Modified preconditioned iterative methods for solving elliptic partial differential equations

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

Additional Information:

- A Doctoral Thesis. Submitted in partial fulfillment of the requirements for the award of Doctor of Philosophy of Loughborough University.

Metadata Record: https://dspace.lboro.ac.uk/2134/10829

Publisher: © C.C. Okeke

Please cite the published version.
This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (https://dspace.lboro.ac.uk/) under the following Creative Commons Licence conditions.

For the full text of this licence, please go to: http://creativecommons.org/licenses/by-nc-nd/2.5/
MODIFIED PRECONDITIONED ITERATIVE METHODS
FOR SOLVING ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

BY

CHRISTOPHER C. OKEKE, B.Sc., M.Sc.

Submitted in partial fulfilment of the requirements
for the award of Doctor of Philosophy
of the University of Technology, Loughborough, 1982.

Supervisor: PROFESSOR D.J. EVANS, Ph.D., D.Sc.
Department of Computer Studies

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.C. OKEKE.
THIS THESIS IS DEDICATED TO

My Father, Christopher

and

My Mummy, Rose
ACKNOWLEDGEMENTS

My earnest thanks and gratitude goes first to my supervisor, Professor D.J. Evans for his excellent guidance and encouragement. He has demonstrated that given good direction, distance cannot be an effective deterrent to research.

I wish also to express my thanks to the present and former Vice-Chancellors of the University of Nigeria, Nsukka, Professor F.N. Ndili and Professor J.O.C. Ezilo, for granting me permission during the long vacations to enable me to execute this research.

For providing me unlimited access to the CDC 7600 in Manchester University and the ICL 1900 machines, I thank the staff of the Computer Centre of Loughborough University.

To my parents and wife I say a hearty "bravo" for their understanding.

I would also like to express my thanks to Miss J.M. Briers for the excellent work done towards the preparation of this thesis.

To my brothers, Dr. Shedrac and Meshac Okeke, I wish to express my profound gratitude for their financial and moral encouragement throughout the duration of this research.
# CONTENTS

## CHAPTER ONE: INTRODUCTION

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Classification and Approximate Solution for Partial</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Differential Equations</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Finite Difference Form for Generalised Dirichlet Problem</td>
<td>8</td>
</tr>
<tr>
<td>1.3</td>
<td>The Model Problem</td>
<td>10</td>
</tr>
<tr>
<td>1.4</td>
<td>Direct Methods for Solving a Set of Linear Equations</td>
<td>12</td>
</tr>
<tr>
<td>1.5</td>
<td>Some Point Iterative Methods</td>
<td>22</td>
</tr>
<tr>
<td>1.6</td>
<td>Variational Method</td>
<td>38</td>
</tr>
<tr>
<td>1.7</td>
<td>Evolution of Solution Techniques for Elliptic P.D.E.s</td>
<td>44</td>
</tr>
</tbody>
</table>

## CHAPTER TWO: MATRIX ANALYSIS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Preliminaries of Matrix Analysis</td>
<td>45</td>
</tr>
<tr>
<td>2.2</td>
<td>Matrix and Vector Norms</td>
<td>49</td>
</tr>
<tr>
<td>2.3</td>
<td>Convergence of Matrix Iterations</td>
<td>51</td>
</tr>
<tr>
<td>2.4</td>
<td>Irreducibility and Weak Diagonal Dominance</td>
<td>52</td>
</tr>
<tr>
<td>2.5</td>
<td>Consistently Ordered Matrices and Ordering Vectors</td>
<td>53</td>
</tr>
<tr>
<td>2.6</td>
<td>Matrices which Have Property A</td>
<td>54</td>
</tr>
</tbody>
</table>

## CHAPTER THREE: DIRECT PRECONDITIONED SOLUTION OF LINEAR SYSTEMS OF EQUATIONS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Error Bounds</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Condition Numbers</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>3.2.1 Condition Numbers for Linear Systems</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Application of Preconditioning Method for Direct Solution</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>of Linear Systems of Equations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.3.1 The First Order Transformation (F.O.T.)</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>3.3.2 The Second Order Transformation (S.O.T.)</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>3.3.3 The Modified Second Order Transformation (M.S.O.T.)</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>3.3.4 Discussion and Observations</td>
<td>84</td>
</tr>
</tbody>
</table>
## CHAPTER FOUR: APPLICATION OF MODIFIED PRECONDITIONING TECHNIQUES FOR THE CONSTRUCTION OF ITERATIVE METHODS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The Modified Preconditioned Jacobi (MPJ) Method</td>
<td>89</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Estimation of Good Bounds on $\xi_{\text{max}}(B)$ and $\xi_{\text{min}}(B)$</td>
<td>91</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Estimation of $s_{\text{max}}(B)$ and $\omega_1$</td>
<td>96</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Numerical Results</td>
<td>105</td>
</tr>
<tr>
<td>4.2</td>
<td>The Modified Preconditioned Simultaneous Displacement (MPSD) Method</td>
<td>110</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Convergence of the MPSD Method</td>
<td>110</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Obtaining Optimal Values for $\tau$ and $\omega$ for the MPSD Method</td>
<td>112</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Numerical Results</td>
<td>117</td>
</tr>
<tr>
<td>4.3</td>
<td>The Modified Successive Overrelaxation (MSOR) Method</td>
<td>124</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Matrices with Diagonal Dominance which are also Irreducible</td>
<td>124</td>
</tr>
<tr>
<td>4.3.2</td>
<td>L Matrices</td>
<td>126</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Positive Definite Matrices</td>
<td>127</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Consistently Ordered Matrices</td>
<td>128</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Numerical Results</td>
<td>140</td>
</tr>
<tr>
<td>4.4</td>
<td>The Modified Gauss-Seidel (MGS) Method</td>
<td>142</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Numerical Results</td>
<td>144</td>
</tr>
<tr>
<td>4.5</td>
<td>Generalised Preconditioned Conjugate Gradient Method</td>
<td>146</td>
</tr>
<tr>
<td>4.5.1</td>
<td>The MPSD Semi-Iterative (MPSD-SI) Method</td>
<td>148</td>
</tr>
<tr>
<td>4.6</td>
<td>Dynamic Acceleration of the Preconditioned Simultaneous Displacement (MPSD) Method</td>
<td>150</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Algorithm 4.6.1</td>
<td>152</td>
</tr>
<tr>
<td>4.7</td>
<td>Block Preconditioned Methods</td>
<td>155</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Introduction</td>
<td>155</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Group MPSD Methods</td>
<td>157</td>
</tr>
</tbody>
</table>

## CHAPTER FIVE: MODIFIED ALTERNATING DIRECTION PRECONDITIONING TECHNIQUE

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>160</td>
</tr>
<tr>
<td>5.2</td>
<td>Self-Adjoint Second Order Elliptic Partial Differential Equations</td>
<td>161</td>
</tr>
<tr>
<td>5.3</td>
<td>The Modified Alternating Direction Preconditioning Method</td>
<td>163</td>
</tr>
<tr>
<td>5.4</td>
<td>Obtaining a Minimal Estimate for the P-Condition Number of $B$</td>
<td>165</td>
</tr>
</tbody>
</table>
5.5 Application of Modified Preconditioned Method to Basic Iterative Methods

5.6 Numerical Results

CHAPTER SIX: SUMMARY AND CONCLUSIONS

REFERENCES
CHAPTER ONE

INTRODUCTION
1.1 CLASSIFICATION AND APPROXIMATE SOLUTION METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

In discussing partial differential equations we shall necessarily be concerned with problems in which the dependent variable is a function of two independent variables. Extensions to three or more variables can also be made, details of which are discussed in Forsythe and Wasow [15].

Three different classes of differential equations are distinguishable and usually different numerical approaches are used for the solution of each class. A class of differential equation may be made to assume a simple form (usually by a change of variable) called the canonical form for the given class. The various forms and their associated classifications are

\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \psi \] (elliptic) \hfill (1.1)

\[ \frac{\partial^2 u}{\partial x^2} = \psi \] (parabolic) \hfill (1.2)

\[ \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = \psi \] (hyperbolic) \hfill (1.3)

where

\[ \psi = \psi(x, y, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}) \]

The most commonly encountered elliptic equation in physical application is

\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \] (Laplace's equation) \hfill (1.4)

and

\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \text{constant } k \] (Poisson's equation) \hfill (1.5)

For higher order equations, the most common include among others, the fourth order biharmonic expression

\[ \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = k \] \hfill (1.6)

The most commonly used parabolic equation has the form

\[ \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial y} (a = \text{constant}) \] \hfill (1.7)
This is called the one dimensional transient diffusion or heat equation with \( y \) representing time.

Hyperbolic equations are often used in the form

\[
\frac{\partial^2 u}{\partial x^2} = \alpha \frac{\partial^2 u}{\partial y^2} \quad (\alpha = \text{constant}) \tag{1.8}
\]

This is called the wave equation.

For solutions of elliptic type equations to exist, there is a necessity that boundary conditions be specified at every point of a well defined (closed) boundary. Any local variation in the boundary conditions suffices to alter the solution over the entire region of the problem. Usually the physical application problem dictate the appropriate boundary conditions.

The number and type of boundary conditions which completely define a problem makes an unique solution dependent on the nature and order of the equation of the problem.

For parabolic and hyperbolic equations it is not in general necessary to have boundary conditions specified at every point of the boundary. Specification of conditions at only a segment of a boundary is often sufficient to determine the solution within a certain portion of the problem region. The boundary itself is not closed.

**Approximate Solution Methods**

There are two general classes of methods that can be used for the solution of partial differential equations. These are the finite difference [4] and finite element approximation methods [41].

The first of these approximation methods essentially entails obtaining finite difference expressions for various partial derivatives or mixed derivatives and so on for the given partial differential equation. For purposes of simplicity, consider equations in two independent variables only of the form,
\[
\psi(u(x,y), \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial x \partial y}, \ldots) = 0 \tag{1.9}
\]

and also derivatives of type
\[
\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial x \partial y} \tag{1.10}
\]

The finite difference approximations for these and others could in general be derived by Taylor series expansion methods. However, other methods exist which include the integration and variational formulation methods [30].

The desired approximation for a given region \( R \) of the \( x,y \) plane is obtained by superimposing a mesh of lines parallel to the reference axes in the given region. Let Figure 1.1 represent any such mesh over the region \( R \).

A regular pattern of square meshes is often used in problems of discretizations of partial differential systems into finite difference forms. Let the point \( 0 \) in the figure designate any point and we shall also assume that it is the origin of \( x \) and \( y \) and that the value \( u(x,y) \) at that point is \( u_0 \). By virtue of Figure 1.1, the following can be shown to be valid:

\[
2h \frac{\partial u}{\partial x} \bigg|_0 \approx u_1 - u_3
\]

\[
2h \frac{\partial u}{\partial y} \bigg|_0 \approx u_2 - u_4
\]
\[ h^2 \frac{\partial^2 u}{\partial x^2} \bigg|_0 = u_1 - 2u_0 + u_3 \]
\[ h^2 \frac{\partial^2 u}{\partial y^2} \bigg|_0 = u_2 - 2u_0 + u_4 \]

and
\[ h^4 \frac{\partial^4 u}{\partial x^4} \bigg|_0 = u_9 - 4u_1 + 6u_0 - 4u_3 + u_{11} \]
\[ h^4 \frac{\partial^4 u}{\partial y^4} \bigg|_0 = u_{10} - 4u_2 + 6u_0 - 4u_4 + u_{12} \]

For the mixed derivatives, we have that since
\[ \frac{\partial^2 u}{\partial x \partial y} = \frac{\partial}{\partial x} \left( \frac{\partial u}{\partial y} \right) = \frac{\partial}{\partial y} \left( \frac{\partial u}{\partial x} \right) \]
then
\[ \frac{\partial^2 u}{\partial x \partial y} \bigg|_0 = \left\{ \frac{\partial u}{\partial y} \bigg|_1 - \frac{\partial u}{\partial y} \bigg|_3 \right\} / 2h \]
and using (1.11), we have
\[ \frac{\partial^2 u}{\partial x \partial y} \bigg|_0 = \left( \frac{u_5 - u_8}{2h} - \frac{u_6 - u_7}{2h} \right) / 2h \]
and so
\[ 4h^2 \frac{\partial^2 u}{\partial x \partial y} \bigg|_0 = u_5 - u_6 + u_7 - u_8 . \]

Similarly,
\[ h^4 \frac{\partial^4 u}{\partial x^2 \partial y^2} \bigg|_0 = u_5 - 2u_2 + u_6 - 2u_1 + 4u_0 - 2u_3 + u_8 - 2u_4 + u_7 . \]

Other methods exist as we mentioned earlier for deriving the finite difference forms but in practice, the finite differences are generated on the grid by using the appropriate computational molecule. Very often the matrix forms of the linearized approximation to differential systems can be associated with known structures and fundamental properties. It is apparent for example that the expression
\[ h^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = \left( u_1 - 2u_0 + u_3 \right) + \left( u_2 - 2u_0 + u_4 \right) \]
\[ = u_1 + u_2 + u_3 + u_4 - 4u_0 \]
(1.11a)
could be expressed in terms of values on a five point computational molecule.

It is apparent that the finite difference expression (1.11a) of the computational molecule as well as the various derivations are only an
approximation of the exact value. There is therefore an associated truncation error for every finite difference discretization of a given derivation.

Let us assume that the solution of a required partial differential equation has the required continuous derivatives of higher order in sufficiently large neighbourhood about a point \((x,y)\) and a constant mesh size \(h\), then by Taylor's expansion theorem the values at the surrounding mesh points can be determined, namely,

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

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

and

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

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

Combining these formulae we have

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

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

and

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

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

Let us denote \(u(x,y)\) by \(u_{i,j}\) for a general mesh point \((x,y) = (ih,jh)\). The Poisson equation for example using (1.11c) and (1.11d) can be replaced at the point \((x_i,y_j)\) by

\[
\frac{1}{h^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}) = g_{i,j} + \frac{h^2}{12} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)_{i,j} + \ldots
\]

or equivalently,

\[
u_{i,j} - u_{i+1,j} + u_{i-1,j} - u_{i,j+1} + u_{i,j-1} = h^2 g_{i,j} - \frac{h^4}{12} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)_{i,j}
\]

As the local truncation error of the above formulae are given by the terms on the right hand side of (1.11f), excluding \(-h^2 g_{i,j}\), while the term
$O(h^4)$ is the principal part of the error. This truncation error can be ignored if $h$ is assumed small.

The second class of methods that could be used to obtain solutions is the **Finite Element method**. This approach is relatively recent and derived originally from solution techniques largely applied in solid mechanics (elasticity, plasticity and structural analysis). This class of method is rapidly gaining ground not only in solid mechanics but also in fluid mechanics, heat transfer problems and other application areas [41]. In the finite element approach the partial differential equations describing a desired quantity (such as displacement) in the continuum are not handled directly. Instead, the continuum is divided into a number of 'strips' or 'finite elements' which are assumed to be joined at a discrete number of points along their boundaries. A functional form is then chosen to represent the variation of the desired quantity over each element in terms of values of this quantity at discrete boundary points of the element. By using physical properties of the continuum and the appropriate physical laws (quite often by use of a minimization principle) a set of simultaneous equations in the unknown quantities at the element boundary points can be obtained. The matrix set is normally large but it generally has a banded structure. Finite element approximation approach seems to have the advantage over the finite element approach with regard to:

(a) The topology of the region over which the problem is defined. Irregularly shaped regions can be handled much more easily than with finite difference method.

(b) Possibility exists for variation of strips or finite elements thus permitting the use of small elements where large variations occur and large elements where gentle variations are expected.
(c) To obtain comparative accuracy, the finite element can usually be much larger than the mesh elements of a finite difference grid. As a result, when elliptic problems are involved the band matrix derived by use of finite element method is usually small enough to permit direct rather than the iterative methods which are commonly used for finite difference methods.
1.2 **FINITE DIFFERENCE FORM FOR THE GENERALISED DIRICHLET PROBLEM**

Let us assume that $R$ is a bounded, connected plane region with a boundary $C$ consisting of one or more differential curves and $\psi(x,y)$ a function defined on $C$. Let $u(x,y)$ be a function continuous in $R+C$, twice differentiable in $R$ and satisfying in $R$ the general second order P.D.E.

\[
\frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2} + d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial y} + fu = h
\]  

(1.12a)

and on $C$ the Dirichlet condition

\[
u(x,y) = \psi(x,y)
\]

(1.12b)

The exact solution $u(x,y)$ to the above P.D.E. in the region $R$ satisfies the equation (1.12a) at every point in $R$ and matches the 'boundary conditions' on $C$. The coefficients of (1.12a) may be constants (including the value zero) or functions of independent variables $x$ and $y$ (in each case) or functions of the dependent variable $u$ and its first derivatives (non-linear case).

Consider the self adjoint partial differential equation

\[
\frac{\partial^2}{\partial x^2} [P \frac{\partial u}{\partial x} + Q \frac{\partial u}{\partial y}] + Zu = G \quad (x,y) \in R
\]

(1.12)

where $P$, $C$, $Z$ and $G$ are assumed to be continuous analytic functions in $x$ and $y$ and satisfy the conditions

\[
P > 0, \quad Q > 0 \quad \text{and} \quad Z \geq 0.
\]

(1.13)

In order to derive a finite difference form for (1.12), we need the imposition of a rectangular spatial mesh consisting of vertical and horizontal lines on the region $R$ of the problem with spacing such that $h = x_{i+1} - x_i$ and $k = y_{j+1} - y_j$. The continuous problem (1.12) can now be replaced by the discrete form for every point $(x,y)$ of $R_h$ by the symmetric difference equation [30].

\[
P(x+h,y)\{u(x+h,y)-u(x,y)\}+Q(x,y+h)\{u(x,y+h)-u(x,y)\}/h^2
\]

\[
+P(x+h,y)\{u(x+h,y)-u(x,y)\}+Q(x,y)u(x,y)=G(x,y)
\]

(1.14)
Relation (1.14) when assembled for each of the node points represents the
discrete generalised Dirichlet problem.

If we multiply (1.14) by $-h^2$, we obtain the difference form

$$u(x,y) = \psi_1(x,y)u(x+h,y) + \psi_2(x,y)u(x,y+h) + \psi_3(x,y)u(x-h,y) + \psi_4(x,y)u(x,y-h) + S(x,y)$$

(1.15)

where

$$\psi_1(x,y) = P(x+h,y)S^{-1}(x,y)$$
$$\psi_2(x,y) = Q(x,y+h)S^{-1}(x,y)$$
$$\psi_3(x,y) = P(x-h,y)S^{-1}(x,y)$$
$$\psi_4(x,y) = Q(x,y-h)S^{-1}(x,y)$$

(1.16)

and

$$B(x,y) = -h^2 G(x,y)/S(x,y)$$

and

$$S(x,y) = P(x+h,y) + P(x-h,y) + Q(x,y+h) + Q(x,y-h) - h^2 Z(x,y).$$

(1.17)

It can be shown Varga [30] that the discrete generalised problem (1.15) can be reduced to the linear form

$$Pu = b,$$

(1.18)

where one equation corresponds to each of the mesh points $(x_i, y_j)$ on $R_h$ where $u_{i,j}$ is unknown. It could also be shown [30] that the matrix $P$ of (1.18) is real, symmetric and has property 'A' as well as being an L-matrix which is irreducible and possesses weak diagonal dominance. (The concept of property A, L-matrix and irreducibility will be defined later). The sparse structure of $P$ necessarily necessitates the use of iterative procedures for the solution of (1.18).
1.3 THE MODEL PROBLEM

Consider the following 2D-model problem:

We seek to determine a continuous, twice differentiable function \( u(x,y) \) in the region, \( \mathbb{R} = (0,1) \times (0,1) \) with boundary \( \mathbb{C} \), satisfying Laplace’s equation
\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 , \quad (x,y) \in \mathbb{R} ,
\]
subject to the Dirichlet boundary conditions
\[
u(x,y) = 0 , \quad (x,y) \in \mathbb{C} .
\]
The region \( \mathbb{R} \) under consideration is covered by a rectilinear net with mesh spacing \( h \) in \( x,y \) directions and mesh points \((x_i, y_j)\), where \( x_i = ih, i=0,1,\ldots,M; y_j = jh, j=0,1,\ldots,M.\)

This problem is a special case \( (C_1(x,y) = C_2(x,y) \equiv 1, P(x,y) = Q(x,y) \equiv 0) \) of the generalised Dirichlet problem given in Section 1.2.

Substituting the finite difference approximations for the derivatives in (1.19), the following five point formula is obtained
\[
4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = 0 .
\]
If we order the \((M-1)^2\) internal mesh points column-wise (Figure 1.2)

![Figure 1.2](image)

the coefficient matrix \( P \) of system (1.18) is a real, square, quindiaagonal, sparse matrix of order \( m^2 = (M-1)^2 \) and of general form
where $I$ is an identity matrix of order $m$ and $P_i, i \in [1,m]$ matrices of order $m$, given by

$$P_i = \begin{bmatrix}
4 & -1 & 0 \\
-1 & 4 & -1 \\
-1 & 4 & -1 \\
0 & -1 & 4
\end{bmatrix}, i \in [1,m].$$

(1.23)

The solution vector $u$ and the right hand side vector $b$ of (1.18) are $(m^2 \times 1)$ column vectors.
1.4 DIRECT METHODS FOR SOLVING A SET OF LINEAR EQUATIONS

Suppose we are interested in solving system (1.18) where $P$ is a square matrix and $b$ is a given column matrix while $u$ is an unknown column matrix. It will be assumed that $P$ is non-singular and so a unique solution exists.

The methods that may be used to obtain the solution of such systems depend mainly upon the structure of the coefficient matrix $P$. For large sparse matrices having a few non-zero elements, iterative methods are normally used. Iterative methods do not change the structure of the original matrix and so sparsity is preserved. The disadvantages of such methods are often the problem of choice of good starting solution vector and that of accuracy of the final solution. The main advantage of iterative methods is that sparse matrix techniques can be used to store only the non-zero elements of the coefficient matrix and hence optimise the amount of storage used.

Direct methods on the other hand are used when the given coefficient matrix is small and dense. The advantage of these methods is that the choice of initial values is not critical and the solution is obtained after a finite number of steps. We shall now describe some well known direct solution methods. In such methods a sequence of operations is performed, in general, once only which results in an approximation to the true result. The approximation enters only because multiplications and divisions are stored and subsequently used with rounding errors.

The 'Gaussian-elimination' method is generally used to solve a system of equations of form (1.18) by reducing $P$ to an upper triangular matrix. The elimination process in the matrix $P$ takes place in $(N-1)$ steps and involves evaluating the matrix $P^{(k)}$, $k=1,2,...,N-1$ where in general after the $(k-1)^{th}$ step of elimination, the matrix $P^{(k-1)}$ has the form:
where the elements $p_{i,k}^{(k-1)}$, $i=k+1,k+2,...,n$ are now eliminated by calculating

$$p_{i,j}^{(k)} = p_{i,j}^{(k-1)} - p_{i,k}^{(k-1)} p_{k,j}^{(k-1)} / p_{k,k}^{(k-1)}, \ i,j=k+1,k+2,...,n. \tag{1.23.2}$$

The accuracy of the final solution depends on the choice of the pivots.

For an accurate result all the pivots have to be selected in such a way that the elements $p_{i,k}^{(k-1)} / p_{k,k}^{(k-1)}$ which represent the multipliers for the different stages of the elimination process (i.e. $k=1,2,...,n-1$) should be less than or equal to unity. To achieve this we consider two of the most important pivotal strategies commonly used.

**Full pivoting.** In using this strategy a search is made at the start of $k^{th}$ ($k=1,2,...,n-1$) step of elimination to find the maximum element in absolute value of the submatrix $(n-k+1) \times (n-k+1)$ which is formed by eliminating the $k^{th}$ row and column $n$ ($k=2,3,...,n-1$) of the matrix and it is used as a pivotal element for the stage.

**Partial pivoting.** In this method a search is made at the $k^{th}$ step of the elimination process only amongst the elements $p_{i,k}^{(k)}$ ($i=k, k+1, ..., n$), and the largest element in absolute value is selected as the pivot.
The Gaussian elimination method is usually programmed to accommodate k right hand side vectors, in which case the storage required is n(n+k) locations plus a possible further nk for the final solution vectors. Finally, the amount of work involved can be seen to be

\[ n \text{ divisions} \]
\[ \frac{1}{3}n^3 + \frac{2}{3}n - \frac{1}{3}n \text{ multiplications} \]  
\[ \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \text{ additions} \]  

In practice, however, in the solution of sets of equations resulting from finite difference approximations to elliptic partial differential equations, this value of work involved is never actually attained as the matrix P is never full. Instead, P is in general a band matrix, i.e.

\[ P = (p_{i,j}) \text{ where} \]
\[ p_{i,j} = 0 \quad \text{if } i-j > m \]
\[ p_{i,j} = 0 \quad \text{if } j-i > m \]

so that the number of non-zero elements in each row is at most 2m+1 the bandwidth of the matrix. In this case, the number of multiplications is of order \( 2m^2n \) in contrast to the factor \( n^3/2 \) for the full matrix.

The 'LU Decomposition' of a matrix P can be obtained by the method of Doolittle which we now illustrate for a general \((n\times n)\) matrix. For the relation \( P = LU \) or

\[
\begin{bmatrix}
    p_{1,1} & \cdots & p_{1,j} & \cdots & p_{1,n} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    p_{i,1} & \cdots & p_{i,j} & \cdots & p_{i,n} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    p_{n,1} & \cdots & p_{n,j} & \cdots & p_{n,n}
\end{bmatrix}
= \begin{bmatrix}
    1 & & & & \\
    \vdots & \ddots & \vdots & & \\
    \vdots & & \ddots & \vdots & \\
    \vdots & & & \ddots & \\
    \vdots & & & & 1
\end{bmatrix}
\begin{bmatrix}
    u_{1,1} & \cdots & u_{1,j} & \cdots & u_{1,n} \\
    0 & \ddots & \vdots & \ddots & \vdots \\
    0 & \ddots & \ddots & \vdots & \vdots \\
    \vdots & \ddots & \ddots & \ddots & \vdots \\
    0 & \cdots & 0 & \cdots & 1
\end{bmatrix}
\]

(L is lower triangular and U is upper triangular) all the coefficients in L and U are initially unknown. They are found by the following rules based on matrix multiplication.
This factorisation method fails only if one of the diagonal elements of $u$ which are used as divisors in the second part of (1.23.4) is zero.

As in the Gaussian elimination process, similar pivotal strategies must be used to ensure sufficiently accurate results. We now illustrate that the LU decomposition uniquely exists if $P$ is a non-singular matrix. For this we have the following theorems.

**Theorem 1.1.1**

The $(n \times n)$ matrix $P$ has an LU decomposition where $L$ and $U$ are non-singular, if the submatrices

\[
P_k = \begin{bmatrix}
  p_{1,1} & \cdots & \cdots & p_{1,k} \\
  \vdots & \ddots & \ddots & \vdots \\
  \vdots & & \ddots & \vdots \\
  p_{k,1} & \cdots & \cdots & p_{k,k}
\end{bmatrix},
\]

are all non-singular ($P_k$ is called the leading principal matrix of order $k$).

**Proof:**

Let us assume an inductive hypothesis such that

\[
P_{k-1} = L_{k-1} U_{k-1},
\]

where $L_{k-1}$ is a unit lower triangular matrix.

Let $P_k$ be partitioned as:

\[
P_k = \begin{bmatrix}
P_{k-1} & b \\
T & \hat{P}_{k,k}
\end{bmatrix},
\]

where $T$ is a matrix and $\hat{P}_{k,k}$ is a matrix.

This factorisation method fails only if one of the diagonal elements of $u$ which are used as divisors in the second part of (1.23.4) is zero.

As in the Gaussian elimination process, similar pivotal strategies must be used to ensure sufficiently accurate results. We now illustrate that the LU decomposition uniquely exists if $P$ is a non-singular matrix. For this we have the following theorems.

**Theorem 1.1.1**

The $(n \times n)$ matrix $P$ has an LU decomposition where $L$ and $U$ are non-singular, if the submatrices

\[
P_k = \begin{bmatrix}
p_{1,1} & \cdots & \cdots & p_{1,k} \\
\vdots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \vdots \\
p_{k,1} & \cdots & \cdots & p_{k,k}
\end{bmatrix},
\]

are all non-singular ($P_k$ is called the leading principal matrix of order $k$).

**Proof:**

Let us assume an inductive hypothesis such that

\[
P_{k-1} = L_{k-1} U_{k-1},
\]

where $L_{k-1}$ is a unit lower triangular matrix.

Let $P_k$ be partitioned as:

\[
P_k = \begin{bmatrix}
P_{k-1} & b \\
T & \hat{P}_{k,k}
\end{bmatrix},
\]
where 
\[ P^T = [p_{k,1}^{\cdot\cdot\cdot}, p_{k,k-1}^{\cdot\cdot\cdot}] \] and 
\[ b = \begin{bmatrix} p_{1,k} \\ \vdots \\ p_{k-1,k} \end{bmatrix} \]
and let \( L_k \) and \( U_k \) be defined by

\[
L_k = \begin{bmatrix} L_{k-1} & 0 \\ \varepsilon^T & 1 \end{bmatrix} \quad \text{and} \quad U_k = \begin{bmatrix} U_{k-1} & U \\ 0^T & u_{k,k} \end{bmatrix}
\]

where \( \varepsilon^T = [\varepsilon_{k,1}^{\cdot\cdot\cdot}, \varepsilon_{k,k-1}^{\cdot\cdot\cdot}] \) and \( u = \begin{bmatrix} u_{1,k} \\ \vdots \\ u_{k-1,k} \end{bmatrix} \)

then \( \varepsilon^T u \) and \( u_{k,k} \) can be uniquely determined such that

\[
L_k U_k = \begin{bmatrix} L_{k-1} U_{k-1} & L_{k-1} u \\ \varepsilon^T & \varepsilon^T u + u_{k,k} \end{bmatrix} = \begin{bmatrix} P_{k-1} & b \\ P^T & p_{k,k} \end{bmatrix} = P_k
\]

This follows since \( L_{k-1} U_{k-1} = p_{k-1} \) by hypothesis and then

\[
\varepsilon^T = P^T U_{k-1}, \quad u = L_{k-1}^{-1} b
\]
and

\[
u_{k,k} = p_{k,k} - \varepsilon^T u
\]
is uniquely determined since \( L_{k-1} \) and \( U_{k-1} \) are non-singular under the hypothesis. Now,

\[
P_1 = [p_{1,1}] = [I][p_{1,1}],
\]
defines the decomposition of \( P_1 \). Hence by induction, \( P = P_n \) has an LU decomposition.

We may note that if \( P \) is a real matrix then \( L \) and \( U \) are also real. Row or column diagonally dominant matrices satisfy the hypotheses of the theorem.
Theorem 1.1.2

If there exists an LU decomposition of a non-singular matrix $P$, where $L$ is unit lower triangular and $U$ is upper triangular matrix, then this decomposition is unique.

Proof:

Let $P = L(1)U(1) = L(2)U(2)$ be two decompositions, where the matrices are non-singular. Since $P$ is non-singular, then

$$L^{-1}(2)L(1) = U(2)U^{-1}(1) = I,$$  \hspace{1cm} (1.23.10)

since $L^{-1}(2)L(1)$ is unit lower triangular and $U(2)U^{-1}(1)$ is upper triangular.

Equivalence of Gaussian Elimination with LU Decomposition

Define the matrix $M_k$ by

$$M_k = \begin{bmatrix}
1 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & -m_{i,k} & \cdots & 1 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \cdots & -m_{n,k} & \cdots & 0
\end{bmatrix}$$  \hspace{1cm} (1.23.11)

(i.e. identity matrix with additional elements $-m_{i,k}$ in $(i,k)^{th}$ position for $k+1 \leq i \leq n$). Then, the premultiplication of a matrix $P$ by $M_k$ has the effect of $-m_{i,k}$ or row $k$ being added to row $i$ for $k+1 \leq i \leq n$.

The Gaussian elimination algorithm without pivoting can then be described theoretically by

$$M_{n-1}^* \cdots M_k^* \cdots M_1^* P = \begin{bmatrix}
\vdots & \vdots & \vdots \\
p_{1,1} & \cdots & p_{1,k} & \cdots & p_{1,n} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
p_{k,1} & \cdots & p_{k,k} & \cdots & p_{k,n} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
p_{n,1} & \cdots & p_{n,k} & \cdots & p_{n,n}
\end{bmatrix} \equiv U, \text{ (say)}$$  \hspace{1cm} (1.23.12)
where 

\[ m_{i,k} = \frac{p_i^{(k)}}{p_{k,k}^{(k)}}. \]

Thus we have

\[ P = M_1^{-1}, ..., M_k^{-1}, ..., M_{n-1}^{-1} * U. \]  

(1.23.13)

But

\[ M_1^{-1}, ..., M_k^{-1}, ..., M_{n-1}^{-1} = \begin{bmatrix}
1 \\
\vdots \\
m_{k,1} & \cdots & -1 \\
m_{i,1} & \cdots & m_{i,k} & 1 \\
m_{n,1} & \cdots & m_{n,k} & \cdots & m_{n,i} \end{bmatrix} \]

\[ \equiv L, \text{ (say)} \]

(1.23.14)

Thus,

\[ P = LU, \]

where \( L \) is a unit lower triangular matrix and \( U \) is upper triangular matrix.

It can be shown that the amount of work in the LU decomposition is the same as for the Gauss method.

However, the disadvantage of the Gaussian elimination process in comparison with the LU decomposition is that each time an element of the reduced matrix is computed and stored a rounding error generally occurs. In the LU decomposition a rounding error can often be avoided by the use of double-precision arithmetic in the calculation of the elements of \( L \) and \( U \) equations (1.23.4). These results may then be rounded to single precision and reduced on completion of each calculation. The removal of the need for computing and recording several intermediate matrices has therefore localised what might otherwise be a significant source of error to a single step in the determination of each element of \( L \) and \( U \). The use of double precision arithmetic in this step leads to a degree of accuracy comparable with that attained if the entire Gaussian elimination process
were carried out with double-precision, this latter is unattractive because it would require twice as much computer storage as the corresponding single precision solution.

The 'Crout Factorisation' is a form of factorisation which is used for symmetric matrices and it will factorise \( P \) in the form \( U^T D U \) in a recursive procedure where at each step of the process, we write \( P \) as

\[
P = P^{(1)} = \begin{bmatrix}
p_1 & r_1^T \\
r_1 & c_1
\end{bmatrix}
\]

(1.23.15)

where \( p_1 \) is a single element, \( r_1 \) is an \((n-1)\times1\) vector and \( c_1 \) is an \((n-1)\times(n-1)\) matrix. The first element of the diagonal matrix \( D \), i.e. \( d_{1,1} \) and the first row of the upper triangular matrix \( U \) is now computed as

\[
u_{1,j} = \left[ r_1^T / p_1 \right] , j = 2, 3, \ldots, n
\]

and

\[d_{1,1} = p_1 \quad (1.23.16)\]

We then compute \( P^{(2)} \) as:

\[
P^{(2)} = \begin{bmatrix}
p_1 & r_1^T \\
r_1 & c_1 - \frac{r_1^* r_1}{p_1}
\end{bmatrix}
\]

(1.23.17)

where the matrix \( \frac{r_1^* r_1}{p_1} \) is subtracted element by element from \( c_1 \).

In general, at the \( k \)th step \((1 \leq k \leq n-1)\) we have \( P^{(k)} \) as

\[
P^{(k)} = \begin{bmatrix}
p_k & B_k^T \\
B_k & c_k - \frac{r_k^* r_k}{p_k}
\end{bmatrix}
\]

(1.23.18)

The 'Choleski Decomposition' (square root method) is used when the matrix is symmetric and positive definite.
Let the matrix $P$ be symmetric and positive definite, then we can factorise $P$ in the form $LL^T$ where $L$ is lower triangular. As an illustration consider the decomposition of a general $n \times n$ symmetric and positive definite matrix in the form

![Matrix Decomposition Diagram]

Then equating coefficients on and below the diagonal we obtain the unknown coefficients of $L$ from the following equation

$$
\begin{align*}
    \ell_{i,i} &= \left[ p_{j,i} - \sum_{k=1}^{j-1} \ell_{j,k}^2 \right]^{1/2} \\
    \ell_{i,j} &= \frac{1}{\ell_{i,i}} \left[ p_{j,i} - \sum_{k=1}^{j-1} \ell_{i,k} \ell_{j,k} \right] \\
    \text{for } j &= 1, 2, \ldots, n
\end{align*}
$$

(1.23.19)

**Theorem 1.1.3**

Let $P$ be a real symmetric positive definite matrix, then there exists unique decompositions of the form

$$
P = LDL^T = LL^T
$$

where $L$ is a real unit lower triangular matrix and $D$ is diagonal with real positive terms and $L$ is a real lower triangular matrix with positive diagonal.
Proof.

If \( P \) is real, symmetric and positive definite then its leading principal matrices \( P_k \) are non-singular and moreover are such that
\[
\det(P_k) > 0 ; \quad k=1,2,\ldots,n.
\]
Thus by Theorem 1.1.3 there exists a unique decomposition
\[
P = LU
\]
where \( L \) is a real lower and \( U \) is a real upper triangular matrix.

Moreover, since
\[
u_{k,k} = \begin{cases} 
p_{1,1} & k=1, \\
\det(P_k)/\det(P_{k-1}) & k>1,
\end{cases}
\]
(1.23.20)
it follows that \( U \) has positive diagonal terms. Now the decomposition may be rewritten in the unique form
\[
P = LDV.
\]
(1.23.21)
where \( D=\text{diag}(\nu_{k,k}) \) is a diagonal matrix with positive terms and \( V=\text{D}^{-1}U \) is a real unit upper triangular matrix. Uniqueness and symmetry then imply that \( V=L^T \). Thus,
\[
P = LDL^T = Lb^2D^{-1}L^T = LL^T,
\]
(1.23.22)
where \( L=lb \) and \( b \), and hence \( L \) can be uniquely chosen to have positive diagonal terms.

Finally, it can be seen that the use of symmetric elimination methods, i.e. the Crout and the Choleski methods give a significant saving in computational cost. Both methods require \( O(1/6 n^3) \) multiplications and \( O(1/6 n^3) \) additions. However, the Choleski method requires \( n \) square roots as well. Therefore, Choleski's method is likely to be less efficient than the Crout's method although it possesses superior accuracy and smaller storage requirements. The pivotal strategies can be applied with the same case and would involve no additional arithmetic. However, extra time is required to search for pivots and to rearrange the matrix accordingly. For symmetric matrices, it should be noted that only the lower or upper part of the matrix need be stored for elimination and backward/forward substitution to solve the linear system.
1.5 SOME POINT ITERATIVE METHODS

If we seek the solution of the linear system (1.18), where the co-efficient matrix $P$ is a large sparse matrix with not more than five non-zero elements in any row, e.g. formula (1.21) then iterative methods of successive approximations are usually used in preference to direct methods such as Gaussian elimination and Cramer's rule.

An iterative process starts with an arbitrary initial approximation to the solution and then it is successively modified in accordance to a pre-defined rule.

Consider the system (1.18) and let $u^{(n)}$ be a sequence such that

$$u^{(n)} \rightarrow p^{-1}b \quad \text{as} \quad n \rightarrow \infty.$$  

Generally, an iterative scheme is said to be of degree $\xi$ if $u^{(n)}$ is a function of $P, b, u^{(n-1)}, u^{(n-2)}, \ldots, u^{(n-\xi)}$.

Since storage requirements need to be minimised, the value of $\xi$ must be reasonably small.

Consequently, for an iteration of first degree ($\xi=1$), we have

$$u^{(n)} = f_n(P, b, u^{(n-1)}). \quad (1.24)$$

The iterative process (1.24) is said to be:

(i) 'linear' if $f_n$ is a function of $u^{(n-1)}$ and

(ii) 'stationary' if $f_n$ is independent of $n$.

A general iterative scheme of form

$$u^{(n+1)} = \theta^{(n)} + f_n(u^{(n)}) ,$$

is said to be 'non-stationary' if either $\theta$ or $f_n$ (or both) are dependent on $n$.

Let us assume a splitting for the matrix $P$ of the form

$$P = E - F ,$$

such that the system (1.18) can be written as

$$Eu = Fu + b . \quad (1.25)$$
Assuming that \( u^{(n)} \) denotes the \( n \)th approximation to the solution \( u \), then the following iterative scheme can be obtained

\[
E u^{(n+1)} = F u^{(n)} + b,
\]

where the initial approximation (starting vector) \( u^{(0)} \) is an arbitrary column vector.

The matrix \( E \) (assuming that \( \det(E) \neq 0 \) and that \( E^{-1} \) exists) is chosen so that \( u^{(n+1)} \) can be easily obtained from (1.26), given that \( u^{(n)} \) is known. It is the structure of \( E \) which determines two important classes of iterative methods, namely:

(i) if \( E \) is a diagonal matrix, then the methods of 'Simultaneous Displacements' [e.g. Jacobi, Richardson, etc.] are developed.

