difference analogues.

In the expression for $d_{i,j}^k$, (5.42), the terms involving $O(\Delta t^3)$ and $O(\Delta t h^2)$ are the same for the three hopscotch methods (odd-even, line and peripheral) as well as for the Crank-Nicolson and Peaceman-Rachford procedures. The only terms which are different are those involving $O(\Delta t^3/h^2)$ in $d_{i,j}^k$. McGuire (1970) investigates the magnitude of the coefficient of the $\Delta t^3/h^2$ term to get a clue to the comparative accuracy of the hopscotch, Peaceman-Rachford and Crank-Nicolson methods. We will perform a similar analysis here on the peripheral hopscotch process to put it in perspective with the afore-mentioned methods.

Now the only terms in the local truncation error of (5.42) which give $O(\Delta t^3/h^2)$ are

\[
\begin{align*}
2\Delta t^3 L_1 L_2 \left. \frac{3\partial}{\partial t} \right|_{i,j}^k \\
2\Delta t^3 L_1 \left( \left. \frac{\partial^{k+1}}{\partial t^k} \right|_{1,j} \frac{3\partial}{\partial t} + \left. \frac{\partial^{k+1}}{\partial t^k} \right|_{4,j} \right)_{i,j}.
\end{align*}
\]
An analogous derivation of such terms is given in subsection 5.4(iv),

From (5.13), \( \phi_{1,j}^{k} = \eta_{i,j}^{k} \), hence from (5.39),

\[
2\Delta t L_{1}\frac{\partial u^{k}}{\partial t} i,j = 2\Delta t \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j
\]

(5.57)

and

\[
2\Delta t L_{1} \left( \phi_{i,j}^{k+1} \delta_{i,j}^{2} + \eta_{i,j}^{k+1} \delta_{i,j}^{2} \right) i,j = 2\Delta t \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j.
\]

(5.58)

We are assuming that \( \phi_{i,j}^{k+1} = \phi_{i,j}^{k+1} \) (for peripheral hopscotch) = 1

(i.e. \((k+1)p\) is even − see (5.13)), since otherwise the terms in

(5.57) and (5.58) are zero. This of course implies from (5.13) that

\( \phi_{i,j}^{k} = \phi_{i,j}^{k} = 0 \) when \((ih,jh)\) lies on the \((p)\)th peripheral.

In general using (5.53), we have

\[
2 \frac{\Delta t}{h^2} \phi_{i,j}^{k+1} L_{1} \phi_{i,j}^{k} L_{1} \frac{\partial u^{k}}{\partial t} i,j = 2 \frac{\Delta t}{h^2} \phi_{i,j}^{k} \left( A_{i,j} \phi_{i,j}^{k} - 1 + \frac{\partial u^{k}}{\partial t} i,j \right)
\]

+ \( B_{i,j} \phi_{i,j}^{k} + C_{i,j} \frac{\partial u^{k}}{\partial t} i,j + D_{i,j} \phi_{i,j}^{k} L_{1} \frac{\partial u^{k}}{\partial t} i,j
\]

+ \( E_{i,j} \phi_{i,j}^{k} L_{1} \frac{\partial u^{k}}{\partial t} i,j + F_{i,j} \phi_{i,j}^{k} L_{1} \frac{\partial u^{k}}{\partial t} i,j\).

\]

(5.59)

Since \( \phi_{i,j}^{k} = \phi_{i,j}^{k} = 0 \) the \((C_{i,j} \phi_{i,j}^{k} L_{1} \frac{\partial u^{k}}{\partial t} i,j\) term disappears. Similarly,

\[
2\Delta t \frac{\delta_{i,j}^{k+1}}{h^2} L_{1} \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j = 2\Delta t \delta_{i,j}^{k+1} L_{1} \left( A_{i,j} \delta_{i,j}^{k+1} - 1 + \frac{\partial u^{k}}{\partial t} i,j \right)
\]

+ \( B_{i,j} \delta_{i,j}^{k+1} + C_{i,j} \frac{\partial u^{k}}{\partial t} i,j + D_{i,j} \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j
\]

+ \( E_{i,j} \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j + F_{i,j} \delta_{i,j}^{k+1} L_{1} \frac{\partial u^{k}}{\partial t} i,j\).

\]

(5.60)

To determine the coefficient of the \(O(\Delta t^3/h^2)\) term in (5.59) and

(5.60) at the point \((ih,jh,k\Delta t)\), we must consider whether \((ih,jh)\),

which we shall denote by \((i,j)\), lies on:

(I) the first or third side of the \((p)\)th peripheral;

(II) the second or fourth side of the \((p)\)th peripheral, or on

(III) a corner of the \((p)\)th peripheral.

We shall be assuming that \((i,j)\) cannot lie on a side and a corner.
Case I

We have that $\theta^k_{i,j} = 0_p = 0$, when $(i,j)$ lies on the $(p)$th peripheral. Hence, if $(i,j)$ lies on the first or the third side of the $(p)$th peripheral, then so do $(i-1,j)$ and $(i+1,j)$. Hence $\theta^k_{i-1,j} = \theta^k_{i+1,j} = 0$, in this case. Since also $(i,j-1),(i,j+1)$ lie on the $(p-1)$th and $(p+1)$th peripheral, or vice-versa, depending upon whether $(i,j)$ lies on the first or third side, we have

from (5.13) that $\theta^k_{i,j-1} = \theta^k_{i,j+1} = 1$.

Thus for Case I, using Taylor series expansions about $y^k_{i,j}$, equation (5.59) becomes

$$2 \frac{\Delta t^3}{h^2} \theta^k_{i+1,j} \theta^k_{i,j} \phi^k_{i,j} = 2 \frac{\Delta t^3}{h^2} \theta^k_{i+1,j} (A_i,j + \frac{3u^k_{i+1,j}}{a_i} + E_{i,j} + \frac{3u^k_{i,j}}{a_i}) + \theta^k_{i+1,j} (\Delta t)^3 .$$

Similarly, we can show for this case, that (5.60) becomes

$$2 \frac{\Delta t^3}{h^2} \theta^k_{i+1,j} \theta^k_{i,j} \phi^k_{i+1,j} = 2 \frac{\Delta t^3}{h^2} \theta^k_{i+1,j} (B_{i,j} + \frac{3g^k_{i+1,j}}{a_i} + E_{i,j} + \frac{3g^k_{i,j}}{a_i}) - \theta^k_{i+1,j} (\Delta t)^3 .$$

Hence for Case I, the $O(\Delta t^3/h^2)$ terms of the local truncation error of the peripheral hopscotch method are, from (5.61) and (5.62)

$$\frac{\Delta t^3}{h^2} \theta^k_{i+1,j} (A_{i,j} + E_{i,j}) (2L e^{3u} - 2g^k_{i,j}) .$$

Case II

In this case, we have $\theta^k_{i,j-1}, \theta^k_{i,j+1} = 0$, and $\theta^k_{i-1,j}, \theta^k_{i+1,j} = 1$.

In a similar manner to Case I, the $O(\Delta t^3/h^2)$ terms for Case II are

$$\frac{\Delta t^3}{h^2} \theta^k_{i+1,j} (B_{i,j} + D_{i,j}) (2L e^{3u} - 2g^k_{i,j}) .$$
\textbf{Case III}

Let the first corner be the bottom left-hand one, and number the other three anti-clockwise. Then

(a) for the first corner, \( \theta_{i,j+1} = \theta_{i+1,j} = 0 \), since \((i,j+1), (i+1,j)\) lie on the \( \{p\}\)th peripheral. Hence, \( \theta_{i-1,j} \theta_{i,j-1} = 1 \). Thus, in an analogous manner to Cases I and II, the \( O(\Delta t^3/h^2) \) terms are

\[
\frac{\Delta t^3}{h^2} \theta_{i,j+1} (A_{i,j} + B_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

(b) Likewise, the \( O(\Delta t^3/h^2) \) terms for the second corner are

\[
2 \frac{\Delta t^3}{h^2} \theta_{i,j+1} (A_{i,j} + B_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

(c) The \( O(\Delta t^3/h^2) \) terms for the third corner are

\[
\frac{\Delta t^3}{h^2} \theta_{i,j+1} (A_{i,j} + B_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

(d) Lastly, for the fourth corner the \( O(\Delta t^3/h^2) \) terms are

\[
\frac{\Delta t^3}{h^2} \theta_{i,j+1} (B_{i,j} + D_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

From McGuire (1970), p.25, the \( O(\Delta t^3/h^2) \) terms of the local truncation error of the odd-even hopscotch method are

\[
- \frac{\Delta t^3}{h^2} \theta_{i,j} c_{i,j} (2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

The \( O(\Delta t^3/h^2) \) terms of the local truncation error of the \( x \)-line hopscotch method are

\[
\frac{\Delta t^3}{h^2} \theta_{i,j} (A_{i,j} + B_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

From McGuire (1970), p.29, the \( O(\Delta t^3/h^2) \) terms of the local truncation error of the \( y \)-line hopscotch method are

\[
\frac{\Delta t^3}{h^2} \theta_{i,j} (B_{i,j} + D_{i,j})(2L \frac{3U_1}{\partial t} - 2 \frac{3g_{j,i}}{\partial t}, i,j) .
\]

Equations (5.69)-(5.71) were obtained in a similar manner to equations (5.59)-(5.68).

McGuire (1970) pp.26-27, has shown that for the Peaceman-Rachford
method, the sum of the terms in (5.56) is

$$\Delta t^3 L_h^{(2)}(2L_h^{(1)} \frac{2u}{\Delta x} + 2 \frac{\partial^2 u}{\partial x^2})_{i,j}.$$  \hspace{1cm} (5.72)

The reasons for this are that \( \theta^k_{i,j} \) and \( \eta^k_{i,j} \) in (5.56) are independent of \( i \) and \( j \) for this process [see equation (2.23)] and that \( L_h^{(1)}, L_h^{(2)} \) are elliptic difference operators.

With the aid of equations (5.63), (5.64)-(5.72) we can obtain a reasonable hierarchy of the previously mentioned methods, with respect to accuracy. The comparative accuracy of the methods depends a great deal on the nature of the elliptic differential operator \( L \), whose finite difference analogue is given by (5.53). In particular, if \( L \) is given by (5.2) and (5.3), we have from (5.53), (5.5)-(5.7),

$$|C_{i,j}^{(1)}| \geq |B_{i,j} + D_{i,j}|$$

and

$$|C_{i,j}^{(2)}| \geq |A_{i,j} + E_{i,j}|,$$

with \( B_{i,j}, D_{i,j} \) of opposite sign to \( C_{i,j}^{(1)} \) and \( A_{i,j}, E_{i,j} \) of opposite sign to \( C_{i,j}^{(2)} \) (see equations (5.6) and (5.7)). Also we have

$$|C_{i,j}^{(1)}| \geq |C_{i,j}^{(1)}|, |C_{i,j}^{(2)}|.$$  \hspace{1cm} (5.74)

It follows from (5.73) and (5.74) that the line and peripheral hopscotch will in general be more accurate than odd-even hopscotch in view of equations (5.63)-(5.71). Also, looking at equations (5.70) and (5.71), it is apparent that an \( x \)-line could be more accurate than a \( y \)-line hopscotch, or vice-versa. Since a peripheral is essentially a combination of \( x \)– and \( y \)-lines, this infers that an \( x \)– or \( y \)-line hopscotch could be more accurate than a peripheral hopscotch process. This is because equations (5.63) and (5.64) (for peripheral hopscotch) are identical to (5.70) (\( x \)-line hopscotch) and (5.71) (\( y \)-line hopscotch) respectively, for identical mesh points \((i,j)\).

From equation (5.72), it is clear that the Peaceman-Rachford procedure will in general be more accurate than/three above hopscotch
methods. The other relative merits of the P-R method with the hopscotch methods, are outlined at the end of Chapter 2.

It would appear then, that the peripheral hopscotch method is comparable in accuracy with the line hopscotch methods, particularly in the case, \( L = \partial^2 / \partial x^2 + \partial^2 / \partial y^2 \), when \( A_{i,j} = B_{i,j} = D_{i,j} = E_{i,j} = -1 \), as will be illustrated experimentally in the next section. However, if for the reasons given above, an x-line hopscotch (say) is more accurate than a y-line, then (again for the reasons given above) it follows that an x-line will most likely be more accurate than a peripheral hopscotch approach.

(vii) Computational experiments

In the following computational experiments we solved the equation

\[
\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + g(x,y,t),
\]

by the peripheral hopscotch method for values of \( U(x,y,t) \) and \( g(x,y,t) \) identical to those specified in McGuire (1970), Chapter 7 and Gourlay and McGuire (1971) pp.224-226, for the purpose of comparison. In each experiment the region of problem solution was the 2x2 square, (0\( \leq \)x \( \leq \)2, 0\( \leq \)y \( \leq \)2).

In experiment 1(a),

\[
\begin{align*}
g(x,y,t) &= \sin x \sin y e^{-t} - 4 \\
U(x,y,t) &= \sin x \sin y e^{-t} + x^2 + y^2
\end{align*}
\]

and in experiment 1(b),

\[
\begin{align*}
g(x,y,t) &= \sin x e^{-t} / (1+y)^2 - 2x - 6xy \\
U(x,y,t) &= \sin x \log (1+y) e^{-t} + x^3 + y^3 + xy^2
\end{align*}
\]

Experiment 2(a)

Comparison of the accuracy of the peripheral with the odd-even and line hopscotch methods (y-line ordering) and the Peaceman-Rachford method. The absolute errors after 100 time steps with \( h=0.1, \Delta t=0.01 \) were calculated for the solution of (5.75), (5.75a). Exact initial and boundary data was used. The results are given in table 1(a).
Examination of these results reveals that, as predicted by subsection 5.2(vi) line and peripheral hopscotch have the same degree of accuracy for this example, and both are about twice as accurate as the odd-even hopscotch.

**Table 2(a)**

The absolute errors for the four methods are displayed in the following manner for a selection of grid points. (1) refers to the odd-even hopscotch method, (2) to line hopscotch, (3) to peripheral hopscotch, and (4) to the Peaceman-Rachford method.

<table>
<thead>
<tr>
<th></th>
<th>(5,15)</th>
<th>(10,15)</th>
<th>(15,15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>8.01 $10^{-4}$</td>
<td>1.23 $10^{-3}$</td>
<td>1.14 $10^{-3}$</td>
</tr>
<tr>
<td>(2)</td>
<td>3.73 $10^{-4}$</td>
<td>5.82 $10^{-4}$</td>
<td>5.29 $10^{-4}$</td>
</tr>
<tr>
<td>(3)</td>
<td>3.59 $10^{-4}$</td>
<td>5.76 $10^{-4}$</td>
<td>5.05 $10^{-4}$</td>
</tr>
<tr>
<td>(4)</td>
<td>7.90 $10^{-5}$</td>
<td>1.27 $10^{-4}$</td>
<td>1.12 $10^{-4}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(5,10)</th>
<th>(10,10)</th>
<th>(15,10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>9.45 $10^{-4}$</td>
<td>1.51 $10^{-3}$</td>
<td>1.28 $10^{-3}$</td>
</tr>
<tr>
<td>(2)</td>
<td>4.37 $10^{-4}$</td>
<td>6.78 $10^{-4}$</td>
<td>5.93 $10^{-4}$</td>
</tr>
<tr>
<td>(3)</td>
<td>4.28 $10^{-4}$</td>
<td>7.02 $10^{-4}$</td>
<td>5.76 $10^{-4}$</td>
</tr>
<tr>
<td>(4)</td>
<td>9.00 $10^{-5}$</td>
<td>1.42 $10^{-4}$</td>
<td>1.22 $10^{-4}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(5,5)</th>
<th>(10,5)</th>
<th>(15,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>6.07 $10^{-4}$</td>
<td>9.43 $10^{-4}$</td>
<td>8.01 $10^{-4}$</td>
</tr>
<tr>
<td>(2)</td>
<td>2.73 $10^{-4}$</td>
<td>4.29 $10^{-4}$</td>
<td>3.70 $10^{-4}$</td>
</tr>
<tr>
<td>(3)</td>
<td>2.72 $10^{-4}$</td>
<td>4.28 $10^{-4}$</td>
<td>3.59 $10^{-4}$</td>
</tr>
<tr>
<td>(4)</td>
<td>5.67 $10^{-5}$</td>
<td>8.90 $10^{-5}$</td>
<td>7.53 $10^{-5}$</td>
</tr>
</tbody>
</table>

The average absolute errors calculated were

(2) $3.33 \times 10^{-4}$, and (3) $3.30 \times 10^{-4}$.

**Experiment 1(b)**

This experiment was just a repeat of experiment 1(a), where in this case we took (5.75) with (5.75b). The comparative accuracy of the methods followed the same trend as in the previous experiment.
Hence, we just give the average absolute errors (at \( T=1.0 \)) for methods (2) and (3):

(2) \( 3.51 \times 10^{-4} \), and (3) \( 3.48 \times 10^{-4} \).

(viii) Comparison of amounts of work

In section 2.10, we gave a guide as to the relative speeds of the odd-even and line hopscotch, and the Peaceman-Rachford procedure. McGuire (1970) pp.55-56 showed by investigating the number of arithmetic operations (multiplications and divisions) required at each point in their (Courlay and McGuire) program for equation (3.75), that line hopscotch was about 3 times slower, and the Peaceman-Rachford method was about \( \frac{4}{3} \) times slower than the odd-even hopscotch process.

The peripheral hopscotch method is similar in design to the line hopscotch process. Hence, we shall compare these two for relative amounts of work, and thus put the peripheral hopscotch procedure in perspective in this respect. Over a double step \( (2\Delta t) \) both methods solve half the points explicitly, and the other half implicitly along lines or peripherals. Since equations (2.29) and (5.15) are the same for both methods, the solution at points using this explicit formula require the same amount of work in each case. For the remaining points we use Algorithm 4.2 and Algorithm 4.3 to solve the systems of equations along alternate lines and peripherals, respectively. From Chapter 4, we know that the latter requires about twice as many arithmetic operations as the former algorithm, to obtain the solution vector. Totalling the arithmetic operations required by both methods over a double step \( 2\Delta t \), to obtain the vector \( u^{2k} \) \( (k=1,2,\ldots) \), it can be seen that the line hopscotch will be roughly \( 1\frac{1}{2} \) times as fast as the peripheral hopscotch procedure (assuming the same number of internal mesh points).

As far as storage is concerned, in view of the nature of Algorithms
4.2 and 4.3, both can be executed using a working space of only one vector, e.g. see Evans and Atkinson (1970). Thus storage is about the same for both methods.

(ii) Applicability of peripheral hopscotch to open regions

From the information gathered so far on the comparative accuracy, the speed and storage of the peripheral hopscotch with other methods mentioned, it would appear that for the problem posed in subsection 5.2(1) (where \( R \) is a closed plane region), the former has no advantages over the line hopscotch, say. However, if the solution to a parabolic equation (such as (5.1)) were required over an open region \( R \), in the \( x-y \) plane, subject to initial and boundary conditions like those in equations (5.4), then peripheral hopscotch would seem to be an ideal method.

The use of "semi-peripheral" orderings on open plane regions such as \( x \geq 0, y \geq 0 \) and \(-\infty < x < \infty, y \geq 0 \) (as illustrated in Figures 5.3 and 5.4, respectively) has indeed been suggested in connection with hopscotch methods by Gourlay (1973). Also, a peripheral need not be square in shape, but could be the shape of a boundary which has sides parallel to the \( x- \) and \( y- \) axes, as illustrated in Figure 5.6. For a more general open region in two space dimensions, with an irregular boundary, however, the standard "square-shaped" peripheral would seem to be the best choice of ordering the mesh points.

Having decided on such an ordering of the unknown mesh points for open regions, the computation could proceed as follows. Using the explicit formula (5.16) at the first time level and (5.15) at subsequent time levels, we would compute the finite difference solution [to (5.1) say] along alternate "semi-peripherals" (as in the case of Figures 5.3 and 5.4) or along alternate peripherals (see Figures 5.5 and 5.6). In the figures, we indicate contours (peripherals/semi-peripherals) with the same \( \rho^k \) value (see equation (5.13)) by the same
symbol. Thus we only require to use two symbols "x" and "0". It is
immaterial which of these corresponds to the value of \( \theta_p^k = 1 \), and which
to the value \( \theta_p^k = 0 \). We then compare the solutions on successive
alternate peripherals until we reach a region of "negligible change".
Then in keeping with the peripheral hopscotch process (see subsection
3.2(iii)), the solution (at the same time level) at mesh points on
the remaining alternate peripherals (semi-peripherals) would be
calculated implicitly using the implicit formula, viz. (5.14) with
\( \theta_p^{k+1} = 1, \theta_p^k = 0 \). In the case of the peripheral (e.g. Figures 5.5 and 5.6)
this would necessitate the solution of periodic tri-diagonal systems
of equations by Algorithm 4.3, while in the case of the semi-peripheral
(see Figures 5.3 and 5.4), we would be required to solve tri-diagonal
systems by Algorithm 4.2. If then we had let the "x" contours
correspond to the explicit method at one time level, then the "0"
contours correspond to the implicit method. By definition of the
hopscotch process, the roles of these two sets of contours will be
reversed at the following time level. This two step process continues
until the desired solution at some time \( t = T \) is reached.

Clearly, for an implicit formula, with an ordering of the mesh
points parallel to one or other of the coordinate axes in Figures 5.3
-5.6, a "point of negligible change" in the solution to (5.1) would
have to be guessed at prior to the start of the calculation. This
indeed does not appear to be such a neat way of tackling the problem
as the peripheral hopscotch approach.

5.3 An explicit finite difference formula for solving heat conduction
problems in polar coordinates (x=9)

In this section we will be concerned with the numerical
solution of the equation

\[
\frac{2u}{\delta r} = \frac{3u}{\delta r^2} + \frac{1}{r} \frac{3u}{\delta r} + \frac{1}{r^2} \frac{3u}{\delta \theta^2} \\
= \frac{1}{r} \left( \frac{3}{9} \left( \frac{2u}{\delta \theta} \right) + \frac{3}{9} \left( \frac{1}{r} \frac{3u}{\delta \theta} \right) \right), u = u(r, \theta, t)
\]

(5.76)
in the cylinder $\mathbb{R} \times [0,\pi]$, where $t$ denotes time, and the region $R$ is either the unit circle ($0 \leq r \leq 1$) or the ring ($1 < r < 2$). We will be given initial data of the form

$$U(r,0) = f(r,t),$$

(5.77)

where $f(r,0)$ is given for all $(r,0) \in \mathbb{R}$, and boundary data of the form

$$U(1,0,t) = h_1(0,t), \quad \text{OstgT},$$

(5.78)

and in the case of the ring ($1 < r < 2$), we will also be given

$$U(2,0,t) = h_2(0,t), \quad \text{OstgT}.$$  

(5.79)

The functions $h_1(0,t), h_2(0,t)$ are prescribed for all $0 \leq t \leq 2\pi$.

With the above conditions, let us firstly consider the solution of (5.76) when $R$ is the ring ($1 < r < 2$). Next let us superimpose a polar grid on the region and adopt a peripheral ordering of the mesh points as illustrated in Figure 5.7.

We shall denote the peripheral of unknown grid points adjacent to the boundary $r=a$, as the first peripheral, and number the remaining peripherals in the direction of $r$ increasing. The mesh points in the $r-\theta$ plane are the points of intersection of the circles (peripherals) $r_i = i \Delta r$ for $i = i_{\min}, i_{\min} + 1, \ldots, i_{\min} + m - 1 = i_{\max}$, where
\[ a = (i_{\text{min}} - 1) \Delta r \]  
and  
\[ 1 = (i_{\text{max}} + 1) \Delta r, \]  
(i.e. the radius of the first and last peripherals are \( i_{\text{min}} \Delta r \) and \( i_{\text{max}} \Delta r \), respectively) and the straight lines \( \theta_j = j \Delta \theta, j = 0, 1, \ldots, n - 1 \) where

\[ \Delta \theta = \frac{2\pi}{n}. \]  

Hence in the case of Figure 5.7, we have \( m = 3, n = 3 \) and \( \Delta r = (1 - a)/(a + 1) \).

Adopting a time step of size \( \Delta t \), let us consider equation (5.76) to be located at time \( t = k \Delta t (k > 0) \). Using analogous central difference formulae to those in equations (2.6) and (2.9), we can derive the following explicit finite difference analogue of (5.76) at the mesh point \( (r_i, \theta_j, t_k) = (i \Delta r, j \Delta \theta, k \Delta t) \), where \( u_{i,j}^k \) denotes the approximate solution \( u(r_i, \theta_j, t_k) \) of the exact solution \( U(r_i, \theta_j, t_k) \):

\[
\frac{u_{i,j}^{k+1} - u_{i,j}^k}{\Delta t} = \left\{ \begin{array}{c}
\frac{(1 - 1/(2l))u_{i-1,j}^k - 2u_{i,j}^k + (1 + 1/(2l))u_{i+1,j}^k}{\Delta r^2} \\
\frac{u_{i,j}^k}{(r_i \Delta \theta)^2}
\end{array} \right. 
\]  

(5.82)

Multiplying through by \( \Delta t \) in (5.82), substituting \( s_1 = \Delta t/\Delta r^2 \), \( s_2 = \Delta t/(r_i \Delta \theta)^2 \) and rearranging, we obtain

\[
u_{i,j}^{k+1} = \left[ 1 - (2s_1 + 2s_2) \right] u_{i,j}^k - \left[ \frac{1 - 1/(2l)}{s_1} u_{i-1,j}^k + \frac{1 + 1/(2l)}{s_1} u_{i+1,j}^k \right] + s_2 u_{i,j-1}^k + s_2 u_{i,j+1}^k + i_{\text{min}} s_i + i_{\text{max}} s_i + 1. \]

(5.83)

The global formulation of equations (5.83) on the ring \((a \leq r \leq b)\), with the above ordering of the internal mesh points, can be written

\[ u^{k+1} = (I - A) u^k + b^k, \]  

(5.84)

where \( b^k \) is independent of the \( u^j \)'s, \( u^k (k > 0) \) is the vector of approximate solution values \( u_{i,j}^k \) and

\[ a = (i_{\text{min}} - 1) \Delta r \]  
and  
\[ 1 = (i_{\text{max}} + 1) \Delta r, \]
Let us consider now the case when $R$ is the unit circle. For reasons which will become apparent later in the Chapter, we choose not to solve equation (5.76) at the centre point of the circle. This is achieved in the following manner. We arrange our "Ax mesh lengths" along each diameter, so that the centre point (of the circle) lies at the midpoint of a "Ax mesh length", e.g. see Figure 5.8.
This means that if \( n \) is the number of internal peripherals, then \( \Delta r = 2/(2n+1) \). For this region, the first peripheral is of radius \( \Delta r/2 \), so that since \( r_{i,\text{min}}^1 = \min \Delta r, i_{\text{min}} = 1 \). Hence using the explicit formula (5.83) on this peripheral, we see that the coefficient of \( u_{i-1,j}^1 \) (namely \( (1-1/r_{i,\text{min}}^1) \)) = 0. Then, ordering the grid points as in Figure 5.8 (which is the same ordering as that given in Figure 5.7) and going over them in that order with formula (5.83), gives us an analogous system of equations to (5.84), where in this case \( i_{\text{min}} = 1 \) (which from (5.80) is equivalent to defining \( s = -\Delta r/2 \)). Hence for the circle, we have \( s_1 = \Delta t/\Delta r^2 \) and \( s_2 = \Delta t/(r_{i,\text{min}}^1 \Delta \theta)^2 = \Delta t/(\Delta r \Delta \theta)^2 \).

**Stability**

It follows from Condition 2.4, that the process (5.84) will be stable if

\[
||I-A|| \leq 1.
\]  
(5.86)

Letting \( || \cdot \|| \) in (5.86) denote \( || \cdot ||_\infty \) from (1.26) we have

\[
||I-A|| = \max_i \left| 1 - 2(s_1 + s_2^i) \right| \leq 2(s_1 + s_2^i) \leq 1,
\]

if \( \max_i \left( 4(s_1 + s_2^i) \right) \leq 2. \)
Since \( \max \left( s_2^i \right) = \max(\Delta t/(r_1 \Delta \theta)^2) = s_{2 \min}^i \) (i.e. \( r_1 = \min r_1 \)), letting \( s_2 = s_{2 \min} \), the stability condition is

\[
s_1 + s_2 \leq 1 \quad (5.87)
\]

When \( R \) is the ring (agr1), (5.87) is equivalent to

\[
\frac{\Delta t}{\Delta r^2} \leq \frac{\Delta t}{[(a+\Delta r)\Delta \theta]^2} \leq 0.5 \quad (5.88)
\]

and when \( R \) is the circle (agr1), (5.87) becomes

\[
\frac{\Delta t}{\Delta r^2} \leq \frac{4\Delta t}{[\Delta r \Delta \theta]^2} \leq 0.5 \quad (5.89)
\]

These two stability conditions are a severe drawback on the use of the explicit formula (5.83), since for \( \Delta r = 0.1 \) (say) it would take at least 200 time steps to advance one unit in time. In fact, this figure would be considerably greater on the circle, in view of (5.89).

**Convergence**

Let

\[
e_{1, j}^k = u_{1, j}^k - u_{1, j}^k \quad (5.90)
\]

denote the difference between the theoretical solutions of the differential and difference equations at the grid point \( (r_{1,0}, t_k) \). Using Taylor series, we have from (5.76), (5.83) and (5.90)

\[
e_{1, j}^{k+1} = e_1^{(1-1/(2!))}e_{1, -1, j}^k + e_1^{(1+1/(2!))}e_{1, +1, j}^k + s_2^1(e_{1, -1, j}^k - e_{1, +1, j}^k)
\]

\[
+(1-2(s_2^i + s_2^j))e_{1, j}^k + \frac{\Delta t}{2} \left( \frac{\partial^2 u}{\partial r^2} \right)_{1, j} - \Delta t \cdot \frac{\Delta r}{12} \left[ \frac{\partial^4 u}{\partial r^4} + \frac{2}{r_1} \frac{\partial^3 u}{\partial r^3} \right]_{1, j}
\]

\[
- \Delta t \cdot \frac{\Delta \theta}{12} \left( \frac{\partial^4 u}{\partial \theta^4} \right)_{1, j} + \ldots \ldots . \quad (5.91)
\]

Letting \( d_{1, j}^k \) denote the local truncation error in (5.91) at the grid point \( (r_{1,0}, t_k) \), we have the global error equation as

\[
e_{1, j}^{k+1} = (I-A)e_{1, j}^k + d_{1, j}^k \quad (5.92)
\]

It is then easily shown that

\[
e_{1, j}^k = (I-A)^{k+1}0 + (d_{1, j}^{k-1} + (I-A)d_{1, j}^{k-2} + \ldots + (I-A)^{k-1}d_{1, j}^0)
\]
Assuming that $||.||$ denotes $||.||_\infty$, we have

$$||e^k||\leq ||e^0||+\max_{0\leq k \leq 1} ||d^k||.$$  \hspace{1cm} (5.93)

Assuming the stability of the explicit formula (5.83), i.e. assuming (5.86), we have from (5.93)

$$||e^k||\leq \max_{0\leq k \leq 1} ||d^k||.$$  \hspace{1cm} (5.94)

Next from (1.20), we have $||d^k||=\max_{i,j}d^k_{i,j}$ and $||e^k||=\max_{i,j}e^k_{i,j}$.

From equation (5.91),

$$d^k_{i,j} = \frac{\Delta t^2}{2} \left\{ \frac{3}{2} + \frac{\Delta t}{\Delta x} \right\} \left\{ \frac{\Delta t^2}{2} + \Delta t \right\} \left( \frac{\Delta t^2}{2} + \Delta t \right).$$  \hspace{1cm} (5.95)

Let us assume that $G$ is an upper bound for the modulus of the partial derivatives in $x, y$ and $t$ in (5.95), for all $i, j, l$. Then

$$\max_{i,j,l} |d^k_{i,j}| \leq \max_{i,j,l} \left\{ \frac{\Delta t^2}{2} + \Delta t \right\} G,$$

and hence

$$\max_{i,j,l} |d^k_{i,j}| \leq G \left[ \frac{\Delta t^2}{2} + \Delta t \right].$$  \hspace{1cm} (5.96)

Thus it follows from (5.94) and (5.96) that

$$||e^k|| \leq G \cdot t_k \left[ \frac{\Delta t^2}{2} + \Delta t \right].$$  \hspace{1cm} (5.97)

Let us now examine the convergence of this method for our two plane regions, $R$. For the first region, i.e. the ring (or oval),

$$r_{i,\min}^2 = \Delta x + \Delta r,$$

from (5.80). Since $a$, the radius of the inner boundary, is a fixed positive number and $r_{i,\min}^2 > a$, the term on the right of the inequality in (5.97) clearly tends to zero as $\Delta t \to 0$, $\Delta r \to 0$, and $\Delta t \to 0$. Hence assuming stability, by definition 2.5 we have
that formula (5.83) for the solution of (5.76) is convergent when \( R \) is the ring.

In the case of the circle, \( \min r_i = 1 \), so that \( r_i = \min r_i = \Delta R/2 \).
Substituting this value into (5.97) gives
\[
||x^k|| \leq C \cdot \frac{\Delta t}{2} + \frac{\Delta x^2}{12} + \frac{\Delta x^3}{3} + \frac{\Delta \theta^2}{3\Delta x^2} \quad \text{(5.98)}
\]

Hence the term on the right of (5.98) will tend to zero as \( \Delta t \to 0 \),
\( \Delta R \to 0, \Delta \theta \to 0 \) if and only if \( (\Delta \theta/\Delta x) \to 0 \). Thus if formula (5.83) is stable
it will be convergent if the last condition holds. It is of course a
very stringent condition, since it implies a very large number of
internal mesh points as compared to a comparative case on the ring.
Also, it now makes the stability condition (5.89) totally restrictive.

Eisen (1966) gives consideration to the stability and convergence
of the explicit difference analogue of \( \partial U/\partial t = \partial^2 U/\partial x^2 + (1/r) \partial U/\partial r \),
\( U=U(r,t), \) \( 0 \leq R \leq r \leq T \). He shows that the process is unstable in the
\( L_2 \)-norm. However, he shows that for grid points \( x_i = i\Delta x \), \( i=0, 3/2, \ldots, m \),
the explicit process is stable and convergent in the \( \infty \)-norm, provided
\( s_i = \Delta x/\Delta x^2 \). It was in fact from the above paper, that we discovered
such an arrangement of the mesh points along each radius of our polar
grid for the "circle problem" discussed above. Smith (1965) has
considered the same problem with the exception that he utilised the
lattice: \( x_i = i\Delta x \), \( i=0, 1, \ldots, m \), for which he was forced to introduce the
condition \( \partial U/\partial r \to 0 \) as \( r \to 0 \). Because of the above restriction on \( s_i \),
authors such as Albasiny (1960) make use of the unconditionally stable
(for all \( s_i > 0 \)) Crank-Nicolson method to approximate the solution of
the above heat conduction problem.

Likewise to Albasiny, because of the severe restrictions on the
mesh ratios \( s_1, s_2 \), we shall turn to implicit methods of solution
(namely Peaceman-Rachford and Hopscotch) which impose far less severe
restrictions on the size of \( s_1 \) and \( s_2 \) for stability and convergence,
particularly as far as the solution of equation (3.76) on the ring
\( (s \ll s_1) \) is concerned.
5.4 The Peaceman-Rachford method in polar coordinates

In this section our main attention will be focused on the finite difference solution of the two problems defined by (5.76)-(5.79), using the Peaceman-Rachford Alternating Direction Implicit method on a polar grid. The method is easily defined for a more general problem in view of the presentation given in section 2.4.

