PRECONDITIONED ITERATIVE METHODS FOR SOLVING ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

by

NIKOLAOS M. MISSIRLIS, B.Sc.

Submitted in partial fulfilment of the requirements
for the award of Doctor of Philosophy
of the Loughborough University of Technology
December, 1978

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.

N.M. MISSIRLIS

DEDICATED TO

My parents

MICHAEL and LEMONIA

ACKNOWLEDGEMENTS

I wish to express my deep gratitude and appreciation to my supervisor, Professor D.J. Evans, Head of Department of Computer Studies, for introducing me to this area of study as well as for his continuous advice, keen interest and encouragement at every stage of my research work.

I would also like to acknowledge the interest and the useful suggestions of Professor A. Hadjidimos at University of Ioannina, Greece, on various aspects of this thesis.

Furthermore, I wish to acknowledge the encouragement and the useful scientific dialogues of my friend, Dr. C. Spyropoulos.

For access to the CDC 7600 computer and its facilities, I thank the staff of the Computer Centre at Loughborough University.

My sincere thanks are extended to Miss Judith Briers for all her help towards the completion of this thesis.

I would also like to express my thanks to Miss Mary Katsiyianni for her interest, help and continuous encouragement during these years.

Finally, I take this opportunity to express my grateful acknowledgements and my deep appreciation to my parents, Michael and Lemonia Missirlis for their unlimited interest, encouragement and help throughout all the years of my studies.

TABLE OF CONTENTS

PAGE

CHAPTER 1: INTRODUCTION 1

1.1 Partial Differential Equations 1

1.2 Discretisation of the Generalised Dirichlet Problem.. 6

CHAPTER 2: MATRIX PRELIMINARIES 10

2.1 Background of Matrix Theory 11

2.2 Positive Definite Matrices 13

2.3 Vector and Matrix Norms 15

2.4 Convergence of Sequences of Matrices 18

2.5 Irreducibility and Weak Diagonal Dominance 19

2.6 Ordering Vectors and Consistently Ordered Matrices 23

2.7 Property A.. 26

CHAPTER 3: STATIONARY AND NON-STATIONARY ITERATIVE METHODS 29

3.1 Introduction 30

3.2 Linear Stationary Iterative Methods 32

3.3 Convergence of Iterative Methods 37

3.4 Rate of Convergence 39

3.5 Some Theorems on the Convergence 41

3.6 Comparison of Reciprocal Rates of Convergence 46

3.7 Semi-Iterative Methods.. 58

3.8 Variable Extrapolation Methods 62

3.9 Second Degree Methods 64

3.10 The Conjugate Gardient Method 66

CHAPTER 4: AN INTRODUCTION TO PRECONDITIONING TECHNIQUES 72

4.1 Introduction 73

4.2 The Preconditioning Technique for the Construction of Iterative Methods 76

4.3 On the Preconditioned Iterative Methods 80

4.3.1 Irreducible Matrices with Weak Diagonal Dominance.. 85

PAGE

4.3.2 Positive Definite Matrices 86
4.3.3 L-Matrices and Related Matrices 88
4.3.4 Consistently Ordered Matrices 90

4.4 The Preconditioned Jacobi Method (PJ Method) 109

4.5 Convergence of the PJ Method.. 111

4.6 Determination of Good Bounds on $\lambda(B_\omega)$ and $\Lambda(B_\omega)$ .. 113

4.7 Determination of $S(\mathcal{H}_{\omega_1})$ and ω_1 118

4.8 Computational Results 124

4.9 The Preconditioned Simultaneous Displacement Method (PSD Method).. 127

4.10 Convergence of the PSD Method 128

4.11 Choice of τ and ω for the PSD Method 129

4.12 Comparison of Reciprocal Rates of Convergence 135

4.13 Computational Results 137

4.14 The Unsymmetric PJ Method (UPJ Method) 147

4.15 The Unsymmetric PSD Method (UPSD Method).. 150

CHAPTER 5: BLOCK PRECONDITIONED ITERATIVE METHODS - ACCELERATED TECHNIQUES 153

Section A: BLOCK PRECONDITIONED ITERATIVE METHODS 154

5.1 Introduction 154

5.2 Group PSD Methods 159

5.3 Comparison of Line and Point PSD Methods.. 162

5.4 Computational Results 165

Section B: ACCELERATED TECHNIQUES 170

5.5 Preconditioned Jacobi-Semi Iterative Method (PJ-SI Method) 170

5.6 Preconditioned Jacobi-Variable Extrapolation Method (PJ-VE Method) 178

5.7 Second Degree-Preconditioned Jacobi Method (SD-PJ Method) 179

5.8 Generalised Conjugate Gradient Method.. 182

5.9 Preconditioned Jacobi-Conjugate Gradient Method (PJ-CG Method) 184

5.10 Comparisons and Computational Results.. 186

PAGE

CHAPTER 6: THE ADAPTIVE ALGORITHM 196

6.1 Introduction 197

6.2 Some Considerations for Choosing the Optimum Parameters.. 199

6.3 Stopping Procedures.. 202

6.4 Computational Procedures and Numerical Results.. .. 205

6.5 The Theoretical Basis for the Adaptive Determination of Parameters 211

6.6 The Adaptive Algorithm.. 221

6.7 Numerical Results 230

CHAPTER 7: ALTERNATING DIRECTION PRECONDITIONING TECHNIQUES FOR THE NUMERICAL SOLUTION OF THE ELLIPTIC SELF-ADJOINT SECOND ORDER AND BIHARMONIC EQUATIONS 243

7.1 Introduction 243

7.2 Some Considerations on the Iterative Scheme (1.11) 248

7.3 The Modified Alternating Direction Preconditioning Method (MADP Method) 254

7.3.1 The Case Where the Eigenvalues of H and V are the Same.. 256

7.3.2 The Case Where the Eigenvalue Ranges of H and V May Be Different.. 261

7.4 Application of the Accelerated Procedures to the MADP Method 266

7.5 The Model Problem - Comparison of Rates of Convergence 269

7.6 Numerical Results 273

7.7 The Biharmonic Equation 277

7.8 The MADP Method for the Numerical Solution of the Biharmonic Equation 279

7.8.1 The Case Where the Eigenvalue Ranges of H and V are the Same 280

7.8.2 The Case Where the Eigenvalue Ranges of H and V May Be Different.. 286

7.9 Rates of Convergence on the Unit Square 287

7.10 Numerical Results 291

CHAPTER 8: SUMMARY AND CONCLUSIONS 294

PAGE

REFERENCES 300

APPENDIX A: ARITHMETIC OPERATION COUNT 313

APPENDIX B: DETERMINATION OF A BOUND ON S(LU) 321

APPENDIX C: CHEBYSHEV MINIMAX THEOREM 325

APPENDIX D: UNIMODULITY OF THE FUNCTION $P(\omega)$ 327

CHAPTER 1

INTRODUCTION

1.1 PARTIAL DIFFERENTIAL EQUATIONS

The majority of the problems of physics and engineering fall into one of three physical categories: equilibrium problems, eigenvalue problems and propagation problems.

The eigenvalue problems may be thought of as extensions of equilibrium problems where critical values of certain parameters are to be determined in addition to the corresponding steady-state configuration. Thus the previous physical classification may be reduced to the two major classes of equilibrium and propagation problems.

These problems are usually represented mathematically by a partial differential equation (or a set of such equations). Such an equation is the linear second-order partial differential

$$A\frac{\partial^2 U}{\partial x^2} + 2B\frac{\partial^2 U}{\partial x \partial y} + C\frac{\partial^2 U}{\partial y^2} + D\frac{\partial U}{\partial x} + E\frac{\partial U}{\partial y} + FU = G, \qquad (1.1)$$

where A,B,C,D,E,F and G are given functions which are continuous in some region in the (x,y) plane.

A characteristic problem is the following: given a region R, finite or infinite, with a boundary ∂R, to find a function U(x,y) which is twice differentiable and satisfies (1.1) in R, which is continuous in R+∂R and satisfies prescribed conditions on ∂R. For example, we might require that

$$U(x,y) = g(x,y) \qquad (1.2)$$

on ∂R, or alternatively, the normal derivative $\frac{\partial U}{\partial n}$ or a linear combination of U and $\frac{\partial U}{\partial n}$ be specified on ∂R. Equations of the form (1.1) may be classified as elliptic, hyperbolic, or parabolic depending upon the behaviour of the coefficients A,B and C. Thus equation (1.1) is said to be

a) elliptic if $B^2-AC<0$ in R,

b) hyperbolic if $B^2-AC>0$ in R,

and c) parabolic if $B^2-AC=0$ in R.

If the quantity B^2-AC changes sign in R, then the equation is said to be

of mixed type. For instance, the differential equation

$$x\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0$$

is elliptic for $x>0$, hyperbolic for $x<0$ and parabolic for $x=0$. Equilibrium or steady state problems are associated with the elliptic equations whereas the governing equations for propagation problems are parabolic or hyperbolic. Representative examples of such equations are:

(i) Poisson's equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = G(x,y) \qquad \text{(elliptic)}. \tag{1.3}$$

If $G(x,y)\equiv 0$, then (1.3) reduces to Laplace's equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0. \tag{1.4}$$

(ii) The vibrating string equation

$$\frac{\partial^2 U}{\partial x^2} - \frac{\partial^2 U}{\partial y^2} = 0 \qquad \text{(hyperbolic)}. \tag{1.5}$$

(iii) The diffusion equation

$$\frac{\partial^2 U}{\partial x^2} - \frac{\partial U}{\partial y} = 0 \qquad \text{(parabolic)}. \tag{1.6}$$

A problem in mathematical physics is called "well posed" if its solution exists, is unique and varies continuously with the boundary data. Let us consider the generalised Dirichlet problem involving a bounded connected region R and a continuous function $g(x,y)$ prescribed on ∂R. The function $U(x,y)$ is required to be continuous in $R+\partial R$, to satisfy (1.1) in R and to satisfy the condition (1.2) on the boundary ∂R. Moreover, if (1.1) is an elliptic equation and if $F\leq 0$ in R, then the generalised Dirichlet problem has a unique solution under fairly general conditions (e.g. see Courant and Hilbert [1962]). A special case of the generalised Dirichlet problem is that which involves Laplace's equation and is a very classical problem in applied mathematics. This problem can be solved analytically in certain special cases. Analytic solutions can be given for

the circle, rectangle and for the half plane (or for certain other regions which can be transformed conveniently by conformal mapping into the previously mentioned regions). However, it is not usually possible that an analytic solution of a problem involving (1.1) can be found under arbitrarily shaped regions and for general boundary conditions. Even if there was no differential equation to be satisfied at all, it would not be easy to find a function defined and continuous in $R+\partial R$ which satisfies (1.2). Thus, one is usually forced to use numerical methods.

Two standard general methods for the numerical solution of elliptic partial differential equations are the method of finite differences (e.g. see Varga [1962], Forsythe and Wasow [1960], Wachspress [1966], Young and Gregory [1973]) and the finite element method (e.g. see Zienkiewicz [1971], George [1971], Strang [1972], Zlamal [1968]).

Recently, the latter method has become a popular and effective procedure (see Kim [1973]). The finite element method is essentially a technique to construct a set of coordinate functions for the Ritz or the Galerkin method (Collatz [1960]). In the finite element method the region R is partitioned into a union of "finite elements", of which commonly used elements are triangles and rectangles. Next, a trial function is constructed with the property that it is a polynomial (not greater degree than three) on each finite element.

Alternatively, in the application of finite difference methods one replaces the region R by a finite set of points R_h where $R_h \subseteq R$ and also replaces the boundary ∂R by a set of points ∂R_h, which may or may not belong to $R+\partial R$. For each point P of R_h we develop a linear relation involving the value of $U(x,y)$ at P and the values of $U(x,y)$ at certain neighbouring points of R_h and at certain points of ∂R_h. If there are N points of R_h, one obtains in this way a system of N linear algebraic equations with N unknowns. If the system of linear equations can be solved uniquely, as is frequently the

case, then the values of U(x,y) at points of R_h are accepted as approximate values of the true solution. Some useful methods for deriving finite difference approximations are based on Taylor's series, integration and the variational technique.

1.2 DISCRETISATION OF THE GENERALISED DIRICHLET PROBLEM

We now describe the procedure for discretising elliptic partial differential equations and show how their solution by finite difference methods often leads to linear systems whose matrices have some properties of fundamental importance.

Let us consider the generalised Dirichlet problem as defined in the previous section and assume that the coefficient B(x,y) in (1.1) of the mixed derivative vanishes identically in R+∂R (one can make a change of independent variables so that the coefficient of the mixed derivative vanishes). Thus, we have that U(x,y) satisfies the linear second-order partial differential equation

$$A\frac{\partial^2 U}{\partial x^2} + C\frac{\partial^2 U}{\partial y^2} + D\frac{\partial U}{\partial x} + E\frac{\partial U}{\partial y} + FU = G \tag{2.1}$$

in R where A,C,D,E,F and G are analytic functions of the independent variables x and y in R and satisfy the conditions $A>0$, $C>0$ and $F\leq 0$. However, if we have

$$\frac{\partial A}{\partial x} = D \qquad \text{and} \qquad \frac{\partial C}{\partial y} = E, \tag{2.2}$$

then instead of (2.1) we consider the self-adjoint differential equation

$$\frac{\partial}{\partial x}\left(A\frac{\partial U}{\partial x}\right) + \frac{\partial}{\partial y}\left(C\frac{\partial U}{\partial y}\right) + FU = G. \tag{2.3}$$

Even if (2.1) is not self-adjoint, it may be possible to obtain a self-adjoint equation by multiplying both sides of (2.1) by an "integrating factor" μ(x,y) so that we have

$$\frac{\partial}{\partial x}(\mu A) = \mu D, \qquad \frac{\partial}{\partial y}(\mu C) = \mu E. \tag{2.4}$$

The function μ(x,y) exists if and only if

$$\frac{\partial}{\partial y}\left(\frac{D-\frac{\partial A}{\partial x}}{A}\right) = \frac{\partial}{\partial x}\left(\frac{E-\frac{\partial C}{\partial y}}{C}\right). \tag{2.5}$$

We will be concerned with the differential equation (2.3) instead of (2.1). If the condition (2.5) is satisfied, then the equation (2.1) is called essentially self-adjoint.

In order to apply the method of finite differences, we superimpose a mesh consisting of a network of horizontal and vertical lines over the region R with a uniform spacing (although this is not necessary) of size $h>0$.

For a given point (x_0,y_0) we consider the set Ω_h which contains all points of the form $(x_0+ih,\ y_0+jh)$ for $i,j=0,\pm1,\pm2,\ldots$. Two points (x,y) and (x',y') of Ω_h are adjacent if $(x-x')^2+(y-y')^2=h^2$, whereas they are properly adjacent if they are adjacent, both are in $R+\partial R$ and the open segment joining them, not necessarily including the end points, is in R. Moreover, we define $R_h=R\cap\Omega_h$ and $\partial R_h=\Omega_h\cap\partial R$. A point P of R_h is regular if the four adjacent mesh points in Ω_h lie in $R+\partial R$ and are properly adjacent to P. In the sequel, we will assume that R and Ω_h are such that all points of R_h are regular points.

Let us now consider the construction of a discrete representation of the differential equation (2.3). For a point (x,y) of R_h the self-adjoint equation given by (2.3) is replaced by the symmetric difference equation

$$\begin{aligned} &h^{-2}\{A(x+\tfrac{1}{2}h,y)[u(x+h,y)-u(x,y)]-A(x-\tfrac{1}{2}h,y)[u(x,y)-u(x-h,y)] \\ &+C(x,y+\tfrac{1}{2}h)[u(x,y+h)-u(x,y)]-C(x,y-\tfrac{1}{2}h)[u(x,y)-u(x,y-h)]\} \\ &+F(x,y)u(x,y) = G(x,y). \end{aligned} \tag{2.6}$$

Thus we have transformed the continuous problem to a discrete generalised Dirichlet problem. That is, we now seek to determine a function $u(x,y)$[†] defined on $R_h\cap\partial R_h$ such that (2.6) is satisfied on R_h and $u(x,y)=g(x,y)$ on ∂R_h.

Multiplying (2.6) by $-h^2$ we obtain the difference equation

$$\begin{aligned} u(x,y) = {} & \beta_1(x,y)u(x+h,y)+\beta_2(x,y)u(x,y+h)+\beta_3(x,y)u(x-h,y) \\ & +\beta_4(x,y)u(x,y-h)+\tau(x,y) \end{aligned} \tag{2.7}$$

where

[†]*The unknown function $u(x,y)$ denotes the finite difference approximation to the exact solution $U(x,y)$.*

$$\left.\begin{aligned} \beta_1(x,y) &= \frac{A(x+\frac{1}{2}h,y)}{S(x,y)}, & \beta_2(x,y) &= \frac{C(x,y+\frac{1}{2}h)}{S(x,y)}, \\ \beta_3(x,y) &= \frac{A(x-\frac{1}{2}h,y)}{S(x,y)}, & \beta_4(x,y) &= \frac{C(x,y-\frac{1}{2}h)}{S(x,y)}, \\ \tau(x,y) &= -h^2G(x,y)/S(x,y) \end{aligned}\right\} \qquad (2.8)$$

and

$$S(x,y) = A(x+\tfrac{1}{2}h,y)+A(x-\tfrac{1}{2}h,y)+C(x,y+\tfrac{1}{2}h)+C(x,y-\tfrac{1}{2}h)-h^2F(x,y). \qquad (2.9)$$

Therefore the problem of solving the discrete generalised Dirichlet problem reduces to the solution of a system of linear algebraic equations of the form

$$Au=b \qquad (2.10)$$

where there is one equation and one unknown for each of the N points of R_h. The row of the matrix corresponding to the point (x,y) has unity as the diagonal element and $\beta_i(x,y)$, i=1,2,3,4 in the column corresponding to a point of R_h properly adjacent to (x,y). Terms of (2.7) which do not involve values of u(x,y) on ∂R_h are brought to the left-hand side of the equation for the point (x,y), while the rest of the terms form the elements of the right hand side vector b in (2.10). Evidently, the order of the matrix A is N, the number of the mesh points in R_h. The matrix A in (2.10) is real and symmetric. Furthermore, it can be shown to be positive definite[†] and to have "Property A". In addition to these properties, it can be verified that A is an L-matrix, is irreducible and has weak diagonal dominance.

If h is very small (as this is the case we will primarily be concerned with), the problem of actually solving (2.10) may present serious practical difficulties even though a unique solution is known to exist. In this case the order of A is about 10^3 to 10^6 and on the other hand as we have seen A is "sparse" i.e., has only a few non-zero elements as compared to the total number of elements of A. These properties lead naturally (but not exclusively)

[†]*For definitions which are not given, see Chapter 2.*

to use iterative techniques for solving such systems of equations since they do not introduce new non-zero elements during the computation and therefore the sparseness of A is preserved. As a result of this, the problem of the accumulation of rounding errors is less serious than for those methods, such as most direct methods, where the matrix A is changed during the computation process.

CHAPTER 2

MATRIX PRELIMINARIES

In this chapter we will present various definitions and theorems, often without proof, from matrix theory which will be useful for reference purposes for our study of iterative methods. We have presupposed a basic knowledge of the general theory of matrices as presented, for instance, in Faddeev and Faddeeva [1963], Bellman [1960], Householder [1964] and Birkhoff and Maclane [1953].

2.1 BACKGROUND OF MATRIX THEORY

Definition 1.1

Given any two vectors $v=(v_1,v_2,\ldots,v_N)^T$ and $w=(w_1,w_2,\ldots,w_N)^T$ we define the inner product of v and w by

$$(v,w) = v^H w = \sum_{i=1}^{N} v_i^* w_i. \tag{1.1}$$

Theorem 1.1

The linear system

$$Au = b \tag{1.2}$$

has a unique solution if and only if A is non-singular.

If A is singular, then (1.2) either has no solution or else it has an infinite number of solutions.

Theorem 1.2

If A is a square matrix of order N with eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_N$, then

$$\det(A) = \prod_{i=1}^{N} \lambda_i, \quad \text{trace}(A) = \sum_{i=1}^{N} \lambda_i. \tag{1.3}$$

Definition 1.2

If A is an N×N matrix, then the spectral radius of A is

$$S(A) = \max_{\lambda \in S_A} |\lambda| \tag{1.4}$$

where S_A is the set of all eigenvalues of A.

We will use the following two theorems from the Perron-Frobenious theory of non-negative matrices.

Theorem 1.3

If $A \geq |B|$, then $S(A) \geq S(B)$.

Theorem 1.4

If $A \geq 0$, then $S(A)$ is an eigenvalue of A and there exists a non-negative eigenvector of A associated with $S(A)$.

Next, we give a useful theorem for determining bounds on the eigenvalues of any Hermitian matrix.

Theorem 1.5

If A is an Hermitian matrix and if λ_1 and λ_N are the largest and the smallest eigenvalues of A, respectively, then

$$\begin{aligned} \lambda_1 &= \max_{v \neq 0} \frac{(v,Av)}{(v,v)} = \frac{(v^{(1)},Av^{(1)})}{(v^{(1)},v^{(1)})}, \\ \lambda_N &= \min_{v \neq 0} \frac{(v,Av)}{(v,v)} = \frac{(v^{(N)},Av^{(N)})}{(v^{(N)},v^{(N)})} \end{aligned} \tag{1.5}$$

where $v^{(1)}$ and $v^{(N)}$ are eigenvalues of A corresponding to λ_1 and λ_N, respectively.

vectors

2.2 POSITIVE DEFINITE MATRICES

The property of a matrix being positive definite is essential in our study, so from the many definitions we will use the following.

Definition 2.1

A matrix A is positive definite if A is Hermitian and

$$(v,Av)>0 \tag{2.1}$$

for all $v\neq 0$. If $(v,Av)\geq 0$ for all v, then A is non-negative definite.

Evidently, one can give similar definitions for negative definite and non-positive definite matrices.

Further, we state a theorem which is sometimes used as a definition of positive (non-negative) definiteness.

Theorem 2.1

A matrix A is positive definite (non-negative definite) if and only if it is Hermitian and all of its eigenvalues are positive (non-negative). A method for constructing a positive definite matrix is given by the following theorem.