(ii) if \( E \) is a lower triangular matrix, the methods of 'Successive Displacements' [e.g. Gauss-Seidel, Successive Overrelaxation (S.O.R.), etc.] are developed.

It should be noted however, that although in the 'Successive Displacement' method the ordering of the mesh points is such that the latest estimate \( u_i^{(n+1)} \) of the components of \( u \) is used in order to determine a new estimate of a component \( u_i^{(n+1)} \), in the 'Simultaneous Displacement' method the order in which the components \( u_i^{(n+1)} \) is obtained has no relevance.

Let us now define some basic methods in the class of 'Simultaneous Displacement' methods.

**Class of Simultaneous Displacement Methods**

Consider without loss of generality that the coefficient matrix \( P \) of (1.18) can be expressed in the form

\[
P = D - L - U,
\]

where \( D \) is a positive diagonal matrix whose elements are the elements of \( P \) and \( L \) and \( U \) are respectively lower and upper triangular matrices with zero diagonal entries.
Since $D$ is positive diagonal matrix, then $D^{-1}$ exists. Let

$$B = D^{-1}(L+U) \quad \text{and} \quad c = D^{-1}b \quad (1.28)$$

Using (1.18), (1.27) and (1.28), we have

$$u = Bu + c, \quad (1.29)$$

The iterative scheme defined by

$$u^{(n+1)} = Bu^{(n)} + c, \quad n \geq 0, \quad (1.30)$$

is defined as the 'Jacobi' iteration and can be equivalently written as

$$u^{(n+1)} = u^{(n)} + (c - D^{-1}Bu^{(n)}). \quad (1.31)$$

The method of 'Simultaneous Displacement' is defined by

$$u^{(n+1)} = u^{(n)} + \beta (c - D^{-1}Bu^{(n)}), \quad (1.32)$$

where $\beta$ is an acceleration parameter (constant) chosen to accelerate the convergence of the iterations.

Assuming, without loss of generality that $D$ may be chosen as the identity matrix $I$, the Jacobi (1.30) and Simultaneous Displacement (1.32) can be rewritten respectively as follows

$$u^{(n+1)} = (L+U)u^{(n)} + b, \quad (1.33)$$

and

$$u^{(n+1)} = u^{(n)} + r^{(n)}, \quad (1.34)$$

or equivalently,

$$u^{(n+1)} = u^{(n)} + r^{(n)}, \quad (1.35)$$

and

$$u^{(n+1)} = u^{(n)} + \beta r^{(n)} = (I - \beta D)u^{(n)} + \beta b, \quad (1.36)$$

where $r$ is the residual vector defined by

$$r^{(n)} = b - Pu^{(n)} \quad (1.37)$$

**Convergence of Simultaneous Displacement Methods**

The iterative methods previously described can be written in the general form

$$u^{(n+1)} = \tilde{M}u^{(n)} + d, \quad n \geq 0, \quad (1.38)$$

where $\tilde{M}$ is called the 'iteration matrix' of the method considered and $d$ is a column vector of constants.
Consider now the system (1.18) and the iterative scheme (1.38).

**Definition 1.1**

The 'consistency condition' for an iterative process is said to be satisfied if the solution \( \tilde{u} \) of system (1.18) is substituted for \( u^{(n)} \) in (1.38) then \( u^{(n+1)} \) is also the solution \( u \) of the system, i.e. the iterative procedure makes no further modification of successive iterates once the solution has been obtained.

**Definition 1.2**

An iterative method is said to be 'convergent' if for all initial vectors \( u^{(0)} \), each component of the successive iterates \( u^{(n)} \) tends to the corresponding component of the solution \( u \) of \( Pu=b \), for any given \( b \).

**Theorem 1.1**

An iterative method which can be expressed in the form of equation (1.38) converges if and only if the spectral radius of \( \tilde{M} \) satisfies

\[ \rho(\tilde{M}) < 1. \]

**Proof**

Let \( e^{(n)} = u^{(n)} - \tilde{u} \) be the error vector after \( n \)-iterative steps. We assume that the iterative method is consistent, i.e.

\[ u = \tilde{M}u + \tilde{d}, \tag{1.39} \]

then using (1.38) and (1.39), we obtain

\[ e^{(n+1)} = \tilde{M}e^{(n)}, \tag{1.40} \]

and furthermore,

\[ e^{(n)} = \tilde{M}e^{(0)}, \tag{1.41} \]

where \( e^{(0)} \) is the error vector associated with the initial vector \( u^{(0)} \).

Assuming that \( e^{(0)} \) is bounded i.e. \( |e^{(0)}_i| < \xi \), \( i \in [1,N] \) for some constant \( \xi \), then from (1.41) it follows that \( e^{(n)} \to 0 \) as \( n \to \infty \) if and only if \( \tilde{M} \to Z \) (where \( Z \) is a null matrix) as \( n \to \infty \).

By a Theorem in [30] (p.82) (stating: if \( P \) is an arbitrary \( n \times n \) matrix
then P is convergent iff $\rho (P) < 1$; this will be true if and only if $\rho (\tilde{M}) < 1$.

It can also be shown [38] that if P has certain properties that the Jacobi method converges.

Rate of Convergence of Simultaneous Displacement Methods

The effectiveness of an iterative method is generally considered from the computational work required per iteration and from the number of iterations required for convergence.

In practice, we assume that an iterative method has converged when
\[ ||e^{(n)}|| < \varepsilon ||e^{(0)}||, \tag{1.42} \]
where $e$ is an error vector, $\varepsilon$ is a preassigned positive factor and $||.||$ denotes the $L_2$ or spectral norm (to be defined later).

From (1.41), we have that
\[ ||e^{(n)}|| = \| \tilde{M}^{(n)} e^{(0)} \| \leq \| \tilde{M}^{n} \| \| e^{(0)} \|, \tag{1.43} \]
(note that in the case of a non-stationary iteration the relationships (1.41), (1.43) become
\[ ||e^{(n)}|| = \prod_{i=0}^{n} \tilde{M}^{(i)} e^{(0)} , \tag{1.44} \]
\[ ||e^{(n)}|| = \prod_{i=0}^{n} \| \tilde{M}^{(i)} \| \| e^{(0)} \|, \tag{1.45} \]
respectively.

It is known that $\| \tilde{M}^{n} \| \to 0$ as $n \to \infty$ iff $\rho (\tilde{M}) < 1$, so (1.42) can be satisfied by $n$ to be large enough such that
\[ \| \tilde{M}^{n} \| \leq \varepsilon . \tag{1.46} \]
Then, for all $n$ sufficiently large that $\| \tilde{M}^{n} \| < 1$, (1.46) is equivalent to
\[ n > \ln \varepsilon / (- \frac{1}{n} \ln \| \tilde{M}^{n} \|) . \tag{1.47} \]

Definition 1.3

Consider the iterative method (1.38), then we define the 'average rate of convergence' by
\[ R_n (\tilde{M}) = \frac{1}{n} \ln \| \tilde{M}^{n} \|. \tag{1.48} \]
Definition 1.4

The 'asymptotic rate of convergence' is defined by

\[ R(\tilde{M}) = \lim_{n \to \infty} R_n(\tilde{M}) = \ln \rho(\tilde{M}). \]  

(1.49)

Henceforth, we shall refer to \( R(M) \) simply as the 'rate of convergence'.

It can be shown \cite{38} (p. 87), that \( \rho(\tilde{M}) = \lim \left( \left| |\tilde{M}_n|\right| \right)^{1/n} \), so equality (1.49) is a direct consequence of (1.48).

We have to note that from the Definitions 1.2 to 1.4 we have that

(a) the number of iterations needed to reduce an initial error by a preassigned factor is approximately inversely proportional to the average rate of convergence.

(b) the rapidity of convergence depends on the rapidity with which the norm \( |\tilde{M}_n| \) tends to zero, as \( n \to \infty \).

(c) the smaller the value of \( \rho(\tilde{M}) \), the greater the rate of convergence \( R(\tilde{M}) \). Consequently iterative processes achieve their fastest rate of convergence when \( \rho(\tilde{M}) \) is reduced to a minimum.

From the simultaneous iteration defined by (1.38), it can be shown that the error vector satisfies:

\[ e^{n+1} = (I-SP)e^n, \]  

(1.50)

or

\[ e^{n+1} = (I-SP)^{(n+1)}e^0. \]  

(1.51)

Note that the error operator is constant throughout the iteration (stationary, linear iterative procedure).

Let us assume that the matrix \((I-P)\) has \( N \) linearly independent eigenvectors \( v_i \), associated with \( N \)-distinct eigenvalues \( \lambda_i \) and let \( \mu_i \), \( i \in [1,N] \) be eigenvalues of \( P \). Given that an arbitrary vector \( e^{(0)} \) can be expressed as

\[ e^{(0)} = \sum_{i=1}^{N} c_i v_i, \]  

(1.52)

where \( c_i, i \in [1,N] \) constants, we have that
\[ e_{(n+1)} = \sum_{i=1}^{N} c_i^2 \lambda_{(n+1)}^i v_i = \lambda_N^{(n+1)} \{ c_1 \left( \frac{1}{N} \right) v_1 + c_2 \left( \frac{2}{N} \right)^{n+1} + \ldots \}. \]

(1.53)

Hence, the iterative scheme converges if \( \rho(I-\beta P) < 1 \), i.e. the modulus of the largest eigenvalue of \( (I-\beta P) \) must be less than unity.

Let \( m, n \) be extreme eigenvalues of \( P \), i.e.,

\[ 0 < m \leq \mu_i \leq M < \infty, \quad i \in [1, N]. \]

(1.54)

Since \( \lambda_i = 1 - \beta \mu_i \), \( i \in [1, N] \),

the necessary and sufficient condition for convergence gives

\[ |1-\beta \mu_i| < 1, \quad i \in [1, N], \]

hence,

\[ 0 < \beta < 2/M. \]

(1.55)

(1.56)

The choice of the parameter \( \beta \) in such a way that \( \rho(I-\beta P) \) is minimised, leads to the fastest rate of convergence. The optimal \( \beta \) is obtained for the smallest of \( \max\{|1-\beta m|, |1-\beta M|\} \) and this value of \( \beta \) is given by

\[ |1-\beta m| = -|1-\beta M|, \]

or equivalently,

\[ \beta = \frac{2}{m+M}. \]

(1.57)

With this choice of \( \beta \), for all \( i \), we have

\[ |1-\beta \mu_i| \leq \frac{M-m}{M+m} < g. \]

(1.58)

From (1.49) the asymptotic rate of convergence \( R_s \) is given by

\[ R_s = -\ln(\rho(I-\beta P)) \]

(1.59)

and so from (1.58), we obtain

\[ R_s > \frac{2}{M} \quad \text{for } \frac{m}{M} > 1; \]

(1.60)

where the ratio \( \frac{M}{m} \) is known as the \( P \)-condition number of \( P \).

Note that the rate of convergence for the Simultaneous Displacement method is independent of the ordering of the points.

If a sequence of different factors \( \beta_n \) (instead of the constant factor \( \beta \)) for each iteration is multiplied by each component of the residual vector \( r^{(n)} \) and then added to each component of \( u^{(n)} \) to give the next value
of $u^{(n+1)}$, the following non-stationary iterative scheme known as

'Richardson's Method' is obtained.

$$u^{(n+1)} = u^{(n)} + \beta \frac{r^{(n)}}{n} = (I - \beta \frac{1}{n} P) u^{(n)} + \beta \frac{b}{n}. \quad (1.61)$$

The constants $\beta$ are either given in terms of extreme eigenvalues of the matrix $P$ or calculated during the iteration from formulae involving $u^{(n)}$.

The error vector for the iterative scheme (1.61) is given by

$$e^{(n+1)} = (I - \beta \frac{1}{n} P)e^{(n)}; \quad (1.62)$$

or

$$e^{(n+1)} = \prod_{i=0}^{n-1} (I - \beta \frac{1}{i} P)e^{(0)}. \quad (1.63)$$

Note that the error operator changes for each iteration (non-stationary iterative method).

Let

$$F_{n+1}(P) = \prod_{i=0}^{n-1} (I - \beta \frac{1}{i} P). \quad (1.64)$$

If $\lambda_i^1, i \in [1,N]$ are the $N$-eigenvalues of the positive definite matrix $P$ and $v_i^1$ are the corresponding eigenvectors, then the eigenvalues and eigenvectors of $F_{n+1}(P)$ are $F_{n+1}(\lambda_i^1)$ and $v_i^1$ respectively. A combination of (1.63), (1.52) gives

$$e^{(n+1)} = F_{n+1}(P)e^{(0)} = \sum_{i=1}^{N} F_{n+1}(\lambda_i^1) c_i v_i^1$$

or

$$e^{(n+1)} = \sum_{i=1}^{N} c_i v_i^1 F_{n+1}(\lambda_i^1). \quad (1.65)$$

The minimization of the error vector is now equivalent to the following problem.

Find a polynomial $F_{n+1}(x)$ of degree $(n+1)$ such that

$$\max_{x \in [M, m]} |F_{n+1}(x)|, \quad (1.66)$$

is minimised under the constraint

$$F_{n+1}(0) = 1. \quad (1.67)$$

The polynomial satisfying (1.66), (1.67) is given in [22a] to be

$$F_{n+1}(x) = \frac{T_{n+1} \frac{M+m-2x}{M-M}}{T_{n+1} \frac{M+m}{M-M}}, \quad (1.68)$$
where $T_n$ is the Chebyshev polynomial of degree $n$ given by

$$T_n(x) = \cos(n\cos^{-1}x) = \frac{1}{2}([x+\sqrt{x^2-1}]^n + [x-\sqrt{x^2-1}]^{-1}),$$  \hspace{1cm} (1.69)$$

adjusted in the interval $[-1,1]$. The optimal choice of parameters $\beta_i$ is such that the zeroes of $F_{n+1}(x)$ are

$$x_i = [\beta_i^{(n+1)}]^{-1}, \ i \in [1,n+1],$$  \hspace{1cm} (1.70)$$

while from (1.68), the zeroes of $T_{n+1}(\frac{M+m-2x}{M-m})$ and consequently of $F_{n+1}(x)$ are:

$$\frac{M+m-2x_i}{M-m} = \cos\left(\frac{(2i-1)\pi}{2(n+1)}\right), \ i \in [1,n+1].$$  \hspace{1cm} (1.71)$$

By equating the zeroes of the polynomials we obtain

$$\beta_i^{(n+1)} = \frac{2}{(M+m)-(M-m)\cos[\frac{(2i-1)\pi}{2(n+1)}]},$$  \hspace{1cm} (1.72)$$

and assuming that only $\nu$-iterations are to be performed then,

$$\beta_i^{(\nu)} = \frac{2}{(M+m)-(M-m)\cos[\frac{(2i-1)\pi}{2}]}, \ i \in [1,\nu]$$  \hspace{1cm} (1.73)$$

Since the maximum absolute value of the numerator of (1.68) is unity, then the maximum value of $F_{n+1}(x)$, as $n \to \infty$, is given by,

$$\max_{x \in [m,M]} |F_{n+1}(x)| = [T_{n+1}(\frac{M+m}{M-m})]^{-1}.$$  \hspace{1cm} (1.74)$$

Consider now the iterative scheme (1.61).

**Definition 1.5**

The average convergence factor for the first $\nu$-iterative steps is defined by the quantity

$$-\ln\left(\left|\frac{|\|e^{(\nu)}\||}{|\|e^{(0)}\||}\right|^{1/\nu}\right).$$

It can be shown [15] that the average rate of convergence is bounded by

$$-\frac{1}{\nu} \ln2 + \ln(Z_O + \sqrt{Z_O^2 - 1}),$$  \hspace{1cm} (1.75)$$

where

$$Z_O = \frac{M+m}{M-m} > 1.$$

The rate of convergence, is the asymptotic value of (1.75) as $\nu \to \infty$ and is given by,
Given that for most problems
\[ Z_0 = 1 + 2 \frac{M}{m} \]
then we obtain
\[ R_R = \frac{2}{\sqrt{\frac{M}{m}}} \quad \text{for} \quad \frac{M}{m} > 1. \quad (1.77) \]

Assuming that the extreme eigenvalues \( m, M \), the value of \( \nu \) and the value of the ratio
\[ \frac{||x^{(0)}||}{||x^{(v)}||} \]
(at which the iteration is to be terminated) are given, the procedure is then simplified to calculate terms of the sequence \( \{ \beta_i \} \), \( i \in [1, \nu] \), which is given by (1.72) and is the best possible choice if only \( \nu \)-iterations are to be performed.

Consider the following linear stationary iteration of second degree known as 'second order Richardson's iterative scheme'
\[ \mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} + a(b-Pu)^{(n)} + \beta (u^{(n)} - u^{(n-1)}) \quad (1.78) \]
where parameters \( a, \beta \) remain constant throughout the iteration and are chosen to provide maximum convergence to the solution. The error vector for (1.78) satisfies
\[ e_i^{(n+1)} = [(1+\beta)I-aP]e_i^{(n)} - \beta e_i^{(n-1)} \quad (1.79) \]
thus for each error mode we have
\[ e_i^{(n+1)} = [1+\beta-a\gamma_i]e_i^{(n)} - \beta e_i^{(n-1)} \quad (1.80) \]
Let \( \gamma_i \) be the eigenvalues of the matrix associated with the iterative process and so
\[ e_i^{(n+1)} = \gamma_i e_i^{(n)} = 2 \gamma_i e_i^{(n-1)} \quad (1.81) \]
From (1.80) and (1.81) we obtain
\[ e_i^{(n+1)} = [1+\beta-a\gamma_i - \frac{\beta}{\gamma_i}] e_i^{(n)} \quad (1.82) \]
and combining (1.81) and (1.82) yields
\[ y_i^2 - (1 + \beta - a_i) y_i + \beta = 0, \quad (1.83) \]

which yields
\[ y_i = \frac{\delta_i \pm \sqrt{\delta_i^2 - 4\beta}}{2} \quad (1.84) \]

where \( \delta_i = 1 + \beta - a_i \).

Choosing the parameters \( a, \beta \) so that \( (\delta_i^2 - 4\beta) < 0 \) for all \( u_i \) (which implies that \( y_i \) will be complex) all \( |y_i| \) will be identical. Then the choices
\[ 1 + \beta - a \cdot m = 2\sqrt{\beta} \quad (1.85) \]
\[ 1 + \beta - a \cdot M = -2\sqrt{\beta} \quad (1.86) \]

lead to
\[ a = \left[ \frac{2}{\sqrt{a+M}} \right]^2, \quad \beta = \left[ \frac{\sqrt{M-M}}{\sqrt{M+M}} \right]^2. \quad (1.87) \]

The above choice of parameters which makes the square root of (1.84) zero or negative, yields,
\[ |y_i| = \sqrt{\beta} \quad (1.88) \]

The eigenvalues \( y_i \) are complex but all have the same absolute value and thus all the error modes are decreased at the same rate. Note that \( |y_i| \) is always smaller than the maximum of \( |\lambda_i|, |\lambda_N| \) and consequently the procedure is always more rapidly convergent than Simultaneous iterations. In the case that the estimates of the extreme eigenvalues \( m, M \) are not exact, then it is advisable to underestimate \( m \) and overestimate \( M \) assuming that \( y_i \) being complex and therefore all modes decaying at the same rate. From (1.78) it can be observed that to commence the iterative process two iterates are required and although one initial iterate is available, only \( |y_i| \) is known and the second initial iterate is unobtainable. Thus the convergence rate indicated by (1.88) is asymptotically only approached as \( i \to \infty \). Given that the spectral radius of the iteration matrix is \( \sqrt{\beta} \), then the rate of convergence is \( \frac{2}{\sqrt{M/m}} \). Note that for each iteration of the method at least two vectors need to be stored.
Another non-stationary degree iteration, less sensitive to round-off errors is the Chebychev second order method defined by

\[ u^{(n+1)} = u^{(n)} + a_n (u^{(n)} - P u^{(n)}) + \beta_n (u^{(n)} - u^{(n-1)}) , \]  

(1.89)

where \( a_n, \beta_n \) are varied with each iterative step.

If the parameters \( a_n \) and \( \beta_n \) have to be computed during the iteration, the technique proves to be inefficient especially for large values of \( N \). For predetermined values of \( a_n, \beta_n \) the computational work per iteration is not much more than that of the iteration for simultaneous displacement method and the storage requirements are no greater also. The values of parameters \( a_n, \beta_n \) are given in [29a] as follows

\[ a_n = \frac{4T}{(M-m)T + (M+m)} \] \[ \beta_n = \frac{T}{M-m} \]  

(1.90)

The error vector is given by

\[ \phi^{(n+1)} = \frac{T_{n+1} (M+m)}{n} \phi^{(n)} \] \[ \phi^{(0)} \]  

(1.91)

which leads to identical results to those obtained using Richardson's method for rate of convergence.

Since \( a_n, \beta_n \) are less than unity round-off errors do not appear and the method is preferred to Richardson's method with the only problem being the computation of the extreme eigenvalues \( m, M \).

Finally for two dimensional model problem it can be shown [15], (p.226) that the simultaneous displacement method of (1.36) is equivalent to the Jacobi method of (1.35).

The processes defined thus far belong to the class of point iterative methods in which each component of \( u^{(n)} \) is expressed 'explicitly' i.e., can be expressed by itself using already computed approximate values of the other unknowns. Grouping the equations of the original system according to a predetermined rule we can obtain the block iterative methods. However, since we are concerned in this thesis with point
iterative methods further discussion of block methods would be discontinued.

Successive Displacement Methods

Assuming that we wish to obtain an iterative solution of the system

\[(1.18)\]

de fined by

\[D_{u}^{(n+1)} = D_{u}^{(n+1)} + \tilde{U}_{u}^{(n)} + b, \quad (1.93)\]

giving

\[(D-L)_{u}^{(n+1)} = D_{u}^{(n)} + b, \quad (1.94)\]

so that

\[u^{(n+1)} = (D-L)^{-1}D_{u}^{(n+1)} + (D-L)^{-1}b, \quad (1.95)\]

\[(I-\omega D)^{-1}L_{u}^{-1}b, \quad (1.96)\]

which shows that the SOR point iteration matrix is given by

\[L_{w} = (I-\omega D)^{-1}L_{u}^{-1}D_{u}^{-1}b, \quad (1.97)\]

when \(w=1\) (i.e. \(L_{1}\)) we obtain the Gauss-Seidel method.

The Optimum Acceleration Parameter for the SOR Method

An essential problem associated with the SOR method is the determination of a suitable value for the acceleration parameter \(w\). Ideally, we would like to have the optimum value of \(w\) denoted by \(w_{b}\) which minimises the spectral radius of the SOR iteration matrix and thereby maximises the rate of convergence of the method. At the present time, no formula exists for the determination of \(w_{b}\) for an arbitrary set of linear equations. But, it can be calculated for many of the difference equations because their matrices are of a special type which possesses property \(A [38]\). It was established in \([38]\) that when a matrix possesses property \(A\) then it can be transformed into what is termed a 'consistently ordered' matrix. Under this condition the eigenvalues \(\lambda\) of the point SOR iteration matrix \(L_{w}\)
associated with \( P \) are related to the eigenvalues \( \mu \) of the corresponding point Jacobi iteration matrix \( B \) of \( P \) by the equation

\[
(\lambda + \omega - 1)^2 = \lambda \omega \mu^2
\]  

(1.98)

From (1.98) we have that

\[
\lambda = \frac{\omega + \sqrt{\omega^2 - 4(\omega - 1)}}{2}
\]  

(1.99)

It can be shown [38], that the rate of convergence is dependent on \( \lambda \) and so to optimise the rate of convergence, \( \lambda \), the eigenvalue of maximum modulus of the SOR iteration matrix \( L \) must be minimised. This is achieved by making the square root equation (1.99) equal to zero for \( \mu \), the spectral radius of the point Jacobi iteration matrix \( B \), i.e.,

\[
\omega^2 = 4(\omega - 1)
\]  

(1.100)

which yields

\[
\omega_b = \frac{2}{1 + \sqrt{1 - \rho(G)}}
\]  

(1.101)

Now, since the Gauss-Seidel method is the same as SOR with \( \omega = 1 \), then from equation (1.99) we have,

\[
\rho(G) = \rho(J)^2
\]  

(1.102)

where \( \rho(G) \) and \( \rho(J) \) are the spectral radii of Gauss-Seidel and point Jacobi iteration matrices respectively. Using (1.102) it is clear from (1.101) that

\[
\omega_b = \frac{2}{1 + \sqrt{1 - \rho(G)}}
\]  

(1.103)

The estimation of \( \omega_b \) depends on whether \( \rho(J) \) or \( \rho(G) \) can be estimated. Several methods have been suggested [30] one of which is the power method which is now described.

Assuming the matrix of the finite difference equations is consistently ordered and has property (A), calculate the sequence of approximations \( u_1^{(1)}, u_2^{(2)}, \ldots, u^{(1)} \) to the solution of the system of equations \( Pu = b \) by the Gauss-Seidel method and then we have

\[
\rho(G) = \lim_{n \to \infty} \frac{||u^{(1)}||}{||u^{(1-1)}||}
\]  

(1.104)
where \( d^{(1)} \) is defined as
\[
d^{(1)} = u^{(1)} - u^{(1-1)},
\]
and
\[
||d^{(1)}|| = \left\{ \sum_{j=1}^{n} (u_j^{(1)} - u_j^{(1-1)})^2 \right\}^{1/2}.
\]

Thus, using the power method we can approximate \( \rho(G) \), which, in turn can be substituted into equation (1.103) to give an estimate of \( \omega_b \), the optimum acceleration factor.

Consistent Ordering

The concept of consistent ordering is central to the theory of the SOR iterative method for solving the equation \( Pu = b \), because the calculation of optimum parameters is currently only possible for consistently ordered matrices. The definitions for consistent ordering with respect to vector and matrix iteration sequences as well as the definition for property A are given in Chapter Two.

Assuming the concepts of property (A) and consistently ordered matrices, the following results can be established \([38]\).

**Theorem 1.2**

\( \mu \) is an eigenvalue of the Jacobi iteration matrix \( B \) if,

1. The coefficient matrix \( P \) is consistently ordered,
2. The acceleration parameter \( \omega \) is in the interval \((0, 2)\),
3. \( \mu \) satisfies the relation
\[
(\lambda + \omega - 1)^2 = \omega \lambda \mu^2,
\]

where \( \lambda \) is a non-zero eigenvalue of the SOR iteration matrix \( L_\omega \).

Conversely, if \( \mu \) is an eigenvalue of \( B \) and if \( \lambda \) satisfies (1.107), then \( \lambda \) is an eigenvalue of \( L_\omega \).

**Theorem 1.3**

If \( P \) is symmetric, positive definite (defined later) and is consistently ordered and if
\[ \omega_b = \frac{2}{1 + \sqrt{1 - \bar{\mu}^2}} , \]  

(1.108)

where \( \bar{\mu} \) is the spectral radius of the Jacobi iteration matrix \( B \) (i.e. \( \bar{\mu} = \rho(B) \)). Then,

\[ \rho(L_{\omega_b}) < \rho(L_{\omega}) , \quad \omega \neq \omega_b , \]  

(1.109)

and

\[ \rho(L_{\omega_b}) = \omega - 1 \text{ for } \omega_b < \omega < 2 . \]  

(1.110)
1.6 VARIATIONAL METHOD

In this section we shall briefly consider the conjugate gradient (CG) method introduced in [42] and [29a] as an iterative method for solving large linear systems [28a].

Specifically, we shall consider the CG method as an acceleration procedure analogous to the SI method with respect to a given iterative method.

Consider the linear system
\[ Au = b , \]  
(1.111)
where \( A \) is an \((n\times n)\) symmetric and positive definite matrix. The quadratic functional related to the system (1.111) is given by,
\[ Q(u) = \frac{1}{2}(u,Au) - (u,b) = \text{const.} \]  
(1.112)
This functional defines a family of ellipsoids in the Euclidean \( N \)-dimensional space, whose common centre is \( A^{-1}b \), the point at which \( Q(u) \) takes its minimum value. For any arbitrary vector \( u^{(n)} \), the residue \( r^{(n)} \) is given by
\[ r^{(n)} = b - Au^{(n)} = -[\nabla Q(u)]^*u^{(n)} , \]  
(1.113)
and it is always normal to the surface of the ellipsoid defined by (1.112). We shall thus attempt to proceed to the solution of \( A^{-1}b \), the centre point of the ellipsoids, by a sequence of vector displacements of the form,
\[ u^{(n+1)} = u^{(n)} + \varepsilon^*_nP^{(n)} , \]  
(1.114)
where \( P^{(n)} \) is an arbitrary direction and \( \varepsilon^*_n \) is an arbitrary constant.

The problem now is to determine \( \varepsilon^*_n \) such that the quadratic function \( Q(u^{(n+1)}) \) will be minimum for a given direction \( P^{(n)} \).

Using (1.112) and (1.114) we have
\[ Q(u^{(n+1)}) = \frac{1}{2}(u^{(n)} + \varepsilon_n^*P^{(n)} , A(u^{(n)} + \varepsilon_n^*P^{(n)} )) - ((u^{(n)} + \varepsilon_n^*P^{(n)} , b) , \]  
(1.115)
* \([\nabla Q(u)]^*u^{(n)}\) represents a vector with components \( \frac{\partial Q(u^{(n)})}{\partial u^i} \), \( i = 1, 2, \ldots, n \).
therefore
\[
\frac{\partial Q(u^{(n+1)})}{\partial \epsilon_n} = (P^{(n)} \cdot Au^{(n)} + \epsilon_n P^{(n)}) - (P^{(n)} \cdot b)
\]
\[
= -(P^{(n)} \cdot r^{(n)}) + (\epsilon_n P^{(n)} \cdot Ap^{(n)}) .
\]  
(1.116)

The optimum value of $\epsilon_n$ is obtained by setting the expression (1.116), equal to zero, which gives

\[
\epsilon_n = \frac{(P^{(n)} \cdot r^{(n)})}{(P^{(n)} \cdot Ap^{(n)})}.
\]  
(1.117)

If we use the definition of $u^{(n+1)}$ from (1.114) and the value of $\epsilon_n$ in (1.117), we have

\[
(P^{(n)} \cdot r^{(n+1)}) = (P^{(n)} \cdot (b - Au^{(n+1)})) = (P^{(n)} \cdot (r^{(n)} - \epsilon_n Ap^{(n)}))
\]
\[
= 0 ,
\]  
(1.118)

which implies that the direction $P^{(n)}$ and the residual $r^{(n+1)}$ are orthogonal.

If we wish to choose $P^{(n)}$ to lie along the line of steepest descent, we simply take

\[
P^{(n)} = r^{(n)}
\]

and by (1.114) and (1.117) we determine the known steepest descent method which results in a very slow convergence in many cases.

A better strategy for choosing the direction $P^{(n)}$ is based on the knowledge that the centre of the ellipsoid lies in the plane conjugate to a given chord. Thus, if we choose vectors $P^{(0)}, P^{(1)}, ..., P^{(n-1)}$ to be pairwise conjugate in the sense that

\[
(P^{(i)} \cdot Ap^{(j)}) = 0 ,
\]  
(1.119)

for $i \neq j$, then by determining $P^{(n+1)}$ by

\[
P^{(n)} = r^{(n)} + \alpha_{n-1} P^{(n-1)} ,
\]  
(1.120)

we can combine (1.120) and (1.119) to obtain

\[
(P^{(n)} \cdot Ap^{(n-1)}) = (r^{(n)} \cdot Ap^{(n-1)}) + (\alpha_{n-1} P^{(n-1)} \cdot Ap^{(n-1)}) = 0
\]  
(1.121)

and finally,

\[
\alpha_{n-1} = \frac{(r^{(n)} \cdot Ap^{(n-1)})}{(P^{(n-1)} \cdot Ap^{(n-1)})}.
\]  
(1.122)
This choice of $p^{(n)}$ results in the Conjugate Gradient iterative scheme which is defined as follows:

$$u^{(n+1)} = u^{(n)} + \varepsilon P^{(n)}, \quad n=0,1,2,\ldots,m-1,$$

$$r^{(n)} = b - Au^{(n)}, \quad n=0,1,2,\ldots,m,$$

$$P^{(n)} = \begin{cases} 0, & n=0 \\ r^{(n)} - \alpha^{(n-1)} P^{(n-1)}, & n=1,2,\ldots,m-1 \end{cases}$$

$$\alpha^{n-1} = \begin{cases} 0, & n=0 \\ \frac{-r^{(n)} P^{(n-1)}}{(r^{(n-1)} P^{(n-1)}), n=1,2,\ldots,m-1} \end{cases}$$

where $m$ is the smallest integer such that

$$r^{(m)} = 0.$$

From (1.127) we can conclude that the CG iterative scheme converges in at most $n$ iterations where $n$ is the order of the matrix $A$. Although the CG method theoretically gives an exact answer in $n$-steps, this in practice is precluded because round-off errors may seriously affect the orthogonality of the residuals.

We summarize below some basic properties of the CG method [3a].

$$(r^{(i)}, r^{(j)}) = 0, \quad i\neq j, \quad i,j=0,1,\ldots,m-1,$$

$$(P^{(i)}, A P^{(i)}) = 0, \quad i\neq j, \quad i,j=0,1,\ldots,m-1,$$

$$(P^{(i)}, \neq 0, \quad i=0,1,\ldots,m-1$$

and

$$\alpha^{n-1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})}, \quad n=1,2,\ldots,m-1.$$

From (1.127) we can conclude that the CG iterative scheme converges in at most $n$ iterations where $n$ is the order of the matrix $A$. Although the CG method theoretically gives an exact answer in $n$-steps, this in practice is precluded because round-off errors may seriously affect the orthogonality of the residuals.

In the recent past a number of modifications and improvements have been made to the CG method [28b],[8a],[28a],[1] and Evans [13a]. One important modification has been the formulation of the method (i.e. determination of $u^{(n+1)}$) in terms of $u^{(n)}$ and $u^{(n-1)}$. By replacing $n$ by $n-1$ in (1.123) we have
\[ u^{(n)} = u^{(n-1)} + \varepsilon_{n-1} p^{(n-1)} \]  
(1.133)

or

\[ \frac{a_{n-1}}{\varepsilon_{n-1}} u^{(n)} = \frac{a_{n-1}}{\varepsilon_{n-1}} u^{(n-1)} + \varepsilon_{n-1} p^{(n-1)} \]  
(1.134)

which by eliminating \( p^{(n-1)} \) using (1.125) becomes

\[ \frac{a_{n-1}}{\varepsilon_{n-1}} u^{(n)} = \frac{a_{n-1}}{\varepsilon_{n-1}} u^{(n-1)} + \varepsilon_{n-1} (p^{(n)} - r^{(n)}) , \]  
(1.135)

and finally eliminating \( p^{(n)} \) by (1.123) we have

\[ u^{(n+1)} = (1 + \frac{\varepsilon_{n}}{\varepsilon_{n-1}} a_{n-1})u^{(n)} - \frac{\varepsilon_{n}}{\varepsilon_{n-1}} a_{n-1} u^{(n-1)} + \varepsilon_{n} r^{(n)} \]  
(1.136)

which can be expressed in the compact form

\[ u^{(n+1)} = \rho_{n+1} (u^{(n)} + \gamma_{n+1} r^{(n)}) + \gamma_{n+1} \rho_{n+1} u^{(n-1)} , \]  
(1.137)

where

\[ \rho_{n+1} = 1 + \frac{\varepsilon_{n}}{\varepsilon_{n-1}} a_{n-1} , \]  
(1.138)

and

\[ \gamma_{n+1} = \frac{\varepsilon_{n}}{\rho_{n+1}} . \]  
(1.139)

We can now simplify the expressions for \( \rho_{n+1} \) and \( \gamma_{n+1} \) by expressing them in terms of certain 'inner products.'

We express (1.137) in terms of certain residuals by using (1.124) hence we have that

\[ r^{(n+1)} = \rho_{n+1} (r^{(n)} - \gamma_{n+1} Ar^{(n)}) + (1 - \rho_{n+1}) r^{(n-1)} . \]  
(1.140)

If we now take the inner products of both sides of (1.140) with \( r^{(n)} \), then by (1.128) we have

\[ 0 = \rho_{n+1} ((r^{(n)}, r^{(n)}) - \gamma_{n+1} (r^{(n)}, Ar^{(n)})) , \]  
(1.141)

and since \( \rho_{n+1} \neq 0 \) we obtain

\[ \gamma_{n+1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n)}, Ar^{(n)})} . \]  
(1.142)

On the other hand, if we take the inner product of both sides of (1.140) with \( r^{(n-1)} \), then,

\[ 0 = \rho_{n+1} (\gamma_{n+1} (r^{(n-1)}, Ar^{(n)})) + (1 - \rho_{n+1}) (r^{(n-1)}, r^{(n-1)}) , \]  
(1.143)
\[ \rho_{n+1} = \left[ 1 + \frac{(r^{(n-1)}, Ar^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \gamma_{n+1} \right]^{-1}. \] (1.144)

Also, by replacing \( n \) by \( n-1 \) in (1.140), we have

\[ r^{(n)} = \rho_n (r^{(n-1)} - \gamma Ar^{(n-1)}) + (1-\rho_n) r^{(n-2)} \]

and if we take inner product of both sides with \( r^{(n)} \) we obtain

\[ (r^{n-1}, Ar^{(n)}) = \frac{(r^{(n)}, r^{(n)})}{\gamma_n \rho_n} \]

thus (1.144) becomes

\[ \rho_{n+1} = \left[ 1 - \frac{\gamma_{n+1}}{\gamma_n} \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \frac{1}{\rho_n} \right]^{-1}. \] (1.145)

Summarising our results the CG method can also be defined as

\[ u^{(n+1)} = \rho_{n+1} (u^{(n)} + \gamma^{n+1} r^{(n)}) + (1-\rho_{n+1}) u^{(n-1)}, \] (1.146)

where \( \rho_1 = 1 \),

\[ \rho_{n+1} = \left[ 1 - \frac{\gamma_{n+1}}{\gamma_n} \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \frac{1}{\rho_n} \right]^{-1}, \quad n=1,2,\ldots \] (1.147)

and

\[ \gamma_{n+1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n)}, Ar^{(n)})}. \] (1.148)

From (1.146) we observe that the CG method has the same form as the SI method (and the second degree method) the difference in this case is that the parameters are variables (whereas in the SI method \( \gamma_1 = \gamma_2 = \ldots = \gamma_\rho \) chosen to minimise the quadratic function \( Q(u) \).

In fact, we expect the CG method to produce a better rate of convergence than with the application of the SI techniques since in the former we have only one additional parameter \( \gamma_{n+1} \) which is a variable instead of being a constant (SI method). In comparing the CG method with the SI method we note that the former requires more computations per iteration, but, on the other hand, it does not require the estimation of the largest and smallest eigenvalues of matrix \( A \). Moreover, it can be proved Young [38a] that for all \( n \) we have

\[ \| u^{(n)} - u^{\ast} \|_A \leq \| u^{(n)} - u \|_A. \] (1.149)
where \( u \) is the exact solution of (1.111), \( \tilde{u}^{(n)} \) the approximate solution obtained by the CG method and \( \hat{u}^{(n)} \) is the approximate solution obtained by the SI method with respect to the basic iterative scheme (1.111).

The relation defined by (1.149) indicates the essential advantage of the CG method over the SI method because with the latter, only the upper and lower bounds for the eigenvalues of the coefficient matrix are used whereas, the former takes advantage of the distribution of the eigenvalues of \( \hat{e} \).

Finally, we observe that the relationship given in (1.149) shows that the CG method is better, in the sense of minimising the \( \lambda^2 \)-norm of error vector than any linear non-stationary second degree method. Since we can obtain estimates for the convergence rate of the SI methods we thus obtain a lower bound on the rapidity of convergence of the CG method. Consequently from (1.149) and from the fact that in the SI method we have

\[
\|u^{(n)} - \bar{u}\|_\lambda \leq \frac{2r^{n/2}}{1+r^n} \|u^{(0)} - \bar{u}\|_\lambda ,
\]

we immediately have

\[
\|\tilde{u}^{(n)} - \bar{u}\|_\lambda \leq \frac{2r^{n/2}}{1+r^n} \|\tilde{u}^{(0)} - \bar{u}\|_\lambda ,
\]

by assuming \( \tilde{u}^{(0)} = u^{(0)} \), where \( r \) is defined by

\[
r^\frac{1}{2} = \frac{1}{a \sqrt{1+\sigma}} \quad \text{and} \quad \sigma = \frac{1}{2} .
\]
1.7 EVOLUTION OF SOLUTION TECHNIQUES FOR ELLIPTIC P.D.E.'S

The coefficient matrices that often result from the discretization of partial differential systems have certain properties which tend to dictate the choice of an appropriate solution technique. The trend in the evolution of such techniques have been influenced primarily by considerations which aim at improving or meeting the limitations of the existing methods or by sheer desire for novelty until recently. Current developments in computer architecture have led to development of computers with multiple processors (SIMD, MIMD, etc. machines) capable of executing several discrete instruction sets in parallel. Increasing interest has developed in examining the adequacy of algorithms for these parallel architectures [23],[25]. The emergence of parallel computers is relevant especially with regard to the solution of partial differential equations where discretization of some complex design problems could give rise to very large linear systems requiring hours of solution on a sequential computer. For the first time the need arises either to evolve techniques which can adapt the large body of scalar algorithms available to the new machines where possible or to evolve new algorithms whose structure could easily permit parallel processing. The explicit preconditioned methods described in this thesis are directed to the latter need.
CHAPTER TWO

MATRIX ANALYSIS
2.1 PRELIMINARIES OF MATRIX ANALYSIS

This chapter provides some of the basics of theorems, definitions and assumptions that have been used in the subsequent chapters. These theorems and definitions are in some cases established by the use of matrix theory.