(4) The finite difference equations

We can in fact derive an analogous formula to (2.19), for the approximate solution of

\[ \frac{\partial u}{\partial t} = L \circ g(r, \theta, t), \quad u = u(r, \theta, t), \]  

(5.99)

where \( L \) is a second order linear, elliptic differential operator in the space variables \( r \) and \( \theta \). Let also (this time) \( L^{(1)}, L^{(2)} \) be operators involving only derivatives with respect to \( r, \theta \) respectively, such that \( L = L^{(1)} + L^{(2)} \). Define \( L_{\Delta r}\Delta \theta \) as the three point difference replacements in the \( r \) and \( \theta \) directions of \( L^{(1)}, L^{(2)} \) respectively, and \( L_{\Delta r} = L^{(1)} + L^{(2)} \) (since \( L = L^{(1)} + L^{(2)} \)) as the finite point difference analogue of \( L \). In particular then, for equation (5.76) we have

\[
L^{(1)}u = \frac{1}{r^2} \frac{\partial}{\partial r} (ru) + \frac{1}{\Delta r} \frac{\partial^2 u}{\partial \theta^2} \]

\[
L^{(2)}u = \frac{1}{\Delta \theta} \frac{\partial u}{\partial \theta} + \frac{1}{\Delta r} \frac{\partial^2 u}{\partial r^2} \]

(5.100)

Both these equations can be obtained from equation (3.2) with \( r \) and \( \theta \) replacing \( x \) and \( y \), respectively.

In the remainder of this chapter, to simplify the notation, we redefine the quantity \( r_j \) as \( r_j = r_{\Delta j} \), so that on the ring (say \( R_j \)), \( r_j \) now equals the radius of the \( j \)th peripheral. Similarly, with \( \Delta r/2 \), \( r_j = (j-\Delta j) \Delta r \) denotes the radius of the \( j \)th peripheral on the unit circle. 

Likewise, the value \( u_{k, j} \) now denotes \( u(r_j, j\Delta \theta, k\Delta \tau) \).

We now describe the general implicit formula similar to (2.19) for
the approximate solution of (5.99); namely
\[ u_{i,j}^{k+1} = u_{i,j}^k + \Delta t \left( \frac{\partial \phi}{\partial x} u_{i,j}^k + \eta_{i,j} \right) + \Delta t \left( \frac{\partial \phi}{\partial y} u_{i,j}^k + \eta_{i,j} \right) \]
\[ + \Delta t \left( \frac{\partial \phi}{\partial x} u_{i,j}^k + \eta_{i,j} \right) + \Delta t \left( \frac{\partial \phi}{\partial y} u_{i,j}^k + \eta_{i,j} \right) \]
\[ \quad + \Delta t \left( \frac{\partial \phi}{\partial x} u_{i,j}^k + \eta_{i,j} \right) + \Delta t \left( \frac{\partial \phi}{\partial y} u_{i,j}^k + \eta_{i,j} \right) \]
(5.101)
with the restrictions
\[ \frac{u_{i,j}^{k+1} + u_{i,j}^k}{2} = 1 \]
\[ \frac{\eta_{i,j}^{k+1} + \eta_{i,j}^k}{2} = 1. \]

for (5.101) to be a consistent approximation to (5.99). In (5.101) we have used the splitting \( g = g^{(1)} + g^{(2)} \). If we substitute the values of \( u_{i,j}^k, \eta_{i,j}^k \) from (2.23) into (5.101), we obtain the local formulation of the Peaceman-Rachford method in polar coordinates, for the solution of (5.99), viz.
\[ u_{i,j}^{k+1} = u_{i,j}^k + \Delta t \left( \frac{\partial \phi}{\partial x} u_{i,j}^k + \frac{\partial \phi}{\partial y} u_{i,j}^k \right) \]
(5.102a)
when \( k \) is even, and
\[ u_{i,j}^{k+1} = u_{i,j}^k + \Delta t \left( \frac{\partial \phi}{\partial x} u_{i,j}^k + \frac{\partial \phi}{\partial y} u_{i,j}^k \right) \]
(5.102b)
when \( k \) is odd. We are assuming a time step of \( 2\Delta t \) over the two-step process.

Let us define
\[ \delta^2 u_{i,j}^k = \left( r_i - \frac{\Delta r}{2} \right) u_{i,j}^k - 2 \frac{\Delta r}{1} u_{i,j}^k + \left( r_i + \frac{\Delta r}{2} \right) u_{i,j}^k \]
(5.103)
and
\[ \delta^2 u_{i,j}^k = u_{i,j}^k - 2 \frac{\Delta r}{1} u_{i,j}^k + u_{i,j}^k + 1. \]

Then the P-R process (5.102a,b) for the solution of (5.70) may be written as
\[ (1 - \delta^2)^{\alpha} u_{i,j}^{k+1} = (1 + \delta^2)^{\alpha} u_{i,j}^k \]
(5.104a)
\[ (1 - \delta^2)^{\alpha} u_{i,j}^{k+1} = (1 + \delta^2)^{\alpha} u_{i,j}^k \]
(5.104b)
where \( u_{i,j}^{(k+1)^a} = u_{i,j}^{k+1} \) is regarded as an intermediate solution (for all \( i,j \)) and \( \alpha_1 = \Delta t/\Delta x^2, \alpha_2 = \Delta t/(r_i \Delta \theta)^2 \).
(ii) The computational procedure

Let us now consider the computational procedure for solving equation (5.76), using the Peaceman-Rachford method given by equations (5.104). Firstly, let us look at the case when the region $R$ is the ring $(a \leq r \leq 1)$, since the process will be shown to be similar for the circle.

Firstly, we superimpose a polar grid on the region, and for the solution of equation (5.104) at each mesh point, we order the unknown grid points as in Figure 5.7. To obtain the solution values on the $(i)$th peripheral at some time, $t_{k+1} = (k+1)\Delta t$, $k=0, 2, 4, \ldots$, requires the implicit solution of periodic tri-diagonal systems of equations of the form

$$A_i u_{i}^{k+1} = b_i^k, \quad i=1, 2, \ldots, m,$$

(5.105)

where $u_{i}^{k+1}$ denotes the vector of unknown solution values $u_{i,j}^{k+1}$ on the $(i)$th peripheral,

$$[A_i]_{i,j} = -s_{2,i}^j u_{i-j-1}^{k+1} + (1+2s_{2,i}^j)u_{i,j}^{k+1} - s_{2,i}^j u_{i,j+1}^{k+1}, \quad j=0, 1, \ldots, n-1,$$

and

$$b_i^k = (1+\frac{1}{r_i^2} \delta_i^2)u_{i,0}^k, \quad j=0, 1, \ldots, n-1.$$

(5.106)

Since each $A_i$ is symmetric and strictly diagonally dominant with positive diagonal entries, it is positive definite by Theorem 1.5,
Corollary. Hence we can use Algorithm 4.3 to solve system (5.105) for each value of $i=1(l)m$.

It is important at this point to notice that on this $(k+1)\text{th}$ or intermediate time level, the boundary data is not used. This is apparent since there are no terms involving $u_{i,j}^{k+1}$ in equation (5.106). For this reason, it is evident that it is unnecessary to compute intermediate boundary values as described in section 2.5.

To solve equations (5.104b) we re-order the grid points along successive radii, commencing with the line $j=0 (t=0)$. This is illustrated in Figure 5.9. To obtain the solution values on the $(k+2)\Delta t$, $(j)\text{th}$ radius at some time $t_{k+2} = /k=0,2,4\ldots$, then requires the implicit solution/tridiagonal systems of equations of the form

$$
c_j u_{j}^{k+2} = u_{j}^{k+1}, \quad j=0,1,\ldots,n-1
$$

(5.107)

where $u_{j}^{k+2}$ denotes the vector of solution values $u_{i,j}^{k+2}$ on the $(j)\text{th}$ radius, and

$$
[C_j u_{j}^{k+2}]_{1,j} = -s_{1x_1}^{-1}(x_1^2 - \Delta x/2)u_{1-1,j}^{k+2} + (1+2s_{1x_1})u_{1,j}^{k+2} - s_{1x_1}^{-1}(x_1^2 + \Delta x/2)u_{1+1,j}^{k+2}
$$

for $i=1,2,\ldots,n$.

Since $C_j$ ($0 \leq j \leq n-1$) is strictly diagonally dominant, we can use Algorithm 4.2 to solve systems (5.107).

This alternating direction implicit process is continued until the solution is obtained at some desired time, $t=T$, say, where $T$ is an even multiple of $\Delta t$.

At this stage it should be pointed out, that in fact the second step, equation (5.104b), can be made more efficient with respect to arithmetic operations. If we add (5.104a) to (5.104b) to eliminate the term $\left[s_{2x_1}^{-1} u_{1,j}^{k+1}\right]$ the result

$$
(1-s_{1x_1}^{-1} \cdot 2^2 s_{x_1}^2)u_{1,j}^{k+2} = 2u_{1,j}^{k+1} - (1+s_{1x_1}^{-1} \cdot 2^2 s_{x_1}^2)u_{1,j}^{k}
$$

(5.108)
is obtained. Hence, this revised version requires less arithmetic operations than (5.104b), since the quantity \((1 + s_i \frac{1}{r_i^2} \delta_i^2)^k u_{i,j}^k, (i,j) \in R\), in (5.108) has already been calculated in (5.104a). An analogous formula has been formulated by Warga (1962) p.240, for the equation \(3U/3t = 2U/3x^2 + 2U/3y^2\), and by Fairweather and Mitchell (1967) p.164. The equations (5.108) will be solved in the same way as equations (5.104b), although of course, the vector \(d_{k+1}^j\) in (5.107) will be defined differently in each case. Thus we propose the use of equations (5.104a) and (5.108) rather than the pair (5.104a) and (5.104b), since the former will be more efficient.

Next let us consider problem (5.76)-(5.78), \([\text{viz. } R \text{ is the unit circle}],\) and its approximate solution by the P-R method, (5.104a), (5.108). As in section 5.3, we choose not to include the centre point of the circle in our calculation (at least, not until the other chosen grid points have been solved for). We arrange our mesh as illustrated in Figure 5.8, for the case \(\Delta r = 2/7, \Delta \theta = \pi/4\). As also was shown in section 5.3, for this region, \(i_{\min} = \frac{\pi}{4}\) giving \(a = -\Delta r/2\) from (5.80). In fact, if \(m\) is the number of peripherals, then the centre point is not a mesh point if \(\Delta r = 2/(2m+1)\), assuming a constant mesh size \(\Delta r\). Hence if \(u_{i,j}^k\) lies on the first peripheral, from (5.103) we have

\[
2 \delta r_{1,j} u_{1,j}^k = -\Delta r_{1,j} u_{1,j}^k + \Delta r_{2,j} u_{2,j}^k \quad (\text{since } r_{1} = \Delta r/2)
\]

In view of this equation, we can order the unknown grid points (see Figure 5.8) in the same way as for the ring to solve equations (5.108a).

Systems of the form (5.105) are then obtained, for \(i=1,2,\ldots,m\).

To solve equations (5.108) at the second step, we re-order the points as shown in Figure 5.10.
Likewise this gives us systems of the form (5.107). As with the ring, we use equation (5.104a) at the intermediate time level, since this avoids the use of boundary correction.

**Computing the solution of (5.76) at the centre point**

Suppose at some time \( T=k\Delta t \) (\( k \) an even integer), we require the solution to (5.76) at the centre of the circle. At this point, \( r=0 \) and hence the right side of equation (5.76) appears to contain singularities. To overcome this difficulty, at \( r=0 \) we replace this equation by its cartesian equivalent, namely,

\[
\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad u = u(x, y, t).
\]

(see, for instance, Smith (1965) p.44). Then, providing \( n/2 = \pi/\Delta \theta \) is even, we can approximate (5.110) at \( r=0 \), with the following Crank-Nicolson formula [see equations (2.19) and (2.22)]:

\[
(1+16s_1)u_{0,j}^{k+2} - 4s_1(u_{-1,j}^{k+2} + u_{1,j}^{k+2} + u_{0,n/4+j}^{k+2} + u_{0,3n/4+j}^{k+2})
\]

\[
= (1-16s_1)u_{0,j}^{k} + 4s_1(u_{-1,j}^{k} + u_{1,j}^{k} + u_{0,n/4+j}^{k} + u_{0,3n/4+j}^{k}), \quad (5.111)
\]

which has a local truncation error of \( O(\Delta t^3 + \Delta t^2 + \Delta x^2) \). Since this equation holds for all \( j=0, 1, 2, \ldots, n-1 \), a better formula may be obtained by adding the \( n \) resulting equations to give
\[(1+16s_1)u_{0,j}^{k+2} - 16s_1u_{1,j}^{k+2} = (1-16s_1)u_{0,j}^k + 16s_1u_{1,j}^k \]  \hspace{1cm} (5.112)

where \(u_{1,j}^k = \frac{1}{n+1} \sum_{j=0}^{n} u_{1,j}^k\). Since, we know a priori, the values \(u_{0,j}^k, u_{1,j}^k\) and \(u_{0,j}^{k+2}, u_{1,j}^{k+2}\), formula (5.112) is essentially explicit.

\((iv)\) Stability

We now consider the stability of the two-step process (5.104), viz. the Peaceman-Rachford finite difference replacement for the problem(s) defined by equations (5.76)-(5.79).

Using equations (5.103), and a peripheral ordering of the grid points, we define
\[
(bu^k)_{i,j} = -\frac{r^2}{r_i} \cdot (\delta^2 u^k)_{i,j} \; ; \\
(vu^k)_{i,j} = -(\delta^2 u^k)_{i,j} \; ;
\]  \hspace{1cm} (5.113)

and let the diagonal matrix \(V\) be defined by
\[
(vu^k)_{i,j} = r_i^2 u^k_{1,i,j} \; .
\]

In each case, \(i=1,2,\ldots,n\) and \(j=0,1,\ldots,n-1\). Let us assume a constant time increment of \(2\Delta t\). Then the global formulation of equations (5.104a,b) can be written as
\[
\begin{bmatrix} I+s_2 \left( V^{-1} W \right) \end{bmatrix} u^{(k+1)a} = \left[ I-s_1 \left( V^{-1} W \right) \right] u^k + \Delta t \cdot b^k \]
\[
\begin{bmatrix} I+s_1 \left( V^{-1} W \right) \end{bmatrix} u^{k+2} = \left[ I-s_2 \left( V^{-1} W \right) \right] u^{(k+1)a} + \Delta t \cdot b^{k+2} \; ,
\]  \hspace{1cm} (5.114)

where \(s_2 = \frac{s_1}{\Delta t / (r_1 \Delta t)^2}, \quad s_1 = \Delta t / \Delta x^2\), and the \(b^k\)s contain values obtained from the boundary(s) of the region.

Eliminating \(u^{(k+1)a}\) from (5.114) gives
\[
u^{k+2} = \tau \cdot u^k + b^{(k+1)a} \; , \quad k=0,2,4,\ldots., \hspace{1cm} (5.115)
\]

where \(b^{(k+1)a}\) is independent of the \(u^k\)s and where
\[
\tau = V^{-1} \left( I+s_1 \bar{V} \right)^{-1} \left( I-s_2 \bar{V} \right) \left( I+s_2 \bar{V} \right)^{-1} \left( I-s_1 \bar{V} \right) F \; .
\]  \hspace{1cm} (5.116)

In (5.116), we have
\[
\bar{V} = V^{-1} \bar{V} \; , \quad \bar{U} = V^{-1} \bar{U} \; .
\]  \hspace{1cm} (5.117)
If
\[
\bar{T} = (I-s_2 \bar{H})(I+s_2 \bar{H})^{-1}(I-s_1 \bar{V})(I+s_1 \bar{V})^{-1},
\]
(5.118)
then
\[
T = F^{-1}(I+s_1 \bar{V})^{-1}.\bar{T}.(I+s_1 \bar{V})F^{-1},
\]
(5.119)
and
\[
T^k = F^{-1}(I+s_1 \bar{V})^{-1}.\bar{T}^k.(I+s_1 \bar{V})F^{-1}.
\]

It is constructive at this point to investigate the properties of \( \bar{H} \) and \( \bar{V} \). This, as will be shown below, is equivalent to doing the same thing for \( H \) and \( V \). Since we are using a peripheral ordering of the unknown mesh points, from (5.113) we have

\[
H = \begin{bmatrix}
H_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & H_m
\end{bmatrix}_{(m \times m)}
\]

\[
\begin{bmatrix}
2 & -1 & \cdots & 0 \\
-1 & 2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 2
\end{bmatrix}_{(n \times n)}
\]

(5.120)

\[
V = \begin{bmatrix}
D_1 & E_1 & \cdots & 0 \\
E_1 & D_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & D_m
\end{bmatrix}_{(m \times m)}
\]

where
\[
D_i = \begin{bmatrix}
2x_i \\
\vdots \\
-1 \\
x_i
\end{bmatrix}_{(i=1,2,\ldots,m)}
\]

(5.121)

and
\[
E_i = \begin{bmatrix}
(x_i + \Delta x/2) \\
\vdots \\
-(x_i + \Delta x/2)
\end{bmatrix}_{(i=1,2,\ldots,m-1)}
\]

(5.211)

By a simple extension to Theorem 1.1, Corollary, we see that each \( H_i \) (1≤i≤m) is irreducible. It then follows from Theorem 1.6, that
$H_i$ $(1 \leq i \leq m)$ has a zero eigenvalue, and so by Theorem 1.2 is non-negative definite. Likewise, from Theorem 1.1, it can be shown that $V$ is irreducible. Theorem 1.5, Corollary then tells us that $V$ is positive definite. Now from (5.117),
\[
\bar{H} = (V_i V_i^{-1})^T (V_i V_i^{-1})^{-1}, \quad \bar{V} = (V_i V_i^{-1})^T (V_i V_i^{-1})^{-1},
\]
thus $\bar{H}$ and $\bar{V}$ are symmetric, and from Theorem 1.3 it follows that they are respectively non-negative and positive definite. It should be stressed, that the above properties of $\bar{H}$ and $\bar{V}$ hold for both regions (i.e. the ring and the circle), in the context of problem (5.76)-(5.79).

Defining the $H_2$ vector norm by (5.27) and the positive diagonal matrix $F$ by (5.113), we define the $F^i$-vector norm as
\[
||u||_{F^i} = ||F^i u||_{H_2}.
\]
It is then easily verified that the subordinate $F^j$-matrix norm is
\[
||c||_{F^j} = ||F^i c F^{-1}||_2.
\]
We now intend to prove the stability of the Peaceman-Rachford formula (5.104a), (5.104b) in this norm, and until stated otherwise, $||-||$ denotes $||-||_{F^i}$.

Hence, from (5.119) we have
\[
||T^k|| = \|[I + s_i \bar{V}]^{-1} T^k (I + s_i \bar{V})\|_2 \\
\quad \leq \|[I + s_i \bar{V}]^{-1}\|_2 ||T||_2 ||I + s_i \bar{V}||_2.
\]
Now, from (5.118),
\[
||I - s_i \bar{V}||_2 \leq \|[I + s_i \bar{V}]^{-1}(I + s_i \bar{V})^{-1}\|_2 ||I - s_i \bar{V}||_2 (I + s_i \bar{V})^{-1}.
\]
Since $\bar{H}, \bar{V}$ are respectively non-negative, positive definite, Lemma 2.2 (p.36) yields
\[
||[I - s_i \bar{V}]^{-1}||_2 \leq 1, \quad ||[I - s_i \bar{V}]||_2 \leq 1.
\]
It therefore follows that
\[
||T||_2 < 1.
\]
In turn, it follows from Lemma 2.1 (p.36) that
\[ ||(I+s_1\overline{v})^{-1}||_2 < 1.\]

We now obtain from (5.125),
\[ ||T^k|| < ||(I+s_1\overline{v})||_2 = \rho(I+s_1\overline{v}). \quad (5.126)\]
(since \(\overline{v}\) is symmetric). Now since \(\overline{v} = F^{-1}VF^{-1} = F^{-1}(F^{-1}v)F^{-1}\) we have \(\rho(\overline{v}) = \rho(F^{-1}v)\). From (5.113) and (5.121), and Theorem 1.4 (Gerschgorin's theorem), we obtain
\[ \rho(\overline{v}) = \rho(F^{-1}v) \leq 4. \quad (5.127)\]

Thus, (5.126) and (5.127) yield
\[ ||T^k|| < 1+4s_1. \quad (5.128)\]
Assuming then that \(s_1=\Delta t/\Delta r^2\) is constant, \(||T^k||<C\), where \(C\) is a constant independent of \(k\). Hence, using Condition 2.2 we can state

**Theorem 5.3**

The Peaceman-Rachford formula (5.104a), (5.104b) for the solution of problem (5.76)-(5.79) is stable in the \(L^2\)-norm if \(s_1=\Delta t/\Delta r^2\) remains constant (bounded).

**(iv) Truncation error analysis for the general formula (5.101)**

We have chosen to investigate the general formula (5.101) in order to provide a basis for the truncation error and convergence analyses of hopscotch methods in polar coordinates (in the next section), as well as to develop similar theory for the present method, i.e. the Peaceman-Rachford method in \((r-\theta)\) geometry. Also from this general analysis, we will be able to make a reasonably good theoretical comparison of the accuracy of the two classes of methods, for solving equation (5.76) in particular.

Since (5.101) is a two-step process, we examine the local truncation errors after the two steps have been completed. Let
\[ L^{(1)}_{\Delta x} u^{1}_{i,j} = (B_{i,j} u^{1}_{i-1,j} + C_{i,j} u^{1}_{i,j} + D_{i,j} u^{1}_{i+1,j}) / \Delta x^2, \]

\[ L^{(2)}_{\Delta \theta} u^{1}_{i,j} = (A_{i,j} u^{1}_{i,j-1} + C_{i,j} u^{1}_{i,j} + E_{i,j} u^{1}_{i,j+1}) / \Delta \theta^2, \]

and

\[ L_{\Delta x \Delta \theta} u^{1}_{i,j} = (L^{(1)}_{\Delta x} + L^{(2)}_{\Delta \theta}) u^{1}_{i,j}. \]

Specifically, if \( L^{(1)}, L^{(2)}, L^{(1)}_{\Delta x} \) and \( L^{(1)}_{\Delta \theta} \) are defined by (5.100) then using Taylor series expansions about \( U(r, \theta, t) \) in a similar fashion to equations (1.2), we obtain,

\[ L^{(1)}_{\Delta x} u^{k}_{i,j} = L^{(1)} u^{k}_{i,j} + \frac{1}{r^4} \cdot \frac{\Delta x^2}{6} \left[ \frac{a^2 u^{k}}{3 \Delta r^2} \right]_{i,j} + \frac{\Delta x^2}{12} \left[ \frac{a^4 u^{k}}{3 \Delta r^4} \right]_{i,j} + \ldots, \]

(5.130)

and

\[ L^{(2)}_{\Delta \theta} u^{k}_{i,j} = L^{(2)} u^{k}_{i,j} + \frac{1}{r^4} \left[ \frac{\Delta \theta^2}{12} \cdot \frac{a^4 u^{k}}{3 \Delta \theta^4} \right]_{i,j} + \ldots, \]

(5.131)

We assume that the functions \( U(r, \theta, t), g(r, \theta, t) \) and \( g^{(1)}(r, \theta, t) \) are sufficiently smooth functions so that they have bounded derivatives of any order required.

The two step process (5.101) is

\[ [1-\Delta t L_{1}] u^{k}_{i,j} = [1+\Delta t L_{2}] u^{k-1}_{i,j} + \Delta t \Phi^{k}_{i,j}, \]

(5.132)

\[ [1-\Delta t L_{2}] u^{k+1}_{i,j} = [1+\Delta t L_{1}] u^{k}_{i,j} + \Delta t \Psi^{k+1}_{i,j}, \]

where

\[ L_{1} = \Theta^{k+1}_{i,j} \Delta t_{x} + \Theta^{k+1}_{i,j} \Delta \theta, \]

(5.133)

\[ L_{2} = \Theta^{k}_{i,j} \Delta t_{x} + \Theta^{k}_{i,j} \Delta \theta, \]

\[ b^{k}_{i,j} = \left( \Theta^{k}_{i,j} u^{(1)}_{i,j} + \Theta^{k}_{i,j} u^{(2)}_{i,j} \right) + \left( \Theta^{k-1}_{i,j} u^{(1)}_{i,j} + \Theta^{k-1}_{i,j} u^{(2)}_{i,j} \right). \]

Eliminating \( u^{k}_{i,j} \) from (5.132) gives,

\[ [1-\Delta t L_{1}] [1-\Delta t L_{2}] u^{k+1}_{i,j} = [1+\Delta t L_{1}] [1+\Delta t L_{2}] u^{k-1}_{i,j} + \Theta^{k}_{i,j}, \]

(5.134)
where
\[ c_{1,j}^k = \Delta t \left[ [1-\Delta t L_1] b_{1,j}^{k+1} \right] + \Delta t \left[ [1+\Delta t L_1] b_{1,j}^k \right]. \]

Again using Taylor series expansions, of a form analogous to equations (2.4) yields
\[ c_{1,j}^k = \Delta t (b_{1,j}^{k+1} + b_{1,j}^k) + \Delta t^2 L_1 (b_{1,j}^k - b_{1,j}^{k+1}) \]
\[ = 2\Delta t (b_{1,j}^k + \eta_{1,j}^{k+1} - \eta_{1,j}^k O(\Delta t^3) + \eta_{1,j}^{k+1} - \eta_{1,j}^k O(\Delta t^3)) \]
\[ - 2\Delta t^3 L_1 \left( \eta_{1,j}^{k+1} \Delta t + \eta_{1,j}^k \Delta t^2 \right). \]

If \( u_{1,j}^k \) denotes the exact solution of the differential equation (5.99) at the point \((r_1, \theta_1, t_k)\), then
\[ \left[ [1-\Delta t L_1] [1-\Delta t L_2] u_{1,j}^{k+1} \right] - \left[ [1+\Delta t L_1] [1+\Delta t L_2] u_{1,j}^{k-1} \right] \]
\[ = u_{1,j}^k - u_{1,j}^{k-1} - \Delta t (L_1 + L_2) (u_{1,j}^{k+1} - u_{1,j}^k) + \Delta t^2 L_1 L_2 (u_{1,j}^{k+1} - u_{1,j}^{k-1}) . \]  

If we return now to our original problem, (5.76)–(5.79) then \( L_1 \) and \( L_2 \) will be defined by equations (5.133) and (5.100). Then making use of equations (5.130) and (5.131), and defining \( c_{1,j}^k \) as in (5.90), leads to the error equation
\[ \left[ [1-\Delta t L_1] [1-\Delta t L_2] c_{1,j}^{k+1} \right] - \left[ [1+\Delta t L_1] [1+\Delta t L_2] c_{1,j}^{k-1} \right] + d_{1,j}^k , \]
where
\[ d_{1,j}^k = 2\Delta t^3 L_1 L_2 \frac{\partial u_{1,j}^k}{\partial t} + \Delta t^3 L_1 (\theta_{1,j}^{k+1} \frac{\partial \theta}{\partial t}^{(1)} + \theta_{1,j}^{k+1} \frac{\partial \theta}{\partial t}^{(2)}), \]
\[ - \frac{\Delta t^3}{3} \frac{\partial u_{1,j}^k}{\partial \theta} \frac{\partial^2 u_{1,j}^k}{\partial \theta^2} + \frac{\Delta t^2}{12} \frac{\partial^2 u_{1,j}^k}{\partial \theta^4} + \frac{1}{12} \frac{\Delta t^2}{\partial \theta^2} \frac{\partial^4 u_{1,j}^k}{\partial \theta^4}. \]

In (5.138), the only terms dependent upon \( \theta_{1,j}^k \) and \( \eta_{1,j}^k \) are
\[ 2\Delta t^3 L_1 L_2 \frac{\partial u_{1,j}^k}{\partial t} , \]
and
\[ 2\Delta t^3 L_1 (\theta_{1,j}^{k+1} \frac{\partial \theta}{\partial t}^{(1)} + \theta_{1,j}^{k+1} \frac{\partial \theta}{\partial t}^{(2)}). \]

In the case of equation (5.76), \( g(r, \theta, t) = 0 \), hence the second term of the above two is zero. Therefore for this particular example
For the Peaceman-Rachford process, by (2.23), \( \Theta^k_{i,j} \) and \( \eta^k_{i,j} \) are functions of \( k \) only, and since \( L_{\Delta x}^{(1)}, L_{\Delta \theta}^{(2)} \) from (5.100) are elliptic difference operators we have the above expression equal to

\[
2\Delta t^3 L_{\Delta \theta}^{(2)} L_{\Delta x}^{(1)} \frac{\partial^k}{\partial t^k_{i,j}}.
\]

which if \( L_{\Delta x}^{(1)}, L_{\Delta \theta}^{(2)} \) are defined by (5.100) becomes

\[
2\Delta t^3 \frac{1}{r^3_1} \frac{2}{\partial t^2_{i,j}} \left( \frac{2}{\partial t} \left( \sum_{i,j} \frac{\partial^k}{\partial t^k} \right) \right) + \text{higher order terms},
\]

using (5.130) and (5.131).

The local truncation error in method (5.104a,b) is then given by (5.138) incorporating (5.142). Let us assume that every partial derivative (mixed or otherwise) of \( U(r, \theta, t) \) is bounded in modulus by a positive constant \( B \), for all \( r, \theta, t \). Then from the afore-mentioned equations, we have

\[
|d^k_{i,j}| \leq 0(\Delta t^3) + 2\Delta t^3 \left( \frac{1}{r^3_1} + 2\Delta t \left( \frac{1}{r^3_1} \frac{\Delta r^2}{6} + \frac{\Delta \theta^2}{12} + \frac{1}{r^3_1} \frac{\Delta \theta^2}{12} \right) \right) B, \tag{5.143}
\]

for all \( i, j, k \). It follows therefore, that since \( r_1 = a + \Delta r (a = 0, 1) \) for the circle), \( r_1 = \min_i r_1 \), yielding:

\[
\max_{i,j,k} |d^k_{i,j}| \leq 0(\Delta t^3) + 2\Delta t^3 \left( \frac{1}{r^3_1} + 2\Delta t \left( \frac{1}{r^3_1} \frac{\Delta r^2}{6} + \frac{\Delta \theta^2}{12} + \frac{1}{r^3_1} \frac{\Delta \theta^2}{12} \right) \right) B. \tag{5.144}
\]

Suppose now, that we superimpose a polar grid on the ring \((\text{arg} r)\) so that all the internal mesh points correspond to identical mesh points on the unit circle \((\text{arg} r = 0)\) and the analytical solution is the same at corresponding mesh points. This of course implies that the mesh sizes \( \Delta r, \Delta \theta \) are the same for both regions; let us also suppose the time increment \( \Delta t \) is the same in both cases.

It thus follows from its definition that the quantity "B" for the ring \((\text{arg} r)\) must be less than or equal to the "B" for the circle \((\text{arg} r)\). Since also for the ring, \( r_1 = a + \Delta r \) \((a > 0)\) and for the circle
\( r_1 = \alpha r/2 \), we have that the term on the right of (5.144) must be larger (more so as "a" becomes larger) for the circle than for the ring, under the above assumptions. This upper bound however may be quite crude since it assumes that all the terms in the local truncation errors are positive (which is rarely the case in practice) and also does not take into account the fact that the derivative terms in the local truncation error [see (5.138) and (5.142)] could decrease as \( r \) decreases. There seems little justification however, in assuming that this latter possibility will in general occur.

Thus, although providing far from conclusive evidence, equation (5.144) gives us a strong indication, that, under the conditions mentioned above, we may expect more accurate results from the approximate solution of problem (5.76)-(5.79) when \( R \) is the ring \((0 \ll \alpha R \ll 1)\) than when \( R \) is the circle \((0 \ll \alpha)\). Numerical evidence to substantiate these remarks is given in subsection 5.4(vi).

(v) Convergence analysis for formula (5.104a,b)

We consider here the convergence of the above formula for the solution of the problem defined by equations (5.76)-(5.79). Again, \( \| \cdot \| \) will denote \( \| \cdot \|_F \).

The global error equation obtained from (5.114) is

\[
\left[ I + s_2 F^{-1} \right] \left[ I + s_1 F^{-1} \right] \tilde{T}^{2k+2} = \left[ I - s_2 F^{-1} \right] \left[ I - s_1 F^{-1} \right] \tilde{T}^{2k+1} + \tilde{T}^{2k+1} \tilde{d}^{2k+1}
\]

(5.145)

for \( k = 0, 1, 2, \ldots \). This may then be written as

\[
\tilde{T}^{2k+2} = \tilde{T}^{2k} + \tilde{T}^{2k+1} + \tilde{d}^{2k+1}
\]

(5.146)

where \( \tilde{T} \) is defined by (5.116) and

\[
\tilde{d}^{2k+1} = F^{-1} \left[ I + s_1 \right]^{-1} \left[ I + s_2 \right]^{-1} \tilde{T}^{2k+1} \tilde{d}^{2k+1}
\]

(5.147)

Now, from (5.146)

\[
\tilde{T}^{2k} = T^{2k} \tilde{T}^{2k} + T^{k-1} \tilde{T}^{k-1} \tilde{T}^{2k} + \ldots \ldots + T^{2k-3} \tilde{T}^{3} + \tilde{T}^{k-1}
\]

\[
+ \tilde{T}^{2k} \tilde{d}^{2k+1} + \tilde{T}^{2k+1}
\]

(5.148)
Thus,

\[ ||a^{2k}|| \leq ||T^{k}|| ||a^{0}|| + ||R_{k}|| \]  

(5.148)

where

\[ ||R_{k}|| = ||T^{k-1} \hat{a} + \ldots + T_{k}^{2k-3} \hat{a}^{2k-1}||. \]

If the process is stable, we know from (5.123) that \( ||T^{k}|| \leq C \) (a positive constant) for all integers \( k(>0) \). In this case

\[ ||R_{k}|| \leq kC \max_{l \leq 2k-1} ||\hat{a}^{l}||. \]  

(5.149)

Now, from (5.147) we obtain

\[ ||\hat{a}^{2k+1}|| = ||(I+a_{21}^{-1})^{-1}||_{2}||a^{2k+1}|| < ||a^{2k+1}||. \]

using Lemma 2.1 (p.36) twice.

It now follows from (5.123) that

\[ ||a^{2k+1}|| = ||a^{2k+1}|| \approx ||a^{2k+1}||_{H_{2}}. \]

From the definition of \( ||.||_{H_{2}} \) [equation (5.27) and (5.113)],

\[ ||y^{1}_{1} a^{2k+1}||_{H_{2}} \leq \max_{i,j} |v_{1}^{1} d^{2k+1}_{i,j}| \]

\[ \leq \left\{ 0(\Delta t^{3})+2 \frac{\Delta x^{3}}{\Delta x^{3}}+2\Delta t(h(1) \Delta \varepsilon^{2}+\Delta \varepsilon^{2}) \right\} B. \]  

(5.150)

for all \( i,j,k \), using (5.143).

Combining equations (5.148)-(5.150) yields

\[ ||a^{2k}|| \leq C ||a^{0}|| + C \cdot t_{2k} \left\{ 0(\Delta t^{3})+ \frac{\Delta \varepsilon^{2}}{\Delta x^{3}} + \Delta \varepsilon^{2} \right\} \]

\[ + \left\{ \frac{1}{r^{1}_{1}} \frac{\Delta x^{2}}{6} + r^{1}_{1} \frac{\Delta x^{2}}{12} + \frac{1}{3} \frac{\Delta x^{2}}{12} \right\} B. \]  

(5.151)

where \( t_{2k} = 2k\Delta t \).