Theorem 2.2

For any matrix A the matrix AA^H is Hermitian and non-negative definite. If A is non-singular, then AA^H is positive definite.

Furthermore, the existence of the positive definite "square root" of a positive definite matrix A is guaranteed from the following theorem.

Theorem 2.3

If A is a positive definite (non-negative definite) matrix, then there exists a unique positive definite (non-negative definite) matrix B (denoted by $A^{\frac{1}{2}}$) such that

$$B^2 = A. \tag{2.2}$$

Definition 2.2

If there exists a non-singular matrix S such that

$$S^H AS = B, \tag{2.3}$$

we say that B is Hermitian congruent to A and that B is obtained from A by a Hermitian-congruence transformation.

It is important to note that the property of a matrix being positive definite is not affected by a congruence transformation.

Theorem 2.4

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

Proof

From Definition 2.2 we have

$$B = S^H AS,$$

also by Theorem 2.3 we can let $A^{\frac{1}{2}}$ be the positive definite matrix whose square is A. Thus, $B=(S^H A^{\frac{1}{2}})(S^H A^{\frac{1}{2}})^H$ and since $A^{\frac{1}{2}}S$ is non-singular, then from Theorem 2.2 it follows that B is positive definite.

Theorem 2.4 can also be proved if we consider the quadratic form of B i.e.,

$$(x,Bx) = (x,S^H ASx) = (Sx,ASx)>0$$

if $x\neq 0$.

2.3 VECTOR AND MATRIX NORMS

Definition 3.1

A vector norm $\|.\|_\alpha$ is a non-negative function on the space C^N, the set of all the vectors, with the following properties:

$$\begin{aligned}
&\text{a)} \quad \|x\|_\alpha > 0, \text{ if } x \neq 0 \\
&\text{b)} \quad \|x\|_\alpha = 0, \text{ if } x = 0 \\
&\text{c)} \quad \|cx\|_\alpha = |c| . \|x\|_\alpha \text{ for any complex number } c \\
&\text{d)} \quad \|x+y\|_\alpha \leq \|x\|_\alpha + \|y\|_\alpha \text{ for all vectors } x, y \in C^N \text{ (triangle inequality).}
\end{aligned} \tag{3.1}$$

There is an infinite number of vector norms and to illustrate this fact we consider the ℓ_p-norms (Hölder norms).

$$\|x\|_p = \begin{cases} \sqrt[p]{\sum_{i=1}^{N} |x_i|^p} & p=1,2,3,\ldots \\ \max_i |x_i| & p=\infty. \end{cases} \tag{3.2}$$

Among these norms the ℓ_1-norm, ℓ_2-norm and ℓ_∞-norm are the most familiar and widely used

$$\|x\|_1 = \sum_{i=1}^{N} |x_i| \tag{3.3}$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2} \tag{3.4}$$

$$\|x\|_\infty = \max_i |x_i| . \tag{3.5}$$

Definition 3.2

The matrix norm $\|.\|_\beta$ is a non-negative function on the space C^{NN}, the set of all the (N×N) matrices, with the following properties:

$$\begin{aligned}
&\text{a)} \quad \|A\|_\beta > 0, \text{ if } A \neq 0 \\
&\text{b)} \quad \|A\|_\beta = 0, \text{ if } A = 0 \\
&\text{c)} \quad \|cA\|_\beta = |c| . \|A\| \text{ for any complex number } c \\
&\text{d)} \quad \|A+B\|_\beta \leq \|A\|_\beta + \|B\|_\beta .
\end{aligned} \tag{3.6}$$

It can be shown that the quantities

$$\|A\|_\infty = \max_i \sum_{j=1}^{N} |a_{i,j}|, \qquad i=1,2,\ldots,N \tag{3.7}$$

$$\|A\|_1 = \max_j \sum_{i=1}^{N} |a_{i,j}|, \qquad j=1,2,\ldots,N \tag{3.8}$$

$$\|A\|_2 = [S(A^H A)]^{\frac{1}{2}} \tag{3.9}$$

and

$$\|A\|_M = N \max|a_{i,j}|, \qquad i,j=1,2,\ldots,N \tag{3.10}$$

are all matrix norms.

Definition 3.3

Given a vector norm $\|.\|_\alpha$, we define the induced matrix norm $\|.\|_{\beta(\alpha)}$ by

$$\|A\|_{\beta(\alpha)} = \max_{v\neq 0} \frac{\|Av\|_\alpha}{\|v\|_\alpha} . \tag{3.11}$$

It is interesting to note that (3.7),(3.8) and (3.9) can be proved to be induced matrix norms corresponding to the vector norms $\|.\|_\infty$, $\|.\|_1$ and $\|.\|_2$, respectively. The main advantage in choosing the induced matrix norm is that the inequality

$$\|Av\|_\alpha \leq \|A\|_\beta \|v\|_\alpha \tag{3.12}$$

is satisfied.

Definition 3.4

If the inequality (3.12) holds, then the vector norm $\|.\|_\alpha$ and the matrix norm $\|.\|_\beta$ are called consistent or compatible. Evidently, any vector norm and the induced matrix norm are consistent.

If the matrix norm $\|.\|_\beta$ and the vector norm are consistent and if for some $v\in C^N$ and $v\neq 0$ we have

$$\|Av\|_\alpha = \|A\|_\beta \|v\|_\alpha , \tag{3.13}$$

then the matrix norm is subordinate to the vector norm. It is obvious now that the induced matrix norm corresponding to a vector norm is subordinate to that vector norm. Throughout our study, we will

frequently use the vector norm $\|.\|_2$ and the corresponding induced matrix norm given by (3.9). Also, when no confusion will arise we will omit the norm suffices.

The following theorem provides a method for determining a good bound on the spectral radius of a matrix.

Theorem 3.1

For any matrix norm $\|.\|_\beta$ we have

$$S(A) \leq \|A\|_\beta . \tag{3.14}$$

Proof

Suppose that λ is an eigenvalue of A and v is an associated eigenvector, then $Av=\lambda v$ and from Definition 3.2 we have

$$\|\lambda v\|_\alpha = \|Av\|_\alpha \leq \|A\|_\beta \|v\|_\alpha .$$

Hence, by (3.1c) we obtain

$$|\lambda| \leq \|A\|_\beta .$$

Since this inequality holds for every eigenvalue, the spectral radius of A is bounded by every norm of A.

Definition 3.5

Given a matrix norm $\|.\|_\beta$ and any non-singular matrix S, then the "β,S-norm" of a matrix A is given by

$$\|A\|_{\beta,S} = \|SAS^{-1}\|_\beta . \tag{3.15}$$

Similarly, we define the "α,S-norm" of a vector v by

$$\|v\|_{\alpha,S} = \|Sv\|_\alpha . \tag{3.16}$$

2.4 CONVERGENCE OF SEQUENCES OF MATRICES

Definition 4.1

A sequence of matrices $A^{(1)}=(a_{i,j}^{(1)})$, $A^{(2)}=(a_{i,j}^{(2)}),\ldots$ converges to a matrix $A=(a_{i,j})$ if

$$\lim_{n\to\infty} (a_{i,j}^{(n)}) = a_{i,j}; \quad i,j=1,2,\ldots,N. \tag{4.1}$$

Theorem 4.1

The sequence $A^{(1)},A^{(2)},\ldots$ converges to a limit A if and only if for every matrix norm $\|.\|_\beta$, we have

$$\lim_{n\to\infty} \|A^{(n)}-A\|_\beta = 0. \tag{4.2}$$

An important condition for the convergence of the sequence of powers of a matrix is given by the following theorem.

Theorem 4.2

Given a matrix A, then $\lim_{n\to\infty} A^n=0$ if and only if

$$S(A)<1. \tag{4.3}$$

Theorem 4.3

The matrix I-A is non-singular and the series $I+A+A^2+\ldots$ converges if and only if $S(A)<1$. Moreover if $S(A)<1$, then

$$(I-A)^{-1} = I+A+A^2+\ldots = \sum_{i=0}^{\infty} A^i. \tag{4.4}$$

2.5 IRREDUCIBILITY AND WEAK DIAGONAL DOMINANCE

As we will see in the next chapter, the matrices which are obtained from the discretisation of certain partial differential equations (see Chapter 1) belong to two important classes of matrices which are considered in this section.

Definition 5.1

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

Another theorem which may well be used as a definition of irreducibility is the following.

Theorem 5.1

The matrix $A=(a_{i,j})$ is irreducible if and only if there does not exist a permutation matrix P such that $P^{-1}AP$ has the form

$$P^{-1}AP = \begin{pmatrix} F & O \\ G & H \end{pmatrix} \tag{5.1}$$

where F and H are square matrices and where O is the null matrix.

The concept of irreducibility is quite important, for by Theorem 5.1 we cannot reduce the matrix system (1.2) to the solution of two lower-order systems which preserve the correspondence between the equations and the unknowns, and which can be solved independently of the original system.

A useful method for verifying irreducibility in practice is given by the following theorem.

Theorem 5.2

A matrix of order N is irreducible if and only if N=1 or, given any two distinct integers i and j with $1\leq i\leq N$, $1\leq j\leq N$, then $a_{i,j}\neq 0$ or there exists $\ell_1,\ell_2,\ldots,\ell_r$ such that

$$a_{i,\ell_1}\, a_{\ell_1,\ell_2} \cdots a_{\ell_r,j} \neq 0. \tag{5.2}$$

Next, we illustrate the use of the geometrical interpretation of the concept of irreducibility by means of graphs. For a given matrix A of order N we consider the distinct points $P_1, P_2, \ldots, P_N$ and we construct the directed graph of A by drawing an arrow from P_i to P_j for each $a_{i,j} \neq 0$. If $a_{i,i} \neq 0$ we draw a small loop containing the point P_i. The matrix is irreducible if N=1 or else there exists a path of arrows from P_i to P_{ℓ_1}, P_{ℓ_1} to $P_{\ell_2}, \ldots, P_{\ell_r}$ to P_j (connected graph). As an example, let us consider the directed graph of a tri-diagonal matrix of order N

$$A = \begin{bmatrix} a_{11} & a_{12} & & & & 0 \\ a_{21} & a_{22} & a_{23} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ 0 & & & a_{N-1,N-2} & a_{N-1,N-1} & a_{N-1,N} \\ & & & & a_{N,N-1} & a_{N,N} \end{bmatrix} \tag{5.3}$$

The directed graph is given in Figure 5.1 where we can readily see that the graph is connected, thus the tri-diagonal matrices are irreducible.

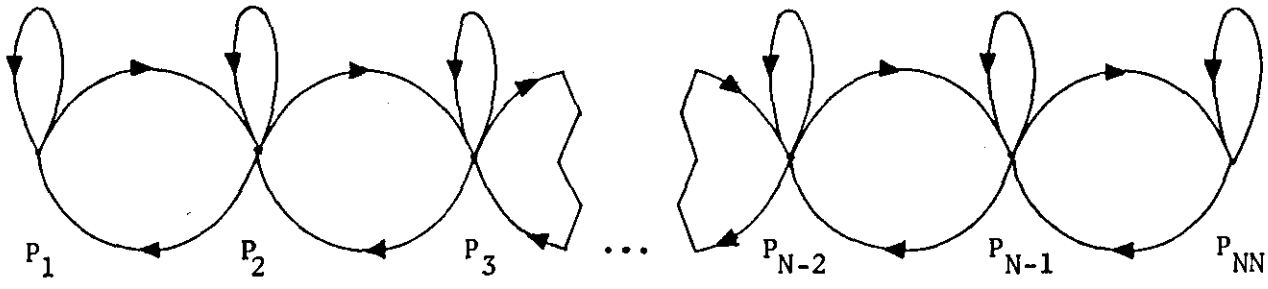


FIGURE 5.1

The other important class of matrices which also appears in the numerical solution of certain partial differential equations are those matrices which have diagonal dominance.

Definition 5.2

A matrix $A=(a_{i,j})$ of order N has weak diagonal dominance if

$$|a_{i,i}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}|, \quad i=1,2,\ldots,N \tag{5.4}$$

and for at least one i

$$|a_{i,i}| > \sum_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}|. \tag{5.5}$$

From Theorem 1.1 we see that when we consider the system (1.2), it is vital to establish whether the matrix A is non-singular. Since in our study it will be, in certain cases, quite difficult to use the criterion of the determinant, we state an alternative criterion given by the following fundamental theorem.

Theorem 5.3

If A is an irreducible matrix with weak diagonal dominance, then det $A\neq 0$ and none of the diagonal elements of A vanish.

Definition 5.3

If the inequality (5.5) holds for every i, then the matrix A has strong diagonal dominance.

Corollary 5.4

If the matrix A has strong diagonal dominance, then det $A\neq 0$.

Next, we give a sufficient condition for an Hermitian matrix to be positive definite using the properties of irreducibility and weak diagonal dominance.

Theorem 5.5

If A is an Hermitian matrix with non-negative diagonal elements and has weak diagonal dominance, then A is non-negative definite. If A is also irreducible or non-singular, then A is positive definite.

Proof

It is known that all the eigenvalues of an Hermitian matrix A are real. Let us consider an eigenvalue λ of A, then

$$\det(A-\lambda I) = 0. \tag{5.6}$$

If we now assume that $\lambda<0$, then the matrix $A-\lambda I$ has strong diagonal dominance, hence by Corollary 5.4, $\det(A-\lambda I)\neq 0$ which contradicts (5.6). Thus, all the eigenvalues of A are non-negative and by Theorem 2.1, A is non-negative definite. If A is irreducible, then from Theorem 5.3 it does not possess the eigenvalue $\lambda=0$. Therefore, if we impose the condition of irreducibility on A, then all its eigenvalues are positive, hence by Theorem 2.2, A is positive definite.

2.6 ORDERING VECTORS AND CONSISTENTLY ORDERED MATRICES

Definition 6.1

Given a matrix $A=(a_{i,j})$ the integers i and j are associated with respect to A if $a_{i,j}\neq 0$ or $a_{j,i}\neq 0$.

Definition 6.2

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 the matrix A of order N if for any pair of associated integers i and j with $i\neq j$ we have $|\gamma_i-\gamma_j|=1$.

Definition 6.3

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

a) $\gamma_i-\gamma_j=1$ if i and j are associated and $i>j$

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

In the above definitions, we have established the concept of the ordering vector and the compatible ordering vector for a given matrix A. Alternatively, we will show that the existence of a compatible ordering or an ordering vector characterises the class of matrices called "consistently ordered" or the wider class, those having "Property A", respectively.

Definition 6.4

The matrix A of order N is consistently ordered if for some t there exist disjoint subsets $S_1,S_2,\ldots,S_t$ of $W=\{1,2,\ldots,N\}$ such that $\sum_{k=1}^{t} S_k=W$ and such that if i and j are associated, then $j\in S_{k+1}$ if $j>i$ and $j\in S_{k-1}$ if $j<i$, where S_k is the subset containing i.

We now consider a matrix where the conditions of the above definition are satisfied. Let the matrix A have the form

$$A=\begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & 0 \\ 0 & a_{23} & a_{33} \end{bmatrix}, \qquad (6.1)$$

if $S_1=\{1\}$, $S_2=\{2\}$, $S_3=\{3\}$, $S_4=\{0\}$, then we see that A is a consistently ordered matrix by Definition 6.4.

The matrix (6.1) belongs to a more general class of matrices which have the block tri-diagonal form

$$\begin{bmatrix} D_1 & F_1 & & & \\ E_1 & D_2 & F_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & E_{m-2} & D_{m-1} & F_{m-1} \\ 0 & & & E_{m-1} & D_m \end{bmatrix} \tag{6.2}$$

where D_i are square diagonal matrices. We will refer to such matrix as a T-matrix.

A useful result concerning the T-matrices is given by the following theorem.

Theorem 6.1

If A is a T-matrix, then A is consistently ordered.

If a matrix is not a T-matrix, then we use the concept of the compatible vector in order to verify whether it is a consistently ordered matrix.

Theorem 6.2

A matrix of order N is consistently ordered if and only if there exists a compatible ordering vector.

If the matrix is consistently ordered, we can construct a compatible ordering vector by letting $\gamma_i=k$ if $i\in S_k$ for $i=1,2,\ldots,N$. Furthermore, if a compatible vector exists, we can construct t disjoint subsets $S_1,S_2,\ldots,S_t$ as follows. First we find $\alpha=\min\{\gamma_1,\gamma_2,\ldots,\gamma_N\}$, $\beta=\max\{\gamma_1,\gamma_2,\ldots,\gamma_N\}$ and we determine $t=\beta-\alpha+1$. Further, we let S_k be the set of all i such that $\gamma_i=\alpha+k-1$.

As an application of this case we can verify the correspondence between the sets $S_1=\{1\}$, $S_2=\{2\}$, $S_3=\{3\}$ and the compatible ordering vector $\gamma=(1,2,3)^T$ for the matrix A given by (6.1).

2.7 PROPERTY A

It is clear from Chapter 1 that we will consider matrices whose non-zero elements form a certain pattern. In this section we will continue to look at such matrices and we will define a wider class of matrices, those having "Property A".

Definition 7.1

A matrix $A=(a_{i,j})$ of order N has Property A if there exist two disjoint subsets S_1 and S_2 of W and such that if $i\neq j$ and if either $a_{i,j}\neq 0$ or $a_{j,i}\neq 0$, then $i\in S_1$ and $j\in S_2$ or else $i\in S_2$ and $j\in S_1$.

In Section 2.6, we have seen the necessity for the existence of a compatible ordering vector for the matrix A to be consistently ordered. Next, we state a theorem which provides a similar criterion for a matrix to have Property A.

Theorem 7.1

There exists an ordering vector for a matrix A if and only if A has Property A. Moreover, if A is consistently ordered, then A has Property A.

The next theorem can be regarded as an alternative definition of Property A.

Theorem 7.2

A matrix A has Property A if and only if A is a diagonal matrix or else there exists a permutation matrix P such that $P^{-1}AP$ has the form

$$A' = P^{-1}AP = \begin{pmatrix} D_1 & H \\ K & D_2 \end{pmatrix} \tag{7.1}$$

where D_1 and D_2 are square diagonal matrices.

The following theorem presents a method for the construction of a consistently ordered matrix which has Property A.

Theorem 7.3

Let A be a matrix with Property A and let γ be any ordering vector for A. There exists a permutation matrix P such that $A'=P^{-1}AP$ is consistently ordered and such that $\gamma'=(\gamma_{\sigma^{-1}(i)})$ is a compatible ordering vector for A', where σ is the permutation corresponding to P.

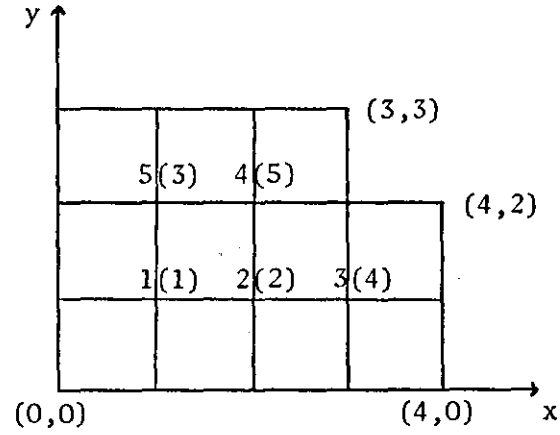


FIGURE 7.1

In order to illustrate the applicability of the above Theorem 7.3 we consider the five point discrete analogue of the Dirichlet problem for the region shown in Figure 7.1 with mesh size h=1. The corresponding matrix A can be readily seen to be the following

$$A = \begin{pmatrix} 4 & -1 & 0 & 0 & -1 \\ -1 & 4 & -1 & -1 & 0 \\ 0 & -1 & 4 & 0 & 0 \\ 0 & -1 & 0 & 4 & -1 \\ -1 & 0 & 0 & -1 & 4 \end{pmatrix} . \tag{7.2}$$

One can verify that $\gamma=(1,2,3,3,2)^T$ is an ordering vector for A and there does not exist a compatible vector for A. Furthermore, we have $\alpha=1$, $\beta=3$ and $t=3$. It is easy now to construct S_k to be the set containing all the i for which $\gamma_i=k$, $k=1,2,3$. Thus we have $S_1=\{1\}$, $S_2=\{2,5\}$, $S_3=\{3,4\}$ and the permutation is $\sigma(1)=1$, $\sigma(2)=2$, $\sigma(5)=3$, $\sigma(3)=4$, $\sigma(4)=5$.

The corresponding permutation matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{7.3}$$

and finally we find

$$A' = P^T A P = \left(\begin{array}{c|cc|cc} 4 & -1 & -1 & 0 & 0 \\ \hline -1 & 4 & 0 & -1 & -1 \\ -1 & 0 & 4 & 0 & -1 \\ \hline 0 & -1 & 0 & 4 & 0 \\ 0 & -1 & -1 & 0 & 4 \end{array}\right) \tag{7.4}$$

which is a consistently ordered matrix by Theorem 6.1.

Finally, the compatible ordering vector is given by $\gamma'=(1,2,2,3,3)^T$. It is evident now that we have to relabel the mesh points of our grid along the diagonals. This has been indicated by relabelling the points in parentheses in Figure 7.1.

A thorough discussion of irreducibility, diagonal dominance, consistently ordered matrices, Property A and their generalisations can be found in Young [1971] and Varga [1962].

Finally, we give some additional definitions which will be used in subsequent chapters and characterise other classes of matrices.

Definition 7.2

A real matrix A of order N is an L-matrix if

$$a_{i,i} > 0, \qquad i=1,2,\ldots,N \tag{7.5}$$

and

$$a_{i,j} \leq 0, \qquad i \neq j, \quad i,j=1,2,\ldots,N. \tag{7.6}$$

Definition 7.3

A real matrix A is a Stieltjes matrix if A is positive definite and if (7.6) holds.

Definition 7.4

A real matrix A is an M-matrix if (7.6) holds, if A is non-singular and if $A^{-1} \geq 0$.