The matrix $B$ is said to be transpose of a matrix $A$ if

$$b_{i,j} = a_{j,i}, \text{ for all } 1 \leq i, j \leq n.$$  \hspace{1cm} (2.1.1)

The matrix $B$ is usually denoted by $A^T$. If the matrix $A$ is symmetric (i.e. $a_{i,j} = a_{j,i}$ for all $i,j$), then,

$$A^T = A.$$ \hspace{1cm} (2.1.2)

The matrix $A$ is said to be orthogonal if

$$A^T A = I$$ \hspace{1cm} (2.1.3)

The determinant of a matrix $A$ will be denoted either by $\det(A)$ or $|A|$. A matrix $A$ is non-singular if and only if $\det(A) \neq 0$.

The conjugate of matrix $A=(a_{i,j})$ is denoted by

$$\overline{A} = (\overline{a}_{i,j}).$$ \hspace{1cm} (2.1.4)

The Hermitian transpose (or conjugate transpose) of a matrix $A$ will be denoted by $A^H$, note that

$$A^H = \overline{A}^T = A^T = (\overline{a}_{j,i}).$$ \hspace{1cm} (2.1.5)

The matrix $A=(a_{i,j})$ is said to be Hermitian matrix if

$$A^H = A.$$ \hspace{1cm} (2.1.6)

The trace of a matrix $A=(a_{i,j})$ is given by

$$\text{Trace } A = \sum_{i=1}^{n} a_{i,i}.$$ \hspace{1cm} (2.1.7)

A permutation matrix $P=(p_{i,j})$ is a matrix with exactly one non-zero element, namely unity in each row and each column. Thus for example

$$P = \begin{bmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{bmatrix}.$$ \hspace{1cm} (2.1.8)
is a permutation matrix of order 4. For any permutation matrix \( P \) we have
\[
p \cdot p^T = p^T \cdot p = I
\] (2.1.9)
\[
p^T = p^{-1}
\] (2.1.10)

**Gerschgorin's Theorem**

If \( \lambda \) is any eigenvalue of \( A \) then
\[
|a_{i,i} - \lambda| \leq \sum_{j \neq i} |a_{i,j}|
\] (2.1.11)
for at least one \( i \) (i.e. every eigenvalue of \( A \) lies in at least one of the
disks with centres \( a_{i,i} \) and radii \( \sum_{j \neq i} |a_{i,j}| \)).

**Proof**

Let \( Ax = \lambda x \) where \( x = [x_1, \ldots, x_n] \) then \( (A - \lambda I)x = 0 \) i.e. \( (a_{i,i} - \lambda)x_i + \sum_{j \neq i} a_{i,j}x_j = 0 \)

thus
\[
|a_{i,i} - \lambda| \leq \sum_{j \neq i} |a_{i,j}| \frac{|x_j|}{|x_i|}
\]
Choosing \( i \) such that \( x_i \) is the element of the largest modulus gives the
result that
\[
|a_{i,i} - \lambda| \leq \sum_{j=1}^{n} |a_{i,j}|
\]
for at least one \( i \).

**Theorem 2.1**

If \( A \) is a Hermitian matrix and if \( \xi_1 \) and \( \xi_N \) are respectively the
largest and smallest eigenvalues of \( A \), then
\[
\xi_1 = \max_{v \neq 0} \frac{(v,Av)}{(v,v)} = \frac{v^{(1)} , Av^{(1)}}{(v^{(1)}, v^{(1)})}
\] (2.1)
\[
\xi_N = \min_{v \neq 0} \frac{(v,Av)}{(v,v)} = \frac{v^{(N)} , Av^{(N)}}{(v^{(N)}, v^{(N)})}
\] (2.2)
where \( v^{(1)} \) and \( v^{(N)} \) are respectively, the eigenvectors of \( A \) corresponding
to \( \xi_1 \) and \( \xi_N \).
Definition 2.1

If A is a \((n \times n)\) matrix, then the spectral radius of \(A\) is given by

\[
S(A) = \max_{\lambda \in \sigma_A} |\lambda| \tag{2.3}
\]

where \(\sigma_A\) represents the eigenvalue spectrum of \(A\).

Using the Perron-Frobenius theory of non-negative matrices, we have the following two theorems.

Theorem 2.2

If \(A \geq 0\), then \(S(A)\) is an eigenvalue of \(A\) and there is an associated eigenvector with \(S(A)\).

Theorem 2.3

For any \((n \times n)\) matrix with eigenvalues \(\xi_1, \xi_2, \ldots, \xi_n\), the trace and determinant are defined by

\[
\text{trace}(A) = \sum_{i=1}^{n} \xi_i, \tag{2.5}
\]
\[
\text{det}(A) = \prod_{i=1}^{n} \xi_i. \tag{2.6}
\]

Theorem 2.4

A given matrix \(A\) is positive definite if and only if it is Hermitian and has only positive eigenvalues.

Theorem 2.5

If the matrix \(A\) is non-singular then \(A^H A\) is Hermitian and positive definite.

Theorem 2.6

If \(A\) is positive definite (non-negative definite) matrix, then there exists a unique positive definite (non-negative definite) matrix \(B\) (denoted by \(A^+)\) such that

\[
B^2 = A.
\]
Definition 2.2

If there exists a non-singular matrix such that

$$S^H A S = B,$$

then we say that $B$ is Hermitian congruent to $A$ and $B$ is obtained from $A$ by Hermitian congruent transformation.

Theorem 2.7

If $A$ is positive definite and $B$ is obtained from $A$ by congruent transformation, then $B$ is also positive definite. Similarly for non-negative and non-positive matrices.

Proof

From Definition 2.2, we have

$$B = S^H A S,$$

also by Theorem 2.6, we let $A^{1/2}$ be positive definite matrix whose square is $A$. Thus

$$B = (S^{1/2} A^{1/2})(S^{1/2} A^{1/2})^H,$$

and since $A^{1/2}$ is non-singular, then from Theorem 2.6, we have that $B$ is positive definite.

Definition 2.3

The matrix $A$ is positive definite if $A$ is Hermitian and $(v, Av) > 0$ for all $v \neq 0$. If $(v, Av) \geq 0$ for all $v$, then $A$ is non-negative definite.
2.2 \textbf{MATRIX AND VECTOR NORMS}

In the following section, we give a few relevant vector and matrix norms and the relation between them.

Let

\[
x = \begin{bmatrix}
x_1 \\
\vdots \\
x_n
\end{bmatrix}
\]

denote a column vector. A 'norm' of vector $x$ is a real number $||x||$ satisfying the axioms:

(i) $||x|| \geq 0$ \hspace{1cm} (2.2.1)

(ii) $||x|| = 0$, if and only if $x = 0$ \hspace{1cm} (2.2.2)

(iii) $||\alpha x|| = |\alpha|||x||$, for any scalar $\alpha$. \hspace{1cm} (2.2.3)

(iv) $||x+y|| \leq ||x|| + ||y||$. \hspace{1cm} (2.2.4)

\textbf{Definition 2.4}

The most commonly used vector norms are defined by

\[
||x||_1 = \sum_{i=1}^{n} |x_i| \hspace{1cm} (1\text{-norm}) \hspace{1cm} (2.2.7)
\]

\[
||x||_2 = \left( \sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \hspace{1cm} (\text{Euclidean vector norm}) \hspace{1cm} (2.2.8)
\]

\[
||x||_\infty = \max_{1 \leq i \leq n} |x_i| \hspace{1cm} (\infty\text{-norm}) \hspace{1cm} (2.2.9)
\]

(2.2.5)-(2.2.7) are special cases of

\[
||x||_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \hspace{1cm} (p\text{-norm}) \hspace{1cm} p \geq 1 . \hspace{1cm} (2.10)
\]

A norm of an $(n \times n)$ matrix $A$ is a real number $||A||$ satisfying the axioms:

(i) $||A|| \geq 0$, \hspace{1cm} (2.2.5)

(ii) $||A|| = 0$, if and only if $A = (0)$. \hspace{1cm} (2.2.6)

(iii) $||\alpha A|| = |\alpha|||A||$, for any scalar $\alpha$, \hspace{1cm} (2.2.7)

(iv) $||A+B|| \leq ||A|| + ||B||$. \hspace{1cm} (2.2.8)

(v) $||A*B|| \leq ||A|| \cdot ||B||$. \hspace{1cm} (2.2.9)
Definition 2.5

Some matrix norms as defined below:

\[ \|A\|_1 = \max_{j} \sum_{j=1}^{n} |a_{ij}|, \quad j=1, \ldots, N. \text{ (row sum-norm)} \]

\[ \|A\|_2 = (S(A^H A))^{1/2} \quad \text{(Euclidean-norm)} \]

\[ \|A\|_m = \max_{i,j} |a_{ij}|, \quad i,j=1,2, \ldots, N. \text{ (maximum-norm)} \]

\[ \|A\|_{\infty} = \max_{i} \sum_{j=1}^{n} |a_{ij}|, \quad i=1,2, \ldots, N. \text{ (column sum-norm)} \]

(2.11)

Definition 2.6

If \( \| \cdot \|_\infty \) is a vector norm, then the induced norm \( \| \cdot \|_{\gamma(\beta)} \) is defined by

\[ \|A\|_{\gamma(\beta)} = \max_{v \neq 0} \frac{\|Av\|_\beta}{\|v\|_\beta} \quad \text{(2.12)} \]

For this norm, the following inequality is valid

\[ \|Av\|_\beta \leq \|A\|_\gamma \|v\|_\beta \quad \text{(2.13)} \]

Definition 2.7

Whenever (2.13) is valid, then the norms \( \| \cdot \|_\gamma \) and \( \| \cdot \|_\beta \) are said to be compatible or consistent.

Also, if the matrix norm \( \| \cdot \|_\beta \) and vector norm are compatible for some \( v \neq 0 \) on space \( \mathbb{C}^N \) and \( v \neq 0 \), then

\[ \|Av\|_\gamma = \|A\|_\beta \|v\|_\gamma \]

and the matrix norm is therefore subordinate to the vector norm.

Theorem 2.7

If \( \| \cdot \|_\gamma \) is a matrix norm then

\[ S(A) \leq \|A\|_\gamma . \quad \text{(2.14)} \]
2.3 **CONDITIONS FOR CONVERGENCE**

**Definition 2.8**

If $Q$ is an arbitrary $(n \times n)$ matrix, then $Q$ is convergent to the null matrix $0$ if the matrix sequences

$$Q, Q^2, Q^3, \ldots, Q^n$$

converges to the null matrix $0$. Otherwise, $Q$ is divergent.

**Theorem 2.8**

Given that $C$ is any matrix then $C$ converges if and only if

$$S(C) < 1.$$ 

**Proof**

Assume that

$$J_1 = \begin{pmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 1 \\ & & & \lambda_1 \end{pmatrix}$$

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

$$\text{diag}(QCQ^{-1}) = J_1, \quad l \leq r,$$

where $\lambda_1$ ($l \leq r$) are distinct eigenvalues of $A$. If each Jordan matrix is raised to the power $n$, then the result tends to the null matrix as $n \to \infty$, if and only if $S(C) < 1$. 


2.4 **Irreducibility and Weak Diagonal Dominance**

**Theorem 2.9**

The matrix $A$ is irreducible if and only if there exists a permutation matrix $P$ such that $P^{-1}AP$ has the form

$$P^{-1}AP = \begin{bmatrix} Q & \phi \\ D & F \end{bmatrix}$$

where $Q$ and $F$ are square matrices and where all elements of $\phi$ vanish.

**Theorem 2.10**

An $(n \times n)$ matrix $A$ has weak diagonal dominance if

$$|a_{i,j}| \leq \sum_{j=1 \atop j \neq i}^{n} |a_{i,j}|, \quad i=1,2,\ldots,N,$$

and for at least one $i$

$$|a_{i,i}| > \sum_{j=1 \atop j \neq i}^{n} |a_{i,j}|.$$

**Theorem 2.11**

The matrix $A$ has weak diagonal dominance and is irreducible if $\det(A) \neq 0$ and not one of the diagonal elements vanish.

**Theorem 2.12**

A matrix $B$ has strong diagonal dominance if for every $i$ we have

$$|b_{i,i}| > \sum_{j=1 \atop j \neq i}^{n} |b_{i,j}|.$$

**Theorem 2.13**

When the matrix $B$ has strong diagonal dominance, then $\det(B) \neq 0$. 
2.5 **CONSISTENTLY ORDERED MATRICES AND ORDERING VECTORS**

**Definition 2.9**

The vector $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_n)^T$, where $\gamma_1, \gamma_2, \ldots, \gamma_n$ are integers, is an ordering vector for a matrix $A$ of order $n$ if for any pair of associated integers $i, j$ with $i \neq j$ the relation $|\gamma_i - \gamma_j|$ is valid.

**Definition 2.10**

An ordering vector $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_n)^T$ for a matrix $A$ of order $n$ is a compatible ordering vector for $A$ if

(i) $\gamma_i - \gamma_j = 1$ and $i$ and $j$ are associated and $i > j$.

(ii) $\gamma_i - \gamma_j = -1$ if $i$ and $j$ are associated and $i < j$.

**Definition 2.11**

A given matrix $A$ is consistently ordered if there exists an ordering vector $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_n)^T$ such that $a_{i,j} \neq 0$, if $i \neq j$. implies

(i) $|\gamma_i - \gamma_j|= 1$

(ii) $\begin{cases} i < j \rightarrow \gamma_i < \gamma_j \\ i > j \rightarrow \gamma_i > \gamma_j \end{cases}$

It follows that if the components of $\gamma$ are in ascending order of magnitude and $\gamma$ is an ordering vector for the matrix $A$ then $A$ is consistently ordered.

**Theorem 2.14**

A matrix of order $n$ is consistently ordered if there exists a compatible ordering vector.
2.6 MATRICES WHICH HAVE PROPERTY A

For the matrix $A$ to have property A, it must be a diagonal matrix or else there must exist a permutation matrix $Q$ such that $Q^{-1}AQ$ has the form

$$
\tilde{A} = Q^{-1}AQ = \begin{bmatrix}
R_1 & 0 \\
0 & R_2
\end{bmatrix}
$$

where $R_1$ and $R_2$ are square or diagonal matrices.

**Theorem 2.16**

A real matrix $A$ of order $n$ is an L-matrix if

$$a_{i,i} > 0, \quad i=1,2,\ldots,N$$

$$a_{i,j} \leq 0, \quad i \neq j, \quad i,j=1,2,\ldots,N.$$ 

**Definition 2.12**

A real matrix $A$ is an M-matrix if

$$a_{i,j} \leq 0, \quad i \neq j, \quad i,j=1,2,\ldots,N \text{ and } A^{-1} \succeq 0. \quad (2.16)$$

**Definition 2.13**

A real matrix $A$ is a Stieltjes matrix if (2.16) is valid and if it is positive definite.
Chapter Three

Direct Preconditioned Solution of Linear Systems of Equations
3.1 ERROR BOUNDS

We shall establish some error bounds by assuming three cases of perturbation of a general linear system.

Consider the linear system

$$Ax = c$$  \hspace{1cm} (3.1)

where $A$ is non-singular.

Case 1:

If we perturb the right hand side of (3.1), we obtain

$$A(x+\delta x) = c + \delta c$$

and

$$A\delta x = \delta c$$

$$\delta x = A^{-1}\delta c.$$  \hspace{1cm} (3.2)

We therefore have the inequality expression

$$||\delta x|| = ||A^{-1}\delta c|| ||A^{-1}|| ||\delta c||.$$  \hspace{1cm} (3.3)

From (3.2), we obtain the bound

$$||\delta x|| \leq ||A^{-1}|| ||\delta c||.$$

Case 2:

Suppose we now perturb $A$ in (3.1), then we have

$$(A+\delta A)(x+\delta x) = c,$$

$$(A+\delta A)\delta x = -\delta A.x$$

Since $A$ is non-singular we can write

$$(A+\delta A) = A(I+A^{-1}\delta A)$$

We have for every eigenvalue $\lambda_i$ of $A^{-1}\delta A$ that

$$||A^{-1}\delta A|| \geq \lambda_i.$$  \hspace{1cm} (3.4)

The eigenvalues of $I+A^{-1}\delta A$ are $1+\lambda_i$ so the matrix $I+A^{-1}\delta A$ is non-singular if

$$||A^{-1}\delta A|| < 1.$$  \hspace{1cm} (3.5)

Assuming that (3.4) is valid, then

$$\delta x = -(A+\delta A)^{-1}\delta A x$$

$$\delta x = -(I+A^{-1}\delta A)^{-1}A^{-1}\delta A x.$$  \hspace{1cm} (3.5)
Let 
\[ Q = (I + A^{-1} \delta A)^{-1}, \] (3.6)

Relation (3.6) implies that
\[ I = Q + A^{-1} \delta AQ \] (3.7)

Also (3.7) implies that
\[ \|I\| \geq \|Q\| - \|A^{-1} \delta A\| \|Q\|. \] (3.8)

and for every subordinate norm, (3.8) gives
\[ 1 \geq \|Q\| - \|A^{-1} \delta A\| \|Q\|. \] (3.9)

From (3.9), we have that
\[ \|Q\| \leq \frac{1}{1 - \|A^{-1} \delta A\|}. \] (3.10)

provided that \( \|A^{-1}\| \cdot \|\delta A\| < 1 \). We have from (3.5) that,
\[ \|\delta x\| = \|QA^{-1} \delta Ax\|, \]

and using (3.10), we obtain,
\[ \|\delta x\| \leq \frac{\|A^{-1} \delta A\| \cdot \|x\|}{1 - \|A^{-1} \delta A\|} \] (3.11)

From (3.11) we have the relative error bound
\[ \frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|A\| \cdot \left( \frac{\|\delta A\|}{\|A\|} / (1 - \|A\| \cdot \|A^{-1}\| \cdot \|\delta A\|) \right)}{\|A\|}. \] (3.12)

**Case 3:**

Next, we consider the simultaneous perturbation of both matrices \( A \) and \( C \).

\[(A + \delta A)(x + \delta x) = c + \delta c\]
\[(A + \delta A)\delta x = \delta c - \delta Ax.\]
\[A(I + A^{-1} \delta A)\delta x = \delta c - \delta Ax\]
\[\delta x = (I + A^{-1} \delta A)^{-1} A^{-1} (\delta c - \delta Ax)\] (3.13)

From (3.13), we have
\[ \|\delta x\| \leq \|I + A^{-1} \delta A\|^{-1} \cdot \|A^{-1}\| \cdot \|\delta c - \delta Ax\|. \] (3.14)

Let us also assume here that
\[ \|A^{-1}\| \cdot \|\delta A\| < 1. \]
and thus

\[ I + A^{-1} \delta A \]

is non-singular and for subordinate norms we have

\[ \frac{1}{1 + \| A^{-1} \delta A \|} \leq \frac{1}{1 - \| A^{-1} \delta A \|} \]

Using (3.14), we have

\[ \| \delta x \| \leq \frac{1}{1 - \| A^{-1} \|. \| \delta A \|} \cdot \| A^{-1} \|. (\| \delta c \| + \| \delta A \|. \| x \|) \] (3.15)

Using (3.15) we obtain the expression for the relative error

\[ \frac{\| \delta x \|}{\| x \|} \leq \frac{\| A \|. \| A^{-1} \| (\frac{\| \delta c \|}{\| A \|. \| x \|} + \frac{\| \delta A \|}{\| A \|})}{1 - \| A^{-1} \|. \| \delta A \|} \] (3.16)

Relations (3.12) and (3.16) have been established to be valid for subordinate norms. However by using

\[ \| A \|_2 \leq \| A \|_E \leq n^\| A \|_2 \]

they can be established for the Euclidean norm as well.
3.2 CONDITION NUMBERS

Suppose a small change in a parameter from $\beta$ to $\beta + \delta \beta$ causes a change of result of a solution vector $x$ to $x + \delta x$, then if $\delta \beta$ is small, we have

$$||\delta x|| = c||\delta x||$$

where $c$ is called the condition number for the absolute changes in $x$ caused by absolute changes in $\beta$.

If $\beta$ is small, we have

$$||\delta x|| = k||\delta \beta||$$

$k$ is the condition number for the relative changes in $x$ caused by relative changes in $\beta$.

We thus have that a condition number is a measure of ill-conditioning of a problem. If the condition number is large, the problem is badly ill-conditioned.

3.2.1 Condition Numbers for Large Linear Systems

We define the following condition numbers which have found much application in matrix analysis.

(a) P-condition number:

The P-condition number of an arbitrary non-singular matrix $A$ is defined by

$$P(A) = \frac{\lambda_1}{\lambda_N}$$

where $\lambda_1$ is the largest eigenvalue of largest modulus and $\lambda_N$ is the smallest eigenvalue of the least modulus of $A$. (where $A$ is symmetric.)

Let $R$ be the inverse of $A$ obtained by Gaussian elimination. Von Neumann and Goldstine (1947) established that if $A$ is symmetric and positive definite, then,

$$||u|| \leq 14.24 P(A)n^2 \epsilon$$  \hspace{1cm} (3.17)

where $u$ is the eigenvalue of the largest modulus of $I - AR$, $n$ is the order of matrix $A$ and $\epsilon$ is the smallest number recognised by the computer
in use (Loughborough University)

\[ c = 2^{-39} \].

When A is symmetric but not positive definite then (3.17) can be expressed as

\[ \|\mathbf{u}\| \leq 36.58 \, P(A^T A)^{1/2} \epsilon \] \hspace{1cm} (3.18)

But it is known that \( P(A^T A) \geq [P(A)]^2 \).

Thus, we have that matrices with large P-condition numbers could be expected to be difficult to invert and it is therefore reasonable to use P-condition number as a "measure" of ill-conditioning of a matrix even if it is not symmetric.

(b) N-condition number:

The N-condition number of a matrix A is defined by

\[ N(A) = \frac{\|A\| \cdot \|A^{-1}\|}{n} \]

where

\[ \|A\| = \left( \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij})^2 \right)^{1/2} \]

(c) M-condition number:

We define the M-condition number of a matrix A by

\[ M(A) = n \max(A) \cdot \max(A^{-1}) \]

where \( \max(A) = (\text{maximum element of matrix } A) \)

\( \max(A^{-1}) = (\text{maximum element of matrix } A^{-1}) \)

(d) K-condition number:

Using matrix norms \( \|\cdot\|_2 \) we define the K-condition number of a matrix A by

\[ K(A) = \|A\|_2 \cdot \|A^{-1}\| \]

where \( \|A\|_2 \) norm is defined in Section 2.2.

We have that the quantity

\[ \|A\| \cdot \|A^{-1}\| \]

is a reasonable measure of ill-condition.
It can be shown that the following relationships hold:

\[
\frac{M(A)}{n} \leq P(A) \leq nM(A)
\]

and

\[
\frac{M(A)}{n^2} \leq N(A) \leq M(A).
\]

It should be noted that the most realistic bound of error is

\[||A||_2 ||A^{-1}||_2\]

and so the K-condition number is a very good measure of ill-conditioning.

However, \(K(A) = ||A||_2 ||A^{-1}||_2\)

\[= (\text{max eigenvalue of } A^T A)^{-\frac{1}{2}}\]

\[= \left(\frac{\lambda_1}{\lambda_N}\right)^{-\frac{1}{2}}\]

where \(\lambda_1\) is the eigenvalue of the largest modulus of \(A^T A\) and \(\lambda_N\) is the eigenvalue of the least modulus of \(A^T A\). So when \(A\) has eigenvalues of very small modulus, it is difficult to calculate the smallest eigenvalue of \(A^T A\), although it is symmetric and positive definite. (When \(A\) is ill-conditioned sometimes \(A^T A\) is near singular).

Another measure of ill-conditioning is the following:

We divide the \(i^{th}\) row of matrix \(A\) by

\[\left(\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}}, \quad i=1,2,\ldots,n.
\]

to obtain a normalised form of the matrix. If the determinant of the normalised matrix is very small compared to \(\pm 1\), the system is ill-conditioned.
3.3 APPLICATION OF PRECONDITIONING METHOD FOR DIRECT SOLUTION OF LINEAR SYSTEMS OF EQUATIONS

Consider the linear system

\[ Ax = b, \quad (3.19) \]

where \( A \) is a non-singular \( n \times n \) matrix which is not necessarily sparse and \( b \) is given.

A general approach to the solution of the above system consists in finding two non-singular matrices \( P \) and \( Q \) such that (3.19) can be replaced by an equivalent system

or

\[ PAQy = d, \quad \text{or} \quad B = PAQ, \quad x = Qy. \quad (3.20) \]

where

The problem then is to find \( P \) and \( Q \) such that the equivalent system (3.20) has a smaller condition number than the original system (3.19).

Minimal Value for the P-Condition Number of the Preconditioned Matrix

We have assumed that the matrix \( A \) is non-singular and we now consider a splitting for this matrix of the form

\[ A = D - L - U \]

where \( D = \text{diag}(A) \) and \( L \) and \( U \) are lower and upper triangular matrices respectively whose entries are the negative of \( A \). We shall constrain the matrices \( P \) and \( Q \) to assume functions of the preconditioning parameter \( \omega \) and so the preconditioned matrix \( B \) of (3.20) would be necessarily a function of \( \omega \). We therefore require a value of this parameter for which the P-condition number of \( B \) is minimal.

3.3.1 The First Order Transformation (F.O.T.)

Consider the transformation defined by
\[ y = (D-\omega U)^{-1}x, \quad (3.21) \]

or, considering the first order expansion of (3.21)
\[ y = (D+\omega U)x, \quad (3.22) \]

If we premultiply (3.22) by \((D+\omega L)\), we obtain
\[ (D+\omega L)Ax = (D+\omega L)b. \quad (3.23) \]

Alternatively, (3.23) can be expressed as
\[ (D+\omega L)A(D+\omega U)(D+\omega U)^{-1}x = (D+\omega L)b \quad (3.24) \]

The expression (3.24) can now be written in the form
\[ B_\omega y = d, \quad (3.25) \]
where
\[ B_\omega = (D+\omega L)A(D+\omega U), \quad (3.26) \]
and
\[ d = (D+\omega L)b \quad (3.27) \]

The effect of the use of preconditioning on the accuracy of direct methods of solutions obtained for linear systems with different assumptions on the structure and properties of the original coefficient matrix \(A\) has been examined extensively in [22]. In [22] also, the relative merits of the use of preconditioning as opposed to an optimal scaling method for solving ill-conditioned linear systems was examined. Our basic aim in this and subsequent sections therefore would be to compare the relative minimizing effect on the \(P\)-condition number of the explicit transformation and that of its implicit variant [22] for any given system coefficient matrix.

We present next, the results obtained by using the first order transformation under certain properties of typical matrices. We shall categorise these matrices into the following cases:

A: Generally, very badly ill-conditioned matrices

B: Matrices with strong diagonal dominance

C: Matrices which have weak or lack diagonal dominance.

Class A: Generally, very badly ill-conditioned matrices

Consider
(i) the 4×4 segment of the Wilson matrix (A1) given by

\[
A = \begin{bmatrix}
5 & 7 & 6 & 5 \\
7 & 10 & 8 & 7 \\
6 & 8 & 10 & 9 \\
5 & 7 & 9 & 10
\end{bmatrix}
\]  \hspace{1cm} (A1)

This matrix is very badly ill-conditioned and has the condition numbers:

P-condition number \( P(A) = 2984 \)

N-condition number \( N(A) = 752 \)

The results obtained for this matrix are given in Table 3.1 and Figures 3.1 and 3.2.

(ii) the matrix \( A_2 \) \([19]\) is given by

\[
A = \begin{bmatrix}
10 & 1 & 4 & 0 \\
1 & 10 & 5 & -1 \\
4 & 5 & 10 & 7 \\
0 & -1 & 4 & 9
\end{bmatrix}
\]  \hspace{1cm} (A2)

The result obtained for this badly ill-conditioned matrix is given in Table 3.2 and Figures 3.3 and 3.4.

**Class B: Matrices with strong diagonal dominance**

Next, we consider the results obtained for a matrix with strong diagonal dominance. Consider the matrix (B1) given by Table 3.2.1 and Figures 3.3.1 and 3.4.1.

\[
A = \begin{bmatrix}
1 & -0.5 & 0 & 0 \\
-0.5 & 1 & -0.5 & 0 \\
0 & -0.5 & 1 & -0.5 \\
0 & 0 & -0.5 & 1
\end{bmatrix}
\]  \hspace{1cm} (B1)

The result for this matrix whose higher orders are badly ill-conditioned is given in Table 3.2.1 and Figures 3.3.1 and 3.4.1.
<table>
<thead>
<tr>
<th>$\omega$</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>41.06687148,10^2</td>
<td>71.93634885,10^2</td>
<td>9.97398364,10^2</td>
<td>16.30601,10^2</td>
<td>0.362</td>
</tr>
<tr>
<td>0.4</td>
<td>19.47767811,10^2</td>
<td>45.89387520,10^2</td>
<td>5.11841814,10^2</td>
<td>9.36720,10^2</td>
<td>0.481</td>
</tr>
<tr>
<td>0.6</td>
<td>13.91412757,10^2</td>
<td>28.34924099,10^2</td>
<td>3.63227939,10^2</td>
<td>7.76174,10^2</td>
<td>0.580</td>
</tr>
<tr>
<td>0.7</td>
<td>15.06183156,10^2</td>
<td>26.78086607,10^2</td>
<td>3.68426264,10^2</td>
<td>8.11502,10^2</td>
<td>0.641</td>
</tr>
<tr>
<td>0.8</td>
<td>17.05046044,10^2</td>
<td>33.76010295,10^2</td>
<td>4.03260035,10^2</td>
<td>11.94827,10^2</td>
<td>0.701</td>
</tr>
<tr>
<td>1.0</td>
<td>23.05568898,10^2</td>
<td>52.53562902,10^2</td>
<td>5.32875970,10^2</td>
<td>18.94077,10^2</td>
<td>0.822</td>
</tr>
<tr>
<td>1.2</td>
<td>32.13554449,10^2</td>
<td>79.2764892,10^2</td>
<td>6.26109315,10^2</td>
<td>29.52765,10^2</td>
<td>0.946</td>
</tr>
<tr>
<td>1.4</td>
<td>46.34723061,10^2</td>
<td>122.90783970,10^2</td>
<td>10.90435109,10^2</td>
<td>44.0054,10^2</td>
<td>0.949</td>
</tr>
<tr>
<td>1.6</td>
<td>71.59997456,10^2</td>
<td>204.12917384,10^2</td>
<td>17.29275963,10^2</td>
<td>62.49825,10^2</td>
<td>0.873</td>
</tr>
<tr>
<td>1.8</td>
<td>121.29046215,10^2</td>
<td>363.16830513,10^2</td>
<td>29.91357133,10^2</td>
<td>85.05411,10^2</td>
<td>0.701</td>
</tr>
<tr>
<td>2.0</td>
<td>220.08699705,10^2</td>
<td>673.26857257,10^2</td>
<td>54.98113570,10^2</td>
<td>116.69128,10^2</td>
<td>0.506</td>
</tr>
</tbody>
</table>

**TABLE 3.1**: Condition numbers and eigenvalues of Matrix A1 (F.O.T.)
FIGURE 3.1: Variation of P-condition Number for Matrix A1 with \( \omega \) (F.O.T.)

FIGURE 3.2: Variation of Maximum and Minimum Eigenvalues of Matrix A1 with \( \omega \) (F.O.T.).
<table>
<thead>
<tr>
<th>( \omega )</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue ( \lambda_{\text{max}} )</th>
<th>Minimum Eigenvalue ( \lambda_{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>25.720120069.10³</td>
<td>30.617285023.10³</td>
<td>70.68941185.10²</td>
<td>15.34428.10²</td>
<td>0.060</td>
</tr>
<tr>
<td>0.4</td>
<td>21.129554323.10³</td>
<td>30.617285026.10³</td>
<td>55.22437120.10²</td>
<td>12.90320.10²</td>
<td>0.071</td>
</tr>
<tr>
<td>0.6</td>
<td>14.089234264.10³</td>
<td>30.617285323.10³</td>
<td>45.6374092.10²</td>
<td>11.56263.10²</td>
<td>0.082</td>
</tr>
<tr>
<td>0.7</td>
<td>13.062131491.10³</td>
<td>30.617286359.10³</td>
<td>42.50430777.10²</td>
<td>11.44072.10²</td>
<td>0.088</td>
</tr>
<tr>
<td>0.8</td>
<td>12.474528836.10³</td>
<td>30.617286571.10³</td>
<td>40.26880503.10²</td>
<td>11.56966.10²</td>
<td>0.093</td>
</tr>
<tr>
<td>1.0</td>
<td>12.141625845.10³</td>
<td>30.617283951.10³</td>
<td>38.07793796.10²</td>
<td>12.25889.10²</td>
<td>0.101</td>
</tr>
<tr>
<td>1.1</td>
<td>12.333855641.10³</td>
<td>30.61728483.10³</td>
<td>38.00051827.10²</td>
<td>12.75622.10²</td>
<td>0.103</td>
</tr>
<tr>
<td>1.3</td>
<td>13.528138066.10³</td>
<td>30.617285778.10³</td>
<td>39.08521583.10²</td>
<td>14.01112.10²</td>
<td>0.104</td>
</tr>
<tr>
<td>1.5</td>
<td>16.187846027.10³</td>
<td>30.617283951.10³</td>
<td>44.88349603.10²</td>
<td>15.60578.10²</td>
<td>0.096</td>
</tr>
<tr>
<td>1.7</td>
<td>21.129554323.10³</td>
<td>30.617295163.10³</td>
<td>54.63517651.10²</td>
<td>17.55125.10²</td>
<td>0.083</td>
</tr>
<tr>
<td>1.9</td>
<td>29.716424834.10³</td>
<td>32.65333620.10³</td>
<td>72.14467135.10³</td>
<td>19.86769.10²</td>
<td>0.067</td>
</tr>
</tbody>
</table>

**TABLE 3.2:** Condition numbers and eigenvalues of matrix A2 (F.O.T.)
FIGURE 3.3: The Variation of P-Condition Number of Matrix $A_2$ with $\omega$ F.O.T.

FIGURE 3.4: The Variation of Maximum and Minimum Eigenvalues of Matrix $A_1$ with $\omega$ (F.O.T.)
<table>
<thead>
<tr>
<th>$\omega$</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>8.05762</td>
<td>9.13600</td>
<td>2.84745</td>
<td>1.67035</td>
<td>0.20730</td>
</tr>
<tr>
<td>0.4</td>
<td>4.98317</td>
<td>7.93600</td>
<td>1.98822</td>
<td>1.33586</td>
<td>0.26807</td>
</tr>
<tr>
<td>0.7</td>
<td>3.16543</td>
<td>7.02400</td>
<td>1.46647</td>
<td>1.0875</td>
<td>0.35027</td>
</tr>
<tr>
<td>1.0</td>
<td>2.30389</td>
<td>6.4000</td>
<td>1.19885</td>
<td>1.05247</td>
<td>0.45682</td>
</tr>
<tr>
<td>1.1</td>
<td>2.17999</td>
<td>6.4000</td>
<td>1.6542</td>
<td>1.08294</td>
<td>0.49676</td>
</tr>
<tr>
<td>1.2</td>
<td>2.10971</td>
<td>6.4000</td>
<td>1.55578</td>
<td>1.12976</td>
<td>0.55661</td>
</tr>
<tr>
<td>1.3</td>
<td>2.13947</td>
<td>6.4000</td>
<td>1.17494</td>
<td>1.190086</td>
<td>0.55661</td>
</tr>
<tr>
<td>1.4</td>
<td>2.39302</td>
<td>6.4000</td>
<td>1.22638</td>
<td>1.26615</td>
<td>0.52910</td>
</tr>
</tbody>
</table>

**TABLE 3.2.1:** Condition numbers and Eigenvalues for matrix Bl (F.O.T.)
FIGURE 3.3.1: The Variation of P-Condition Number of Matrix B1 with $\omega$ (F.O.T.)

FIGURE 3.4.1: The Variation of Maximum and Minimum Eigenvalues of Matrix B1 with $\omega$ (F.O.T.)
Class C: Matrices which have weak or lack diagonal dominance

Consider:

(i) the $20 \times 20$ matrix $(C1)$ given by

\[
A = \begin{bmatrix}
\end{bmatrix}
\]

In the unpreconditioned form, matrix $(C1)$ has the condition numbers,

\[
P(A) = 7390.92 \\
M(A) = 6640.99 \\
N(A) = 870.52
\]

For the above matrix, the following optimal values for the condition numbers were obtained using the First Order Transformation (F.O.T.)

\[
P(A) = 2019.4072 \quad (\omega=1.0) \\
M(A) = 2968.7246 \\
N(A) = 186.9185
\]

This transformation has thus produced a minimization which approximates about 1/3 of the original condition numbers of the matrix $(C1)$.

(ii) the matrix $(C2)$ defined by

\[
A = \begin{cases}
\gamma_{ij}, & \text{if } i \leq j \\
\gamma_i, & \text{if } i > j
\end{cases}
\]

which lacks diagonal dominance. For $n=10$, we obtained the segment,
Without preconditioning, this matrix has the condition numbers:

\[ P(A) = 1750 \]
\[ M(A) = 200 \]
\[ N(A) = 33.45 \]

However, the results obtained by use of the First Order Transformation (F.O.T.) does not suggest the adequacy of this transformations for matrices which lack diagonal dominance.

### 3.3.2 The Second Order Transformation (S.O.T.)

We can express (3.21) as a second order expansion of form

\[ y = (D + \omega L + \omega^2 L^2)x \]

(3.28)

With a similar approach as in Section 3.3.1, we premultiply (3.19) by

\[ (D + \omega L + \omega^2 L^2)Ax = (D + \omega L + \omega^2 L^2)b \]

(3.29)

However, (3.29) can be shown to represent the preconditioned form

\[ \tilde{B}_\omega y = \tilde{d} \]

(3.30)

where

\[ \tilde{B}_\omega = (D + \omega L + \omega^2 L^2)A(D + \omega U + \omega^2 U^2) \]

(3.31)

and

\[ \tilde{d} = (D + \omega L + \omega^2 L^2)b \]

(3.32)

while \( y \) is defined by the transformation defined by (3.28).

Using the Second Order Transformation, we shall next obtain results for the various classes of matrices considered in the previous section.
Class A: Generally, very badly ill-conditioned matrices

Consider:

(i) The matrix (A1) defined in Section 3.3.1. The results obtained for this matrix do not appear sufficiently encouraging to warrant extensive investigations.

(ii) The matrix (A2) defined also in Section 3.3.1 also gave results which appear inferior to those of the first order transformation.

The two matrices considered in class A seem to suggest that there are some inherent restrictions in the effectiveness of the Second Order Transformation. This will be examined later in Lemma 3.2.1.

We consider next,

Class C: Matrices which are weak or lack diagonal dominance

We have:

(i) For the matrix (C1) which has weak diagonal dominance, the following condition numbers were obtained by using the Second Order Transformation (S.O.T.).

\[ P(A) = 5422.0149 \]
\[ M(A) = 4963.3285 \]
\[ N(A) = 421.2125 \]

(ii) For the matrix (C2) which lacks diagonal dominance, it will be observed that the Second Order Transformation was as ineffective as the First Order Transformation (F.O.T.)

The apparent weakness of the Second Order Transformation with regard to the two classes of matrices we have just discussed prompted investigations on necessary conditions that would permit its effectiveness.
Lemma 3.2.1

A necessary condition for the effectiveness of the S.O.T. is that there exist some norm $M$ of matrix $A$ such that

$$||M|| = ||-(L_{L}+U_{U})|| < 1.$$  \hfill (3.33)

Before proceeding with the proof of the above Lemma, we first need to make the following observations.

Since the choice of using a direct as against an iterative solution procedure is mainly determined by size and structure of the original system coefficient matrix, we shall consider the solution of the preconditioned system (3.30) by a general iterative procedure of the form

$$x^{(n+1)} = Mx^{(n)} + c$$  \hfill (3.34)

where $M$ is the amplification or iteration matrix and $c$ a constant vector.

The general convergence conditions for iterative process (3.34) are known. We shall assume a splitting of $A$ of form

$$A = I+L_{L} + U_{U},$$  \hfill (3.35)

where $A$ has the constraints defined on (3.19). $I$ is an identity matrix and $L_{L}$ and $U_{U}$ are lower and upper triangular matrices respectively.