**Case 2:** \( R \) is the ring (after)

Firstly in (5.151) we have \( ||a^{0}||=0 \) (same initial data for difference and differential equations). Now in case 1, "a" is a fixed positive number strictly greater than zero. Hence

\( r^{1}_{1}=a+\Delta x>>0 \). The constant terms in (5.151) are \( C, t_{2k} \) and \( B \).

Hence from Definition 2.5, we have
Theorem 5.4

The Peaceman-Rachford formula \((5.104a,b)\) for the solution of problem \((5.76)-(5.79)\) when \(R\) is the \(\frac{\pi}{2}\) circle \((a \neq 1)\), is convergent in the \(H^2\)-norm if the process is stable (in the \(H^1\)-norm).

Case 3: \(R\) is the unit circle

In this case \(\tau_j = \Delta \tau/2\). Substitution into \((5.151)\) gives

\[
||e^{2k}\| \leq C \tau^k \left\{ 0(\Delta \tau^2) + 4/\sqrt{2} \frac{\Delta \tau^2}{3^{3/2}} \left( \frac{\Delta \tau^2}{3^{3/2}} \right) \right. + \frac{\Delta \tau^2}{12^{3/2}} \left. + \frac{\Delta \tau^2}{3^{3/2} \Delta \tau^{3/2}} \right\}.
\]

Hence,

\[
||e^{2k}\| \to 0 \text{ if } \left( \frac{-\Delta \tau}{\Delta \tau^{3/2}} \right) \to 0 \text{ and } \left( \frac{-\Delta \tau}{\Delta \tau^{3/2}} \right) \to 0, \text{ as } \Delta \tau \to 0, \Delta \tau \to 0, \Delta \theta \to 0. \tag{5.152}
\]

We notice that the second condition is similar to the sufficient condition (assuming stability) given for the convergence in the \(\infty\)-norm of the explicit method \((5.83)\) for the solution of this problem. We can now state

Theorem 5.5

The Peaceman-Rachford formula \((5.104a,b)\) for the solution of problem \((5.76)-(5.79)\) when \(R\) is the (unit) circle is convergent (in the \(H^2\)-norm) if the process is stable (in the \(H^1\)-norm) and if conditions \((5.152)\) are satisfied.

It also follows from the above arguments and Definition 2.6, that formula \((5.104a,b)\) will be a consistent process in the \(H^1\)-norm for case 1. It will also be consistent for case 2 if conditions \((5.152)\) hold.

(vi) Computational experiments

To verify experimentally some of the points brought out in the previous subsections, several experiments were undertaken. These consisted of solving a Dirichlet problem for the heat equation

\[
\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial y^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + g(r, \theta, t), \tag{5.153}
\]
in the cylinder $R^4(0.5rsl.0]$, $R$ being the ring (asrs1),
the quarter ring $(0.5rs1, 0.5s2)$ or the unit circle. In each
case, the analytical solution $U(x,y,t)$ was given, so that exact
initial and boundary conditions were used in the form of equations
(5.77)-(5.79). In each case the exact solution $U(x,y,t)$ chosen
was such that

$$\frac{\partial^4 U}{\partial t^4} = x.f(x,y,t), \quad (5.154)$$

where $f(x,y,t)$ was bounded for all $x,y,t$. In this instance the

$$\left\{ \begin{array}{l}
-1 \Delta t^2 \frac{\partial^4 U}{\partial t^4} \\
\frac{12}{\Delta t^2} \frac{\partial^4 U}{\partial t^4}
\end{array} \right\}_{1,j}$$

becomes

$$\left\{ \begin{array}{l}
\frac{1}{12} \frac{\Delta t^2}{\Delta t^2} f(x,y,t)
\end{array} \right\}, \quad (5.138)$$

in each case the exact solution of (5.77)-(5.79) obtained from all
solutions of (5.78)-(5.79) of the form (5.154). This was done both for simplicity, and also
because we wish in the next experiment to compare accuracy with the
P-R method in cartesian coordinates [see equations (2.30), (2.31),]
used to solve the same problem. Both the unit circle and the ring
(asrs1) had their centres at the point $(x=1.0, y=1.0)$. 

**Experiment 2**

Comparison of the accuracy of the Pecceadi-Rachford procedure
(5.102a,b) for solving equation (5.153) on the ring and on the circle.
In each case the exact solution of (5.153) was obtained from an exact
solution of (5.153) in cartesian coordinates (namely, $3U/3x=3^2U/3x^2+$
$3^2U/3y^2+g(x,y,t))$. This was done both for simplicity, and also
because we wish in the next experiment to compare accuracy with the
P-R method in cartesian coordinates [see equations (2.30), (2.31),]
(a) \[ g(r, \theta, t) = 0 ; \quad U(r, \theta, t) = U(x, y, t) = \sin x \sin y e^{-2t} \]
\[ = \sin(1 + r \cos \theta) \sin(1 + r \sin \theta) e^{-2t} \] .

We chose \( \Delta r = 2/(2m+1) = 2/15 \) (i.e. \( m = 7 \) peripheres), \( \Delta \theta = \pi/22 \) and \( \Delta t = 0.01 \). With these same mesh sizes, the ring \((a \theta_0 l)\) had \( a = 7/15 \), so that all its mesh points where contained in the unit circle as well. For both regions the absolute errors given by
\[ |u_{i,j}^k - u_{i,j}^*| \]
were calculated after \( k = 100 \) time steps, vis. at \( t = 1.0 \), for all grid points \((r_i, \theta_j) \in \mathbb{R} \). In Table 1(a) the minimum, average and maximum absolute errors denoted by min.a.e., av.a.e. and max.a.e. were obtained at time \( t = 1.0 \).

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>( R = \text{RING}(7/15\theta_0 l) )</th>
<th>( R = \text{unit CIRCLE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>2.08 ( 10^{-7} )</td>
<td>3.90 ( 10^{-6} )</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>1.57 ( 10^{-5} )</td>
<td>1.44 ( 10^{-4} )</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>4.07 ( 10^{-5} )</td>
<td>4.25 ( 10^{-4} )</td>
</tr>
</tbody>
</table>

(b) \[ g(r, \theta, t) = \sin x e^{-t} / (1 + y)^2 = \sin(1 + r \cos \theta) e^{-t} / (2 + r \sin \theta)^2 ; \quad U(r, \theta, t) \]
\[ = \sin x \log(1 + y) e^{-t} = \sin(1 + r \cos \theta) \log(2 + r \sin \theta) e^{-t} . \]

With this new definition of \( g(r, \theta, t) \), an identical experiment to 1(a) was carried out. The results are displayed below.

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>( R = \text{RING}(7/15\theta_0 l) )</th>
<th>( R = \text{unit CIRCLE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>5.13 ( 10^{-7} )</td>
<td>6.36 ( 10^{-6} )</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>4.00 ( 10^{-5} )</td>
<td>9.11 ( 10^{-4} )</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>1.25 ( 10^{-4} )</td>
<td>3.42 ( 10^{-4} )</td>
</tr>
</tbody>
</table>
In both experiments, we notice an appreciable difference in the overall accuracy of the method on the two regions in question. These results uphold the remarks made at the end of subsection 5.4(iv).

Experiment 2

Comparison of the accuracy of the Peaceman-Rachford formula \((5.104ab)\) in \((r-\theta)\) geometry used to solve the heat conduction equation \((5.76)\) with the equivalent method in \((x,y)\) geometry [namely, equations \((2.47)\)] used to approximate the solution of \((5.76)\) in cartesian coordinates viz.

\[
\frac{\partial U}{\partial t} = \frac{1}{\partial x^2} + \frac{1}{\partial y^2}, \quad U=U(x,y,t). \tag{5.157}
\]

The latter method has been described in section 2.4.

With regard to the circle, a square mesh of size \(h=0.1\) was adopted on the cartesian grid, giving 303 internal grid points. On the polar grid, \(\Delta r\) and \(\Delta \theta\) were the same as before, resulting in 303 internal mesh points. In both cases, \(\Delta t=0.01\).

The results for the P-R method in \((x,y)\) geometry on the circle (and later on the quarter ring) were obtained from a general program supplied, together with all the necessary starting data, by Dr. A.R. Courlay (1974). As with the previous experiment, in both methods the absolute errors were calculated after \(k=100\) steps \((T=1.0)\), together with the min.a.e., av.a.e. and max.a.e. which are abbreviations defined just before Table 1(a). In both cases, the unit circles had their centres at the point \((x=1.0, y=1.0)\).

The first analytical solution for both cases was

(a) that given by \((5.155)\). In Table 2(a) below, we give the absolute errors at \(T=1.0\) at points \((x,y)\) common to both grids (polar and cartesian) for

I1) the Peaceman-Rachford method in \(x-y\) geometry and

II) the Peaceman-Rachford method in \(r-\theta\) geometry.
Henceforth, to be referred to as methods (1) and (2).

Table 2(a)

<table>
<thead>
<tr>
<th>(0.4, 1.0)</th>
<th>(0.8, 1.0)</th>
<th>(1.0, 1.0)</th>
<th>(1.2, 1.0)</th>
<th>(1.6, 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 2.185 $10^{-5}$</td>
<td>(1) 4.826 $10^{-5}$</td>
<td>(1) 4.803 $10^{-5}$</td>
<td>(1) 4.843 $10^{-5}$</td>
<td>(1) 3.309 $10^{-5}$</td>
</tr>
<tr>
<td>(2) 1.412 $10^{-4}$</td>
<td>(2) 2.994 $10^{-4}$</td>
<td>(2) -</td>
<td>(2) 9.840 $10^{-5}$</td>
<td>(2) 9.094 $10^{-5}$</td>
</tr>
</tbody>
</table>

(1) Max.a.e. = 4.976 $10^{-5}$; av.a.e. = 2.2 $10^{-5}$; min.a.e. = 5.639 $10^{-8}$.
(2) Max.a.e. = 4.252 $10^{-4}$; av.a.e. = 1.44 $10^{-4}$; min.a.e. = 3.904 $10^{-6}$.

(b) An identical experiment was then carried out with a new solution of $U(x, y, t) = \cos x \cos y e^{-2\theta\cos(1+\cos\theta)\cos(1+\sin\theta)e^{-2t}}$. Since, as above, the solution along the line $x=1.0$ is the same as that along $y=1.0$, we just give the errors at the points on the line $y=1.0$.

Table 2(b)

<table>
<thead>
<tr>
<th>(0.4, 1.0)</th>
<th>(0.8, 1.0)</th>
<th>(1.0, 1.0)</th>
<th>(1.2, 1.0)</th>
<th>(1.6, 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 1.695 $10^{-5}$</td>
<td>(1) 2.160 $10^{-5}$</td>
<td>(1) 1.982 $10^{-5}$</td>
<td>(1) 1.603 $10^{-5}$</td>
<td>(1) 5.700 $10^{-6}$</td>
</tr>
<tr>
<td>(2) 2.272 $10^{-5}$</td>
<td>(2) 1.850 $10^{-5}$</td>
<td>(2) -</td>
<td>(2) 1.825 $10^{-4}$</td>
<td>(2) 7.298 $10^{-5}$</td>
</tr>
</tbody>
</table>

(1) Max.a.e. = 2.369 $10^{-5}$; av.a.e. = 9.269 $10^{-6}$; min.a.e. = 2.739 $10^{-8}$.
(2) Max.a.e. = 2.814 $10^{-4}$; av.a.e. = 7.817 $10^{-5}$; min.a.e. = 5.464 $10^{-7}$.

In both experiments 2(a) and 2(b) it was noticed that while the errors in method (1) [at $T=1.0$] were fairly uniform over the grid [as indicated by the results in Tables 2(a) and 2(b)], the same was not true for method (2). An example of this effect for the latter method is shown in Table 2(c) where the absolute errors at mesh points along the radius $j=0$ of the polar grid ($y=1.0$, $1.0\leq j\leq 2.0$ on the cartesian grid) are displayed.

Table 2(c)

<table>
<thead>
<tr>
<th>(1/15, 0)</th>
<th>(3/15, 0)</th>
<th>(5/15, 0)</th>
<th>(7/15, 0)</th>
<th>(9/15, 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2) 2.505 $10^{-4}$</td>
<td>(2) 1.825 $10^{-4}$</td>
<td>(2) 1.432 $10^{-4}$</td>
<td>(2) 1.075 $10^{-4}$</td>
<td>(2) 7.298 $10^{-5}$</td>
</tr>
<tr>
<td>(2) -</td>
<td>(2) 4.137 $10^{-5}$</td>
<td>(2) 1.586 $10^{-5}$.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The first value in \((r, \theta)\) denotes the distance \(r\) from the centre of the circle, and "0" denotes the radius \(\theta=0(j=0)\). This variation in the error over the grid was present throughout, although it was not always monotonically increasing as \(r\to 0\) as in Table 2(c). However, on the whole it was noticed that the smallest errors were at points close to the boundary, and the largest at points near the centre of the circle. This is caused by the factors \(x_i^{-1}, x_i^{-2}\) and \(x_i^{-3}\) in the local truncation error of \((5.138)\) with \((5.142)\), and in view of the condition \((5.154)\) which holds for experiment 2(a) and 2(b), it is evident from the former two equations that in general the errors might well be worse for method (2).

Experiments 2(a) and 2(b) provide fairly conclusive evidence of the superiority, accuracy-wise, of method (1) over method (2) for the approximate solution of problem \((5.76)-(5.79)\), when \(R\) is the unit circle. This is despite the fact that boundary correction (see section 2.5) was not used in method (1), and that the \(O(\Delta t^2)\) error term at regular mesh points reduces to \(O(\Delta t)\) at irregular mesh points adjacent to the boundary (see, for instance, formula \((2.16)\)). It would appear that these reductions in the accuracy of method (1) near the boundary are outweighed by the reductions in accuracy of method (2) near the vicinity of the centre of the circle caused by the factors \(x_i^{-1}, x_i^{-2}\) and \(x_i^{-3}\) in the local truncation error \((5.138),(5.142)\).

(e) The next experiment was similar to experiment 2(a), except that this time the region \(R\) was the quarter ring \((0.5r\leq 1, 0.5\theta\leq \pi/2)\), where the origin \((0,0)\) of the polar coordinate system was at the point \((x=1.0, y=1.0)\) again. For method (1), \(h=0.1\) again, but for method (2) we took \(\Delta r=0.1, \Delta \theta=\pi/28\) (and \(\Delta t=0.01\) for both procedures), so that there were exactly 32 internal mesh points on both grids. We chose
a quarter rather than a full ring, since the program we were using for method (1) [Gourlay, 1975] required that each \( x \)- and \( y \)-line of the network crossed the boundary at only two points. Exact initial and boundary data were now given on the lines \( \theta = \frac{\pi}{2} \) and \( \theta = \pi \), \( 0 < y \leq 1 \), \( 0 < x \leq 1 \), \( 0 < y \leq 1 \), and \( x = 1 \), \( 0 < y \leq 1 \).

It will be noticed that method (2) now only requires tri-diagonal inversions along \( \theta \)-lines (or "semi-peripheral") at the first step [see equation (5.104a)], and similar inversions along the radii at the second step [see equation (5.104b)]. For this problem, boundary correction would be necessary along the lines \( \theta = 0 \) and \( \theta = \frac{\pi}{2} \). However, this was omitted in method (2), and also in method (1). The stability and convergence of method (2) follow in an analogous manner to the ring (agr4).

The following table shows the comparative maximum, average and minimum absolute errors at time \( T=1.0 \) (after 50 time steps of \( 2\Delta t=0.02 \)) for the two methods.

### Table 2(d)

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Method (1)</th>
<th>Method (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>7.822 \times 10^{-11}</td>
<td>2.688 \times 10^{-6}</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>2.987 \times 10^{-6}</td>
<td>6.513 \times 10^{-6}</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>6.075 \times 10^{-6}</td>
<td>1.089 \times 10^{-5}</td>
</tr>
</tbody>
</table>

(d) The above experiment was then repeated in exactly the same way except that the solution was this time given by \( U(x,y,t) = \cos x \cos y \cdot e^{-2t} \).

### Table 2(e)

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Method (1)</th>
<th>Method (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>5.058 \times 10^{-8}</td>
<td>6.823 \times 10^{-6}</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>4.613 \times 10^{-7}</td>
<td>1.218 \times 10^{-5}</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>1.780 \times 10^{-6}</td>
<td>1.659 \times 10^{-5}</td>
</tr>
</tbody>
</table>
We notice that in both the above two experiments the errors are much smaller than in the comparative experiments on the unit circle. This is largely due to the factors \( r^{-1}, r^{-2} \) and \( r^{-3} \) being considerably smaller on the ring, and the values of \( \alpha \) and \( \Delta \theta \) are also smaller on this quarter ring.

It is clear from the above results, how the behaviour of the solution can have a lot to do with accuracy, and since methods (1) and (2) contain different derivative terms in their local truncation errors, this fact may also be partly responsible for the difference in accuracy for a particular problem.

We can only conclude that for experiments 2(c) and 2(d) the Peaceman-Rachford method in \((r-\theta)\) geometry appears to give reasonably accurate results, even though they are not quite as good as for the Peaceman-Rachford method in \((x-y)\) geometry.

**Experiment 3**

This last experiment was carried out using method (2) on the (complete) ring \((0.5\pi \leq \theta \leq \pi)\), for the solution of problem \((5.76)-(5.79)\). In subsection 5.4(ii), we showed that by solving equations \((5.104a)\) at the intermediate time level boundary correction was unnecessary. In experiments 3(a) and 3(b) below, we illustrate the drop in accuracy if we use formula \((5.104b)\) at the intermediate time level without boundary correction, vis. so that the computational effort is identical in both cases.

We computed the max.a.e. (maximum absolute error) and av.a.e. (average absolute error) after 50 full time steps of \(2\Delta t=0.02\), with \( \Delta r=0.1 \) and \( \Delta \theta=\pi/28 \), using

- \((a)\) formula \((5.104a)\) at the intermediate time level, and using
- \((aa)\) formula \((5.104b)\) at the intermediate time level.
(a) Firstly, we chose $U(r, \theta, t) = \sin(1 + r \cos \theta) \sin(1 + r \sin \theta) e^{-2t}$, i.e. with the centre of the ring at $(x=1.0, y=1.0)$

<table>
<thead>
<tr>
<th>errors</th>
<th>*</th>
<th>**</th>
</tr>
</thead>
<tbody>
<tr>
<td>max.a.e.</td>
<td>$1.92 \times 10^{-5}$</td>
<td>$3.95 \times 10^{-5}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>$7.27 \times 10^{-6}$</td>
<td>$1.41 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

(b) Secondly, we chose $U(r, \theta, t) = \cos(1 + r \cos \theta) \cos(1 + r \sin \theta) e^{-2t}$.

<table>
<thead>
<tr>
<th>errors</th>
<th>*</th>
<th>**</th>
</tr>
</thead>
<tbody>
<tr>
<td>max.a.e.</td>
<td>$1.50 \times 10^{-5}$</td>
<td>$4.68 \times 10^{-5}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>$5.03 \times 10^{-6}$</td>
<td>$2.39 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

5.5 Hopscotch methods in polar coordinates

(i) The hopscotch categories

In section 2.4, we defined the odd-even and line hopscotch processes with respect to the general formula (2.19). In the same way, we can define our hopscotch methods in polar coordinates with respect to equation (5.101). We shall continue to use the definitions of $r_{i,j}^{k}$ and $u_{i,j}^{k}$ given after equations (5.100).

Thus, using this notation, we shall give three definitions of $\theta_{i,j}^{k}$ for incorporation with the general formula (5.101), which will in fact be the same as those given in equations (2.24) and (2.25). Hence, the value

$$\theta_{i,j}^{k} = \eta_{i,j}^{k} = \begin{cases} 1 & \text{if } k+i+j \text{ is even} \\ 0 & \text{if } k+i+j \text{ is odd} \end{cases}$$

(5.158)

gives the odd-even hopscotch method;

$$\theta_{i,j}^{k} = \eta_{i,j}^{k} = \varepsilon_{i}^{k} = \begin{cases} 1 & \text{if } k+i \text{ is even} \\ 0 & \text{if } k+i \text{ is odd} \end{cases}$$

(5.159)

gives the peripheral hopscotch method:

$$\theta_{i,j}^{k} = \eta_{i,j}^{k} = \varepsilon_{j}^{k} = \begin{cases} 1 & \text{if } k+j \text{ is even} \\ 0 & \text{if } k+j \text{ is odd} \end{cases}$$

(5.160)
gives the radial hopscotch method. All three methods refer of course to a polar grid.

(ii) The computational procedure:

The above three hopscotch processes in polar coordinates for the solution of the general equation (3.99) can then be written as

\[ u_{i,j}^{k+1} - \Delta \theta_{i,j} L_{\theta} u_{i,j}^{k+1} = u_{i,j}^k + \Delta \phi_{i,j} L_{\phi} u_{i,j}^k + \Delta \theta_{i,j} L_{\theta} u_{i,j}^k + \Delta \phi_{i,j} L_{\phi} u_{i,j}^k \]

(3.161)

If in particular \( L = V^2 \), then \( L_{\theta \theta} \) is defined by equations (5.100).

For this example, and our two regions of problem solution, namely the ring \((a \leq r \leq b)\) and the unit circle, the computational procedures for the above three methods are analogous to the odd-even and line hopscotch processes described in section 2.4.

For odd-even hopscotch on a polar grid we can use either a peripheral ordering (see Figures 5.7 and 5.8) or a radial ordering (see Figures 5.9 and 5.10) of our internal mesh points. We can in fact derive an identical formula to (2.29), in the case when \( k+i+j \) is odd, namely,

\[ u_{i,j}^{k+2} = 2u_{i,j}^{k+1} - u_{i,j}^k, \quad k=0,1,2,\ldots \]  

(5.162)

The computation now follows an identical pattern to that described for its counterpart in \((x-y)\) geometry (see section 2.4).

As mentioned above, the peripheral and radial hopscotch methods on a polar grid are analogous to the \(x\)- and \(y\)-line hopscotch methods on a square grid. The difference in the case of peripheral hopscotch is that having solved for \( u_{i,j}^{k+1} \) on alternate peripherals using the explicit formula \( u_{i,j}^{k+1} = 1, u_{i,j}^{k+1} = 0 \) in (3.161) when \( k=1 \), and formula (3.162) otherwise), the difference solutions at points on the remaining alternate peripherals are obtained by solving periodic tri-diagonal systems of equations, by Algorithm 4.3. For this method, we are of course using a peripheral ordering of the mesh.
points, as exemplified in Figures 5.7 and 5.8.

For the radial hopscotch method we adopt a radial ordering of the grid points as depicted in Figures 5.9 and 5.10. This is just like an x- or y-line method requiring the solution of tri-diagonal systems of equations by Algorithm 4.2 on alternate radii, having computed the values \( u_{i,j}^{k+1} \) on the other radii using one of the afore-mentioned explicit formulae.

(iii) Stability

In this section we look at the global form of the two-step process (5.161) used to solve problem (5.76)-(5.79), and hence discuss its stability. In this case, \( g = -g(1) - g(2) = 0 \) in (5.161), and in the same equation, \( L_{2\Delta r}^{(1)} \) and \( L_{2\Delta \theta}^{(2)} \) are defined by (5.100).

Let us specify a peripheral ordering of our unknown grid points. We will therefore be considering the stability of the odd-even (with this ordering) and peripheral methods for the numerical solution of (5.161). The stability analysis for the radial hopscotch method (using of course a radial ordering of the unknown grid points) follows along similar lines, and we shall thus discuss only the former two methods.

First, we introduce some notation. Let the matrices \( H, V \) and \( F \) be defined by (5.113), and the matrix \( A \) by

\[
A = (s_2 H + s_1 V) \quad (5.163)
\]

Let us assume a constant time increment of \( 2\Delta t \) so that \( s_1 = \Delta t / \Delta r^2 \) and \( s_2 = \Delta t / (r \Delta \theta)^2 \) remain unaltered for fixed \( \Delta r, \Delta \theta \). Further, let the diagonal matrices \( I_1, I_2 \) be defined by

\[
[I_1 u_{i,j}^{2k}]_{i,j} = \theta_{1,i,j}^{2k} u_{i,j}^{2k} \quad ; \quad I_2 = I - I_1 \quad (5.164)
\]

where \( \theta_{1,i,j}^{k} \) is defined by (5.158) and (5.159) for the odd-even and peripheral hopscotch methods, respectively. With these definitions of \( I_1 \) and \( I_2 \), the global formulation of the odd-even and peripheral
hopsotch processes for the solution of (5.76) can be derived
from (5.161) as:

\[ |I+I_2F^{-1}(a_2u+a_1v)|u^k = |I-I_2F^{-1}(a_2u+a_1v)|u^{k-1} + \Delta b^k, \]
\[ |I+I_2F^{-1}(a_2u+a_1v)|u^{k+1} = |I-I_1F^{-1}(a_2u+a_1v)|u^k + \Delta b^k, \]  
(5.165)

where \( k = 1, 3, 5, \ldots \), and the vectors \( b^k \) are known, containing the
initial and boundary values. Both \( u^k \) and \( b^k \) are \( (m+1) \)-vectors.

Eliminating \( u^k \) from (5.165) gives

\[ u^{k+1} = Tu^{k-1} \quad \Delta b^k, \quad k = 1, 3, 5, \ldots \]  
(5.166)

where \( \Delta b^k \) is independent of the \( u^k \)s and where

\[ T = F^{-1} |I+I_2\bar{A}|^{-1} \bar{T}|I+I_2\bar{A}|, \]
\[ \bar{A} = F^{-1} AF^{-1} \]
and \( \bar{T} = |I-I_1\bar{A}||I+I_1\bar{A}|^{-1} |I-I_2\bar{A}||I+I_2\bar{A}|^{-1} \)  
(5.167)

In deriving these expressions, we have made use of the fact that
since \( F, I_1 \) and \( I_2 \) are diagonal matrices, they commute.

We now examine the matrix \( \bar{A} \). From equations (5.163), (5.120)
and (5.121) we see that \( A \) is symmetric, and being the sum of a
non-negative and positive definite matrix is itself positive
definite. By similar arguments to those given before, \( \bar{A} = F^{-1} AF^{-1} \)
is also positive definite (since \( F \) is a positive diagonal matrix).

Letting \( ||u||_{H_2} \) be defined as before (see (5.27)), we define

\[ ||u||_{A^2F} = ||A^2F^{1/2}u||_{H_2}, \]  
(5.168)

as the \( A^2F^{1/2} \)-vector norm, and its subordinate matrix norm as

\[ ||u||_{A^2F} = ||A^2F^{1/2}G^{1/2}A^{-1}||_2. \]  
(5.169)

Until indicated otherwise, we now let \( ||.|| \) denote \( ||.||_{A^2F} \).

From equation (5.167) we have

\[ \bar{T}^k = F^{-1} |I+I_2\bar{A}|^{-1} \bar{T}^k |I+I_2\bar{A}|, \]
so that

\[ ||\bar{T}^k|| = ||A^2|I+I_2\bar{A}|^{-1} \bar{T}^k |I+I_2\bar{A}|^{-1}||_2 \]

\[ = |||I+I_2\bar{A}|^{-1} \bar{T}^k |I+I_2\bar{A}|||_{A^2F}. \]  
(5.170)
where \( |\lambda| \), for positive definite \( A \) is defined by (5.29). We have now arrived at an almost identical situation to the one encountered in subsection 5.2(iv).

To avoid repetition, we make use of the results in subsection 5.2(iv), to show, that from (5.170),
\[
||T^k|| \leq 1 + \rho(I_2 F^{-1} A).
\]  
(5.171)
since \( F^{-1} (I_2 A) F^\dagger = F^{-1} (I_2 F^{-1} A) F^\dagger = I_2 F^{-1} A \). \( \{I_2 F^{-1} A\} \) because \( F^{-1}, I_2 \) are diagonal matrices. Now \( A \) is defined by equations (5.163), (5.120), (5.121) and \( F \) by equation (5.113). Next, \( I_2 (F^{-1} A) \) consists of rows from \( (F^{-1} A) \) and null rows, so it follows from Theorem 1.4, Corollary that
\[ \rho(I_2 F^{-1} A) \leq 4(s_1 + s_2). \]  
(5.172)
Thus if \( s_1 = \Delta t/\Delta r^2 \) and \( s_2 = \Delta t/(r_1 \Delta r)^2 \) remain constant and/or bounded, (5.171) and (5.172) yield:
\[ ||T^k|| \leq C, \text{ independent of } k. \]

If \( R \) is the range (agreement) then \( r_1 = a + \Delta r \) (a fixed), so that \( s_2 < \Delta t/(a \Delta r)^2 \) for all \( \Delta r > 0 \). Hence the above condition on \( s_1 \) and \( s_2 \) will not be too restrictive. However, if \( R \) is the unit circle, \( r_1 = \Delta r/2 \), so that \( s_2 = 4 \Delta t/(\Delta r \Delta r)^2 \). Since for convergence it will be shown (as in the previous section) that \( \Delta r \ll \Delta t, \Delta t \) will have to be very small for \( s_2 \) to remain constant and/or bounded. This of course would be a severe restriction on the practical use of the method.

Thus we can state:

**Theorem 5.6**

The Odd-even and peripheral hopscotch processes given by equation (5.161) and equations (5.158) and (5.159) respectively, are stable in the \( \{A F^\dagger \}_1 \)-norm for the solution of problem (5.76)-(5.79), if \( s_1 \) and \( s_2 \) remain constant.

N.B. From the above remarks, it is easily shown that the same
conditions ensure the stability of the radial hopscotch method.

(iv) Truncation error analysis for formula (5.101)

The local truncation error $\delta_{1,j}^{k}$ in the general formula (5.101) for the solution of equation (5.76) has already been given by equation (5.138). As regards the odd-even, radial, and peripheral hopscotch processes in equation (5.101), (see (5.61)) the only terms dependent upon $\theta_{1,j}^{k}$ in (5.138) are given by equations (5.139) and (5.140). The remaining terms in the "general local truncation error" are the same as for the Peaceman-Rachford method of the previous section. To work out the terms in (5.139) and (5.140), we use an identical approach to that given in subsection 5.2(vi) to determine the local truncation error in the peripheral and other hopscotch methods on a square cartesian grid. This time, instead of $i$ and $j$ in $\theta_{1,j}^{k}$ going along $x$- and $y$-lines, they go along $r$- and $\theta$-lines or along radii and peripherals, respectively.

Using the definitions of $L_{\Delta r}$, $L_{\Delta \theta}$ and $L_{\Delta r \Delta \theta}$ from equations (5.129), and those of $L_{1}$ and $L_{2}$ from equations (5.133) and since for the three hopscotch methods in question

$$\theta_{1,j}^{k} = \Phi_{1,j}^{k},$$

we have

\[
2\Delta r^{3} L_{1} L_{1} \Phi_{1,j}^{k} = 2\Delta r^{3} \theta_{i,j}^{k+1} L_{\Delta r} \theta_{i,j}^{k} L_{\Delta \theta} \theta_{i,j}^{k} \Phi_{1,j}^{k},
\]

and

\[
2\Delta r^{3} L_{1} \left( \Phi(1)_{i,j} + \delta_{i,j}^{k+1} \theta_{(2)}_{i,j}^{k} \right) = 2\Delta r^{3} \theta_{i,j}^{k+1} L_{\Delta r} \theta_{i,j}^{k} \delta_{i,j}^{k+1} \Phi_{1,j}^{k}.
\]

In a similar manner to the truncation error analysis of subsection 5.2(iv), we can derive simpler expressions for the sum of the two quantities in (5.173). Hence, in this present context, the sum of the two terms in (5.173), for odd-even hopscotch is

\[
2\Delta r^{3} \left[ C_{1}^{(1)} \frac{\delta_{i,j}^{k}}{\Delta t} + C_{1}^{(2)} \frac{\delta_{i,j}^{k}}{\Delta \theta} \right] \left( -\frac{\Delta r}{\Delta \theta} \frac{\Phi_{i,j}^{k}}{\Phi_{i,j}^{k}} + \frac{\Delta \theta}{\Delta r} \frac{\Phi_{i,j}^{k}}{\Phi_{i,j}^{k}} \right),
\]

where $\theta_{1,j}^{k}$ is defined by (5.158).
For peripheral hopscotch this sum is

$$2\Delta t^3 \frac{\partial^2}{\partial x^2} \theta_{i,j}^{k+1} \left( b_{i,j} + d_{i,j} \right) \left( \frac{\partial^2 u}{\partial t^2} - \frac{2g}{\alpha} k \right)_{i,j} \quad (5.175)$$

where \( \theta_{i,j} \) is defined by (5.159).

Likewise, for the radial hopscotch process, the sum of the terms in (5.173) is

$$2\Delta t^3 \frac{\partial^2}{\partial \theta^2} \theta_{i,j}^{k+1} \left( a_{i,j} + e_{i,j} \right) \left( \frac{\partial^2 u}{\partial t^2} - \frac{2g}{\alpha} k \right)_{i,j} \quad (5.176)$$

where \( \theta_{i,j} \) is defined by (5.160).

As we did with the Peaceman-Rachford process in the previous section, let us now return to our original problem defined by equations (5.76)-(5.79). In this case \( L^{(1)} \text{ and } L^{(2)} \) and therefore \( a_{i,j}, b_{i,j}, c_{i,j}, d_{i,j}, e_{i,j} \) are given by equations (5.100). Hence, in this case, equations (5.174)-(5.176) become (and for this problem \( g=0 \))

$$4\theta_{i,j}^{k+1} \frac{\Delta t^3}{\partial x^2} \left\{ \frac{1}{\Delta x^2} + \frac{1}{(r_{i,j})^2} \right\} \left( \frac{\partial^2 u}{\partial t^2} - \frac{2g}{\alpha} k \right)_{i,j} \quad (5.177)$$

for the odd-even hopscotch;

$$4\theta_{i,j}^{k+1} \frac{\Delta t^3}{\partial \theta^2} \left( \frac{\partial^2 u}{\partial t^2} - \frac{2g}{\alpha} k \right)_{i,j} \quad (5.178)$$

for the peripheral hopscotch;

$$4\theta_{i,j}^{k+1} \frac{\Delta t^3}{(r_{i,j})^2} \left( \frac{\partial^2 u}{\partial t^2} - \frac{2g}{\alpha} k \right)_{i,j} \quad (5.179)$$

for the radial hopscotch method.

It is now apparent that in modulus this part of the total local truncation error in the peripheral and radial hopscotch processes is less than the corresponding term given by (5.177) for the odd-even hopscotch. Also, as \( r_{i,j} \) decreases and if \( \Delta x \) and \( \Delta \theta \) are approximately equal (or indeed if the latter is less than the former) the third term (5.179) becomes larger than the second, (5.178). In such a situation, it is likely (taking into account that the remaining terms in the local truncation error are the same for all three methods) that the peripheral hopscotch
will be more accurate than the radial hopscotch, and both will be more accurate than the odd–even hopscotch procedure.

The corresponding term to equations (5.177)-(5.179) in the local truncation error of the Peaceman–Rachford process (5.104) for the solution of the same problem is given by equation (5.142); namely

\[ 2\Delta t^3 \frac{1}{r_i^3} \frac{\partial^2}{\partial \theta^2} \left[ \frac{2}{\kappa \partial \theta} \right] \frac{1}{\partial^2} \]  

\[ \text{For the ring (5\pi \xi 1), } r_i = e^{i\Delta r}, \]  

(1\text{sign}), so for \( a \) sufficiently large, although the derivative terms in this error are different from those in (5.178), say, we may expect more accurate results from the P-R method than from the peripheral hopscotch process. For the circle, \( r_i = (1-\Delta r) \), (1\text{sign}), so in this case, the difference in accuracy of these two methods, is likely to be less marked than in the previous case.