CHAPTER 3

LINEAR STATIONARY AND NON-STATIONARY ITERATIVE METHODS

3.1 INTRODUCTION

We have seen (Chapter 1) that the use of finite difference methods for solving the generalised Dirichlet problem may lead to a linear system of the form

$$Au = b \tag{1.1}$$

where A is a given N×N matrix and b is a given vector. The order of the matrix equals the number of interior mesh points and may be so large that it may be impractical to store the matrix even in a large computer or to solve the system by direct methods. On the other hand, since the matrix is sparse it is usually possible to store all of the non-zero elements and apply various iterative methods.

Definition 1.1

The sequence of functions $q_0(A,b)$, $q_1(u^{(0)};A,b)$, $q_2(u^{(0)},u^{(1)};A,b),\ldots,$ $q_k(u^{(0)},u^{(1)},\ldots,u^{(k-1)};A,b)$, where

$$\begin{aligned} u^{(0)} &= q_0(A,b), \\ u^{(n+1)} &= q_{n+1}(u^{(0)},u^{(1)},\ldots,u^{(n)};A,b), \end{aligned} \tag{1.2}$$

is said to be an iterative method.

We call the iterative method stationary if q_n is independent of n for all n>m, where n,m are positive integers, otherwise it is non-stationary. If $u^{(n+1)}=q(u^{(n-1)},u^{(n-2)},\ldots,u^{(n-m)};A,b)$, then the degree of the method is m. Finally, if q_n is a linear function of $u^{(0)},u^{(1)}$, $u^{(2)},\ldots,u^{(n-1)}$, then the method is called linear, otherwise it is non-linear. The form of a linear stationary iterative method of first degree is

$$u^{(n+1)} = Gu^{(n)}+k. \tag{1.3}$$

Furthermore, it is desirable for any iterative method to satisfy the following requirements:

a) If at any stage we obtain a solution of (1.1), then the subsequent iterants remain unchanged (consistency).

b) If the sequence of vectors defined by (1.3) converges, then it converges to a solution of (1.1) (reciprocally consistent).

The conditions under which the above restrictions hold, are given by the following theorems.

Theorem 1.1

If A is non-singular, then the iterative process (1.3) is consistent with (1.1) if and only if

$$k = (I-G)A^{-1}b. \tag{1.4}$$

Theorem 1.2

If I-G is non-singular, then the method is reciprocally consistent if and only if

$$b = A(I-G)^{-1}k. \tag{1.5}$$

If both our requirements are valid then the iterative method is completely consistent with the system (1.1) in the sense that the only solution of the related linear system

$$u = Gu + k \tag{1.6}$$

is the solution $\bar{u}$ of (1.1).

Theorem 1.3

If A is non-singular, then the iterative method (1.3) is completely consistent with (1.1) if and only if it is consistent and I-G is non-singular. If I-G is non-singular, then complete consistency holds if and only if the method (1.3) is reciprocally consistent and A is non-singular.

Finally, a more general case is covered by the following theorem.

Theorem 1.4

The method (1.3) is completely consistent with (1.1) if and only if a non-singular matrix Q exists such that

$$G = I-QA, \quad k=Qb. \tag{1.7}$$

3.2 LINEAR STATIONARY ITERATIVE METHODS

In this section we will consider known linear stationary methods which will be used later for comparison purposes. Let us seek the solution of the linear system (1.1) where it is assumed that the diagonal elements of A are non-zero. The matrix A can be expressed as the matrix sum

$$A = D - C_L - C_U \tag{2.1}$$

where D=diag A and C_L, C_U are respectively strictly lower and upper triangular N×N matrices, whose entries are the negatives of the entries of A below and above its main diagonal of A, respectively. We can rewrite (1.1) by substituting A from (2.1) as follows

$$Du = (C_L+C_U)u + b. \tag{2.2}$$

Since the diagonal elements of A do not vanish, then D^{-1} exists, thus we can replace the system (2.2) by the equivalent system

$$u = Bu + c \tag{2.3}$$

where
$$B = D^{-1}C = L + U, \tag{2.4}$$

$$c = D^{-1}b \tag{2.5}$$

and
$$L = D^{-1}C_L,\ U = D^{-1}C_U \tag{2.6}$$

with
$$D = \text{diag } A,\quad C = D - A. \tag{2.7}$$

The Jacobi method (J method) can now be defined by commencing with an arbitrary vector $u^{(0)}$ and then computing a sequence of vectors $u^{(1)}, u^{(2)}, \ldots$ from the relationship

$$u^{(n+1)} = Bu^{(n)} + c. \tag{2.8}$$

An examination of this iterative method indicates the requirement to save the vector $u^{(n)}$ while computing $u^{(n+1)}$. The J method is consistent with (1.1) since by Theorem 1.1 and the relationships (2.4),(2.5), (2.7) we have

$$I-B = I-D^{-1}C = D^{-1}(D-C) = D^{-1}A \tag{2.9}$$

thus
$$(I-B)A^{-1}b = D^{-1}b = c. \tag{2.10}$$

If I-B is a non-singular matrix, then from Theorem 1.2 the J method

is completely consistent. Hence by (2.9) we require A to be a non-singular matrix. The assumption of the existence of a unique solution and therefore, by Theorem 2-1.1, the requirement of A being non-singular, is intuitively connected with the concept of complete consistency. We will therefore assume in the remainder of this thesis that A is non-singular (although we may often recall this assumption for emphasis). A modified version of the J method is the simultaneous overrelaxation method (JOR method) which is defined with the introduction of a real parameter ω by

$$u^{(n+1)} = \omega(Bu^{(n)}+c) + (1-\omega)u^{(n)} \tag{2.11}$$

or, equivalently

$$u^{(n+1)} = u^{(n)} + \omega(Bu^{(n)}+c-u^{(n)}). \tag{2.12}$$

A more compact form is given by

$$u^{(n+1)} = B_\omega u^{(n)} + \omega c \tag{2.13}$$

where

$$B_\omega = \omega B + (1-\omega)I. \tag{2.14}$$

Evidently,

$$I - B_\omega = \omega D^{-1}A \tag{2.15}$$

hence $I-B_\omega$ is non-singular, if $\omega \neq 0$.

Furthermore, by (2.14) we have

$$(I-B_\omega)A^{-1}b = \omega c \tag{2.16}$$

thus by Theorem 1.1 the JOR method is completely consistent with (1.1) for $\omega \neq 0$. If $\omega=1$, we see that the JOR method coincides with the J method. From (2.12) we note that the JOR method can be regarded as a form of extrapolation of the J method, at least when $\omega>1$. The role of the parameter ω will be considered later in Section 3.6. Both the J and the JOR methods are clearly independent of the order in which the mesh points are scanned, since the arithmetic is not affected by the different orderings.

It would seem to be more attractive to use the latest estimates of the components of $u^{(n+1)}$ as soon as they are available instead of $u^{(n)}$ in (2.8). This results in the following iterative scheme

$$u^{(n+1)} = Lu^{(n+1)} + Uu^{(n)} + c. \tag{2.17}$$

By definition we have $\det(I-L)=1$, therefore $(I-L)^{-1}$ exists and we can solve (2.17) for $u^{(n+1)}$ obtaining

$$u^{(n+1)} = Lu^{(n)} + \ell \tag{2.18}$$

where $$L = (I-L)^{-1}U \tag{2.19}$$

and $$\ell = (I-L)^{-1}c. \tag{2.20}$$

The above iterative procedure is known as the Gauss-Seidel method (GS method). If we now apply the same technique to extrapolate the GS method as we did with the J method, then we can produce the following iterative scheme which is known as the successive overrelaxation method (SOR method)

$$u^{(n+1)} = u^{(n)} + \omega(Lu^{(n+1)}+Uu^{(n)}+c-u^{(n)}) \tag{2.21}$$

where ω is a real parameter known as the relaxation factor. The matrix $I-\omega L$ is non-singular, hence $(I-\omega L)^{-1}$ exists and (2.21) can be rewritten to yield

$$u^{(n+1)} = L_\omega u^{(n)} + \ell_\omega \tag{2.22}$$

where $$L_\omega = (I-\omega L)^{-1}(\omega U+(1-\omega)I) = I-\omega(I-\omega L)^{-1}D^{-1}A \tag{2.23}$$

and $$\ell_\omega = \omega(I-\omega L)^{-1}c. \tag{2.24}$$

Evidently,

$$I-L_\omega = \omega(I-\omega L)^{-1}D^{-1}A \tag{2.25}$$

hence $I-L_\omega$ is non-singular if $\omega\neq 0$. From (2.25) we have

$$(I-L_\omega)A^{-1}b = \ell_\omega \tag{2.26}$$

which by Theorem 1.1, indicates that the SOR method is completely consistent with (1.1) for $\omega\neq 0$. If $\omega=1$, then we see that the SOR method reduces exactly to the GS method. Finally, we note that unlike the J and JOR methods, the GS and SOR methods depend upon the order in which the points are scanned in the mesh (Young [1954]).

Another iterative scheme which involves the residual vector

$$r^{(n)} = b-Au^{(n)} \tag{2.27}$$

is the Generalised Simultaneous Displacement method (GSD method) defined by

$$u^{(n+1)} = u^{(n)} + R(b-Au^{(n)}) \tag{2.28}$$

where R is any non-singular diagonal matrix.

The GSD method can be written in a more compact form to yield

$$u^{(n+1)} = R_x u^{(n)} + Rb, \tag{2.29}$$

where

$$R_x = I - RA. \tag{2.30}$$

By Theorem 1.4 and the relationships (2.29) and (2.30) we can easily verify that the GSD method is completely consistent. A useful observation is that if $R=D^{-1}$, then the GSD method reduces exactly to the J method. Further, if we let $R=\hat{\alpha}I$, then we obtain the Simultaneous Displacement method (SD method) (Forsythe and Wasow [1960])

$$u^{(n+1)} = u^{(n)} + \hat{\alpha}(b-Au^{(n)}) \tag{2.31}$$

or

$$u^{(n+1)} = R_{\hat{\alpha}} u^{(n)} + \hat{\alpha}b, \tag{2.32}$$

where

$$R_{\hat{\alpha}} = I - \hat{\alpha}A.$$

By Theorem 1.4 we have that the SD method is completely consistent with (1.1) if $\hat{\alpha}\neq 0$. This method was first considered by Richardson [1960] but $\hat{\alpha}$ was varied in each iteration $\hat{\alpha}=\hat{\alpha}_n$ resulting in a non-stationary iterative scheme. By letting $R=\omega D^{-1}$ it follows that the GSD method degenerates to the JOR method. It has therefore been verified that the GSD method is a generalisation of the J, SD and JOR methods for various forms of the matrix R.

Next, we consider a modification of the SOR method which results in the symmetric SOR method (SSOR method). Each iteration of the SSOR method consists of two half-iterations. The first half is just the ordinary SOR iteration while the second is an SOR iteration which scans the mesh in reverse order. Consequently, the SSOR iterative scheme is defined by

$$u^{(n+1/2)} = u^{(n)} + \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n)} + c - u^{(n)}) \tag{2.33}$$

and

$$u^{(n+1)} = u^{(n+\frac{1}{2})} + \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n+1)} + c - u^{(n+\frac{1}{2})}) \tag{2.34}$$

where again ω is a real parameter and $u^{(n+\frac{1}{2})}$ is an intermediate approximation to the solution.

Evidently, (2.33) and (2.34) can be written alternatively to yield

$$u^{(n+\frac{1}{2})} = L_\omega u^{(n)} + \omega(I-\omega L)^{-1}c \tag{2.35}$$

and

$$u^{(n+1)} = U_\omega u^{(n+\frac{1}{2})} + \omega(I-\omega U)^{-1}c, \tag{2.36}$$

where L_ω is given by (2.23) and

$$U_\omega = (I-\omega U)^{-1}(\omega L+(1-\omega)I) = I-\omega(I-\omega U)^{-1}D^{-1}A. \tag{2.37}$$

Finally, the SSOR method can be written in a more compact form by eliminating $u^{(n+\frac{1}{2})}$ from (2.35) and (2.36) to yield

$$u^{(n+1)} = \xi_\omega u^{(n)} + k_\omega \tag{2.38}$$

where

$$\begin{aligned}\xi_\omega = U_\omega L_\omega &= (I-\omega U)^{-1}(\omega L+(1-\omega)I)(I-\omega L)^{-1}(\omega U+(1-\omega)I) \\ &= I-\omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \end{aligned} \tag{2.39}$$

and

$$k_\omega = \omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}c \tag{2.40}$$

It can be easily verified that the above method is completely consistent with (1.1) if $\omega \neq 0,2$ and A is non-singular. Since one SSOR iteration combines two SOR iterations we note that the SSOR process is dependent upon the order in which the points are scanned in the mesh. Further, we see from (2.33) and (2.34) that the SSOR method requires twice as much work as the SOR method. However, it can be shown (Niethammer [1964], Conrad and Wallach [1977]) that the work can be reduced to become identical with the SOR method. The SSOR method was first considered by Sheldon [1955] and it is a generalisation of the Aitken method (Aitken [1950]).

3.3 CONVERGENCE OF ITERATIVE METHODS

In this section we will consider under which conditions the sequence of vectors $u^{(0)},u^{(1)},\ldots$ produced by the iterative process (1.3) converges for any arbitrary starting vector $u^{(0)}$.

Definition 3.1

The iterative method (1.3) is convergent if the sequence $u^{(0)},u^{(1)},\ldots$ converges to a limit $\bar{u}$ for all the starting vectors $u^{(0)}$.

Theorem 3.1

The iterative method (1.3) converges if and only if

$$S(G) < 1. \tag{3.1}$$

Proof

Let us assume that the sequence $u^{(0)},u^{(1)},\ldots$ produced by the iterative method (1.3) converges to the solution vector $\bar{u}$, then from (1.3) and (1.6) we have

$$\varepsilon^{(n+1)} = G\varepsilon^{(n)} \tag{3.2}$$

where

$$\varepsilon^{(n)} = u^{(n)} - \bar{u}. \tag{3.3}$$

Evidently, from (3.2) we have the relationship

$$\varepsilon^{(n)} = G^n \varepsilon^{(0)}. \tag{3.4}$$

Moreover, $\lim_{n\to\infty} \varepsilon^{(n)}$ exists if and only if $\lim_{n\to\infty} u^{(n)}$ exists and $\lim_{n\to\infty} u^{(n)}=\bar{u}$ if and only if $\lim_{n\to\infty} \varepsilon^{(n)}=0$. Thus, by Theorem 2-4.2, we have that $\lim_{n\to\infty} \varepsilon^{(n)}=0$ if and only if the inequality (3.1) is satisfied. Let us now assume that (3.1) holds, then I-G is a non-singular matrix, hence the iterative method (1.3) is consistent and by Theorem 2-4.2 we have that $\lim_{n\to\infty} \varepsilon^{(n)}=0$, hence the proof of the theorem is complete.

In order to prove that an iterative method is convergent, it is preferable to consider other conditions besides (3.1) since it is sometimes laborious to evaluate the spectral radius of a matrix. An alternative condition for convergence is given by the following theorem, where A is a positive definite matrix.

Theorem 3.2

If A is a positive definite matrix and if the iterative method (1.3) is completely consistent with (1.1), then the method is convergent if R is a non-singular matrix and

$$M = R + R^T - A \tag{3.5}$$

is positive definite, where R satisfies the relationship

$$G = I - R^{-1}A. \tag{3.6}$$

Moreover, we have

$$\|G\|_{A^{\frac{1}{2}}} < 1. \tag{3.7}$$

Conversley, if (3.7) holds, then M is a positive definite matrix.

3.4 RATE OF CONVERGENCE

In order to evaluate the effectiveness of an iterative method we have to consider both the work required per iteration and the rate for the sequence of vectors to converge to the required solution.

In practice a "measure" for the latter is defined by requiring the norm of $\varepsilon^{(n)}$ to be reduced to a fraction ρ of the norm of the original vector $\varepsilon^{(0)}$. From (3.4), using matrix norms, we have the inequality

$$\|\varepsilon^{(n)}\| \leq \|G^n\| \, \|\varepsilon^{(0)}\|. \tag{4.1}$$

Then if $u^{(0)} \neq \bar{u}$ we have

$$\|\varepsilon^{(n)}\| / \|\varepsilon^{(0)}\| \leq \|G^n\|. \tag{4.2}$$

Since it is assumed that the method (1.3) is convergent, we require

$$\|\varepsilon^{(n)}\| \leq \rho \|\varepsilon^{(0)}\| \tag{4.3}$$

and we select n such that the following inequality is satisfied

$$\|G^n\| \leq \rho. \tag{4.4}$$

In order for (4.4) to hold for all n sufficiently large such that $\|G^n\|<1$, it follows that (4.4) is equivalent to

$$n \geq -\log\rho / (-\tfrac{1}{n} \log\|G^n\|). \tag{4.5}$$

From the inequality (4.5) we obtain a lower bound on the number of iterations for the iterative method (1.3). Furthermore, from (4.5) we conclude that the number of iterations n depends inversely on the expression $\frac{1}{n}\log\|G^n\|$ and therefore this quantity serves as a basis of comparison for the different iterative schemes.

Definition 4.1

For any convergent iterative method of the form (1.3) the quantity

$$R_n(G) = -\tfrac{1}{n} \log\|G^n\| \tag{4.6}$$

is the average rate of convergence.

It can be shown (Varga [1962], Young [1971]) that

$$S(G) = \lim_{n\to\infty} (\|G^n\|)^{1/n}, \tag{4.7}$$

hence we have the following definition.

Definition 4.2

For any convergent iterative method of the form (1.3) the quantity

$$R(G) = \lim_{n\to\infty} R_n(G) = -\log S(G) \tag{4.8}$$

is the asymptotic average rate of convergence or simply the rate of convergence. Finally, we define the quantity

$$RR(G) = 1/R(G) \tag{4.9}$$

as the reciprocal rate of convergence of the method (1.3). From (4.5) the number of iterations required for convergence is approximately proportional to the reciprocal rate of convergence.

3.5 SOME THEOREMS ON THE CONVERGENCE

In this section we will state some known theorems, often without proof, which provide certain conditions for the convergence of the iterative schemes developed in Section 3.2. First , we impose the restriction on the matrix A of the system (1.1) that all its diagonal elements are different from zero, i.e. diagA is non-singular. Then we can determine the range of ω for the JOR method to converge from the following theorem.

Theorem 5.1

If the J method converges, then the JOR method converges for $0<\omega\leq 1$.

Another theorem which concerns the SOR method and has been proven by Kahan [1958], is the following.

Theorem 5.2

If L_ω is defined by (2.23), then,

$$S(L_\omega) \geq |\omega-1| \tag{5.1}$$

for all real ω, with equality only if all the eigenvalues of L_ω are of modulus $|\omega-1|$. Moreover, if the SOR method converges, then

$$0 < \omega < 2. \tag{5.2}$$

If we require the matrix A to be non-singular, then from Theorem 2-5.3 the sufficient condition for this restriction is to assume that the matrix A is irreducible and has weak diagonal dominance. Under these properties, Geiringer [1949] has proved the following theorem.

Theorem 5.3

If A is an irreducible matrix and has weak diagonal dominance, then

(i) the J method converges and the JOR method converges for $0<\omega\leq 1$;

(ii) the GS method converges and the SOR method converges for $0<\omega\leq 1$.

Let us proceed in our study of the conditions under which certain iterative methods converge and suppose that A is a positive definite matrix.

In particular, let us consider the application of Theorem 3.2 to the iterative methods considered in Section 3.2, then the following theorem can be verified (Young [1971]).

Theorem 5.4

If A is a positive definite matrix and D=diagA, then

(a) $\|B\|_{A^{\frac{1}{2}}}<1$ if 2D-A is positive definite;

(b) $\|B_\omega\|_{A^{\frac{1}{2}}}<1$ if $\frac{2}{\omega}D-A$ is positive definite or, equivalently, if

$$0<\omega<2/(1-m(B))\leq 2 \tag{5.3}$$

where $m(B)\leq 0$ is the smallest eigenvalue of B;

(c) $\|L\|_{A^{\frac{1}{2}}}<1$;

(d) $\|L_\omega\|_{A^{\frac{1}{2}}}<1$ if $0<\omega<2$;

(e) $\|R\|_{A^{\frac{1}{2}}}<1$ if $2R^{-1}-A$ is positive definite;

(f) $\|R_{\hat{\alpha}}\|_{A^{\frac{1}{2}}}<1$ if $0<\hat{\alpha}<2/M(A)\leq 2$ (5.4)

where M(A) is the largest eigenvalue of A.

Stein and Rosenberg [1948] have developed an analysis on the convergence of the J,JOR,GS and SOR methods when the matrix A is an L-matrix. Their analysis is summarised in the following theorem.

Theorem 5.5

If A is an L-matrix and if $0<\omega\leq 1$, then

(a) $S(B)<1$ if and only if $S(L_\omega)<1$.

(b) $S(B)<1$ (and $S(L_\omega)<1$) if and only if A is an M-matrix;

if $S(B)<1$, then

$$S(L_\omega)\leq 1-\omega+\omega S(B) \tag{5.5}$$

(c) if $S(B)\geq 1$ and $S(L_\omega)\geq 1$, then

$$S(L_\omega)\geq 1-\omega+\omega S(B)\geq 1. \tag{5.6}$$

On the other hand, the conditions under which the SSOR method converges have been summarised in the following theorem.

Theorem 5.6

Let A be a symmetric matrix with positive diagonal elements, then the eigenvalues of $\mathcal{L}_\omega$ are real and non-negative for any real ω. Moreover, if A is positive definite and $0<\omega<2$, then

$$\|\mathcal{L}_\omega\|_{A^{\frac{1}{2}}} = S(\mathcal{L}_\omega) = \|L_\omega\|^2_{A^{\frac{1}{2}}} < 1. \tag{5.7}$$

Conversely, if $S(\mathcal{L}_\omega)<1$, then $0<\omega<2$ and A is positive definite.