Proof

The solution of (3.30) is equivalent to the solution of the system

$$Ay = d_{1}.$$  \hfill (3.36)

Using (3.35), we can express (3.36) in the form (3.34). Thus

$$y^{(n+1)} = My^{(n)} - d_{1}$$  \hfill (3.37)

where

$$M = -(L_{L}+U_{U}).$$  \hfill (3.38)

It is known that a necessary condition for the effective convergence of iterative processes of form (3.37) is that

$$||M|| = ||-(L_{L}+U_{U})|| = \sum_{i=1}^{n} \left| a_{ij} \right| < 1$$  \hfill (3.39)
Now, relation (3.39) is always satisfied if the original matrix $A$ has strong diagonal dominance. This completes the proof of the Lemma.

Thus, considering the splitting of $A$ defined by (3.35) and the restrictions implied in (3.39) we next applied the Second Order Transformation using the normalised forms of the various classes of matrices discussed in Section 3.3.1.

In considering the class $A$ matrices, we obtained for the normalized matrix ($A_1$) the results given in Table 3.3 and Figures 3.5 and 3.6.

Also the results obtained for the normalised matrix ($A_2$) are given in Table 3.4 and Figure 3.7.

For Class $C$ matrices with weak diagonal dominance, the results obtained for the normalised matrix ($C_1$) are given in Table 3.5 and Figure 3.8.

However, the unsatisfactory results obtained for the normalised matrix ($C_2$) in this class lacking diagonal dominance does demonstrate for such cases the inability of the normalization to approximate the requirement defined by (3.39).

For purposes of comparison we now present some results obtained for normalised Class $A$ and $C$ matrices using First Order Transformation (F.O.T.).

The results obtained for the normalised Class $A$ matrix ($A_2$) are given in Table 3.6 and Figures 3.9 and 3.10.

For the normalised matrix ($C_2$) with weak diagonal dominance, the following condition numbers were obtained using the First Order Transformation:

\[
\begin{align*}
\rho(A) &= 2019.4072 \\
M(A) &= 2968.7246 \\
N(A) &= 186.9185
\end{align*}
\]

$(\omega=1.0)$

We consider next a transformation which is a variant of the Second Order Transformation which we had earlier examined. It is hoped that the use of this variant would better approximate the boundedness condition defined by (3.39) for all classes of matrices hitherto considered.
<table>
<thead>
<tr>
<th>$\omega$</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>$4.171724473 \times 10^3$</td>
<td>$7.185011227 \times 10^3$</td>
<td>$10.11085700 \times 10^2$</td>
<td>$5.8185 \times 10^{-2}$</td>
<td>$1.4 \times 10^{-5}$</td>
</tr>
<tr>
<td>0.4</td>
<td>$2.922546587 \times 10^3$</td>
<td>$4.54679126 \times 10^3$</td>
<td>$5.23248040 \times 10^2$</td>
<td>$3.45995 \times 10^{-2}$</td>
<td>$1.7 \times 10^{-5}$</td>
</tr>
<tr>
<td>0.6</td>
<td>$1.14569946 \times 10^3$</td>
<td>$2.73955473 \times 10^3$</td>
<td>$3.18642522 \times 10^2$</td>
<td>$2.3963 \times 10^{-2}$</td>
<td>$2.1 \times 10^{-5}$</td>
</tr>
<tr>
<td>0.7</td>
<td>$1.057768462 \times 10^3$</td>
<td>$2.083612854 \times 10^3$</td>
<td>$2.7778934 \times 10^2$</td>
<td>$2.7078 \times 10^{-2}$</td>
<td>$2.3 \times 10^{-5}$</td>
</tr>
<tr>
<td>0.9</td>
<td>$1.04199214 \times 10^3$</td>
<td>$1.689361755 \times 10^3$</td>
<td>$2.4919304 \times 10^2$</td>
<td>$3.0329 \times 10^{-2}$</td>
<td>$2.9 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.1</td>
<td>$1.052114854 \times 10^3$</td>
<td>$1.781482294 \times 10^3$</td>
<td>$2.4320137 \times 10^2$</td>
<td>$3.8508 \times 10^{-2}$</td>
<td>$3.7 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.3</td>
<td>$1.050275533 \times 10^3$</td>
<td>$1.748567936 \times 10^3$</td>
<td>$2.39583868 \times 10^2$</td>
<td>$4.8281 \times 10^{-2}$</td>
<td>$4.6 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.5</td>
<td>$1.028133577 \times 10^3$</td>
<td>$1.648668303 \times 10^3$</td>
<td>$2.35490855 \times 10^2$</td>
<td>$5.8866 \times 10^{-2}$</td>
<td>$5.7 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.6</td>
<td>$1.012108693 \times 10^3$</td>
<td>$1.607829606 \times 10^3$</td>
<td>$2.34249154 \times 10^2$</td>
<td>$6.4212 \times 10^{-2}$</td>
<td>$6.3 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.7</td>
<td>$9.95603947 \times 10^2$</td>
<td>$1.592102397 \times 10^3$</td>
<td>$2.4632756 \times 10^2$</td>
<td>$7.9437 \times 10^{-2}$</td>
<td>$7.2 \times 10^{-5}$</td>
</tr>
<tr>
<td>1.9</td>
<td>$9.80910087 \times 10^2$</td>
<td>$1.873912065 \times 10^3$</td>
<td>$2.4632756 \times 10^2$</td>
<td>$7.9437 \times 10^{-2}$</td>
<td>$8.5 \times 10^{-5}$</td>
</tr>
<tr>
<td>2.0</td>
<td>$1.001583131 \times 10^3$</td>
<td>$2.188268787 \times 10^3$</td>
<td>$2.61660573 \times 10^2$</td>
<td>$8.4040 \times 10^{-2}$</td>
<td>$8.3 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

**TABLE 3.3:** Condition numbers and eigenvalues for normalised matrix $A_l$ (S.O.T.)
FIGURE 3.5: The Variation of P-Condition Number for Normalised Matrix A1 with \( \omega \) (S.O.T.)

FIGURE 3.6: Variation of Maximum and Minimum Eigenvalues of Normalised Matrix A1 with \( \omega \) (S.O.T.)
<table>
<thead>
<tr>
<th>ω</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue ( \lambda_{\text{max}} )</th>
<th>Minimum Eigenvalue ( \lambda_{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>25.711726284.10^3</td>
<td>30.61728308.10^3</td>
<td>70.51301697.10^2</td>
<td>11.5.10^-2</td>
<td>0.44725674.10^-5</td>
</tr>
<tr>
<td>0.4</td>
<td>18.141715620.10^3</td>
<td>30.617287623.10^3</td>
<td>54.54374601.10^2</td>
<td>9.7.10^-2</td>
<td>0.53467931.10^-5</td>
</tr>
<tr>
<td>0.6</td>
<td>13.448987759.10^3</td>
<td>30.617284310.10^3</td>
<td>44.1544556.10^2</td>
<td>8.5.10^-2</td>
<td>0.63201784.10^-5</td>
</tr>
<tr>
<td>0.8</td>
<td>10.927489776.10^3</td>
<td>30.617285550.10^3</td>
<td>37.67598269.10^2</td>
<td>8.0.10^-2</td>
<td>0.732273299.10^-5</td>
</tr>
<tr>
<td>0.9</td>
<td>10.197454687.10^3</td>
<td>30.617286040.10^2</td>
<td>30.617286040.10^2</td>
<td>7.9.10^-2</td>
<td>0.774703124.10^-5</td>
</tr>
<tr>
<td>1.1</td>
<td>9.376724108.10^3</td>
<td>30.617284760.10^3</td>
<td>32.91894715.10^2</td>
<td>8.1.10^-2</td>
<td>0.863841135.10^-5</td>
</tr>
<tr>
<td>1.3</td>
<td>9.217039067.10^3</td>
<td>30.617285393.10^3</td>
<td>32.09326329.10^2</td>
<td>8.5.10^-2</td>
<td>0.922050562.10^-5</td>
</tr>
<tr>
<td>1.5</td>
<td>9.692603413.10^3</td>
<td>30.617285145.10^3</td>
<td>32.78010532.10^2</td>
<td>9.2.10^-2</td>
<td>0.949177389.10^-5</td>
</tr>
<tr>
<td>1.7</td>
<td>10.90892734.10^3</td>
<td>30.617285730.10^3</td>
<td>34.99924817.10^2</td>
<td>10.2.10^-2</td>
<td>0.878007362.10^-5</td>
</tr>
<tr>
<td>1.9</td>
<td>13.097840629.10^3</td>
<td>30.714154415.10^3</td>
<td>39.10017516.10^2</td>
<td>11.5.10^-2</td>
<td>0.845103283.10^-5</td>
</tr>
<tr>
<td>2.0</td>
<td>14.672762780.10^3</td>
<td>33.156663742.10^3</td>
<td>42.67435020.10^2</td>
<td>12.4.10^-2</td>
<td>0.81053283.10^-5</td>
</tr>
</tbody>
</table>

**TABLE 3.4:** Condition numbers and Eigenvalues for normalised matrix A2 (S.O.T.)
FIGURE 3.7: The Variation of P-Condition Number of Matrix A2 with $\omega$ (S.O.T.)
<table>
<thead>
<tr>
<th>ω</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5.6607570.10³</td>
<td>6.0253131.10³</td>
<td>5.233355.10²</td>
<td>9.558.10⁻³</td>
<td>2.10⁻⁶</td>
</tr>
<tr>
<td>0.3</td>
<td>3.2630344.10³</td>
<td>5.0037145.10³</td>
<td>3.52498.10²</td>
<td>6.646.10⁻³</td>
<td>2.10⁻⁶</td>
</tr>
<tr>
<td>0.5</td>
<td>2.294065.10³</td>
<td>4.1996355.10³</td>
<td>2.518880.10²</td>
<td>5.578.10⁻³</td>
<td>2.10⁻⁶</td>
</tr>
<tr>
<td>0.7</td>
<td>1.9109284.10³</td>
<td>3.5584966.10³</td>
<td>1.95112.10²</td>
<td>5.490.10⁻³</td>
<td>3.10⁻⁶</td>
</tr>
<tr>
<td>0.9</td>
<td>1.7219407.10³</td>
<td>3.0415571.10³</td>
<td>1.65378.10²</td>
<td>5.798.10⁻³</td>
<td>3.10⁻⁶</td>
</tr>
<tr>
<td>1.1</td>
<td>1.6199786.10³</td>
<td>2.6206756.10³</td>
<td>1.496406.10²</td>
<td>6.345.10⁻³</td>
<td>4.10⁻⁶</td>
</tr>
<tr>
<td>1.3</td>
<td>1.5658716.10³</td>
<td>2.2765148.10³</td>
<td>1.416658.10²</td>
<td>7.086.10⁻³</td>
<td>5.10⁻⁶</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5401556.10³</td>
<td>2.0785451.10³</td>
<td>1.386842.10²</td>
<td>7.999.10⁻³</td>
<td>5.10⁻⁶</td>
</tr>
<tr>
<td>1.7</td>
<td>1.6864524.10³</td>
<td>2.8848930.10³</td>
<td>1.961762.10²</td>
<td>9.081.10⁻³</td>
<td>5.10⁻⁶</td>
</tr>
</tbody>
</table>

**TABLE 3.5:** Condition numbers and eigenvalues for normalised matrix Cl (S.O.T.)
FIGURE 3.8: The Variation of P-Condition Number of $\omega$ for Normalised Matrix C1 (S.O.T).
<table>
<thead>
<tr>
<th>$\omega$</th>
<th>P-Condition Number</th>
<th>M-Condition Number</th>
<th>N-Condition Number</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>25.7201328.10$^3$</td>
<td>30.6172856.10$^3$</td>
<td>70.689413.10$^2$</td>
<td>11.42255.10$^{-2}$</td>
<td>0.000004</td>
</tr>
<tr>
<td>0.4</td>
<td>18.2552087.10$^3$</td>
<td>30.6172854.10$^3$</td>
<td>55.22437.10$^2$</td>
<td>9.6078.10$^{-2}$</td>
<td>0.000005</td>
</tr>
<tr>
<td>0.6</td>
<td>14.0892324.10$^3$</td>
<td>30.6172866.10$^3$</td>
<td>45.637742.10$^2$</td>
<td>8.6096.10$^{-2}$</td>
<td>0.000006</td>
</tr>
<tr>
<td>0.7</td>
<td>13.0621302.10$^3$</td>
<td>30.6172845.10$^3$</td>
<td>42.504305.10$^2$</td>
<td>8.5118.10$^{-2}$</td>
<td>0.000007</td>
</tr>
<tr>
<td>0.8</td>
<td>12.474528.10$^3$</td>
<td>30.6172854.10$^3$</td>
<td>40.268803.10$^2$</td>
<td>8.6149.10$^{-2}$</td>
<td>0.000007</td>
</tr>
<tr>
<td>1.0</td>
<td>12.14162566.10$^3$</td>
<td>30.6172868.10$^3$</td>
<td>38.079941.10$^2$</td>
<td>9.1984.10$^{-2}$</td>
<td>0.000008</td>
</tr>
<tr>
<td>1.1</td>
<td>12.333856.10$^3$</td>
<td>30.6172850.10$^3$</td>
<td>38.000518.10$^2$</td>
<td>9.4984.10$^{-2}$</td>
<td>0.000008</td>
</tr>
<tr>
<td>1.2</td>
<td>12.782411.10$^3$</td>
<td>30.6172840.10$^3$</td>
<td>38.580109.10$^2$</td>
<td>9.9340.10$^{-2}$</td>
<td>0.000008</td>
</tr>
<tr>
<td>1.4</td>
<td>14.634399.10$^3$</td>
<td>30.6172860.10$^3$</td>
<td>41.903332.10$^2$</td>
<td>10.9946.10$^{-2}$</td>
<td>0.000008</td>
</tr>
<tr>
<td>1.6</td>
<td>18.303420.10$^3$</td>
<td>30.6172863.10$^3$</td>
<td>49.02033.10$^2$</td>
<td>13.8956.10$^{-2}$</td>
<td>0.000005</td>
</tr>
<tr>
<td>1.8</td>
<td>24.85495.10$^3$</td>
<td>30.6172863.10$^3$</td>
<td>62.158718.10$^2$</td>
<td>13.8956.10$^{-2}$</td>
<td>0.000005</td>
</tr>
<tr>
<td>2.0</td>
<td>36.008192.10$^3$</td>
<td>35.51605522.10$^3$</td>
<td>85.285587.10$^2$</td>
<td>15.765.10$^{-2}$</td>
<td>0.000004</td>
</tr>
</tbody>
</table>

**TABLE 3.6:** Condition numbers and eigenvalues for normalised matrix A2 (F.O.T.)
FIGURE 3.9: The Variation of P-Condition Number for Normalised Matrix A2 with $\omega$ (S.O.T.)

FIGURE 3.10: The Variation of Maximum and Minimum Eigenvalues of Normalised Matrix A2 with $\omega$ (F.O.T.)
3.3.3 Modified Second Order Transformation (M.S.O.T.)

For this transformation, we have a preconditioned matrix $B_\omega$ of form

$$
B_\omega = (\omega_1 I + \omega_2 U + \omega_3 U^2 ) A (\omega_1 I + \omega_2 U + \omega_3 U^2 )
$$

(3.40)

where $I$ is an identity matrix and $\omega_1, \omega_2$ and $\omega_3$ are preconditioning parameters arbitrarily chosen and tested to ascertain combinations that would produce optimal results.

We now present the results obtained by using this transformation for the following classes of matrices already considered in Section 3.3.1.

Class A: Generally, very badly ill-conditioned matrices

For this class we present the result for matrix $(A1)$ for which the following optimal values were obtained experimentally:

- $\omega_1 = 1.0$
- $\omega_2 = 0.10$
- $\omega_3 = 0.01$

$$
P(A) = 439.660304
$$

$$
M(A) = 668.40001
$$

$$
N(A) = 105.464256
$$

Class C: Matrices which are weak or lack diagonal dominance

In this class, the results obtained for the matrix $(C2)$ which lack diagonal dominance are given below:

- $\omega_1 = 0.5$
- $\omega_2 = 0.03$
- $\omega_3 = 0.0009$

$$
P(A) = 39.656109
$$

$$
M(A) = 200.0001
$$

$$
N(A) = 6.803921
$$

These results suggest that the M.S.O.T. could prove very effective for matrices in classes A and C.
3.3.4 Discussion and Observations

The relative P-condition minimizing effect of the First Order, Second Order and the implicit transformation [22], for an arbitrary matrix which is very badly conditioned is portrayed in Figure 3.11. From this Figure we observe that although the bound on the P-condition number obtained for the S.O.T. is slightly better than that for the F.O.T., they both approximate the bound obtained by the implicit transformation. The observation is valid for most Class A matrices.

From the results indicated in Sections 3.3.1 and 3.3.2 for Class C matrices, we deduce that the modified transformations (S.O.T. and F.O.T.) are inferior to the implicit transformation especially when the matrices lack diagonal dominance.

If we now compare the modified form of the Second Order Transformation (M.S.O.T.) and the implicit transformation, we observe for Class A matrices that the M.S.O.T. is competitive. For example, we have for matrix (A2), the results given in Table 3.8.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>P-Condition Number</th>
<th>N-Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpreconditioned</td>
<td>2984.10</td>
<td>752</td>
</tr>
<tr>
<td>Implicit Trans. [22]</td>
<td>358.610</td>
<td>105.46</td>
</tr>
<tr>
<td>M.S.O.T.</td>
<td>439.660</td>
<td>96.95</td>
</tr>
</tbody>
</table>

**TABLE 3.8: Matrix A2**

Also, for Class C matrices lacking diagonal dominance, the M.S.O.T. and the implicit transformation are fairly competitive. The results for the matrix (C2) are given in Table 3.9.

<table>
<thead>
<tr>
<th>Transformation</th>
<th>P-Condition Number</th>
<th>N-Condition Number</th>
<th>M-Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpreconditioned</td>
<td>1750</td>
<td>200</td>
<td>33.45</td>
</tr>
<tr>
<td>Implicit Trans. [22]</td>
<td>10</td>
<td>100</td>
<td>2.5</td>
</tr>
<tr>
<td>M.S.O.T.</td>
<td>39.65</td>
<td>200</td>
<td>6.8</td>
</tr>
</tbody>
</table>

**TABLE 3.9: Matrix C2**
FIGURE 3.11: Comparative Bounds on P-Condition Number of Matrix $A_1$

- Implicit transformation
- S.O.T. Normalised $A_1$
- F.O.T. Normalised $A_1$
From the foregoing discussions we conclude that generally for matrices with strong diagonal dominance, the modified transformations are competitive with the implicit transformation. However, for matrices which lack diagonal dominance, the implicit transformation in the original form introduced by Evans [10] should be preferred.

Computationally, however, the modified transformations are less laborious and this constitutes the unique merit they may possess over the implicit variant since the latter requires some initial matrix inversions in the formation of the preconditioned matrix.
CHAPTER FOUR

APPLICATION OF MODIFIED PRECONDITIONING

TECHNIQUES FOR THE CONSTRUCTION OF ITERATIVE METHODS
Given the general system
\[ Ax = b \]  
(4.1)
where \( A \) is a sparse matrix and \( b \) is given. A solution \( x \) to (4.1) may be obtained by evaluating
\[ x = A^{-1}b, \]
where \( A^{-1} \) is the inverse of \( A \). It is known however, that for certain problems, the computation of \( A^{-1} \) can be difficult especially when \( A \) is near singular. However, a more common solution approach is to multiply (4.1) by a non-singular matrix \( R^{-1} \) which is an approximate inverse of \( A \). Thus, the given system can be transformed to the form
\[ R^{-1}Ax = R^{-1}b, \]
(4.2)
The matrix \( R \) in (4.2) therefore assumes the role of a 'conditioning matrix'. Traditionally, certain restrictions are placed on the choice of \( R \) to allow for easy computation as well as ensuring that the condition of \( R^{-1}A \) is smaller than that of \( A \). For the modified methods discussed in this and subsequent sections, it would be unnecessary to emphasize any requirement relating to practical computation of \( R^{-1} \) since this inverse would not be computed by an implicit process.

Having formed (4.2), we can now define iterative schemes with their associated \( R \) as well as the preconditioning and acceleration parameters \( \omega \) and \( \tau \). Assume now that \( A \) has a splitting of the form
\[ A = D + L + U, \]
(4.3)
where \( D \) is in general a block diagonal matrix and \( L \) and \( U \) are block matrices such that \((D+L)\) and \((D+U)\) are easily invertible. For this form of splitting \( R \) can normally take the form,
\[ R = (D+L_1)D^{-1}(D+U_1), \]
(4.4)
where \( L_1 \) and \( U_1 \) are functions of \( L \) and \( U \). By selecting \( L_1 \) and \( U_1 \) to have
simple forms such as

\[ L_1 = \omega_1 L \quad \text{and} \quad U_1 = \omega_2 U \]

where \( \omega_1 \) and \( \omega_2 \) are real parameters to be chosen later. Various iterative schemes can now be constructed with respect to the preconditioned system (4.2) as follows

\[ x^{(n+1)} = x^{(n)} + \tau R^{-1}(b - Ax^{(n)}) \]  

(4.5)

The iterative scheme (4.5) represents all known first degree iterative methods. The various methods are illustrated in Figure 4.1. These methods are based on common premise that the errors in the initial estimate of the solution vector are progressively minimized as the iterative procedure continues.

<table>
<thead>
<tr>
<th>Preconditioning Parameters</th>
<th>Acceleration Parameter</th>
<th>Conditioning Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega_1 ) ( \omega_2 )</td>
<td>( \tau ) ( \tau_0 ) ( \omega ) ( \omega(2-\omega) ) ( \tau )</td>
<td>( R^{-1} ) ( D ) ( D ) ( D(I-\omega L)^{-1} ) ( D(I-\omega L)^{-1} ) ( D(I-\omega L)^{-1} ) ( D(I-\omega L)^{-1} )</td>
</tr>
</tbody>
</table>

TABLE 4.1

It is apparent from Table 4.1 that the central issue in the construction of new iterative schemes is the need to improve the condition of the original system by using different forms of conditioning matrix with their various acceleration and preconditioning parameters \( \tau \) and \( \omega \).
4.1 THE MODIFIED PRECONDITIONED JACOBI (MPJ) METHOD

By considering (4.5) and Table 4.1, it is easy to show that the MPJ iterative scheme has the form,

\[ x^{(n+1)} = x^{(n)} + (I - \omega U^{-1})(I - \omega U)D^{-1}D(Ax^{(n)}) \]  

(4.6)

In a more compact form, (4.6) can be represented as

\[ x^{(n+1)} = \mathcal{M}_\omega x^{(n)} + C_\omega, \]

(4.7)

where

\[ \mathcal{M}_\omega = I - (I - \omega U)^{-1}(I - \omega U)^{-1}D^{-1}A \]  

(4.8)

and

\[ C_\omega = (I - \omega U)^{-1}(I - \omega U)^{-1}D^{-1}b. \]  

(4.9)

The preconditioned matrix \( B_\omega \) of this scheme is given by

\[ B_\omega = (I - \omega U)^{-1}(I - \omega U)^{-1}(I - \omega U)^{-1}D^{-1}A. \]  

(4.10)

Next, we shall attempt to obtain expressions for the maximal and minimal eigenvalue spectrum of \( B_\omega \) since these bounds would enable us to obtain expressions for the maximal and minimal eigenvalue spectrum of \( B_\omega \) since these bounds would enable us to obtain a good estimate for \( S(\mathcal{M}_\omega) \) and to examine the latter with \( \omega \). In general, an approach requiring the determination of eigenspectrum bounds on \( B_\omega \) should be preferred to individual estimation of the eigenvalues since the latter is often considered impractical.

Now we have that

\[ B_\omega = (I - \omega U)^{-1}(I - \omega U)^{-1}B_\omega [(I - \omega U)^{-1}(I - \omega U)^{-1}]^{-1}. \]  

(4.10a)

From (4.10) and (4.10a), we obtain

\[ B_\omega = (I - \omega U)^{-1}(I - \omega U)^{-1}D^{-1}A(I - \omega U)^{-1}(I - \omega U)^{-1} \]

\[ = D^{-1}(I - \omega U)^{-1}(I - \omega U)^{-1}A[(I - \omega U)^{-1}(I - \omega U)^{-1}]T. \]  

(4.11)

From the relation (4.11), we have that \( B_\omega \) is similar to \( A \) since it was derived from the latter by a congruent transformation and since \( (I - \omega U)^{-1}(I - \omega U)^{-1} \) is non-singular. Also, \( B_\omega \) is positive definite since
it is similar to $\tilde{B}_w$.

If $\lambda_i$ and $\xi_i$ are real eigenvalues of $m_Jf$ and $B_w$ respectively, then, they are related by

$$\lambda_i = 1 - \xi_i . \quad (4.12)$$

**Convergence of the MPJ Method**

By virtue of Theorem 2.7 and (4.12) we have that the MPJ scheme converges if and only if

$$0 < \xi_i < 2 . \quad (4.13)$$

**Theorem 4.1**

Given that the matrix $A$ is symmetric and has positive definite entries, then

$$S(\tilde{m}_w) < 1 , \quad (4.14)$$

if and only if,

(i) $A$ is positive definite

(ii) $\omega_a < \omega < \omega_f < 2(\omega_A$ and $\omega_f$ are maximum and minimum $\text{bom}(\omega, \omega)$). \quad (4.15)

where $\omega_a = 1 - \sqrt{6}/2$ and $\omega_f = 1 + \sqrt{6}/2$.

**Proof**

From [24], we have for the SSOR scheme that $\xi$ has the form

$$\xi = I - \omega (2 - \omega) (I - \omega L)^{-1} (I - \omega U)^{-1} A . \quad (4.16)$$

We shall however prefer (4.16) in the form

$$\xi = I - \omega (2 - \omega) (I - \omega L)^{-1} (I - \omega U)^{-1} (I - \omega^2 L^2) D^{-1} A \quad (4.17)$$

From (4.10) and (4.17) we have

$$\xi = I - \omega (2 - \omega) B_w . \quad (4.18)$$

Now, (4.18) is equivalent to

$$\omega (2 - \omega) B_w = I - \xi \quad (4.19)$$

and if and only if $\xi_i$ and $\lambda_i$ are eigenvalues of $B_w$ and $\xi_w$ respectively, then we have the relation

$$\omega (2 - \omega) \xi_i = 1 - \lambda_i . \quad (4.20)$$
If we assume also that $A$ is positive definite and so (4.15) is valid, then from theorem 2.4 we have that $\lambda_i$ lies within the interval $(0,1)$ which by (4.20) implies that

$$\xi_i \in (0, \frac{1}{\omega(2-\omega)}) \quad (4.21)$$

But since it is obvious from (4.15) that

$$\frac{1}{\omega(2-\omega)} < 2 \quad (4.22)$$

then, it follows that (4.13) holds and consequently (4.14) is established.

Since we have from (4.21) that $\xi_i > 0$, then by (4.11) we have that $A$ is positive definite.

We have already shown by virtue of theorem 2.4 that $0 \leq \xi \leq 1$ when $0 < \omega < 2$. Thus (4.15) is established and the proof of the theorem is complete.

4.1.1 Estimation of Good Bounds on $\xi_{\max}, \xi_{\omega}(B)$ and $\xi_{\min}, \xi_{\omega}(B)$

In order to determine the bounds of spectrum of $B_\omega$ we shall adopt an analytical approach analogous to that of Habetler and Wachpress [21] which involves the computation of the eigenvalues of $B_\omega$ in terms of certain inner products.

If $\xi$ is an eigenvalue of $B_\omega$ and $v$ is an associated eigenvector, then

$B_\omega v = \xi v \quad (4.23)$

From expressions (4.10) and (4.23), we have

$$(I-\omega U)^{-1}(I-\omega U^2)(I-\omega L)^{-1}(I-\omega L^2)D^{-1} Av = \xi v \quad (4.24)$$

Alternatively, from (4.24) we have

$$Av \preceq \xi D(I-\omega U)(I+\omega U^2)(I-\omega L)(I+\omega L^2)v \quad (4.25)$$

Taking the inner products of both sides of (4.25) with respect to $v$ yields

$$(v, Av) = \xi (vD(I-\omega L+\omega L^2)+\omega L^2+U^2+LU))v \quad (4.26)$$

From (4.26), we obtain

$$\xi = \frac{v.Av}{(v, Dv)-\omega(v, Dv)+\omega L^2+U^2+LU} \quad (4.27)$$
Also, (4.24) can be expressed in the form,

\[ \xi = \frac{(v, Dv) - (v, DBv)}{(v, Dv) - \omega(v, DBv) + \omega^2((L+U)^2 - LU)v} \]  

(4.28)

Dividing the numerator and denominator of (4.28) by \((v, Dv)\neq 0\), we have the expression

\[ \xi = \frac{1 - \hat{a}(v)}{1 - \omega \hat{a}(v) + \omega^2 (\hat{a}^2(v) - \hat{\beta}(v))} \]  

(4.29)

where

\[ \hat{a}(v) = \frac{(v, DBv)}{(v, Dv)} \]  

(4.30)

and

\[ \hat{\beta}(v) = \frac{(v, DLUv)}{(v, Dv)} \]

Expression (4.29) indicates that the maximal and minimal eigenvalues of \( B \) could be determined if we could obtain their associated eigenvectors \( \lambda \) respectively. We shall therefore use the bounds on \( \hat{a}(v) \) and \( \hat{\beta}(v) \) to determine some reasonable bounds on \( \xi_{\text{max}}(B) \) and \( \xi_{\text{min}}(B) \).

Now, since \( B \) is similar to a symmetric matrix, we have by Theorem 2.1 for any \( v \neq 0 \) that

\[ \frac{1}{1 - \omega \hat{a}(v) + \omega^2 (\hat{a}^2(v) - \hat{\beta}(v))} \leq \xi_{\text{max}}(B) \]

(4.31)

Before proceeding to determine the bounds on \( \xi_{\text{min}}(B) \) and \( \xi_{\text{max}}(B) \), we shall first establish the following relationships with respect to the quantities \( \hat{a}(v) \) and \( \hat{\beta}(v) \).

**Lemma 4.1**

If the eigenvalues \( \mu \) of \( B \) lie in the range

\[ m(B) = m \leq \mu \leq M = M(B) \]  

(4.32)

then \( \hat{a}(v) \) and \( \hat{\beta}(v) \) are also bounded by

\[ m = m(B) \leq \hat{a}(v) \leq M = M(B) \]  

(4.33a)

and

\[ 0 \leq \hat{\beta}(v) \leq S(LU) \]  

(4.33)

**Proof**

We have from (4.30) that
\[ a(v) = \frac{(v, DBv)}{(v, Dv)} = \frac{(D^t (D^t BD^{-1}) D^t v)}{(D^t v, D^t v)} = \frac{(E, BE)}{(E, E)} \]  

where we have set \( E = D^t v \) and \( \tilde{B} = D^t BD^{-1} \).

We have therefore that \( a(v) \) is a Rayleigh quotient with respect to \( \tilde{B} \) which is similar to \( B \). By virtue of Theorem 2.1, (4.33a) is established.

Similarly, from (4.30), we have that

\[ \hat{b}(v) = \frac{(v, DLUv)}{(v, Dv)} = \frac{(D^t v, (D^t LUD^{-1}) D^t v)}{(D^t v, D^t v)} = \frac{(E, \tilde{L}E)}{(E, E)} \]  

where we have set \( E = D^t v \) and \( \tilde{L} = \tilde{U}^T = D^t LD^{-1} \).

Thus, \( \hat{b}(v) \) is also a Rayleigh quotient with respect to the symmetric and positive definite matrix \( L \) and the proof is complete.

We have assumed that \( A \) is positive definite and so \( B \) is similar to the matrix

\[ \tilde{B} = D^t BD^{-1} \]

\[ = I - D^{-1} AD^{-1} \]

It is known that

\[ \sum_{i=1}^{N} \mu_i = \text{trace}(B) = 0 \]  

and so,

\[ m(B) < 0 < M(B) \]  

where \( m(B) \) and \( M(B) \) are the minimal and maximal eigenvalues of \( B \) respectively. From (4.36) we therefore have that

\[ m(B) < \mu < M(B) \]

and so

\[ M(\tilde{B}) = M(B) = M < 1 \]

Also, since \( \tilde{B} = \tilde{L} + \tilde{U} \) is symmetric and similar to \( B \), we have then that

\[ S(B) = S(\tilde{B}) = S(\tilde{L} + \tilde{U}) \]

\[ = \| \tilde{L} + \tilde{U} \| \leq \| \tilde{L} \| + \| \tilde{U} \| \]
= |\|L\| + |\|L^T\| |
= 2|\|L\| = 2\sqrt{S(LU)}
= 2\sqrt{S(LU)} \leq 2\sqrt{\beta} \quad (4.38)

where \( S(LU) \leq \beta \) \quad (4.39)

Using (4.38), we have
\(-m \leq 2\sqrt{\beta} \)
\(M \leq 2\sqrt{\beta} \)

which indicates that when the bounds exceed \(-m\) and \(M\), then \(M\) should be replaced by \(2\sqrt{\beta}\) and \(-m\) by \(-2\sqrt{\beta}\).

We conclude therefore that the following inequalities apply:
\[-2\sqrt{\beta} \leq m \leq 0 \leq M \leq \min(1, 2\sqrt{\beta}) \]

The following theorem establishes the upper bound for \(\xi_{\max}(B)\).

**Theorem 4.2**

Let \(A\) be a positive definite matrix, then
\[\xi_{\max}(B) = \frac{1}{\omega(2-\omega)} \quad (4.40)\]

where \(0 < \omega < 2\).

**Proof**

We have from (4.26) that
\[(v,Av) = \xi (v.D(I-\omega U-\omega L + \omega^2 L^2 + \omega U^2 + \omega^2 LU)v\]
and thus
\[\xi = \frac{(v,Av)}{(v,D[(I-\omega U)(I-\omega L)]v + vD\omega^2 ((L+U)^2 - LU)v} \quad (4.41)\]

It can be established that
\[D(I-\omega L)(I-\omega U) = \omega(2-\omega)A + D[(1-\omega)I+\omega L][1-(1-\omega)I+\omega U] \quad (4.42)\]
Also, \(vD\omega^2 ((L+U)^2 - LU)v = vD\omega^2 (B^2 - LU)v\).

Now since,
\[(v, DB^2v) = (D^\top, DB^2D^\top)D^\top v \geq 0 \quad (4.43)\]
and
\[(v, DLUv) = (D^\top v, DLU^2D^\top v \geq 0) \]

we have using (4.41) and (4.42) that

\[ \xi = \frac{(v,Av)}{(v,\omega(2-\omega)Av)+v,D[(1-\omega)I+\omega L][(1-\omega)I+\omega L]v+(v,\mathcal{B}^2)v+(v,DLUV)} \]

Thus

\[ \xi \leq \frac{(v,Av)}{(v,\omega(2-\omega)Av)} \]  

(4.43a)

and this is established for all eigenvalues of \( B \). It is known however that \( B \) is positive definite, therefore by (4.43a), (4.40) is established with the proviso however that the interval \( 0<\omega<2 \) for \( \omega \) remains valid.

Next, we shall attempt to determine the lower bound on \( \xi_{\min} \) by examining the behaviour of (4.40) with respect to \( \hat{\alpha}(v) \) and \( \hat{\beta}(v) \). Now consider the following theorem.

**Theorem 4.3**

Given that \( \beta, m, M \) are numbers such that the following are valid:

\[-2\sqrt{\beta} \leq m \leq m(B) \]
\[ M(B) \leq M \leq \min(1,\sqrt{2}) \]

and

\[ S(LU) \leq \beta \]  

(4.44)

then a minimal bound on \( \xi_{\min} \) is given by

\[ \xi_{\min} \geq \begin{cases} 
\frac{1-M}{1-\omega M+\omega^2(3\beta)} = Z_1(\omega) & \text{if } \beta \leq \beta_0 \text{ or if } \beta_0 \leq \beta \\
\frac{1-M}{1-\omega M+\omega^2(3\beta)} = Z_2(\omega) & \text{if } \beta_0 \leq \beta \text{ and } \omega > \omega^* 
\end{cases} \]  

(4.45)

where \( \beta_0 \), we define \( \omega^* \) by

\[ \omega^* = \frac{2}{1+\sqrt{1+3\beta-2M}} \]  

(4.46)

Proof

The eigenvalue expression (4.29) is a function of the variables \( \omega, \hat{\alpha} \) and \( \hat{\beta} \) and we shall obtain a minimal bound for this expression by examining its behaviour with respect to \( \hat{\alpha} \) and \( \hat{\beta} \). We therefore need to solve the
related problem defined by
\[ \xi(B) \geq \min_{\hat{\alpha}, \hat{\beta}} \xi(\omega, \hat{\alpha}, \hat{\beta}) = \min_{\hat{\alpha}, \hat{\beta}} \left( \frac{1-\hat{\alpha}}{1-\omega \hat{\alpha} + \omega^2 (3 \hat{\beta})} \right) \] (4.47)

For \( \omega \) and \( \hat{\alpha} \) constant, we have
\[ \text{sgn} \left( \frac{\partial}{\partial \hat{\beta}} \xi(\omega, \hat{\alpha}, \hat{\beta}) \right) \]
\[ = \text{sgn} \left( \frac{-(1-\hat{\alpha})3\omega^2}{(1-\omega \hat{\alpha} + 3\omega^2 \hat{\beta})^2} \right) \leq 0 , \] (4.48)
since \( \omega > 0 \) and \( \hat{\alpha} \leq 1 \).

From (4.48), we have that
\[ \xi(\omega, \hat{\alpha}, \hat{\beta}) \] is a decreasing function of \( \hat{\beta} \) and thus using (4.44), (4.47), (4.33a) and (4.33), the problem reduced to
\[ \xi(B) \geq \xi(\omega, \hat{\alpha}, \hat{\beta}) . \] (4.49)

Furthermore, if we now have that \( \omega \) is constant, then
\[ \text{sgn} \left( \frac{\partial}{\partial \hat{\alpha}} \xi(\omega, \hat{\alpha}, \hat{\beta}) \right) = \text{sgn}(-3\omega^2 \hat{\beta} + \omega - 1). \]

Thus, we can now construct Table 4.2 which establishes (4.45). The proof of the theorem is now complete.

<table>
<thead>
<tr>
<th>( \overline{\beta} )</th>
<th>( \omega )</th>
<th>( 3\omega^2 \hat{\beta} - \omega + 1 )</th>
<th>Bound on ( \xi(B) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \overline{\beta} &gt; \hat{\beta} )</td>
<td>( 0 &lt; \omega &lt; 2 )</td>
<td>( &gt; 0 )</td>
<td>( Z_1(\omega) = \xi(\omega, M, \hat{\beta}) )</td>
</tr>
<tr>
<td>( 0 &lt; \omega &lt; \omega^* )</td>
<td>( &gt; 0 )</td>
<td>( Z_1(\omega) = \xi(\omega, M, \hat{\beta}) )</td>
<td></td>
</tr>
<tr>
<td>( \omega = \omega^* )</td>
<td>( = 0 )</td>
<td>( Z_1(\omega) = \xi(\omega, \hat{\alpha}, \hat{\beta}) )</td>
<td></td>
</tr>
<tr>
<td>( \omega^* &lt; \omega &lt; 2 )</td>
<td>( &lt; 0 )</td>
<td>( Z_2(\omega) = \xi(\omega, m, \beta) )</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 4.2:** The variation of \( \xi(\omega, \hat{\alpha}, \hat{\beta}) \) as a function of \( \hat{\alpha} \)

**4.1.2 Estimation of \( S(\mathcal{K}_m, \omega_1) \) and \( \omega_1 \)**

From the preceding sections we have
\[ S(\mathcal{K}_m, \omega) = \max_\lambda \left( \lambda (\mathcal{K}_m), |\lambda (\mathcal{K}_m)| \right) , \]
(4.50)
where $\lambda_{\text{max}}(\mathcal{H})$ and $\lambda_{\text{min}}(\mathcal{H})$ are the maximal and minimal eigenvalues of $\mathcal{H}$ respectively.

From (4.12) and (4.29) we can deduce that

$$\lambda_{\text{max}}(\mathcal{H}) = 1 - \xi(\omega) ,$$

and

$$|\lambda_{\text{min}}(\mathcal{H})| = \frac{1}{\omega(2-\omega)} - 1 .$$

(4.51)

Next, consider the variation with $\omega$ of the function

$$\lambda(\omega, \hat{\alpha}, \hat{\beta}) = 1 - \xi(\omega, \hat{\alpha}, \hat{\beta}) ,$$

in the interval $\omega \in (0, 2)$. We have that

$$\text{sgn} \left( \frac{\partial}{\partial \omega} \lambda(\omega, \hat{\alpha}, \hat{\beta}) \right) = \text{sgn}(6\omega\hat{\beta} - \hat{\alpha}) .$$

(4.52)

(4.53)

Now, (4.53) suggests that we have to examine the behaviour of $(6\omega\hat{\beta} - \hat{\alpha})$.