In subsection 5.5(vi), we give numerical results to substantiate the above remarks.

(v) \textbf{Convergence}  

We consider first the convergence of the odd–even and peripheral hopscotch processes. From equations (5.165), the global error equation may be written as

\[ [I + I_1(F^{-1}A)] [I + I_2(F^{-1}A)] \Delta^{2k+2} = [I - I_1(F^{-1}A)] [I - I_2(G^{-1}A)] \Delta^{2k} \]  

\[ + \Delta^{2k+1}, \ k=0,1,2,...(5.180) \]  

(A similar error equation can be derived for the radial hopscotch process). Letting \( ||.|| \) denote the \( A^{-1}F \) matrix norm (see (5.169)), the analysis follows an almost identical pattern to that given for the convergence of hopscotch methods in (x–y) geometry in subsection 5.2(v) – see equations (5.43)-(5.47). Thus to avoid repetition, we utilise these latter results to obtain in the present context

\[ ||\Delta^{2k}|| \leq ||R_k|| \leq \text{C. max } ||\Delta^{2k-1}||, \]  

(5.181)  

where \( C \) satisfies the stability condition \( ||T^k|| \leq C, \) for all integer \( k \geq 0. \)
Now for this particular problem, the local truncation errors for the odd-even and peripheral hopscotch methods (and radial hopscotch) are given by equation (5.138), where $g_{g(1)} g_{g(2)} = 0$ and where the term $2 \Delta t^3 \sum_{l_1} \sum_{e_{i,j}} \frac{3 u}{e_{i,j}^3}$ is evaluated for the respective processes (again with $g=0$) by equations (5.177)-(5.179). In particular, for odd-even hopscotch we have

$$d_{i,j}^{k+1} = \frac{4 \Delta t^3}{1} \left\{ \frac{1}{(\Delta x^2)} + \frac{1}{(\Delta \theta^2)} \right\} \frac{3 u^k_{i,j}}{\Delta t^2} + O(\Delta t^3)$$

$$-2 \Delta t \left( \frac{1}{r_1^2} \frac{\Delta x^2}{6} \frac{3 u}{\Delta x^2} + \frac{\Delta \theta^2}{12} \frac{3 u}{\Delta \theta^2} + \frac{1}{r_1^2} \frac{\Delta \theta^2}{12} \frac{3 u^k_{i,j}}{\Delta \theta^2} \right). \quad (5.182)$$

Hence, if $B$ is defined as before,

$$\max_{1,j,k} |d_{i,j}^{k+1}| \leq \left[ O(\Delta t^3) + \frac{4 \Delta t^3}{1} \left\{ \frac{1}{(\Delta x^2)} + \frac{1}{(\Delta \theta^2)} \right\} \right]$$

$$\leq 2 \Delta t \left( \frac{1}{r_1^2} \frac{\Delta x^2}{6} + \frac{\Delta \theta^2}{12} \right) B = \Delta t (\Delta x, \Delta \theta, \Delta t), \quad (5.183)$$

Now,

$$||d|| = ||d_{i,j}^{k+1}||_{A_F^1} = ||A^1_F d^k||_{A^1} \leq ||A^1||_{A^2} ||d^k||_{A^2}$$

$$\leq ||A^1||_2 ||d^k||_{A^2} \leq \max_{1,j,k} |d_{i,j}^{k+1}| \quad [\text{from (5.27)}]$$

$$\leq ||A^1||_2 r_1^4 B (\Delta x, \Delta \theta, \Delta t). \quad [\text{from (5.182)}] \quad (5.184)$$

Since $A^1$ is symmetric, $||A^1||_2 = \rho(A^1) = \frac{1}{4} \Delta t^3$. Now $F^{-1} A_F = F^{-1} (F^{-1} A F) F = F^{-1} A$, hence $\rho(A) = \rho(F^{-1} A)$. Now it follows in exactly the same way that equation (5.172) was obtained, that $\rho(F^{-1} A) \leq 2 (s_1 + s_2)$. Therefore

$$||d|| \leq \rho(A^1) \leq 2 (s_1 + s_2). \quad (5.185)$$

Substituting (5.185) into (5.184) we obtain

$$||d|| \leq 2 (s_1 + s_2)^{\frac{1}{4}} r_1^4 B (\Delta x, \Delta \theta, \Delta t), \quad \text{for all } k \geq 0. \quad (5.186)$$
Thus equations (5.181) and (5.186) yield
\[ |e^{2k_1}| e^{2k_2} \sigma_{k_2} \cdot C \cdot (e_{1} + e_{2})^{\frac{1}{2}} \cdot \frac{D(\Delta \xi, k_{2})}{\Delta t}, \quad (5.187) \]
where \( k_{2} \equiv 2k_{1} \).

Hence, \( |e^{2k_1}| \to 0 \) as \( \Delta r \to 0, \Delta \theta \to 0, \Delta t \to 0 \) if \( e_{1} \) and \( e_{2} \) remain constant (bounded) and if \( \{ e_{1} \frac{D(\Delta r, \Delta \theta, \Delta t)}{2 \Delta r} \} \to 0 \). From (5.183), the last condition will be satisfied if \( e_{1} = \Delta t / \Delta r^{2} \) and \( e_{2} = \Delta t / (r_{1} \Delta \theta) \) remain constant, and if \( (\Delta \theta / r_{1}^{2}) \to 0 \) (as \( \Delta \theta \to 0, \Delta r \to 0 \)). When \( R \) is the ring (agree), \( r_{1} = a + \Delta r \) ("a" fixed), so the last condition is automatically satisfied. When \( R \) is the unit circle, however, \( r_{1} = \Delta r / 2 \), so \( (\Delta \theta / r_{1}^{2}) = (\Delta \theta / \Delta r / 2) \) must tend to zero as \( \Delta r \to 0, \Delta \theta \to 0 \).

Analogous remarks can be made concerning the convergence of the peripheral and radial hopscotch processes. The only difference being that the term in the curly brackets in (5.182) and (5.183) is replaced by \( (1/\Delta r^{2}) \) [see (5.178)] and \( (1/(r_{1} \Delta \theta)^{2}) \) [see (5.179)], respectively.

We can now state the following two convergence theorems.

**Theorem 5.7**

The odd-even, peripheral and radial hopscotch procedures are convergent in the \( A^{1} \mu^{1} \)-norm for the solution of problem (5.76)-(5.79) when \( R \) is the ring (agree), if the processes are stable, and if \( e_{1} \) and \( e_{2} \) remain constant.

**Theorem 5.8**

The above three methods are convergent (in the same norm) for the solution of problem (5.76)-(5.79) when \( R \) is the unit circle if they are stable, if \( e_{1} \) and \( e_{2} \) remain constant, and if \( \Delta \theta / \Delta r^{3} \to 0 \) as \( \Delta r \to 0, \Delta \theta \to 0 \).

As usual, it is now easily verified [from (5.186)] that the above three methods for the problem in question, will be consistent (in the \( A^{1} \mu^{1} \)-norm) if \( e_{1} \) and \( e_{2} \) remain constant, and if \( R \) is the circle, then the last condition in Theorem 5.8 must hold.
When \( R \) is the ring \((a<r<1)\), assuming \( a\gg 0 \), the condition that \( s_1 \) and \( s_2 \) must be kept constant is not too restrictive. The reason being that in this case, \( \Delta t = s_1 \Delta r^2 = s_2 x^2 \Delta \theta^2 = \frac{s_2}{a}(s+\Delta r)^2 \Delta \theta^2 < s_2 (a \Delta \theta)^2 \), so for the above condition to hold, \( \Delta t \) must be proportional to \( \Delta r^2 \) and \( \Delta \theta^2 \).

When \( R \) is the circle, however, this condition on \( s_1 \) and \( s_2 \) implies \( \Delta t \) must be proportional to \( (\Delta r \Delta \theta)^2 \), since \( s_1 = \Delta r / 2 \). This is of course, very restrictive, since in general it necessitates very small values of \( \Delta t \), and hence a very large number of steps to advance one unit (say) in time. What with these very stringent conditions, and the tendency for a substantial drop in accuracy in a close neighbourhood to the centre of the circle, leads us to conclude that in general the above three hopscotch methods (like the Peaceman-Rachford process) will not be practical in this case.

Since also for these hopscotch processes, the local truncation errors have all but one term in common with the Peaceman-Rachford method [see (5.138)] analogous remarks to those made at the end of subsection 5.4(iv) can be made regarding the comparative accuracy of the hopscotch methods when \( R \) is a ring and when \( R \) is a circle. This is illustrated by numerical experiment below.

(vi) Computational results

In this subsection we describe experiments carried out using the odd-even, radial and peripheral hopscotch methods on a polar grid. The problems we solved were exactly the same as those described in Experiment 1 of subsection 5.4(vi).

Experiment 1(a) \( (\partial U/\partial t = \partial^2 U/\partial r^2 + (1/r) \partial U/\partial r + (1/r^2) \partial^2 U/\partial \theta^2) \)

In the table below are displayed the minimum, average and maximum absolute errors at \( T=1.0 \) after 50 time steps with \( 2\Delta t=0.02 \). To complete the picture we also include the results of the Peaceman-
Rachford procedure (P.R.) from subsection 5.4(vi).

Table I(a)

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Odd-even</th>
<th>Radial</th>
<th>Peripher</th>
<th>P.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>1.73 $10^{-6}$</td>
<td>1.15 $10^{-6}$</td>
<td>6.07 $10^{-7}$</td>
<td>2.08 $10^{-7}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>1.17 $10^{-4}$</td>
<td>8.16 $10^{-5}$</td>
<td>3.06 $10^{-5}$</td>
<td>1.57 $10^{-5}$</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>2.73 $10^{-6}$</td>
<td>1.89 $10^{-4}$</td>
<td>8.84 $10^{-5}$</td>
<td>4.07 $10^{-5}$</td>
</tr>
</tbody>
</table>

$R$ is the ring (7/15$r/4$)

Table I(b)

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Odd-even</th>
<th>Radial</th>
<th>Peripher</th>
<th>P.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>5.56 $10^{-4}$</td>
<td>1.30 $10^{-4}$</td>
<td>2.94 $10^{-5}$</td>
<td>3.90 $10^{-6}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>2.86 $10^{-3}$</td>
<td>2.41 $10^{-3}$</td>
<td>3.35 $10^{-4}$</td>
<td>1.44 $10^{-4}$</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>4.86 $10^{-3}$</td>
<td>4.18 $10^{-3}$</td>
<td>5.65 $10^{-4}$</td>
<td>4.25 $10^{-4}$</td>
</tr>
</tbody>
</table>

Experiment I(b)

In this case we solved $\frac{\partial u}{\partial r} = \frac{\partial^2 u}{\partial r^2} + (1/r) \frac{\partial u}{\partial r} + (1/r^2) \frac{\partial^2 u}{\partial r^2} + \sin(1+\cos\theta) e^{-t}/(2+\sin\theta)^2$.

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Odd-even</th>
<th>Radial</th>
<th>Peripher</th>
<th>P.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>5.10 $10^{-7}$</td>
<td>8.07 $10^{-7}$</td>
<td>2.23 $10^{-6}$</td>
<td>5.13 $10^{-7}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>8.21 $10^{-5}$</td>
<td>6.33 $10^{-5}$</td>
<td>4.46 $10^{-5}$</td>
<td>4.00 $10^{-5}$</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>1.93 $10^{-4}$</td>
<td>1.69 $10^{-4}$</td>
<td>1.55 $10^{-4}$</td>
<td>1.25 $10^{-4}$</td>
</tr>
</tbody>
</table>

$R$ is the ring (7/15$r/4$)

$R$ is the unit circle

<table>
<thead>
<tr>
<th>Absolute errors</th>
<th>Odd-even</th>
<th>Radial</th>
<th>Peripher</th>
<th>P.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>min.a.e.</td>
<td>5.50 $10^{-6}$</td>
<td>4.95 $10^{-4}$</td>
<td>9.90 $10^{-8}$</td>
<td>6.36 $10^{-8}$</td>
</tr>
<tr>
<td>av.a.e.</td>
<td>3.10 $10^{-3}$</td>
<td>2.86 $10^{-3}$</td>
<td>1.69 $10^{-4}$</td>
<td>9.11 $10^{-5}$</td>
</tr>
<tr>
<td>max.a.e.</td>
<td>5.83 $10^{-6}$</td>
<td>5.48 $10^{-3}$</td>
<td>2.99 $10^{-4}$</td>
<td>3.42 $10^{-4}$</td>
</tr>
</tbody>
</table>
As predicted by the theory, in both experiments we have more accurate results on the ring than on the circle, for all four methods. As also predicted, the peripheral hopscotch is more accurate in these experiments than the radial hopscotch (particularly on the circle), which in turn is more accurate than the odd-even hopscotch. It was pointed out in subsection 5.4(vi) that in these experiments the term $3^4U/30^4$ satisfies condition (5.154), so that in general, the results could be significantly less accurate (for the same mesh sizes) on the circle.

Next we consider the comparative efficiency of the four methods.

For the three hopscotch methods, let us look at the amount of work required to obtain the solution vector $u^{k+1}$ at time $t_{k+1} = (k+1)\Delta t$. As we know from subsection 5.5(ii), in each method half of the solutions at the mesh points are calculated by the same explicit formula. The remaining half, assuming it is equation (5.76) we are solving, are obtained from the fully implicit formula,

\[
\begin{split}
(1+2(s_1+s_2))u^{k+1}_{i,j} - s_1u^{k+1}_{i-1,j} - s_1u^{k+1}_{i+1,j} - \frac{\Delta t}{r_i^2}(r_i - \frac{\Delta r}{2})u^{k+1}_{i-1,j} - \frac{\Delta t}{r_i^2}(r_i + \frac{\Delta r}{2})u^{k+1}_{i+1,j} = u^k_{i,j}.
\end{split}
\]

We will assume that $k>1$, so that the coefficient terms in (5.188) have already been calculated and stored after completion of the first time step.

It is the way in which formula (5.188) is implemented on the grid which results in one method being more efficient than the other. In the odd-even hopscotch method the value $u^{k+1}_{i,j}$ is calculated explicitly (since $u^{k+1}_{i-1,j}, u^{k+1}_{i+1,j}, u^{k+1}_{i,j-1}, u^{k+1}_{i,j+1}$ are known) in 3 multiplications, 1 division and 4 additions. In the radial hopscotch process tri-diagonal inversions along alternate radii require the same amount of work as the previous method to obtain the solution values $u^{k+1}_{i,j}$ - see Algorithm 4.2. Similar operation counts for the peripheral hopscotch, which uses Algorithm 4.3 to solve for the
\[ u_{k+1}^{1,j} \text{ along alternate peripherals, reveals that this process requires approximately } 1 \frac{1}{3} \text{ times the work of the previous two, to obtain } u^{(k+1)} \text{.} \]

A similar investigation of the arithmetic operations involved in the calculation of \( u^{2k} \) \((k>1)\) by the Peaceman-Rachford method \((5.104a)\) \((5.104b)\) reveals that this method is about three times slower than the radial (and odd-even) hopscotch process.

We now have a hierarchy of methods for the solution of problem \((5.76)-(5.79)\) similar to that given for the hopscotch and P.R. methods in x-y geometry, at the end of section 2.10. Firstly, the odd-even hopscotch is the easiest to program and requires the least storage of the four methods. However, it is no more efficient than radial hopscotch, which is not difficult to program, and is more accurate than odd-even hopscotch. Peripheral hopscotch is less efficient than the former, but in situations where \( \Delta r \gg r, \Delta \theta \) for all \( i \), it follows from equations \((5.178)\) and \((5.179)\), that a peripheral will usually be more accurate than a radial hopscotch process (as illustrated in the above experiments). Finally, the P.R. method is generally the most accurate of the four, but at the expense of development and running time.

### 5.6 Conclusions

To be specific, let our concluding remarks refer to the numerical solution of

\[
\frac{3u}{2t} = v^2 u, \quad (5.189)
\]

(\( v^2 \) is the Laplacian operator in two space dimensions) using the finite difference methods we have described throughout this chapter.

Let us consider first, the case when \( v^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \). In section 5.2, we applied the method of peripheral hopscotch to the finite difference solution of equation \((5.2)\) of which equation \((5.189)\) is a particular example, on a closed plane region. We discovered
that for the solution of the latter equation, peripheral hopscotch has the same degree of accuracy (in general) as line hopscotch. However, the line hopscotch is about 14 as fast as the peripheral hopscotch for obtaining the solution vector \( u^{2k} \) (\( k=0,1,\ldots \)). The former therefore appears to be the most desirable of the two methods in this context.

However, the usefulness of a peripheral ordering was shown in subsection 5.2(ix), where we considered the problem of solving equations such as (5.189), in an open region of the x-y plane. In this case, the roles of the two orderings of the grid points were reversed in their suitability to this particular problem. These remarks would now suggest that the main area of application of a peripheral ordering of the mesh points for the solution of equation (5.189) say, is on an open plane region.

Secondly, let us consider the case when \( v^2 = \frac{\partial^2}{\partial \theta^2} + \frac{\partial}{\partial r} \left( \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \right) \) (\( r, \theta \) denoting the polar coordinate directions. With this definition, we applied the Peaceman-Rachford (A.D.I.) and hopscotch methods to the approximate solution of equation (5.189) on the (a.s.1) and the unit circle. The accuracy and convergence in the latter case where shown to be very poor for the above two classes of methods, mainly due to the behaviour of the functions \( r^{-1}, r^{-2} \) and \( r^{-3} \) near the origin of the r-\( \theta \) coordinate system. Our experimental results showed that the F.R. method in x-y geometry, solving \( \partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 \) on the circle, was about 10 times as accurate as the analogous method in r-\( \theta \) geometry. We must therefore conclude that for the circle or indeed any plane region which contains the origin of the r-\( \theta \) system, a square cartesian grid is preferable to a polar grid for the solution of equation (5.189).
In the case of the ring (a4r51), where "a" was bounded away from r=0, accuracy improved substantially on the polar grid. This again seemed to be caused by the functions $r^{-1}, r^{-2}, r^{-3}$, being notably smaller in this case. In one of the experiments, the accuracy of the Peaceman-Rachford methods in x-y and r-θ geometry on the quarter-ring $(\alpha, 0.5\alpha, \theta, 0.5\theta/2)$, for an identical number of internal mesh points in each case, did not vary to any great degree. It is difficult however to come to any firm conclusions with regard to comparative accuracy in this instance; suffice it to say that from the two experiments carried out, both methods gave acceptable accuracy. Certainly, the latter method is easier to program than the former on the regions in question, and requires no logical decision making with regard to the boundary of the region.

Of the three hopscotch methods we considered on a polar grid, namely odd-even, radial and peripheral, the last one tended to be the most accurate for the solution of equation (5.189) on the ring (a4r51) and the unit circle, but required roughly $1\frac{1}{2}$ times the amount of work as the previous two. As with the peripheral hopscotch method in cartesian coordinates, the peripheral hopscotch process in polar coordinates would be well suited to the problem of solving the heat conduction equation (5.189) on a region external to the unit circle, say. The computational procedure would be carried out in the same manner as described for its counterpart on an (x-y) grid [see subsection 5.2(1x)], and for such a problem the factors $r^{-1}, r^{-2}, r^{-3}<1$ in the local truncation error. Accuracy-wise we would then have a similar situation to the ring (a4r51).
CHAPTER 6

THE PEACEMAN-RACHFORD ITERATIVE METHOD

IN (r,θ) GEOMETRY
6.2 Introduction

In this chapter, we investigate the use of the Peaceman-Rachford Alternating-Direction Implicit iterative method for computing an approximate solution to Laplace's equation in \( (r, \theta) \) geometry. The two regions of problem solution will be the ring \((a, b)\) and the unit circle. The problems are detailed in section 6.2, and like the one described in section 3.2, can be regarded as model problems.

Considerable use is made of the information in section 3.5, where the equivalent method in cartesian coordinates is described. In particular, it was pointed out in that section that while we can obtain theoretical rates of convergence using an "optimum" set of positive acceleration parameters for the "commutative case" (see pages 59-62) we cannot do likewise for the "non-commutative case" (see pages 62-64). However, although the former case is more attractive from a theoretical point of view, we learned that it occurs in only a relatively small class of problems.

For the present problem we find that it is necessary to normalise the difference equations to achieve a "commutative case". This version of our Peaceman-Rachford iterative methods is discussed in section 6.5, and in section 6.7 we investigate a "non-commutative case" arising from the original difference equations. In section 6.8, experimental results reveal that it is the latter which gives the superior rates of convergence for our two model problems, especially on the unit circle.

We will be particularly interested in comparing rates of convergence with Successive "Line" Over-relaxation techniques used to solve identical problems. These methods are analogous to the S.L.O.R. method on a cartesian grid (described in section 3.4), where in the present context a "line" is either a circumference...
(peripheral) of constant radius or a radius on the polar grid.

Benson (1969, 1973) and Benson and Evans (1972) have applied both S.O.R. methods to the two model problems in question, and refer to them respectively as Successive Peripheral Over-relaxation (S.P.O.R.) and Successive Radial Over-relaxation (S.R.O.R.). These methods will be discussed in more detail at the beginning of section 6.8.

MacMillan (1964) has also considered this problem in a pie-shaped region of the (r,θ) plane (viz. a region bounded by two radii and the circular arc between them) using the two above S.O.R. methods and the commutative (normalised) version of our Peaceman-Rachford method (see section 6.5).

6.3 Statement of the problem and formation of the difference equations

We consider the solution of

\[ \frac{1}{r^2} \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = 0, \quad U = U(r, \theta) \quad (6.1) \]

(r,θ) ∈ R, where R is the ring (a≤r≤b) or the unit circle (0≤r≤1), subject to the (Dirichlet) boundary conditions

\[ U(a, \theta) = f(\theta); \quad U(1, \theta) = g(\theta). \quad (6.2) \]

Both equations in (6.2) hold for the former region, while only the second equation in (6.2) is specified for the latter region.

We superimpose a polar grid on the ring and the circle, and arrange the meshes as in Figures 5.7(p.97) and 5.8(p.100) respectively. The finite difference analogue of (6.1) to be applied at the unknown mesh points \((r_i, j, \theta)\), where \(r_i = a + i\Delta r (a - \Delta r/2 \text{ for the circle})\), \(i = 1, 2, \ldots, n\) and \(j = 0, 1, \ldots, n-1\), is

\[ \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta r^2} + \frac{1}{r_i} \cdot \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta r} + \frac{1}{r_i^2} \cdot \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta \theta^2} = 0. \quad (6.3) \]
Equation (6.3) has been obtained by substituting \( h = \Delta r, x = r \) in equations (1.3) and (1.4), and \( h = \Delta \theta, y = \theta \) in (1.6) and amalgamating the new equations. Multiplying (6.3) through by \(-\Delta r^2\) and re-arranging yields,

\[
\left[ \frac{2u_{i,j} - u_{i,j-1} - u_{i,j+1}}{(1 \Delta \theta)^2} \right] \cdot \left[ 2u_{i,j} - (1 - \frac{1}{2\Delta \theta})u_{i-1,j} - (1 + \frac{1}{2\Delta \theta})u_{i+1,j} \right] = 0. \tag{6.4}
\]

where in this case \( i = i_{\text{min}}' + 1, \ldots, i_{\text{min}} + u - 1, i_{\text{min}} \) being defined by (5.80). Alternatively, multiplying (6.3) through by \((-\Delta \theta \Delta r \Delta \theta)\) and re-arranging gives

\[
\frac{\Delta r}{r_{i,j}} \left( 2u_{i,j} - u_{i,j-1} - u_{i,j+1} \right) \cdot \frac{\Delta \theta}{\Delta r} \left( 2r_{i,j} - (r_{i,j} - \frac{\Delta r}{2})u_{i-1,j} - (r_{i,j} + \frac{\Delta r}{2})u_{i+1,j} \right)
\]

\[= 0. \tag{6.5}\]

6.3 The coefficient matrices of the difference equations

Following along similar lines to section 3.5, we let the terms involving the brackets in (6.5) respectively define \( \mathbf{H} \) and \( \mathbf{V} \), in the sense that if \( [H_{i,j}] \) denotes the component of the vector \( \mathbf{H} \) corresponding to the spatial mesh point \((r_{i,j})\), then

\[
[H_{i,j}] = \frac{\Delta r}{r_{i,j}} \left( 2u_{i,j} - u_{i,j-1} - u_{i,j+1} \right), \tag{6.6a}
\]

\[
[V_{i,j}] = \frac{\Delta \theta}{\Delta r} \left( 2r_{i,j} - (r_{i,j} - \frac{\Delta r}{2})u_{i-1,j} - (r_{i,j} + \frac{\Delta r}{2})u_{i+1,j} \right), \tag{6.6b}
\]

when all the mesh points \((r_{i,j} \pm \Delta r, 0 \pm \Delta \theta)\) are mesh points of unknowns.

Adopting a peripheral ordering of these grid points on the ring \((a < r < 1)\) and the unit circle, as illustrated by Figures 5.7 (p.97) and 5.8 (p.100) respectively, yields a system of \( nm \) linear simultaneous equations, represented in matrix notation as

\[
\mathbf{A}_u = (\mathbf{H} + \mathbf{V})u = \mathbf{b}. \tag{6.7}
\]

The vectors \( u \) and \( \mathbf{b} \) are defined in the usual way, and from equations (6.6)
\[
\begin{bmatrix}
H_1 \\
H_2 \\
\vdots \\
H_m
\end{bmatrix}
\] (n×m)

\[
H_{ij} = \frac{\Delta r}{\tau_i \Delta \theta}
\]

\[
2 -1 \\
-1 -1
\] (n×n)

for \(i=1,2,\ldots,m\); and

\[
V =
\begin{bmatrix}
D_1 & E_1 \\
E_1 & E_{m-1}
\end{bmatrix}
\]

\[
D_i = 2r \frac{\Delta \theta}{\Delta r} I_{(n\times n)}
\]

\[
E_i = -\frac{\Delta \theta}{\Delta r} (r_i + \frac{\Delta r}{2}) I_{(n\times n)}
\]

By inspection we see that the matrices \(H\) and \(V\) satisfy property (3.6b).

Further, a simple extension to Theorem 1.1, Corollary reveals that

\(H_j(I_{(j)\times m})\) is irreducible, and since each \(H_j\) satisfies the hypothesis

of Theorem 1.6, it follows from this theorem and Theorem 1.2, that

the \(H_j(I_{(j)\times m})\) are non-negative definite. Next, by Theorem 1.1, the

matrix \(V\) is irreducible, so by Theorem 1.4, Corollary, it is also

positive definite. Thus, as \(A\) in (6.7) is the sum of a positive

definite and non-negative definite matrix, \(A\) is positive definite

(hence non-singular), and the solution of the matrix equation \(A\)\textsuperscript{un}=\textsuperscript{b}

is unique.

In an analogous manner, equations (6.4) yield the system of

equations

\[
\hat{A}u = (\hat{u} + \hat{v})u = \hat{b}
\]

(6.9)
where
\[ \hat{A} = \left( \frac{\Delta x}{\Delta y} \right)^{-1} A, \quad \hat{u} = \left( \frac{\Delta x}{\Delta y} \right)^{-1} u, \quad \hat{v} = \left( \frac{\Delta x}{\Delta y} \right)^{-1} v; \]

\[
F = \begin{bmatrix} r_1^I \\ r_2^I \\ \vdots \\ r_m^I \end{bmatrix}, \quad (m \times m)
\]

and I is the unit matrix of order n.

Returning to equation (6.6b), we see that by re-ordering the mesh points along successive radii of the grid, as in Figures 5.9 (p.106) and 5.10 (p.109), the matrix V can be made tri-diagonal. Hence using this and the information above, we can state that there exist m×m permutation matrices \( P_1 \) and \( P_2 \) such that \( P_1^{HT} \) and \( P_2^{HT} \) are the direct sum of non-negative definite periodic tri-diagonal matrices and the direct sum of positive definite tri-diagonal matrices, respectively. We remind the reader that these properties hold for both regions of problem solution under consideration.

6.4 The iterative process

We define the Peaceman-Rachford (A.D.I.) iterative method for the solution of system (6.7) say, in exactly the same way as the process was treated in section 3.5. So in this case, we have

\[
(p_{k+1} I + H)u^{(k+1)} = (p_{k+1} I - V)u^{(k)} + b, \quad k \geq 0, \quad (6.11a)
\]

\[
(p_{k+1} I + V)u^{(k+1)} = (p_{k+1} I - H)u^{(k+1)} + b. \quad (6.11b)
\]

The \( p_{k+1} (>0) \) are the positive acceleration parameters, \( u^{(k+1)} \) the auxiliary vector, and \( u^{(0)} \) some arbitrary initial vector approximation of the unique solution of (6.7).

Since \( (p_{k+1} I + H) \) and \( (p_{k+1} I + V) \) are, by suitable rearrangement
of their rows and corresponding columns (or equivalently, a re-ordering of the mesh points from a peripheral to a radial ordering) direct sums of positive definite periodic tri-diagonal and tri-diagonal matrices respectively, systems (6.11a) and (6.11b) can be solved directly, for each value of \( k \), using Algorithms 4.3 and 4.2 respectively.

Of course, an analogous method to (6.11) can be used to solve system (6.9), where in this instance \( \hat{U} \hat{V} \) would replace \( U \hat{V} \) in equations (6.11). In either case, the computation continues until the initial error has been reduced by some factor \( \varepsilon \), say (see section 3.3).

In section 3.5, it was pointed out that A-D-I. iterative methods converge appreciably faster when using a sequence of iteration parameters, \( p_{n+1} = k_{0}, 1, \ldots, k_{n} \). It was also mentioned in that section, that at present a theoretical analysis for obtaining the "optimum" parameters, only exists when the matrices \( \hat{U} \) and \( \hat{V} \) commute, in view of Theorem 3.4. Since from (6.8) and (6.10), \( \hat{U} \) and \( \hat{V} \) and \( \hat{U} \) and \( \hat{V} \) do not commute, we make use of the Hackerg and Habetler variant of the F-R method (see p.63) to construct matrices \( \hat{H} = \hat{V} \hat{U} \hat{V} \), \( \hat{V} \hat{U} \hat{V} \) which do commute. This method of approach will come under the following heading.

6.5 The commutative case

In the context of equations (3.97) and (3.98), the positive diagonal matrix \( D \) in this case is \( D = D^{-1} \), where \( D \) is defined by (6.10). An equivalent definition has been given by Macmillan (1964). Then defining \( \hat{F} = \hat{W} \hat{U} \hat{V} \) (i.e. \( \hat{F} = \hat{U}_{1}^{\hat{V}} \hat{u}_{1,j} \)) the normalised version of the finite difference equations (6.5) is given by

\[
\frac{\Delta x}{\Delta y} \left[ 2\tilde{u}_{i,j} - \tilde{u}_{i+1,j} - \tilde{u}_{i-1,j} \right] + \frac{\Delta \hat{y}}{\Delta x} \left[ 2\hat{x}_{i}^{2} \hat{u}_{i,j} \right] - (\hat{x}_{i} - \frac{\Delta \hat{y}}{2} \sqrt{\hat{x}_{i}^{2} + \Delta \hat{y}^{2}}) \hat{u}_{i+1,j} = 0.
\]

(6.12)

Then letting \( \hat{H} \) and \( \hat{V} \) be defined by the first and second bracketed terms respectively, in the same way as before (or following equations
(3.99) and (3.100), we obtain

\[ (p_{k+1} I + v) \tilde{u}^{(k+1)} = (p_{k+1} I - \tilde{v}) \tilde{u}^{(k)} + \tilde{b}, \quad (6.13a) \]

\[ (p_{k+1} I - \tilde{v}) \tilde{u}^{(k+1)} = (p_{k+1} I - \tilde{v}) \tilde{u}^{(k)} + \tilde{b}, \quad (6.13b) \]

Method (6.13) is then carried out in the same manner as described for method (6.11).

In an analogous way to equations (5.104a), (5.104b) and (5.108), we can make the system (6.13b) more efficient. This is achieved by eliminating \( \tilde{u}^{(k+1)} \) from (6.13b) to obtain

\[ (p_{k+1} I - \tilde{v}) \tilde{u}^{(k+1)} = 2p_{k+1} \tilde{u}^{(k)} - (p_{k+1} I - \tilde{v}) \tilde{u}^{(k)}, \quad (6.13b)' \]

which requires less arithmetic operations than (6.13b) since the vector \( (p_{k+1} I - \tilde{v}) \tilde{u}^{(k)} \) in (6.13b)' has already been calculated in (6.13a).

When all the parameters \( p_{k+1} \) equal a fixed constant \( p > 0 \), (6.13a) and (6.13b) (or (6.13b)') can be combined to give

\[ \tilde{u}^{(k+1)} = \tilde{T}_{p} \tilde{u}^{(k)} + \tilde{g}_{p}(\tilde{b}), \quad k \geq 0, \quad (6.14) \]

where

\[ \tilde{T}_{p} = (p I + \tilde{v})^{-1} (p I - \tilde{v}) (p I + \tilde{H}) (p I - \tilde{H})^{-1} (p I - \tilde{v}), \quad (6.15) \]

By Theorem 3.1, the method (6.14) will be convergent (see Definition 3.2) if and only if \( \rho(\tilde{T}_{p}) < 1 \). Now the matrix \( \tilde{T}_{p} \) defined as

\[ \tilde{T}_{p} = (p I + \tilde{v}) \tilde{T}_{p}^{-1} (p I + \tilde{v})^{-1} \]

\( \tilde{v} \) will be shown to be positive definite, so \( (p I + \tilde{v})^{-1} \) exists, is similar to \( \tilde{T}_{p} \) and thus has the same eigenvalues. Hence

\[ \tilde{T}_{p} = [(p I - \tilde{H}) (p I + \tilde{H})^{-1}] [(p I - \tilde{v}) (p I + \tilde{v})^{-1}] \]

and using the \( L_{2} \) (spectral) norm of (1.25) and property (1.23), we have

\[ \rho(\tilde{T}_{p}) = \rho(\tilde{T}_{p}) \leq ||\tilde{T}_{p}|| \]

\[ \leq ||(p I - \tilde{H}) (p I + \tilde{H})^{-1}|| ||(p I - \tilde{v}) (p I + \tilde{v})^{-1}||, \quad (6.16) \]

Since

\[ \tilde{v} = \tilde{v}^{T} \tilde{y}, \quad \tilde{v} = (\tilde{v}^{T} \tilde{y})^{T} = ((\tilde{v}^{T} \tilde{y})^{T} \tilde{v}^{T} \tilde{y})^{T} \tilde{y} \tilde{y}^{T}, \quad (6.16)' \]
it follows from Theorem 1.3 that \( \tilde{H} \) and \( \tilde{V} \) are respectively non-negative and positive definite matrices.

Hence, from the above remarks, if \( \lambda_i, v_i \) (isign) are the eigenvalues of \( \tilde{H}, \tilde{V} \), then \( \lambda_i \geq 0 \) with equality for at least one \( i \), and

\[ v_i > 0 \] for all isign. Since \( \tilde{V} \) is symmetric, it is readily verified that the matrix \( (pI-\tilde{V})(pI+\tilde{V})^{-1} \) is symmetric, with eigenvalues

\[ \frac{p-v_i}{p+v_i} \], isign . \tag{6.17} \]

Thus from property (1.27), we can write

\[ ||(pI-\tilde{V})(pI+\tilde{V})^{-1}|| = \max_{isign} \left| \frac{p-v_i}{p+v_i} \right| < 1 . \tag{6.18} \]

Similarly, we have

\[ ||(pI-\tilde{H})(pI+\tilde{H})^{-1}|| = \max_{isign} \left| \frac{p-v_i}{p+v_i} \right| = 1 \] (since \( \lambda_1 \) (say)=0). \tag{6.19} \]

Therefore from equations (6.16)-(6.19), we conclude that \( \rho(\tilde{T}_p) < 1 \), so the method (6.13) converges. Since this result would hold for a positive definite \( \tilde{H} \) also, we have proved Theorem 3.3 (p.57).