Proof

The matrix $\mathcal{L}_\omega$ is similar to

$$\tilde{\mathcal{L}}_\omega = (I-\omega U)\mathcal{L}_\omega(I-\omega U)^{-1} \tag{5.8}$$

and from (2.39) $\tilde{\mathcal{L}}_\omega$ is equivalent to

$$\begin{aligned}\tilde{\mathcal{L}}_\omega &= (\omega L+(1-\omega)I)(I-\omega L)^{-1}(\omega U+(1-\omega)I)(I-\omega U)^{-1} \\ &= [(I-\omega L)^{-1}(\omega L+(1-\omega)I)][(I-\omega L)^{-1}(\omega L+(1-\omega)I]^T\end{aligned} \tag{5.9}$$

which by Theorem 2-2.2 implies that $\tilde{\mathcal{L}}_\omega$ is a non-negative definite matrix. Thus, all the eigenvalues of $\mathcal{L}_\omega$ are real and non-negative for any real ω. Moreover, if we make the assumption that A is positive definite and $0<\omega<2$, then by Theorem 2-2.3 there exists $A^{\frac{1}{2}}$, the unique positive definite matrix such that $(A^{\frac{1}{2}})^2=A$. From (2.39),(2.6) we can also have

$$\begin{aligned}\mathcal{L}_\omega &= I-\omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \\ &= I - \frac{2-\omega}{\omega}(\tfrac{1}{\omega}D-C_U)^{-1}D(\tfrac{1}{\omega}D-C_L)^{-1}A\end{aligned} \tag{5.10}$$

and

$$\mathcal{L}'_\omega = A^{\frac{1}{2}}\mathcal{L}_\omega A^{-\frac{1}{2}} = I-\frac{2-\omega}{\omega}\{[A^{\frac{1}{2}}(\tfrac{1}{\omega}D-C_U)^{-1}D^{\frac{1}{2}}][A^{\frac{1}{2}}(\tfrac{1}{\omega}D-C_U)^{-1}D^{\frac{1}{2}}]^T\}. \tag{5.11}$$

Since $0<\omega<2$ and $A^{\frac{1}{2}}(\frac{1}{\omega}D-C_U)^{-1}D^{\frac{1}{2}}$ is a non-singular matrix, then by Theorem 2-2.2 the second term in the right hand side of (5.11) is a positive definite matrix. Hence, from (5.11) all the eigenvalues of $\mathcal{L}'_\omega$ and therefore of $\mathcal{L}_\omega$ are less than unity. Moreover, if

$$L'_\omega = A^{\frac{1}{2}}L_\omega A^{-\frac{1}{2}} \quad \text{and} \quad U'_\omega = A^{\frac{1}{2}}U_\omega A^{-\frac{1}{2}}, \tag{5.12}$$

then by (2.23) and (2.37) we can easily verify that

$$(L'_\omega)^T = U'_\omega. \tag{5.13}$$

Therefore, from the above analysis and (2.39) we have

$$\|L_\omega\|^2_{A^{\frac{1}{2}}} = \|L'_\omega\|^2 = S(L'_\omega(L'_\omega)^T) = S(L'_\omega U'_\omega)$$

$$= S(U'_\omega L'_\omega) = S(\xi'_\omega) = \|\xi'_\omega\| = \|\xi_\omega\|_{A^{\frac{1}{2}}} < 1 \tag{5.14}$$

and the first part of the theorem has been proved.

Let us now make the assumption that $S(\xi_\omega)<1$. If λ_i $i=1,2,\ldots,N$ are the eigenvalues of ξ_ω, then by Theorem 2-1.2 we have the following result

$$\prod_{i=1}^{N}\lambda_i = \det(\xi_\omega) = \det((I-\omega U)^{-1}(\omega L+(1-\omega)I)(I-\omega L)^{-1}(\omega U+(1-\omega)I))$$
$$= (1-\omega)^{2N}. \tag{5.15}$$

Thus, we finally obtain

$$|1-\omega|^{2N} = \left|\prod_{i=1}^{N}\lambda_i\right| \leq \prod_{i=1}^{N}|\lambda_i| \leq S(\xi_\omega)^N$$

or

$$S(\xi_\omega) \geq |1-\omega|^2 \tag{5.16}$$

Evidently, for (5.7) to hold it follows from (5.16) that we must have $|1-\omega|<1$ or $0<\omega<2$ since ω is real. Furthermore, we seek to prove that A is positive definite. By (5.8) and (5.10) we have

$$\tilde{\xi}_\omega = I-\omega(2-\omega)D^{\frac{1}{2}}(D-\omega C_L)^{-1}A(D-\omega C_U)^{-1}D^{\frac{1}{2}}$$
$$= I-\omega(2-\omega)A^*, \tag{5.17}$$

where
$$A^* = D^{\frac{1}{2}}(D-\omega C_L)^{-1}A(D-\omega C_U)^{-1}D^{\frac{1}{2}}. \tag{5.18}$$

If ν_i are the eigenvalues of A*, then by (5.17) we have the following eigenvalue relationship

$$\lambda_i = 1-\omega(2-\omega)\nu_i. \tag{5.19}$$

Since now

$$0 \leq \lambda_i < 1, \tag{5.20}$$

then from (5.19),(5.20) we obtain

$$\nu_i > 0 \tag{5.21}$$

which by Theorem 2-2.1 implies that A* is positive definite. Moreover, from Theorem 2-2.1 there exists $v\neq 0$ such that

$$(v,A^*v) = (w,Aw) > 0, \tag{5.22}$$

where $$w = (D-\omega C_U)^{-1} D^{\frac{1}{2}} v.$$

Evidently, from (5.22) it follows that the matrix A is positive definite, (see Definition 2-2.1) hence the proof of the theorem is complete.

Corollary 5.7

Under the hypotheses of Theorem 5.6 we have that

$$S(L) \leq \|L\|_{A^{\frac{1}{2}}} < 1. \tag{5.23}$$

3.6 COMPARISON OF RECIPROCAL RATES OF CONVERGENCE

In this section, we will be concerned with the task of reducing the reciprocal rate of convergence of the iterative schemes considered so far assuming that A is symmetric and positive definite. It can be seen (from (4.5) and (4.9)) that the number of iterations required to achieve a certain level of convergence is proportional to the reciprocal rate of convergence. This will be the basis for our comparison apart from the work involved in each iteration. First it can be noted that the matrix B defined by (2.4) is similar to the matrix

$$\tilde{B} = D^{\frac{1}{2}}BD^{-\frac{1}{2}} = I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}. \tag{6.1}$$

Consequently, the eigenvalues of $\tilde{B}$ and hence those of B are real and less than unity since $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is a symmetric and positive definite matrix. By the definition of the matrix B we have that its eigenvalues μ_i satisfy the relationship

$$\sum_{i=1}^{N} \mu_i = \text{trace}(B) = 0 \tag{6.2}$$

and therefore

$$m(B) \leq 0 \leq M(B), \tag{6.3}$$

where m(B) and M(B) are the smallest and largest eigenvalues of B, respectively. From (6.3) we conclude that

$$m(B) \leq \mu \leq M(B) \tag{6.4}$$

where μ is an eigenvalue of B.

Let us consider the JOR method and determine the role of the real parameter ω so as to minimise the spectral radius of $S(B_\omega)$. From (2.14) we have that $S(B_\omega)$ is given by the expression

$$S(B_\omega) = \max_{m(B)\leq\mu\leq M(B)} |\omega\mu+1-\omega|. \tag{6.5}$$

Evidently, for all μ the function $S(B_\omega)$ attains its maximum at the end points of the range (6.4). Thus (6.5) yields the expression

$$S(B_\omega) = \max(|\omega m(B)+1-\omega|, |\omega M(B)+1-\omega|). \tag{6.6}$$

Finally, from the above analysis we can readily verify the following theorem.

Theorem 6.1

If

$$\bar{\omega} = \frac{2}{2-m(B)-M(B)}, \tag{6.7}$$

then $$S(B_{\bar{\omega}}) = \frac{M(B)-m(B)}{2-M(B)-m(B)} < 1 \tag{6.8}$$

and if $\omega \neq \bar{\omega}$, then

$$S(B_{\omega}) > S(B_{\bar{\omega}}).$$

If we let

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{6.9}$$

then by (6.1) we have the eigenvalue relationships

$$\begin{aligned} m(B) &= 1-M(\hat{A}) \\ \text{and} \quad M(B) &= 1-m(\hat{A}). \end{aligned} \tag{6.10}$$

By substituting these values of m(B) and M(B) into (6.7) and (6.8) we obtain

$$\bar{\omega} = \frac{2}{M(\hat{A})+m(\hat{A})} \tag{6.11}$$

and

$$S(B_{\bar{\omega}}) = \frac{M(\hat{A})-m(\hat{A})}{M(\hat{A})+m(\hat{A})} = \frac{P(\hat{A})-1}{P(\hat{A})+1}, \tag{6.12}$$

where

$$P(\hat{A}) = \frac{M(\hat{A})}{m(\hat{A})} \tag{6.13}$$

is the P-condition number† of the matrix $\hat{A}$.

The relationship (6.12) is an alternative expression of the $S(B_{\bar{\omega}})$ in terms of $P(\hat{A})$. From Definition 4.2 the rate of convergence for the JOR method is given by

$$R(B_{\bar{\omega}}) = -\log S(B_{\bar{\omega}}) = -\log \frac{P(\hat{A})-1}{P(\hat{A})+1} \sim \frac{2}{P(\hat{A})} \tag{6.14}$$

for $P(\hat{A}) \gg 1$. Evidently, the reciprocal rate of convergence is given by the expression

$$RR(B_{\bar{\omega}}) = \frac{1}{R(B_{\bar{\omega}})} \sim \frac{P(\hat{A})}{2}. \tag{6.15}$$

From (6.15) we see that $P(\hat{A})$ is proportional to the reciprocal rate

---

†*In general one can define the spectral condition number of a non-singular matrix A by* $k(A)=\|A\|.\|A^{-1}\|$. *However, if A is positive definite, then k(A) becomes the P-condition number denoted by P(A) and is given by the expression* $P(A)=M(A)/m(A)$.

of convergence which in turn is proportional to the number of iterations required for the JOR method to converge. This observation shows the dependence of the effectiveness of the JOR method upon the condition of the matrix A.

As intimated earlier, the rate of convergence of the SOR, GS and SSOR method depends upon the order in which the points are scanned in the mesh. However, if the matrix A resulting from a certain ordering is consistently ordered, then as we will see from the following theorems,the spectral radii of the GS and SOR methods remain constant.

Theorem 6.2

If A is a T-matrix with non-vanishing diagonal elements and B the matrix as defined in (2.4), then

(a) If μ is any eigenvalue of B of multiplicity p, then $-\mu$ is also an eigenvalue of B of multiplicity p.

(b) λ satisfies

$$(\lambda+\omega-1)^2 = \omega^2\mu^2\lambda \tag{6.16}$$

for some eigenvalue μ of B if and only if λ satisfies

$$\lambda+\omega-1 = \omega\mu\lambda^{\frac{1}{2}} \tag{6.17}$$

for some eigenvalue μ of B.

(c) If λ satisfies either, and hence both of the relations (6.16) and (6.17), then λ is an eigenvalue of L_ω.

(d) If λ is an eigenvalue of L_ω, then there exists an eigenvalue μ of B such that (6.16) and (6.17) hold.

Corollary 6.3

Under the same hypotheses of Theorem 6.2, the set of eigenvalues of L includes the number zero together with the numbers $\mu_1^2,\mu_2^2,\ldots,\mu_q^2$, where $\pm\mu_1,\pm\mu_2,\ldots,\pm\mu_q$ are the non-zero eigenvalues of B.

As a result of Theorem 2-6.1 and Theorem 6.2 we have the following generalisation.

Theorem 6.4

If the matrix A is a consistently ordered matrix, with non-vanishing diagonal elements, then the conclusion of Theorems 6.2 and Corollary 6.3 are valid.

Furthermore, from the above theorems we have the extended conclusion.

Theorem 6.5

If the matrix A is a symmetric consistently ordered matrix with positive diagonal elements, then $\bar{\mu}=S(B)<1$ if and only if A is positive definite.

The above theorem is valid, since by Corollary 5.7 the GS method converges when A is symmetric and has positive diagonal elements if and only if A is positive definite. But by Theorem 6.3 the GS method converges if and only if $\bar{\mu}<1$.

From the above analysis of the SOR method and Section 2-2.6 we see the requirement of having a certain technique to scan the mesh in such a way so that the resulting matrix has a compatible ordering vector. From Theorem 2-7.1 we have that if $(x_0+p_ih,\ y_0+q_ih)$, $i=1,2,\ldots,N$ is a given set of mesh points, then the resulting matrix A has Property A if and only if there exist at least one ordering vector. Two ordering vectors $\gamma^{(0)}$ and $\gamma^{(1)}$ are the following:

$$\gamma^{(0)} = \begin{cases} 1 & \text{if } p_i+q_i \text{ is even} \\ 2 & \text{if } p_i+q_i \text{ is odd} \end{cases} \tag{6.18}$$

$$\gamma^{(1)} = p_i+q_i. \tag{6.19}$$

The following methods of relabelling the mesh points guarantee that the resulting matrices have $\gamma^{(0)}$ or $\gamma^{(1)}$ as a compatible ordering vector and hence they all lead to consistently ordered matrices.

1. A point $(x_0+ph,\ y_0+qh)$ occurs before $(x_0+p'h,\ y_0+q'h)$ if $q<q'$ or if $q=q'$ and $p<p'$. This ordering is known as the σ_2 or the natural ordering since the mesh is scanned from left to right and from bottom to top.

2. All points (x_0+ph, y_0+qh) with p+q even (or red) occur before those with p+q odd (or black). This ordering is called the red-black, or σ_1, or checker-board ordering. In this ordering all red points are scanned before the black points are updated, or vice versa.
3. A point (x_0+ph, y_0+qh) occurs before $(x_0+p'h, y_0+q'h)$ if $p+q<p'+q'$. This ordering is known as "ordering by diagonals" since we scan the mesh along the diagonals.

We now give some theorems, which form the basis of the theory of the SOR method when A belongs to the class of consistently ordered matrices and B has real eigenvalues.

Theorem 6.6

If A is a consistently ordered matrix with non-vanishing diagonal elements such that $B=I-(\text{diag } A)^{-1}A$ has real eigenvalues, then

$$S(L_\omega) = \bar{S}(L_\omega)^{\dagger} < 1$$

if and only if

$$0 < \omega < 2 \tag{6.20}$$

and

$$S(B) < 1. \tag{6.21}$$

For the determination of the optimum value ω in the SOR method we have the following theorem (Young [1954,1971]).

Theorem 6.7

If A is a consistently ordered matrix with non-vanishing diagonal elements such that $B=I-(\text{diag } A)^{-1}A$ has real eigenvalues and such that $\bar{\mu}=S(B)<1$ and if ω_b is defined by

$$\omega_b = \frac{2}{1+\sqrt{1-\bar{\mu}^2}} = 1 + \left[\frac{\bar{\mu}}{1+\sqrt{1-\bar{\mu}^2}}\right]^2, \tag{6.22}$$

$^{\dagger}\bar{S}(L_\omega)$ *denotes the virtual spectral radius of* L_ω *(see Young [1971] p.170).*

then

$$\bar{S}(L_{\omega_b}) = S(L_{\omega_b}) = \omega_b - 1 = \frac{1-\sqrt{1-\bar{\mu}^2}}{1+\sqrt{1-\bar{\mu}^2}} \tag{6.23}$$

and if $\omega \neq \omega_b$, then

$$\bar{S}(L_{\omega}) = S(L_{\omega}) > S(L_{\omega_b}). \tag{6.24}$$

Moreover, for any ω in the range $0<\omega<2$, we have

$$\bar{S}(L_{\omega}) = S(L_{\omega}) = \begin{cases} \left[\frac{\omega\bar{\mu}+(\omega^2\bar{\mu}^2-4(\omega-1))^{\frac{1}{2}}}{2}\right]^2, & \text{if } 0<\omega\leq\omega_b \\ \omega-1, & \text{if } \omega_b\leq\omega<2. \end{cases} \tag{6.25}$$

From the above theorem we can show that if the eigenvalues of B are real, then the rate of convergence of the SOR method using the optimum relaxation factor ω_b is greater by an order of magnitude than the rate of convergence of the JOR method.

Theorem 6.8

Under the hypotheses of Theorem 6.7 and for ω_b, $\bar{\omega}$ given by (6.22) and (6.7) respectively, then

$$\lim_{\bar{\mu}\to 1} \frac{RR(L_{\omega_b})}{\sqrt{RR(B_{\bar{\omega}})}} = \frac{1}{2\sqrt{2}} \quad . \tag{6.26}$$

Proof

By Theorem 6.4 we have that

$$S(B) = M(B) = -m(B) < 1. \tag{6.27}$$

whereas by (6.7) and (6.27) it follows that

$$\bar{\omega} = 1. \tag{6.28}$$

Consequently, under the hypotheses of the theorem, the JOR method coincides with the J method i.e.,

$$B_{\bar{\omega}} = B. \tag{6.29}$$

By (6.23) the rate of convergence for the SOR method is given by

$$R(L_{\omega_b}) = -2\log\left(\frac{\bar{\mu}}{1+\sqrt{1-\bar{\mu}^2}}\right), \tag{6.30}$$

whereas by (6.29) the rate of convergence for the JOR method is

$$R(B) = -\log\bar{\mu}. \tag{6.31}$$

By applying L'Hospital's rule we successively obtain the result

$$\lim_{\bar{\mu}\to 1^-} \frac{R(L_{\omega_b})}{\sqrt{R(B_{\bar{\omega}})}} = \lim_{\bar{\mu}\to 1^-} \frac{\frac{d}{d\bar{\mu}}[R(L_{\omega_b})]}{\frac{d}{d\bar{\mu}}[R(B_{\bar{\omega}})]} = \lim_{\bar{\mu}\to 1^-} \frac{2\sqrt{2}(-2\log\bar{\mu})^{\frac{1}{2}}}{1-\bar{\mu}^2} = 2\sqrt{2}. \tag{6.32}$$

Hence, by the definition of the reciprocal rate of convergence the proof of the theorem is complete.

From (6.32) we see that there is an order of magnitude improvement of the SOR method over the JOR method. This improvement also remains over the GS method, as it can be seen from Corollary 6.3. On the other hand, it is known (Young [1971], Chapter 7) that the Jordan canonical form of L_{ω_b} is not diagonal and hence the gain in convergence rate is somewhat less than expected.

Furthermore, we can express the spectral radius of the SOR method in terms of the P-condition number of $\hat{A}$. By (6.13) and (6.22)† we have

$$\omega_b = 1 + \left[\frac{\sqrt{P(\hat{A})}-1}{\sqrt{P(\hat{A})}+1}\right]^2 , \tag{6.33}$$

hence (6.23) yields the following expression

$$S(L_{\omega_b}) = \left[\frac{\sqrt{P(\hat{A})}-1}{\sqrt{P(\hat{A})}+1}\right]^2 . \tag{6.34}$$

Finally from (6.34) we obtain successively

$$R(L_{\omega_b}) = -\log S(L_{\omega_b}) = -2\log\left[\frac{\sqrt{P(\hat{A})}-1}{\sqrt{P(\hat{A})}+1}\right] \sim \frac{4}{\sqrt{P(\hat{A})}} \tag{6.35}$$

for $P(\hat{A})>>1$.

Consequently, (Evans [1973]) we have shown the following.

Corollary 6.9

Under the hypotheses of Theorem 6.7 and ω_b satisfying (6.22), then

$$RR(L_{\omega_b}) \sim \frac{\sqrt{P(\hat{A})}}{4} . \tag{6.36}$$

The same result could have been obtained more simply by using (6.26) and

---

†*From (6.1) and (6.13) we note that* $P(\hat{A})=(1+S(B))/(1-S(B))$.

(6.19). Finally, from (6.35) we can state our main conclusion, that the number of iterations for the SOR method is proportional to the square root of the P-condition number of the original matrix $\hat{A}$.

As we have seen so far, if A is a positive definite consistently ordered matrix, then a substantial improvement can be achieved using the SOR iterative scheme with an appropriate relaxation factor as compared with the J, GS, JOR and GSP methods. It is interesting to note that we can actually relax the conditions on the matrix A. It is known (Kahan [1958], Varga [1959]) that the theory of the SOR method holds with approximately the same results if the matrix A is a Stieltjes matrix. For such matrices we have again that $S(B)<1$, thus one can compute ω_b by (6.22).

Theorem 6.10

If A is a Stieltjes matrix and if ω_b is given by (6.22), then the following inequalities hold

$$\omega_b - 1 \leq S(\mathcal{L}_{\omega_b}) \leq (\omega_b - 1)^{\frac{1}{2}}. \tag{6.37}$$

A comparison of the SOR method with the JOR method can be carried out, in the case where A is a Stieltjes matrix. From (6.3) we have that

$$m(B) \leq M(B) = \bar{\mu}. \tag{6.38}$$

It is evident that (6.8) is minimised if $m(B)=0$, hence

$$S(B_{\bar{\omega}}) \geq \frac{M(B)}{2-M(B)} \quad . \tag{6.39}$$

From (6.39) we can work analogously towards the proof of Theorem 6.8 and derive the following:

Lemma 6.11

If A is a Stieltjes matrix, then we have

$$\lim_{\bar{\mu}\to 1^-} \frac{R(B_{\bar{\omega}})}{R(B)} = 2. \tag{6.40}$$

On the other hand, from (6.37), (6.23) and (6.39) we easily prove:

Corollary 6.12

Under the hypotheses of Theorem 6.10 we have

$$RR(\mathcal{L}_{\omega_b}) \lesssim \sqrt{RR(B_{\bar{\omega}})}. \tag{6.41}$$

By comparing the relationships (6.26) and (6.41) we conclude that even though the reciprocal rate of convergence for the SOR method, when A is a Stieltjes matrix, may be greater than in the consistently ordered case, we still have an order of magnitude improvement over the JOR method.

As we have seen the SOR method is not affected by the different consistent orderings mentioned earlier. This is not the case with the SSOR method where these consistent orderings yield a convergence rate which differs by an order of magnitude even though more work is required per iteration. Indeed, with σ_1-ordering the SSOR method is no better than the GS method which converges with an order of magnitude slower than the SOR method (D'Sylva and Miles [1963]). This is shown in the following theorem (Wachspress [1966], Young [1971]).