This is illustrated in Table 4.3.

<table>
<thead>
<tr>
<th>$\omega$-Domain</th>
<th>$6\omega\hat{\beta} - \hat{\alpha}$</th>
<th>$\lambda(\omega, \hat{\alpha}, \hat{\beta})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega &gt; \frac{\hat{\alpha}}{6\hat{\beta}}$</td>
<td>$&gt;0$</td>
<td>Increasing</td>
</tr>
<tr>
<td>$\omega = \frac{\hat{\alpha}}{6\hat{\beta}}$</td>
<td>$=0$</td>
<td>Stationary</td>
</tr>
<tr>
<td>$\omega &lt; \frac{\hat{\alpha}}{6\hat{\beta}}$</td>
<td>$&lt;0$</td>
<td>Decreasing</td>
</tr>
</tbody>
</table>

TABLE 4.3: Variation of $\lambda(\omega, \hat{\alpha}, \hat{\beta})$ as a function of $\omega$

By virtue of Table 4.3, we can examine the behaviour with respect to $\omega$ of the functions

$$\gamma_1(\omega) = \lambda(\omega, \hat{M}, \hat{\beta})$$

$$\gamma_2(\omega) = \lambda(\omega, \hat{\alpha}, \hat{\beta}) .$$

(4.54)

We have therefore that the critical point of $\gamma_1(\omega)$ is

$$\omega_{\hat{M}} = \frac{M}{28} .$$

(4.55)

But since $6\omega\hat{\beta} - \hat{\alpha} > 0$ for any value of $\omega$ in the range $0 < \omega < 2$, then $6\omega\hat{\beta} - \hat{\alpha}$ is an increasing function in the interval $\omega \in (0, 2)$. From the foregoing analysis, we can establish the following Table 4.4 and Figure 4.1.
TABLE 4.4: Variation of $\gamma_1(\omega)$ as a function of $\omega$

<table>
<thead>
<tr>
<th>Interval of $\omega$</th>
<th>$6\omega \beta - M$</th>
<th>$\gamma_1(\omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; \omega &lt; \omega_M$</td>
<td>$&lt; 0$</td>
<td>Decreasing</td>
</tr>
<tr>
<td>$\omega = \omega_M$</td>
<td>$= 0$</td>
<td>Stationary</td>
</tr>
<tr>
<td>$\omega_M &lt; \omega$</td>
<td>$&gt; 0$</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

**Diagram Description:**
- The graph shows the variation of $\gamma_1(\omega)$ as a function of $\omega$.
- The function $\gamma_1(\omega)$ decreases in the interval $0 < \omega < \omega_M$.
- It is stationary at $\omega = \omega_M$.
- It increases in the interval $\omega_M < \omega$.

**Graph Key:**
- $\gamma_1(\omega)$ is plotted against $\omega$.
- The intervals are marked on the horizontal axis.
- The vertical axis represents the value of $\gamma_1(\omega)$.
FIGURE 4.1: Variation of $\gamma_2(\omega)$ as a function of $\omega$

It can be verified that the crossover point $\hat{\omega}$ is given by $\gamma_2(\omega)=0$ or $f(\omega)=3\omega^2 \hat{\beta}^2 - \omega m + \gamma = 0$ when

$$\omega = \hat{\omega} = \frac{-m + \sqrt{m^2 - 12\hat{\beta}m}}{6\hat{\beta}}$$

(4.56)

Also, since we have that $\text{sgn}(f(0))$ and $f(1)<0$, and $m/\hat{\beta}<0$, then $\hat{\omega} \in (0,1)$.

Considering our earlier analysis we can establish the following three different cases:

The General Case A: $\beta \geq \hat{\beta}$, $0<\omega<2$.

For this case we have that

$$\gamma_1(\omega) \geq \gamma_2(\omega)$$

(4.57)

Case B: $0 \leq \beta < \hat{\beta}$, $0<\omega<\omega^*$

We have here that

$$\gamma_1(\omega) \geq \gamma_2(\omega)$$

(4.58)
Case C: \( 0 \leq \phi \leq \pi, 0 < \omega^* < \omega \).

In this case we have that

\[ \gamma_1(\omega) \leq \gamma_2(\omega). \tag{4.59} \]

The variation of \( \gamma_1(\omega) \) and \( \gamma_2(\omega) \) as a function of \( \omega \) is illustrated in Table 4.5.
<table>
<thead>
<tr>
<th>Interval</th>
<th>$\omega$- Interval</th>
<th>$\gamma_{\max} (\mathcal{H})$ Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho &gt; 1/12$</td>
<td>$0 &lt; \omega &lt; 2$</td>
<td>$\gamma_1 (\omega)$</td>
</tr>
<tr>
<td>$0 &lt; \rho &lt; 1/12$</td>
<td>$0 &lt; \omega &lt; \omega^*$</td>
<td>$\gamma_1 (\omega)$</td>
</tr>
<tr>
<td>$M = 12 \rho$</td>
<td>$M = 12 \rho$</td>
<td></td>
</tr>
<tr>
<td>$0 &lt; \rho &lt; 1/12$</td>
<td>$\omega^* &lt; \omega &lt; 2$</td>
<td>$\gamma_2 (\omega)$</td>
</tr>
<tr>
<td>$M &gt; 12 \rho$</td>
<td>$M &gt; 12 \rho$</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 4.5:** Relation between $\gamma_1 (\omega)$ and $\gamma_2 (\omega)$
We shall next employ the analysis we have developed so far to determine \( S(\mathcal{K}) \) by examining the relation of \( \lambda_{\max}(\mathcal{K}) \) and \( |\lambda_{\min}(\mathcal{K})| \). This is portrayed in Figure 4.2.

Consider now Case A, then we have from Table 4.5 that \( \gamma_1(\omega) \) dominates over \( \gamma_2(\omega) \) for all values of \( \omega \) in the interval \( 0 \leq \omega \leq 2 \). The value of \( \omega \) in this case which best minimizes the bound on \( S(\mathcal{K}) \) can be obtained from the relation

\[
\lambda_{\max}(\mathcal{K}) = |\lambda_{\min}(\mathcal{K})|.
\]  

(4.60)

Using (4.51) and (4.45) we have that (4.60) reduces to

\[
1 - \frac{1-M}{1-\omega M + 3\omega B} = \frac{1}{\omega(2-\omega)} - 1.
\]  

(4.61)

Equivalently, (4.61) can be expressed as,

\[
P(\omega) = 6B\omega^4 - 2(M+6B)\omega^3 + (1+5M+3B)\omega^2 - (3M+2)\omega + 1 = 0
\]  

(4.62)

Next, we consider the problem of the existence and uniqueness of the
roots of (4.62). We have from expression (4.62) that \( \text{sign}(p(0), p(1)) > 0 \) and \( \text{sign}(p(1) \text{ and } p(2)) > 0 \). We therefore deduce from the Descartes rule of signs that either there are no roots at all in the interval \( 0 \leq \omega \leq 2 \) or there are at least two positive roots \( \omega_1 \) and \( \omega_2 \) such that \( 0 < \omega_1 < 1 \), and \( 1 < \omega_2 < 2 \). A unique value of this root lies in the interval \( 1 < \omega < 2 \) as shown in Figure 4.2. This value can be determined numerically by solving (4.62) by any root finding algorithm. It is then possible thereafter to obtain the bound on \( S(\mathcal{K}) \) from

\[
S(\mathcal{K}) = \frac{1}{\omega_2 (2-\omega_2)} - 1.
\]  

(4.63)

If we were to solve (4.63) we would obtain \( \omega_2 \) and compare it with \( \omega^* \). However, we can set

\[
\omega_f \leq \omega^*,
\]

(4.64)

or using (4.46),

\[
\omega_f \leq \omega \leq \frac{2}{1+\sqrt{1-12\beta}}.
\]

(4.65)

Thus, if \( \beta > \frac{2}{\omega_f} = 0.24263 \), \( \omega_2 \) is a good choice of \( \omega \).

We can now summarize by observing that

\[
\omega_1 = \begin{cases} 
\omega_2, & \text{if } \beta \geq 0.24263 \text{ or if } \omega_2 < \omega^* \\
\omega^*, & \text{if } \omega^* < \omega_2
\end{cases}
\]

(4.66)

and the corresponding bound on \( S(\mathcal{K}) \) is given by

\[
S(\mathcal{K}) \leq \begin{cases} 
\frac{1}{\omega_2 (2-\omega_2)} - 1 & \text{if } \beta \geq 0.24263 \text{ or if } \omega < \omega^* \\
\frac{\omega^*-1}{\omega^*} & \text{if } \beta < 0.24263 \text{ or if } \omega^* < \omega_2
\end{cases}
\]

(4.67)

(4.68)

The behaviour of \( S(\mathcal{K}) \) when the bound is defined by (4.68) is portrayed in Figure 4.3.
FIGURE 4.3: The variation of $S_{m, \omega}$ with $\omega$ when $\beta<0.24263$ and $\omega^*<\omega_2$. 
4.1.3 Numerical Results

Table 4.6 shows the experimental values obtained for $S(\mathcal{K})$ in the Laplace problem on a unit square for various mesh sizes by the power method.

<table>
<thead>
<tr>
<th>$\omega$</th>
<th>$h^{-1}=10$</th>
<th>$h^{-1}=20$</th>
<th>$h^{-1}=40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.7807</td>
<td>0.8020</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.5138</td>
<td>0.5120</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.2311</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.3127</td>
<td>0.3115</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.1308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.1854</td>
<td>0.1850</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.0865</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.1215</td>
<td>0.1252</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.0950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.1163</td>
<td>0.1250</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>0.2055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>0.1693</td>
<td>0.1848</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>0.3896</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>0.2911</td>
<td>0.2911</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.6346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td>0.5078</td>
<td>0.5079</td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>0.94768</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td>0.7958</td>
<td>0.7990</td>
<td></td>
</tr>
<tr>
<td>1.9</td>
<td>0.7958</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 4.6**: Variation of $S(\mathcal{K})$ with $\omega$ for the Laplace Equation

Using (4.40) and (4.12), we can compare the theoretical upper bound on $S(\mathcal{K})$ and the experimentally computed values given in Table 4.6. From this we can deduce that there is some measure of agreement.

It is clear that Figure 4.4 depicts the behaviour of $S(\mathcal{K})$ for various mesh sizes $h^{-1}$ with the Laplace equation.
We observe for the general interval $0 \leq \omega \leq 2$ that $S_{m \omega}$ is less than unity for $\omega = 0.0, 0.1, 0.2$ which contradicts Theorem 4.1. However, $S_{m \omega}$ is fairly adequately approximated by $\lambda_{\text{max}} (\mathcal{M})$ for $0 \leq \omega \leq 1$. From these results, we have therefore that there is some agreement between theoretical and experimental results for the required bound especially in the interval $0 \leq \omega \leq 1$.

It is observed also that as $h \rightarrow 0$, the value of $\omega_1$ tends to $\omega_F$ while the spectral radius becomes less sensitive to the choice of $\omega$ in the vicinity of the optimal value. We can therefore define the convergence interval for the MPJ method by

$$0 \leq \omega \leq 1 + \frac{\sqrt{3}}{2}.$$  \hspace{1cm} (4.69)

This justifies our assumptions on the behaviour of $S_{m \omega}$ as depicted in Figures 4.2 and 4.3.

As $\omega$ increases from $\omega_F$, $S_{m \omega}$ increases rather rapidly and assumes a value near unity as $\omega$ approaches 2 and this implies a much slower convergence in this interval for the MPJ method.

Next, we compare the numerical results obtained for $S_{m \omega \text{opt}}$ by the MPJ method and its implicit variant the PJ (Preconditioned Jacobi) [10] method. Results were obtained for the Laplace equation in a unit square. The optimal parameter $\omega_{\text{opt}}$ and $S_{m \omega \text{opt}}$ were computed by the power method.

The spectral radius $S_{m \omega \text{opt}}$ is given by

$$\lambda^{(k+1)} = \frac{\mathcal{M} w^{(k)}}{w^{(k)}}$$

$$\lambda^{(k)} = \frac{(w^{(k)}, w^{(k+1)})}{(w^{(k)}, w^{(k)})}, \quad k > 0$$

for $w^{(0)} \neq 0$.

$w^{(k)}$ approximates the normalised eigenvector associated with $S_{m \omega \text{opt}}$ and $\lambda^{(k)}$ converges to $S_{m \omega \text{opt}}$ as $n$ tends to infinity [20].

Also, since we can assume that $w^{(0)}$ has a non-zero component in the
FIGURE 4.4: The variation of $S(\frac{\mathcal{H}}{m \omega})$ with $\omega$ for the Laplace Equation
direction of the dominant eigenvector, we can apply the Fibonacci search sequence [40] to obtain \( \omega^{\text{opt}} \) and \( S_{m} \omega^{\text{opt}} \).

For the given Laplace model problem, the MPJ as well as the PJ [10] methods involved the use of a starting vector \( u^{(0)} \) and were terminated when the prescribed convergence criterion \( |u^{(n)}| < 10^{-6} \) was satisfied. The results are exhibited in Table 4.7.

Table 4.7 indicates that the results for the PJ method better approximates the theoretical bound on \( S_{m} \omega^{\text{opt}} \) while those of the MPJ method reflects approximation to the bound on \( \omega^{\text{opt}} \).

The seemingly anomalous but smaller iteration count for the determination of \( S_{m} \omega^{\text{opt}} \) as \( h \rightarrow 0 \) indicates the relative convergence advantage in the computation of this bound when \( S(LU) \) takes the new bound defined for the MPJ method.
<table>
<thead>
<tr>
<th>Iterative Method</th>
<th>$n^{-1}$</th>
<th>Spectral Radius</th>
<th>$\omega_{\text{opt}}$</th>
<th>Number of iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$S(H_{\omega_{\text{opt}}})$</td>
<td>$S(H_{m\omega_{\text{opt}}})$</td>
<td>$\eta_{\text{PJ}}$</td>
</tr>
<tr>
<td>PJ (Preconditioned Jacobi) [10]</td>
<td>20</td>
<td>0.7147</td>
<td>-</td>
<td>1.6456</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.8883</td>
<td>-</td>
<td>1.6859</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.9435</td>
<td>-</td>
<td>1.6967</td>
</tr>
<tr>
<td>MPJ (Modified Preconditioned Jacobi)</td>
<td>20</td>
<td>-</td>
<td>0.11351</td>
<td>1.0710</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>-</td>
<td>0.125038</td>
<td>0.9000</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>-</td>
<td>0.116829</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**TABLE 4.7**: Comparative bounds on $\omega_{\text{opt}}$ and $S(H_{\omega_{\text{opt}}})$ obtained by the PJ and MPJ iterative methods for Laplace equation.
4.2 THE MODIFIED PRECONDITIONED SIMULTANEOUS DISPLACEMENT (MPSD) METHOD

Consider the iterative scheme defined by

\[ y^{(k+1)} = y^{(k)} + \tau(I - \omega U)^{-1} (I - \omega L)^{-1} (h - Ay) \]  

(4.70)

where the real variables \( \tau \) and \( \omega \) are used respectively as the acceleration and preconditioning parameters.

In more compact notation, the MPSD (Modified Preconditioned Simultaneous Displacement) method (4.70) can be expressed as

\[ y^{(k+1)} = D_{m, \tau, \omega} y^{(k)} + C_d \]  

(4.71)

where the associated iteration matrix is given by

\[ D_{m, \tau, \omega} = I - \tau(I - \omega U)^{-1} (I - \omega L)^{-1} (I - \omega L^2) D^{-1} A, \]  

(4.72)

and

\[ C_d = \tau(I - \omega U)^{-1} (I - \omega L)^{-1} (I - \omega L^2) b. \]  

(4.73)

From (4.72) we have that if \( A \) is non-singular and \( \tau \neq 0 \), then \( I - D_{m, \tau, \omega} \) will be non-singular.

In studying the MPSD iterative scheme (4.70), we shall first examine its convergence and related properties and by using a model problem we shall compare the results of this method with those of its implicit variant, the PSD (Preconditioned Simultaneous Displacement) method. The comparison is also extended to the SOR method by the use of some relationships established in Evans [4] for a similar problem.

4.2.1 Convergence of the MPSD Method

Given some assumptions on the matrix \( A \), we shall show that the MPSD method converges and that this convergence is valid for certain intervals of the preconditioning and acceleration parameters.

Theorem 4.2.1

If the matrix \( A \) is symmetric and positive definite, then the eigenvalues of \( D_{m, \tau, \omega} \) are real for all choices of \( \omega \) and \( \tau \). Also,
if and only if \( 0 < \omega < 2 \)
and \( 0 < \tau < 2w(2-\omega) \).

Proof

Since \( D^{-1} \) exists, we have that \( \tilde{D}_{m,\omega} \) of (4.72) is similar to

\[
\tilde{D}_{m,\omega} = D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2) D^{-1} (I-\omega^2 U^2)^{-1} \nonumber
\]

\[
= I-\tau (D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2) A (D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2))^T \nonumber
\]

We have therefore that

\[
\tilde{D}_{m,\omega} = I-\tau \tilde{B}_\omega \nonumber
\]

where the preconditioned matrix \( \tilde{B}_\omega \) is given by

\[
\tilde{B}_\omega = (D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2) A (D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2))^T \nonumber
\]

From (4.79) we have that if \( A \) is symmetric, then \( \tilde{B}_\omega \) is also symmetric.

Now, since \( \tilde{B}_\omega \) is obtained from \( A \) by congruent transformation, then by

Theorem 2.7, \( \tilde{B}_\omega \) is positive definite since \( D^{-1} (I-\omega U)^{-1} (I-\omega^2 U^2) \) is non-singular.

Now, if we designate the eigenvalues of \( \tilde{D}_{m,\omega} \) and \( B_\omega \) by \( \lambda_i \) and \( \xi_i \) respectively, then, they are related by

\[
\lambda_i = 1-\xi_i \nonumber
\]

where

\[
\xi_i > 0 \nonumber
\]

For the iterative scheme (4.71) to be convergent, we require that

\[
S(\tilde{D}_{m,\omega}) < 1 \nonumber
\]

But (4.81) can only be valid if

\[
0 < 1 - \tau \xi_i < 1 \nonumber
\]

From Young [38], we have for any eigenvalue \( \lambda \) of \( \tilde{D}_{m,\omega} \) and \( \xi \) of \( B_\omega \) that the following relationship is valid,

\[
\lambda + \omega - 1 = \omega \xi \nonumber
\]

\[
\lambda + \omega - 1 = \omega \xi \nonumber
\]
The relationship (4.83) can be expressed as a quadratic in \( \lambda \). Thus
\[
\lambda = \omega \xi \lambda^2 + \omega - 1 = 0
\] (4.84)
or
\[
\lambda = -b\lambda + c = 0
\]
where the constant \( b \) and \( c \) in (4.84) has the values
\[
b = -\omega \xi \quad \text{and} \quad c = \omega - 1.
\]
It is established in Young [38] also that if \( b \) and \( c \) are real, then they satisfy the inequalities
\[
|b| < 1 \\
|c| < 1
\]
if and only if
\[
|\omega - 1| < 1 \quad \text{and} \quad |\omega \xi| < 1 + \omega - 1
\] (4.85)
From (4.82), (4.84) and (4.85) we have (4.75) and (4.76) and so the theorem is proved.

4.2.2 Obtaining Optimal Values of \( \tau \) and \( \omega \) for the MPSD Method

Consider now the related problem of obtaining good estimates for the parameters \( \tau \) and \( \omega \) and the spectral radius, \( D_{m\tau,\omega} \) of the MPSD scheme (4.71). We shall be concerned with the general case when the matrix \( A \) is not consistently ordered.

Theorem 4.2.2

If \( \bar{\beta}, m \) and \( M \) are numbers such that
\[
-2\sqrt{\beta} \leq m \leq m(B) \\
M(B) \leq M \leq \min(1, 2\sqrt{\beta})
\]
and
\[
S(LU) \leq \bar{\beta}
\] (4.86)
then, a minimal bound on the P-condition number \( P(B_{\omega}) \) of \( B_{\omega} \) can be obtained by
\[
P(B_{\omega}) \leq \begin{cases} 
\frac{1 - \omega M + 3\omega}{2 - \omega} = \eta_1(\omega), & \text{if } \bar{\beta} > \beta \text{ or if } \bar{\beta} < \beta \text{ and } \omega > \omega^* \\
\frac{1 - \omega M + 3\omega}{2 - \omega} = \eta_2(\omega), & \text{if } \bar{\beta} < \beta \text{ and } \omega > \omega^*. 
\end{cases}
\] (4.87)
where for $\beta < \frac{1}{4}$, we define $\omega^*$ by

$$\omega^* = \frac{2}{1 + \sqrt{1 - 2\beta}}$$

(4.88)

The bound on the $P$-condition number (4.87) is minimised if $\omega$ takes the value

$$\omega_1 = \begin{cases} \frac{2}{1 + \sqrt{1 - 2\beta}} & \text{for } M < 12\beta \\ \frac{2}{1 + \sqrt{1 - 12\beta}} & \text{for } M > 12\beta \end{cases}$$

The corresponding bound on $P(B)$ is given by

$$P(B) = \begin{cases} \frac{1}{1 - M} & \text{if } M < 12\beta \\ \frac{1}{2 - \omega^*} & \text{if } M > 12\beta \end{cases}$$

(4.89)

**Proof**

The validity of (4.87) is apparent from Theorems (4.1) and (4.2).

We shall therefore be concerned with the variation with $\omega$ of bound on $P(B)$ defined by (4.87). Now, we have that

$$P(\omega, A, B) = \frac{1}{(2\omega - \omega^2)(\omega, A, B)}$$

(4.90)

We can see that (4.90) is a consequence of the definition for $P$.

Using relation (4.91) for $\xi(\omega, A, B)$, (4.90) becomes

$$P(\omega, A, B) = \frac{1 - \omega^2 + 3\omega^2 B}{2(2\omega - \omega^2)(1 - A)}$$

(4.91)

We can therefore isolate the functions

$$P(\omega, A, B) = \eta_1(\omega) \text{ and } P(\omega, M, B) = \eta_2(\omega).$$

Using (4.91), we have that

$$\text{sgn}\left(\frac{3}{\omega^2} P(\omega, A, B)\right) = \text{sgn}(\omega^2(6\beta - A) + 2\omega - 2).$$

Thus, we have that the critical points for $\eta_1(\omega)$ and $\eta_2(\omega)$ in the relevant interval $0 < \omega < 2$ are

$$\omega_M = \frac{2}{1 + \sqrt{1 - 2M + 12\beta}}$$

and

$$\omega_m = \frac{2}{1 + \sqrt{1 - 2m + 12\beta}}$$

(4.92)
It is now possible to establish the Tables 4.2.1 and 4.2.2.

<table>
<thead>
<tr>
<th>Interval</th>
<th>$P(\omega, M, \beta)$</th>
<th>$\eta_1(\omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0&lt;\omega&lt;\frac{\omega}{M}$</td>
<td>positive (&gt;0)</td>
<td>Decreasing</td>
</tr>
<tr>
<td>$\omega=\frac{\omega}{M}$</td>
<td>= zero</td>
<td>Stationary</td>
</tr>
<tr>
<td>$\omega &lt; \omega &lt; 2$</td>
<td>negative</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

**TABLE 4.2.1:** Variation of $\eta_1(\omega)=P(\omega, M, \beta)$ as a function of $\omega$

<table>
<thead>
<tr>
<th>Interval</th>
<th>$P(\omega, m, \beta)$</th>
<th>$\eta_2(\omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0&lt;\omega&lt;\frac{\omega}{m}$</td>
<td>positive (&gt;0)</td>
<td>Decreasing</td>
</tr>
<tr>
<td>$\omega=\frac{\omega}{m}$</td>
<td>= zero</td>
<td>Stationary</td>
</tr>
<tr>
<td>$\omega &lt; \omega &lt; 2$</td>
<td>negative</td>
<td>Increasing</td>
</tr>
</tbody>
</table>

**TABLE 4.2.2:** Variation of $\eta_2(\omega)=P(\omega, m, \beta)$ as a function of $\omega$

Certainly we have that

\[ \omega_m < \omega_M \]

since $m<M$. Also, when $\beta \geq m$ then $\omega \leq 1$ and using (4.92) we have $\omega_m \leq 1$. We therefore have

\[ \omega_m < \omega^* \]  

We also have that

\[ \text{sgn} \left( \frac{2}{3} P(\omega, \hat{\alpha}, \beta) \right) = \text{sgn}(3\omega^2 \hat{\beta} - \omega + 1) \]

and so we have Table 4.2.3.

<table>
<thead>
<tr>
<th>$2\omega^2 \beta - \omega + 1$</th>
<th>Relation</th>
<th>Bound on $P(B)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt;0$</td>
<td>$\eta_1(\omega) &gt; \eta_2(\omega)$</td>
<td>$\eta_1(\omega)$</td>
</tr>
<tr>
<td>$=0$</td>
<td>$\eta_1(\omega) = \eta_2(\omega)$</td>
<td>$\eta_1(\omega)$ or $\eta_2(\omega)$</td>
</tr>
<tr>
<td>$&lt;0$</td>
<td>$\eta_1(\omega) &lt; \eta_2(\omega)$</td>
<td>$\eta_2(\omega)$</td>
</tr>
</tbody>
</table>

**TABLE 4.2.3:** Relation between $\eta_1(\omega)$ and $\eta_2(\omega)$

If we consider that a similar table to Table 4.1 also applies
for \( p(\omega, \hat{a}, \hat{b}) \), then on the basis of Table 4.2.2 as well, we have that for \( \beta > \eta \), \( \eta_1(\omega) > \eta_2(\omega) \) for all \( \omega \) and therefore from Table 4.2.3,

\[
\omega_1 = \omega_M.
\]

On the other hand, if \( \beta \leq \eta \), then we can distinguish two possibilities, namely,

Case A: the range \( \omega_m \leq \omega \leq \omega_M \).

Case B: the range \( \omega_M \leq \omega \leq \omega^* \).

The following Figures 4.2.1 and 4.2.2 exhibit these possibilities.

For Case A, \( \omega = \omega^* \) while \( \omega_1 = \omega_M \) in Case B. If \( \beta > \eta \), and \( M < 1 \), then these simply imply that \( M > \eta \).

For Case A, we have that \( \beta > \eta \) and \( \omega_M > \omega^* \) and so \( M > 12 \beta \). We therefore have that (4.88) represents a good estimation of \( \omega \) while the corresponding good bound on \( P(B_{\omega_1}) \) is given by using (4.88) in (4.87).

The method we have just employed for obtaining \( \omega_1 \) and \( P(B_{\omega_1}) \) is similar to that used by Young [37] and Axelsson [1] for obtaining good bounds on \( S(\varepsilon_\omega) \).

Using a similar approach as for PSD method [24], we can establish that the optimal bound on the value of \( \tau \) is given by

\[
\tau_1 = \frac{2\omega_1(2-\omega_1)}{1+P(B_{\omega_1})} \quad (4.93)
\]
FIGURE 4.2.1: Case B: $\omega_m \leq \omega_M \leq \omega^*$

FIGURE 4.2.2: Case A: $\omega_M \leq \omega^* \leq \omega_m$
4.2.3 Numerical Results

In this section we present results obtained for a model problem by use of the MPSD and the PSD (Preconditioned Simultaneous Displacement) methods. For purposes of comparison, the results of the unpreconditioned case are also given as well as those of the SOR method.

The model problem considered was the solution of the Laplace equation on a unit square for mesh sizes $h^{-1} = 5, 10, 20$.

Table 4.2.4 shows the maximum and minimum eigenvalues as well as the P-condition number obtained by applying the modified method with the power method for the case when $h^{-1} = 5$, while Figure 4.2.3 exhibits the variation of these eigenvalues with $\omega$ for the MPSD and PSD [10] methods.

Table 4.2.5 gives the results obtained in the solution of the model problem with mesh sizes $h^{-1} = 5, 10, 20$ using the MPSD and PSD [10] methods with a first order acceleration technique.

Table 4.2.6 gives similar results as Table 4.2.5 but with second order Richardson's acceleration technique applied to the given methods.

From Table 4.2.5, we observe that when used with first order acceleration technique, the MPSD is inferior to the PSD method while Table 4.2.6 indicates that the methods are fairly competitive when they are used with the second degree Richardson's acceleration technique. Generally, however, the PSD method has slightly better convergence properties than the MPSD method especially for small $h$.

From Evans [11], we have for the SOR method that the convergence rate $R$ is given by

$$ R = \frac{4}{\sqrt{P}} \quad (4.94) $$

where $P$ is the P-condition number for the case when the solution is obtained for the model problem without use of preconditioning.
Using the relation (4.94) and Table 4.2.4, we obtain Table 4.2.6 which exhibits the convergence and relative ratio of convergence for the MPSD and SOR methods for the given mesh sizes of the model problem. The results of this table indicate that even with the relatively slow first degree acceleration, the MPSD has comparable convergence rates with the SOR method.

However, it may be observed that the MPSD method has an explicit structure which lends it more readily susceptible to parallel processing. This implies that significant computational advantages would be evidenced by the use of MPSD method over the implicit methods if the given model problem were solved using a parallel or array processor computer.
FIGURE 4.2.3: Plot of maximum and minimum eigenvalues of $\mathbf{B}_\omega$ matrix with power method
<table>
<thead>
<tr>
<th>Preconditioning Parameter $\omega$</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
<th>P-Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.58062</td>
<td>0.16174</td>
<td>9.77259</td>
</tr>
<tr>
<td>0.4</td>
<td>1.36247</td>
<td>0.19491</td>
<td>6.99025</td>
</tr>
<tr>
<td>0.6</td>
<td>1.20243</td>
<td>0.20368</td>
<td>5.07697</td>
</tr>
<tr>
<td>0.8</td>
<td>1.09520</td>
<td>0.28888</td>
<td>3.79119</td>
</tr>
<tr>
<td>1.0</td>
<td>1.04062</td>
<td>0.35243</td>
<td>2.952699</td>
</tr>
<tr>
<td>1.2</td>
<td>1.07822</td>
<td>0.42986</td>
<td>2.508305</td>
</tr>
<tr>
<td>1.4</td>
<td>1.19882</td>
<td>0.52221</td>
<td>2.295660</td>
</tr>
<tr>
<td>1.6</td>
<td>1.37344</td>
<td>0.39137</td>
<td>3.509313</td>
</tr>
<tr>
<td>1.8</td>
<td>1.60043</td>
<td>0.27042</td>
<td>5.9183122</td>
</tr>
<tr>
<td>2.0</td>
<td>1.88468</td>
<td>0.17963</td>
<td>10.9201136</td>
</tr>
</tbody>
</table>

**TABLE 4.2.4:** The P-condition number and maximum and minimum eigenvalues of Preconditioned matrix $(B_\omega)$ obtained by MPSD and power method $(h^{-1}=5)$
<table>
<thead>
<tr>
<th>Method</th>
<th>Mesh Size $h^{-1}$</th>
<th>Preconditioning Parameter $\omega$</th>
<th>Acceleration Parameter $\alpha$</th>
<th>Maximum Eigenvalue $\lambda_{\text{max}}$</th>
<th>Minimum Eigenvalue $\lambda_{\text{min}}$</th>
<th>P-Condition Number $P$</th>
<th>Rate of Convergence $R$</th>
<th>Number of Iterations $\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simultaneous Displacement</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1.80902</td>
<td>0.19098</td>
<td>9.473</td>
<td>0.211</td>
<td>35</td>
</tr>
<tr>
<td>(unpreconditioned)</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1.95106</td>
<td>0.04894</td>
<td>39.866</td>
<td>0.0502</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>1.98769</td>
<td>0.01231</td>
<td>161.470</td>
<td>0.0124</td>
<td>480</td>
</tr>
<tr>
<td>PSD [10]</td>
<td>5</td>
<td>1.3</td>
<td>1.1</td>
<td>1.09882</td>
<td>0.66383</td>
<td>1.655</td>
<td>1.2083</td>
<td>8</td>
</tr>
<tr>
<td>(Preconditioned Simultaneous Displacement)</td>
<td>10</td>
<td>1.6</td>
<td>0.95</td>
<td>1.55512</td>
<td>0.55218</td>
<td>2.816</td>
<td>0.7101</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.75</td>
<td>0.75</td>
<td>2.28502</td>
<td>0.43322</td>
<td>5.274</td>
<td>0.3792</td>
<td>21</td>
</tr>
<tr>
<td>MPSD</td>
<td>5</td>
<td>1.5</td>
<td>1.0</td>
<td>1.9882</td>
<td>0.52221</td>
<td>2.29566</td>
<td>0.87120</td>
<td>17</td>
</tr>
<tr>
<td>(Modified Preconditioned Simultaneous Displacement)</td>
<td>10</td>
<td>1.69</td>
<td>1.13</td>
<td>1.086568</td>
<td>0.249685</td>
<td>4.35175</td>
<td>0.45958</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.701</td>
<td>1.24</td>
<td>1.116367</td>
<td>0.090471</td>
<td>12.8623</td>
<td>0.15549</td>
<td>110</td>
</tr>
</tbody>
</table>

**TABLE 4.2.4:** Comparison of the PSD [10] and MPSD methods
<table>
<thead>
<tr>
<th>Method</th>
<th>Mesh Size h^{-1}</th>
<th>Preconditioning Parameter ω</th>
<th>Acceleration Parameters</th>
<th>Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simultaneous Displacement</td>
<td>5</td>
<td>0</td>
<td>1.25 0.25</td>
<td>8</td>
</tr>
<tr>
<td>(unpreconditioned)</td>
<td>10</td>
<td>0</td>
<td>1.53 0.53</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>1.73 0.73</td>
<td>65</td>
</tr>
<tr>
<td>PSD [10]</td>
<td>5</td>
<td>1.3</td>
<td>1.15 0.0175</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.6</td>
<td>1.00 0.0641</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.75</td>
<td>0.85 0.155</td>
<td>13</td>
</tr>
<tr>
<td>MPSD</td>
<td>5</td>
<td>1.3</td>
<td>1.20 0.10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.61</td>
<td>1.43 0.20</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.60</td>
<td>1.402 0.50</td>
<td>36</td>
</tr>
</tbody>
</table>

**TABLE 4.2.5**: Comparison of the MPSD and PSD 2nd-order Richardson's Acceleration methods
<table>
<thead>
<tr>
<th>Method</th>
<th>Mesh Size $h^{-1}$</th>
<th>Rate of Convergence</th>
<th>Relative Rate of Convergence</th>
<th>$R_{\text{MPSD}}$</th>
<th>$R_{\text{SOR}}$</th>
<th>$R_{\text{MPSD}} / R_{\text{SOR}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPSD</td>
<td>5</td>
<td>-</td>
<td>0.87120</td>
<td>-</td>
<td>0.6703</td>
<td>0.87120</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-</td>
<td>0.45958</td>
<td>-</td>
<td>0.7254</td>
<td>0.45958</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>-</td>
<td>0.15549</td>
<td>-</td>
<td>0.4939</td>
<td>0.15549</td>
</tr>
<tr>
<td>SOR</td>
<td>5</td>
<td>1.29962</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.633517</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.31478</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**TABLE 2.4.6:** Convergence and relative convergence ratio for the MPSD and SOR methods in the solution of the Laplace Equation in a unit square.
4.3 THE MODIFIED SUCCESSIVE OVERRELAXATION (MSOR) METHOD

We wish to solve

$$Ax = b$$  \hfill (4.95)

where A is non-singular and sparse matrix and b is given.

Let A have a splitting of the form

$$A = I - L - U$$

where L and U are strictly lower and upper triangular parts of A.

We now have

$$x^{(n+1)} = x^{(n)} + \omega (Lx^{(n)} + Ux^{(n)} - x^{(n)} - b)$$  \hfill (4.96)

The form (4.96) can be expressed as

$$(I - \omega L)x^{(n+1)} = Lx^{(n)} + (1 - \omega)Ix^{(n)} + \omega b$$  \hfill (4.97)

or equivalently,

$$x^{(n+1)} = (I + \omega L + 2L^2) (\omega U + (1 - \omega)I)x^{(n)} + [I + \omega L + 2L^2] \omega b$$  \hfill (4.98)

A more compact form of (4.98) is given by

$$x^{(n+1)} = L_{1,\omega} x^{(n)} + C$$  \hfill (4.99)

where

$$L_{1,\omega} = (I + \omega L + 2L^2) (\omega U + (1 - \omega)I)$$  \hfill (4.100)

and

$$C = (I + \omega L + 2L^2) \omega b$$  \hfill (4.101)

The expression (4.99) defines the modified successive overrelaxation (MSOR) method.

Next, we shall make various assumptions on A and determine sufficient and necessary conditions under which the scheme converges and the interval of parameter $\omega$ for which this convergence is valid.

4.3.1 Matrices with Diagonal Dominance which are also Irreducible

It has been shown in Theorem 2.11 that irreducible matrices with weak diagonal dominance are non-singular. We now prove the following theorem on the convergence of the scheme defined by (4.99).
Theorem 4.3.1

If A is an irreducible matrix with weak diagonal dominance, then the MSOR converges in the interval

\[ 0 \leq \omega \leq 1. \] (4.102)

or equivalently, if \( S(L_{1,\omega}) < 1 \), then

\[ 0 \leq \omega \leq 1. \]

Proof

Let \( 0 \leq \omega \leq 1 \) and let us assume that

\[ S(L_{1,\omega}) \geq 1. \] (4.103)

From (4.103) it follows that there is at least one value in the eigenvalue spectrum \( \lambda \) of \( L_{1,\omega} \) for which

\[ |\lambda| > 1. \] (4.104)

The eigenvalues of \( L_{1,\omega} \) are the roots of the characteristic polynomial

\[ \det(\lambda I - L_{1,\omega}) = 0 \] (4.105)

Using (4.100) and (4.105) we have

\[ \det\{\lambda^{2}(I+\omega L+\omega^{2}L^{2})-\omega I\} = 0 \] (4.106)

Let

\[ \lambda = I+\omega L+\omega^{2}L^{2} \]

and since \( \lambda \) is non-singular, it can be verified that \( \det \lambda = 1 \). It can also be shown (see Varga, 1969, pp.76) that (4.105) is equivalent to

\[ \det\{(\lambda+\omega-1)I-\omega L-\omega U\} = 0 \] (4.107)

From (4.107) we have

\[ \det\left\{ I - \frac{\lambda}{\lambda+\omega-1} L - \frac{\omega}{\lambda+\omega-1} U \right\} = 0 \] (4.108)

Let

\[ \phi = I - \frac{\lambda}{\lambda+\omega-1} L - \frac{\omega}{\lambda+\omega-1} U \]

or

\[ \phi = I - aL - \beta U \] (4.109)

where

\[ a = \frac{\lambda}{\lambda+\omega-1} \quad \text{and} \quad \beta = \frac{\omega}{\lambda+\omega-1} \]

Assuming first that matrix A is real and using (4.104) and (4.102) we have that
\( a \leq 1 \) and \( b \leq 1 \)

Let the given matrix \( A \) have a splitting of form

\[
A = (D-L-U) = (I-L^*-U^*)D
\]

therefore

\[
AD^{-1} = I-L^*-U^*.
\]

Now, if \( A \) is irreducible and has weak diagonal dominance, then \( AD^{-1} \) or \( \phi \) is also irreducible and has weak diagonal dominance. But since \( \phi \) is also irreducible we have by Theorem 2.11 that it is non-singular and so

\[
det(\phi) \neq 0.
\] \hspace{1cm} (4.110)

But (4.110) contradicts (4.108), therefore \( S(L_{l,\omega}) < 1 \).

Similarly, if \( A \) is complex we can assume complex eigenvalues \( \xi \) of \( L_{l,\omega} \) and proceed as above to obtain (4.110). Thus the proof of the theorem is complete.