Since \( \tilde{H} \) and \( \tilde{V} \) commute and are both symmetric, it follows from Theorem 3.4, that they have a common set of orthonormal eigenvectors \( Y_i \) (say), isignum. Hence, we have from (6.15) that

\[ \tilde{T}_p Y_i = \begin{bmatrix} p-\lambda_i & p-v_i \\ p+v_i & p+\lambda_i \end{bmatrix} Y_i \], isignum . \tag{6.20} \]

The spectral radius of \( \tilde{T}_p \) is then given by

\[ \rho(\tilde{T}_p) = \max_{isign} \left| \begin{bmatrix} p-\lambda_i & p-v_i \\ p+v_i & p+\lambda_i \end{bmatrix} \right| 
= \max_{isign} \left| \frac{p-v_i}{p+v_i} \right| \tag{6.21} \]

making use of equation (6.19).

Now the matrix \( \tilde{V} \) has been defined from equation (6.12) second bracketed term. If we re-order the mesh points along successive radii of the polar grid (see Figures 5.9 and 5.10), as we would in
the second step (6.13b) of the iterative method, then the matrix \( \bar{V}_T \)
defined by the second bracketed term in (6.12) is

\[
\bar{V}_T = \left( \frac{\Delta \theta}{\Delta r} \right)
\begin{bmatrix}
\phi^0 \\
\phi^1 \\
\vdots \\
\phi^n
\end{bmatrix}
\]

(6.22)

where

\[
\bar{\phi}^n = \begin{bmatrix}
2x_1^2 - (x_1 + \Delta r/2)\sqrt{(x_1 x_2)} \\
-(x_1 + \Delta r/2)\sqrt{(x_1 x_2)} \\
2x_2^2 - (x_2 + \Delta r/2)\sqrt{(x_2 x_3)} \\
-(x_2 + \Delta r/2)\sqrt{(x_2 x_3)} \\
\vdots \\
-(x_{m-1} + \Delta r/2)\sqrt{(x_{m-1} x_m)} \\
2x_m^2
\end{bmatrix}
\]

(6.23)

Hence if \( \bar{\phi}_C \) (lgtm) are the eigenvalues of \( \bar{V}_C \), then \( \frac{\Delta \theta}{\Delta r} \bar{\phi}_C \) (lgtm) are the eigenvalues of \( \bar{V}_C \), with multiplicity \( n \). Since we already know that \( \bar{V}_T = P_2 \bar{V}_C P_2^{-1} \) (where \( P_2 \) is a permutation matrix) we conclude that \( \frac{\Delta \theta}{\Delta r} \bar{\phi}_C, \ t=1,2,\ldots, m, \) are the eigenvalues of \( \bar{V}_C \), also.

As is evident from Definition 3.4, to optimise the rate of convergence, we must minimise \( p(\bar{\phi}_C) \) with respect to \( p \), which from (6.21) and the information just given, means minimising

\[
\max_{\text{lgtm}} \left| \frac{p - \Delta \theta \bar{\phi}_C}{p + \Delta \theta \bar{\phi}_C} \right|
\]

(6.24)

From equations (3.76)-(3.79) the "optimum" value of \( p \) which minimises this expression is given by

\[
\bar{p} = \sqrt{\varepsilon_C}
\]

(6.25)
where the eigenvalues \( v_t = (\Delta R/A) v_{\xi} \) of \( \tilde{V} \) satisfy

\[
0 < v_t = (\Delta R/A) v_{\xi} \leq (\Delta R/A) \beta^t = \beta_t, \quad t = 1, 2, \ldots, m. \tag{6.26}
\]

And from equation (3.79) we obtain

\[
\min_{p > 0} \rho(\tilde{V}) = \rho(\tilde{V}_p) = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}} = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}}. \tag{6.27}
\]

Hence, the rate of convergence of this optimised one parameter P.R. (A.D.I.) method depends solely on the ratio of the spectral bounds \( \alpha, \beta \) for the eigenvalues of the positive definite matrix \( \tilde{V} \) [see (6.23)]. We notice in particular from (6.23), that the matrix \( \tilde{V} \) is independent of \( n \), the number of radii on the polar grid, and hence of \( \Delta \theta \) since \( \Delta \theta = 2\pi/n \). Thus the spectral bounds \( \alpha, \beta \) and likewise from (6.27) the optimum rate of convergence are also independent of \( n \) (the optimum rate of convergence being given by \(-\log(\rho(\tilde{V}))\) using Definition 3.4).

**A sequence (cycle) of acceleration parameters**

We now consider the use of the iterative method (6.13) with a sequence of \( M \) positive parameters \( p_j \). The iteration matrix in this case is then given by

\[
\tilde{V}^{(k+1)} = \left( \prod_{j=1}^{M} \tilde{V}^{(k)} \right) \tilde{V}^{(k)} + \tilde{G}(p_1, p_2, \ldots, p_M, b), \quad k = 0 \pmod{M}. \tag{6.28}
\]

Then since \( \tilde{H} \) and \( \tilde{V} \) commute, it follows from Theorem 3.4 and using (6.20) that

\[
\begin{bmatrix}
\prod_{j=1}^{M} \tilde{V}^{(k)} \\
\end{bmatrix} \tilde{Y}_1 = \begin{bmatrix}
\prod_{j=1}^{M} p_j \lambda_j^{1} \\
\prod_{j=1}^{M} p_j \lambda_j^{1} \\
\end{bmatrix} \begin{bmatrix}
\prod_{j=1}^{M} p_j \lambda_j^{1} \\
\prod_{j=1}^{M} p_j \lambda_j^{1} \\
\end{bmatrix} \begin{bmatrix}
\tilde{Y}_1 \\
\end{bmatrix} = \text{Id} + \text{degm}, \tag{6.29}
\end{bmatrix}
\]

where \( \lambda_j^{1} v_i^{1} \) are the eigenvalues of the maxima matrices \( \tilde{H}_j \tilde{V} \). It now follows from (6.29) that

\[
\rho \left[ \prod_{j=1}^{M} \tilde{V}^{(k)} \right] \leq \max_{1 \leq j \leq M} \left( \frac{\prod_{j=1}^{M} p_j \lambda_j^{1}}{\prod_{j=1}^{M} p_j \lambda_j^{1}} \right) \max \left( \frac{\prod_{j=1}^{M} p_j \lambda_j^{1}}{\prod_{j=1}^{M} p_j \lambda_j^{1}} \right) \tag{6.30}
\]

In a similar manner to equation (6.19), we have
\[
\max_{1 \leq j \leq M} \frac{\left| \frac{P_j^{-\lambda_1}}{P_j^\ast \lambda_1} \right|}{\left| \frac{P_j^{-\lambda_2}}{P_j^\ast \lambda_2} \right|} = 1.
\]

Also, similarly to equation (6.18),
\[
\max_{1 \leq j \leq M} \frac{\left| \frac{\nu_j P_j^{-\lambda_1}}{P_j^\ast \nu_j} \right|}{\left| \frac{\nu_j P_j^{-\lambda_2}}{P_j^\ast \nu_j} \right|} < 1,
\]
so from (6.30),
\[
\rho \left( \max_{1 \leq j \leq M} \frac{\nu_j P_j^{-\lambda_1}}{P_j^\ast \nu_j} \right) < 1. \tag{6.31}
\]

Therefore by Theorem 3.1, the method (6.28) is convergent. From (6.31), if \(\alpha\) and \(\beta\) are defined as in (6.26), it is clear that we now want to minimise
\[
\max_{1 \leq j \leq M} \frac{\nu_j P_j^{-\lambda_1}}{P_j^\ast \nu_j}. \tag{6.32}
\]

There are various techniques described in the literature for obtaining positive parameters to either exactly minimise the expression (6.32), or to approximately minimise it. Two of these approximate methods are given by Warga (1962) p.226. W.E. Jordan has given an exact analysis for finding acceleration parameters \(\nu_j(1 \leq j \leq M)\) for any integer \(M > 0\) to minimise (6.32). This method is described in Wachpress (1966) pp.185-193. The latter author (1962) gives an exact determination of the \(\nu_j\) for \(M = 2^k\), \(k > 0\). He tells us that in numerical application, it is often not unduly restrictive to limit the number of iterations to an integral power of two. His algorithm is very simple to implement (more so than Jordan's algorithm) requiring no approximations other than those involved in finding square roots, and is especially useful when \(M\) is small.

For these reasons, it has been decided to use solely these last mentioned "optimum" parameters. For brevity, we shall usually refer to them as "The parameters".

The solution to this min-max problem (6.32), is given by equations
(3.86)-(3.89), using the specified parameters, which are derived from the algorithm mentioned above. If then, \( d_h[a, b] \) from (3.88) denotes the minimum of the expression (6.33), we have from (6.31),

\[
\rho \left[ \prod_{j=1}^{N} \frac{\tilde{c}_j}{p_j} \right] \leq d_h[a, b] . \tag{6.33}
\]

where \( \bar{p}_j \) (16x54M) denote the \( N=2^k \) (k>0) optimum parameters. 

Hence, the average spectral radius satisfies

\[
\rho \left[ \prod_{j=1}^{N} \frac{\tilde{c}_j}{p_j} \right] \leq \left( d_h[a, b] \right)^{1/N} . \tag{6.34}
\]

and likewise, the average rate of convergence satisfies

\[
\rho \left[ \prod_{j=1}^{N} \frac{\tilde{c}_j}{p_j} \right] \geq \frac{-\log \{ d_h[a, b] \}}{N} . \tag{6.35}
\]

6.6 Estimation of eigenvalue bounds

It is now apparent that the calculation of the optimum parameters \( \bar{p}_j \) (16x54M) and average rates of convergence depend on the values of the spectral bounds \( a, b \) for the eigenvalues of the positive definite matrix \( \tilde{v} \). From previous arguments, this is equivalent to determining \( a', b' \) for the matrix \( \tilde{v}' \) (defined by (6.23)). Unfortunately, the author has discovered no exact formula for the eigenvalues of this matrix, or for determining the lower bound \( a' \). Estimates of the upper bound \( b' \), can be made by using Theorem 1.4, Corollary. Authors such as Varga (1962) p.288 and Wachpress (1966) p.223, tell us that this estimate is adequate for obtaining reasonable optimum parameters, and does not have a critical affect on rates of convergence. However, a method of obtaining a slightly more accurate upper bound \( b' \), will be described below. Estimates of the lower bound \( a' \), however, do have a critical effect on rates of convergence, especially when \( a' \) is close to zero. This indeed proves to be true for our case, when the region of problem solution is the unit circle, as will be explained later.

To obtain an estimate for \( a' \), we follow the analysis of Varga (1962)
pp. 288-289. Firstly, we notice from Theorem 1.1, Corollary and (3.65a),
that \( \tilde{V}' \) is an irreducible tri-diagonal Stieltjes matrix. It can then
be shown (Varga (1962) p. 85) that \((\tilde{V}')^{-1} > 0 \). From this it follows that

**Lemma 6.1**

If \( \tilde{V}' \) is an irreducible \( m \times m \) Stieltjes matrix with eigenvalues
\( 0 < v_1, v_2, \ldots, v_m \), then for all \( 0 < \sigma < v_1 \),
\[ \left( \tilde{V}' - \sigma I \right)^{-1} > 0. \]

**Proof**

For any \( \sigma \) with \( 0 < \sigma < v_1 \), \( \tilde{V}' - \sigma I \) is also an irreducible Stieltjes matrix
(by Theorem 1.1, Corollary and (3.65a)), and the result again follows
from Varga (1962) p. 85.

To find rigorous upper and lower bounds for \( v_1 \), the minimum
eigenvalue of \( \tilde{V}' \), Varga (1962) makes use of the Perron-Frobenius theory
of non-negative matrices, which he discusses in his second chapter. Let
\[ \sigma_0 = 0, \quad \bar{\pi}^{(0)} = (1, \ldots, 1)^T, \]
and define \( \bar{\pi}^{(k)} \), \( \sigma_k \) recursively by
\[ (\tilde{V}' - \sigma_k I) \bar{\pi}^{(k+1)} = \bar{\pi}^{(k)}, \quad k = 0, 1, \ldots; \]
\[ \sigma_{k+1} = \nu_{\min} \quad k = 0, 1, \ldots; \]
where
\[ \nu_{\min} = \sigma_k + \min_{\text{lsigm}} \frac{x_i^{(k)}}{x_i^{(k+1)}}. \]

Since \( \tilde{V}' \) is tridiagonal and non-singular, we can directly solve (by
Algorithm 4.2) for the vector \( \bar{\pi}^{(k+1)} \), which from Lemma 6.1 has positive
components. Now it may be shown (see Varga (1962), Theorem 22, p. 32)
that
\[ \rho(\tilde{V}' - \sigma_k I)^{-1} = \frac{1}{\nu_1 - \sigma_k} \max_{\text{lsigm}} \frac{x_i^{(k+1)}}{x_i^{(k)}}. \]

If we use the Rayleigh quotient to underestimate \( \rho(\tilde{V}' - \sigma_k I)^{-1} \),
[see for example, Theorem 11.5, p. 88 of Faddeev and Faddeeva (1963)]
we are led to the inequality
The iteration procedure may be continued until either 
\[ \nu_{\text{min}}^{(k)} = \nu_{\text{max}}^{(k)} \] 
for some \( k \) or until the convergence criterion

\[ \left| \frac{\nu_{\text{max}}^{(k)} - \nu_{\text{min}}^{(k)}}{\nu_{\text{min}}^{(k)}} - \delta \right| < \delta \]

is satisfied.

Varga (1962) p.289 proves under the hypothesis of Lemma 6.1 that the sequences of positive numbers \( \{\nu_{\text{min}}^{(k)}\}_{k=0}^{\infty} \) and \( \{\nu_{\text{max}}^{(k)}\}_{k=0}^{\infty} \) are respectively non-decreasing and non-increasing, and have the common limit \( \nu_1 \).

Spanier (1967) uses this technique in a general program called "HOT-1" and in that he chooses \( \delta \) to be 0.2. In our own context, we have chosen \( \delta = 0.005 \) (which is probably unnecessarily small), and in almost every case convergence was achieved in three iterations. The value \( \nu_{\text{min}}^{(k)} \) so determined will be a rigorous lower bound for the smallest eigenvalue \( \nu_1 \) of \( \vec{\nu} \), i.e. \( \alpha^1 \) takes the value \( \nu_{\text{min}}^{(k)} \).

As stated earlier, a sufficiently accurate upper bound for \( \nu_{n}^1 \) the largest eigenvalue of \( \vec{\nu} \), can be obtained from Theorem 1.4, Corollary. If \( \alpha^1 \) denotes this upper bound, we have

\[ \nu_{\text{max}}^{(k)} = (\nu_{n-2} + \delta^2) / \nu_{n-2} \nu_{n-1}^2 + 2 \nu_{n-1}^2 + (\nu_{n-1} + \delta^2) / \nu_{n-1} \nu_n \]  

from (6.23). However, a more accurate upper bound may be obtained as follows (and this is the one we have chosen to use). Firstly, we make use of the following result; see Varga (1962), p.47, Exercises 2 and 3.
Lemma 6.2

Let $A \geq 0$ be any non-negative $m \times m$ matrix, $x > 0$ any positive vector.

Then
\[
\min \frac{(Ax)_i}{\log \|x\|} \leq \rho(A) \leq \max \frac{(Ax)_i}{\log \|x\|}.
\]

Furthermore, suppose $A > 0$, whenever $x > 0$, and let
\[
\lambda^{(k)} = A^{(k-1)} \ldots A^{(1)} x^{(0)} > 0
\]
and
\[
\lambda^{(k)} = \max \left\{ \frac{[x^{(k)}]_1}{[x^{(k-1)}]_1}, \ldots, \frac{[x^{(k)}]_1}{[x^{(k-k)}]_1} \right\}, \quad \lambda^{(k)} = \min \left\{ \frac{[x^{(k)}]_1}{[x^{(k-1)}]_1}, \ldots, \frac{[x^{(k)}]_1}{[x^{(k-k)}]_1} \right\}.
\]

Then
\[
\lambda^{(0)} \leq \lambda^{(1)} \leq \ldots \leq \rho(A) \leq \ldots \leq \lambda^{(1)} \leq \lambda^{(0)}.
\]

Let $D$ denote the $(m \times m)$ orthogonal, diagonal matrix defined by
\[
D = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & -1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}.
\]

The following results are easily established from (6.23) and (6.38).

(a) $D^T D = [\tilde{v}]_1$, where $[\tilde{v}]_1$ is the matrix whose elements are $|(\tilde{v})_{1j}|$.

(b) Let $\tilde{v}^{(0)} = (1, -1, \ldots)^T$ be a $(m \times 1)$ vector of alternating signs. Then
$$(\tilde{v}^{(0)})_1$$ and $$(\tilde{v}^{(0)})_1$$ have the same sign.

We now apply Lemma 6.2 to the matrix $[\tilde{v}]_1$. Since $[\tilde{v}]_1 > 0$ and $[\tilde{v}]_1$ has at least one non-zero entry in every row, $[\tilde{v}]_1 > 0$ whenever $x > 0$.

Therefore the conditions of Lemma 6.2 apply. Let $z^{(k)} = (1, 1, \ldots, 1)^T (m \times 1)$, and let $[\tilde{v}]_1 = [\tilde{v}]_1^{(k-1)} = [\tilde{v}]_1^{(k)}$, $k = 0, 1, \ldots, k$. Then, since $\tilde{v}$ is similar to $[\tilde{v}]_1$ (because $D = D^{-1}$),
\[
\rho(\tilde{v}) = \rho([\tilde{v}]_1) \leq \lambda^{(k-1)} \max_{i} \frac{[\tilde{v}]_1^{(k-1)}_i}{[x^{(k-1)}]_1}.
\]
We observe that
\[
|\tilde{v}|^k \Sigma^{(o)} = D(\tilde{v})^k \Sigma^{(o)} = D(\tilde{v})^k \Sigma^{(o)}
\]
where \(\Sigma^{(o)} = (1, -1, \ldots, 1)^T\). Then
\[
\begin{align*}
\frac{[x^{(k)}]_i}{|x^{(k-1)}|_i} &= \frac{[\tilde{v}^k \Sigma^{(o)}]_i}{|\tilde{v}^{k-1} \Sigma^{(o)}|_i} \\
&= \frac{[\tilde{v}^k \Sigma^{(o)}]_i}{[\tilde{v}^{k-1} \Sigma^{(o)}]_i}
\end{align*}
\]
so that the following process is actually used to estimate \(\rho(\tilde{v})\). With \(\Sigma^{(o)}\) defined as above, and \(\Sigma^{(k)} = (\tilde{v})^k \Sigma^{(o)}\) we therefore use the following inequality to estimate \(\rho(\tilde{v})\):
\[
\rho(\tilde{v}) \leq \max_{1 \leq i \leq m} \frac{|x_i^{(k)}|}{|x_i^{(k-1)}|} = \lambda^{(k-1)}_{\max}. \quad (6.39a)
\]

The above technique for finding an estimate of the spectral radius of a tri-diagonal matrix, is identical to the one presented by Spanier (1967) pp.227-228, for incorporation in his "BCT-1" program. In this program, he says that the value \(\hat{k}=6\) in (6.39a) was found to give adequately sharp estimates of the spectral radius. In our case, it was indeed found that \(\hat{k}\) need be no larger than 10. Alternatively, a convergence criterion of the form
\[
|\lambda^{(k)}_{\max} - \lambda^{(k-1)}_{\max}| < \varepsilon, \quad k=1, 2, \ldots,
\]
could be used, where \(\varepsilon\) is some error reduction factor. Hence, the upper bound \(\beta'\), we require for the spectral radius of \(\tilde{v}\) is given by
\[
\beta' = \lambda^{(k-1)}_{\max} \text{ for some } k=10. \quad (6.39b)
\]

Having now obtained estimates for \(\alpha'\) and \(\beta'\), and hence for \(\alpha\) and \(\beta\), we can proceed to evaluate \(\tilde{\rho}\) from (6.25) for the one parameter case of the Pasceman-Rachford method (6.13a) and (6.13b). [Or rather (6.13b)' in practice], and also to evaluate the sequence of parameters \(\tilde{p}_j (k \leq j \leq M=2^k)\) from Nachpress (1962). Further, we can compute the
bounds for the average spectral radii and average rates of convergence given by (6.34) and (6.35) respectively. These computations, along with experimental results using the iterative method in question, are given in section 6.8.

6.7 A non-commutative case

As mentioned earlier, Macmillan (1964) has considered the solution of (6.1) on a pie-shaped region in the (r=0) plane, and for the A.D.I. (Peaceman-Rachford) iterative method recommends normalising the difference equation (6.5) to obtain equation (6.12), so that the difference operators in the two coordinate directions commute with one another. Largely because of Theorem 3.4, this approach is the most attractive one theoretically. However, we will provide computational evidence which suggests that for the problem defined in section 6.2, the most rapid convergence in practice (especially on the circle) can be obtained by applying the above iterative method to the finite difference equations (6.4), given in matrix form by (6.9). As stated at the end of section 6.4, the \( \hat{H} \) and \( \hat{V} \) matrices defined by (6.10) and (6.8) do not commute, and thus the bulk of the theory in section 6.3 is not theoretically applicable in this section. In what is to follow, to try and overcome this difficulty, we will make some intuitive assumptions.

The Peaceman-Rachford method for the matrix problem (6.9) can be derived in the same way as (6.13a) and (6.13b), so that we obtain

\[
(p_{k+1} I-\hat{H})u^{(k+1)} = (p_{k+1} I-\hat{V})u^{(k)} + b, \quad k \geq 0 \tag{6.40a}
\]

\[
(p_{k+1} I-\hat{V})u^{(k+1)} = (p_{k+1} I-\hat{H})u^{(k+1)} + b. \tag{6.40b}
\]

In an analogous way, a more efficient system of the form (6.13b)' can replace (6.40b), so for this case we have

\[
(p_{k+1} I-\hat{V})u^{(k+1)} = 2p_{k+1}u^{(k+1)} - (p_{k+1} I-\hat{V})u^{(k)}. \tag{6.40b}^'
\]
Likewise the P.R. iteration matrix for method (6.40), with a constant parameter \( p_{k+1} = p \) (for all \( k \)) is

\[
\hat{\mathbf{p}}_n = (p_1 + \hat{\mathbf{v}})^{-1} (p_1 - \mathbf{u}) (p_1 + \hat{\mathbf{u}})^{-1} (p_1 - \mathbf{u}) \]

\[
= F^{-1} (p_1 + \hat{\mathbf{v}})^{-1} (p_1 - \mathbf{u}) (p_1 + \hat{\mathbf{u}})^{-1} (p_1 - \mathbf{u}) p_1 \]

\[
= F^{-1} \hat{\mathbf{p}}_n F (\text{say}) ,
\]

where

\[
\hat{\mathbf{u}} = \frac{\Delta T}{A_0} F^{-1} \mathbf{u} F^{-1} , \quad \mathbf{v} = \frac{\Delta T}{A_0} F^{-1} \mathbf{v} F^{-1} \quad \text{see (6.8)-(6.10)} .
\]

From (6.41) we have

\[
\rho(\hat{\mathbf{p}}_n) = \rho((p_1 + \hat{\mathbf{v}})^{-1} (p_1 - \mathbf{u}) (p_1 + \hat{\mathbf{u}})^{-1} (p_1 - \mathbf{u})) = \rho(\hat{\mathbf{p}}_n) .
\]

By a similar argument to that regarding equations (6.16), it can be shown that \( \hat{\mathbf{u}} \), \( \mathbf{v} \) are respectively non-negative, positive definite matrices. Hence by Theorem 3.3, which was proved in section 6.5, we have

\[
\rho(\hat{\mathbf{p}}_n) = \rho(\hat{\mathbf{p}}_n) < 1, \text{ for all } p > 0 .
\]

Using the same argument as that given in section 6.5, viz. equations (6.15)-(6.19), we obtain

\[
\rho(\hat{\mathbf{p}}_n) \leq \max_{1 \leq i \leq n} \left| \frac{p - \mathbf{v}_1}{p - \mathbf{v}_1} \right| ,
\]

where \( \mathbf{v}_1 \) (1 value) are the eigenvalues of \( \mathbf{v} \), which from (6.42) has the same eigenvalues as \( \hat{\mathbf{v}} = F^{-1} \mathbf{v} F^{-1} = (\frac{\Delta T}{A_0}) F^{-1} \mathbf{v} \); see (6.9).

Hence we have shown that the P.R. method (6.40), for solving problem 6.2, is convergent for a single parameter, \( p > 0 \). Because \( \hat{\mathbf{u}} \) and \( \hat{\mathbf{v}} \) have not a common set of eigenvectors, a similar analysis to that in section 6.5 for proving convergence and obtaining a sequence of "optimum" parameters, is impossible.

If we define \( T_{p_j} \) in terms of two general matrices \( \mathbf{u} \) and \( \mathbf{v} \) as

\[
T_{p_j} = (p_j I + \mathbf{v})^{-1} (p_j I - \mathbf{u}) (p_j I + \mathbf{u})^{-1} (p_j I - \mathbf{v}) ,
\]

then the following two theorems can be proved for the non-commutative
case. We state them only, and refer the reader to Wachpress (1966) pp. 210-213 for the proofs. The first was originally demonstrated by Pearcy (1962), and although we have changed the hypothesis slightly (to cope with our particular problem), the proof follows in a similar manner.

**Theorem 6.1**

Let $H$ and $V$ in (6.45) be respectively non-negative and positive definite. Order the iteration parameters in a monotonically non-increasing sequence:

$$\hat{s} \geq p_1 > p_2 > \ldots > p_k > \hat{a},$$

where $\hat{s}$ and $\hat{a}$ are the real positive eigenvalue bounds of $V$. Then there exists an $n_0$ such that for all $t > n_0$ the spectral radius of

$$\prod_{j=1}^{t} T_{P_j}$$

is less than unity.

**Theorem 6.2**

Let $H$ and $V$ be real and symmetric and $A = H + V$ (the coefficient matrix from the system of equations $A\mathbf{x} = \mathbf{b}$) be positive definite. Let the iteration parameters be determined by Jordan's method (see for example, Wachpress (1966) section 6.2.3) as though $H$ and $V$ commuted. Apply the $p_k$ in monotonically decreasing order. Then there exists an $n_0$ such that for all $t > n_0$, the iteration matrix $(\prod_{j=1}^{t} T_{P_j})$ has a spectral radius less than $\mu(n_0) < 1$, where $\mu(n_0)$ may be as small as we choose.

The matrices $H$, $V$ defined by (6.42) and $A = H + V$ satisfy the hypotheses in Theorems 6.1 and 6.2. Hence if the parameters $p_{k+1}$ ($k > 0$) satisfy the remaining hypotheses in these two theorems, the iterative method (6.40) will converge.

Now in (6.41), it was shown that $T_{Pj} = F^{-1} T_{Pj} F$, and in a similar it can be shown that

$$T^{(t)} = \prod_{j=1}^{t} T_{P_j} = F^{-1} \left( \prod_{j=1}^{t} T_{P_j} \right) F = F^{-1} T_{H(t)} F,$$  (6.46)
Hence, \( \rho(\frac{M}{\prod_{j=1}^{N} P_j}) < 1 \) implies \( \rho(\frac{M}{\prod_{j=1}^{N} P_j}) < . \) Therefore, in applying the above two theorems to the iterative method (6.40), we require the real eigenvalue bounds \( \tilde{a}, \tilde{b} > 0 \) for the eigenvalues of \( \tilde{V} \) (whose eigenvalues are also those of \( \tilde{V} \)).

Theorem 6.2 then gives a method of choosing a sequence of parameters in the non-commutative case. Below, we shall apply this technique to method (6.40) for the solution of problem 6.2, where \( R \) is the unit circle, and subject to the specified boundary conditions. As in section 6.5, we shall use Wachpress parameters of \( \lambda = 2^k \) type, and since these are identical to the Jordan parameters for these values of \( N \), provided we choose \( N \) large enough and use the parameters in monotonically decreasing order, Theorem 6.2 assures convergence.

The Unit Circle

For the circle, \( \hat{H} \) and \( \hat{V} \) in equations (6.40) are defined by (6.10) and (6.8), and \( r_i = (i+1)\Delta r, i = 1, 2, \ldots, m \). The iteration matrix of method (6.40) is given by (6.46). Since \( \hat{T}^{(N)} \) is similar to \( \hat{T}^{(M)} \) we choose to investigate \( \rho(\hat{T}^{(N)}) \).

Because \( \hat{H} = \frac{A_r}{20} F^{-1} H F^{-1} \), \( \hat{V} = \frac{A_r}{20} F^{-1} V F^{-1} \), they are respectively non-negative and positive definite, since \( H \) and \( V \) have these respective properties (see previous arguments). Thus \( \hat{A} = \hat{H} + \hat{V} \) is positive definite, so that the first part of the hypothesis in Theorem 6.2 is satisfied.

Next, we assume that \( \hat{H} \) and \( \hat{V} \) commute, so in this case we can proceed to derive our sequence of parameters in the same way as we did for the true commutative case of section 6.5.

Repeating then the analysis of section 6.5 with \( \hat{H}, \hat{V} \) replacing \( \hat{U}, \hat{V} \), we arrive at the point where we must obtain the spectral bounds \( \overrightarrow{a}, \overrightarrow{b} \) for the positive real eigenvalues of \( \hat{V} \). In a similar manner to the way \( \phi \) in equation (6.22) was derived, the matrix \( \hat{V} \) (assuming we again have a peripheral ordering of our internal mesh points) can be shown to be similar to
\[
\begin{bmatrix}
\n-\frac{(r_1+\Delta r/2)}{\sqrt{r_1 r_2}} & 0 & 0 & \cdots & 0 \\
0 & -\frac{(r_2+\Delta r/2)}{\sqrt{r_2 r_3}} & \frac{(r_3-\Delta r/2)}{\sqrt{r_3 r_1}} & \cdots & 0 \\
0 & 0 & \cdots & \cdots & 0 \\
\end{bmatrix}
\]

By Theorem 1.1, Corollary and (3.65a), \(\mathbf{V}^t\) is an irreducible Stieltjes matrix. Thus, Lemma 6.1 is applicable to this matrix.

Similarly, the matrix \(|\mathbf{V}^t| = D\mathbf{V}^t D\) (where \(D\) is defined by (6.38)) satisfies the hypothesis of Lemma 6.2. Thus we can use the information in section 6.6 to estimate the eigenvalue bounds \(\bar{a}', \bar{b}'\) for the eigenvalues of the matrix \(\mathbf{V}^t\). Alternatively, to estimate \(\bar{a}'\), we could use Theorem 1.4, Corollary to show, that since \(\bar{V} = (\Delta r/\Delta \theta) \mathbf{V}^{-1} \mathbf{V}^t = \mathbf{V}^{-1} \mathbf{V}^t\) from (6.10) and (6.48)

\[
\bar{b}' = \rho(\bar{V}) = \rho(\mathbf{V}^t) \leq 4.
\]

Having obtained our spectral bounds \(\bar{a} = \bar{a}', \bar{b} = \bar{b}'\) we can now determine our parameters, \(p_j, 1 \leq j \leq 2^k, k \geq 0\), to be used in monotonically decreasing order with method (6.40).

**Computing the solution at the centre point**

The solution of (6.1) at the centre of the circle is found by replacing the polar coordinate form of Laplace's equation by

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad u = u(x, y).
\]
Then adopting a similar procedure to that given on pages 109-110 for the solution of $\frac{\partial^2 U}{\partial t^2} = V^2 U$ at the centre point, yields,

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 4\left(U_M - U_0\right)/\left(\Delta r\right)^2 + O(\Delta r^2),$$

(6.50)

where $U_M$ is the mean value of $U$ round the peripheral of radius $\Delta r/2$, i.e. $U_M = \left(\sum_{j=1}^{n} U_{1,j}\right)/n$. Hence, neglecting the local truncation error term, (6.50) yields

$$U_0 = U_M,$$

(6.51)

where $U_0$ is the required finite difference solution of (6.1) at the centre of the circle. The values $U_{1,j}$ (1$sj$n) have been obtained from the main iterative process, and we have assumed that they are sufficiently accurate approximations to the finite difference solutions.

**The ring (axial)**

Next, we consider the solution of problem 6.2 on the ring (axial) with the specified boundary conditions. We again use the F.R. iterative method (6.40) to solve the matrix problem (6.9), which represents the set of difference equations (6.4) approximating the solution of the differential equation (6.1). In this context, the variables $r_i$ are defined by $r_i = r + i\Delta r, \ i=1,2,\ldots,m$.

For this problem, we intend to use a different approach to estimate rates of convergence and to obtain acceleration parameters $p_j (j=1,2,\ldots,k)$ for the iterative process given by equations (6.40). Firstly, we point out that we can choose a rectangle which is topologically equivalent to some ring (axial). Let us then consider the solution of Laplace's equation,

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0, \ U = U(x,y), \ (x,y) \in R,$$

(6.52)

where $R$ is the rectangular region

$$0 \leq x \leq l_1, \ 0 \leq y \leq l_2,$$

(6.53)

with a periodicity condition in the $x$-direction.
\[ U(x_2, y) = U(x+\delta x_1, y) \quad \delta = \pm 1, \]

and Dirichlet boundary conditions

\[
\begin{align*}
U(x, y) &= f_1(x), \quad 0 \leq y < 0; \\
U(x, y) &= f_2(x), \quad 0 < y < l_2.
\end{align*}
\]

We superimpose a rectangular grid with mesh lengths \( \Delta x = \frac{1}{n} \) in the \( x \)-direction and \( \Delta y = \frac{l_2}{(n+1)} \) in the \( y \)-direction, such that \( \Delta y = \Delta x \), where \( \Delta x \) is the mesh length in the \( r \)-direction of the polar grid superimposed on the ring (as before). We also stipulate that the number of \( x \)-lines on the rectangular grid should correspond to the number of \( \theta \)-lines (peripherals) on the polar grid. In other words, we have \( l_2 = 1-a \).

If we could bend the rectangle so that the sides \( x=0 \) and \( x=1 \) joined together, a ring would be formed, and this would make sense, since \( U(0, y) = U(l_1, y) \). Thus the problem on the rectangle has a connection with the one of solving Laplace's equation on a ring. Because of this analogy between the two problems, which henceforth we shall refer to as the "rectangle" and "ring" problems, it has been decided to investigate the analogous Peaceman-Rachford iterative method for solving the former problem. The purpose being, to derive a set of Wachpress \((k=2^k)\) parameters theoretically for the former case, which parameters we intend to use in connection with the P.R. method (6.40) applied to the "ring" problem. Later, we shall give numerical evidence, which substantiates the above analogy drawn between the two problems.

To obtain our finite difference equations for the "rectangle" problem, we use equations of the form (1.4) and (1.6), with \( \Delta x \) replacing \( h \) in the former, and \( \Delta y \) replacing \( h \) in the latter. This yields

\[
\left[ \frac{u_{i+1, j} - 2u_{i, j} + u_{i-1, j}}{\Delta x^2} \right] + \left[ \frac{u_{i, j+1} - 2u_{i, j} + u_{i, j-1}}{\Delta y^2} \right] = 0. \quad (6.55)
\]

We are now at the same stage as we were when differencing equation (6.1) to obtain (6.3). Since in that instance we multiplied (6.3) by \(-\Delta x^2\) to
give us equation (6.4), here we multiply (6.55) by \( -\Delta y^2 \) (since \( \Delta x = \Delta y \)) to obtain

\[
\left[ -\theta x u_{i-1,j} + 20 x_{i,j} - x_{i+1,j} \right] + \left[ -u_{i,j} - 20 u_{i,j} - x_{i-1,j} \right] = 0, \tag{6.56}
\]

where \( \theta x = \Delta y^2 / \Delta x^2 \).