Theorem 6.13

Let A be a positive definite matrix of the form (2-7.1), then

$$S(\mathcal{G}_\omega) = S(\mathcal{L}_{\omega(2-\omega)}) \lesssim 1 - \tfrac{1}{2}\omega^2(2-\omega)^2(1-\bar{\mu})^2 \tag{6.42}$$

and unless $\omega=1$, we have

$$S(\mathcal{G}_\omega) > S(\mathcal{G}_1) = S(\mathcal{L}) = \bar{\mu}^2. \tag{6.43}$$

From the above theorem we conclude that there is no justification whatever for using σ_1-ordering with SSOR. We will therefore consider the case where the natural ordering is used i.e., the matrix A is not of the form (2-7.1).

The analysis for the determination of a "good" value for ω in the SSOR method can be summarised in the following theorem (Young [1974]).

Theorem 6.14

Let $\bar{\beta}$, M and m be numbers such that

$$S(LU) \leq \bar{\beta} \tag{6.44}$$

and

$$-2\sqrt{\bar{\beta}} \leq m \leq m(B) \leq 0 \leq M(B) \leq M \leq \min(1, 2\sqrt{\bar{\beta}}), \tag{6.45}$$

then a good bound on $S(\mathcal{G}_\omega)$ is given by

$$S(\mathcal{G}_\omega) \leq \begin{cases} 1 - \dfrac{\omega(2-\omega)(1-M)}{1-\omega M+\omega^2\bar{\beta}}, & \text{if } \bar{\beta} \geq \frac{1}{4} \text{ or if } \bar{\beta} < \frac{1}{4} \text{ and } \omega \leq \omega^* \\ 1 - \dfrac{\omega(2-\omega)(1-m)}{1-\omega m+\omega^2\bar{\beta}}, & \text{if } \bar{\beta} < \frac{1}{4} \text{ and } \omega > \omega^*. \end{cases} \tag{6.46}$$

Here, for $\bar{\beta} < \frac{1}{4}$ we define ω^* by

$$\omega^* = \frac{2}{1+\sqrt{1-4\bar{\beta}}}. \tag{6.47}$$

Moreover, the above bound is minimised when

$$\omega_1 = \begin{cases} \dfrac{2}{1+\sqrt{1-2M+4\bar{\beta}}} = \omega_M, & \text{if } M \leq 4\bar{\beta} \\ \dfrac{2}{1+\sqrt{1-4\bar{\beta}}} = \omega^*, & \text{if } M \geq 4\bar{\beta}. \end{cases} \tag{6.48}$$

The corresponding value of $S(\mathcal{G}_{\omega_1})$ is then given by

$$S(\mathcal{G}_{\omega_1}) \leq \begin{cases} \dfrac{1 - \dfrac{1-M}{\sqrt{1-2M+4\bar{\beta}}}}{1 + \dfrac{1-M}{\sqrt{1-2M+4\bar{\beta}}}}, & \text{if } M \leq 4\bar{\beta} \\ \dfrac{1-\sqrt{1-4\bar{\beta}}}{1+\sqrt{1-4\bar{\beta}}} = \omega^* - 1, & \text{if } M \geq 4\bar{\beta}. \end{cases} \tag{6.49}$$

The above bounds can be modified to yield the expressions

$$S(\mathcal{G}_{\omega_1}) \leq \begin{cases} \dfrac{1-\sqrt{1-M}}{1+\sqrt{1-M}}, & \text{if } \bar{\beta} \leq \frac{M}{4} \\ \dfrac{1-\sqrt{\frac{1-M}{2}}}{1+\sqrt{\frac{1-M}{2}}}, & \text{if } \frac{M}{4} \leq \bar{\beta} \leq \frac{1}{4} \\ \dfrac{1-\gamma\sqrt{\frac{1-M}{2}}}{1+\gamma\sqrt{\frac{1-M}{2}}}, & \text{if } \bar{\beta} \geq \frac{1}{4} \end{cases} \tag{6.50}$$

where
$$\gamma = \left[1+\frac{2(\bar{\beta}-\frac{1}{4})}{1-M}\right]^{-1/4}. \tag{6.51}$$

We note that for the case $S(LU)\leq\frac{1}{4}$, the results of Theorem 6.14 and the formulae (6.48),(6.49) were obtained independently by Axelsson [1972]. This problem had also been considered earlier by Habelter and Wachspress [1961] using variational techniques. They developed an implicit equation for determining ω, (it involves the eigenvector of $S(\mathcal{L}_{\omega_0})$, where ω_0 is the optimum value of ω). This equation was used by Evans and Forrington [1963] to develop an iterative scheme for determining the optimum ω of the SSOR method for the model problem.

A comparison between the asymptotic bounds on $RR(\mathcal{L}_{\omega_1})$ and $RR(B_{\bar{\omega}})$, using the relationships (6.39),(6.29) and (6.50), is given by the following table.

Range of $\bar{\beta}$	Asymptotic Bounds on $RR(\mathcal{L}_{\omega_1})/\sqrt{RR(B_{\bar{\omega}})}$	
	General Case	Property A
$\bar{\beta}\leq\frac{M}{4}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$
$\frac{M}{4}\leq\bar{\beta}\leq\frac{1}{4}$	1	$\frac{1}{\sqrt{2}}$
$\bar{\beta}>\frac{1}{4}$	γ^{-1}	$\frac{1}{\sqrt{2}}\gamma^{-1}$

TABLE 6.1

Finally, from (5.7) and Theorem 2-3.1 we have the result
$$S(L_{\omega_1}) \leq \|L_{\omega_1}\|_{A^{\frac{1}{2}}} = \sqrt{S(\mathcal{L}_{\omega_1})} \tag{6.52}$$
which gives
$$RR(L_{\omega_1}) \leq 2RR(\mathcal{L}_{\omega_1}). \tag{6.53}$$

From (6.53) it follows that we can obtain bounds on $RR(L_{\omega_1})$ in terms of $RR(B_{\bar{\omega}})$ for the various cases (Table 6.1). However, it is expected that in general these bounds will not be as good as the ones

given by (6.26) and (6.41). On the other hand, for the case of a consistently ordered matrix the bound (6.26) for $RR(L_{\omega_b})$ is smaller than the best possible bound on $RR(\mathcal{L}_{\omega_1})$, namely $\frac{1}{2}\sqrt{RR(B_{\bar{\omega}})}$. This observation suggests that even if we employ Niethammer's scheme (Niethammer [1964]) to reduce the work per iteration of the SSOR method to that of the SOR method there is little to be gained by using the former method. However, the eigenvalues of the SSOR iteration matrix $\mathcal{L}_{\omega}$ are real and non-negative (see Theorem 5.6) and under these conditions, it is possible to accelerate the SSOR method by an order of magnitude by means of semi-iteration (Varga [1957], Golub and Varga [1961]). This approach is precluded for SSOR with optimum $\omega=\omega_b$ since the eigenvalues of L_{ω_b} are complex.

3.7 SEMI-ITERATIVE METHODS

Let us again consider the completely consistent linear stationary iterative method defined by

$$u^{(n+1)} = Gu^{(n)} + k \tag{7.1}$$

where I-G is non-singular and $k=(I-G)A^{-1}b$. From the theory of summability of sequences we can often develop another sequence of vectors $v^{(0)},v^{(1)},\ldots$ such that either the new sequence converges when the old one does not or else the new sequence converges faster than the old one. The new sequence can be considered as a linear combination of the old one i.e.

$$v^{(n)} = \sum_{k=0}^{n} \alpha_k(n) u^{(k)}. \tag{7.2}$$

Our object here is to determine the constants $\alpha_k(n)$ such that the rate of convergence of the new iterative procedure is greater than the one given by (7.1). It can be stated that in general (7.2) is a non-stationary method associated with a linear stationary iterative method of first degree.

The new process defined by (7.2) is known as the semi-iterative method (Varga [1957]) with respect to the linear stationary method of (7.1). A natural restriction on the coefficients $\alpha_k(n)$ is

$$\sum_{k=0}^{n} \alpha_k(n) = 1, \quad n=0,1,2,\ldots . \tag{7.3}$$

If we let

$$\tilde{\varepsilon}^{(n)} = v^{(n)} - \bar{u} \tag{7.4}$$

where $\bar{u}$ is the true solution of (7.1), then from (7.2) and (3.3) we have

$$\tilde{\varepsilon}^{(n)} = \sum_{k=0}^{n} \alpha_k(n) \varepsilon^{(k)}, \tag{7.5}$$

hence by (3.4)

$$\tilde{\varepsilon}^{(n)} = \left(\sum_{k=0}^{n} \alpha_k(n) G^k\right) \varepsilon^{(0)}. \tag{7.6}$$

Furthermore, by defining the polynomial $P_n(x)$ to be

$$P_n(x) = \sum_{k=0}^{n} \alpha_k(n) x^k, \tag{7.7}$$

we can write (7.6) in the form

$$\tilde{\varepsilon}^{(n)} = P_n(G)\varepsilon^{(0)} = P_n(G)\tilde{\varepsilon}^{(0)}. \tag{7.8}$$

In addition, we assume that for some real numbers α and β with $\alpha \leq \beta < 1$ the eigenvalues μ of G are real and lie in the interval

$$\alpha \leq \mu \leq \beta < 1. \tag{7.9}$$

Then from (7.8) it follows that

$$\|\tilde{\varepsilon}^{(n)}\| \leq \|P_n(G)\| \, \|\varepsilon^{(0)}\|. \tag{7.10}$$

Thus we are naturally led to the minimisation of

$$\max_{\alpha \leq \mu \leq \beta} |P_n(\mu)|, \tag{7.11}$$

where $P_n(1)=1$.

In order for the above problem to be reduced to a standard one we map the interval $\alpha \leq \mu \leq \beta$ onto the interval $-1 \leq \gamma \leq 1$ by the transformation

$$\gamma = \frac{2\mu-(\beta+\alpha)}{\beta-\alpha} \tag{7.12}$$

and

$$\mu = \frac{1}{2}[(\beta-\alpha)\gamma+(\beta+\alpha)]. \tag{7.13}$$

If we now let

$$Q_n(\gamma) = P_n(\tfrac{1}{2}[(\beta-\alpha)\gamma+(\beta+\alpha)]), \tag{7.14}$$

the problem is reduced to finding the polynomial $Q_n(\gamma)$ of degree n or less such that $Q_n(z)=1$, where

$$z = \gamma(1) = \frac{2-(\alpha+\beta)}{\beta-\alpha} \tag{7.15}$$

and

$$\max_{-1 \leq \gamma \leq 1} |Q_n(\gamma)| \tag{7.16}$$

is minimised. The solution of this problem is known (Markoff [1892], Flanders and Shortley [1950]) and is given by

$$Q_n(\gamma) = \frac{T_n(\gamma)}{T_n(z)}. \tag{7.17}$$

Moreover, from the relationships

$$\max_{-1 \leq \gamma \leq 1} |Q_n(\gamma)| = 1/T_n(z) = 1/T_n(\frac{2-(\beta+\alpha)}{\beta-\alpha}), \tag{7.18}$$

we have

$$P_n(\mu) = Q_n(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}) = T_n(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha})/T_n(\frac{2-(\beta+\alpha)}{\beta-\alpha}). \tag{7.19}$$

Finally, by (7.8) and using the three term recurrence relation of the Chebyshev polynomials we obtain

$$\tilde{\varepsilon}^{(n+1)} = 2\left[\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\right]\frac{T_n(z)}{T_{n+1}(z)}\tilde{\varepsilon}^{(n)} - \frac{T_{n-1}(z)}{T_{n+1}(z)}\tilde{\varepsilon}^{(n-1)} \tag{7.20}$$

which by (7.4) can be written in terms of the new vectors $v^{(n)}$ as

$$v^{(n+1)} = \rho_{n+1}[\bar{\rho}(Gu^{(n)}+k)+(1-\bar{\rho})u^{(n)}]+(1-\rho_{n+1})u^{(n-1)} \tag{7.21}$$

where

$$\bar{\rho} = \frac{2}{2-(\alpha+\beta)}\ , \tag{7.22}$$

$$\left.\begin{aligned} \rho_1 &= 1\ , \\ \rho_2 &= \left(1-\frac{\sigma^2}{2}\right)^{-1}\ , \\ \rho_{n+1} &= \left(1-\frac{\sigma^2}{4}\rho_n\right)^{-1}\ , \quad n=2,3,\ldots \end{aligned}\right\} \tag{7.23}$$

with

$$\sigma = \frac{\beta-\alpha}{2-(\beta+\alpha)}\ . \tag{7.24}$$

We notice that (7.21) is obtained as a combination of a two stage acceleration of (7.1). Firstly, we could consider the JOR version applied to (7.1), which is defined by

$$u^{(n+1)} = \bar{\rho}(Gu^{(n)}+k)+(1-\bar{\rho})u^{(n)} \tag{7.25}$$

and then for further acceleration we could consider the Chebyshev semi-iterative method with respect to (7.25), as defined in Golub and Varga [1961], in order to obtain (7.21). From (7.9) we note that we do not always require the basic iterative scheme (7.1) to be convergent. Also, the recursive relation (7.21) shows that it is unnecessary to form the auxilary vector iterates $u^{(n)}$ in order to determine the vectors $v^{(n)}$. Finally, (7.21) requires an additional vector of storage over (7.1), which can be of considerable weight in practice, if computer storage is limited. However, as we will see the application of the semi-iterative process (7.21) can often give a convergence rate accelerated by an order of magnitude.

We can rewrite (7.21) in the form

$$u^{(n)} = P_n(G)u^{(0)} + k_n, \tag{7.26}$$

where $P_n(G)$ is a certain polynomial in G and k_n is a suitable vector. The virtual spectral radius can be obtained by (7.16), (7.18) and is given by

$$\bar{S}(P_n(G)) = \max_{\alpha \leq \mu \leq \beta} |P_n(\mu)| = 1/T_n(z). \tag{7.27}$$

By the definition of Chebyshev polynomial it can be shown (Young [1971] that

$$\bar{S}(P_n(G)) = \frac{2r^{n/2}}{1+r^n}, \tag{7.28}$$

where

$$r^{\frac{1}{2}} = \frac{\sigma}{1+\sqrt{1-\sigma^2}} \quad \text{and} \quad \sigma = \frac{1}{2}. \tag{7.29}$$

On the other hand, the average rate of convergence

$$R_n(P_n(G)) = -\frac{1}{n} \log \bar{S}(P_n(G)) \tag{7.30}$$

approaches the asymptotic average rate of convergence, or simply the rate of convergence

$$R_\infty(P_n(G)) = -\frac{1}{2} \log r \tag{7.31}$$

as $n \to \infty$.

An analogous result to (6.26) can be obtained, if one follows the proof of Theorem 6.8. Indeed, for n sufficiently large we have

$$\lim_{\sigma \to 1^-} \frac{R_\infty(P_n(G))}{\sqrt{R_1(P_1(G))}} = \sqrt{2}, \tag{7.32}$$

where

$$R_1(P_1(G)) = R(G_{\bar{\rho}}) = -\log \frac{r^{\frac{1}{2}}}{1+r} = -\log\sigma \tag{7.33}$$

and

$$G_{\bar{\rho}} = \bar{\rho}G + (1-\bar{\rho})I. \tag{7.34}$$

From (7.33) and (7.34) one can verify that for n=1 the semi-iterative process degenerates to (7.25). The relationship (7.32) establishes the fact that by using optimum semi-iterative techniques based on a given method, the reciprocal asymptotic average rate of convergence of the SI method is improved by an order of magnitude over than the optimum extrapolated one.

3.8 VARIABLE EXTRAPOLATION METHODS

In the previous section it was shown how we are able to construct an effective iterative process using the concept of semi-iterative techniques. It was also mentioned that in the new procedure each vector $v^{(n+1)}$ requires the computation of two previous vectors $v^{(n)}$ and $v^{(n-1)}$. If computer storage is limited, then we can consider another form of accelerating the basic iterative process (7.1). This can be achieved by allowing the parameter $\bar{\rho}$ in (7.25) to vary in each iteration, hence we have

$$u^{(n+1)} = \theta_{n+1}(Gu^{(n)}+k) + (1-\theta_{n+1})u^{(n)}, \tag{8.1}$$

where $\theta_1,\theta_2,\ldots$ are iteration parameters.

This idea was presented by Richardson [1910] and applied to a certain method of the form (7.1). The iteration parameters θ_n are selected in the cyclic order $\theta_1,\theta_2,\ldots,\theta_m,\theta_1,\theta_2,\ldots$ where m is an integer. From (8.1) we have that given $\theta_1,\theta_2,\ldots,\theta_m$, then

$$u^{(m)} = P_m(G)u^{(0)}+k_m, \tag{8.2}$$

for a suitable vector k_m and $P_m(G)$ is the polynomial

$$P_m(G) = \prod_{k=1}^{m}(\theta_k G+(1-\theta_k)I). \tag{8.3}$$

If one now follows the analysis of the previous section, then it can be easily concluded that the minimised polynomial $P_m(\mu)$ is given by

$$P_m(\mu) = \frac{T_m\left(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}\right)}{T_m\left(\frac{2-(\beta+\alpha)}{\beta-\alpha}\right)}. \tag{8.4}$$

The iteration parameters θ_k can be determined by equating the roots of (8.3) and (8.4). Thus we obtain values for the parameters θ_k of the form

$$\theta_k = \frac{2}{2-(\beta-\alpha)\cos\frac{(2k-1)\pi}{2m}-(\beta+\alpha)}, \quad k=1,2,\ldots,m. \tag{8.5}$$

The virtual spectral radius of (8.1) can be verified by (8.4) to be

$$\bar{S}(P_{\ell m}(G)) = \left(\frac{2r^{m/2}}{1+r^m}\right)^{\ell}. \tag{8.6}$$

where ℓ is an integer determining the number of cycles. It can be seen from (8.6) and (7.28) that as m increases, then the rapidity of convergence tends to the one given by the semi-iterative method.

However, numerical experiments (Young [1954a,1956], Young and Warlick [1953]) show that for large m numerical instability may occur. Also, it is undesriable to select m very large because convergence is expected after ℓm iterations.

3.9 SECOND-DEGREE METHODS

An accelerated scheme similar to (7.21) can be produced by considering constant iteration parameters throughout the process. By rewriting (7.21) in the form

$$u^{(n+1)} = u^{(n)}+(\rho_{n+1}-1)(u^{(n)}-u^{(n-1)})+\frac{2\rho_{n+1}}{2-(\alpha+\beta)}(Gu^{(n)}+k-u^{(n)}) \tag{9.1}$$

we can obtain the linear second-degree method

$$u^{(n+1)} = u^{(n)}+\xi(u^{(n)}-u^{(n-1)})+\eta(Gu^{(n)}+k-u^{(n)}) \tag{9.2}$$

where we replaced $\rho_{n+1}-1$ by ξ and $\frac{2\rho_{n+1}}{2-(\alpha+\beta)}$ by η.

The form of (9.2) is a special case of the linear stationary iterative method of second-degree given by

$$u^{(n+1)} = G_1u^{(n)}+H_1u^{(n-1)}+k_1. \tag{9.3}$$

We can see (Golub and Varga [1961]) that (9.3) can be written as

$$\begin{pmatrix} u^{(n)} \\ u^{(n-1)} \end{pmatrix} = \begin{pmatrix} 0 & I \\ H_1 & G_1 \end{pmatrix} \begin{pmatrix} u^{(n-1)} \\ u^{(n)} \end{pmatrix} + \begin{pmatrix} 0 \\ k_1 \end{pmatrix}. \tag{9.4}$$

The iterative process (9.4) is convergent if and only if

$$S(M) < 1 \tag{9.5}$$

where

$$M = \begin{pmatrix} 0 & I \\ H_1 & G_1 \end{pmatrix}. \tag{9.6}$$

Thus, if λ is an eigenvalue of M, then the roots of

$$\det(\lambda^2 I-\lambda G_1-H_1) = 0 \tag{9.7}$$

must be less than unity in modulus for (9.5) to hold. In the case of (9.2), it is easily seen that (9.7) becomes

$$\det(\lambda^2 I-\lambda(\eta G+(1-\eta+\xi)I)+\xi I) = 0, \tag{9.8}$$

hence if μ is an eigenvalue of G, then the following relationship holds

$$\lambda^2-\lambda(\eta\mu+1-\eta+\xi)+\xi = 0. \tag{9.9}$$

For fixed ξ the root radius i.e., $\max_{\mu}|\lambda|$ is minimised when

$$(\eta\mu+1-\eta+\xi)^2 = 4\xi \tag{9.10}$$

thus (see (7.9)) we have the relationships,

$$\eta(\beta-1)+1+\xi = 2\xi^{\frac{1}{2}} \tag{9.11}$$

and

$$\eta(\alpha-1)+1+\xi = -2\xi^{\frac{1}{2}}. \tag{9.12}$$

Consequently, by adding (9.11) and (9.12) we can determine η from the relationship

$$\eta = \frac{2(1+\xi)}{2-(\beta+\alpha)} . \tag{9.13}$$

Moreover, from (9.13) and either of (9.11),(9.12) the best choice of ξ is given by

$$\xi_0 = \hat{\omega}_0 - 1 \tag{9.14}$$

where

$$\hat{\omega}_0 = \frac{2}{1+\sqrt{1-\sigma^2}} \tag{9.15}$$

and σ is defined in (7.24).

Finally, from (9.14) and (9.13) we obtain the best value of η by the expression

$$\eta_0 = \frac{2\hat{\omega}_0}{2-(\beta+\alpha)} . \tag{9.16}$$

From (9.14),(9.15),(9.9) and (7.29) the spectral radius of M is given by

$$S(M) = (\hat{\omega}_0-1)^{\frac{1}{2}} = r^{\frac{1}{2}} , \tag{9.17}$$

thus the rate of convergence

$$R(M) = -\frac{1}{2} \log r \tag{9.18}$$

is comparable with the one obtained by semi-iterative techniques. Also, by (9.17) and (7.31) we conclude that the rate of convergence of semi-iterative and second degree methods depends on the same quantity r. On the other hand, it can be proved (Young and Kincaid [1969]) that the semi-iterative method yields greater acceleration than the second-degree methods as expected, since in the latter the coefficients are constants whereas in the former they are variables. However, as with semi-iterative methods, we need to store two vector iterants for each iteration and this storage requirement can be severe for large systems of equations, or computers with limited storage capacity.