### 4.3.2 \( L \) Matrices

**Theorem 4.3.2**

If \( A \) is an \( L \)-matrix of order \( N \), and if \( \omega \leq 1 \), then

\[
S(B) < 1 \quad \text{if and only if} \quad S(L_{l,\omega}) < 1.
\]

**Proof**

Since \( L \) is lower triangular, it can be verified that

\[
I + \omega L + \omega^2 L^2 \geq 0 \quad \text{(4.111)}
\]

Also,

\[
(I - \omega^2 L^2) (I + \omega U) \geq 0 \quad \text{(4.112)}
\]

Using (4.111) and (4.112) we have

\[
L_{l,\omega} = (I + \omega L + \omega^2 L^2) (I + \omega U + (1-\omega) L) \geq 0 \quad \text{(4.113)}
\]

But \( L_{l,\omega} \) is non-negative by Theorem 2.11. If \( \xi = S(L_{l,\omega}) \) is an eigenvalue of \( L_{l,\omega} \), then there is an associated eigenvector such that

\[
L_{l,\omega} v = \xi v \quad \text{(4.114)}
\]
or  
\[ (I+\omega L+\omega^2 L^2) (\omega U+(1-\omega) L)v = \xi v \]  
(4.115)

Expanding and ignoring second order terms we have

\[ (1-\omega)+(1-\omega)L+\omega U) v = \xi v \]  
(4.116)

or equivalently,

\[ \{\omega(l-w)L+U\} v = \xi (l-w) L+U \]  
(4.117)

From (4.116), we have that  \( \frac{\xi-1+w}{\omega} \) is an eigenvalue of the matrix 
\( \{\omega(l-w)L+U\} \) and we have therefore the following inequality

\[ \frac{\xi-1+w}{\omega} \leq S\left(\frac{\omega-(1-\omega)}{\omega} L+U\right) \]  
(4.118)

Now, if \( \xi \leq 1 \), then we have that

\[ S\left(\frac{\omega-(1-\omega)}{\omega} L+U\right) \leq S(L+U) = S(B) = \mu \]

By using (4.118) we have

\[ \frac{\xi-1+w}{\omega} \leq \mu \]

or

\[ \xi \leq \omega(\mu-1) + 1 \]  
(4.119)

Relation (4.119) implies that if \( \xi \leq 1 \), then \( \mu \leq 1 \) which proves the theorem.

4.3.3 Positive Definite Matrices

**Theorem 4.3.3**

If the matrix \( A \) is positive definite, then

\[ \|m_{l,w}\| \leq 1 \]  
(4.120)

if the iterative method (4.5) is completely consistent with (4.1) and if

\[ M = R + R^T - A \]  
(4.121)

is positive definite and \( R \) satisfies the relationship

\[ m_{l,w} = 1-R^{-1}_A. \]  
(4.122)

**Proof**

Since \( A \) is real and positive definite we have by Theorem 2.6 that there exists a real positive definite matrix, say \( A^\frac{1}{2} \), whose square root is \( A \).
Now, let
\[ L' = A^{-1} \left( M_{1,\omega} \right) A^{-1} = I - A^{-1} R^{-1} A^{-1} \]  \hspace{1cm} (4.123)

Following Wachpress [32], we have
\[ L'(L')^T = I - A^{-1} R^{-1} M(R')^{-1} A^{-1} \]  \hspace{1cm} (4.124)

Assuming that \( M \) is real and positive definite, then
\[ Z = M^{-1} \]
exists. We have therefore that
\[ T = A^{-1} R^{-1} M(R')^{-1} A^{-1} = (A^{-1} R^{-1} Z) (A^{-1} R^{-1} Z) \]  \hspace{1cm} (4.125)

Since \( A^{-1} R^{-1} Z \) is non-singular because \( A^{-1} \), \( R \) and \( Z \) are non-singular, then by Theorem 2.7 \( T \) is positive definite. Using (4.123), we have that the eigenvalues of \( L'(L')^T \) are less than unity, thus
\[ \| L' \| = \| L'(L')^T \| = \| S(L'(L')^T) \| < 1 \]  \hspace{1cm} (4.126)
and (4.120) follows. The proof of the theorem is now complete.

4.3.4 Consistently Ordered Matrices

We state without proof the following theorem which will be relevant to our examination of convergence of the MSOR method for consistently ordered matrices.

Theorem 4.3.4

Given that \( A \) is a consistently ordered matrix with non-zero diagonal entries and let \( B = I - (\text{diag} A)^{-1} \). If \( \mu \) is an eigenvalue of \( B \) and satisfies the relation
\[ (1 - \lambda)^2 = \mu^2 (1 - \lambda \omega) \]  \hspace{1cm} (4.127)
then \( \lambda \) is an eigenvalue of the matrix
\[ \psi = (I + \omega L + \omega^2 L)^{-1} A. \]  \hspace{1cm} (4.128)

We now prove the following theorem which relates to the interval of \( \omega \) for which the MSOR scheme is convergent.
Theorem 4.3.5

If the matrix A is consistently ordered with non-zero diagonal entries, and if matrix B has real eigenvalues, then the MSOR scheme converges if and only if

\[ \bar{\mu} = S(B) < 1 \]  \hspace{1cm} (4.129)

and the parameter \( \omega \) lies in the interval

\[ 0 < \omega < 2. \]  \hspace{1cm} (4.130)

Proof

Because A is consistently ordered, we have that Theorem 4.3.4 holds. However, (4.127) can be expressed in the form

\[ \lambda^2 - (2 - \mu^2) \lambda + (1 - \mu^2) = 0 \]  \hspace{1cm} (4.131)

By virtue of Theorem 2.8, we have that the MSOR converges if and only if

\[ S_{\omega}(\mathbf{L}_1, \omega) < 1. \]  \hspace{1cm} (4.132)

If we assume that \( a + ib \) is an eigenvalue of \( \Psi_{\omega} \), then \( 1 - (a + ib) \) is an eigenvalue of \( \mathbf{L}_1, \omega \) with modulus

\[ \left[ (1-a)^2 + b^2 \right]^{\frac{1}{2}}. \]  \hspace{1cm} (4.133)

Using (4.133), we now have from (4.132) that

\[ a^2 + b^2 < 2a, \]  \hspace{1cm} (4.134)

and from (4.134) we deduce that

\[ a > 0. \]  \hspace{1cm} (4.135)

By virtue of Theorem 4.3.4 we have that the eigenvalues of \( \Psi_{\omega} \) are given by the roots of (4.131) and so the eigenvalue with the maximum real part is given by

\[ \lambda_{\text{max}}(\omega, \mu^2) = \text{Re}\left\{ \frac{1 - \omega^2}{2} + \sqrt{\frac{\mu^2 + 4(\omega - 1)}{2}} \right\}. \]  \hspace{1cm} (4.136)

We assume first that \( \omega^2 + 4(\omega - 1) < 0 \), then \( \omega > 1 \) and \( \omega^2 - 4(\omega - 1) < 0 \) for all \( \mu^2 \) such that \( 0 < \mu^2 \). Also, we have that the modulus of the eigenvalues of \( \Psi_{\omega} \) is given by

\[ |\gamma(\omega, \mu^2)| = \sqrt{1 - \mu^2}. \]  \hspace{1cm} (4.137)
Now, (4.137) implies that

$$1 - \mu^2 > 0$$

which implies that (4.129) is valid. Now since we have shown that $a > 0$, then

$$2 - \mu^2 > 0$$

(4.137a)

and so we have that the parameter $\omega$ lies in the interval

$$0 \leq \omega \leq 2 < \frac{\omega^2}{\mu^2}$$

(4.138)

and so (4.130) is established.

Now, we need to determine a value of $\omega$ for which $S_{m_1, \omega}$ is minimal. This can be obtained when

$$| (1 - \bar{a}) + \bar{b} | = | (1 - \bar{a}) + b |$$

given that

$$\bar{a} \leq a \leq \bar{a}$$

and

$$b \leq \bar{b} \leq \bar{b}$$

(4.139)

where $\bar{a}$ and $\bar{a}$ could take positive or negative values. If we set

$$\bar{b} = \bar{b} = 0$$

(4.139a)

in (4.139), then $S_{m_1, \omega}$ attains a minimal value and this is given by

$$S_{m_1, \omega} = \frac{K(\Psi_\omega) - 1}{K(\Psi_\omega) + 1}$$

(4.140)

where $K(\Psi_\omega)$ which is called the 'virtual condition' number is defined by

$$K(\Psi_\omega) = \frac{\bar{a}}{a}$$

(4.141)

**Theorem 4.3.6**

If $A$ is consistently ordered with non-zero diagonal entries such that the matrix $B$ has real eigenvalues with $\bar{u} = S(B) < 1$, then given any value of $\omega$ in the interval $0 \leq \omega < 2$, the virtual condition number $K(\Psi_\omega)$ is given by

$$K(\Psi_\omega) = \begin{cases} 
\frac{2 - \omega^2}{2 - \omega^2}, & \text{if } 0 \leq \omega < 1 \\
\frac{2 - \omega^2}{2 - \omega^2 - 4(\omega - 1)}, & \text{if } 1 \leq \omega \leq \omega^* \\
\frac{2}{2 - \omega^2}, & \text{if } \omega^* < \omega \leq 2 
\end{cases}$$

(4.142)
and $K(\omega)$ is a strictly decreasing function of $\omega$ in the interval $0 < \omega < \omega^*$. Also, a minimal value of $K(\omega)$ is obtained if we set

$$\omega = \omega^*$$

and this value is given by

$$K(\omega^*) = \frac{1}{\sqrt{1 - \mu^2}} \tag{4.145}$$

A minimal value for the spectral radius is obtained by using (4.139) and this is given by

$$S_m(\omega) = \begin{cases} \frac{1}{\sqrt{1 - \mu^2}} & \omega < \omega^* \\ 1 & \omega = \omega^* \\ \frac{1}{\sqrt{1 - \mu^2}} & \omega > \omega^* \end{cases} \tag{4.146}$$

Proof

By Theorem 4.3.4, the eigenvalues of

$$\Psi_\omega = (I + \omega \mathcal{L} + \omega^2 \mathcal{L}^2) \mathcal{D}^{-1} \mathcal{A}$$

are the roots of

$$\lambda^2 - (2 - \mu \omega) \lambda + (1 - \mu^2) = 0,$$

and are given by

$$r(\omega, \mu^2) = \frac{2 - \mu \omega \pm \sqrt{(\mu^2 \omega^2 - 4(\omega - 1))}}{2} \tag{4.147}$$

and given that $\Psi_\omega$ are complex eigenvalues then

$$a = \frac{2 - \mu \omega}{2}$$

and

$$b = \frac{(\mu^2 \omega^2 - 4(\omega - 1))}{2} \tag{4.148}$$

Relation (4.148) suggests that

$$b = \min_{0 < \mu \leq \mu^*} b = 0 \tag{4.149}$$

and

$$\bar{b} = \max_{0 < \mu \leq \mu^*} b = \frac{(\mu^2 \omega^2 - 4(\omega - 1))}{2} \tag{4.150}$$

and so by virtue of (4.138a) we have

$$\omega^2 - 4(\omega - 1) = 0 \tag{4.151}$$

From (4.150), the parameter $\omega$ takes one of the values

$$\omega = \omega^* = \frac{2 + 2\sqrt{1 - \mu^2}}{\mu^2}$$

$$\omega = \omega^{**} = \frac{2 - 2\sqrt{1 - \mu^2}}{\mu^2}$$
or \[ \omega = \omega^* = \frac{2 - 2\sqrt{1 - \mu^2}}{\mu^2} \] (4.152)

and it is apparent that
\[ 1 \leq \omega^* < 2 < \omega^{**} \] (4.153)

We next determine a value for \( K(V_\omega) \) for the intervals \( 0 < \omega \leq 2 \). Given that \( 0 < \omega \leq 2 \), then we have from (4.135) that
\[ E_{\text{max}}(\omega, \mu^2) = \frac{2 - \omega \mu^2}{2} \] (4.154)

which implies that
\[ \max E_{\text{max}}(\omega, \mu^2) = E_{\text{max}}(\omega, 0) = 1 \] (4.155)

If we now consider from (4.150) that
\[ \omega \mu^2 - 4(\omega - 1) > 0, \]
then we define \( \mu_0^2 \) by
\[ \mu_0^2 = \begin{cases} 4(\omega - 1)/\omega^2 & \text{if } \omega > 1 \\ 0 & \text{if } \omega < 1. \end{cases} \] (4.156)

Again, if \( \mu_0^2 \leq \mu^2 \), then
\[ \omega \mu^2 - 4(\omega - 1) \leq 0 \]
and
\[ \max E_{\text{max}}(\omega, \mu^2) = 1. \]

Also, for \( 0 < \mu^2 < \mu_0^2 \), the function \( E_{\text{max}}(\omega, \mu^2) \) is given from (4.135) by
\[ E_{\text{max}}(\omega, \mu^2) = \frac{2 - \omega \mu^2 + (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}}{2} \]

and it can thus be shown that \( E_{\text{max}}(\omega, \mu^2) \) is an increasing function of \( \mu^2 \)

and so we have
\[ \max E_{\text{max}}(\omega, \mu^2) = 2 - \omega \mu^2 + (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2} \] (4.157)

Also
\[ E_{\text{max}}(\omega, \mu^2) < E_{\text{max}}(\omega, 1) = 2 - \omega \] (4.158)

We can therefore deduce that
\[ \max E_{\text{max}}(\omega, \mu^2) = \begin{cases} E_{\text{max}}(\omega, \mu^2) < 2 - \omega, & \text{if } 0 < \omega < 1 \\ 1, & \text{if } 1 < \omega < 2. \end{cases} \] (4.159)

assuming that \( \omega > 0 \).

From (4.154) we have therefore that the eigenvalues with the maximum real part is given by,
max \( E_{\text{max}}(\omega, \mu^2) \) = \[
\begin{cases} 
2-\omega \mu^2 + (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}, & \text{if } \xi \omega < 1 \\
1, & \text{if } 1 \leq \xi \omega \leq 2.
\end{cases}
\] (4.160)

To obtain eigenvalue with the minimum real part, we define the function
\[ \beta(\omega, \mu^2) = \Re \left\{ \frac{2 - \omega \mu^2 - (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}}{2} \right\} \] (4.161)
and we therefore have
\[ \beta(\omega, \mu^2) = \Re \left\{ \frac{2 - \omega \mu^2 - (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}}{2} \right\} \] (4.162)

We therefore establish,

**Lemma 4.3.1**

By virtue of Theorem 4.3.6, we have that
\[ \min \beta(\omega, \mu^2) = \beta(\omega, \mu^2) \]
\[ \frac{2-\omega \mu^2}{2}, \text{if } \xi \mu^2 - 4(\omega - 1) < 0 \]
\[ \xi \mu^2 - 4(\omega - 1) < 0 \text{ for all } \mu^2 \text{ such that } \xi \mu^2 \geq \mu^2 \text{ and so from (4.162) we have that} \]
\[ \min \beta(\omega, \mu^2) = \beta(\omega, \mu^2) = \frac{2-\omega \mu^2}{2} \] (4.164)

Also, if \( \mu^2 < 0 \), then \( \omega \mu^2 - 4(\omega - 1) < 0 \) and \( \min \beta(\omega, \mu^2) \) is given by (4.164).

However, if \( \xi \mu^2 - 4(\omega - 1) < 0 \), then
\[ \beta(\omega, \mu^2) = \frac{2-\omega \mu^2 - (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}}{2} \] (4.165)
which is a decreasing function of \( \mu^2 \) since
\[ \operatorname{sgn} \left( \frac{\partial \beta(\omega, \mu^2)}{\partial \mu^2} \right) = -\operatorname{sgn} \left( (\omega \mu^2 - 4(\omega - 1) + \omega \mu^2) \right) \] (4.166)
and we have
\[ \min \beta(\omega, \mu^2) = \beta(\omega, \mu^2) = \frac{2-\omega \mu^2 - (\mu^2 [\omega \mu^2 - 4(\omega - 1)])^{1/2}}{2} \] (4.167)

Therefore for all \( \mu^2 \) such that \( \xi \mu^2 \geq \mu^2 \) we prove that
\[ \min \beta(\omega, \mu^2) = \beta(\omega, \mu^2) \] (4.168)
Also, since \(4(\omega-1)/\omega^2\) is an increasing function of \(\omega\) in the interval \(0<\omega<2\) and using (4.150), we have that
\[
4(\omega^*-1)/(\omega^*)^2 = \frac{-2}{\mu^2}
\]
and it implies that if \(0<\omega<\omega^*\), then
\[
\frac{-2}{\mu^2} > 4(\omega-1)/\omega^2
\]
and \(\min \beta(\omega,\mu^2)\) is given by (4.167) while if \(\omega^*<\omega<2\), then
\[
\frac{-2}{\mu^2} \leq 4(\omega-1)/\omega^2
\]
and \(\min \beta(\omega,\mu^2)\) is given by (4.164), and therefore (4.163) is valid and this completes the proof of the Lemma.

From (4.160), (4.163) and (4.140) we have that \(K(\psi_\omega)\) is given by (4.141) and (4.143). From this we have therefore that \(K(\psi_\omega) > K(\psi_{\omega^*})\) if \(\omega^*<\omega<2\).

Also, we need to show that if \(0<\omega<\omega^*\), then \(K(\psi_\omega)\) is a decreasing function of \(\omega\). But for \(0<\omega<1\) we have
\[
\text{sgn}\left(\frac{3}{2\omega}(K(\psi_\omega))-\text{sgn}(\mu-1)\right) = -1
\]
and for \(1<\omega<\omega^*\)
\[
\text{sgn}\left(\frac{3}{2\omega}(K(\psi_\omega))\right) = \text{sgn}\left(2-\omega_\mu^2 - (\mu^2 \omega^2 - 4(\omega-1))\right)^{1/2}
\]
where \(\left(2-\omega_\mu^2 \right)^2 = \omega^2_\mu^2 - 4\omega_\mu^2 + 4\)
and \(\left((\mu^2 \omega^2 - 4(\omega-1))^2\right)^{1/2} = \omega^2_\mu^2 - 4\omega_\mu^2 + 4\mu^2\)
and so
\[
\text{sgn}\left(\frac{3}{2\omega}(K(\psi_\omega))\right) = -1
\]
Therefore if \(\omega=\omega^*\), then \(K(\psi_\omega)\) achieves its minimum and this is given by (4.145). Using (4.139) and (4.143) we establish that the relation (4.146).

We have established expressions for the virtual condition number and spectral radius for the MSOR scheme when the matrix \(A\) is consistently ordered and with a given property of the Jacobi iteration matrix \(B\). The expressions for \(K(\psi_\omega)\) and \(S(L_{1,\omega})\) have been shown to be dependent on \(\mu^2\). Similar expressions could equally be shown to be valid for the SOR scheme.
For the latter iterative scheme, the acceleration parameter could take the value 1 which is fixed for the MSOR method and so it is expected that the two methods should have fairly approximate optimal rates of convergence for the case considered. However, the explicit structure of the MSOR method is more likely to render it less sensitive for matrices with wide distribution of eigenvalues especially when \( \mu \to 0 \).

In subsequent theorems, we shall obtain similar expressions as before for consistently ordered matrices but with Jacobi iteration matrix which has real eigenvalues \( \mu_i \) such that \( \mu = \min \mu_i \neq 0 \).

**Theorem 4.3.7**

Given that \( A \) is consistently ordered with non-zero diagonal entries such that matrix \( B = I - D^{-1} A \) has real eigenvalues \( \mu_i, i=1, 2, \ldots, N \) with

\[
\mu = \min_i |\mu_i| \neq 0
\]

and

\[
\overline{\mu} = \max_i |\mu_i| \quad (4.171)
\]

then the MSOR converges if and only if

\[
\overline{\mu} = S(B) < 1
\]

and if

\[
(1 - \overline{\mu}^2)^{\frac{1}{2}} < 1 - \mu^2 \quad (4.172)
\]

and if

\[
(1 - \overline{\mu}^2)^{\frac{1}{2}} < 1 - \mu^2 \quad (4.173)
\]

then

\[
\max_{\omega, \mu} E_{\max} (\omega, \mu^2) = \begin{cases} 
2 - \omega \mu + (\overline{\mu}^2 - 2(\omega - 1))^{\frac{1}{2}} \quad \text{if } 0 \leq \omega \leq \omega^*(\overline{\mu}) \\
2 - \omega \mu^2 - 2 \quad \text{if } \omega^*(\overline{\mu}) \leq \omega \leq 2.
\end{cases} \quad (4.174)
\]

where

\[
\omega^*(\overline{\mu}) = \frac{2}{1 + (1 - \overline{\mu}^2)^{\frac{1}{2}}} \quad (4.175)
\]

**Proof**

We could first assume that the real parts of the eigenvalues of \( \varphi \) are positive and so from (4.136a) we have

\[
\omega < \frac{2}{\mu^2}
\]

and the eigenvalue of \( \varphi \) with maximum real part is given by
\[ E_{\text{mx}}(\omega, \mu^2) = \text{Re} \left\{ \frac{2-\omega \mu^2 + (\mu^2 [\omega^2 \mu^2 - 4(\omega-1)])^{1/2}}{2} \right\} \] (4.177)

We have also in this case that the eigenvalues relationship (4.129) is satisfied and as such we have (4.136) which implies that (4.172) is valid.

Using (4.171), we have that
\[ \mu^2 \leq \mu^2 \leq \bar{\mu}^2 \] (4.178)

From (4.135), we have that
\[ E_{\text{mx}}(\omega, \mu^2) = \frac{2-\omega \mu^2}{2} \] (4.179)

and so
\[ \max E_{\text{mx}}(\omega, \mu^2) = E_{\text{mx}}(\omega, 0) = 1 \] (4.180)

If the case where \( \omega \mu^2 - 4(\omega-1) > 0 \) is considered, then we could define \( \mu_1^2 \) by
\[ \mu_1^2 = \begin{cases} 
\frac{4(\omega-1)/\omega^2}{2}, & \text{if } \omega \geq 1 \\
0, & \text{if } \omega < 1 
\end{cases} \] (4.181)

This implies that in order to examine \( \mu_1^2 \) with respect to \( \mu^2 \) and \( \bar{\mu}^2 \) we shall identify the following cases:

(a) \( 0 < \mu^2 \leq \mu_1^2 \)

(b) \( 0 < \mu_1^2 < \mu^2 \)

and

(c) \( 0 < \mu_1^2 < \mu^2 \).

Case (a): \( 0 < \mu^2 \leq \mu_1^2 \)

For this case we have that \( \omega \mu^2 - 4(\omega-1) < 0 \) and so \( \omega > 1 \) and \( \omega \mu^2 - 4(\omega-1) < 0 \) for all \( \mu^2 \) satisfying (4.178). Also, we have from (4.177) that
\[ E_{\text{mx}}(\omega, \mu^2) = \frac{2-\omega \mu^2}{2} \] (4.182)

and so
\[ \max_{\mu_1^2 \leq \mu^2} E_{\text{mx}}(\omega, \mu^2) = \frac{2-\omega \mu_1^2}{2} \] (4.183)

where \( \omega \) lies in the interval
\[ 1 < \omega < 2 < \frac{2}{\mu_1^2} \] (4.184)

Case (b): \( 0 < \mu_1^2 \leq \mu^2 \)

Here, we need to identify two subcases.
(i) \( \mu_1^2 \leq \mu_2^2 \)

For this subcase \( \mu_2^2 - 4(\omega - 1) \geq 0 \) and so from (4.177) we have

\[
E_{\text{max}}(\omega, \mu) = \frac{2 - \omega \mu^2 + (\mu^2 [\omega^2 \mu - 4(\omega - 1)])^2}{2}
\]

Relation (4.185) can be shown to be an increasing function of \( \mu^2 \) since

\[
\text{sgn} \left( \frac{3E_{\text{max}}(\omega, \mu^2)}{3\mu^2} \right)
= \text{sgn} \left( \left[ \omega^2 \mu - 4(\omega - 1) \right]^4 \right) - \omega \mu^2
\]

Thus,

\[
\max E_{\text{max}}(\omega, \mu^2) = E_{\text{max}}(\omega, \mu^2)
\]

\[
\mu^2 \leq \mu < \mu_2^2
\]

\[
= 2 - \omega \mu^2 + (\mu^2 [\omega^2 \mu - 4(\omega - 1)])^2
\]

where it can be shown that

\[
E_{\text{max}}(\omega, \mu^2) < E_{\text{max}}(\omega, 1) = 2 - \omega
\]

(ii) \( \mu^2 \leq \mu_1^2 \)

This case is similar to case (a) since \( \omega^2 \mu^2 - 4(\omega - 1) \geq 0 \).

Case (c): \( \omega^2 \mu_1^2 \leq \mu_2^2 \)

For all \( \mu^2 \) which satisfy (4.178), we have that \( \omega^2 \mu^2 - 4(\omega - 1) \geq 0 \) and so the results of subcase (i) are valid.

From case (a) we have that if \( \omega^*(\mu) \leq \omega \), then \( \max E_{\text{max}}(\omega, \mu^2) \) is given by (4.183) and for case (c) if \( \omega^2 \mu_1^2 \leq \omega \), then \( \max E_{\text{max}}(\omega, \mu^2) \) is given by (4.187).

We have therefore that for \( \omega \) in the range \( \omega^*(\mu) \leq \omega \leq \omega^*(\mu) \) that \( \max E_{\text{max}}(\omega, \mu^2) \) could be given by either of the quantities (4.183) or (4.187). We shall therefore need to designate (4.183) as \( \eta_1 \) and (4.187) as \( \eta_2 \) and examine the quantity \( \eta_1 - \eta_2 \). However, the sign \( \eta_2 - \eta_1 \) can be shown to be

\[
\text{sgn}(\eta_2 - \eta_1) = \text{sgn}(\tilde{\omega} - \omega)
\]

where \( \tilde{\omega} \) is defined by

\[
\tilde{\omega} = \frac{\omega^2}{2 - \mu}
\]
and so for \( w \) in the interval \( \omega^*(\mu) \leq \omega \leq \omega^*(\hat{\mu}) \) we have that

\[
\max E_{mx}(\omega, \mu^2) = \begin{cases} 
\eta_2, & \text{if } w < \bar{\omega} \\
\eta_1, & \text{if } w > \bar{\omega}. 
\end{cases} \tag{4.189}
\]

Also, if (4.173) is valid, then \( \mu \leq \omega^*(\hat{\mu}) \). This implies that if (4.173) is valid, then \( \max E_{mx}(\omega, \mu^2) \) is obtained by use of (4.174).

**Lemma 4.3.2**

By virtue of hypothesis of Theorem 4.3.7 and given that \( \bar{\mu} = S(B) < 1 \),

then for \( \bar{\omega} \leq \omega \leq 2 \)

\[
\min \beta(\omega, \mu^2) = \beta(\omega, \mu^2) \frac{2}{2 - 2} \leq \mu \leq \bar{\mu}
\]

where

\[
\beta(\omega, \mu^2) = \Re \left\{ \frac{1 - \omega^2}{\omega (1 - 4\omega^2)} \right\} \tag{4.190}
\]

Also, for any \( \omega \) in the range \( \omega \leq \omega \leq 2 \), we obtain

\[
\beta(\omega, \mu^2) = \begin{cases} 
\frac{2 - \omega^2}{2} - (\omega^2 \mu^2 - 4(\omega^2 - 1))^{\frac{1}{2}}, & \text{if } \omega \leq \omega^*(\bar{\mu}) \\
\frac{1}{2} - \omega^2, & \text{if } \omega^*(\bar{\mu}) < \omega \leq 2. 
\end{cases} \tag{4.191}
\]

**Proof**

By following steps analogous to that of Lemma 4.3.1 this Lemma can be established.

**Theorem 4.3.8**

Given that \( A \) is a consistently ordered matrix with non-zero diagonal entries such that the matrix \( B = I - D^{-1}A \) has real eigenvalues \( \mu_i \) for \( i = 1, 2, \ldots, N \)

with

\[
\bar{\mu} = \min_i |\mu_i| \neq 0
\]

and

\[
\bar{\omega} = \max_i |\mu_i|
\]

and such that \( \bar{\mu} = S(B) < 1 \).

For \( \omega \) in the interval \( \omega \leq \omega \leq 2 \), we have that if

\[
(1 - \bar{\mu}^2)^{\frac{1}{2}} < 1 - \mu^2 \tag{4.193}
\]
then

\[
K(\Psi_0) = \begin{cases} 
\frac{2-\omega u^2 + (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])}{2-\omega u^2 - (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])}, & \text{if } 0 < \omega < \omega^*(\bar{\mu}) \\
\frac{2-\omega u^2}{2-\omega u^2}, & \text{if } \omega^*(\bar{\mu}) \leq \omega \leq 2
\end{cases}
\]  

(4.194)

otherwise,

\[
K(\Psi_0) = \begin{cases} 
\frac{2-\omega u^2 + (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])}{2-\omega u^2 - (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])}, & \text{if } 0 < \omega < \omega^*(\bar{\mu}) \\
\frac{2-\omega u^2}{2-\omega u^2}, & \text{if } \omega^*(\bar{\mu}) \leq \omega \leq 2
\end{cases}
\]  

(4.195)

where \(\bar{\omega}\) and \(\omega(\bar{\mu})\) are defined respectively by (4.188) and (4.144). Also, \(K(\Psi_0)\) is a decreasing function of \(\omega\) for \(0 < \omega < \omega^*(\bar{\mu})\) and if we set

\[
\omega = \omega(\bar{\mu}) = \omega^*
\]

(4.195a)

then we obtain a minimal value for \(K(\Psi_0)\) and this value is given by

\[
K(\Psi_0^*) = \frac{1 - (1 - \mu^2)}{1 - (1 - \mu^2)^2}
\]

(4.196)

An optimal value for the spectral radius \(S_{m_1, \omega}\) is obtained as

\[
S_{m_1, \omega} = \frac{1 - \mu^2}{1 - \mu^2}
\]

(4.197)

Proof

By virtue of Theorem 4.3.7 and Lemma 4.3.2, (4.194) and (4.195) can be established. It is obvious that \(K(\Psi_0^*) > K(\Psi_0^*)\) for the interval \(\omega^*(\bar{\mu}) < \omega < 2\).

Next, we show that if \(0 < \omega < \omega^*(\bar{\mu})\), then \(K(\Psi_0)\) is a decreasing function of \(\omega\).

We have that

\[
\text{sgn} \left\{ a \frac{2-\omega u^2 + (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])}{2-\omega u^2 - (\bar{\mu}^2 [2 \omega^2 - 4(\omega - 1)])} \right\} = \text{sgn} (\bar{\mu} - 1)
\]

and
and

\[ \text{sgn}\left\{ \frac{2-\omega \mu^2}{2-\omega \mu^2 - \left( \frac{2-\mu^2}{\mu^2 \left( 2-\mu^2 - 4(\omega-1) \right)} \right)^2} \right\} \]

\[ = \text{sgn}\left\{ -\mu^2 \left( 2-\omega \mu^2 - \left( \frac{2-\mu^2}{\mu^2 \left( 2-\mu^2 - 4(\omega-1) \right)} \right)^2 \right\} \right\} \]

where \( 2-\omega \mu^2 > 0 \) for all \( \omega \) in the range \( 0 < \omega < \omega^* \).

We have therefore from (4.194) and (4.195) that in the range \( 0 < \omega < \omega^* \), \( K(\omega) \) is a decreasing function of \( \omega \) and its value is minimised when \( \omega \) takes the value (4.195a). Using either (4.194) or (4.195a) we have for this value of \( \omega \) that \( K(\omega) \) is given by

\[ K(\omega^*) = \frac{2-\mu^2 \omega^*}{2-\mu^2 \omega^*} \]

and using (4.144) we have (4.196). By using (4.139), we obtain (4.197) as the minimal value for \( S(\lambda, L_1, \omega) \). The expression (4.197) for \( S(\lambda, L_1, \omega) \) suggests a dependence on \( \mu \). It is therefore apparent that in order to achieve optimum results for the MSOR scheme, a knowledge of \( \mu \) would be necessary. The added effort in the computation of \( \mu \) would suggest an extra work when compared to the SOR scheme.

For general systems where the matrix \( A \) is not necessarily consistently ordered, the MSOR theory is expected to hold approximately.

4.3.5 Numerical Results

We present in this section the results obtained for the model problem discussed in Section 4.1.3 for various mesh sizes \( h^{-1} = 20, 40, 60 \).
FIGURE 4.4a: Comparison of the MSOR, SOR and SSOR iterative schemes for a model problem
4.4 THE MODIFIED GAUSS-SEIDEL (MGS) METHOD

Let \( A \) be a non-singular sparse matrix and we seek iteratively the solution of the linear system
\[
Ax = b,
\]
where \( b \) is given. We shall assume that the diagonal entries of \( A \) are non-zero.

Let us assume a splitting of \( A \) such that
\[
A = I - L - U
\]
where \( L \) and \( U \) are respectively strictly lower and upper triangular parts of \( A \). We can express (4.202) using (4.203) as
\[
(I-L)x^{(n+1)} = Ux^{(n)} + b,
\]
and since \( (I-L) \) is non-singular and lower triangular then
\[
x^{(n+1)} = (I-L)^{-1}(Ux^{(n)} + b)
\]
For this method \( \omega = \tau = 1 \) (see Table 4.1). We can express (4.205) as
\[
x^{(n+1)} = m_{1,1}x^{(n)} + c
\]
where
\[
m_{1,1} = (I+L+L^2)U
\]
and
\[
c = (I+L+L^2)b
\]
For the convergence of the MGS method, we now prove the following:

**Theorem 4.4.1**

Let \( A \) be a consistently ordered matrix with non-zero diagonal entries and let \( B = I - D^{-1}A \) has real eigenvalues then
\[
S_{m_{1,1}} \leq S_{m_{1,1}} < 1
\]
if and only iff
\[
\bar{\mu} = S(B) < 1.
\]

**Proof**

Let the preconditioned matrix of the MGS method be represented by
\[
\psi = (I+L+L^2)U
\]
and if \( \overline{\lambda} \) and \( \overline{\mu} \) are the eigenvalues of \( \psi \) and \( B \) respectively, then the eigenvalue relationship
\[
\bar{\xi} = 1 - \mu^2 \quad (4.211)
\]
can be shown to hold (see Young ([38], p. 143)).

Since \( B \) is positive definite, then

\[
0 \leq \mu^2 \leq \bar{\mu}^2 \quad (4.212)
\]

Using (4.211), we have that

\[
\max_{0 \leq \mu^2 \leq \bar{\mu}^2} \{ \xi \} = \bar{\xi} = 1 \quad (4.213)
\]

and

\[
\min_{0 \leq \mu^2 \leq \bar{\mu}^2} \{ \xi \} = \underline{\xi} = 1 - \bar{\mu}^2 \quad (4.214)
\]

Since we have that \( \bar{\xi} > 0 \), then the MGS scheme will converge if and only if

\[
\underline{\xi} > 0 \quad (4.125)
\]

and so we have

\[
\underline{\xi} = (1 - \bar{\mu}^2) > 0
\]

which implies that \( \bar{\mu} = S(B) < 1 \).

This completes the proof of the theorem.

Conversely, we can prove,

**Theorem 4.4.2**

Given that \( A \) is consistently ordered with non-vanishing diagonal elements such that \( B \) has real eigenvalues with

\[
\bar{\mu} = S(B) < 1
\]

then

\[
S(L_{1,1}) = \frac{S(L_{1,1})}{m_{1,1}} < 1.
\]

**Proof**

We obtain an expression for P-condition number \( P(\psi) \) of \( \psi \) in terms of the eigenvalues \( \mu \) of \( B \) using (4.213) and (4.214). Thus

\[
P(\psi) = \frac{\bar{\xi}}{\underline{\xi}} = \frac{1}{1 - \bar{\mu}^2} \quad (4.216)
\]

The spectral radius \( S(\frac{L_{1,1}}{m_{1,1}}) \) of the MGS scheme can be expressed in the well known form Evans [11].
Using (4.126) and (4.127) we have

\[ S(\mathbf{L}_{1,1}^{1,1}) = \frac{P(\psi) - 1}{P(\psi) + 1} \tag{4.127} \]

Using (4.126) and (4.127) we have

\[ S(\mathbf{L}_{1,1}^{1,1}) = \frac{1}{1 - \mu^2} - 1 \]
\[ = \frac{1}{1 - \mu^2} + 1 \]
\[ = \frac{\mu^2}{2 - \mu^2} < 1, \]

since \( \mu < 1 \). Thus the theorem is proved.

4.4.1 Numerical Results

We now present the results for a model problem discussed in Section 4.1.3 with grid sizes \( h^{-1} = 20, 40, 60 \).
FIGURE 4.4: Comparison of the MGS, SOR and SSOR iterative schemes for a model problem.
4.5 GENERALISED PRECONDITIONED CONJUGATE GRADIENT METHOD

In the introductory Section 1.6 we showed how the CG method can be taken as an iterative procedure similar to the SI method. Let us now apply the concept, in more detail with a view to applying the CG method to the MPSD and thus produce a very powerful iterative scheme.

Consider the iterative scheme

$$u^{(n+1)} = Gu^{(n)} + k$$

where

$$k = (I-G)A^{-1}b,$$

with $A=I-L-L^T$.

We shall assume that the amplification matrix $G$ has the form

$$G = I-R^{-1}A$$

where $R$ the conditioning matrix is the product of a matrix and its transpose $R$ is defined by its explicit form as given earlier, i.e.,

$$R = (I+\omega L)(I+\omega L)^T$$

where $(I+\omega L)$ is non-singular.

We can develop a version of CG procedure with respect to the iterative scheme (4.128) in a manner similar to that followed for the SI method. We shall adopt a preconditioning concept similar to that originally introduced in Evans [10]. From (4.128) we have

$$(I+\omega L)A(I+\omega L)^T[(I+\omega L)^T]^{-1}u = (I+\omega L)b$$

or

$$\hat{A}\hat{u} = \hat{b}$$

where

$$\hat{A} = (I+\omega L)A(I+\omega L)^T$$

$$\hat{u} = [(I+\omega L)^T]^{-1}u = (I+\omega L)^T u$$

$$\hat{b} = (I+\omega L)b.$$
If we use the non-stationary degree version developed in Section 1.6, then by (1.146) we have the iterative scheme

\[ u(n+1) = \rho_{n+1}(u(n) + \gamma_{n+1} \hat{r}(n)) + (1-\rho_{n+1})\hat{u}(n-1) \quad (4.134) \]

where

\[ \hat{r}(n) = b-A\hat{u}(n) = (I+wL)r(n) \quad (4.135) \]

and

\[ r(n) = b-Au(n). \quad (4.136) \]

Using (4.132) and (4.135), (4.134) becomes,

\[ u(n+1) = \rho_{n+1}(u(n) + \gamma_{n+1}(I+wL)^T(I+wL)r(n) + (1-\rho_{n+1})u(n-1) \quad (4.137) \]

If we observe that

\[ (I+wL)^T(I+wL)r(n) = (I+wL)^T(I+wL)(b-Au(n) \]

then we have the compact form,

\[ u(n+1) = \rho_{n+1}(u(n) + \gamma_{n+1}u(n) + (1-\gamma_{n+1})u(n-1) + (1-\rho_{n+1})u(n-1) \quad (4.139) \]

Now, if we consider the expressions for the parameters \( \rho_{n+1} \) and \( \gamma_{n+1} \) we have from (1.148) that

\[ \gamma_{n+1} = \frac{(\hat{r}(n), \hat{r}(n))}{(\hat{r}(n), A\hat{r}(n))} \]

\[ = \frac{(\hat{r}(n), (I+wL)A(I+wL)^T\hat{r}(n))}{(\hat{r}(n), A(I+wL)^T\hat{r}(n))} \]

\[ = \frac{(\hat{r}(n), \hat{r}(n))}{(I+wL)^T\hat{r}(n), A(I+wL)^T\hat{r}(n)} \]

\[ = \frac{(\hat{s}(n), \hat{s}(n))}{(\hat{s}(n), A\hat{s}(n))} \quad (4.138) \]

where

\[ \hat{s}(n) = (I+wL)^T\hat{r}(n) = (I+wL)^Tr(n) \]

\[ = Gu(n) + k - u(n). \quad (4.139) \]

Finally from (1.147) we have the following expression for \( \rho_{n+1} \)

\[ \rho_{n+1} = \left[ \frac{\gamma_{n+1}}{\gamma_n} \frac{(\hat{r}(n), \hat{r}(n))}{(\hat{r}(n-1), \hat{r}(n-1))} \frac{1}{\rho_n} \right]^{-1} , \quad n=1,2,... \quad (4.140) \]

The relations (4.137), (4.138) and (4.140) define the preconditioned CG method with respect to the iterative scheme (4.128).
4.5.1 The MPSD Semi-iterative (MPSD-SI) Method

Let us now consider the use of the CG method developed thus far and employ it to accelerate the Modified Preconditioned Displacement (MPSD) method.

The basic problem is to obtain a solution to the system of form

$$ Au = b \quad (4.141) $$

where A is assumed to be a sparse, real and non-singular matrix with non-zero diagonal entries and also has the splitting

$$ A = I - L - U $$

By virtue of the analysis in the previous sections, the iterative scheme (4.128) can be written in the form

$$ u^{(n+1)} = D_{\tau, \omega} u^{(n)} + \delta_{\tau, \omega} \hspace{0.5cm} (4.142) $$

where

$$ D_{\tau, \omega} = I - \tau B = I - \tau (I + \omega U) (I + \omega L) D^{-1} b \hspace{0.5cm} (4.143) $$

and

$$ \delta_{\tau, \omega} = \tau n \omega = \tau (I + \omega U) (I + \omega L) D^{-1} b \hspace{0.5cm} (4.144) $$

We can now express (4.142) as a semi-iterative process based on the conjugate gradient method. Thus,

$$ u^{(n+1)} = (1 - \rho_{n+1}) u^{(n-1)} + \rho_{n+1} (D_{\rho, \omega} u^{(n)} + \delta_{\rho, \omega}) \hspace{0.5cm} (4.145) $$

where

$$ \rho = 2 / [\xi_{\text{max}} (B) + \xi_{\text{min}} (B)] $$

$$ \rho_1 = 1, \hspace{0.5cm} \rho_2 = (1 - \sigma^2/2)^{-1} $$

$$ \rho_{n+1} = (1 - \sigma \rho_n/4)^{-1}, \hspace{0.5cm} n = 2, 3, 4, ... \hspace{0.5cm} (4.146) $$

and

$$ \sigma = [P(B) - 1] / [P(B) + 1] \hspace{0.5cm} (4.147) $$

and P(B) denotes the P-condition number of B.\omega,

$$ 0 < \xi_{\text{min}} (B) \leq \xi \leq \xi_{\text{max}} (B) \hspace{0.5cm} (4.148) $$

and \( \xi \) denotes an eigenvalue of \( B \).