Using the same approach as that described in section 3.5, we let the bracketed terms define the matrices \( V_T \) and \( V_1 \) respectively. The Peaceman-Rachford (A.D.I.) iterative method is then defined as usually by (3.68) with \( r_k = r_{k-1} \). Since this \( H_T \) and \( V_1 \) \emph{commute}, and hence have a common set of eigenvectors \( Y_s \) \emph{say}, \( \text{Ogse} = 1 \), \( \text{lgstam} \), each \( Y_s \) is an eigenvector of the matrix \( T_p \), given by (3.70) and

\[
T_p Y_s = \begin{bmatrix} p - 40 x \sin^2(y/n) \frac{1}{p + 40 x \sin^2(y/n)} & p - 4 \sin^2[y/(2(m+1))] \frac{1}{p + 4 \sin^2[y/(2(m+1))]} \end{bmatrix} Y_s, \tag{6.57}
\]

where \( \lambda_{(1)} = 40 x \sin^2(y/n) \), \( v_{(1)} = 4 \sin^2[y/(2(m+1))] \), \( \text{Ogse} = 1 \), \( \text{lgstam} \), are the respective eigenvalues of the matrices \( H_T \) and \( V_1 \).

For optimun rate of convergence, we must choose \( p \) to minimise

\[
\rho(T_p) = \max_{\text{Ogse} = 1} \left( \max_{\text{lgstam}} \begin{bmatrix} p - 40 x \sin^2(y/n) \frac{1}{p + 40 x \sin^2(y/n)} & p - 4 \sin^2[y/(2(m+1))] \frac{1}{p + 4 \sin^2[y/(2(m+1))]} \end{bmatrix} \right) < 1. \tag{6.58}
\]

From equations (3.76)-(3.79) we thus have

\[
\begin{align*}
\min_{p > 0} \rho(T_p) &= p(p-1) = \frac{1 - \tan[y/(2(m+1))]}{1 + \tan[y/(2(m+1))]} \tag{6.59}.
\end{align*}
\]

[where \( p = 2 \sin(y/(m+1))] \), \emph{independent of} \( n \), the number of \emph{y}-lines on the grid. For large \( n \),

\[
\min_{p > 0} \rho(T_p) = 1 - \frac{1}{m+1}, \tag{6.60}
\]

\( \text{where} \ m \ \text{is the number of} \ x\text{-lines (which correspond to the number of peripherals on the polar grid on the ring)}. \)
The rate of convergence for small $n$, from (6.39) is given by

$$R(\text{P.R. fixed parameter}) = -\log \left[ \frac{1-\tan \frac{\alpha}{2}}{1+\tan \frac{\alpha}{2}} \right],$$

(6.61)

where $\alpha = \pi/(2(m+1))$. For large $n$, from (6.60), the rate of convergence is

$$R_\infty(\text{P.R. fixed parameter}) = -\log \left[ 1- \frac{n}{m+1} \right] = \frac{n}{m+1}.$$

(6.62)

We now consider the product ($\prod_{j=1}^{N} P_j$) of $N$ P.R. matrices corresponding to $N$ iterations of the method with a sequence of $N$ positive parameters $p_j$. Hence if $\{ \lambda_d \}$ is a common orthonormal basis of eigenvectors of $H_1$ and $V_1$ (see Theorem 3.4) we have

$$\left[ \prod_{j=1}^{N} P_j \right] \lambda_d = \left[ \prod_{j=1}^{N} \begin{bmatrix} p_j - \lambda_d^{(1)} \\ p_j + \lambda_d^{(1)} \\ p_j^2 + v_d^{(1)} \end{bmatrix} \right] \lambda_d,$$

where $\lambda_d^{(1)}$, $v_d^{(1)}$ are respectively the eigenvalues of the $(nm) \times (nm)$ matrices $H_1$ and $V_1$ corresponding to the eigenvectors $\lambda_d$. Then

$$\rho \left[ \prod_{j=1}^{N} P_j \right] = \max_{1 \leq d \leq n} \left| \frac{p_j - \lambda_d^{(1)}}{p_j + \lambda_d^{(1)}} \right| \left| \frac{p_j^2 - v_d^{(1)}}{p_j^2 + v_d^{(1)}} \right|,$$

(6.63)

Since $\lambda_0 = 4 \sin^2 (\pi/n)$, $\cos \pi/n = 1$, we have $\lambda_0 = 0$, so that

$$\max_{1 \leq d \leq n} \left| \frac{p_j - \lambda_d^{(1)}}{p_j + \lambda_d^{(1)}} \right| = 1.$$

Hence, if $\alpha^{(1)}$, $\beta^{(1)}$ are the spectral bounds for the eigenvalues $\lambda_d^{(1)} (1 \leq d \leq n)$ of the positive definite matrix $V_1$, we want to minimise

$$\max_{1 \leq d \leq n} \left| \frac{p_j - \lambda_d^{(1)}}{p_j + \lambda_d^{(1)}} \right|.$$

(6.64)

Since $v_d^{(1)} = 4 \sin^2 \left[ \pi/(2(m+1)) \right]$, $1 \leq d \leq n$. 


\begin{align}
\alpha^{(1)} &= 4\sin^2\left(\frac{\pi}{2(\omega+1)}\right), \\
\beta^{(1)} &= 4\cos^2\left(\frac{\pi}{2(\omega+1)}\right).
\end{align} \tag{6.65}

Knowing \(\alpha^{(1)}\) and \(\beta^{(1)}\), we can again use Wachpress's method \((1962)\) to give us an exact determination of the \(\bar{\varphi}_j\) for \(\mu = 2^k, \ k = 0,1,2,\ldots\). \(\ldots\)

The solution to the \textit{min-max} problem \((6.64)\) is given by equations \((3.86)-(3.89)\). The average spectral radius has an upper bound given by \((6.34)\). Similarly, the average rate of convergence has a lower bound given by \((6.35)\).

If we put \(k=1\), i.e. take two parameters alternately, we have from \((3.89)\),

\begin{align}
\alpha_1 &= 2\sin(\frac{\pi}{\omega+1}) \ , \ \beta_1 = 2 .
\end{align} \tag{6.66}

Substituting equations \((6.66)\) in \((3.88)\), we obtain

\begin{align}
\sqrt[2]{[\alpha^{(1)},\beta^{(1)}]} &= \sqrt{1-\frac{\sin[\pi/(\omega+1)]}{1+\sin[\pi/(\omega+1)]}} \tag{6.67a}
\end{align}

\begin{align}
\geq 1 - \frac{\pi}{\omega+1} , \text{ for } \omega \text{ large.} \tag{6.67b}
\end{align}

Thus, given the values of \(\alpha^{(1)}\) and \(\beta^{(1)}\) from \((6.65)\) we have been able to obtain \(\mu = 2^k (k\geq 0)\) positive parameters \(\bar{\varphi}_j\) to minimise \((6.64)\), and we obtain bounds on the average spectral radii and average rates of convergence. In each case, the only variable involved is "\(n\)|, the \textit{number of x-lines} on the rectangular grid. As mentioned earlier, "\(n\)" also denotes the \textit{number of peripherals} on the polar grid (on the ring). So what we propose to do, is to use these self-same parameters \(\bar{\varphi}_j\) \((1\leq j\leq 2^k)\) just derived for the "rectangle" problem, for the "ring" problem, i.e. in equations \((6.40a)\) and \((6.40b)\).

In section 6.8, we give numerical results showing the success of this technique. In particular, a comparison of convergence rates with the S.F.O.R. method, shows a marked similarity with the comparison of convergence rates given by Wood \((1971)\) between the analogous F.R. (A.D.I.) and one x-line S.O.R. methods on the "rectangle" problem.
6.9 Computational results

As stated in section 6.1, having investigated the use of the Peaceman-Rachford iterative method in \((r-\theta)\) geometry, it is now our intention to compare rates of convergence with block S.O.R. techniques for the same problem. Benson (1969) and Benson and Evans (1972) have studied Successive Peripheral Over-Relaxation (S.P.O.R.) and Successive Radial Over-Relaxation (S.R.O.R.) for the solution of Laplace's equation (in \((r-\theta)\) geometry) on the ring \((a<r<b)\) and on the unit circle subject to Dirichlet boundary conditions. Both methods are like the 1-line S.O.R. process described in section 3.4, in that the first requires the implicit solution (by Algorithm 4.3) of periodic tridiagonal systems of equations along successive peripherals (or circumferences) of the polar grid, and the second, the implicit solution (by Algorithm 4.2) of tridiagonal systems along successive radii of the polar grid. The mesh points for these two methods will of course be ordered as in Figures 5.7, 5.9 and 5.8, 5.10, respectively. The finite difference analogue they use in both cases is given by (6.4). They showed that with a peripheral ordering of the internal grid points, the coefficient matrix of the difference equations possesses block Property A and is block consistently ordered (Arms et al., 1956) whereas with a radial ordering, the coefficient matrix has only the first property. Hence, the block S.O.R. theory is only valid for S.P.O.R. They also demonstrated by numerical experiment that for the problem(s) mentioned above, S.P.O.R. converged considerably faster than S.R.O.R., e.g. for the ring with \(a=0.2\), \(b=1/20\), \(A\theta=2\pi/80\), i.e. 15 peripherals and 80 radii on the polar grid, the number of iterations required for convergence for the same error reduction factor \(e\) [see (3.28)] was 20 for S.P.O.R. and 60 for S.R.O.R. The difference in the number of iterations increased (decreased) as the number of peripherals on the grid increased (decreased). For more detail see Benson and Evans (1972), p.78.
Macmillan (1964) reaches the same conclusion as the above two authors after considering S.O.R. techniques for solving Laplace's equation in a pie-shaped region of the \((r=0)\) plane. He uses finite difference equations of the form (6.5). By assuming separability, i.e. \(U(r,0) = f(r).g(0)\), and the condition \(\partial U/\partial r = 0\) at \(r=0\), he is able to develop the theory for the equivalent of the S.F.O.R. method on his region (i.e. an ordering of the mesh points along successive \(r\)-lines). For S.F.O.R. he has to assume that the region of problem solution is bounded away from the origin (of the \((r=0)\) coordinate system) by the circle \(r=r_{min}\) where \(r_{min}\) is small, and that \(U(0)/\partial r = 0\). With these assumptions he shows that S.F.O.R. (or \(a\text{muthal line over-relaxation}\) as he calls it) converges much faster than S.R.O.R. (= \(r\text{adial line over-relaxation}\)). By considering an equivalent commutative version of the P.R. (A.D.I.) method to that given by equations (6.13), he comes to the conclusion that this technique is comparable to S.R.O.R. with regard to rates of convergence. He also concludes that the slow convergence of these two methods "is caused by the nature of the difference equations near the origin of the \((r=0)\) coordinate system".

In Figure 6.3 and Table 4, we given numerical evidence which shows that when the value "a" (inner radius of the ring) decreases, with "a" remaining constant, then the rate of convergence of the P.R. method (6.13) decreases. In otherwords, it would appear that it is the value of \(\min r_1 = r_1^a + Dr_1\), particularly, as well as \(m_1\), the number of peripherals (or the order of the matrix \(V^\prime\)), that govern the rate of convergence of this method.

In view of the conclusions of the above three authors, we will make comparisons of rates of convergence just between the P.R. (A.D.I.) method and S.F.O.R. We will not make any of the assumptions made by Macmillan, but will give experimental evidence from a particular "model problem" to compare them. We will also give encouraging results for
the non-commutative version of the P.R. method (discussed in section 6.7) for these problems. This does not appear to have been considered before.

The problem(s) we are interested in are given in section 6.2. The two regions of problem solution are the unit circle and the ring (as real). We will consider the former region first.

The unit circle

In sections 6.5 and 6.7 we have considered a "commutative case" and a "non-commutative case" respectively, of the P.R. (A.D.I.) method applied to this problem. We will now give some results to compare the relative merits of the two cases.

Let us first assume that from section 6.6 we have obtained the eigenvalue bounds $\alpha', \beta'$ (commutative case) and $\bar{\alpha}, \bar{\beta}$ (non-commutative case) for the matrices $\tilde{V}$ (6.23) and $\tilde{V}$ or $\bar{V}$ (6.47), using the methods (6.13) and (6.40). Then for one fixed parameter, (6.27) gives an expression for the minimum spectral radius (commutative case). The bounds $\alpha', \beta'$ obtained from section 6.6 are not exact, but are very close approximations to the true values, and we will therefore assume that with these computed eigenvalue bounds, (6.27) is "very nearly" satisfied. From (6.44), for a single parameter in the non-commutative case, we have

$$\min_{p>0} \rho(p) \leq \min_{p>0} \left\{ \max_{1 \leq i \leq m} \left| \frac{p - \tilde{v}_i}{p + \tilde{v}_i} \right| \right\} = \frac{\sqrt{\bar{\alpha} \bar{\beta}} - \bar{\alpha}}{\sqrt{\bar{\alpha} \bar{\beta}} + \bar{\alpha}}$$  \hspace{1cm} (6.68)

from equations (3.76)-(3.79) ("\tilde{\alpha}\) denotes "approximately equal to", in view of the fact that $\bar{\alpha}, \bar{\beta}$ although very good approximations to the true eigenvalue bounds, are not necessarily exact). The "optimum" value of $p$ which minimizes the quantity on the left of the inequality in (6.68) is given by

$$p = \sqrt{\bar{\alpha} \bar{\beta}} .$$  \hspace{1cm} (6.69)
We can then re-write (6.68) as

\[
\min_{\rho > 0} \frac{\rho(\tau) \cdot \frac{\Delta^2}{\rho \Delta \dot{\omega}^2}}{1 + \sqrt{\frac{a}{\beta}}} = \frac{1 - \sqrt{\frac{a}{\beta}}}{1 + \sqrt{\frac{a}{\beta}}}. \tag{6.70}
\]

In Table 1 below, we compare the spectral bounds \((a', \beta')\) and \((\tilde{a}, \tilde{\beta})\) for the eigenvalues of the \((n \times n)\) matrices \(\tilde{V}'\) and \(\tilde{V}\). As usual, "n" denotes the number of peripherals (unknowns per radius) on the polar grid.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(a')</th>
<th>(\tilde{a})</th>
<th>(\beta')</th>
<th>(\tilde{\beta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.00629</td>
<td>0.18761</td>
<td>1.68166</td>
<td>3.81321</td>
</tr>
<tr>
<td>10</td>
<td>0.00123</td>
<td>0.05218</td>
<td>2.34790</td>
<td>3.97544</td>
</tr>
<tr>
<td>15</td>
<td>0.00047</td>
<td>0.02401</td>
<td>2.67560</td>
<td>3.99940</td>
</tr>
<tr>
<td>20</td>
<td>0.00024</td>
<td>0.01374</td>
<td>2.87651</td>
<td>4.00000</td>
</tr>
</tbody>
</table>

We notice in particular from this table, that the \((a')\)'s are quite close to zero (more so as \(n\) becomes larger) and indeed much smaller than the \(\tilde{\omega}\)'s. This accounts for the big difference in the rates of convergence of the two cases. Using (6.27) and (6.70) with Definition 3.4, Table 2 gives a comparison of the rates of convergence using the results of Table 1.

<table>
<thead>
<tr>
<th>(n)</th>
<th>Rate of convergence</th>
<th>Lower bound for rate of convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.12251</td>
<td>0.45112</td>
</tr>
<tr>
<td>10</td>
<td>0.04579</td>
<td>0.23014</td>
</tr>
<tr>
<td>15</td>
<td>0.02659</td>
<td>0.15527</td>
</tr>
<tr>
<td>20</td>
<td>0.01827</td>
<td>0.11735</td>
</tr>
</tbody>
</table>
Again, the middle column will not be exact values since we are using approximate bounds \((a', b')\). However, because the latter are very sharp estimates, then so too are the rates of convergence above. It is fairly clear then from Table 2, that for \(m=5(5)20\) the non-commutative case has faster rates of convergence with a single parameter than the corresponding commutative case. The above table was taken from a larger one with \(m=3(1)20\), giving the same trend as that above.

However, it is the use of a sequence of positive acceleration parameters \(p_j\) and the corresponding rates of convergence that we are mainly interested in. As stated earlier, we will use Wachpress parameters of the type \(M=2^k\) \((k>0)\) throughout \((M=\text{number of parameters})\), and from now on will refer to them as "the parameters". As far as the non-commutative case is concerned, we cannot extend the theoretical treatment of rates of convergence of the one parameter case to several parameters. We can however obtain some results on average rates of convergence for the commutative case from equation (6.35) and equations (3.88) and (3.89). Table 3 gives lower bounds on the average rates of convergence of method (6.13) applied to the problem in question, using \(M=2, 4, \) and \(8\) parameters, a different set for each value of \(m\) (since \(\tilde{v}'\) and therefore \(b', a'\) change with \(m\) — see (6.23) and Table 1).

<table>
<thead>
<tr>
<th>(m)</th>
<th>(M=2)</th>
<th>(M=4)</th>
<th>(M=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.38447</td>
<td>0.53431</td>
<td>0.62094</td>
</tr>
<tr>
<td>10</td>
<td>0.21728</td>
<td>0.37877</td>
<td>0.46534</td>
</tr>
<tr>
<td>15</td>
<td>0.16483</td>
<td>0.31907</td>
<td>0.40354</td>
</tr>
<tr>
<td>20</td>
<td>0.13599</td>
<td>0.28492</td>
<td>0.37124</td>
</tr>
</tbody>
</table>

The trend shown in the above table was the same for similar sets of values corresponding to \(m=3(1)20\).
We now come to the experimental work carried out. The problem posed on the unit circle was to solve equation (6.1), subject to the boundary conditions

\[ U(1, \theta) = 0 . \]

The commutative and non-commutative versions of the P.R. (A.D.T.) iterative method described in sections 6.3 and 6.7 respectively, using \( 1, 2, 4 \) and \( 8 \) parameters, were compared with the S.P.O.R. technique using the optimum \( w \) at every stage. The latter results were provided by Dr. A. Benson (1973). The initial guess (approximation) to the finite difference solution in each case was

\[ u^{(0)}_{i, j} = u^{(0)}(x_i, \theta_j) = 1.0, \quad i=1, 2, \ldots, n; j=0, 1, \ldots, n-1. \]

The iterations were continued until

\[ \max_{0 \leq j \leq n-1} |u^{(k+1)}_{i, j} - u^{(kM)}_{i, j}| \leq 5 \times 10^{-6}, \]

in each case (\( M=1 \) in the case of S.P.O.R.). Also, for each method, \( \Delta r \) and \( \Delta \theta \) were chosen to be approximately equal (viz. if \( \Delta r = 2/(2m+1) \), then \( n \) was chosen to be the nearest integer to \( (2m+1)v \), since \( \Delta \theta = 2\pi/n \)).

Figure 6.1 shows the number of iterations \( N \) plotted against \( m \), the number of peripherals, for \( m=3(1)19 \), using the commutative version of the P.R. method \((N=1, 2, 4 \text{ and } 8)\) given by equations (6.13) and using S.P.O.R. The results for the P.R. procedure are consistent with those given in Table 3. Although the use of 4 and 8 parameters gives a vast improvement in convergence rates over the one parameter case, we see from Figure 6.1 that the number of iterations is only smaller than for S.P.O.R. when \( m>18 \). Extra computations performed using 4 and 8 parameters for \( m=18(1)30 \), showed that 8 parameters gave the fastest convergence with \( N=8 \) for this range of values of \( m \).

As stated on page 48, we must compare the amount of work required per iteration, as well as the number of iterations required for convergence.
In implementing the S.P.O.R. method, one must solve a periodic tri-diagonal system of equations at each iteration, using Algorithm 4.3 (say). Similar systems of equations are solved at the first step (6.13a) of the P.R. (A.D.I.) method. In both methods the matrix elements remain unchanged after the first iteration, so that the amount of work reduces to that given at the end of Algorithm 4.3. Hence, we can state that the first step (6.13a) of the P.R. (A.D.I.) method requires approximately the same amount of work as one iteration using S.P.O.R.. Although for the experimental results given here, we used (6.13b) as the second step of the former process, it is more efficient to use (6.13b)''. In that case, we can use Algorithm 4.2 to solve the tri-diagonal system arising from it. Again the amount of work reduces after the first iteration to that given at the end of the algorithm, i.e. approximately half that required by Algorithm 4.3. Hence for each iteration, the P.R. procedure requires approximately \( \frac{1}{2} \) times as much work (arithmetic operations) as the S.P.O.R. technique.

Extending the "line" for S.P.O.R. in Figure 6.1 from \( w=19 \) to \( w=30 \), reveals that with 8 parameters, the P.R. procedure is only faster (in terms of total arithmetic operations) than S.P.O.R. for approximately \( w>28 \), i.e. \( A=\frac{2}{1+2w} \). We would therefore have to require quite an accurate solution of this problem to justify the use of method (6.13) (with 8 parameters) in preference to S.P.O.R.

Figure 6.2 shows the non-commutative version of the P.R. procedure \( (N=1,2,4 \) and 8) given by equations (6.40) against S.P.O.R.\( _i \) using the same axes as Figure 6.1 and solving the same problem as before. For the one parameter case we used (6.69) in equations (6.40). The value of \( \hat{p} \) obtained from (6.69) was compared with the value of \( p \) which in experiments carried out with \( w=3(2)9 \), gave the least number of iterations.
for convergence. In each case the former was within 0.05 of the
optimum experimental parameter obtained, and the respective number of
iterations differed by at most two. The sequences of parameters (n=2,4
and 8) were obtained in the manner described for "The unit circle" in
section 6.7. In view of Theorems 6.1 and 6.2, these parameters were
used in a \textit{monotonically decreasing order}. We point out however, that the
required minimum number of parameters to ensure convergence (from the
criterion of the above two theorems) was \textit{not} ascertained, since our
main aim was to achieve convergence \textit{experimentally} in the smallest
number of iterations, in order to obtain a realistic comparison with
the S.P.O.R. method.

Looking at Figure 6.2, we see that on the whole the fastest
convergence is obtained, with regard to A.D.I., using \(M=4\) parameters.
We also notice that for \(n>9\), A.D.I. (\(M=4\)) converges in fewer iterations
than S.P.O.R. However, we must take into account the comments made
above regarding the comparative amounts of work per iteration for the
two methods. Analogous remarks apply to this case, i.e. the P.R.
procedure (6.40a) and (6.40b) requires approximately \(1 \frac{1}{4}\) times as
much work (arithmetic operations) per iteration as S.P.O.R. With
this in mind, it would appear from Figure 6.2, that A.D.I. (\(M=4\))
will give the required solution faster (in terms of total arithmetic
operations) than S.P.O.R. for approximately \(n>16\), i.e. \(\Delta r(=2/(2n+1)<2/33\).
This is a considerable improvement over the results obtaining using
the \textit{commutative version} of the P.R. (A.D.I.) process. The results for
the respective one parameter A.D.I. procedures given in Figures 6.1
and 6.2 bear out the comparisons given in Table 2. From these two
figures we must therefore conclude that the \textit{non-commutative version}
of our P.R. (A.D.I.) methods given by equations (6.40a) and (6.40b)' is
preferable, as far as rates of convergence are concerned, to the
\textit{commutative case}, given by equations (6.13a) and (6.13b)'.

179
The ring (aqrst)

Next, we consider the case where the region of problem solution \( R \) in section 6.2 is the ring (aqrst). We investigate first the commutative case of our P.R. iterative procedures given by (6.13a) and (6.13b)'.

In Table 4 below, we have given representative values of the lower bound \( \alpha' \) and upper bound \( \beta' \) for the positive real eigenvalues of the (m-m) matrix \( \tilde{V} \) defined by (6.23). As illustrated in section 6.3, the average rates of convergence depend on the ratio \( \alpha'/\beta' \). (The "ring" (aqrst) is taken to be the unit circle with the solution to (6.1) given at its centre, as well as on the boundary \( r=1 \).)

<table>
<thead>
<tr>
<th>( n )</th>
<th>( a=0.0 )</th>
<th>( a=0.25 )</th>
<th>( a=0.5 )</th>
<th>( a=0.75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha' )</td>
<td>( \beta' )</td>
<td>( \alpha' )</td>
<td>( \beta' )</td>
</tr>
<tr>
<td>5</td>
<td>0.02208</td>
<td>1.79930</td>
<td>0.07899</td>
<td>2.08537</td>
</tr>
<tr>
<td>10</td>
<td>0.00466</td>
<td>2.39005</td>
<td>0.02373</td>
<td>2.63256</td>
</tr>
<tr>
<td>15</td>
<td>0.00184</td>
<td>2.69982</td>
<td>0.01125</td>
<td>2.90293</td>
</tr>
<tr>
<td>20</td>
<td>0.00094</td>
<td>2.89235</td>
<td>0.00654</td>
<td>3.06939</td>
</tr>
</tbody>
</table>

The above results show that as "a" becomes smaller (a fixed) so does \( \alpha' \). We notice that the value of \( \beta' \) does not vary appreciably for \( a=0.0(0.25)0.75 \), and hence it is mainly the decrease in the value of \( \alpha' \) over this range of values of "a", that causes a corresponding decrease in the ratio \( \alpha'/\beta' \). The trend above was also found for \( a=3(1)20 \).

We illustrate this effect experimentally, by considering the following problem. We wish to solve equation (6.1) subject to the boundary conditions

\[
U(a, \theta), U(1, \theta) = 1.0, \quad a = 0.0(0.25)0.75.
\] (6.71)

In each case the initial approximation to the finite difference solution was

\[
u^{(0)}(r_j, \lambda \theta) = 0.2, \quad r_j = a + j \Delta r \quad (j = 1, 2, \ldots, \lambda) \]
\[
u^{(0)}(r_j, \lambda \theta) = 0.2, \quad j = 0, 1, \ldots, n=1.
\] (6.72)
The iterations were continued until

$$\max_{i,j} |u_{i,j}^{(k+1)} - u_{i,j}^{(k)}| \leq 5 \times 10^{-6}$$

where (6.73)

in this experiment a parameter given by (6.25).

The results are displayed in Figure 6.3, where $N$, the number of iterations, is plotted against $m$, the number of peripherals (unknowns per radius). They are consistent with those results given in Table 4 in that the number of iterations required for convergence increases (for $m$ fixed) as "a" decreases. Since we showed in section 6.5 that the rate of convergence was independent of $\Delta \theta$ (or equivalently $n$ = number of radii), in the above experiment it retained the value $\Delta \theta = \pi/30$ for each value of $m$ and "a".

In Figure 6.4, we give results for the solution of the problem specified above on the ring $(0.5 \leq r \leq 1)$. The axes are the same as before, and we have used the commutative version of the P.R. (A.D.I.) process given by (6.13) with $M=1, 2, 4$ and 8 parameters. These results are compared with those obtained from the S.P.O.R. method for the same problem using the optimum $\omega$ at every stage. The latter results were again provided by Dr. A. Benson (1973). We see that on the whole for $m=3(1)19$, 4 parameters appear to give the swiftest convergence of the $M=1, 2, 4$ and 8 cases. The former converges faster than S.P.O.R. for $m=10$, but in view of the 1/4 fold increase in work per iteration over the latter, P.R. (A.D.I.) $M=4$, will only be quicker in terms of total arithmetic operations for approximately $m>13$, i.e. $\Delta r=1/33$. Again, as we knew from section 6.5 that rates of convergence were independent of $\Delta \theta$, it was kept fixed at the value $\Delta \theta = \pi/30$ [this applies to the S.P.O.R. as well as to the A.D.I. method (6.13)].

We now come to the non-commutative case of our P.R. (A.D.I.) iterative methods, given by equations (6.40), applied to the present problem. Our proposed technique for deriving a set of acceleration
parameters has been described under the title "the ring" in section 6.7.

A study of S.O.R. and P.R. (A.D.I.) iterative methods of solution of the finite difference equations derived from the "rectangle problem" (see sec.6.7) has been given by Wood (1971) and Benson (1969). The former author has shown that the 1 x-line S.O.R. process (which corresponds to S.P.O.R. for the "ring" problem) converges faster than the 2-parameter P.R. (A.D.I.) method when \( m \) (the number of x-lines (peripherals) on the rectangular (polar) grid) is 9, but for \( m \geq 10 \) the roles are reversed. It is also interesting to note that the optimum \( m \) obtained experimentally by the latter author (1973) using S.P.O.R. for solving a "ring problem", agreed to within 0.01, for \( m = 3(1)20 \), with the theoretical value \( \bar{m} = 2/(1+\sin(\pi/(m+1))) \) given by Wood (1971) for the 1 x-line S.O.R. procedure.

Using the "non-commutative" method (6.40) for the approximate solution of equation (6.1), we repeated the experiment described by (6.71)-(6.73) for \( \alpha = 0.3 \). Firstly, the value of \( \bar{p} \) given by (6.59) was compared with the value of \( p \) which in experiments carried out, gave the least number of iterations for convergence. Agreement was obtained to within a margin of 0.05, and the respective number of iterations varied by no more than two in the cases \( m = 3(1)8 \). This is initial evidence to support the afore-mentioned analogy drawn between the "ring" and "rectangle" problems.

For the same problem, Figure 6.5 shows the 2-, 4- and 8-parameter P.R. (A.D.I.) method (6.40) adopting the parameters obtained from the analogous "rectangle problem". Also displayed again are the results of the S.P.O.R. process. From the figure, we see that indeed the 2-parameter P.R. (A.D.I.) converges faster than S.P.O.R. for approximately \( m \leq 10 \), agreeing with the theoretical results mentioned above comparing 1x1S.O.R. and 2-parameter A.D.I. on the "rectangle problem". Using
4 parameters then, it would appear from Figure 6.5, that P.R. (A.D.I.) is faster than S.P.O.R., in terms of total computational effort, for approximately \( m \approx 16 \).

Two runs were completed for this experiment with \( m = 1, 2, 4 \) and 8 and with \( \Delta \theta = \pi/24 \) and \( \Delta \theta = \pi/30 \). In each case the results were the same. This suggests that the rate of convergence is independent of \( m \), the number of radii on the grid, which would agree with the case of the "rectangle problem" where rates of convergence for P.R. (fixed parameter) are independent of \( m \), the corresponding number of \( y \)-lines (lines parallel to the \( y \)-axis) on the grid — see equations (6.61) and (6.62). The acceleration parameters were also applied in monotonically decreasing order but again the results were identical. This again provides evidence that our "ring problem" is like the commutative one on the rectangle.

Lastly, using the single parameter \( p \) given by (6.59), we used the P.R. method (6.40) to solve the above problem on three rings, namely (0.25grsl) (0.5grsl) and (0.75grsl) with \( \Delta \theta \) fixed at \( \Delta \theta = \pi/30 \). In each case, for \( m = 3(1)20 \), the number of iterations required for convergence were almost identical, showing again how the rate of convergence seems to be dependent solely on the variable \( m \).

6.9 Conclusions

With the unit circle and ring (aggrsl) as our regions of problem solution, we have considered a "commutative version" of the P.R. (A.D.I.) method given by equations (6.13) and a "non-commutative version" given by equations (6.40), both for approximating the solution of the differential equation (6.1). Because of Theorem 3.4, the former version is preferable theoretically, in that we can prove convergence for any number of positive parameters \( p_j \), and having obtained the spectral bounds \( a', b' \) for the eigenvalues of \( \bar{V} \) (see (6.23)) from section 6.6, we can rigorously minimise the respective upper bounds.
for the spectral radii of the iteration matrices \( \prod_{j=1}^{M-1} T_j \) for \( M=2^k \), \( k>0 \), using Wachpress's method (1962). As far as the "non-commutative cases" are concerned, Theorems 6.1 and 6.2 are probably among the most general theorems for proving convergence with a sequence of positive parameters. As far as choosing the "optimum" parameters is concerned we are left with the choice of assuming commutativity and then following the "commutative" analysis, or in the case of the ring, obtaining parameters from the analogous "rectangle" problem.

Despite its theoretical appeal, we have seen from Figure 6.1 and Tables 3 and 2 that the convergence rates of the "commutative version" (6.13) on the unit circle are very unattractive. The "non-commutative version" (6.40) however gives far better results as is clear from Table 2 and Figure 6.2, and although we have not conclusively proved convergence, it would appear for the range of values of \( m \) considered, that 4 parameters (used in monotonically decreasing order) is the "best" number to use. As already shown, this latter method (with \( M=4 \)) requires less total computational effort than the corresponding S.P.O.R procedure to obtain the "same" difference solution on the circle, when \( m \) is approximately greater than 16, i.e. \( \Delta r<2/33 \).

For the ring (n\&rs1), the "commutative version" is more attractive, and we have seen from Figure 6.3 that using the optimum single parameter, the number of iterations for \( m \) fixed, decreases as "\( a \)" increases. In particular for \( a=0.3 \), we see from Figures 6.4 and 6.5 that this version with \( M=2, 4 \) and 8 parameters has comparable convergence rates with the "non-commutative case" using parameters from the analogous "rectangle problem". However, the advantage the latter has over the former is that the rate of convergence (for one parameter at least) appears not to vary with the value of "\( a \)". Also, the bounds \( (a^{(1)}, \beta^{(1)}) \)
need to compute the sequences of parameters are given explicitly by (6.65). It is thus for the user to decide whether he wants to use the "commutative case" (6.13) which is more attractive in its theoretical development and not unattractive computationally (for $\alpha > 0.5$, say), or the "non-commutative case" (6.40), which is preferable computationally speaking, but not from a theoretical point of view. For the ring $(0.5 < \alpha < 1)$, 4 parameters appear to give the fastest convergence in each case, for $n=3(1)19$. The "commutative case" required less total computational effort than S.P.O.R. for $n>18$ (approximately), and the "non-commutative case" likewise for $n>16$ (approximately), from Figures 6.4 and 6.5 respectively.
CHAPTER 7

ALTERNATING-DIRECTION PRECONDITIONING TECHNIQUES
IN THE ITERATIVE SOLUTION OF CERTAIN
ELLIPTIC DIFFERENCE EQUATIONS
7.1 Introduction

Suppose that we are required to solve a linear system of equations given in matrix form by $A\mathbf{u} = \mathbf{b}$, where $A$ and $\mathbf{b}$ are known. Then it was shown earlier, in section 3.3, that the optimum rate of convergence for both the method of Simultaneous Displacements (3.15) and Richardson's second degree method (3.46) (for the iterative solution of this matrix problem) even with the correct ("optimum") choice of parameters given by (3.39) and (3.47) respectively, depended upon the value of the P-condition number (see Definition 3.5) of the coefficient matrix $A$.

Consequently, if some form of "preconditioning" to the original system can be effected so as to transform it into a new system whose coefficient matrix has a P-condition number which can be minimised (to less than that of the P-condition number of the original coefficient matrix $A$), then these two basic methods will have improved rates of convergence.