3.10 THE CONJUGATE GRADIENT METHOD

In this section we will briefly consider the conjugate gradient method (CG method) introduced firstly by Hestenes and Stiefel [1952], Stiefel [1952] as an iterative method for solution of large sparse systems (Reid [1971]).

As a matter of fact we will consider the CG method as an acceleration procedure analogous to the SI method with respect to the iterative method (1.3).

Let us again consider the linear system

$$Au = b \tag{10.1}$$

where A is an N×N symmetric and positive definite matrix. The quadratic functional related to the system (10.1) is given by

$$Q(u) = \frac{1}{2}(u,Au)-(u,b) = \text{const.} \tag{10.2}$$

This functional defines a family of similar 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

$$r^{(n)} = b-Au^{(n)} = -[\text{Grad } Q(u)]^{\dagger}u^{(n)} \tag{10.3}$$

and it is always normal to the surface of the ellipsoids defined by (10.2). Thus, we attempt to proceed to the solution $A^{-1}b$, the centre point of the ellipsoids, by a sequence of vector displacements of the form

$$u^{(n+1)} = u^{(n)} + \varepsilon_n p^{(n)} \tag{10.4}$$

where $p^{(n)}$ is an arbitrary direction and ε_n is an arbitrary constant. The problem now is to determine ε_n such that the quadratic function $Q(u^{(n+1)})$ will be minimum for a given direction $p^{(n)}$.

From (10.2) and (10.4) we have that $Q(u^{(n+1)})$ is given by the expression

$$Q(u^{(n+1)}) = \frac{1}{2}((u^{(n)}+\varepsilon_n p^{(n)}),A(u^{(n)}+\varepsilon_n p^{(n)}))-((u^{(n)}+\varepsilon_n p^{(n)}),b) \tag{10.5}$$

---

†*where* $[Grad\ Q(u)]u^{(n)}$ *represents a vector with components* $\frac{\partial Q(u^{(n)})}{\partial u_i}$, $i=1,2,\ldots,n$.

hence

$$\frac{\partial Q(u^{(n+1)})}{\partial \varepsilon_n} = (p^{(n)},A(u^{(n)}+\varepsilon_n p^{(n)}))-(p^{(n)},b)$$

$$= -(p^{(n)},r^{(n)})+(\varepsilon_n p^{(n)},Ap^{(n)}). \tag{10.6}$$

The optimum value of ε_n is obtained by setting the expression (10.6) equal to zero, which immediately gives

$$\varepsilon_n = \frac{(p^{(n)},r^{(n)})}{(p^{(n)},Ap^{(n)})}. \tag{10.7}$$

Also, by using the definition of $u^{(n+1)}$ from (10.4) and the value we have just obtained for ε_n, we have

$$(p^{(n)},r^{(n+1)}) = (p^{(n)},(b-Au^{(n+1)})) = (p^{(n)},(r^{(n)}-\varepsilon_n Ap^{(n)})) = 0 \tag{10.8}$$

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

The choice of the direction vector $p^{(n)}$ differentiates many methods which are all convergent for any given $p^{(n)}$. If we wish to choose $p^{(n)}$ to lie along the line of steepest descent, we simply take $p^{(n)}=r^{(n)}$ and by (10.4) and (10.7) we immediately 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 the vectors $p^{(0)},p^{(1)},\ldots,p^{N-1}$ to be pairwise conjugate in the sense that

$$(p^{(i)},Ap^{(j)}) = 0 \tag{10.9}$$

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

$$p^{(n)} = r^{(n)}+\alpha_{n-1}p^{(n-1)} \tag{10.10}$$

we can combine (10.9) and (10.10) to obtain

$$(p^{(n)},Ap^{(n-1)}) = (r^{(n)},Ap^{(n-1)})+(\alpha_{n-1}p^{(n-1)},Ap^{(n-1)}) = 0 \tag{10.11}$$

and finally

$$\alpha_{n-1} = \frac{(r^{(n)},Ap^{(n-1)})}{(p^{(n-1)},Ap^{(n-1)})}. \tag{10.12}$$

This choice of $p^{(n)}$ results in the Conjugate Gradient iterative scheme which is defined as follows

$$u^{(n+1)} = u^{(n)} + \varepsilon_n p^{(n)}, \quad n=0,1,2,\ldots,m-1 \tag{10.13}$$

$$r^{(n)} = b - Au^{(n)}, \quad n=0,1,2,\ldots,m \tag{10.14}$$

$$p^{(n)} = \begin{cases} 0, & n=-1 \\ r^{(n)} + \alpha_{n-1} p^{(n-1)}, & n=0,1,2,\ldots,m-1 \end{cases} \tag{10.15}$$

$$\alpha_{n-1} = \begin{cases} 0, & n=0 \\ -\dfrac{(r^{(n)}, Ap^{(n-1)})}{(p^{(n-1)}, Ap^{(n-1)})}, & n=1,2,\ldots,m-1 \end{cases} \tag{10.16}$$

where m is the smallest integer such that

$$r^{(m)} = 0. \tag{10.17}$$

We summarise below some basic properties of the CG method (see Beckmann [1960])

$$(r^{(i)}, r^{(j)}) = 0 \qquad i \neq j, \quad i,j=0,1,\ldots,m-1 \tag{10.18}$$

$$(p^{(i)}, Ap^{(j)}) = 0 \qquad i \neq j, \quad i,j=0,1,\ldots,m-1 \tag{10.19}$$

$$p^{(i)} \neq 0, \quad i=0,1,\ldots,m-1 \tag{10.20}$$

$$m \leq N \tag{10.21}$$

and

$$\alpha_{n-1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})}, \quad n=1,2,\ldots,m-1. \tag{10.22}$$

From (10.17) we easily 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 is not what actually happens in practice, where the round-off errors may strongly affect the orthogonality of the residuals.

In the past few years a number of modifications and improvements have been made to the CG method (see Rutishauer [1959], Daniel [1967], Reid [1971], Axelsson [1974] and Evans [1973a]). One important modification has been the formulation of the method as a second degree method, i.e. the determination of $u^{(n+1)}$ in terms of $u^{(n)}$ and $u^{(n-1)}$.

By replacing n by n-1 in (10.13) we have

$$u^{(n)} = u^{(n-1)} + \varepsilon_{n-1} p^{(n-1)} \tag{10.23}$$

or

$$\frac{\alpha_{n-1}}{\varepsilon_{n-1}} \varepsilon_n u^{(n)} = \frac{\alpha_{n-1}}{\varepsilon_{n-1}} \varepsilon_n u^{(n-1)} + \varepsilon_n \alpha_{n-1} p^{(n-1)} \tag{10.24}$$

which by eliminating $p^{(n-1)}$ using (10.15) becomes

$$\frac{\alpha_{n-1}}{\varepsilon_{n-1}} \varepsilon_n u^{(n)} = \frac{\alpha_{n-1}}{\varepsilon_{n-1}} \varepsilon_n u^{(n-1)} + \varepsilon_n (p^{(n)} - r^{(n)}) \tag{10.25}$$

and finally eliminating $p^{(n)}$ by (10.13) we obtain

$$u^{(n+1)} = (1 + \frac{\varepsilon_n}{\varepsilon_{n-1}} \alpha_{n-1}) u^{(n)} - \frac{\varepsilon_n}{\varepsilon_{n-1}} \alpha_{n-1} u^{(n-1)} + \varepsilon_n r^{(n)} \tag{10.26}$$

which can be written in the more compact form

$$u^{(n+1)} = \rho_{n+1}(u^{(n)} + \gamma_{n+1} r^{(n)}) + (1-\rho_{n+1}) u^{(n-1)} \tag{10.27}$$

where

$$\rho_{n+1} = 1 + \frac{\varepsilon_n}{\varepsilon_{n-1}} \alpha_{n-1} \tag{10.28}$$

and

$$\gamma_{n+1} = \frac{\varepsilon_n}{\rho_{n+1}} . \tag{10.29}$$

We now proceed to simplify the expressions of ρ_{n+1} and γ_{n+1} by expressing them in terms of certain inner products.

We express (10.27) in terms of residuals by using (10.14) hence it follows that

$$r^{(n+1)} = \rho_{n+1}(r^{(n)} - \gamma_{n+1} A r^{(n)}) + (1-\rho_{n+1}) r^{(n-1)} \tag{10.30}$$

If we now take the inner product of both sides of (10.30) with $r^{(n)}$, then by (10.18) we get

$$0 = \rho_{n+1}((r^{(n)}, r^{(n)}) - \gamma_{n+1}(r^{(n)}, A r^{(n)})) \tag{10.31}$$

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

$$\gamma_{n+1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n)}, A r^{(n)})} . \tag{10.32}$$

On the other hand, if we take the inner product of both sides of (10.30) with $r^{(n-1)}$ yields

$$0 = \rho_{n+1}(-\gamma_{n+1}(r^{(n-1)}, A r^{(n)})) + (1-\rho_{n+1})(r^{(n-1)}, r^{(n-1)}) \tag{10.33}$$

or

$$\rho_{n+1} = \left[1+\frac{(r^{(n-1)},Ar^{(n)})}{(r^{(n-1)},r^{(n-1)})}\gamma_{n+1}\right]^{-1} . \tag{10.34}$$

Furthermore, by replacing n by n-1 in (10.30) we have

$$r^{(n)} = \rho_n(r^{(n-1)}-\gamma_n Ar^{(n-1)}) + (1-\rho_n)r^{(n-2)} \tag{10.35}$$

and if we take the inner product of both sides with $r^{(n)}$ yields

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

thus (10.34) takes the form

$$\rho_{n+1} = \left[1-\frac{\gamma_{n+1}}{\gamma_n}\ \frac{(r^{(n)},r^{(n)})}{(r^{(n-1)},r^{(n-1)})}\ \frac{1}{\rho_n}\right]^{-1} . \tag{10.36}$$

Summarising our results the CG method can also be defined as

$$u^{(n+1)} = \rho_{n+1}(u^{(n)}+\gamma_{n+1}r^{(n)})+(1-\rho_{n+1})u^{(n-1)} \tag{10.37}$$

where

$$\rho_1 = 1,$$

$$\rho_{n+1} = \left[1-\frac{\gamma_{n+1}}{\gamma_n}\cdot\frac{(r^{(n)},r^{(n)})}{(r^{(n-1)},r^{(n-1)})}\cdot\frac{1}{\rho_n}\right]^{-1}, \quad n=1,2,\ldots \tag{10.38}$$

and

$$\gamma_{n+1} = \frac{(r^{(n)},r^{(n)})}{(r^{(n)},Ar^{(n)})} . \tag{10.39}$$

From (10.37) we note that the CG method is of the same form as the SI method (and the second degree method) the only difference being that here the parameters are variables (whereas in SI method we have $\gamma_1=\gamma_2=\ldots=\bar{\rho}$) chosen to minimise the quadratic function Q(u).

As a matter of fact, we expect the CG method to produce a better rate of convergence than with the application of SI techniques since in the former we have one additional parameter γ_{n+1} which is variable instead of being 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 the matrix A. Moreover, it can be proved (Young [1975]) that for all n we have

$$\|\tilde{u}^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}} \leq \|u^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}} \tag{10.40}$$

where $\bar{u}$ is the exact solution of (10.1), $\tilde{u}^{(n)}$ is the approximate solution obtained by the CG method and $u^{(n)}$ is the approximate solution obtained by the SI method with respect to the basic iterative scheme (10.1).

The relationship (10.40) indicates an 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 G. Finally, we note that the relationship given by (10.40) shows that the CG method is better, in the sense of minimising the $A^{\frac{1}{2}}$-norm of the 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 (10.40) and from the fact that in the SI method we have

$$\|u^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}} \leq \frac{2r^{n/2}}{1+r^{n}} \|u^{(0)}-\bar{u}\|_{A^{\frac{1}{2}}} \tag{10.41}$$

we immediately obtain

$$\|\tilde{u}^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}} \leq \frac{2r^{n/2}}{1+r^{n}} \|\tilde{u}^{(0)}-\bar{u}\|_{A^{\frac{1}{2}}} \tag{10.42}$$

by assuming $\tilde{u}^{(0)}=u^{(0)}$, where r is given by (7.29).

CHAPTER 4

AN INTRODUCTION TO PRECONDITIONING TECHNIQUES

4.1 INTRODUCTION

As it was shown in the previous chapter, the rate of convergence R(G) of all the iterative schemes considered so far depends inversely upon the P-condition number of the scaled coefficient matrix

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}. \tag{1.1}$$

Therefore any attempt to improve these iterative methods has clearly to apply some form of "preconditioning" to the original system of equations in order to minimise the P-condition number of the coefficient matrix and hence increase the rate of convergence of the considered iterative procedure. This idea was first introduced by Evans [1968]†, where it was applied to the solution of large linear systems such as those described in section 1.2.

The earlier work on the Extrapolated Modified Aitken method (EMA method) defined by

$$(I-\omega L)(I-\omega U)u^{(n+1)} = [(1-\omega)I+\omega^2 LU]u^{(n)}+\omega c \tag{1.2}$$

(Evans [1963,1964]) and the effectiveness of the SSOR method created the strong feeling of "striving to obtain left hand sides such as in (1.2)" (Evans [1973]). Thus the preconditioning theory was affected by this previous suggestion and was developed by Evans [1968] as follows.

Let us assume, without loss of generality, that the non-singular coefficient matrix A has the splitting

$$A = I-L-U \tag{1.3}$$

where I is the identity matrix and the matrices L,U are defined as in (3-2.6). Further, let us also assume that A is a symmetric and positive definite matrix. Next, we let v be an intermediate transformation vector given by

$$v = (I-\omega U)u \tag{1.4}$$

where ω is a real parameter to be defined later. If we pre-multiply the original system (3-1.1) by the non-singular matrix $(I-\omega L)^{-1}$, then the

---

†*See also Evans [1973,1974].*

transformed system is equivalent to

$$(I-\omega L)^{-1}A(I-\omega U)^{-1}[(I-\omega U)u] = (I-\omega L)^{-1}b \tag{1.5}$$

which can be written in the more compact form as

$$G_\omega^T A G_\omega v = d_\omega \tag{1.6}$$

or

$$\tilde{B}_\omega v = d_\omega \tag{1.7}$$

where

$$\tilde{B}_\omega = G_\omega^T A G_\omega \tag{1.8}$$

$$d_\omega = (I-\omega L)^{-1}b$$

and

$$G_\omega = (I-\omega U)^{-1}. \tag{1.9}$$

Consequently, the original system (3-1.1) has been transformed into the preconditioned system (1.7), where the matrix $\tilde{B}_\omega$ is symmetric congruent to A (see Definition 2-2.2). Since A is positive definite, then by Theorem 2-2.4 it follows that $\tilde{B}_\omega$ is also a positive definite matrix. By inspection, we see that for $\omega=0$ the new system (1.7) reverts back to its original form (3-1.1). Thus, by introducing the transformation above, we allow ω to play the role of a preconditioning parameter such that as ω varies, we hope we can obtain a value of $P(\tilde{B}_\omega)$ which is less than $P(A)$.

After a tedious analysis Evans [1968] showed that there is a value of ω in the range $1<\omega<2$ for which $P(\tilde{B}_\omega)$ achieves its minimum value. (An alternative proof is given in Hatzopoulos [1974]). However, the value of the optimum preconditioning parameter ω_0 was given only for the model problem by an implicit expression involving the eigenvectors associated with the maximum and minimum eigenvalues of $\tilde{B}_{\omega_0}$. Once system (3-1.1) was preconditioned and brought into the form (1.6), Evans was able to introduce new iterative schemes analogous to the already known ones i.e.,

$$v^{(n+1)} = v^{(n)} + (d_\omega - \tilde{B}_\omega v^{(n)}) \tag{1.10}$$

$$v^{(n+1)} = v^{(n)} + \hat{\alpha}(d_\omega - \tilde{B}_\omega v^{(n)}) \tag{1.11}$$

and their accelerated versions of Richardson's, second order Richardson and Chebyshev semi-iterative methods. Finally, it was shown (Evans [1973]) that for the model problem the P-condition number of the matrix $\tilde{B}_{\omega_0}$ was

approximately equal to the square root of $P(\tilde{B}_{\omega=0})$ which established an order of magnitude improvement on the convergence rate of (1.11) over the basic iterative method i.e., when $\omega=0$. This was also verified by a number of numerical examples for the Laplace, General Diffusion ($-\nabla^2\phi+A\phi=S$) and Biharmonic equations.

The establishment of the advantages of the preconditioning techniques was also confirmed by their use in the direct methods of solution for ill-conditioned systems of linear equations (Hatzopoulos [1974]).

In the remainder of this chapter we will consider the idea of preconditioning in a more general concept which will allow us to associate the most effective iterative schemes, similar to the already known ones, with respect to any splitting of the coefficient matrix A. As a result of this, it will help us to examine the known iterative methodsas special cases and their mathematical formulation from a different viewpoint. Also, we would be able to extract some conclusions about the effectiveness of the type of "preconditioning" being used which will provide a guide for future development in the area of the iterative procedures.

4.2 THE PRECONDITIONING TECHNIQUE FOR THE CONSTRUCTION OF ITERATIVE METHODS

We note that if we premultiply system (3-1.1) by A^{-1}, then we are able to obtain immediately its solution $\bar{u}=A^{-1}b$. On the other hand, we have seen (see section 1-1.2) that there are certain difficulties for computing A^{-1}, so instead we consider the case where system (3-1.1) is premultiplied by a non-singular matrix R^{-1}, where R^{-1} is an approximate inverse of A, thus transforming the original system into the following preconditioned form

$$R^{-1}Au = R^{-1}b. \tag{2.1}$$

The matrix R will be referred to as the conditioning matrix and according to our previous observation we first require that this matrix should approach A. Of course, it would be desirable to select R in such a form so that the property of positive definteness is also retained for the preconditioned matrix $R^{-1}A$, although this is not a necessity. The second requirement on R is to possess such a form so that it is possible to compute its inverse relatively easily. Summarising, we require R to satisfy the following properties:

(a) The spectral condition number of the matrix $R^{-1}A$ to become smaller than the spectral condition number of the original matrix A.

(b) For any vectors s and t it is "convenient" to solve the system Rs=t for s.

After we form system (2.1), then we can define a version of the GSD method (see (3-2.28)) with respect to the preconditioned system (2.1) as follows

$$u^{(n+1)} = u^{(n)} + \tau R^{-1}(b-Au^{(n)}) \tag{2.2}$$

where the role of the real parameter τ is similar to the one of $\hat{\alpha}$ in the SD method (see (3-2.31)) and it will be considered later.

From Theorem 3-1.4 we can verify that the constructed iterative

scheme (2.2) is completely consistent with (2.1) if and only if R is a non-singular matrix and $\tau \neq 0$.

Let us now consider the possible forms which the matrix R can possess within the contents of the above restrictions (a) and (b).

The form of R will be closely associated with the splitting of the matrix A. Thus let us define

$$A = D+P+Q \tag{2.3}$$

to be a splitting of A where in general D is a block diagonal matrix and P,Q are also block matrices. If the matrices D+P and D+Q are computationally easily invertible, then the conditioning matrix R could have the form

$$R = (D+P_1)D^{-1}(D+Q_1) \tag{2.4}$$

where P_1, Q_1 may be functions of P,Q respectively.

On the other hand, R could in general take the form

$$R = f(D,P_1,Q_1) \tag{2.5}$$

where $f(D,P_1,Q_1)$ is any function of the matrices D,P_1,Q_1 approaching A.

The form (2.4) of the matrix R gives by (2.2) the following iterative scheme

$$u^{(n+1)} = u^{(n)}+\tau(D+Q_1)^{-1}D(D+P_1)^{-1}(b-Au^{(n)}) \tag{2.6}$$

or

$$u^{(n+1)} = u^{(n)}+\tau(I+D^{-1}Q_1)^{-1}(I+D^{-1}P_1)^{-1}D^{-1}(b-Au^{(n)}). \tag{2.7}$$

From the form of the iterative scheme (2.6) we see that we have to use the following two-level fractional step method (Marchuk [1971]) where we work with vector corrections $\zeta^{(n)}$

$$\left.\begin{aligned}(D+P_1)\zeta^{(n+\frac{1}{2})} &= b-Au^{(n)}\\ (D+Q_1)\zeta^{(n+1)} &= D\zeta^{(n+\frac{1}{2})}\\ u^{(n+1)} &= u^{(n)}+\zeta^{(n+1)}.\end{aligned}\right\} \tag{2.8}$$

If the matrix A is split up into more than three matrices, we can follow the same idea and end up with a multi-level fractional step method similar to (2.8). By restricting ourselves to the form (2.4) of R we observe that we can commence to develop various iterative schemes which are

associated with the splitting (2.3) of A depending upon the different forms of P_1, Q_1. In particular, by selecting P_1 and Q_1 to have simple forms such as

$$P_1 = \omega_1 P, \qquad Q_1 = \omega_2 Q \tag{2.9}$$

where ω_1, ω_2 are real "preconditioning" parameters, then it follows that all the known linear first degree iterative schemes have the form (2.2) or (2.7),(2.8), for specific values of the involved parameters τ, ω_1, ω_2 and certain forms of the matrices D,P,Q. Therefore if we assume that A has the form

$$A = D - C_L - C_U \tag{2.10}$$

where D, C_L and C_U are defined as in (3-2.6) and (3-2.7), then we can easily verify from (2.2),(2.3),(2.4) and (2.6), Table 2.1.