Since (4.145) is independent of \( \tau \), it also represents the Modified Preconditioned Jacobi-SI method (MPJ-SI method) and can alternatively be written as
where

\[ u^{(n+1)} = \rho_{n+1} [\mathcal{K} u^{(n)} + \eta_\omega ] + (1-\bar{\rho}) u^{(n)}] + (1-\rho_{n+1}) u^{n-1} \]  

(4.149)

\[ \mathcal{K} = D_{m,\omega} = I - B \]  

(4.150)

In order to apply the accelerated procedure (4.145) or (4.149) we needed to have the optimum values of parameters \( \omega_0 \) and \( \rho(\omega_0) \). Since some times the computation of optimal of the parameters could require as many iterations as the solution of (4.141) itself, we now seek an adaptive approach which simultaneously computes and updates the parameter values while the iterative procedure progresses.
4.6 Dynamic Acceleration of the Preconditioned Simultaneous Displacement (MPSD) Method

In this section we shall outline a procedure for an algorithm which dynamically or adaptively determines a sequence of the required acceleration parameters and our approach will follow the lines proposed by Missirlis and Evans [24b], Benokraitis [3b] and Hageman [21a].

The Adaptive Algorithm

We now define an algorithm which employs the MPSD-SI method and simultaneously updates the values of $\omega$ and $P(B)$. By virtue of Theorem 4.3 which gives an upper bound $\phi(\omega, v)$ for the smallest eigenvalue $\xi_{\min}(B)$ we deduce that we can estimate a lower bound for the P-condition number $P(B)$ from the relationship

$$p(\omega, v) \leq P(B) \omega$$

where

$$p(\omega, v) = \frac{1}{\omega(2-\omega)\phi(\omega, v)}$$

(4.151)

(4.152)

The lower bound $p(\omega, v)$ of $P(B)$ defined by (4.152) suggests that it should be possible to approximate $P(B)$ using the MPSD-SI method. In this section we shall re-designate the quantities $\hat{a}(v)$ and $\hat{b}(v)$ in (4.30) as $a$ and $\beta$ respectively and in order to clarify their role we shall examine the behaviour of $\phi(\omega, v)$ with respect to these quantities. From (4.40) and (4.45), we have that

$$\phi(\omega, v) = \frac{1-\omega+3\omega^2}{1-\alpha} = \omega + \frac{1-\omega+3\omega^2}{1-\alpha}$$

(4.153)

and we can thus construct the Table 4.5.1.

<table>
<thead>
<tr>
<th>$\beta$-Domain</th>
<th>$\omega$-Domain</th>
<th>$3\omega^2 \beta - \omega^2 + 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta \leq \frac{\omega}{2}$</td>
<td>$0 &lt; \omega \leq \omega^*$</td>
<td>$&gt;0$</td>
</tr>
<tr>
<td></td>
<td>$\omega = \omega^*$</td>
<td>$=0$</td>
</tr>
<tr>
<td></td>
<td>$\omega &lt; \omega^* \leq \omega$</td>
<td>$&lt;0$</td>
</tr>
<tr>
<td>$\beta &gt; \frac{\omega}{2}$</td>
<td>$0 &lt; \omega &lt; 2$</td>
<td>$&gt;0$</td>
</tr>
</tbody>
</table>

**Table 4.5.1:** The behaviour of $3\omega^2 \beta - \omega^2 + 1$ as a function of $\omega$
where
\[ \omega^* = \frac{2}{1+\sqrt{1-12\beta}} \]  
(4.154)

Now (4.152), (4.153) and Table 4.5.1 reveals that for \( \beta \leq 1/12 \) we have,

(i) if \( \omega \leq \omega^* \), then \( p(\omega, v) \) is maximised,

(ii) if \( \omega = \omega^* \), then \( p(\omega, v) = \frac{1}{2 - \omega^*} \) and,

(iii) if \( \omega \geq \omega^* \), then \( p(\omega, v) \) is maximised when \( \alpha \) is maximised for \( 0 < \omega < 2 \).

Also, if we maximise \( \phi(\omega, v) \) then we have an approximation to \( P(B, \omega) \) which can be minimised with respect to \( \omega \).

Next, let us assume that an approximation or an initial guess \( v = v^{(1)} \neq 0 \) to the eigenvector of \( \mu_{\min}(B, \omega) \). If \( A \) is an L-matrix, the quantity \( \alpha \) is maximised for \( \nu \geq 0 \) (i.e. all the components of \( v \) are non-negative). Thus \( \beta \leq 1/12 \) and \( \omega \leq \omega^* \) or \( \beta > 1/12 \), where \( p(\omega, v) \) is maximised if \( \alpha \) is maximised, we always let \( v \) have the form
\[ v = v^{(1)} = (v^{(1)}_1, v^{(1)}_2, v^{(1)}_3, \ldots, v^{(1)}_n)^T \]
where \( v^{(1)}_1 \geq 0 \). This choice of \( v \) gives
\[ \alpha_1 = \frac{(v^{(1)}^T, Bv^{(1)})}{(v^{(1)}_1, v^{(1)}_1)} \geq 0 \]  
(4.155)

if \( A \) is an L-matrix. On the other hand, it is required that the quantity \( \alpha \) be minimised (if \( \omega^* \leq \omega \) and \( \beta \leq 1/12 \)) which can be achieved if one chooses \( v \) to have the alternative form \( v = v^{(2)} = (v^{(2)}_1, v^{(2)}_2, v^{(2)}_3, \ldots, v^{(2)}_n)^T \), where
\[ v^{(2)}_k = \begin{cases} v^{(1)}_k, & \text{on even points} \\ -v^{(1)}_k, & \text{on odd points} \end{cases} \]

The above choice of \( v \) gives
\[ \alpha_2 = \frac{(v^{(2)}_1, Bv^{(2)}_1)}{(v^{(2)}_1, v^{(2)}_1)} \leq 0 \]  
(4.156)

which tends to maximise \( p(\omega, v) \) if \( \omega^* \leq \omega \) and \( \beta \geq 1/12 \).

Finally, we see that if \( A \) is an L-matrix with Property A, then
\[ \beta = \frac{(v^{(1)}_1, Luv^{(1)}_1)}{(v^{(1)}_1, v^{(1)}_1)} = \frac{(v^{(2)}_1, Luv^{(2)}_1)}{(v^{(2)}_1, v^{(2)}_1)} = \beta_2 \]  
(4.157)
Consequently, a lower bound on \( P(B) \) is given by the following expression

\[
P(B) \geq \begin{cases} 
\frac{1-M}{1-M+3\omega^2} & \text{if } \omega \leq \omega^* \\
\frac{1-M}{1-M+3\omega^2} & \text{if } \omega > \omega^*
\end{cases}
\]  

where

\[
M = \alpha_1 > 0 \\
m = \alpha_2 < 0 \\
\beta = \beta_1 = \beta_2
\]

and

\[
\omega_1 = \begin{cases} 
\frac{2}{1+\sqrt{1-2M+12\beta}} = \omega_M^* & \text{if } M < 12\beta \\
\frac{2}{1+\sqrt{1-2\omega}} = \omega^* M > 12\beta,
\end{cases}
\]

while the corresponding bound on \( P(B) \) is given by

\[
P(B) \geq \begin{cases} 
\frac{\omega_1(1+\sqrt{1-2M+12\beta})}{1-M} & \text{if } M < 12\beta \\
\frac{1}{2-M^*} & \text{if } M > 12\beta.
\end{cases}
\]

Thus, we have a good choice of the preconditioning parameter \( \omega \) in the sense of minimising the bound (4.158) given by

\[
\omega_1 = \frac{2}{1+\sqrt{1-2M+12\beta}}
\]

A sequential description of the adaptive algorithm is given in Section 4.6.1.

### 4.6.1 Algorithm 4.6.1

1. Select an initial approximation \( u^{(0)} \) such that \( \|u^{(0)} - u\|_A \leq \epsilon \) and choose a convergence tolerance \( \epsilon \). Also, let

\[
v = (v_1, v_2, v_3, \ldots, v_n)^T = (1, 1, 1, \ldots)^T
\]

and

\[
\hat{v} = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n)^T
\]

where

\[
\hat{v}_k = \begin{cases} 
v_k & \text{on even points} \\
-v_k & \text{on odd points}
\end{cases}
\]

Set \( i = 1 \).

2. For the latest two vectors, \( v, \hat{v} \) compute

\[
\alpha_i = \frac{(v, Bv)}{(v, v)}, \quad \alpha_{i+1} = \frac{(\hat{v}, B\hat{v})}{(\hat{v}, \hat{v})}
\]

\[
\beta_i = \frac{(v, Luv)}{(v, v)}, \quad \beta_{i+1} = \frac{(\hat{v}, LU\hat{v})}{(\hat{v}, \hat{v})}
\]
if they have not been previously computed.

(3) Use a Fibonacci search technique to determine \( \omega \) by minimising the function

\[
P(\omega) = \max_i \left\{ \frac{1-\omega a_i + 3\omega^2 b_i}{(\omega(2-\omega)(1-a_i))} \right\}
\]

for all pairs \((a_i, b_i)\). Moreover, compute the corresponding value \( P_E(B) \) from the expression

\[
P_E(B) = P(\omega)
\]

(4) Select \( q \) to be the least integer \( q \) which satisfies the inequality

\[
\frac{1}{n} \log \frac{2^n}{1+\bar{r}} \geq -0.9 \log \bar{r},
\]

where

\[
\bar{r} = \frac{\sqrt{P(E)} - 1}{\sqrt{P(E)} + 1}
\]

(5) Iterate \( q \) times with the MPSD-SI method using the latest parameters \( \omega \) and \( P_E(B) \). After each iteration, check for convergence by computing the residual vector

\[
\delta^{(n)} = (I+\omega U)(I+\omega L)(b-Au^{(n)}), \quad n < q
\]

and testing the stopping procedure whether or not

\[
\frac{P_E||B||||\delta^{(n)}||_A}{||u^{(n)}||_A} \leq \varepsilon
\]

If (4.168) is satisfied, terminate the algorithm, otherwise continue to the next step.

(6) In this step we test whether to update the parameters \( \omega P_E(B) \) or not. From the previous step we have obtained the pseudo-residual vector \( \delta^{(n)} \), thus we now compute

\[
a = \frac{n}{(\delta q, B\delta q)^n}, \quad \beta = \frac{n}{(\delta q, L\delta q)^n}
\]

and

\[
P = \frac{1-\omega a + 3\omega^2 b}{\omega(2-\omega)(1-a)}
\]
If the following inequality is satisfied

\[ P \leq P_{E}(B) \tag{4.171} \]

then go to step 5 and note that the next MPSD-SI iteration can be computed from

\[ u^{(n+1)} = (1-P_{n+1})u^{(n-1)} + P_{n+1}(u^{(n)} + \rho \delta^{(n)}) \tag{4.172} \]

where \( \delta^{(n)} \) has already been computed in step 5. Otherwise, continue to the next step before altering the parameters.

(7) In order not to waste the computational work for the determination of \( \delta^{(n)} \) in (4.167), apply an MPSD-SI iteration using (4.172) with the old parameters \( P_{E}(B) \). Also, let

\[ v = \delta^{(n)} = (\delta_{1}^{(n)}, \delta_{2}^{(n)}, \ldots, \delta_{n}^{(n)})^{T} \]

and

\[ \hat{\nu} = (\hat{\nu}_{1}, \hat{\nu}_{2}, \ldots, \hat{\nu}_{n})^{T}, \]

where

\[ \hat{\nu}_{k} = \begin{cases} 
\delta_{k}^{(n)}, & \text{on even points} \\
-\delta_{k}^{(n)}, & \text{on odd points}
\end{cases} \]

Then set \( i = i + 2 \) and go to step 2 to compute new quantities \( \alpha \) and \( \beta \) in order to update \( P_{E}(B) \). In step 2 however, \( \alpha_{i} \) and \( \beta_{i} \) have already been computed from (4.169) in step 6.
4.7 BLOCK PRECONDITIONED METHODS

4.7.1 Introduction

In the earlier chapters we had considered various iterative schemes which entailed the determination of each component of \( u^{(n)} \) by using already computed approximate values of the other unknowns. These schemes are called point methods in order to be distinguished from the group iterative methods. In the latter case, we assign the equations to groups and then solve the group equations for the corresponding unknowns \( u_i \), regarding the other values of \( u_j \) as unknown.

A special case of grouping is a partitioning where some integers \( n_1, n_2, \ldots, n_q \) satisfying \( 1 \leq n_1 < n_2 < \ldots < n_q = N \) the equations for \( i=1,2,\ldots,n_1 \) belong to the first group, those for \( i=n_1+1,n_1+2,\ldots,n_2 \) belong to the second group, etc. The methods which are based on partitioning are known as block methods. Early work on block methods could be found in [29b] and [19a]. The first generalised S.O.R. block method is treated in [la] while the analysis of its convergence rate is given in [15a]. In Varga [30a] it was also shown that the rate of convergence of the two-line S.O.R. method with optimum \( \omega \) is approximately twice that of the point S.O.R. while Parter [27a] showed that the \( k \)-line S.O.R. method with optimum \( \omega \) converges approximately \((2k)^{1/2}\) as fast as the point S.O.R. In [14a],[14b], the line SSOR for the five-point discrete Dirichlet problem was shown to converge faster than the point SSOR method.

In this section we shall extend the preconditioning techniques we have developed in the previous chapter and show how it could be applied to construct and develop the corresponding group methods similar to the iterative procedures we had discussed earlier. First we give a few basic definitions and a theorem.
Definition 4.7.1

An ordered grouping \( \pi \) of \( w = \{1, 2, \ldots, N\} \) is a subdivision of \( w \) into disjoint subsets \( R_1, R_2, \ldots, R_q \) such that \( R_1 + R_2 + \ldots + R_q = w \).

We let \( \pi \) denote the ordered grouping defined by \( R_k = \{k\} \), \( k = 1, 2, \ldots, N \).

Given a matrix \( A \) and an ordered grouping \( \pi \) we define the submatrices \( A_{r,s} \) for \( r, s = 1, 2, \ldots, q \) by deleting from \( A \) all rows except those corresponding to \( R_r \) and all columns except those corresponding to \( R_s \). We can now generalise the concepts of Property A and consistently ordered matrices.

Given a matrix \( A \) and an ordered grouping \( \pi \), with \( q \) groups, we define the \( q \times q \) matrix \( Z = (z_{r,s}) \) by

\[
Z_{r,s} = \begin{cases} 
0, & \text{if } A_{r,s} = 0 \\
1, & \text{if } A_{r,s} \neq 0
\end{cases}
\]

(4.173)

Definition 4.7.2

The matrix \( A \) has property A \((\pi)\) if \( Z \) has property A.

Definition 4.7.3

The matrix \( A \) is a \( \pi \)-consistently ordered matrix (\( \pi \)-CO-matrix) if \( Z \) is consistently ordered.

Definition 4.7.4

A matrix is a generalised \( \pi \)-consistently ordered matrix (\( \pi \)-GCO-matrix) if

\[
A = \det(A_C^{(\pi)} + \alpha I_C^{(\pi)} - kD^{(\pi)})
\]

(4.174)

where

\[
A_C^{(\pi)} = D^{(\pi)} - C_L^{(\pi)} - C_U^{(\pi)}
\]

is independent of \( \alpha \) for all \( \alpha \neq 0 \) and for all \( k \).

Here \( D^{(\pi)} \) is the matrix formed from \( A \) by replacing with zero all \( a_{i,j} \) unless \( i \) and \( j \) belong to the same group, whereas \( C_L^{(\pi)} \) and \( C_U^{(\pi)} \) are formed from \( A \) by replacing all elements of \( A \) by zero except those \( a_{i,j} \) such that \( i \) and \( j \) belong to different groups and such that the group containing \( i \) comes after and before, respectively the groups containing \( j \).
Theorem 4.7.1 (Arms, Gates and Zondek [10])

If \( A \) is a \( \pi \)-GCO-matrix such that \( D(\pi) \) is non-singular, then the conclusions of Theorem 4.3.7 are valid if we replace \( B \) by \( B(\pi) \) and \( L_\omega \) by \( L(\pi)_\omega \).

From the above analysis we note that the definition of the group methods is based on the splitting (4.174). Following the previous analysis of the preconditioning techniques we can regard (4.174) as another splitting of \( A \) and in an analogous way we can develop the group versions of the preconditioned methods defined in Chapter 4.

If we let the conditioning matrix have the form

\[
R = D(\pi)
\]  

for any ordered grouping \( \pi \), then we define the group SD method using (4.5) and (4.175) by

\[
u(n+1) = u(n) + \tau (D(\pi))^{-1}(b - Au(n))
\]  

where \( \tau \neq 0 \) is a real parameter and \( D(\pi) \) is a non-singular matrix. We therefore see that the rate of convergence of the group SD method depends upon the grouping \( \pi \), since if \( D(\pi) = A \), then we solve our system immediately. On the other hand, the inversion of \( D(\pi) \) by direct methods is a limit to the above observation [8b].

4.7.2 Group MPSD Methods

For any ordered grouping \( \pi \), we let the preconditioned matrix have the form

\[
R^{-1} = (D(\pi))^{-1}(I + \omega U(\pi))(I + \omega L(\pi))
\]  

and thus we define the group PSD method by

\[
u(n+1) = u(n) + \tau (I + \omega U(\pi))(I + \omega L(\pi))(D(\pi))^{-1}(b - Au(n))
\]  

where \( \tau, \omega \) are real parameters as previously defined.

The iterative scheme (4.178) can be replaced by the computable form
\[ D'(\pi)(n+1) = \omega C_L(\pi)(n+1) + r(n) \]
\[ D(\pi)(n+1) = \omega C_U(\pi)(n+1) + D(\pi)(n+1) \]
\[ u(n+1) = u(n) + r(\pi)(n+1) \]  
(4.179)

where \[ r(n) = b - Au(n) \].

Let us write (4.179) in the form
\[ u(n+1) + D(\pi)u(n) + \delta(\pi) \]  
(4.180)

where
\[ D(\pi) = I - \tau(I + \omega U(\pi))(I + \omega L(\pi))(D(\pi))^{-1}A \]  
(4.181)

and
\[ \delta(\pi) = \tau(I + \omega U(\pi))(I + \omega L(\pi))b. \]  
(4.182)

By (4.181) and (4.182) we see that the group PSD method is completely consistent if \( D(\pi) \) is non-singular and \( \tau \neq 0 \). Most of the analysis given for the point methods in the previous sections can also apply to the group MPSD methods. Let us consider the determination of the spectral radius of the MPSD method applied to a system derived from \( Au = b \).

Theorem 4.7.2

If \( A \) is a symmetric and positive definite matrix and if \( A^* \) is obtained from \( A \) by deleting rows and the corresponding columns of \( A \), then,
\[ P(B^*_\omega) < P(B^*_\omega) \]  
(4.183)

where
\[ B^*_\omega = (I + \omega U)(I + \omega L)D_A^{-1}. \]  
(4.184)

Proof

Let \( \xi_{\min}(B^*_\omega) \) and \( \xi_{\max}(B^*_\omega) \) denote the smallest and largest eigenvalue of
\[ B^*_\omega = (I + \omega U)(I + \omega U)^{-1}A. \]  
(4.185)

From (4.10a) we have that \( B^*_\omega \) is similar to the symmetric matrix
\[ B^*_\omega = D_A^{-1}(D + \omega U \omega L)A(D + \omega U \omega L)^{-1} \]  
(4.186)

If now, \( v_* \) is an eigenvector associated with \( \xi_{\min}(B^*_\omega) \), then
\[ \xi_{\min}(B^*_\omega) = \frac{(v_*^*, B^*_\omega v_*^*)}{(v_*^*, v_*^*)} = \frac{(w_*^*, Aw_*^*)}{(v_*^*, v_*^*)} \]  
(4.187)
where  \[ w^* = (D_* + \omega U_*) D_* v_* \neq 0 \] (4.188)

or  \[ D_* v_* = (D_* + \omega U_*) D_* v_* \] (4.189)

Next, we augment \( w_* \) with zero components (at the positions which were deleted from \( A \) to form \( A_* \)) to form \( w \) and \( v \) such that
\[ D^t v = (D + \omega U) w . \] (4.190)

Certainly, from the definitions of \( w \) and \( v \) we have
\[ \lambda(B_{w_*}) = \frac{(w_* , A_* w_*)}{(v_* , v_*)} \geq \frac{(w_* , A_* w_*)}{(v_* , v_*)} \leq \frac{(w_* , A_* w_*)}{(v_* , v_*)} = \xi_{\min}(B_{w_*}) \] (4.191)

since
\[ (w_* , A_* w_*) = (w_* , A_* w_*) \]
and the effect of the added rows and columns in \( A \) is annihilated by zero components of \( w \). Also by the definition of \( w \), the right hand member of (4.190) has identical components as the right hand side of (4.189), plus additional ones. Now since \( D^t \) is diagonal, the components of \( v \) are identical as the components of \( v_* \) plus additional ones. If \( \xi_{\min}(B_{w}) \)
denotes the smallest eigenvalue of \( B_{w} \), then we have (see Theorem 4.1.3)
\[ \xi_{\min}(B_{w}) \leq \frac{v_B v}{(v,v)} = \frac{(w_* , A_* w_*)}{(v_* , v_*)} \leq \frac{(w_* , A_* w_*)}{(v_* , v_*)} = \xi_{\min}(B_{w}) \] (4.192)

Similarly, we can establish that
\[ \xi_{\max}(B_{w}) \leq \xi_{\max}(B_{w}) \] (4.193)
where \( \xi_{\max}(B_{w}) \) and \( \xi_{\max}(B_{w}) \) denote the largest eigenvalues of \( B_{w} \) and \( B_{w} \), respectively.

Thus, if \( A \) is positive definite then (4.183) follows from (4.192) and (4.193) and the proof of the theorem is complete.

A similar result has been established in [14a]. The analysis for the determination of good estimates for \( \tau \), the preconditioning parameter \( \omega \) and the spectral radius \( D_{\tau,\omega}^{(\pi)} \) is similar to ones earlier developed for the MPSD method.
CHAPTER FIVE

MODIFIED ALTERNATING DIRECTION
PRECONDITIONING TECHNIQUES
5.1 INTRODUCTION

The present section is concerned with the numerical solution of linear systems of form

\[ Au = b, \]  

(I.1)

where \( A \) is assumed sparse and generated from the discretization by finite difference approach of boundary value problems of elliptic partial differential equations. The vector \( b \) is given and \( u \) is to be determined.

Current literature abounds with techniques and methods often utilized for the solution of such problems. Some of these methods include those described in [27],[30]:and [32].

A new form of solution by 'modified preconditioning' is presented and it is sequel to the original version introduced in Evans [10] and amplified in [14]. Two classical iterants, the Simultaneous Displacement method and Richardson's second degree method [15] will form the basis of this discussion.

We shall establish using the modified preconditioning approach that a significant reduction in the P-condition number of the original system coefficient matrix can be attained and this implies that the acceleration of the convergence rates of the basic iterants is possible. Subsequently, we shall outline our computational strategy and in the numerical section compare the modified preconditioning method with its preconditioned counterpart [14], the S.O.R. and the ADI-PR methods.
5.2 SELF ADJOINT SECOND ORDER ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

Consider the problem which seeks the solution of the differential equation defined by

$$\frac{KU}{E_2(x)E_1(y)} = \frac{1}{E_2(x)} \frac{\partial}{\partial x} (E_1(x) \frac{\partial u}{\partial x}) + \frac{1}{E_1(y)} (F_2(y) \frac{\partial u}{\partial y})$$  \hspace{1cm} (5.1)$$

for \((x,y) \in \mathbb{R}_q\), where \(U = u(x,y)\) and \(\mathbb{R}_q\) is a rectangular region such that \(0 < x < L_x\) and \(0 < y < L_y\).

The boundary conditions require that

\(U(x,y) = S(x,y), \quad (x,y) \in \mathbb{R}_q,\)

where \(S(x,y)\) is a specified boundary constant. The following conditions apply:

\(E_1(x), F_1(y), E_2(x), F_2(y) > 0\) for all \(K > 0\), for all \((x,y) \in \mathbb{R}\).

Discretization of the Problem

The discretization of the problem is normally achieved by imposing a regular square grid of size \(h\) over the domain \(\mathbb{R}_q\) such that

\[(n+1)\Delta x = L_x\]

and

\[(m+1)\Delta y = L_y.\]

The components of the equation (5.1) are then approximated by using central differences in the \(x\) and \(y\) directions by

$$\frac{\Delta x \Delta y}{E_2(x)} \frac{\partial}{\partial x} (E_1(x) \frac{\partial u}{\partial x}) = [B_u] (x,y)$$

$$= \frac{\Delta y}{E_2(x)} (E_1(x+h/2)u(x+h,y)-(E_1(x+h/2)u(x,y)+$$

\[+E_1(x-h/2))u(x,y)+E_1(x-h/2)u(x-h,y)) \hspace{1cm} (5.2)$$
\[
\frac{\Delta x \Delta y}{F_1(y)} \frac{\partial}{\partial x} \left( F_2(y) \frac{\partial u}{\partial y} \right) = [\mathcal{V}u](x,y)
\]

\[
= \frac{(\Delta x/\Delta y)}{F_1(y)} \left( F_2(y+h/2)u(x,y+h) - F_2(y+h/2)u(x,y) + F_2(y-h/2)u(x,y-h) \right)
\]

\[
u(x,y) + F_2(y-h/2)u(x,y-h)
\]

(5.3)

and

\[Q(x,y) = [\Sigma u](x,y) = \Delta x \Delta y \ K u(x,y).\] (5.4)

The given problem (5.1) has thus been decomposed into a system of finite difference equations which gave rise to a system of linear equations in \(u\) with \(H, V\) and \(L\) as shown.

\[Au = (H+V+L)u = (H_1 + V_1)u = b\] (5.5)

where \(H_1 = H + 4\Sigma\)

and \(V_1 = V + 4\Sigma\).

The matrices \(H\) and \(V\) are real symmetric \(n \times n\) matrices [30]. The matrix \(\Sigma\) is a diagonal matrix whose diagonal entries are non-negative. The matrices \(H\) and \(V\) are also Stieltjes matrices and by means of appropriate simultaneous row and column permutations they can be brought into tridiagonal form. We have by [5], that the matrices \(H_1\) and \(V_1\) are commutative.
5.3 THE MODIFIED ALTERNATING DIRECTION PRECONDITIONING METHOD

We can express the matrix $A$ in (3.5) in the form

$$ A = \hat{H} \hat{V} + I, $$

(5.6)

where

$$ \hat{H} = H - I, $$

(5.7)

and

$$ \hat{V} = V - I. $$

(5.8)

By premultiplying (5.6) by $(I + \omega \hat{H})^{-1}$ and using (5.5), we have

$$ (I + \omega \hat{H})^{-1} A (I + \omega \hat{V})^{-1} (I + \omega \hat{V}) u = (I + \omega \hat{H}) b. $$

(5.9)

Alternatively, (5.8) can be expressed as

$$ B \gamma = d, $$

(5.10)

where

$$ B_\omega = (I + \omega \hat{H})^{-1} A (I + \omega \hat{V})^{-1} $$

(5.11)

and

$$ \gamma = (I + \omega \hat{V}) u. $$

(5.12)

From (5.10), we have by expanding the Neumann series

$$ B_\omega = (I + \omega \hat{H} + \omega^2 \hat{H}^2 + \cdots) A (I - \omega \hat{V} + \omega^2 \hat{V}^2 + \cdots). $$

(5.13)

We can set

$$ -\omega \hat{H} = -\omega H + \omega^2 \hat{H}^2 $$

and

$$ -\omega \hat{V} = -\omega V + \omega^2 \hat{V}^2. $$

(5.14)

Also, (5.13) can take the form

$$ \hat{H} = \hat{H} - \omega \hat{H} $$

$$ \hat{V} = \hat{V} - \omega \hat{V}. $$

(5.15)

Using (5.14), (5.12) becomes

$$ B_\omega = [(I - (I - \omega \hat{H}) \omega \hat{H}) A [I - (I - \omega \hat{V}) \omega \hat{V}]] $$

$$ = [(I - (I + \omega \hat{H})^{-1} \omega \hat{H}) A [I - (I + \omega \hat{V})^{-1} \omega \hat{V}]]. $$

(5.16)

Expression (5.15) now takes the form

$$ B_\omega = (I + \omega \hat{H})^{-1} [1 + \omega \hat{H} - \omega \hat{H}] A [1 + \omega \hat{V} - \omega \hat{V}] (I + \omega \hat{V})^{-1} $$

$$ = (I + \omega \hat{H})^{-1} A (I + \omega \hat{V})^{-1}. $$

(5.17)

We can now use an analytical approach similar to those developed for the implicit schemes [30] and [32].
Now, we have that since \( A, \overline{H} \) and \( \overline{V} \) of (5.16) are symmetric and pairwise commutative, since \( A \) and \( v \) are symmetric and pairwise commutative, that by virtue of the theorem of Frobenius ([30], p.220), that they have a common basis of orthonormal eigenvectors.

We thus have that we can express the eigenvalues \( \lambda \) of \( B_\omega \) in (5.16) as

\[
\lambda = \frac{\overline{u} + \overline{v}}{(1-1/2\omega + \overline{u})(1-1/2\omega + \overline{v})}
\]

(5.17)

where \( \overline{u} \) and \( \overline{v} \) are respectively, the real positive eigenvalues of \( \overline{H}_1 \) and \( \overline{V}_1 \) given from (5.5) as

\[
\overline{H}_1 = \overline{H} + \frac{1}{2} \Sigma
\]

and

\[
\overline{V}_1 = \overline{V} + \frac{1}{2} \Sigma.
\]

(5.18)

It can be verified from (5.17) that \( \lambda > 0 \) as long as \( 0 < \omega \leq 2 \) for all \( h \).

Using ([32], p.9) we have that \( B_\omega \) is positive definite given that \( 0 < \omega \leq 2 \).

We have obtained (5.17) in the form developed for implicit methods [30].

The eigenvalue \( \overline{u} \) and \( \overline{v} \) used simply indicate that the modified preconditioning method has been derived from the implicit preconditioning version [14] by a slight translation of the origin of \( x \) and \( y \) in the negative direction and we designate this shifted origin by \( \tilde{x} \) and \( \tilde{y} \).
5.4 OBTAINING A MINIMAL ESTIMATE FOR THE P-CONDITION NUMBER OF B

By virtue of the definition of the P-condition number as the ratio of the maximal to the minimal eigenvalue, we need first to obtain optimal expressions for the minimal and maximal eigenvalues of $B_\omega$ before considering $P(B_\omega)$ in the relevant interval $0 \leq \omega \leq 2$.

**Minimal Eigenvalue of $B_\omega$**

We shall assume that the eigenvalues $\tilde{u}$ and $\tilde{v}$ have a common range, so that

$$\lambda_{\min}(\tilde{u}, \tilde{v}, \omega) = \min_{\theta_1 \leq u, v \leq \theta_2} \lambda(\tilde{u}, \tilde{v}, \omega)$$

and

$$\lambda_{\max}(\tilde{u}, \tilde{v}, \omega) = \max_{\theta_1 \leq u, v \leq \theta_2} \lambda(\tilde{u}, \tilde{v}, \omega)$$

(5.19)

Consider now the continuous function

$$Z(\tilde{x}, \tilde{y}, \omega) = \frac{\tilde{x} + \tilde{y}}{(1-\tilde{y}+\omega \tilde{x}) (1-\tilde{x}+\omega \tilde{y})}$$

(5.20)

for $\theta_1 \leq \tilde{x}, \tilde{y} \leq \theta_2$.

The expression (5.20) can take the form

$$Z(\tilde{x}, \tilde{y}, \omega) = P(\omega). (1-Q(\tilde{x}, \tilde{y}, \omega))$$

(5.21)

where

$$P(\omega) = \frac{1}{\omega (2-\omega)}$$

and

$$Q(\tilde{x}, \tilde{y}, \omega) = \frac{(1-\tilde{y}+\omega \tilde{x}) (1-\tilde{x}+\omega \tilde{y})}{(1-\tilde{y}+\omega \tilde{x})(1-\tilde{x}+\omega \tilde{y})}$$

(5.22)

If we set $r = \frac{1-\tilde{y}}{\omega}$ and $g(\tilde{x}, r) = \frac{\tilde{x}-\tilde{y}}{r+\tilde{y}}$ (for $r > 0$) in $Q(\tilde{x}, \tilde{y}, \omega)$ of (5.22), then (5.21) becomes,

$$Z(\tilde{x}, \tilde{y}, \omega) = P(\omega). (1-g(\tilde{x}, r).g(\tilde{y}, r)), \quad 0 < \omega < 2,$$

(5.24)

Obviously, $Z(\tilde{x}, \tilde{y}, \omega)$ is minimized when

$$|g(\tilde{x}, r)g(\tilde{y}, r)|$$

(5.24a)
in (5.24) has its minimum value. Thus,
\[ \min_{\tilde{x}, \tilde{y} \leq \theta_2} Z(\tilde{x}, \tilde{y}, \omega) = P(\omega) \cdot \{1 - \max_{\theta_1 \leq \tilde{x}, \tilde{y} \leq \theta_2} |g(\tilde{x}; \omega) g(\tilde{y}; \omega)| \} \]  
(5.25)
where \( \omega \in (0, 2) \).

From (5.25), we have
\[ \min_{\tilde{x}, \tilde{y} \leq \theta_2} Z(\tilde{x}, \tilde{y}, \omega) = P(\omega) \cdot \{1 - \max_{\theta_1 \leq \tilde{x} < \theta_2} \{|g(\tilde{x}; r)|\}^2 \} \]  
(5.26)

The right hand side of (5.26) is based on the assumption that \( \tilde{x}, \tilde{y} \) have a common range \( \theta_1, \theta_2 \).

Now, the derivative of \( g(\tilde{x}; r) \) in (5.26) with respect to \( x \) is negative for all \( \tilde{x} > 0 \) and so the maximum of
\[ |g(\tilde{x}; r)| \]
is given by one end of the interval \( (\theta_1, \theta_2) \). Thus
\[ \max_{\tilde{x} \leq \theta_2} |g(\tilde{x}; r)| = \max_{\tilde{x} \leq \theta_2} \left\{ \frac{r - \theta_1}{r + \theta_1}, \frac{r - \theta_2}{r + \theta_2} \right\} \]  
(5.27)

Using the results of ([30], p.215), we obtain
\[ \max_{\tilde{x} \leq \theta_2} |g(\tilde{x}; r)| = \begin{cases} \frac{\theta_2 - r}{\theta_2 + r}, & 0 < r \leq (\theta_1 \theta_2)^{1/2} \\ \frac{r - \theta_1}{r + \theta_1}, & r > (\theta_1 \theta_2)^{1/2} \end{cases} \]  
(5.28)

From (5.28), (5.26) and (5.25), we obtain
\[ \min_{\tilde{x}, \tilde{y} \leq \theta_2} Z(\tilde{x}, \tilde{y}, \omega) = \begin{cases} P(\omega) \cdot \{1 - \left(\frac{\theta_2 - r}{\theta_2 + r}\right)^2\}, & 0 < r \leq (\theta_1 \theta_2)^{1/2} \\ P(\omega) \cdot \{1 - \left(\frac{r - \theta_1}{r + \theta_1}\right)^2\}, & r > (\theta_1 \theta_2)^{1/2} \end{cases} \]  
(5.29)

By using (5.23) in (5.29) we obtain a final minimal expression for \( Z(\tilde{x}, \tilde{y}, \omega) \).

Thus,
\[ \min_{\tilde{x}, \tilde{y} \leq \theta_2} Z(\tilde{x}, \tilde{y}, \omega) = \begin{cases} 2\theta_2 \cdot (1 - \omega \theta_2)^{-2}, & 0 \leq \omega \leq \tilde{\omega}_1 \\ 2\theta_2 \cdot (1 - \omega \theta_2)^{-2}, & \omega \leq \tilde{\omega}_2 \end{cases} \]  
(5.30)

where
\[ \tilde{\omega} = 2 \cdot (1 + 2/\theta_1 \theta_2). \]  
(5.31)
Maximal Eigenvalue of B

For the preconditioned matrix $B$, we shall obtain expressions for the maximal eigenvalue as $\omega$ varies in the following intervals which are defined in terms of its eigenvalue bounds:

(i) $r \geq \theta_2$ or using (5.23), $0 < \omega \leq \frac{2}{1+2\theta_2}$

(ii) $1 \leq r \leq \theta_2$, or $\frac{2}{1+2\theta_2} < \omega \leq \frac{2}{1+2\theta_1}$

(iii) $0 < r < \theta_1$, or $\frac{2}{1+2\theta_1} < \omega < 2$

(5.32)

Consider first the interval (i) $r \geq \theta_2$; $0 < \omega \leq \frac{2}{1+2\theta_2}$. Using (5.24), we have,

$$
\max_{\tilde{X}, \tilde{Y} \leq \Theta_2} Z(\tilde{X}, \tilde{Y}, \omega) = F(\omega)(F+1)
$$

(5.33)

where

$$
F = \max_{\tilde{X}, \tilde{Y} \leq \Theta_2} (-g(\tilde{X}; r).g(\tilde{Y}; r))
$$

(5.34)

For the above interval, we have that $g(\tilde{X}; r)$ and $g(\tilde{Y}; r) \geq 0$ for all $\theta_1 \leq \tilde{X}, \tilde{Y} \leq \Theta_2$ since $r \geq \theta_2$. Thus, (5.34) assumes the form

$$
F = -(\min_{\theta_1 \leq \tilde{X} \leq \theta_2} g(\tilde{X}; r)).\min_{\theta_1 \leq \tilde{Y} \leq \theta_2} g(\tilde{Y}; r)
$$

(5.35)

$$
= (\min_{\theta_1 \leq \tilde{X} \leq \theta_2} g(\tilde{X}; r))^2
$$

(5.36)

The expression (5.36) is based on the fact that $\tilde{X}$ and $\tilde{Y}$ span a common interval.

Now sign $\frac{\partial}{\partial \tilde{X}} g(\tilde{X}; r)$ is negative for all $\tilde{X} \geq 0$ and thus

$$
\min_{\theta_1 \leq \tilde{X} \leq \theta_2} g(\tilde{X}; r) = g(\theta_2; r)
$$

(5.37)

and (5.36) therefore becomes

$$
F = -(g(\theta_2; r))^2
$$

Finally, from (5.33) we have

$$
\lambda_{\max}(\tilde{U}, \tilde{V}, \omega) = \max_{\tilde{X}, \tilde{Y} \leq \Theta_2} Z(\tilde{X}, \tilde{Y}, \omega)
$$

(5.38)
(ii) Next, we consider the interval

\[ \theta_1 \leq r \leq \theta_2 : \frac{2}{1+2\theta_1} \leq \omega \leq \frac{2}{1+2\theta_2} \]

For this interval, \( g(\widetilde{x};r) \) and \( g(\widetilde{y};r) \) takes positive and negative values for \( \theta_1 \leq \widetilde{x}, \widetilde{y} \leq \theta_2 \). We thus have from (5.33) that

\[
Z \max(\widetilde{x}, \widetilde{y}, \omega) = P(\omega) \left( 1 + \left( -\min_{\theta_1 \leq x \leq \theta_2} g(\widetilde{x};r) \cdot \max_{\theta_1 \leq y \leq \theta_2} g(\widetilde{y};r) \right) \right) \\
= P(\omega) \left( 1 - \max_{\theta_1 \leq x \leq \theta_2} g(\widetilde{x};r) \cdot \left( -\min_{\theta_1 \leq y \leq \theta_2} g(\widetilde{y};r) \right) \right) \tag{5.39}
\]

Since \( \widetilde{x} \) and \( \widetilde{y} \) lie in the same interval, their roles can be reversed. Thus, we have for this interval that

\[
\min_{\theta_1 \leq \widetilde{x} \leq \theta_2} g(\widetilde{x},r) = g(\theta_2;r) \\
\max_{\theta_1 \leq \widetilde{y} \leq \theta_2} g(\widetilde{y},r) = g(\theta_1;r) \tag{5.39a}
\]

Hence from (5.39), we obtain

\[
\max \lambda(\widetilde{u}, \widetilde{v}, \omega) = \max_{\theta_1 \leq \widetilde{x}, \widetilde{y} \leq \theta_2} Z(\widetilde{x}, \widetilde{y}, \omega) \\
= Z(\theta_2, \theta_1, \omega) = Z(\theta_1, \theta_2, \omega) \tag{5.40}
\]

Finally, we consider the interval.