This idea was first conceived by Evans (1967) [see also Evans (1974)] who applied it to the iterative solution of linear equations with symmetric positive definite coefficient matrices. This method of approach is as follows. Suppose that $\mathbf{y}$ is an intermediate transformation vector given by

$$\mathbf{y} = (I - wL^T)\mathbf{u} \quad (7.1)$$

where $w$ is an acceleration parameter to be defined later, and the coefficient matrix $A$ is written as

$$A = I - L - L^T \quad (7.2)$$

where $I$ is the unit matrix, and $L, L^T$ are respectively lower and upper triangular matrices with null diagonals. After multiplication by $(I - wL)^{-1}$ the original system $(A\mathbf{u} = \mathbf{b})$ may be written as

$$(I - wL)^{-1}A(I - wL^T)^{-1}(I - wL^T)\mathbf{u} = (I - wL)^{-1}\mathbf{b}.$$  

This equation can be written as

$$C_w^TA \mathbf{G}_w \mathbf{y} = \mathbf{d} \quad (7.3)$$

or

$$B_w \mathbf{y} = \mathbf{d}.$$
The matrix \( B = (I-\omega L)^{-1} A (I-\omega L^T)^{-1} \) can easily be shown to be positive definite if \( A \) is positive definite. The optimum acceleration parameter, \( \omega_o \), is now defined as that value of \( \omega \) for which the P-condition number of \( B_{\omega} \) is minimised. We can now apply the above two iterative methods to the preconditioned system (7.3). Thus the method of "preconditioned Simultaneous Displacements" corresponding to equation (3.15) is

\[
\chi^{(k+1)} = \chi^{(k)} + \alpha (d-B_{\omega} \chi^{(k)}) ,
\]

and the "preconditioned Richardson's second degree method" corresponding to (3.46) is

\[
\chi^{(k+1)} = \chi^{(k)} + \alpha (d-B_{\omega} \chi^{(k)}) + \beta (\chi^{(k)} - \chi^{(k-1)}) .
\]

The optimum values of \( \alpha \) and \((\alpha, \beta)\) are again given respectively by (3.39) and (3.47) except that this time \((\alpha, \beta)\) denote the eigenvalue bounds for the eigenvalues of the positive definite matrix \( B_{\omega} \). The iterations proceed in the \( y \)-variable until a specified degree of accuracy is achieved. Evans (1967) gives numerical results for the model problem of section 3.2 (Laplace's equation), illustrating that although the application of pre-conditioning involves approximately a two-fold increase in work over the original two iterative methods this is more than outweighed by the greatly increased convergence rates. We will refer to the above preconditioning technique as "Triangular Preconditioning" or "T.P." for short.

In the remainder of this chapter, we consider two new forms of preconditioning, closely related to each other. The first, discussed in section A, is in connection with the iterative solution of linear systems of equations arising from the discretisation of self-adjoint second order elliptic partial differential equations. The second, considered in section B, is applied to the iterative solution of the linear system obtained from discretising the Biharmonic equation.

In both section A and section B, the partial differential equations
will be in two space dimensions.

Section A Self-adjoint second order elliptic partial differential equations

7A.1 Formulation of the problem

We consider the solution of

\begin{equation}
\frac{3}{3x} (P(x,y) \frac{\partial u}{\partial x}) + \frac{3}{3y} (Q(x,y) \frac{\partial u}{\partial y}) + \sigma(x,y)U(x,y) + f(x,y) = 0, \quad (x,y) \in \Omega,
\end{equation}

where \( \Omega \) is the rectangular region

\[ \Omega = [x_1, x_2] \times [y_1, y_2]. \]

For boundary conditions, we assume that

\[ U(x,y) = g(x,y), \quad (x,y) \in \partial \Omega, \]

where \( g(x,y) \) is a prescribed function on the boundary \( \partial \Omega \) of \( \Omega \). We also have the conditions that \( P, Q, \sigma, f \) are continuous in \( \bar{\Omega} \) and satisfy

\[ P(x,y) > 0; \quad \sigma(x,y) > 0, \quad \forall (x,y) \in \Omega. \]

7A.2 The difference equations

We superimpose a rectangular grid (of length \( Ax \) in the \( x \)-direction and \( Ay \) in the \( y \)-direction) on \( \Omega \), such that

\begin{equation}
(n+1)Ax = l_1, \quad (m+1)Ay = l_2.
\end{equation}

We consider the numerical solution of the problem defined in section 7A.1 associated with the set of mesh points on the rectangular grid.

Defining the quantities \( H_u(x,y), V_u(x,y) \) by (3.61), (3.62) respectively where this time

\begin{align*}
a &= (\Delta y/\Delta x)P(x+h/2,y), \quad b = (\Delta y/\Delta x)P(x-h/2,y), \quad 2b = a + c \\
\gamma &= (\Delta x/\Delta y)Q(x,y+h/2), \quad \beta = (\Delta x/\Delta y)Q(x,y-h/2), \quad 2\beta = \alpha + \delta
\end{align*}

we see that they are respectively central difference approximations to \( \{ -\Delta x \Delta y \frac{3}{3x} (P(x,y) \frac{\partial u}{\partial x}) \} \) and \( \{ -\Delta x \Delta y \frac{3}{3y} (Q(x,y) \frac{\partial u}{\partial y}) \} \). Ordering the unknown mesh points along successive rows (\( x \)-lines) of the grid, the finite difference equations approximating (7.6) at each of these mesh points may then be written as

\[ A u = (H + V) u = (H_1 + V_1) u = b, \]

where

\[ A = \begin{bmatrix}
H & V \\
V & H
\end{bmatrix}, \quad \begin{bmatrix}
H_1 & V_2 \\
V_2 & H_1
\end{bmatrix} = \begin{bmatrix}
H + V \\
V + H
\end{bmatrix}, \quad \begin{bmatrix}
H_2 & V_1 \\
V_1 & H_2
\end{bmatrix} = \begin{bmatrix}
H + V \\
V + H
\end{bmatrix} + \begin{bmatrix}
H_1 & V_1 \\
V_1 & H_1
\end{bmatrix}, \quad \begin{bmatrix}
H_2 & V_2 \\
V_2 & H_2
\end{bmatrix} = \begin{bmatrix}
H + V \\
V + H
\end{bmatrix} + \begin{bmatrix}
H_1 & V_1 \\
V_1 & H_1
\end{bmatrix} + \begin{bmatrix}
H_2 & V_2 \\
V_2 & H_2
\end{bmatrix}. \]
where $H_1$ and $V_1$ are defined by (3.67). The matrices $A, H_1$ and $V_1$ have the properties given immediately after equation (3.64).

In order that $H_1$ and $V_1$ commute the differential equation (7.6) must be of the form (3.96). Hence, as this property is essential to our theoretical analysis to follow, we assume from now on that the differential equation is of the latter form.

7A.3 The Alternating Direction Preconditioning technique

Let us firstly rewrite the coefficient matrix $A$ as

$$A = I + (H_1 - I/2) + (V_1 - I/2)$$

$$= I + \hat{H} + \hat{V}, \quad (7.12)$$

where

$$\hat{H} = H_1 - I/2, \quad \hat{V} = V_1 - I/2. \quad (7.13)$$

If (7.11) is premultiplied by $(I + \hat{H})^{-1}$ then it is equivalent to

$$(I + \hat{H})^{-1}A(I + \hat{V})^{-1}(I + \hat{V})u = (I + \hat{H})^{-1}b \quad (7.14a)$$

or

$$B_\omega \cdot \gamma = d \quad (7.14b)$$

where

$$B_\omega = (I + \hat{H})^{-1}A(I + \hat{V})^{-1}; \quad \gamma = (I + \hat{V})u; \quad d = (I + \hat{H})^{-1}b. \quad (7.15)$$

The object now is identical to that described for Triangular Preconditioning after equation (7.3). It is obvious that because $H_1, V_1$ are symmetric and commute, then the same applies to $\hat{H}, \hat{V}$. Hence, $B_\omega$ is symmetric also.

Because $A, \hat{H}$ and $\hat{V}$ are pairwise commutative, and all symmetric, they have a common set of orthonormal eigenvectors, by Theorem 3.4. Thus, from (7.15) the eigenvalues $\lambda$, of $B_\omega$ can be expressed as

$$\lambda = \frac{\mu + \nu}{(1 - \omega/2 + \omega)(1 - \omega/2 + \omega)} \quad (7.16)$$

where $\mu, \nu$ are the real positive eigenvalues of $H_1, V_1$ respectively. We see from (7.16) that if $0 < \omega < 2$, then $\lambda > 0$ for all $\lambda$. Hence with $\omega$ in this range we can immediately state that the matrix $B_\omega$ is positive definite (from Theorem 1.2).
7A.4 *Minimisation of \( P \)-condition number of coefficient matrix*

As mentioned in section 7.1, the object of preconditioning the original system is to obtain a new system, viz. (7.14), whose coefficient matrix varies with a parameter \( \omega \). We then minimise the \( P \)-condition number of \( B_\omega \), denoted by \( P(B_\omega) \), with respect to \( \omega \), for some "optimum" value \( \omega = \omega_2 \).

In order to investigate the \( P \)-condition number of \( B_\omega \), we must first examine its smallest eigenvalue given by \( \min \lambda(\mu, v, \omega) = \lambda_{\min}(\mu, v, \omega) \) and then its largest eigenvalue given by \( \max \lambda(\mu, v, \omega) = \lambda_{\max}(\mu, v, \omega) \) from (7.16).

We have assumed here that the eigenvalues \( \mu \) of \( B \) and \( v \) of \( V \) lie in the ranges

\[
0 < \mu < \beta, \quad 0 < \omega < 2 . \tag{7.17}
\]

To achieve these ends, let us consider the following continuous function

\[
f(x, y, \omega) = \frac{x + y}{(1 - \omega/2 + \omega x)(1 - \omega/2 + \omega y)}, \quad 0 < \omega < 2 \quad 0 < x < 2 . \tag{7.18}
\]

Let us first consider the case where \( \mu \) and \( v \) (and hence \( x \) and \( y \)) lie in the same range, so that

\[
a = a, \quad b = b . \tag{7.19}
\]

(i) *Minimum eigenvalue of \( B_\omega \) for \( 0 < \omega < 2 \).*

Now \( f(x, y, \omega) \) from (7.18) can be re-written as

\[
f(x, y, \omega) = \frac{1}{u(2-u)} \left[ 1 - \frac{(1-u/2+\omega x)(1-u/2+\omega y)}{(1-u/2+\omega x)(1-u/2+\omega y)} \right], \quad 0 < \omega < 2 . \tag{7.20}
\]

Letting \( r = \frac{(1-u/2)}{\omega} \) and \( g(x; r) = \frac{x-r}{r+x} \),

from (7.20) we have

\[
f(x, y, \omega) = \frac{1}{u(2-u)} \left[ 1 - g(x; r) \cdot g(y; r) \right], \quad 0 < \omega < 2 . \tag{7.22}
\]

Hence,

\[
\min_{x \in \mathbb{R}, y \in \mathbb{R}} f(x, y, \omega) = \frac{1}{u(2-u)} \left[ 1 - \max_{x \in \mathbb{R}, y \in \mathbb{R}} \left| g(x; r) \cdot g(y; r) \right| \right], \quad 0 < \omega < 2 . \tag{7.23}
\]
Since \( x \) and \( y \) lie within the same range,

\[
\max_{\alpha \beta} \left| g(x \mid r), g(y \mid r) \right| = \left( \max_{\alpha \beta} \left| g(x \mid r) \right| \right)^2 , \quad 0 < \omega < 2.
\]

(7.24)

From equations (3.76)-(3.78) we have

\[
\max_{\alpha \beta} \left| g(x \mid r) \right| = \begin{cases} \frac{\theta - r}{\beta + r} , & 0 < r < \sqrt{\alpha \beta} \\ \frac{r - a}{r + a} , & r \geq \sqrt{\alpha \beta} \end{cases}
\]

(7.25)

From equations (7.23)-(7.25) we obtain

\[
\min_{\alpha \beta} f(x, y, w) = \begin{cases} \frac{1}{\omega(2 - \omega)} \left[ 1 - \left( \frac{\theta - r}{\beta + r} \right)^2 \right] , & 0 < r < \sqrt{\alpha \beta} \\ \frac{1}{\omega(2 - \omega)} \left[ 1 - \left( \frac{r - a}{r + a} \right)^2 \right] , & r \geq \sqrt{\alpha \beta} \end{cases}
\]

(7.26)

Substituting (7.21) in (7.26) yields

\[
\min_{\alpha \beta} f(x, y, w) = \begin{cases} \frac{1}{\omega(2 - \omega)} \left[ 1 - \frac{(1 - \omega/2 - \omega)^2}{(1 - \omega/2 + \omega)^2} \right] , & 0 < \omega < \frac{2}{1 + 2/\alpha \beta} \\ \frac{1}{\omega(2 - \omega)} \left[ 1 - \frac{(1 - \omega/2 - \omega)^2}{(1 - \omega/2 + \omega)^2} \right] , & \frac{2}{1 + 2/\alpha \beta} \leq \omega < 2. \end{cases}
\]

(7.27)

\[
\frac{2 \alpha}{(1 - \omega/2 + \omega)^2} , \quad 0 < \omega \leq \omega_a
\]

(7.28)

\[
\frac{2 \beta}{(1 - \omega/2 + \omega)^2} , \quad \omega > \omega_a
\]

(7.29)

where,

\[
\omega_a = \frac{2}{1 + 2/\alpha \beta}.
\]

Thus, returning to equation (7.16), we have that

\[
\min_{\alpha \beta} \lambda(\mu, v, w)
\]

is given by (7.28).

\[\text{(ii) Maximum eigenvalue of } B_{w} \text{ for } 0 < \omega < 2.\]

From (7.22) we have

\[
\max_{\alpha \beta} f(x, y, w) = \frac{1}{\omega(2 - \omega)} \left[ 1 + \omega \right].
\]

(7.30)
where

\[ M = \max_{x,y} \{-g(x,r), g(y,r)\} \]  
(7.31)

(a) For \( r > 0 \), \( g(x,r), g(y,r) > 0 \), for all \( a \leq x, y \leq b \). Hence for \( r \) in this range

\[ M = - \{ \min_{a \leq x \leq b} g(x,r) \} \{ \min_{a \leq y \leq b} g(y,r) \} \]  
(7.32)

Clearly, for \( r > 0 \),

\[ \min_{a \leq x \leq b} g(x,r) = g(b,r) \]  
(7.33)

since the derivative of \( g(x,r) \) with respect to \( x \) is negative for all \( x > 0 \).

Hence for \( r > 0 \) or equivalently \( 0 < a \leq \frac{2}{1+2r} \) [from (7.21)],

\[ \max_{a \leq x, y \leq b} \lambda(a, \beta, \gamma) = \max_{a \leq x, y \leq b} f(x, y, \omega) = f(\beta, \alpha, \omega) \]  
(7.34)

(b) For \( a \leq x, y \leq b \), \( g(x,r) \) and \( g(y,r) \) will take on positive and negative values for \( 0 < a \leq y \leq b \). Hence

\[ M = \{-\min_{a \leq x \leq b} g(x,r)\} \{ \max_{a \leq y \leq b} g(y,r)\} = \{ \max_{a \leq x \leq b} g(x,r)\} \{-\min_{a \leq y \leq b} g(y,r)\} \]  
(7.35)

Clearly for \( a \leq x, y \leq b \),

\[ \min_{a \leq x \leq b} g(x,r) = g(b,r) ; \quad \max_{a \leq y \leq b} g(y,r) = g(a,r) \]  
(7.36)

Since \( x \) and \( y \) are in the same range, their roles can be reversed in (7.36).

Thus for \( a \leq x, y \leq b \), or equivalently \( \frac{2}{1+2r} \), \( 0 < a \leq \frac{2}{1+2a} \) [from (7.21)],

\[ \max_{a \leq x, y \leq b} \lambda(a, \beta, \gamma) = \max_{a \leq x, y \leq b} f(x, y, \omega) = f(\beta, \alpha, \omega) = f(\alpha, \beta, \omega) \]  
(7.37)

(c) Lastly, for \( 0 < a \leq \alpha, \beta \) both \( g(x,r) \) and \( g(y,r) \) will be non-positive for all \( a \leq x, y \leq b \). Hence

\[ M = \{-\max_{a \leq x \leq b} g(x,r)\} \{ \max_{a \leq y \leq b} g(y,r)\} \]
Now for $0 < \varrho \varphi$, 
\[
\max_{\varphi \varphi} g(\chi; \tau) = g(\alpha; \tau). 
\] (7.39)

Therefore, for $0 < \varrho \varphi$ or equivalently $\frac{2}{1+2\varphi} \varphi \varphi < 2$,
\[
\max_{\alpha \varphi \varphi} \lambda(\mu; \nu; \omega) = \max_{\varphi \varphi} f(x; y; \omega) = f(\alpha; \alpha; \omega). 
\] (7.40)

Collecting together equations (7.34), (7.37) and (7.40) and using (7.16), yields
\[
\max_{\alpha \varphi \varphi} \lambda(\mu; \nu; \omega) = \lambda(\alpha; \alpha; \omega) = \lambda(\beta; \beta; \omega), \quad 0 < \omega < \frac{2}{1+2\varphi}. 
\] (7.41)

(iii) The $P$-condition number of $B_\omega$ for $0 < \omega < 2$.

It now follows from equations (7.28) and (7.41) that we must examine $P(B_\omega)$ in four successive intervals namely, $0 < \omega < 2/(1+2\varphi), \quad 2/(1+2\varphi) < \omega < 2/(1+2\sqrt{\varphi}), \quad 2/(1+2\sqrt{\varphi}) < \omega < 2/(1+2\alpha)$ and $2/(1+2\alpha) < \omega < 2$. (N.B. $0 < \omega < \sqrt{\alpha} \beta$ since $\alpha < \beta$.)

Since $B_\omega$ ($0 < \omega < 2$) is positive definite, then $P(B_\omega) > 0$ for $0 < \omega < 2$.

(a) When $0 < \omega < 2/(1+2\varphi)$, we have from (7.28) and (7.41)
\[
P(B_\omega) = \frac{\beta(1-\omega/2+\omega)^2}{\alpha(1-\omega/2+\omega)^2}. 
\] (7.42)

Therefore, in this range,
\[
\frac{dP}{d\omega} = \frac{2\beta}{\alpha} \frac{(\alpha-\beta)(1-\omega/2+\omega)^2}{(1-\omega/2+\omega)^3} < 0, \quad \text{since } \beta > \alpha. 
\] (7.43)

Hence, for $0 < \omega < 2/(1+2\varphi)$,
\[
\min_{\omega} P(B_\omega) = P(B_{\omega=2/(1+2\varphi)}) . 
\] (7.44)
(b) When $2/(1+2\beta) \leq 2/(1+2\alpha)$ equations (7.28) and (7.41) yield

$$P(B_\omega) = \frac{(\alpha\beta)(1-\omega/2+\omega\beta)}{2\alpha(1-\omega/2+\omega\beta)}$$  \hspace{1cm} (7.45)

Thus, in this range

$$\frac{dP}{d\omega} = \frac{-(\alpha\beta)}{2\alpha(1-\omega/2+\omega\beta)} < 0, \text{ since } \alpha < \beta$$  \hspace{1cm} (7.46)

Hence, for $2/(1+2\beta) \leq 2/(1+2\alpha)$,

$$\min_{\omega} P(B_\omega) = P(B_{\omega=2/(1+2\sqrt{\alpha\beta})})$$  \hspace{1cm} (7.47)

(c) When $2/(1+2\sqrt{\alpha\beta}) \leq 2/(1+2\alpha)$ we obtain from (7.28) and (7.41),

$$P(B_\omega) = \frac{(\alpha\beta)(1-\omega/2+\omega\beta)}{2\beta(1-\omega/2+\omega\alpha)}$$  \hspace{1cm} (7.48)

So in this interval

$$\frac{dP}{d\omega} = \frac{2\alpha(\beta-\omega)(1-\omega/2+\omega\beta)}{2\beta(1-\omega/2+\omega\alpha)^3} > 0, \text{ since } \beta > \alpha$$  \hspace{1cm} (7.49)

Therefore, when $2/(1+2\sqrt{\alpha\beta}) \leq 2/(1+2\alpha)$,

$$\min_{\omega} P(B_\omega) = P(B_{\omega=2/(1+2\sqrt{\alpha\beta})})$$  \hspace{1cm} (7.50)

(d) Lastly, when $2/(1+2\alpha) \leq 2$, we have from (7.28) and (7.41)

$$P(B_\omega) = \frac{\alpha(1-\omega/2+\omega\beta)^2}{\beta(1-\omega/2+\omega\alpha)^2}$$

In this interval

$$\frac{dP}{d\omega} = \frac{2\alpha(\beta-\omega)(1-\omega/2+\omega\beta)}{2\beta(1-\omega/2+\omega\alpha)^3} > 0, \text{ since } \beta > \alpha$$  \hspace{1cm} (7.51)

Hence, for $2/(1+2\alpha) \leq 2$,

$$\min_{\omega} P(B_\omega) = P(B_{\omega=2/(1+2\alpha)})$$  \hspace{1cm} (7.52)

It is now apparent from equations (7.44), (7.47), (7.50) and (7.52) that

$$\min_{0<\omega<2} P(B_\omega) = P(B_{\bar{\omega}}) = \frac{\alpha\beta}{2\sqrt{\alpha\beta}}$$  \hspace{1cm} (7.53)

where $\bar{\omega}=2/(1+2\sqrt{\alpha\beta})$, and $P(B_{\bar{\omega}})$ is given by (7.45) or (7.48).

We shall refer to $\bar{\omega}$ as the optimum preconditioning parameter.

(iv) Case when eigenvalue ranges of $H_1$ and $V_1$ are different

Up to this point, we have assumed that $0<\alpha\mu<\beta$ and $0<\alpha\nu<\beta$. 
are such that $a=\alpha$ and $b=\beta$. This will not always occur in practice, e.g. when solving Laplace's equation on a rectangle with adjacent sides of different length, using a square mesh.

As far as the minimum eigenvalue of $B_\omega$ is concerned an approximate analysis can be achieved by noticing that $0<\alpha^t\xi_\mu,v_\eta\beta^t$, where $\alpha^t=\min(a,\alpha)$ and $\beta^t=\max(b,\beta)$. Then $\min_{a\xi_\mu,v_\eta\beta^t} \lambda(\mu,v,\omega)\geq \min_{a\xi_\mu,v_\eta\beta^t} \lambda(\mu,v,\omega)$, and the value $\min_{a\xi_\mu,v_\eta\beta^t} \lambda(\mu,v,\omega)$ of the function on the right can be determined from subsection 7A.4(i) for $0<\omega<2$.

Next, we have

$$\max_{a\xi_\mu,v_\eta\beta^t} \lambda(\mu,v,\omega) \leq \max \lambda(\mu,v,\omega) ,$$

and the function on the right for $0<\omega<2$ can be evaluated from subsection 7A.4(ii). Hence from subsections 7A.4(i),(ii) we obtain

$$\min_{0<\omega<2} P(B_\omega) \leq \min_{0<\omega<2} \left\{ \max_{a\xi_\mu,v_\eta\beta^t} \lambda(\mu,v,\omega) \right\} = P(B_{\omega^t}) ,$$

where $\omega^t=2/(1+2\sqrt{\alpha^t\beta^t})$.

7A.5 Application of the Pre-conditioning technique to the basic iterative methods

We will now apply the results of the previous section to modify the two original basic iterative methods given by equations (3.15) and (3.46), i.e. the Simultaneous Displacements and Richardson's second degree methods, respectively. It follows immediately that if the techniques used in Section 7A.4 are shown to apply to the above two methods, then an improvement in convergence rate must follow.

We now proceed from equation (7.14), and develop similar iterative processes to the above two methods by working throughout in the transformed variable $y$. From equation (7.14) then, equations describing iterative processes similar to (3.15) and (3.46) are

$$\gamma^{(k+1)} = \gamma^{(k)} + \alpha \left( \frac{\gamma^{(k)}}{d-B^{(k)}} \right)$$

(7.54)
and
\[ \chi^{(k+1)} = \chi^{(k)} + \alpha (d - B \chi^{(k)}) + \beta (\chi^{(k)} - \chi^{(k-1)}), \]  
(7.55)
respectively. The iterations proceed in the \( y \) variable until a specified degree of accuracy is achieved. The final solution is then obtained by one application of the formula,
\[ u = (I + \omega \chi)^{-1} \chi. \]  
(7.56)
The optimum value of \( \alpha = \overline{\alpha} \) for method (7.54) is given from (3.39) by
\[ \overline{\alpha} = \frac{2}{\alpha + b} \]  
(7.57)
where \( 0 < \alpha \leq \lambda_1(\omega) \leq b \), and the \( \lambda_1(\omega) \) are the positive real eigenvalues of \( B_\omega \). With this choice of \( \alpha \), the spectral radius of the iteration matrix, 
\[ (I - \overline{\alpha} B_\omega) \]  
is
\[ \max_{i} \left| 1 - \overline{\alpha} \lambda_i(\omega) \right| \leq \frac{\overline{b} - \alpha}{\overline{b} + a} = \frac{P(B_\omega) - 1}{P(B_\omega) + 1}. \]  
(7.58)
The values \( \overline{\alpha} = \min_{\alpha \in \mu, \nu \in \beta} \lambda(\mu, \nu, \omega) \) and \( \overline{b} = \max_{\alpha \in \mu, \nu \in \beta} \lambda(\mu, \nu, \omega) \) are given by equations (7.28) and (7.37) respectively. From (7.58), the rate of convergence is
\[ R(I - \overline{\alpha} B_\omega) = -\log \frac{P(B_\omega) - 1}{P(B_\omega) + 1}. \]  
(7.59)
Similarly, for equation (7.55) (preconditioned Richardson's second degree method) we have from (3.47),
\[ \overline{\alpha} \left[ \frac{2}{\sqrt{a + \beta}} \right]^2 \text{ and } \overline{\beta} = \left[ \frac{\sqrt{a} - \sqrt{b}}{\sqrt{a + \beta}} \right]^2. \]  
(7.60)
and the spectral radius of the iteration matrix from (3.48) is
\[ \sqrt{\overline{\beta}}, \]  
(7.61)
so that the rate of convergence satisfies
\[ -\log(\sqrt{\overline{\beta}}) = -\log \frac{\sqrt{P(B_\omega) - 1}}{\sqrt{P(B_\omega) + 1}}. \]  
(7.62)
There are other methods to which we could apply this new preconditioning technique such as the Chebyshev acceleration of (7.55), which with \( \overline{\alpha}_k, \overline{\beta}_k \) replacing \( \overline{\alpha}, \overline{\beta} \) respectively, is defined by that equation. This is a
non-stationary method, since the acceleration parameters \( \overline{a}_k, \overline{a}_k \) vary at each iteration. Chebyshev polynomials are used to optimise \( \overline{a}_k, \overline{a}_k \) with the result that we again have an asymptotic rate of convergence equal to \( 2/\sqrt{F(\omega)} \). However, in what is to follow, we restrict our attention to methods (7.54) and (7.55).

**The computational procedure**

We will consider the computational procedure for method (7.54) only, since it is analogous for method (7.55). Let us assume for problem 7A.1, that we have ordered our grid points horizontally along successive x-lines of the grid. We also assume that we have first obtained our optimum parameters \( \hat{a}, \hat{a} \) from (7.29), (7.37) respectively. Initially, from (7.15) we have

\[
(I + \omega \hat{H}) \hat{d} = \hat{b} \quad (\text{b known})
\]  

(7.63)

Since we have ordered the mesh points horizontally, we see from (3.61) that \( \hat{H} = \hat{H}_1 - I/2 = (\hat{H} + \hat{J}/2) - I/2 \) is tri-diagonal. Thus since \( I + \omega \hat{H} \) is positive definite \( (0 < \omega < 2) \) we can solve (7.63) directly for \( \hat{d} \) using Algorithm 4.2. Next, from (7.54) and (7.15) we see that at each iteration we must evaluate

\[
(I + \omega \hat{H})^{-1} A(I + \omega \hat{V})^{-1} \Gamma^{(k)}, \quad k > 0
\]  

(7.64)

Let \( \chi^{(k)} = (I + \omega \hat{V})^{-1} \Gamma^{(k)}, \) i.e. \( (I + \omega \hat{V})\chi^{(k)} = \Gamma^{(k)} \). Because of the present ordering of the unknown mesh points, we see from (3.62) that \( \hat{V} = \hat{V}_1 - I/2 = (\hat{V} + \hat{J}/2) - I/2 \) is a block tri-diagonal matrix, where each block element is a diagonal matrix. By re-ordering the elements of the vector \( \Gamma^{(k)} \) and thus \( \chi^{(k)} \) also, to correspond with a vertical ordering of the grid points along successive y-lines, a new system is derived namely,

\[
(I + \omega \hat{V}_T) \chi_T^{(k)} = \chi_T^{(k)}
\]  

(7.65)

From (3.62) we see that \( \hat{V}_T \) is a block diagonal matrix, each block being a tri-diagonal matrix, whose size corresponds to the number of unknown mesh points on each y-line. Since for \( 0 < \omega < 2 \), \( I + \omega \hat{V}_T \)
is positive definite (and diagonally dominant), we can solve (7.65) directly for $\mathbf{x}_k^{(k)}$ using Algorithm 4.2. Next we must solve

$$\mathbf{x}_k^{(k)} = A_k^{(k)} \mathbf{x}_k^{(k)}, \quad k > 0$$

(7.66)

The vector $\mathbf{x}_k^{(k)}$ can be generated, by remembering that the matrix $A$ of (7.11) was formed from applying the five point difference approximation of (7.6) at each unknown grid point with a horizontal ordering. Thus (7.66) can be evaluated by applying the same difference analogue at each mesh point in the same order with $\mathbf{r}_k^{(k)}$ (not $\mathbf{x}_k^{(k)}$) replacing $\mathbf{u}$ (see (7.11)). Finally, we must solve

$$\mathbf{q}_k^{(k)} = (I + \omega h)^{-1} \mathbf{x}_k^{(k)}, \quad \text{i.e.} \quad (I + \omega h) \mathbf{q}_k^{(k)} = \mathbf{x}_k^{(k)}, \quad k > 0.$$  

(7.67)

System (7.67) is solved in exactly the same way as system (7.63), to obtain the vector $\mathbf{q}_k^{(k)}$ which is the value(s) of the expression in (7.64).

Method (7.54) can now be written as

$$\mathbf{y}_k^{(k+1)} = \mathbf{y}_k^{(k)} + \mathbf{u}_k^{(k)} (\mathbf{d} + \mathbf{q}_k^{(k)}), \quad k > 0,$$

(7.68)

which is easily solved for $\mathbf{y}_k^{(k+1)}$, $k > 0$ since each vector on the right hand side is known. This process continues until convergence (to the desired accuracy) is achieved. The original solution vector $\mathbf{u}$ is then found from (7.56) in the same way as system (7.65) was derived and then solved for. We can now see the reason for calling this an "Alternating Direction Preconditioning (A.D.P.) technique".

7A.6 The model problem: comparative rates of convergence

The model problem we are considering has been defined in section 3.2. For this problem we have $\Sigma = 0$ in (7.11). Also, dividing each difference equation by 4, and denoting the new coefficient matrix again by $A$, we have

$$A = \mathbb{H} + \mathbb{V} (\mathbb{H} + \mathbb{V}), \quad \text{or} \quad A = \mathbb{H} + \mathbb{V} (\mathbb{H} + \mathbb{V}),$$

(7.69)
where

$$(\mathcal{H}_1 \mathbf{H}) = \begin{bmatrix} 
B & B \\
B & (n^2 \times n^2)
\end{bmatrix}$$

and

$$(\mathcal{V}_1 \mathbf{V}) = \begin{bmatrix} 
1/2 & -1/4 \\
-1/4 & 1/2 \\
-1/4 & 1/2 \\
-1/4 & 1/2 \end{bmatrix}$$

where \( I \) is the unit matrix of order \( n \)

$$(7.70)$$

Here, we have assumed a square mesh size of \( h=1/(n+1) \) so that there are \( n^2 \) internal mesh points. If again \( \lambda, \mu \) denote the eigenvalues of \( H(=\mathcal{H}_1), V(=\mathcal{V}_1) \) respectively, then

$$
u_1 = \sin^2 \left( \frac{iw}{2(n+1)} \right); \quad \nu_j = \sin^2 \left( \frac{jw}{2(n+1)} \right), \quad 1 \leq i, j \leq n. \tag{7.72}$$

Hence, if \( a = \nu_1 \), \( b = \nu_j \), \( 1 \leq i, j \leq n \), then

$$
a = a = \sin^2 \left( \frac{w}{2(n+1)} \right); \quad b = \beta = \cos^2 \left( \frac{w}{2(n+1)} \right). \tag{7.73}$$

Thus for the model problem, equation (7.29) yields

$$
\frac{2}{1+\sin(w/(n+1))} = \frac{2}{1+\sin(wh)} \tag{7.74}
$$

Substituting the values of \( a, \beta \) from (7.73) in equation (7.53) yields

$$
\min_{0<\omega<2} P(B_\omega) = P(B_{\bar{\omega}}) = \frac{1}{\sin(w/(n+1))} = \frac{1}{\sin(wh)} \tag{7.75}
$$

Using the optimum acceleration parameter \( \bar{\omega} \) defined by (7.57), \( \bar{a} \) and \( \bar{b} \) defined by (7.28) and (7.37) for \( \omega = \bar{\omega} \) from (7.58) the

Alternating Direction Preconditioned (A.D.P.) Simultaneous Displacements method has a spectral radius given by
\[
\max_{1 \leq i,j \leq n} \left| 1 - \tilde{\omega} \lambda_{i,j}(\omega) \right| = \frac{1 - \sin(\pi h)}{1 + \sin(\pi h)} \quad (7.76)
\]

But this expression is actually quite familiar to us. From equation (3.81) we see that

\[
\min_{\omega} \rho(\text{A.D.P. Simultaneous Displacements}) = \\
\omega, \alpha
\]

\[
= \min_{\omega} \rho(\text{point S.O.R.)}) = \min_0(\text{one parameter P.R. (A.D.I.)}) (7.77)
\]

Hence, as optimised one parameter iterative methods, the Peaceman-Rachford method, the point S.O.R. process and the A.D.P. Simultaneous Displacements method have \textit{identical} asymptotic rates of convergence for all \(h>0\) for the \textit{model problem}. In fact, in each case from (7.76) and Definition 3.4,

\[
R_\omega \approx \frac{2\pi}{n+1} = 2\pi h, \text{ as } h \to 0. \quad (7.78)
\]

We also notice in this context, that the optimum preconditioning parameter \(\bar{\omega}\) given by (7.74) is identical to the optimum over-relaxation parameter defined by (3.52) for the point S.O.R. method.

In view of equation (7.62), improved rates of convergence can be obtained by using the Alternating Direction Preconditioned second degree Richardson method defined by (7.55). Thus using the optimum preconditioning factor \(\bar{\omega}\) from (7.74) and the optimum parameters \(\bar{\alpha}\) and \(\bar{\beta}\) in (7.60), we have from (7.62)

\[
R_\omega (\text{A.D.P. second degree Richardson method}) = 2\sqrt{\sin(\pi h)} \approx 2\sqrt{\pi h} = 2\frac{\sqrt{\pi h}}{\sqrt{n+1}} \text{ as } n \to \infty. \quad (7.79)
\]

We see from (3.94) that this quantity is identical to the asymptotic average rate of convergence of the 2-parameter P.R. (A.D.I.) iterative method (where we are using the \(M=2^k(\geq 0)\) type parameters derived by Wachpress (1962)). For large \(n\), from (3.95), (3.88) and (3.89) increased asymptotic average rates of convergence can be obtained by using \(M>>2\) parameters.
Hence for the *model problem* our two new preconditioned methods (7.54) and (7.55) would appear to offer no gains in rates of convergence over existing methods such as the $M=2^k$-parameter P.R. (A.D.I.) iterative procedure.