Preconditioning Parameters		Acceleration Parameter	Conditioning Matrix	Iterative Method
ω_1	ω_2	τ	R	
0	0	1	D	J
0	0	τ_0	D	JOR
0	0	τ_0	†I	SD
1	0	1	D(I-L)	GS
ω	0	ω	D(I-ωL)	SOR
ω	ω	$\omega(2-\omega)$	D(I-ωL)(I-ωU)	SSOR
ω	ω	ω	D(I-ωL)(I-ωU)	EMA

TABLE 2.1

The optimum value of the acceleration parameter τ is given by $\tau_0 = \frac{2}{\bar{a}+\bar{b}}$ where $\bar{a}$ and $\bar{b}$ are the minimum and maximum eigenvalues of the preconditioned matrix $R^{-1}A$, respectively i.e.,

$$\bar{a} \leq \lambda(R^{-1}A) \leq \bar{b}. \tag{2.11}$$

From the previous considerations we conclude that the iterative

---

† *The form of R is as in (2.4) but D=I.*

schemes in Table 2.1 have been based on the idea of improving the "condition" of the original system by using different types of conditioning matrices. Also, if we observe carefully the J and the JOR schemes, then we can immediately predict that although both methods have identical conditioning matrices, the optimum value τ_0 for the parameter τ assures an improvement in the rate of convergence of the JOR over the J method.

From the previous observation, it follows that given the conditioning matrix R, the most effective iterative scheme is obtained if τ takes its optimum value τ_0. However, this does not seem to be the case (at least at this primary stage) for the GS,SOR,SSOR and EMA iterative methods (see Table 2.1).

We can therefore clearly realise the strong need for a reconstruction of the iterative schemes illustrated in Table 2.1.

4.3 ON THE PRECONDITIONED ITERATIVE METHODS

In this section we consider the splitting (2.10) of A and the form of R given by (2.4). For R=D the produced iterative scheme from (2.2) has been studied (see Chapter 3), hence we will concentrate on the following cases which are not covered in Table 2.1.

CASE	ω_1	ω_2	τ
I	1	0	τ_0
II	ω	0	1
III	ω	0	τ_0

TABLE 3.1

Since we intend to make a thorough analysis when A is not consistently ordered in the general case III, we will assume in cases I and II that A is a consistently ordered matrix with non-vanishing diagonal elements.

Case I

In this case we have from (2.2) the iterative scheme

$$u^{(n+1)} = u^{(n)} + \tau(I-L)^{-1}D^{-1}(b-Au^{(n)}) \tag{3.1}$$

or

$$u^{(n+1)} = Lu^{(n+1)}+(1-\tau)(I-L)u^{(n)}+\tau Uu^{(n)}+\tau c \tag{3.2}$$

which can be written in the more compact form

$$u^{(n+1)} = L_{\tau,1}u^{(n)}+\tau(I-L)^{-1}c \tag{3.3}$$

where

$$L_{\tau,1} = I-\tau(I-L)^{-1}D^{-1}A. \tag{3.4}$$

This scheme is an extrapolated version of the GS method since for $\tau=1$ the two methods coincide. Thus, we will refer to (3.3) as the Extrapolated GS method (EGS method).

An obvious restriction is $\tau\neq 0$ for the EGS method to be completely consistent. Next, we prove the following theorem concerned with the convergence of the EGS method.

Theorem 3.1

If A is a consistently ordered matrix with non-vanishing diagonal elements such that $B=I-D^{-1}A$ has real eigenvalues, then $S(L_{\tau,1})=\bar{S}(L_{\tau,1})<1$ if and only if

$$0 < \tau < 2 \tag{3.5}$$

and

$$\bar{\mu} = S(B) < 1. \tag{3.6}$$

Proof

The preconditioned matrix of the EGS method is given by the expression

$$\Lambda_1 = (I-L)^{-1}D^{-1}A. \tag{3.7}$$

If μ,λ are the eigenvalues of B and Λ_1, respectively, then by working in a similar way towards the proof of Theorem 3-6.2 (see Young [1971],p.143) we obtain the following eigenvalue relationship

$$\lambda = 1-\mu^2. \tag{3.8}$$

In order to determine the range of the parameter τ so that the EGS method is convergent we have to determine the maximum and minimum eigenvalue of Λ_1. But we have that

$$0 \leq \mu^2 \leq \bar{\mu}^2 \tag{3.9}$$

hence from (3.8) we obtain

$$\max_{0\leq\mu^2\leq\bar{\mu}^2} \{\lambda\} = \bar{\lambda} = 1 \text{ and } \min_{0\leq\mu^2\leq\bar{\mu}^2} \{\lambda\} = \underline{\lambda} = 1-\bar{\mu}^2. \tag{3.10}$$

Since $\bar{\lambda}>0$, then the EGS method converges if and only if

$$\underline{\lambda} > 0 \tag{3.11}$$

and

$$0 < \tau < 2/\bar{\lambda}\ . \tag{3.12}$$

Evidently, from (3.10), (3.11) and (3.12) it follows that (3.5), (3.6) hold and therefore the proof of the theorem is complete.

The determination of the optimum value of τ is given by the following theorem.

Theorem 3.2

Let A be a consistently ordered matrix with non-vanishing diagonal elements such that the matrix B has real eigenvalues with

$$\bar{\mu} = S(B) < 1\ . \tag{3.13}$$

If we let

$$\tau = \tau_0 = 2/(2-\bar{\mu}^2)\ , \tag{3.14}$$

then $S(L_{\tau,1})$ is minimised and its corresponding value is given by the expression

$$\bar{S}(L_{\tau,1}) = S(L_{\tau,1}) = \tau_0\bar{\mu}^2/2. \tag{3.15}$$

Proof

The optimum value of τ is given by the formula

$$\tau_0 = 2/(\bar{\lambda}+\underline{\lambda})$$

which yields (3.14) by substituting $\bar{\lambda}$ and $\underline{\lambda}$ from (3.10).

Further, $P(\Lambda_1)$ is evaluated by the expression

$$P(\Lambda_1) = \bar{\lambda}/\underline{\lambda} = 1/(1-\bar{\mu}^2) \tag{3.16}$$

and if we calculate $S(L_{\tau,1})$ from the formula

$$S(L_{\tau,1}) = \frac{P(\Lambda_1)-1}{P(\Lambda_1)+1} \tag{3.17}$$

we obtain (3.15) and the proof of the theorem is complete.

Theorem 3.3

Under the hypotheses of Theorem 3.2 and if τ_0 is defined by (3.14), then

$$\lim_{\bar{\mu}\to 1^-} \frac{R(L_{\tau_0,1})}{R(L_{1,1})} = 2 \tag{3.18}$$

where $L_{1,1}=L$ (see (3-2.19)).

Proof

Similar to the one followed in Theorem 3-6.8.

Case II

In this case from (2.2) we have the following iterative scheme

$$u^{(n+1)} = u^{(n)} + (I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{3.19}$$

or

$$u^{(n+1)} = \omega Lu^{(n+1)} + (1-\omega)Lu^{(n)} + Uu^{(n)} + c \tag{3.20}$$

which can be written in the more compact form

$$u^{(n+1)} = L_{1,\omega}u^{(n)} + (I-\omega L)^{-1}c \tag{3.21}$$

As can be seen, (3.19) is a different form of extrapolating the GS method, which is considered if $\omega=1$. Note that in this case, there is

only one parameter ω and its optimum value will be determined in a similar way to τ_0. If we follow the proof of Theorem 3-6.6 (see Young [1971]p.172), then we have the following theorem concerned with the convergence of (3.21).

Theorem 3.4

If A is a consistently ordered matrix with non-vanishing diagonal elements such that B has real eigenvalues, then

$$S(L_{1,\omega}) = \bar{S}(L_{1,\omega}) < 1 \tag{3.22}$$

if and only if

$$\omega_1^* = -\frac{1-\bar{\mu}^2}{2\bar{\mu}^2} < \omega < \frac{1+\bar{\mu}^2}{\bar{\mu}^2} = \omega_2^* \tag{3.23}$$

and

$$\bar{\mu} = S(B) < 1. \tag{3.24}$$

Proof

Since the matrices A,B fulfill the requirements of Theorem 3-6.2, then we can find the eigenvalue relationship

$$\mu(\lambda\omega+1-\omega)^{\frac{1}{2}} = \lambda \tag{3.25}$$

where μ,λ are eigenvalues of B and $L_{1,\omega}$, respectively. The relationship (3.25) can be written as a quadratic in λ to yield the equation

$$\lambda^2-\mu^2\omega\lambda+\mu^2(\omega-1) = 0. \tag{3.26}$$

A sufficient and necessary condition for the convergence of (3.21) is the roots of (3.26) to be less than one in modulus, or equivalently (see Young [1971]p.171)

$$|\mu^2(\omega-1)| < 1 \quad \text{and} \quad |\mu^2\omega| < 1+\mu^2(\omega-1). \tag{3.27}$$

After some algebraic manipulation the relationships (3.27) can be shown to be equivalent to the inequalities (3.23) and (3.24) hence the proof of the theorem is complete.

Finally, we note that as $\bar{\mu}\to 1^-$, then (3.23) yields the range $\omega\in(0,2)$.

In order to complete the analysis on the iterative scheme (3.21) we prove the following theorem (analogous to Theorem 3-6.7) concerning the determination of the optimum value of the parameter ω.

Theorem 3.5

Let A be a consistently ordered matrix with non-vanishing diagonal elements such that the matrix B has real eigenvalues and such that $\bar{\mu}=S(B)<1$.

If ω_b is defined by (3-6.22), then

$$\bar{S}(\mathcal{L}_{1,\omega_b}) = S(\mathcal{L}_{1,\omega_b}) = \bar{\mu}(\omega_b-1)^{\frac{1}{2}} \tag{3.28}$$

and if $\omega\neq\omega_b$, then

$$\bar{S}(\mathcal{L}_{1,\omega}) = S(\mathcal{L}_{1,\omega}) > S(\mathcal{L}_{1,\omega_b}). \tag{3.29}$$

Moreover, for any ω in the range (3.23) we have

$$\bar{S}(\mathcal{L}_{1,\omega}) = S(\mathcal{L}_{1,\omega}) = \begin{cases} \left|\dfrac{\omega\bar{\mu}^2+[\bar{\mu}^2(\omega^2\bar{\mu}^2-4(\omega-1))]^{\frac{1}{2}}}{2}\right| & , \text{ if } \omega_1^*<\omega\leq\omega_b \\ \bar{\mu}(\omega-1)^{\frac{1}{2}} & , \text{ if } \omega_b\leq\omega<\omega_2^*. \end{cases} \tag{3.30}$$

where ω_1^* and ω_2^* are defined by (3.23).

Finally, if $\omega_1^*<\omega<\omega_b$, then $S(\mathcal{L}_{1,\omega})$ is a strictly decreasing function of ω.

Proof

From equation (3.26) we have that the maximum of the moduli of the values λ is given by the expression

$$r(\omega,\mu) = \left|\frac{\omega\mu^2+\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right| . \tag{3.31}$$

Evidently, we can easily apply the analysis of the SOR theory to (3.31) (see proof of Theorem 6-2.3 and Lemma 6-2.4, Young [1971]) and obtain the relationships (3.28),(3.29) and (3.30), thus completing the proof of the theorem.

From (3.28) and (3-6.23) it follows that we have the following relationship of spectral radii

$$S(\mathcal{L}_{1,\omega_b}) = S(B)S(\mathcal{L}_{\omega_b,\omega_b})^{\frac{1}{2}} \tag{3.32}$$

or in terms of rates of convergence

$$R(\mathcal{L}_{1,\omega_b}) = R(B) + \tfrac{1}{2}R(\mathcal{L}_{\omega_b,\omega_b}). \tag{3.33}$$

Therefore, we see that as $\bar{\mu}=S(B)\to 1^-$, then the rate of convergence of the optimum iterative scheme (3.21) is approximately half of the SOR.

Further, we note that the optimum value of the parameter ω is identical for the SOR and the iterative scheme (3.21).

Case III

In this case we have the iterative scheme

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{3.34}$$

or equivalently

$$u^{(n+1)} = \omega L u^{(n+1)} + [(1-\tau)I + (\tau-\omega)L + \tau U]u^{(n)} + \tau c \tag{3.35}$$

which can be written in the compact form

$$u^{(n+1)} = L_{\tau,\omega}u^{(n)} + \ell_{\tau,\omega} \tag{3.36}$$

where

$$L_{\tau,\omega} = (I-\omega L)^{-1}[(1-\tau)I + (\tau-\omega)L + \tau U] \tag{3.37}$$

and

$$\ell_{\tau,\omega} = \tau(I-\omega L)^{-1}c. \tag{3.38}$$

It can be noted that for certain values of the parameters τ and ω we obtain the previous considered iterative schemes. Also, the above introduced iterative scheme (3.36) is likely to produce a more improved rate of convergence, than any other iterative scheme which possesses the same conditioning matrix. Therefore, it is expected that in general (3.36) will be faster than the SOR method. The iterative method (3.36) will be referred to as the Extrapolated Successive Overrelaxation method (ESOR method). In the remainder of this section we will attempt to find under certain assumptions on the matrix A, what restrictions are imposed on the parameters τ and ω so that the ESOR method converges. We will also determine the optimum values of these parameters so that ESOR attains its maximum rate of convergence.

4.3.1 Irreducible matrices with weak diagonal dominance

We have already seen (Theorem 2-5.3) that irreducible matrices with weak diagonal dominance are non-singular. If A has these properties, then the following theorem can be proved.

Theorem 3.1.1

Let A be an irreducible matrix with weak diagonal dominance. Then,

(a) The GS method converges and the EGS method converges for $0<\tau\leq 1$.

(b) The iterative scheme (3.21) converges for $0\leq\omega\leq 1$ and the ESOR method converges for $0<\tau\leq 1$ and $0\leq\omega\leq 1$.

Proof

We assume that $0<\tau\leq 1$, $0\leq\omega\leq 1$ and that $S(L_{\tau,\omega})\geq 1$. Then for some eigenvalue λ of $L_{\tau,\omega}$ we have $|\lambda|\geq 1$.

Furthermore, we have

$$\det(L_{\tau,\omega}-\lambda I) = \det Q = 0 \tag{3.1.1}$$

where

$$Q = I - \left(\frac{\tau-\omega+\omega\lambda}{\lambda+\tau-1}\right) L - \left(\frac{\tau}{\lambda+\tau-1}\right) U. \tag{3.1.2}$$

We let $\lambda^{-1}=qe^{i\theta}$ where q and θ are real and $0<q\leq 1$, hence we have

$$\left|\frac{\tau-\omega+\omega\lambda}{\lambda+\tau-1}\right| = \left[\frac{(\tau-\omega)^2q^2+2\omega q(\tau-\omega)\cos\theta+\omega^2}{1-2q(1-\tau)\cos\theta+q^2(1-\tau)^2}\right]^{\frac{1}{2}} \leq \frac{\omega+q(\tau-\omega)}{1-q(1-\tau)} \tag{3.1.3}$$

since $q\leq 1$, $0<\tau\leq 1$ and $0\leq\omega\leq 1$. But

$$1 - \frac{\omega+q(\tau-\omega)}{1-q(1-\tau)} = \frac{(1-q)(1-\omega)}{1-q(1-\tau)} \geq 0 \tag{3.1.4}$$

and hence

$$\left|\frac{\tau}{\lambda+\tau-1}\right| \leq \left|\frac{\tau+\omega(\lambda-1)}{\lambda+\tau-1}\right| \leq 1. \tag{3.1.5}$$

Since A is irreducible and has weak diagonal dominance, $D^{-1}A=I-L-U$ possesses also the same properties. From (3.1.2) and (3.1.5) it follows that Q has weak diagonal diminance. However, Q is also irreducible and by Theorem 2-5.3 it follows that $\det Q\neq 0$ which contradicts (3.1.1) and therefore $S(L_{\tau,\omega})<1$. This completes the proof of the theorem.

4.3.2 Positive definite matrices

If we now apply Theorem 3-3.2, then we prove:

Theorem 3.2.1

Let A be a positive definite matrix and let D=diagA. Then

$$\|L_{\tau,\omega}\|_{A^{\frac{1}{2}}} < 1$$

if the matrix $\tau^{-1}[(2-\tau)D+(\tau-\omega)(C_L+C_U)]$ is positive definite or, equivalently, if

$$0 < \tau \leq \omega < 2, \tag{3.2.1}$$

where $\mu_{max} \geq 0$ is the largest eigenvalue of B.

Proof

By Theorem 3-3.2 we need only to show that the matrix

$$M = \frac{1}{\tau}[2D-\omega(C_L+C_U)] - A \tag{3.2.2}$$

is positive definite.

It can be easily shown that the above matrix is equivalent to

$$M = \frac{1}{\tau}[(2-\tau)D+(\tau-\omega)(C_L+C_U)]. \tag{3.2.3}$$

If now M is positive definite, then it has positive diagonal elements, hence from (3.2.3), since D is positive definite, we have that

$$0 < \tau < 2. \tag{3.2.4}$$

On the other hand, if μ_i are the eigenvalues of B, then from (3.2.3) we obtain

$$\frac{1}{\tau}[(2-\tau)+(\tau-\omega)\mu_i] > 0 \tag{3.2.5}$$

for all eigenvalues μ_i of B.

If we now assume

$$\omega < \tau, \tag{3.2.6}$$

then since

$$\mu_{min} \leq 0 \leq \mu_{max} \tag{3.2.7}$$

we obtain from (3.2.5) that

$$(2-\tau) + (\tau-\omega)\mu_{min} > 0 \tag{3.2.8}$$

or by (3.2.4)

$$(2-\omega)\mu_{min} > 0$$

which contradicts (3.2.6), hence

$$\tau \leq \omega. \tag{3.2.9}$$

By (3.2.9) and (3.2.5) we have

$$(2-\tau) + (\tau-\omega)\mu_{max} > 0 \tag{3.2.10}$$

thus

$$\text{if } \mu_{max} \leq 1, \text{ then } 0<\omega<2. \tag{3.2.11}$$

From (3.2.4), (3.2.9) and (3.2.11) it follows that (3.2.1) holds and the proof of the theorem is complete.

4.3.3 L-matrices and related matrices

We now prove an analogous theorem to Theorem 3-5.5 concerning the ESOR method.

Theorem 3.3.1

If A is an L-matrix of order N and if $0 \leq \omega \leq \tau \leq 1$ $(\tau \neq 0)$ then

(a) $S(B)<1$ if and only if $S(L_{\tau,\omega})<1$.

(b) $S(B)<1$ (and $S(L_{\tau,\omega})<1$) if and only if A is an M-matrix;

if $S(B)<1$, then

$$S(L_{\tau,\omega}) \leq 1-\tau+\tau S(B).$$

Proof

Evidently, if $S(L_{\tau,\omega})<1$, then $\bar{\mu}=S(B)<1$. Since now L is a strictly lower triangluar matrix, then $L^N=0$ and because of our assumptions we can easily verify that

$$(I-\omega L)^{-1} = I+\omega L+\omega^2 L^2+\ldots+\omega^{N-1}L^{N-1} \geq 0 \tag{3.3.1}$$

and also that

$$(1-\tau)I+(\tau-\omega)L+\tau U \geq 0. \tag{3.3.2}$$

Thus, from (3.3.1) and (3.3.2) it follows that

$$L_{\tau,\omega} = (I-\omega L)^{-1}((1-\tau)I+(\tau-\omega)L+\tau U) \geq 0. \tag{3.3.3}$$

Since $L_{\tau,\omega}$ is a non-negative matrix, by Theorem 2-1.4 we have that $\bar{\lambda}=S(L_{\tau,\omega})$ is an eigenvalue of $L_{\tau,\omega}$ and that there exist an eigenvector $v>0$ such that

$$L_{\tau,\omega}v = \bar{\lambda}v \tag{3.3.4}$$

or

$$\left(\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+U\right)v = \frac{\bar{\lambda}-1+\tau}{\tau}v \tag{3.3.5}$$

which implies that $\frac{\bar{\lambda}-1+\tau}{\tau}$ is an eigenvalue of the matrix $\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+U$.

Therefore the following inequality holds

$$\frac{\bar{\lambda}-1+\tau}{\tau} \leq S\left(\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+U\right). \tag{3.3.6}$$

If we now assume $\bar{\lambda}\leq 1$, then by Theorem 2-1.3 we have

$$S\left(\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+U\right) \leq S(L+U) = S(B) = \bar{\mu} \tag{3.3.7}$$

and by (3.3.6) we obtain

$$\bar{\lambda}\leq\tau\bar{\mu}+1-\tau\ . \tag{3.3.8}$$

On the other hand, if $\bar{\lambda}\geq 1$, then

$$\frac{\bar{\lambda}-1+\tau}{\tau} \leq S\left(\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+U\right) \leq S\left(\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}L+\frac{\tau+\omega(\bar{\lambda}-1)}{\tau}U\right)$$

$$= \frac{\tau+\omega(\bar{\lambda}-1)}{\tau}\bar{\mu} \tag{3.3.9}$$

thus we finally have

$$\bar{\mu} \geq \frac{\bar{\lambda}-1+\tau}{\tau+\omega(\bar{\lambda}-1)} = 1 + \frac{(1-\omega)(\bar{\lambda}-1)}{\tau+\omega(\bar{\lambda}-1)} \geq 1. \tag{3.3.10}$$

If we summarise our results from the above analysis, then we have shown:

(i) if $\bar{\lambda}\leq 1$, then $\bar{\lambda}\leq\tau\bar{\mu}+1-\tau$

(ii) if $\bar{\lambda}\geq 1$, then $\bar{\mu}\geq 1$

which imply

(iii) if $\bar{\mu}<1$, then $\bar{\lambda}<1$ and we have proved (a). Furthermore, by (i) and Theorem 2-7.2 of Young [1971] we have (b) and the proof of Theorem 3-3.1 is complete.

Theorem 3.3.2

If A is an M-matrix and if

$$1\leq \tau \leq \omega < \frac{2}{1+S(B)}, \tag{3.3.11}$$

then $S(\mathcal{L}_{\tau,\omega})<1$.