(iii) \( \theta_1 \leq r \leq \theta_2 : \frac{2}{1+2\theta_1} \leq \omega \leq \frac{2}{1+2\theta_2} \).

In this interval, both \( g(\widetilde{x};r) \) and \( g(\widetilde{y};r) \) will be negative for all \( \theta_1 \leq \widetilde{x}, \widetilde{y} \leq \theta_2 \). We therefore have from (5.39) that

\[
Z \max(\widetilde{x}, \widetilde{y}, \omega) = P(\omega) \left( 1 + \left( -\max_{\theta_1 \leq x \leq \theta_2} g(\widetilde{x};r) \cdot \max_{\theta_1 \leq y \leq \theta_2} g(\widetilde{y};r) \right) \right) \\
= P(\omega) \left( 1 + \left( -\max_{\theta_1 \leq \widetilde{x} \leq \theta_2} g(\widetilde{x};r) \cdot \max_{\theta_1 \leq \widetilde{y} \leq \theta_2} g(\widetilde{y};r)^2 \right) \right)
\]

But since for this interval

\[
\max_{\theta_1 \leq \widetilde{x} \leq \theta_2} g(\widetilde{x};r) = g(\theta_1;r), \text{ then} \\
\lambda_{\max}(\widetilde{u}, \widetilde{v}, \omega) = Z_{\max}(\widetilde{x}, \widetilde{y}, \omega) = Z(\theta_1, \theta_1, \omega) \tag{5.41}
\]
The $P$-Condition Number of $B_w$

We have obtained expressions for the minimal eigenvalue of $B_w$ in two successive intervals of $\omega$ and these are given by (5.30). We obtained also expressions for the maximal eigenvalue of $B_w$ and these are given by (5.38), (5.40) and (5.41) for the various intervals defined in (5.32). We shall therefore examine the sign of the derivative with respect to $\omega$ of the ratio of the maximal to minimal eigenvalue in the following successive intervals:

$$
\begin{align*}
\text{(a)} \quad & 0 < \omega \leq \frac{2}{1+2\theta_2} \\
\text{(b)} \quad & \frac{2}{1+2\theta_2} < \omega \leq \frac{2}{1+2\sqrt{\theta_1\theta_2}} \\
\text{(c)} \quad & \frac{2}{1+2\sqrt{\theta_1\theta_2}} < \omega \leq \frac{2}{1+2\theta_1} \\
\text{(d)} \quad & \frac{2}{1+2\theta_1} < \omega < 2
\end{align*}
$$

(We have assumed since $\theta_1 < \theta_2$ that $0 < \omega \leq \sqrt{\theta_1\theta_2}$).

In the interval,

(a) $0 < \omega \leq \frac{2}{1+2\theta_2}$, we have by using (5.30) and (5.38) that

$$P(B_\omega) = \left(\frac{\theta_2}{\theta_1}\right) \cdot \frac{(1-\theta_1\omega_2)^2}{(1-\theta_1\omega_1)^2}.$$ 

Now, sign $\frac{3}{\delta_\omega} P(B_\omega)$ is negative, we have therefore that the minimum takes a value of one end of the interval. Thus

$$\min_\omega P(B_\omega) = P(B_\omega = 2/(1+2\theta_2))$$

(b) In the interval $\frac{2}{1+2\theta_2} < \omega \leq \frac{2}{1+2\sqrt{\theta_1\theta_2}}$, we have using (5.30) and (5.40) that

$$P(B_\omega) = \left(-\frac{\theta_1+\theta_2}{2\theta_1}\right) \cdot \frac{(1-\theta_1\omega_1\omega_2)}{(1-\theta_1\omega_1\omega_2)^2}.$$ 

Again sign $\frac{3}{\delta_\omega} P(B_\omega)$ is negative in this interval and thus we have that
\[
\min P(B_\omega) = P(B_\omega = 2/(1+2\sqrt{\theta_1 \theta_2})) \quad (5.44)
\]

The sign of the derivative of \( P(B_\omega) \) with respect to \( \omega \) has not changed in the above two successive intervals. We consider the next interval:

(c) \[
\frac{2}{(1+2\sqrt{\theta_1 \theta_2})} \leq \omega \leq \frac{2}{(1+2\theta_1)}
\]

\[
P(B_\omega) = \frac{\theta_1 + \theta_2}{\theta_2} \cdot \frac{(1-\omega + \omega \theta_2)^2}{(1-\omega + \omega \theta_1)^2}
\]

Here, \( \text{sign} \left( \frac{\partial}{\partial \omega} P(B_\omega) \right) \) is positive, therefore

\[
\min P(B_\omega = 2/(1+2\sqrt{\theta_1 \theta_2}))
\]

In the interval:

(d) \[
\frac{2}{1+2\theta_1} \leq \omega < 2,
\]

we have from (5.30) and (5.41) that

\[
P(B_\omega) = \frac{\theta_1 + \theta_2}{\theta_2} \cdot \frac{(1-\omega + \omega \theta_2)^2}{(1-\omega + \omega \theta_1)^2}
\]

Here also, \( \text{sign} \left( \frac{\partial}{\partial \omega} P(B_\omega) \right) \) is positive, and so

\[
\min P(B_\omega) = P(B_\omega = 2/(1+2\theta_1))
\]

Considering the sign of \( \frac{\partial P}{\partial \omega} \) in the various successive intervals, we have that

\[
\min P(B_\omega) = \frac{\theta_1 + \theta_2}{2\sqrt{\theta_1 \theta_2}} \quad (5.46)
\]

where we have used

\[
\omega = 2/(1+2\sqrt{\theta_1 \theta_2})
\]

in (5.44) and (5.45).

The Case When the Eigenvalues Ranges of \( H_1 \) and \( V_1 \) are Different

We had earlier assumed that the eigenvalue spectrum of both \( \tilde{u} \) and \( \tilde{v} \)

are respectively bounded by \( \theta_1 \) and \( \theta_2 \). If however we assume different bounds for these eigenvalues, then (5.19) takes the form

\[
\min \lambda(\tilde{u}, \tilde{v}, \omega) = \min \lambda(\tilde{u}, \tilde{v}, \omega) \quad (5.47)
\]

\[
\begin{array}{c}
\theta_1 \leq \gamma \leq \theta_2 \\
\theta_1 \leq \gamma \leq \theta_2
\end{array}
\]
where \( \gamma = \min(\theta_1, \tilde{\theta}_1) \) and \( \rho = \max(\theta_2, \tilde{\theta}_2) \).

Similarly, we have

\[
\max_{\theta_1 \leq u \leq \theta_2, \gamma \leq \tilde{u} \leq \rho} \lambda (u, v, w) \leq \max_{\gamma \leq \tilde{u} \leq \rho} \lambda (u, \tilde{v}, w).
\]

Following a similar approach as in preceding sections, we have

\[
\min_{0 < \omega < 2} P(B_\omega) \leq \min_{0 < \omega < 2} \frac{\max_{\gamma \leq \tilde{u}, \tilde{v} \leq \rho} \lambda (\tilde{u}, \tilde{v}, \omega)}{\min_{\gamma \leq \tilde{u}, \tilde{v} \leq \rho} \lambda (\tilde{u}, \tilde{v}, \omega)} = P(B_\omega)
\]

where \( \tilde{\omega} = 2/(1+2\sqrt{\gamma \rho}) \).
5.5 APPLICATION OF MODIFIED PRECONDITIONING METHOD TO BASIC ITERATIVE METHODS

In (5.46) we obtained an expression for the minimal value of the P-condition number of the preconditioned matrix in terms of the eigenvalue bounds of $\theta_1$ and $\theta_2$ of the eigenspectrum of matrices $\bar{H}$ and $\bar{V}$. From (5.46) we have

$$P(B_{\omega}) = \frac{1}{4} \left\{ \frac{\theta_1}{\theta_2} + \frac{\theta_2}{\theta_1} \right\}$$

(5.48)

By virtue of the definition of P-condition number, we have

$$P(A)_{\omega} = \frac{\theta_1}{\theta_2}$$

(5.49)

and thus

$$P(B_{\omega}) = \frac{1}{4} \left\{ \frac{\theta_1}{\theta_2} + \frac{\theta_2}{\theta_1} \right\} \left\{ \frac{\theta_2}{\theta_1} \right\} = P(A)_{\omega}$$

(5.50)

From (5.50), it becomes apparent that a significant reduction in the condition number of the coefficient matrix of the original system can be achieved by use of the modified preconditioning approach.

We shall now attempt to solve (5.9) by means of two iterative techniques and the acceleration of these techniques is bound to follow since this is implicit on the reduction of the condition of the preconditioned system given by (5.50).

Consider the iterative processes defined by

$$y^{(k+1)} = y^{(k)} + \alpha_1 (d-B_{\omega} y^{(k)})$$

and

$$y^{(k+1)} = y^{(k)} + \alpha (d-B_{\omega} y^{(k)}) + \beta (y^{(k)} - y^{(k-1)})$$

(5.51) (5.52)

where $\alpha$ and $\beta$ are optimal values of the acceleration parameters defined respectively by

$$\alpha_1 = \frac{2}{(\bar{\lambda}_1 + \bar{\lambda}_n)}$$

(5.53)

and

$$\alpha = \left( \frac{2}{(\bar{\lambda}_1)_{\omega} + (\bar{\lambda}_n)_{\omega}} \right)^2$$

$$\beta = \left( \frac{(\bar{\lambda}_1)_{\omega} - (\bar{\lambda}_n)_{\omega}}{(\bar{\lambda}_1)_{\omega} + (\bar{\lambda}_n)_{\omega}} \right)^2$$

(5.54)

where $\bar{\lambda}_1$ and $\bar{\lambda}_n$ are eigenvalue bounds of the $B_{\omega}$ matrix.
One practical significance of the theory developed so far in the previous section lies in the estimation of parameters defined by (5.53) and (5.54). A knowledge of the eigenvalue bounds of preconditioned matrix is therefore necessary in order to obtain estimates of $\tilde{a}_1$ and $\tilde{\beta}$ for the iterative processes (5.51) and (5.52). However, these parameters could also be obtained by a non-analytical or experimental approach.

Comparison of Rates of Convergence for the Modified and Implicit Methods

The first degree Richardson iterant (5.51) given by

$$y^{(k+1)} = y^{(k)} + \tilde{a}_1 \left( d - B \right) y^{(k)}$$

can be expressed in the form

$$y^{(k+1)} = My^{(k)} + c,$$

where

$$M = I - \tilde{a}_1 B$$

and

$$c = \tilde{a}_1 d.$$

The spectral radius of (5.55) is given by

$$S(M) = \max \left| 1 - \tilde{a}_1 \lambda_i \right|,$$

where $\lambda_i$ is an eigenvalue of $B_{\omega}^{-1}$.

We have that

$$\max_i \left| 1 - \tilde{a}_1 \lambda_i \right| \leq \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1},$$

where $\lambda_n$ and $\lambda_1$ are the maximal and minimal eigenvalue bounds of $B_{\omega}$ obtained from (5.40) and (5.30) respectively.

Since by definition $\tilde{\lambda}_n = P(B_{\omega})$, we can express the right hand member of (5.58) in terms of $P(B_{\omega})$. Thus

$$\frac{\tilde{\lambda}_n - \lambda_1}{\tilde{\lambda}_n + \lambda_1} = \frac{P(B_{\omega}) - 1}{P(B_{\omega}) + 1}.$$  

Using (5.58), we therefore have

$$\max_i \left| 1 - \tilde{a}_1 \lambda_i \right| \leq \frac{P(B_{\omega}) - 1}{P(B_{\omega}) + 1}.$$  

If we assume equality in (5.60) we can use (5.46) and (5.60) to obtain
\[
S(I - \frac{\theta_1 + \theta_2}{2\sqrt{\theta_1 \theta_2}}B_{\omega}) = \frac{\theta_1 + \theta_2}{2\sqrt{\theta_1 \theta_2}} - 1
\]
\[
= \frac{\theta_1 + \theta_2 - 2(\theta_1 \theta_2)^{\frac{1}{2}}}{\theta_1 + \theta_2 + 2(\theta_1 \theta_2)^{\frac{1}{2}}}
\]
\[
= \frac{(\theta_2)^{\frac{1}{2}} - (\theta_1)^{\frac{1}{2}}}{(\theta_2)^{\frac{1}{2}} + (\theta_1)^{\frac{1}{2}}}
\]
\[
= \left(\frac{1 - (\theta_1 / \theta_2)^{\frac{1}{2}}}{1 + (\theta_1 / \theta_2)^{\frac{1}{2}}}\right)^2
\]
(5.61)

It has been established in ([30], p.215), that the spectral radius of the PR (Peaceman-Rachford) iterative method applied to system of equations (5.5) has the form given in (5.60). For the model problem it is known that the PR and point SOR iterative methods as optimised one parameter iterative methods have asymptotically identical convergence rates for all grid sizes h. It would thus appear that the modified scheme would have identical convergence rates as well. This is however precluded by the fact that \( \theta_1 \) and \( \theta_2 \) in (5.60) are the eigenvalues of \( \bar{H} \) and \( \bar{V} \) defined by (5.14) and so are marginally different from the equivalent bounds of \( \bar{H} \) and \( \bar{V} \) that applies for the SOR, the PR and the preconditioned implicit methods [14].

We therefore deduce that the modified preconditioned simultaneous displacement method would have convergence rates which would closely approximate those of the unmodified methods especially for large h.

The second degree Richardson iterant (5.52) is given by
\[
y^{(k+1)} = y^{(k)} + \alpha_d (d - B_{\omega}) y^{(k)} + \beta (y^{(k)} - y^{(k-1)})
\]
(5.62)

It has been established by Frankel [17], that the spectral radius of (5.62) is equal to
\[
\left( \frac{(P(B_{\omega}) - 1)}{\omega} \right)^{\frac{1}{2}} = \left( \frac{\theta_1}{\theta_2} \right)^{\frac{1}{2}}
\]
(5.63)

We therefore have from (5.46) and (5.63) the following expression for the
It may be observed that (5.64) is equivalent to the upper bound given in [33] for the spectral radius of the 2-parameter PR method. This suggests that iterant (5.53) would not achieve better convergence rates than the PR method and therefore for large number of grid points would be slower than the \( M=2^k \) \((k>1)\) parameter PR method.

It can be shown that the use of the modified alternating preconditioning with iterants (5.51) and (5.52) has some computational advantage in terms of arithmetic operation counts over its implicit preconditioned variant [18]. Besides, the explicit structure of MADP is readily ammenable to parallel processing and so significant gain in computational speed over the ADI PR method or its preconditioned variant would be expected if parallel computation is employed.

Computational Approach

Consider the application of the modified preconditioning to the iterative processes (5.51) and (5.52) with regard to the solution of (5.9). For these two methods, common computational sequences are employed and these entail the following procedures:

(i) First compute

\[
d = (I + wH + w^2H^2)b, \quad (b \text{ is given by a row ordering of grid points}).
\]

Considering only the iterant (5.52), we observe from (5.9) that since we would compute \( B_w \) at each iteration we needed to form

\[
q(k) = (I + wH + w^2H^2)A(I + wV + w^2V^2)y(k).
\]

(ii) Then, we next compute

\[
x(k) = (I + wV + w^2V^2)y(k).
\]
by vertical ordering of grid points. Subsequently, we perform

\[(iii)\] \( Z^{(k)} = AX^{(k)} \) \hspace{1cm} \text{(5.65)}

The vector \( Z^{(k)} \) is generated directly from matrix \( A \) of (5.5) by applying
the five point difference approximation of equation (5.1) to each unknown
grid point with row ordering. We next compute

\[(iv)\] \( q^{(k)} = (I + \omega H + \omega H^2) Z^{(k)} \)

also by row ordering of the grid points.

\[(v)\] We now introduce any of the iterants (5.52) or (5.53) and solve for
each \( y^{(k+1)} \) until a specified accuracy tolerance has been met.

\[(vi)\] Finally, we compute the original solution vector in \( u \) from (5.11)
using a vertical ordering of grid points.

It may be emphasised however that the entire computational procedure
is executed using the optimal values of the preconditioning and acceleration
parameters defined by (5.31) and (5.54) for iterant (5.53).

The name 'Alternating Direction' should be apparent from the alternate
switching of the ordering of grid points along vertical and horizontal
directions.
5.6 **NUMERICAL RESULTS**

We shall discuss in this section the results obtained for the solution of a model problem [39] for which commutative condition normally imposed on H and V do not necessarily apply. We shall compare our results with those obtained from other methods [14],[39],[38] and these are shown in Figure 5.1.

The problem requires the solution of

\[ U_{xx}(x,y) + U_{yy}(x,y) = 0 \]

(where \( U_{xx} = \frac{\partial^2 u}{\partial x^2} \) and \( U_{yy} = \frac{\partial^2 u}{\partial y^2} \))

on the boundary \( B \), the unit square 4/10, 4/10 square removed from the centre.

The following boundary conditions apply.

\[ U(x,0) = U(1,y) = U(x,1) = U(0,y) = 0, \]

for \( 0 \leq x, y \leq 1 \).

Also,

\[ U(x,0.3) = U(0,7,y) = U(x,0.7) = U(0.3,y) = 1, \]

\((0.3 \leq x, y \leq 0.7)\)

The solution is obtained by imposing a square grid of size \( h \) on the region \( B \) with various values of \( h \).

\( h^{-1} = 5,10,20,40. \)

**Observations**

The Figure (5.1) gives \( \log(N) \) the number of iterations plotted against \( \log(h^{-1}) \) for the model problem with the various grid sizes.

We observe that the modified preconditioned simultaneous displacement first degree Richardson method is not as good as its implicit counterpart, the ADP Simultaneous Displacement method, the SOR nor ADI PR (M=1) method. The modified preconditioned second degree Richardson's method however appears competitive with the PR (M=3) and ADP Second Order Richardson's method.
If we consider as we have earlier mentioned, that the explicit structure of the modified preconditioning method renders it more readily ammenable to parallel processing than the other methods, then there is a possibility of significant speed ups of the convergence rates of the basic iterants when used on a parallel computer.

The application of the MADP can be extended to the fourth order equations and thus be used for the solution of Biharmonic equations.
FIGURE 5.1: Comparison of the modified ADP methods with SOR, ADP Simultaneous Displacement and ADP second degree Richardson method.
CHAPTER SIX

SUMMARY AND CONCLUSIONS
In this study we have applied the concept of explicit preconditioning for direct as well as the iterative solution of linear systems that could arise in the solution of partial differential equations.

Much of the explicit preconditioning techniques developed in this study derive from similar concepts to the implicit methods originally developed in Evans (1968). The explicit methods developed would naturally not only complement the existing body of implicit methods but would constitute it is hoped, a much more effective technique for the currently evolving generation of parallel computers.

Our study commenced in Chapter 3 by the development of three new explicit transformations for the direct solution of ill-conditioned linear systems. The 'condition minimising' effect of these transformations were investigated using various classes and categories of ill-conditioned matrices. A comparative appraisal of the relative effectiveness of the explicit as opposed to implicit preconditioning technique for different structure and properties of matrices was also undertaken.

In the first section of this chapter, we developed the First Order Transformation (F.O.T.) and the Second Order Transformation was developed in Section Two while the modified form of the Second Order Transformation was developed in Section Three. In these three sections we also investigated the effect of the use of normalised as opposed to un-normalised matrices.

Numerical results suggest that for Class C matrices which have weak or lack diagonal dominance, the F.O.T. transformation is inferior to the implicit transformation but has the advantage of simplicity. The use of normalised matrices is assumed for the implicit method but the use of normalisation for the F.O.T. has little or no effect on the result for this class of matrices. Also, the S.O.T. transformation appeared equally
less effective than the implicit transformation for the same Class C matrices although the use of normalised matrix forms for the S.O.T. appeared to have some significance. This weakness of the F.O.T. and S.O.T. transformations for Class C matrices is implicit on the boundedness condition defined in equation (3.33). However, for Class A and B matrices, the F.O.T. and S.O.T. fairly approximate the implicit transformation and are competitive for all classes of matrices considered.

In computational terms, however, the explicit methods appear to be more attractive. Besides, a strong possibility exists for the possible use in parallel processing for the explicit methods with significant computational advantage.

In Chapter 4, we examined some new explicit iterative schemes as well as the associated theory. In the first part of this Chapter, the Modified Preconditioned Jacobi (MPJ) method was developed. Part of the theoretical work for this new method consisted in obtaining valid expressions for the bounds on \( \hat{\xi}_{\max} (B) \) and \( \hat{\xi}_{\min} (B) \). It was shown that the upper bound on \( \xi_{\max} (B) \) is similar to the equivalent bound established in the classical implicit theory. However, a new bound on \( \xi_{\min} (B) \) and a new range of were derived which were essentially different from the bound and range hitherto established in the classical theory. Some associated parameters were also established for this new explicit scheme.

A comparison of the theoretical and numerical estimates indicate that the implicit method better approximates the theoretical bound on the optimal spectral radius while the explicit method gives a better approximation to the bound on \( \omega_{\text{opt}} \).

In the second part of Chapter 4, the Modified Preconditioned Simultaneous Displacement (MPSD) method was developed as well as determining certain optimal parameters. Numerical results for the model problem indicate
that the MPSD method is fairly competitive with its implicit counterpart (PSD) method especially for large $h^{-1}$.

In Section Three of Chapter 4, we considered the Modified Successive Overrelaxation (MSOR) method. We established convergence conditions and obtained expressions for the optimal spectral radius for some classes of matrices. Results indicate that this method compares well with the SOR method.

In Section Four of Chapter 4, we considered the Modified Gauss-Seidel (MGS) method. This method can be viewed as a special case of the MSOR method when the preconditioning parameter is unity. Thus the discussions of Section Three could apply to this method but with the given restriction on $\omega$. For this method we established a convergence condition and obtained expressions for the optimal spectral radius for a given class of matrices. The results also appear comparable to that of the SOR method.

In Section 4.5, we defined an approach to the use of explicit second degree iterative schemes based on the Conjugate Gradient method while an outline for the dynamic acceleration of the MPSD method was presented in Section Six of this chapter.

We concluded this chapter by considering the block preconditioned methods and showed that the concepts applicable to the point explicit methods can be extended to the block explicit methods.

In Chapter 5, a novel explicit scheme, the Modified Alternating Direction Preconditioning (MADP) method was developed and applied for the solution of Self Adjoint Second Order Partial Differential Equations. The relevant MADP theory was developed. This theory could be viewed as a modified form of the related classical theory in which the equation variables are assumed to have been slightly perturbed about the origin. The theoretical results established in this chapter indicate that the MADP
method could prove effective for accelerating the convergence of some basic iterants two of which formed the basis of the numerical work. Numerical results for the first degree iterant, the MADP first degree Richardson’s method suggests that for first order acceleration, the explicit method does not compare favourably with its implicit counterparts nor with the ADI-PR method. However, with second degree acceleration the explicit method appears competitive with its implicit counterpart as well as the ADI-PR method.

For the basic iterants considered, it may be noticed that the explicit method has a structure which is readily amenable to parallel processing. Thus, there is a strong possibility of significant speed-ups of the iterants when multiple array processors are used.

The concepts and techniques of the MADP method can be extended to the solution of fourth order or Biharmonic equations.

Finally, it may be stressed that one unique merit which most of the direct and iterative solution schemes developed in this study possess is that their structure would readily permit parallel processing as opposed to any of their preconditioned or classical implicit counterparts. The present study could therefore constitute a basis for the development of fast algorithms for use in some newly emerging parallel computers. It is hoped that an investigation in this direction would form a logical sequel and an extension of this study in the near future.
REFERENCES


[24b] MISSIRLIS, N.M. and EVANS, D.J., "On dynamic acceleration of the
preconditioned simultaneous displacement method", Inter.J.


Cliffs, New Jersey.

[26a] NIETHAMER, W., "Relaxation by Kimpexen matrizen", Math.Zietsch, 86,

[27] PEACEMAN, D.W. and RACHFORD (Jr.) H.H., "The numerical solution of

[27a] PARTER, S.V., "Multi-line iterative methods for elliptic difference
equations and fundamental frequencies", Numer.Math., 3, pp.305-


[28a] REID, J.K., "On the method of conjugate gradients for solution of
large sparse systems of linear equations", in 'Large Sets of
Linear Equations', Ed. by J.K. Reid, Academic Press, London,
231-254 (1971).

Iterative Methods for Computation of Solution and Eigenvalues
of Self-Adjoint Boundary Value Problems', by Engeli, M.,
[29] RICHARDSON, L.F., "Approximate arithmetic solution of finite
differences of physical problems involving differential equations
with application to stresses in masonry dam", Phil. Trans. Roy.


[29b] SOUTHWELL, R.V., "Relaxation methods in theoretical physics",

[30a] VARGA, R.S., "Factorization and normalised overrelaxation scheme",


[32] WACHSPRESS, E.L., "Iterative solution of elliptic schemes and
application to neutron diffusion equations of reactor physics",

[33] WACHSPRESS, E.L., "Optimum alternating direction iteration parameters
(1960).

[34] WILKINSON, J., "Algebraic eigenvalue problems", Oxford University
Press (1965).

[35] WILKINSON, J., "Rounding errors in algebraic processes", Notes in
Applied Science, No. 32, National Physical Laboratory.


MASTER MPJ1

N=PRECONDITIONING CONSTANT
BETA AND T ACCELERATION CONSTANTS
X,Y,Z,X1,Z2, AND D ARE REAL ARRAYS

DIMENSION X(45,45),Y(45,45),Z(45,45),X1(45,45),Z2(45,45),D(45,45)

****************************************************************
THE MPJ METHOD
SECOND-DEGREE RICHARDSON ITERANT
****************************************************************

N=20
W=1.60
BETA=0.50
T=1.4030
ITER=9
M=N+4
ZA=1.0/8.0
ZC=1.0/16.0
PP=1.0/4.0
NH=M-2

SET ZEROES ON X,Y MESHES

DO 5 I=1,N
DO 5 J=1,N
X(I,J)=0
X1(I,J)=0
Y(I,J)=0
Z(I,J)=0
Z2(I,J)=0
D(I,J)=0
5 CONTINUE
SET ONES ON INNER MESHES POINTS

DO 3 I=3,IM
DO 2 J=3,MM
X(I,J)=1.0
3 CONTINUE
DO 31 J=3,MM
X(I,J)=1.0
31 CONTINUE
COUNT=1
ITER=ITER+1

FORM (I+U+U*U*U*U)(X(N))
DO 23 I=3,IM
DO 23 J=3,MM
Y(I,J)=X(I,J)+W*PP*(X(I+1,J)+X(I,J+1))+W*V*(ZQ*X(I+1,J)+ZT*X(I+1,J +1))+ZQ*X(I+2,J)
23 CONTINUE

FORM A(I+U+U*U*U*U)X(N)
DO 24 I=3,IM
DO 24 J=3,MM
Z(I,J)=Y(I,J)-PP*(Y(I+1,J)+Y(I-1,J)+Y(I,J+1)+Y(I,J-1))
24 CONTINUE

FORM (I+L+L*U*U*U*U)A(I+U+U*U*U*U)X(N)
DO 25 I=3,IM
DO 25 J=3,MM
Z2(I,J)=Z(I,J)+PP*(Z(I-1,J)+Z(I,J-1))+V*V*(ZQ*Z(I-2,J)+ZT*Z(I-1, J-1))+ZQ*Z(I,J-2)
25 CONTINUE

FORM X(N+1)=X(N)+T*Z2
DO 27 I=3,IM
DO 27 J=3,MM
X(I,J)=X(I,J)-T*Z2(I,J)+BETA*(X(I,J)-D(I,J))
27 CONTINUE

TEST OF CONVERGENCE
DO 60 I=3,MM
DO 60 J=3,MM
EPS=0.000005
IF(ABS(X1(I,J)-X(I,J)) .GT. EPS) GOTO 30
ICOUNT=ICOUNT+1
30 CONTINUE
60 CONTINUE
31 CONTINUE
JB=**2
IF(ICOUNT.EQ.JB) GOTO 32
DO 61 I=3,MM
DO 61 J=3,MM
D(I,J)=X(I,J)
DO 61 CONTINUE
GOTO 31
32 CONTINUE
WRITE(2,49) ITER
WRITE(2,56) U
WRITE(2,41) BETA
WRITE(2,72) T
33 CONTINUE
42 FORMAT(1H ,24H NO OF POINTS CONVERGED=,I13/)
49 FORMAT(1H ,I10/)
56 FORMAT(1H ,"EXTRAP FACTOR",F5.2/)
41 FORMAT(1H ,"BETA= ",F7.2/)
71 FORMAT(1H ,"KB= ",I4)
72 FORMAT(1H ," ALPHA =" ,F7.4/)
STOP
END
FINISH
**MASTER MRJ1**

DIMENSION D(45,45),U(45,45),Y(45,45),X(45,45),Z(45,45)
DIMENSION Q(45,45),Y1(45,45),D(45,45),XX(45,45),YY(45,45)

******************************************************************
MODIFIED ALTERNATING DIRECTION PRECONDITIONING (MADP) METHOD

* X X X X X X X X X X X X X X X

N=20
W=1.61
BETA=0.51
T=1.32
ITER=0
NM=4
NA=NM-1
PP=1.9/4.0

SET ZEROES ON ALL MESH POINTS (INNER)
DO 5 I=1,NM
DO 5 J=1,NM
D(I,J)=0.0
U(I,J)=0.0
Y(I,J)=0.0
X(I,J)=0.0
Z(I,J)=0.0
Q(I,J)=0.0
Y1(I,J)=0.0
B(I,J)=0.0
XX(I,J)=0.0
YY(I,J)=0.0
5 CONTINUE

SET ONES ON INNER MESHES OF U AND B
DO 3 I=2,NA
DO 8 J=2,NA
U(I,J)=1.0
8 CONTINUE

GOTO 18

COMPUTE D=(I+J+I*2+J*1)B BY R-ORDERING

DO 11 J=2,NA
DO 11 I=2,NA
XX(I,J)=I*PP*(B(I,J+1)+B(I,J-1))
11 CONTINUE

DO 12 J=2,NA
DO 12 I=2,NA
YY(I,J)=I*PP*(XX(I,J+1)+XX(I,J-1))
12 CONTINUE

DO 17 J=2,NA
DO 17 I=2,NA
D(I,J)=B(I,J)+XX(I,J)+YY(I,J)
17 CONTINUE
18 CONTINUE

COMPUTE Y=(I+y)U......BY..V...ORDERING

DO 19 I=2,NA
DO 19 J=2,NA
XX(I,J)=I*PP*(U(I-1,J)+U(I+1,J))
19 CONTINUE

DO 21 I=2,NA
DO 21 J=2,NA
Y(I,J)=U(I,J)*XX(I,J)
21 CONTINUE
22 CONTINUE
23 CONTINUE
24 CONTINUE

COMPUTE X=(I+WH2*J+1)Y BY ....V...ORDERING

DO 26 I=2,NA
DO 26 J=2,NA
XX(I,J)=I*PP*(Y(I-1,J)+Y(I+1,J))
26 CONTINUE
YY(I,J)=w*PP*(XX(I-1,J)+XX(I+1,J))

DO 28 I=2,NA
DO 28 J=2,NA
X(I,J)=Y(I,J)+XX(I,J)+YY(I,J)

COMPUTE Z=AX......BY..R...ORDERING

DO 30 J=2,NA
DO 30 I=2,NA
Z(I,J)=X(I,J)-PP*(X(I,J-1)+X(I,J+1)+X(I-1,J)+X(I+1,J))

COMPUTE Q=(1+WH+2*H)*Z......BY..R...ORDERING

DO 32 J=2,NA
DO 32 I=2,NA
XX(I,J)=r*PP*(Z(I,J+1)+Z(I,J-1))

INTRODUCE THE REQD ITERANT

DO 36 J=2,NA
DO 36 I=2,NA
Y(I,J)=Y(I,J)+XX(I,J)+YY(I,J)

TEST OF CONVERGENCE

DO 40 I=2,NA
DO 40 J=2,NA
EPS=0.000005
IF(ABS(Y(I,J)-Y(I,J)).GE.EPS)GOTO 50
ICOUNT=ICOUNT+1

IF(ICOUNT.EQ.NBB)GOTO 52

DO 46 I=2,NA
DO 46 J=2,NA
D(I,J)=Y(I,J)
Y(I,J)=Y(I,J)

GOTO 7
```
52 CONTINUE

COMPUTE U=(I+4*H+7*2*H**H)Y1.....BY....V...ORDERING

DO 44 I=2,NA
DO 44 J=2,NA
XX(I,J)=4*PP*(Y1(I-1,J)+Y1(I+1,J))
44 CONTINUE

DO 43 I=2,NA
DO 43 J=2,NA
YY(I,J)=4*PP*(XX(I-1,J)+XX(I+1,J))
43 CONTINUE

DO 54 I=2,NA
DO 54 J=2,NA
U(I,J)=Y1(I,J)+XX(I,J)+YY(I,J)
54 CONTINUE

WRITE(2,56)W
WRITE(2,41)BETA
41 FORMAT(1H , 'BETA= ', F7.4)
WRITE(2,78)T
78 FORMAT(1H , 'T= ', F7.4)
WRITE(2,49)ITER

42 FORMAT(1H , 'NO OF POINTS CONVERGED= ', I10)
49 FORMAT(1H , 'NO OF ITERATIONS = ', I10)
56 FORMAT(1H , 'EXTRAP FACTOR= ', F7.4)
77 FORMAT(1H , '5F15.6,' )

332 CONTINUE
33 CONTINUE
STOP
END
```

DIRECT PRECONDITIONED MINIMIZATION OF CONDITION OF MATRIX SYSTEM

A is a 4*4 segment of an ill-conditioned matrix

N = 4
AQ = 1.0
IQ = N
IQ = N
ID = N
NC = N
NA = N
NJ = N

READ(1, 7)((A(I, J), J=1, N), I=1, N)
WRITE(2, 71)
WRITE(2, 33)((A(I, J), J=1, N), I=1, N)

DO 4 I=1, N
DO 4 J=1, N
AA(I, J) = A(I, J)
4 CONTINUE

DO 88 K=1, 20, 1

DO 81 I=1, N
DO 81 J=1, N
A(I, J) = AA(I, J)
81 CONTINUE

IFAIL = 1
DO 8 I=1,N
DO 8 J=1,N
D(I,J)=0.0
L(I,J)=0.0
U(I,J)=0.0
8 CONTINUE

C
DO 18 I=1,N
D(I,I)=A(I,I)
18 CONTINUE
WRITE(2, 33)((D(I,J), J=1,N), I=1,N)

C
DO 19 I=1,N
DO 19 J=1,N
IF(J.GE.I)GOTO 19
L(I,J)=A(I,J)
L(I,J)=-L(I,J)
19 CONTINUE

C
WRITE(2, 33)((L(I,J), J=1,N), I=1,N)
CALL F01CKF(IL1,L,L,IQ,IY,ID,Z,1,1,IFAIL)
DO 20 I=1,N
DO 20 J=1,N
IF(J.LE.I)GOTO 20
U(I,J)=A(I,J)
U(I,J)=-U(I,J)
20 CONTINUE

C
WRITE(2, 33)((U(I,J), J=1,N), I=1,N)
CALL F01CKF(UU1,U,U,IQ,IY,ID,Z,1,1,IFAIL)

C
IW=0.1*K
C THIS MULTIPLIES THE ARRAYS
AIW=IW*IW
DO 70 I=1,N
DO 70 J=1,N
70 IL(I,J)=IW*L(I,J)+AIW*LL1(I,J)

C
THIS COPIES THE ARRAYS
DO 72 I=1,N
DO 72 J=1,N
72 V(I,J)=IL(I,J)
C
THIS MULTIPLIES THE ARRAYS
DO 74 I=1,N
DO 74 J=1,N
74 IU(I,J)=IW*U(I,J)+AIW*UU1(I,J)
C
THIS COPIES THE ARRAYS
C
DO 76 I=1,N
DO 76 J=1,N
76 U(I,J)=IU(I,J)
C
DO 78 I=1,N
DO 78 J=1,N
DA(I,J)=D(I,J)
M=N
THIS PERFORMS ARRAY ADDITION
CALL F01CDF(S,D,V,M,N,IFAIL)
IF(IFAIL.EQ.2)GOTO 101
CALL F01CDF(Q,D,U,M,N,IFAIL)
IF(IFAIL.EQ.1)GOTO 101
IF(IFAIL.EQ.2)GOTO 102
GOTO 111
101 CONTINUE
WRITE(2,103)
102 CONTINUE
WRITE(2,104)
111 CONTINUE
FORMS MATRIX PRODUCTS
CALL F01CKF(R,S,A,IQ,IX,ID,Z,1,1,IFAIL)
IF(IFAIL.NE.0)GOTO 105
NQ=N
NY=N
ND=N
CALL F01CKF(B,R,Q,M,NY,ND,Z,1,1,IFAIL)
DO 400 I=1,N
DO 400 J=1,N
DAL(I,J)=B(I,J)
400 CONTINUE
ALLOP=IW
231 CONTINUE
GO TO 113
105 CONTINUE
WRITE(2,106)
113 CONTINUE
WRITE(2,15)
DO 13 I=1,N
13 WRITE(2,33)(B(I,J),J=1,N)
THIS EVALUATES THE EIGENVALUES
IA=N
CALL F02AAF(B,IA,N,EVAL,WK,IFAIL)
IF(IFAIL.EQ.2)GOTO 107
GOTO 115
107 CONTINUE
WRITE(2,108)
115 CONTINUE
WRITE(2,86)
WRITE(2,33)(EVAL(I),I=1,N)
WRITE(2,240)
WRITE(2,46)ALLOP
WRITE(2, 44)
BVAL = EVAL(N)/EVAL(1)
WRITE(2, 46) BVAL

C
THIS COPIES ARRAYS TO ANOTHER OF SIZE =(N+1,N)
DO 304 I = 1, N
DO 304 J = 1, N
BB(I, J) = B(I, J)
304 CONTINUE
IIA = N + 1
CALL F01ABF(BB, IIA, N, BINV, N, WK, IFAIL)

C
COPY OF INVERSE TO UPPER PART
DO 376 I = 1, N
DO 376 J = 1, N
IF(BINV(I, J).EQ.0) GOTO 375
IF(BINV(J, I).EQ.0) GOTO 375
375 CONTINUE
376 CONTINUE
C
COPY ENDS HERE
316 CONTINUE
C = 0
319 CONTINUE
NZ = N
DO 320 I = 1, NZ
NCA = N - 1
DO 320 J = 1, NCA
QA = BINV(I, J)
QB = BINV(I, J + 1)
IF(ABS(QB).LT.ABS(QA)) GOTO 324
GOTO 326
324 CONTINUE
BINV(I, J + 1) = QA
BINV(I, J) = QB
326 CONTINUE
320 CONTINUE
C = C + 1
QQ = N
IF(C.LT.QQ) GOTO 319
CONTINUE
DO 334 I = 1, N
VV(I, 1) = BINV(I, N)
334 CONTINUE
VV(I, 1) = QP
VV(I + 1, 1) = QR
346 CONTINUE
340 CONTINUE
HHAB = VV(N, 1)

C
DO 356 I = 1, N
DO 356 J = 1, N
DD(I, J) = DAL(I, J)
356 CONTINUE
NCA = N - 1
DO 358 I=1,NZ
DO 358 J=1,NCA
QQA=DD(I,J)
QQB=DD(I,J+1)
IF(ABS(QQB).LT.ABS(OOA))GOTO 370
GOTO 373
370 CONTINUE
DD(I,J+1)=QQA
DD(I,J)=QQB
373 CONTINUE

358 CONTINUE
CC=CC+1
QQQ=N
IF(CC.LT.QQQ)GOTO 358
CONTINUE
NSS=N
DO 362 I=1,NSS
AV(I,1)=DD(I,N)
362 CONTINUE

364 CONTINUE
QQR=AV(I,1)
QQP=AV(I+1,1)
IF(ABS(QQP).LT.ABS(QQR))GOTO 372
GOTO 374
372 CONTINUE
AV(I,1)=QQP
AV(I+1,1)=QQR
374 CONTINUE

AHFABEL=AV(N,1)
WRITE(2,366)
MCOND=ABS(MCOND)
WRITE(2,46)MCOND

C

THIS EVALUATES THE N-CONDITION NUMBER N
SU=0
DO 380 I=1,N
DO 380 J=1,N
SU=SU+DAL(I,J)**2
380 CONTINUE
TA=1.0/2.0
SQRS=SU**TA
SS=0
DO 384 I=1,N
DO 384 J=1,N
SS=SS+BIN(V(I,J)**2
384 CONTINUE
AT=1.0/2.0
SSQSS=SS**AT
MCOND=(SQRS*SSQSS)/N
WRITE(2,388)
WRITE(2,46)MCOND
88 CONTINUE
STOP
END
FINISH