7A.7 *Numerical results*

The two problems we solved were taken from Young and Ehrlich (1960). Both involved the approximate solution of Laplace's equation. The A.D.P. Simultaneous Displacements and A.D.P. second degree Richardson methods given by (7.54) and (7.55) were compared with the point S.O.R. process and the Peaceman Rachford procedure using a fixed number $M=1,2$ and 3) of parameters (the results for the last two methods being taken from the previously mentioned reference). By "fixed" number of parameters, we mean that $M$ is independent of $h$, while the values of the parameters themselves vary with $h$. Varga (1962), p.229 shows that with $M>1$, this technique converges faster than the point S.O.R. method with optimum $\omega$, when $h^{-1}$ is large (i.e. for the model problem). It should be noted that these type of parameters will tend to give slower convergence than the Wachpress ($M=2^k$) parameters, since the former were obtained using a much cruder analysis than the latter (e.g. see Varga (1962) pp.226-229).

The first region chosen was the **unit square**. The boundary conditions were

$$U(x,1)=1, \quad 0 \leq x \leq 1; \quad U(0,y)=U(1,y)=U(x,0)=0, \quad 0 \leq x, y \leq 1. \quad (7.80)$$

For the former two methods mentioned above $h^{-1}$ took the values 5(5)30, whereas the latter two had $h^{-1}=5,10,20$ and 40. The convergence criteria were

$$\max_{1 \leq i,j \leq n} |y_{i,j}^{(k+1)} - y_{i,j}^{(k)}| < 10^{-6}, \quad (7.81)$$
Figure 7.1
for the former two methods, and

$$\max_{|i,j| \leq n} |u_{i,j}^{k+1} - u_{i,j}^{(k)}| < 10^{-6}, \quad k=0, \ldots, (7.81b)$$

for the latter two. The respective iterative processes started with

$$y_{i,j}(0) = 0$$

for all $1 \leq i,j \leq n$. The results are displayed in Figure 7.1, the logarithms of $N$, the number of iterations, being plotted against $\log(h^{-1})$. They bear out the theoretical comparison made between the rates of convergence of the optimised one parameter A.D.P. Simultaneous Displacements, point S.O.R. and P.R. (A.D.I.) methods, i.e. they are the same. The figure also shows the expected increase in rate of convergence over the previous three just mentioned, of the A.D.P. second degree Richardson method.

The second region chosen was the unit square with a $0.4 \times 0.4$ square removed from the centre. For the first two methods mentioned at the beginning of the section, $h^{-1} = 10, 20, 30, 40$ and for the second pair, $h^{-1} = 10, 20, 40$. The boundary conditions were

$$U(x,0) = U(1,y) = U(x,1) = U(0,y) = 0, \quad 0 \leq x, y \leq 1; (7.82)$$

$$U(x,0.3) = U(0.7,y) = U(x,0.7) = U(0.3,y) = 1, 0.3 \leq x, y \leq 0.7$$

The results, using the same axes as before, are given in Figure 7.2. For this problem, the matrices $H$ and $V$ and therefore $\hat{H}$ and $\hat{V}$ do not commute, hence the coefficient matrix $B_{n}$ of the pre-conditioned system (7.14) is non-symmetric, and thus could possess complex eigenvalues. A theoretical proof of convergence and theoretical treatment of optimum parameters and rates of convergence has not yet been devised for this "non-commutative case". In this respect, we have inherent difficulties analogous to those of the "non-commutative case" of the P.R. (A.D.I.) iterative method (see sections 3.5 and 6.7).

The above problem is intended to serve as an example of the workability of our pre-conditioned iterative methods in practice, when "model problem" conditions do not prevail. Emulating Young and Ehrlich
(1960), it was decided in all methods, to use the respective "optimum" parameters obtained for the first problem i.e. on the unit square. Whilst, from Figure 7.2, the rates of convergence are seen to be worse for the P.R. (A.D.I.) method (M=1,2,3) as compared to those given for the unit square (see Figure 7.1), the two pre-conditioned methods give almost identical results on the two regions in question.

Lastly, we compared the P-condition numbers of the A.D.P. and T.P. techniques for the model problem of section 3.2. The latter procedure has been outlined in section 7.1. The results are tabulated in Table 7.1, below, the T.P. results being taken from Evans (1967).

Table 7.1 (P-condition numbers)

<table>
<thead>
<tr>
<th>h^-1</th>
<th>Triangular Preconditioning</th>
<th>Alternating Direction Preconditioning</th>
</tr>
</thead>
<tbody>
<tr>
<td> </td>
<td> </td>
<td> </td>
</tr>
<tr>
<td>5</td>
<td>1.3</td>
<td>1.655</td>
</tr>
<tr>
<td>10</td>
<td>1.6</td>
<td>2.816</td>
</tr>
<tr>
<td>20</td>
<td>1.75</td>
<td>5.274</td>
</tr>
</tbody>
</table>

We see that for h^-1=5,10,20, Triangular Preconditioning (T.P.) has smaller P-condition numbers than A.D.P. Thus in view of equations (7.59) and (7.62) we can expect faster rates of convergence from the former technique applied to the two iterative methods discussed in section 7A.5.

7A.8 Conclusions

For the model problem, we have shown that the A.D.P. Simultaneous Displacements method has identical rates of convergence with the optimised point S.O.R. and one parameter P.R. (A.D.I.) iterative procedures. Likewise, the A.D.P. second order Richardson process was shown to have the same rate of convergence (for the model problem) as the 2 parameter P.R. (A.D.I.) using M=2^k type parameters (i.e.,k=1).
Since for the model problem, the amount of work per iteration for the last two methods can be shown from sections 3.5 and 7A.5 to be about the same, the equality in rates of convergence of the respective methods above, must very nearly apply to total computational effort as well. Hence methods (7.54) and (7.55) provide no improvement over existing procedures (for the model problem) and indeed as h→0, will be considerably less efficient than the H(≫2)-parameter P.R. (A.D.I.) process.

An advantage this new preconditioning technique has over the old (see section 7.1) is that we have been able to develop an exact theoretical treatment of rates of convergence and optimum parameters, for the commutative case. This has not so far been possible for Triangular Preconditioning. However, provided the coefficient matrix A is positive definite, then for T.P., B_0 is positive definite also, and our two preconditioned methods (7.54) and (7.55) (using T.P.) will therefore converge, by (7.58) and (7.61). For A.D.P. this result has only been improved when H and V commute. If they do not, the matrix B_0 (for A.D.P) is not symmetric, and therefore could possess complex eigenvalues. Also in the non-commutative case, a theoretical analysis of the eigenvalues of B_0 seems improbable in view of Theorem 3.4. Nevertheless the problem relating to Figure 7.2 is a "non-commutative one", but the rates of convergence using the parameters from the unit square (for the model problem) are almost identical to those convergence rates on the square. Little or nothing is known however, about the behaviour of these two A.D.P. iterative methods on more general regions in the (x-y) plane.
Section B The Biharmonic Equation

7B.1 Statement of the problem and formation of the difference equations

We consider the solution of
\[ \nabla^4 u(x, y) = \frac{\partial^4 u}{\partial x^4} + 2 \left( \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} \right) = f(x, y), \quad (x, y) \in R \tag{7.83} \]
where \( R \) is the rectangular region defined by (7.7). With \( f(x, y) = 0 \)
for example, the biharmonic equation (7.83), together with appropriate
boundary conditions, describes the small deflections in a thin
homogeneous plate under various forces exerted on the boundary of the
plate (rectangle).

The particular boundary conditions we shall be considering, in
connection with the approximate solution of equation (7.83), are
given by,
\[
\begin{aligned}
U(x, y) &= e(x, y) & \text{for all } (x, y) \in B, \\
\frac{\partial^2 U}{\partial n^2}(x, y) &= g(x, y) 
\end{aligned}
\tag{7.83}^
\]
where \( B \) is the boundary of \( R \), and \( \partial \nabla n \) is the normal derivative to \( B \).

Using Taylor series, we obtain the following finite difference
analog to (7.83) at the point \((x+iAx, y+jAy)\)
\[
\left\{ \frac{(\Delta y)^2}{(Ax)^2} [u_{i+2,j} - 4u_{i+1,j} + 6u_{i,j} - 4u_{i-1,j} + u_{i-2,j}] \right\}
+ \left\{ \frac{(\Delta x)^2}{(Ay)^2} [u_{i,j+2} - 4u_{i,j+1} + 6u_{i,j} - 4u_{i,j-1} + u_{i,j-2}] \right\}
+ 2 \left\{ u_{i-1,j-1} - u_{i+1,j-1} + u_{i-1,j+1} - u_{i+1,j+1} \right\}
- 2 \left\{ u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} \right\} = (Ax\Delta y)^2 f_{i,j} \cdot (7.84)
\]
Similar expressions to equations (1.4) and (1.6) with \( Ax \) and \( Ay \) re-
placing \( h \), respectively, are used to approximate \( \partial U/\partial n \) on the boundary
Then, superimposing a rectangular grid (as defined by (7.9)) on the
region \( R \), and ordering the mesh points horizontally, yields the system
of equations
\[ Au = Hu + Vu + 2Mu = h \tag{7.85} \]
where the matrices \( H, V, M \) correspond to the curly brackets in (7.84) in an analogous manner to the derivation of equations (3.61) and (3.62). More precisely, if

\[
H_L = \begin{bmatrix}
2 & -1 \\
-1 & -1 \\
-1 & 2
\end{bmatrix}_{(n \times n)}
\]  

then

\[
H = \left( \frac{\partial y}{\partial x} \right)^2 \text{ diagonal } \{ H_L \} \text{ in block } n \times n \text{ form.}
\]  

Also, if

\[
V_L = \begin{bmatrix}
2I & -I \\
-1 & -1 \\
-1 & -I \\
-I & 2I
\end{bmatrix}_{(nm \times nm)}
\]  

\( I = \text{unit matrix of order } n \), then

\[
V = \left( \frac{\partial y}{\partial x} \right)^2 V_L^2
\]  

and

\[
M = V_L \cdot \text{ diagonal } \{ H_L \}.
\]  

It is easily verified that the matrices \( H, V \) and \( M \) are symmetric and commute, and so by Theorem 3.4 have a common set of orthonormal eigenvectors. In particular from equations (7.85)-(7.90) we see that

\[
A = \left[ \left( \frac{\partial y}{\partial x} \right) \text{ diagonal } \{ H_L \} \left( \frac{\partial x}{\partial y} \right) V_L \right]^2
\]  

It thus follows from (7.91) and Theorem 1.2 that \( A \) is symmetric and positive definite, so system (7.85) has a unique solution.

75.3 The Alternating Direction Preconditioning \( (A.D.P.) \) technique

Let us re-write the coefficient matrix \( A \) as

\[
A = I + \hat{H} + \hat{V} + 2H
\]
where
\[ \hat{H} = H - I/2, \quad \hat{V} = V - I/2. \]  
(7.92)

If (7.86) is pre-multiplied by \((I+w\hat{H})^{-1}\) then it is equivalent to
\[ (I+w\hat{H})^{-1}A(I+w\hat{V})^{-1}(I+w\hat{V})u = (I+w\hat{H})^{-1}b, \]  
(7.93a)
or
\[ B_\omega \cdot \hat{V} = d, \]  
(7.93b)

where \(B_\omega \cdot \hat{V}\) and \(d\) (with \(H, V\) specified by (7.92)) are defined by (7.15).

The objectives now are exactly the same as those in section A, i.e. to minimise \(P(B_\omega)\) with respect to \(O_{\omega_2}\). Since \(H, V\) and \(M\) commute, it follows from (7.92), that \(H, V\) and \(A\) are pairwise commutative. Thus if \(\mu, \nu\) denote the real positive eigenvalues of \((\Delta y/\Delta x)\) diagonal \((H_L)\) and \((\Delta x/\Delta y) V_L\) respectively, with
\[ 0 < \omega_1, \quad 0 < \omega_2, \]  
(7.94)

we have from equations (7.87)-(7.93)
\[ \lambda = \frac{(\mu+\nu)^2}{(1-w/2+\omega^2)(1-w/2+\omega v^2)}, \]  
(7.95)

where \(\lambda\) is an eigenvalue of \(B_\omega\). We see that if \(0 < \omega_2\), then \(\lambda > 0\) for all \(\lambda\), so it follows from Theorem 1.2 that \(B_\omega\) is positive definite.

Also, for the problem in question, the eigenvalues \(\mu\) and \(\nu\) are given by
\[ \mu_i = \frac{\Delta y}{\Delta x} \cdot \frac{4}{\sin^2 \frac{\pi}{2(m+1)}}, \]  
for \(i=1,2,\ldots,n\);
\[ \nu_j = \frac{\Delta x}{\Delta y} \cdot \frac{4}{\sin^2 \frac{\pi}{2(m+1)}}, \]  
for \(j=1,2,\ldots,m\).
(7.96)

So from (7.94) we have
\[ a = \frac{\Delta y}{\Delta x} \cdot \frac{4}{\sin^2 \frac{\pi}{2(m+1)}}, \quad b = \frac{\Delta y}{\Delta x} \cdot \frac{4}{\cos^2 \frac{\pi}{2(m+1)}}, \]  
(7.97)
\[ c = \frac{\Delta x}{\Delta y} \cdot \frac{4}{\sin^2 \frac{\pi}{2(m+1)}}, \quad d = \frac{\Delta x}{\Delta y} \cdot \frac{4}{\cos^2 \frac{\pi}{2(m+1)}}. \]

To simplify the argument, let us firstly assume that the region of problem solution is the UNIT SQUARE, so that \(\Delta x=\Delta y=b\) and \(m=n\). If now follows for this region that \(a=c\) and \(b=d\) from (7.97).
7B.3 Minimisation of P-condition number of coefficient matrix

To examine the maximum and minimum eigenvalues of $\lambda(B_\omega)$ given by (7.95), with respect to $\mu_i$ and $\nu_j$ (for $i, j \in \mathbb{N}$) and then the $\min P(B_\omega)$, consider the continuous function

$$ f(x, y, \omega) = \frac{(x+y)^2}{(1-\omega/2+wx^2)(1-\omega/2+wy^2)} $$

$$ = g(x, y, \omega) + h(x, y, \omega), \quad 0 < \alpha \leq x, y \leq \beta \quad (7.98) $$

where

$$ g(x, y, \omega) = \frac{x^2+y^2}{(1-\omega/2+wx^2)(1-\omega/2+wy^2)} \quad (7.99) $$

and

$$ h(x, y, \omega) = \frac{2xy}{(1-\omega/2+wx^2)(1-\omega/2+wy^2)} \quad (7.100) $$

Since $0 < \omega < 2$, and $0 < \alpha \leq x, y \leq \beta$,

$$ g(x, y, \omega), \quad h(x, y, \omega) > 0. \quad (7.100) $$

(i) Minimum eigenvalue of $B_\omega$ for $0 < \omega < 2$.

From (7.98) and (7.100) we have

$$ \min f(x, y, \omega) = \min_{x, y \leq \beta} (g(x, y, \omega) + h(x, y, \omega)) $$

$$ \geq \min_{0 < \omega} g(x, y, \omega) + \min_{0 < \omega} h(x, y, \omega), \quad 0 < \omega < 2. \quad (7.101) $$

By (7.27) and (7.18),

$$ \min_{0 < \omega} g(x, y, \omega) = \begin{cases} g(\alpha, \alpha, \omega), & 0 < \omega < \overline{\omega} \\ g(\beta, \beta, \omega), & \overline{\omega} < 2 \end{cases} \quad (7.102) $$

where, in this case $\overline{\omega} = 2/(1+2\alpha \beta)$ since $0 < \alpha^2 \leq x, y \leq \beta^2$.

Letting

$$ p(x) = \frac{\sqrt{2x}}{(1-\omega/2+wx^2)} \quad (7.103) $$

we have from (7.99),

$$ h(x, y, \omega) = p(x)p(y). \quad (7.104) $$
Next, we obtain

\[
\frac{dp}{dx} = \sqrt{2} \frac{u((1-u/2)/u-x^2)}{(1-u/2+ux^2)^2}
\]

\[
\begin{cases}
\leq 0 \text{ when } x^2 \geq \frac{(1-u/2)}{u} \\
\geq 0 \text{ when } x^2 \leq \frac{(1-u/2)}{u}
\end{cases}
\]  

(7.105)

Hence if \(a^2 \geq (1-u/2)/u\), i.e. \(2/(1+2a^2) \leq u<2\), then from (7.105), \(dp/dx \leq 0\). Since \(p(x)>0 \) for all \(0<\alpha \leq \beta\), \(0<\omega<2\), we have

\[
\min_{\alpha \leq \beta} p(x) = p(\beta)
\]

and

\[
\max_{\alpha \leq \beta} p(x) = p(\alpha)
\]

min \(p(x) = p(\beta)\) \(\alpha \leq \beta\)

\[
\alpha \leq \beta
\]

(7.106)

 Likewise, if \(\beta^2 \leq (1-u/2)/u\), i.e. \(0<\omega \geq 2/(1+\beta^2)\), then from (7.105), \(dp/dx \leq 0\). Therefore

\[
\min_{\alpha \leq \beta} p(x) = p(\alpha)
\]

and

\[
\max_{\alpha \leq \beta} p(x) = p(\beta)
\]

(7.107)

Consider \(2/(1+\beta^2) \geq 2/(1+2a^2)\). In this range,

\[
\frac{dp}{dx} = 0 \text{ when } x = \sqrt{\frac{(1-u/2)}{u}}
\]

(7.108)

It is easily verified that this value of \(x\) gives a maximum i.e.

\[
\max_{\alpha \leq \beta} p(x) = p\left(\sqrt{\frac{(1-u/2)}{u}}\right)
\]

(7.109)

when \(\frac{2}{1+2a^2} \geq \frac{2}{1+2\beta^2}\).

Hence for \(\omega\) in this range,

\[
\min_{\alpha \leq \beta} p(x) = p(\alpha) \text{ or } p(\beta)
\]

\(0<\omega \leq \beta\)

It is then also easily shown that

\[
\min_{\alpha \leq \beta} p(x) = \begin{cases}
p(\alpha), & \frac{2}{1+2a^2} \geq \omega \\
p(\beta), & \omega<2
\end{cases}
\]

(7.110)
where $\bar{w}=2/(1+2\alpha)$. Hence from equations (7.104), (7.106), (7.107) and (7.110) we obtain

$$
\min_{a \in \mathbb{R}, y \in \mathbb{R}} h(x, y, w) = \begin{cases}
    h(z, c, \omega), & 0 < \omega \leq \bar{w} \\
    h(z, c, \omega), & \omega \leq 2.
\end{cases} \quad (7.111)
$$

Combining this result with (7.102) we see that equality holds in (7.101), i.e.

$$
\min_{a \in \mathbb{R}, y \in \mathbb{R}} f(x, y, w) = \begin{cases}
    f(z, c, \omega), & 0 < \omega \leq \bar{w} \\
    f(z, c, \omega), & \omega \leq 2.
\end{cases} \quad (7.112)
$$

\begin{enumerate}
\item[(ii)] Maximum eigenvalue of $B_w$ for $0 < \omega < 2$.
\end{enumerate}

From (7.98) and (7.100) we have

$$
\max_{a \in \mathbb{R}, y \in \mathbb{R}} g(x, y, w) + \max_{a \in \mathbb{R}, \in \mathbb{R}} h(x, y, w) = 0 < \omega < 2. \quad (7.113)
$$

By (7.41) and (7.18),

$$
\max_{a \in \mathbb{R}, y \in \mathbb{R}} g(x, y, w) = \begin{cases}
    g(z, c, w), & 0 < \omega \leq 2/(1+2a^2), \\
    g(z, c, w), & 2/(1+2a^2) \leq \omega \leq 2.
\end{cases} \quad (7.114)
$$

Next, from equations (7.103)-(7.109) we have

$$
\max_{a \in \mathbb{R}, y \in \mathbb{R}} h(x, y, w) = \begin{cases}
    h(z, c, \omega), & 0 < \omega \leq 2/(1+2a^2), \\
    h(z, c, \omega), & 2/(1+2a^2) \leq \omega < 2.
\end{cases} \quad (7.115)
$$

where $\bar{x} = \sqrt{(1-w^2)/w}$. Combining equations (7.113)-(7.115) yields

$$
\max_{a \in \mathbb{R}, y \in \mathbb{R}} f(x, y, w) = \begin{cases}
    f(z, c, \omega), & 0 < \omega \leq 2/(1+2a^2) \\
    f(z, c, \omega), & 2/(1+2a^2) \leq \omega \leq 2.
\end{cases} \quad (7.116)
$$

\begin{enumerate}
\item[(iii)] The $P$ condition number of $B_w(0 < \omega < 2)$.
\end{enumerate}

From (7.95) we see that

$$
\min_{a \in \mathbb{R}, y \in \mathbb{R}} \lambda(\nu, \omega) = \min_{a \in \mathbb{R}, y \in \mathbb{R}} f(x, y, w) \text{ given by (7.112)}. \quad (7.117)
$$
Next, \( \max_{\omega} \lambda(\mu, \nu, \omega) \) is given approximately by (7.116).

Combining equations (7.112) and (7.116) we obtain

\[
P(\omega) = \max_{\omega} \lambda(\mu, \nu, \omega) \quad \text{min}_{\omega} \lambda(\mu, \nu, \omega)
\]

\[
= \frac{f(\beta, \beta, \omega)}{f(\alpha, \alpha, \omega)}, \quad 0 < \omega < \frac{2}{1+2\beta^2},
\]

\[
= \frac{g(\alpha, \beta, \omega) + h(\bar{x}, \bar{x}, \omega)}{f(\alpha, \alpha, \omega)} \cdot \frac{2}{1+2\beta^2} \omega^2,
\]

\[
\leq \frac{g(\alpha, \beta, \omega) + h(\bar{x}, \bar{x}, \omega)}{f(\beta, \beta, \omega)} \cdot \omega^2 \frac{2}{1+2\beta^2} \omega < 2. \quad (7.118)
\]

\((\omega = 2/(1+2\alpha \beta))\). Using exactly the same approach as that used in subsection 7A.4(iii), we arrive at

\[
\min_{\omega} P(\omega) \leq \frac{g(\alpha, \beta, \omega) + h(\bar{x}, \bar{x}, \omega)}{f(\alpha, \alpha, \omega)} \quad (7.119)
\]

We also have

\[
\bar{x}(\omega) = \sqrt{\frac{-\omega^2}{\omega^2}} = \sqrt{\beta^2} \quad (7.120)
\]

Substituting for \( \bar{x} \) and \( \omega \) in equations (7.98)-(7.99), (7.119) becomes

\[
\min_{\omega} P(\omega) \leq \frac{(3\beta^2 + 3\beta^4 + 2\alpha \beta)}{8 \alpha \beta} \quad (7.121)
\]

In fact, from (7.98) it is easily verified that

\[
\max_{\omega} f(x, y, \omega) = f(\bar{x}, \bar{x}, \omega) \quad (7.122)
\]

This yields from (7.118) and (7.112),

\[
\min_{\omega} P(\omega) \leq \frac{f(\bar{x}, \bar{x}, \omega)}{f(\alpha, \alpha, \omega)} = \frac{(\alpha + \beta)^2}{4 \alpha \beta} \quad (7.123)
\]

a smaller upper bound than in (7.121), provided \( \alpha < \beta \).
Considering specifically the case when $R$ is the unit square, from (7.97)

$$\alpha = 4\sin^2(wh/2), \quad \beta = 4\cos^2(wh/2), \quad h = 1/(n+1). \quad (7.124)$$

Therefore,

$$\hat{\omega} = \frac{2}{1+2\alpha \beta} = \frac{2}{1+8\sin^2(wh)} \quad (7.125)$$

and from (7.123),

$$\min_{0<\omega<2} P(B_\omega) \leq \frac{(\alpha + \beta)^2}{4\alpha \beta} = \frac{1}{\sin^2(wh)} \quad (7.126)$$

When $\omega = 0$, we have from (7.95),

$$P(0) = P(B_{\omega=0}) = \cot^2(wh/2) \gg 1/(\sin^2 wh) \quad (7.127)$$

If the eigenvalues $\lambda_{i,j}(a,b) = (i,j=1,2,\ldots,n)$ of $B_\omega$ are such that

$$0 < \lambda_{i,j}(a,b) \leq \hat{\omega} \quad (7.128)$$

then from (7.112) and (7.122) we have

$$\tilde{a} = \min_{1 \leq i,j \leq n} \lambda_{i,j}(\omega) = f(\alpha,\alpha,\omega) = f(\beta,\beta,\omega) \quad (7.129)$$

$$\bar{b} = \max_{1 \leq i,j \leq n} \lambda_{i,j}(\omega) \leq \max_{a \leq x,y \leq b} f(x,y,\omega) = f(x,\bar{x},\omega) \quad (7.129)$$

where from (7.120), $\bar{x}(\omega) = \sqrt{\omega \beta} = 2\sin(wh)$.  

(iv) **Case when eigenvalue ranges of $u,v$ are different**

The approximate analysis for this case follows along similar lines to subsection 7A.4(iv).  If $u,v$ are such that $a \leq u \leq b$ and if $\lambda(u,v,\omega)$ is an eigenvalue of $B_\omega$ (defined by (7.93)) then

$$\min_{0<\omega<2} P(B_\omega) \leq \min_{0<\omega<2} \left\{ \max_{a \leq x,y \leq b} \lambda(u,v,\omega) \right\} \leq \min_{0<\omega<2} \left\{ \max_{a \leq x,y \leq b} \lambda(u,v,\omega) \right\} \leq P(\tau) \quad (7.129)$$

Here we have $\alpha = \min(a,\alpha)$, $\beta = \max(b,\beta)$ and $\tau = 2/(1+2\alpha \beta')$.  The above result follows from section 7B.3.

Such a case with $a \leq u$ and $b \leq \beta$ occurs for the solution of problem 7B.1 on the rectangle with $Ax = Ay$ and $\tau$: see equations (7.97).
7.4 Application of the A.D.P. technique to two basic iterative methods

Having obtained our optimum (or "near" optimum) preconditioning parameter \( \omega = \omega^* \) and the spectral bounds \( \bar{\omega}, \underline{\omega} \), for the positive real eigenvalues \( \lambda_{i,j} (1 \leq i, j \leq n) \) of \( B^{-1} \), we can again use the Simultaneous Displacements method and Richardson's second degree method to solve system (7.93).

The Computational procedure

This proceeds in an analogous manner to the procedure outlined in section 7A.5. Since the two are so similar, we will not repeat it in bulk, but point out the differences. We first assume that for this case, \( \bar{\omega}, \underline{\omega} \) and \( \hat{H} \) and \( \hat{V} \) are defined by (7.125, (7.57), (7.91), (7.93) and (7.92) respectively. The only difference between the two cases is that for the present one, we now have quin-diagonal systems of equations to solve instead of tri-diagonal ones (this is apparent from the definitions of \( \hat{H} \) and \( \hat{V} \) in (7.32) and (7.86)-(7.89).

Let us assume then that we have ordered the unknown grid points along successive \( \omega \)-lines of the grid. The matrix \((I+\hat{\omega})\) is then symmetric, positive definite and quin-diagonal, so that we can use Algorithm 4.4 to solve system (7.63). Then to evaluate the first part of (7.64), i.e. \( \Sigma^{(k)} = (I+\hat{\omega})^{-1} \Sigma^{(k)} \), we re-order both vectors to correspond to a column (or \( y \)-line) ordering of the mesh points. A new system, \((I+\hat{\omega})X^{(k)} = Z^{(k)} \) results, the coefficient matrix being quin-diagonal and positive definite (this follows particularly from (7.84)), so we can again use Algorithm 4.4 to solve this system. System (7.66) is evaluated in an analogous manner to before, except we now have a 13-point difference formula (from (7.84)) instead of a 5-point one. Expression (7.64) is finally evaluated by solving (7.67) in the manner already illustrated. We now solve for \( \Sigma^{(k+1)} \) from equation (7.68) the process continuing until convergence (to the desired accuracy) in \( \Sigma \) is obtained. The original solution vector \( u \) is then obtained from \((I+\hat{\omega})u = y \) in the way described above.
7B.5 Rates of convergence on the square

The rate of convergence of the present A.D.P. Simultaneous Displacements method for solving problem 7B.1, is given by (7.59). Substituting \( P(\beta) = 1/\sin^2(\theta h) \) from (7.126) in the former equation yields

\[
R_\infty (I - \overline{\alpha}^2) = 2w_h^2, \quad \text{as } h \to 0. \tag{7.130}
\]

Substituting the same value of \( P(\beta) \) in equation (7.62) gives the asymptotic rate of convergence of the A.D.P. second degree Richardson iterative process for approximating the solution of problem 7B.1, namely,

\[
R_\infty = 2w_h^2, \quad \text{as } h \to 0. \tag{7.131}
\]

We see from (7.78) that this is identical to the asymptotic rate of convergence of the "tri-diagonal" A.D.P. Simultaneous Displacements procedure for solving the model problem of section 3.2.

7B.6 The Douglas-Rachford (A.D.I.) iterative method

The method used by Conte and Dames (1958) to approximate the solution of problem 7B.1, is the A.D.I. scheme which corresponds to the Douglas-Rachford method for second order equations. Applied to the iterative solution of system (7.85) this is

\[
(r_{k+1}I+H)u(k+1) = (r_{k+1}I-2I-V)u(k), \quad k \geq 0 \tag{7.132a}
\]

and

\[
(r_{k+1}I+V)u(k+1) = Vu(k) + \sum_{k+1}^u(k+1), \tag{7.132b}
\]

where \( r_{k+1} \) is a positive parameter chosen to accelerate convergence.

The recommended set of iteration parameters given by the above two methods are

\[
r_{k+1} = 16x_k^k, \quad k = 1, 2, \ldots, \| \tag{7.133}
\]

for \( 0 < x < 1 \). The number \( M \) of parameters for a given value of the mesh size \( h \) (in D, their region of problem solution was the unit square) satisfies

\[
M \geq 1 + \frac{4 \log(\sin(wh/2))}{\log x}. \tag{7.134}
\]
They come to the conclusion, that for the above scheme (7.132) to give the best results, 0 < x < 0.2.

We notice from (7.132) and (7.86)-(7.89) that the coefficient matrices on the right of equations (7.132a) and (7.132b) are both direct sums of positive definite symmetric quidiagonal matrices after suitable permutations of their rows and columns, so that we can solve both systems directly using Algorithm 4.4.

7B.7 Numerical Results

The problem solved was the biharmonic equation

\[ \nabla^4 U(x,y) = 0, \quad (x,y) \in \mathbb{R}, \]

where the region \( \mathbb{R} \) was the unit square. The boundary conditions were

\[ U(x,y) = \frac{\partial^2 U}{\partial n^2}(x,y) = 0, \quad \text{for all } (x,y) \in \mathbb{B}, \]

where \( \mathbb{B} \) is the boundary of the square, and \( \partial/\partial n \) is the normal derivative to \( \mathbb{B} \). The number of iterations which were required to make

\[ \max_{1 \leq i,j \leq n} |y_{i,j}^{(k+1)} - y_{i,j}^{(k)}| < 10^{-6} \]

for the two above preconditioned iterative methods, and

\[ \max_{1 \leq i,j \leq n} |u_{i,j}^{(k+1)} - u_{i,j}^{(k)}| < 10^{-6}, \]

(\( k \) = number of acceleration parameters) for the D.R. (A.D.I.) process, starting with \( y_{i,j}^{(0)} = 1, u_{i,j}^{(0)} = 1, 1 \leq i,j \leq n \), are given in Figure 7.3. The logarithm of \( N \), the number of iterations are plotted against \( \log(h^{-1}) \).

For the A.D.P. Simultaneous Displacements and A.D.P. second degree Richardson methods the optimum preconditioning parameter \( \bar{\omega} \) used, was that specified in (7.125), while the optimum acceleration parameters used are those defined by (7.57) and (7.60) respectively, using the values of \( \bar{\omega} \) from (7.129).

Regarding the Douglas-Rachford A.D.I. scheme (or D.R. (A.D.I.) for short), we at first chose \( x = 0.2 \) in (7.136). The number of parameters then to be used cyclically, for each value of \( h \), was taken to be the smallest integer satisfying (7.134). The results from Figure 7.3 show
Figure 7.3
that the A.D.P. Simultaneous Displacement method converges (as expected) much slower than the other two. We also notice that for this value of \( x \), the A.D.P. second order Richardson scheme converges in less iterations than D.R. (A.D.I.) for approximately \( h^{-1} \approx 12 \), and in more for \( h^{-1} > 12 \) (i.e. the "cross over point" is at \( h^{-1} = 12 \)).

In view of the fact that Conte and Dames (1958) recommended \( x < 0.2 \), it was decided to try values of \( x = 0.1(0.01)0.2 \) in (7.133) and (7.134) for the same values of \( h^{-1} \) as before, to find the 'best' values. These were found to be \( x_{\text{opt}} \) (say) = 0.15, 0.16, 0.17, 0.18 and 0.18 for \( h^{-1} = 6, 10, 15, \) and 30, respectively. The "cross over point" now for the two afore-mentioned methods is seen to be about \( h^{-1} = 6 \), although there is no significant difference in number of iterations until \( h^{-1} \gg 10 \).

We now look at the comparative amounts of work per iteration of the above two iterative methods. On the square, the D.R. (A.D.I.) method (7.132a), (7.132b) requires the implicit solution of \( 2n \) quidiagonal systems of equations by Algorithm 4.4. The coefficient matrices have the form (4.16), and for this particular problem (in that equation)

\[
d_i = b_{i+1} = i, \quad i = 1, 2, \ldots, n-1; \quad c_i = a_{i+2} = i, \quad i = 1, 2, \ldots, n-2
\]

(7.135)

Making use of relations (7.135) for one complete iteration (i.e. a double sweep of the grid) the D.R. (A.D.I.) method requires approximately \( 13n^2 \) multiplications, \( 6n^2 \) divisions and \( 30n^2 \) additions (see Conte and Dames (1958), p.204).

The A.D.P. second order Richardson method is defined by equation (7.55) where for this problem \( B_{\alpha} \), in particular, is defined by (7.93). From the latter equation we can write

\[
B_{\alpha}^{(k)} = \left[ I(1-\omega/2)+\frac{\gamma}{2} \right]^{-1} A \left[ I(1-\omega/2)+\omega \frac{\gamma}{2} \right]^{-1}
\]

\[
= \omega^{-2} \left[ \frac{\gamma}{2} I + \frac{\gamma}{2} \right]^{-1} A \left[ \frac{\gamma}{2} I + \frac{\gamma}{2} \right]^{-1}
\]

(7.136)
where \( r = \frac{(1-\omega/2)}{\omega} \). Adopting the same procedure as that described in section 7B.4, we can save \( n^2 \) multiplications per iteration by expressing \( B_{12}^{(k)} \) as in (7.135) and making use of the last relation in (7.135). Method (7.55) then requires approximately \( 12n^2 \) multiplications, \( 2n^2 \) divisions and \( 24n^2 \) additions, per iteration. In order to compare this amount of work with that given for D.R. (A.D.L) we ignore the additions in both and regard a multiplication and a division as being about the same, in terms of computing time. Thus the respective amounts of work are now approximately \( 19n^2 \) multiplications and \( 14n^2 \) multiplications. In other words, D.R. (A.D.L) requires about \( \frac{1}{3} \) times the work per iteration of the A.D.L second degree Richardson method.

Thus returning our attention to Figure 7.3, it would appear that with \( x(\text{fixed}) = 0.2 \) in (7.133), the former process is more efficient (in terms of total arithmetic operations) than the latter for approximately \( h^{-1} > 20 \). For \( h^{-1} \approx 20 \) (approximately) the roles are reversed. In the case when the "optimum" value of \( x \) (say \( x_{\text{opt}} \)) is used in (7.133), as mentioned above, the "cross over point" in terms of overall efficiency will be about \( h^{-1} = 12 \).
REFERENCES


Benson A. (1973), Private communication


Bickley and Thompson (1964), Matrices, their meaning and manipulation, London, EUP.