Proof

If we now have $1 \leq \tau \leq \omega$, then the matrix

$$T_{\tau,\omega} = (I-\omega L)^{-1}[(\tau-1)I+(\omega-\tau)L+\tau U] \tag{3.3.12}$$

is non-negative and we also have

$$|L_{\tau,\omega}| \leq T_{\tau,\omega}. \tag{3.3.13}$$

If we let $\bar{\gamma}=S(T_{\tau,\omega})$, then since $T_{\tau,\omega} \geq 0$, by Theorem 2-1.4 there exists $v \neq 0$ such that $T_{\tau,\omega} v=\bar{\gamma}v$ which implies that

$$(\tau U+(\omega-\tau+\bar{\gamma}\omega)L)v = (\bar{\gamma}+1-\tau)v. \tag{3.3.14}$$

If $\bar{\gamma} \geq 1$, then $\frac{\omega-\tau+\bar{\gamma}\omega}{\tau} \geq 1$ and from Theorem 2-1.3 we obtain

$$\bar{\gamma}+1-\tau \leq (\omega-\tau+\bar{\gamma}\omega)\bar{\mu} \tag{3.3.15}$$

or

$$\omega \geq (\bar{\gamma}+1)/(1+\bar{\gamma}\bar{\mu}) \geq \frac{2}{1+\bar{\mu}} \tag{3.3.16}$$

Therefore, if (3.3.11) holds, then we must have $\bar{\gamma}<1$. By Theorem 2-1.3 and (3.3.13), it follows that $S(L_{\tau,\omega}) \leq \bar{\gamma}<1$ and the proof of Theorem 3.3.2 is complete.

4.3.4 Consistently ordered matrices

In this section we assume that A is a consistently ordered matrix. By working in an analogous way towards the proof of Theorem 3-6.2 the following can be shown.

Theorem 3.4.1

Let A be a consistently ordered matrix with non-vanishing diagonal elements and let $B=I-(\text{diag}A)^{-1}A$. If μ is an eigenvalue of B and satisfies the relationship

$$(1-\lambda)^2 = \mu^2(1-\lambda\omega), \tag{3.4.1}$$

then λ is an eigenvalue of the matrix

$$\Lambda_\omega = (I-\omega L)^{-1}D^{-1}A \tag{3.4.2}$$

and vice versa.

For the convergence of the ESOR method we prove:

Theorem 3.4.2

If A is a consistently ordered matrix with non-vanishing diagonal

elements and if the matrix B has real eigenvalues, then the ESOR method converges if and only if

$$\bar{\mu} = S(B)<1 \tag{3.4.3}$$

and the parameters τ and ω lie in either of the following ranges:

for $\omega \geq 0$,

$$\left.\begin{array}{ll} 0<\tau<1 & \text{and } 0\leq\omega<1 \\ \text{or} & \\ 1\leq\tau<2 & \text{and } 1\leq\omega\leq 2 \end{array}\right\} \tag{3.4.4a}$$

while for $\omega \leq 0$, the ranges of τ remain the same but the corresponding ranges of ω are the following:

$$\begin{array}{l} -1\leq\omega\leq 0, \\ -2\leq\omega\leq -1, \end{array} \tag{3.4.4b}$$

Proof

Since the matrix A satisfies the requirements of Theorem 3.4.1 it follows that (3.4.1) holds. But (3.4.1) can be written alternatively to yield the quadratic

$$\lambda^2-(2-\mu^2\omega)\lambda+(1-\mu^2) = 0. \tag{3.4.5}$$

On the other hand, from Theorem 3-3.1 it follows that the ESOR method converges if and only if

$$S(L_{\tau,\omega}) < 1. \tag{3.4.6}$$

By assuming a+ib to be an eigenvalue of Λ_ω, then $1-\tau(a+ib)$ is an eigenvalue of $L_{\tau,\omega}$ with modulus

$$[(1-\tau a)^2+\tau^2 b^2]^{\frac{1}{2}}$$

thus (3.4.6) becomes

$$\tau^2(a^2+b^2)<2\tau a. \tag{3.4.7}$$

From this inequality we see that we always have

$$\tau a > 0 \tag{3.4.8}$$

which indicates that we have to distinguish the following cases.

Case I: $a>0$ and Case II: $a<0$, that is, the real parts of the eigenvalues of Λ_ω to be either positive or negative.

Case I

In this case we have $a>0$ and $\tau>0$, thus from (3.4.7) it follows that

$$\tau < \frac{2a}{a^2+b^2} . \tag{3.4.9}$$

But from Theorem 3.4.1 we have that the eigenvalues of Λ_ω are the roots of (3.4.5) so the eigenvalue with the maximum real part is given by the expression

$$\Gamma(\omega,\mu^2) = \mathrm{Re}\left\{\frac{|2-\omega\mu^2|+\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right\} \tag{3.4.10}$$

hence the range of τ for the ESOR method to converge is the following

$$0 < \tau < 2/\max_{0\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2). \tag{3.4.11}$$

Our problem therefore is to find the quantity $\max_{\mu^2} \Gamma(\omega,\mu^2)$.

We first assume that $\omega^2\bar{\mu}^2-4(\omega-1)<0$, then $\omega>1$ and $\omega^2\mu^2-4(\omega-1)<0$ for all μ^2 such that $0\leq\mu^2\leq\bar{\mu}^2$. In addition, we have that the modulus of the eigenvalues of Λ_ω is given by the expression

$$|r(\omega,\mu^2)| = \sqrt{1-\mu^2} \tag{3.4.12}$$

which implies that

$$1-\mu^2 > 0 \tag{3.4.13}$$

and therefore (3.4.3) holds. In this case $a>0$, which implies that

$$2-\mu^2\omega > 0 \tag{3.4.14}$$

hence the parameter ω lies in the following range

$$1<\omega\leq 2<2/\bar{\mu}^2. \tag{3.4.15}$$

On the other hand, from (3.4.10) it follows that

$$\Gamma(\omega,\mu^2) = \frac{2-\omega\mu^2}{2} \tag{3.4.16}$$

thus

$$\max_{0\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \Gamma(\omega,0) = 1. \tag{3.4.17}$$

Let us now consider the case where $\omega^2\bar{\mu}^2-4(\omega-1)>0$, then we define μ_0^2 by

$$\mu_0^2 = \begin{cases} 4(\omega-1)/\omega^2 & \text{if } \omega\geq 1 \\ 0 & \text{if } \omega\leq 1. \end{cases} \tag{3.4.18}$$

Next, if $\mu^2 \leq \mu_0^2$, then $\omega^2\mu^2 - 4(\omega-1) \leq 0$ and $\max_{\mu^2} \Gamma(\omega,\mu^2) = 1$. Moreover, for $0 \leq \mu_0^2 \leq \bar{\mu}^2$ the function $\Gamma(\omega,\mu^2)$ is given by the expression (see (3.4.10))

$$\Gamma(\omega,\mu^2) = \frac{2-\omega\mu^2+\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2} . \tag{3.4.19}$$

It can be easily verified that $\Gamma(\omega,\mu^2)$ is an increasing function of μ^2, thus

$$\max_{0\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \frac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2} = \Gamma(\omega,\bar{\mu}^2) \tag{3.4.20}$$

and also

$$\Gamma(\omega,\bar{\mu}^2) < \Gamma(\omega,1) = 2-\omega. \tag{3.4.21}$$

Summarising our results we have that

$$\max_{0\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \begin{cases} \Gamma(\omega,\bar{\mu}^2) < 2-\omega & , \text{ if}^{\dagger}\ 0\leq\omega<1 \\ 1 & , \text{ if } 1\leq\omega\leq 2. \end{cases} \tag{3.4.22}$$

By combining (3.4.22) and (3.4.11) we readily see that the relationships (3.4.4a) hold.

Case II

In this case we have that $a<0$ and $\tau<0$, thus from (3.4.7) it follows that

$$\frac{2a}{a^2+b^2} < \tau. \tag{3.4.23}$$

By following a similar analysis as in Case I we can show that (3.4.3) holds. In addition, since in this case

$$2-\omega\mu^2 < 0 \tag{3.4.24}$$

we conclude that the range for the preconditioning parameter ω is

$$2/\mu^2 < \omega < \infty. \tag{3.4.25}$$

or

$$\max_{0\leq\mu\leq\bar{\mu}^2}\{2/\mu\} < \omega < \infty \tag{3.4.26}$$

implying that $\omega \to \infty$. On the other hand, we have

$$\Gamma(\omega,\mu^2) = \mathrm{Re}\left\{\left(\omega\mu^2 - 2 + \sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}\right)/2\right\} \tag{3.4.27}$$

---

†*Here we assume $\omega\geq 0$. We discuss later the case $\omega\leq 0$.*

which is an increasing function of ω. This implies that if $\omega \to \infty$, then $\tau \to 0^-$. Therefore, in this case the ESOR method does not converge.

Further, if we assume that $\omega \lesssim 0$, then we let

$$\hat{\omega} = -\omega \tag{3.4.28}$$

and define the ESOR method as

$$u^{(n+1)} = u^{(n)} + \tau(I-\hat{\omega}L)^{-1}D^{-1}(b-Au^{(n)}) \tag{3.4.29}$$

where we see that we can apply the same theory as for the case $\omega \gtrsim 0$. Therefore, from (3.4.4a) and (3.4.4b) we obtain that

$$\begin{aligned} &0<\tau<1 \quad \text{and} \quad 0\lesssim\hat{\omega}<1 \\ \text{or} \\ &1\lesssim\tau<2 \quad \text{and} \quad 1\lesssim\hat{\omega}\lesssim 2, \end{aligned} \tag{3.4.30}$$

where if we use (3.4.28), then we obtain (3.4.4b) and the proof of the theorem is complete.

We now seek to determine the optimum values of τ and ω such that $S(L_{\tau,\omega})$ is minimised. This is achieved when

$$|(1-\tau\bar{a})+i\tau\bar{b}| = |(1-\tau\underline{a})+i\tau\underline{b}| \tag{3.4.31}$$

with $\quad \underline{a} \lesssim a \lesssim \bar{a}\,, \quad \underline{b} \lesssim b \lesssim \bar{b}$

where $\underline{a}$ and $\bar{a}$ are either positive or negative values. From (3.4.31) we see that if

$$\bar{b} = \underline{b} = 0 \tag{3.4.32}$$

and

$$\tau = \tau_0 = \frac{2}{\bar{a}+\underline{a}}\,, \tag{3.4.33}$$

then $S(L_{\tau,\omega})$ attains its minimum value which is given by the expression

$$S(L_{\tau_0,\omega}) = \frac{k(\Lambda_\omega)-1}{k(\Lambda_\omega)+1} \tag{3.4.34}$$

where the quantity $k(\Lambda_\omega)$ is defined as

$$k(\Lambda_\omega) = \frac{\bar{a}}{\underline{a}} \tag{3.4.35}$$

and will be referred to as the virtual condition number of the matrix Λ_ω. Evidently, the optimum value of ω will be determined such that (3.4.32) to hold and such that $k(\Lambda_\omega)$ attains its minimum value.

Theorem 3.4.4

Let A be a consistently ordered matrix with non-vanishing diagonal elements such that the matrix B has real eigenvalues with $\bar{\mu}=S(B)<1$.

(i) for any ω in the range $0\leq\omega\leq2$ the virtual condition number of Λ_ω is given by

$$k(\Lambda_\omega) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}, & \text{if } 0\leq\omega<1 \quad (3.4.36) \\ \dfrac{2}{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}, & \text{if } 1\leq\omega\leq\omega_b' \quad (3.4.37) \\ \dfrac{2}{2-\omega\bar{\mu}^2}, & \text{if } \omega_b'\leq\omega\leq2 \quad (3.4.38) \end{cases}$$

where

$$\omega_b' = \frac{2}{1+\sqrt{1-\bar{\mu}^2}} \qquad (3.4.39)$$

and $k(\Lambda_\omega)$ is a strictly decreasing function of ω for $0<\omega<\omega_b'$. Moreover, $k(\Lambda_\omega)$ is minimised if we let

$$\omega = \omega_b' \qquad (3.4.40)$$

and its corresponding value is given by

$$k(\Lambda_{\omega_b'}) = 1/(1-\bar{\mu}^2)^{\frac{1}{2}}. \qquad (3.4.41)$$

On the other hand, if we also let

$$\tau = \tau_0 = \omega_b' , \qquad (3.4.42)$$

then the spectral radius $S(L_{\tau,\omega})$ attains its minimum value which is given by the expression

$$S(L_{\tau_0,\omega_b'}) = S(L_{\omega_b',\omega_b'}) = \frac{1-\sqrt{1-\bar{\mu}^2}}{1+\sqrt{1-\bar{\mu}^2}} = \omega_b'-1 . \qquad (3.4.43)$$

The proof of this theorem follows on the next page.

Proof

The eigenvalues of $\Lambda_\omega = (I-\omega L)^{-1}D^{-1}A$ are the roots of (3.4.5) and are given by the expression

$$r(\omega,\mu^2) = \frac{2-\omega\mu^2 \pm \sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2} \qquad (3.4.44)$$

hence in the case where Λ_ω has complex eigenvalues we have

$$a = \frac{2-\omega\mu^2}{2} \quad \text{and} \quad b = \sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}/2. \qquad (3.4.45)$$

From (3.4.45) it follows that

$$\underline{b} = \min_{0\leq\mu^2\leq\bar{\mu}^2} b = 0 \qquad (3.4.46)$$

and

$$\bar{b} = \max_{0\leq\mu^2\leq\bar{\mu}^2} b = \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}/2 \qquad (3.4.47)$$

which imply that for (3.4.32) to be satisfied we must have

$$\omega^2\bar{\mu}^2-4(\omega-1) = 0 \qquad (3.4.48)$$

or the preconditioning parameter ω to take either the value

$$\omega = \omega_b' \qquad (3.4.49)$$

or

$$\omega = \omega_b'' \qquad (3.4.50)$$

where it can be readily verified that

$$1\leq\omega_b'<2<\omega_b'' \; . \qquad (3.4.51)$$

Next, we seek to determine $k(\Lambda_\omega)$ for $0\leq\omega\leq 2$ and for $2<\omega<\infty$.

If $0\leq\omega\leq 2$, then we recall from (3.4.22) that the eigenvalue of Λ_ω with the maximum real part is given by

$$\max_{0\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}, & \text{if } 0\leq\omega<1 \\ 1, & \text{if } 1\leq\omega\leq 2. \end{cases} \tag{3.4.52}$$

In order to determine the eigenvalue with the minimum real part, we define the function

$$\gamma(\omega,\mu^2) = \text{Re}\left\{\frac{|2-\omega\mu^2|-\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right\} \tag{3.4.53}$$

thus in this case we have

$$\gamma(\omega,\mu^2) = \text{Re}\left\{\frac{2-\omega\mu^2-\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right\} \quad . \tag{3.4.54}$$

Moreover, we prove:

Lemma 3.4.5

Under the hypotheses of Theorem 3.4.4 we have

$$\min_{0\leq\mu^2\leq\bar{\mu}^2} \gamma(\omega,\mu^2) = \gamma(\omega,\bar{\mu}^2) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}, & \text{if } 0\leq\omega\leq\omega_b' \\ \dfrac{2-\omega\bar{\mu}^2}{2}, & \text{if } \omega_b'\leq\omega<2 \end{cases} \tag{3.4.55}$$

Proof

If $\omega^2\bar{\mu}^2-4(\omega-1)<0$, then $\omega>1$ and $\omega^2\mu^2-4(\omega-1)<0$ for all μ^2 such that $0\leq\mu^2\leq\bar{\mu}^2$, hence from (3.4.54) we have that

$$\min_{0\leq\mu^2\leq\bar{\mu}^2}\gamma(\omega,\mu^2) = \gamma(\omega,\bar{\mu}^2) = \frac{2-\omega\bar{\mu}^2}{2} \quad . \tag{3.4.56}$$

Next, if $\mu^2\leq\mu_0^2$, then $\omega^2\mu^2-4(\omega-1)\leq 0$ and $\min_{\mu^2}\gamma(\omega,\mu^2)$ is again given by (3.4.56). Alternatively, if $0\leq\mu_0^2\leq\mu^2\leq\bar{\mu}^2$, then

$$\gamma(\omega,\mu^2) = \frac{2-\omega\mu^2-\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2} \tag{3.4.57}$$

which is a decreasing function of μ^2 since

$$\text{sign}\left\{\frac{\partial\gamma(\omega,\mu^2)}{\partial\mu^2}\right\} = -\text{ sign}[(\sqrt{\omega^2\mu^2-4(\omega-1)}+\omega\mu)^2] \tag{3.4.58}$$

and we have that

$$\min_{0\leq\mu^2\leq\bar{\mu}^2} \gamma(\omega,\mu^2) = \gamma(\omega,\bar{\mu}^2) = \frac{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}. \quad (3.4.59)$$

Consequently, for all μ^2 such that $0\leq\mu^2\leq\bar{\mu}^2$ we prove that

$$\min_{\mu^2} \gamma(\omega,\mu^2) = \gamma(\omega,\bar{\mu}^2). \quad (3.4.60)$$

Since the function $4(\omega-1)/\omega^2$ is an increasing function of ω in the range $0<\omega<2$ and from (3.4.48) we have $4(\omega_b'-1)/(\omega_b')^2=\bar{\mu}^2$, it follows that if $0<\omega\leq\omega_b'$, then $\bar{\mu}^2\geq 4(\omega-1)/\omega^2$ and $\min_{\mu^2} \gamma(\omega,\mu^2)$ is given by (3.4.59) whereas if $\omega_b'\leq\omega<2$, then $\bar{\mu}^2\leq 4(\omega-1)/\omega^2$ and $\min_{\mu^2} \gamma(\omega,\mu^2)$ is given by (3.4.56), hence (3.4.55) holds and the proof of Lemma 3.4.5 is complete.

From (3.4.52) and (3.4.55) we readily see that $k(\Lambda_\omega)$ is given by (3.4.36), (3.4.37), (3.4.38). From this it follows that $k(\Lambda_\omega)>k(\Lambda_{\omega_b'})$ if $\omega_b'<\omega<2$.

Next, we seek to show that if $0<\omega<\omega_b'$, then $k(\Lambda_\omega)$ is a decreasing function of ω. But for $0\leq\omega\leq 1$ we find

$$\text{sign}\left[\frac{\partial}{\partial\omega}k(\Lambda_\omega)\right] = \text{sign}(\bar{\mu}-1) = -1 \quad (3.4.61)$$

and for $1\leq\omega<\omega_b'$

$$\text{sign}\left[\frac{\partial}{\partial\omega}k(\Lambda_\omega)\right] = -\text{sign}(2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]})$$

where

$$(2-\omega\bar{\mu}^2)^2 = \omega^2\bar{\mu}^4-4\omega\bar{\mu}^2+4$$

and

$$(\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]})^2 = \omega^2\bar{\mu}^4-4\omega\bar{\mu}^2+4\bar{\mu}^2$$

hence

$$\text{sign}\left[\frac{\partial}{\partial\omega}k(\Lambda_\omega)\right] = -1. \quad (3.4.62)$$

Consequently, if $\omega=\omega_b'$, then $k(\Lambda_\omega)$ is minimised and its corresponding value is given by (3.4.41). For this value of ω we have from (3.4.55) that

$$\gamma(\omega_b',\bar{\mu}^2) = (1-\bar{\mu}^2)^{\frac{1}{2}} \quad (3.4.63)$$

hence the optimum value of τ is determined by (3.4.33), (3.4.52), (3.4.63) and is given by (3.4.42). Finally, from (3.4.34) and (3.4.38) we easily prove the validity of (3.4.43).

From the above analysis we see that if $\omega \leq 0$, then by using (3.4.28) we can develop the same theory as for $\omega \geq 0$ and obtain the same results with some evident modifications $(\hat{\omega}_b' = -\omega_b')$.

A corollary from Theorem 3.4.4 is that although the ESOR scheme is different from SOR (when $\tau \neq \omega$), the two methods have the same rate of convergence at the optimum stage. However, as it will be seen from the following analysis, this does not happen in the more general case where the matrix A is consistently ordered and the Jacobi iteration matrix B has real eigenvalues μ_i such that $\underline{\mu} = \min_i |\mu_i| \neq 0$. On the contrary it is expected to obtain a greater rate of convergence for ESOR than the SOR method since we let τ take its optimum value, whereas this is precluded in the latter iterative scheme.

Theorem 3.4.6

If A is a consistently ordered matrix with non-vanishing diagonal elements such that the matrix $B=I-D^{-1}A$ has real eigenvalues μ_i, $i=1,2,\ldots,N$ with

$$\underline{\mu} = \min_i |\mu_i| \neq 0 \quad \text{and} \quad \bar{\mu} = \max_i |\mu_i| \ , \tag{3.4.64}$$

then the ESOR method converges if and only if

$$\bar{\mu} = S(B) < 1 \tag{3.4.65}$$

and either (I)

$$0<\tau<2/\max_{\mu^2} \Gamma(\omega,\mu^2) \tag{3.4.66}$$

where if

$$\sqrt{1-\bar{\mu}^2} < 1-\underline{\mu}^2, \tag{3.4.67}$$

then

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}, & \text{if } 0\leq\omega\leq\hat{\omega} \\ \dfrac{2-\omega\underline{\mu}^2}{2}, & \text{if } \hat{\omega}\leq\omega\leq 2 \end{cases} \tag{3.4.68}$$

with

$$\hat{\omega} = \frac{2}{2-\underline{\mu}^2} \tag{3.4.69}$$

otherwise

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}, & \text{if } 0\leq\omega\leq\omega_b'(\bar{\mu}) \\ \dfrac{2-\omega\underline{\mu}^2}{2}, & \text{if } \omega_b'(\bar{\mu})\leq\omega\leq 2 \end{cases} \tag{3.4.70}$$

where

$$\omega_b'(\bar{\mu}) = \frac{2}{1+\sqrt{1-\bar{\mu}^2}}, \tag{3.4.71}$$

or (II) $\max_{\mu^2} \Gamma(\omega,\mu^2) < \tau < 0$ where if $1-\underline{\mu}^2 \leq \sqrt{1-\bar{\mu}^2}$, then

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \begin{cases} \dfrac{\omega\bar{\mu}^2-2}{2}, & \text{if } 2/\underline{\mu}^2\leq\omega\leq\omega_b''(\bar{\mu}) \\ \dfrac{\omega\bar{\mu}^2-2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}, & \text{if } \omega_b''(\bar{\mu})\leq\omega<\infty, \end{cases} \tag{3.4.72}$$

with

$$\omega_b''(\bar{\mu}) = \frac{2}{1-\sqrt{1-\bar{\mu}^2}} \tag{3.4.73}$$

otherwise

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma_2(\omega,\mu^2) = \frac{\omega\bar{\mu}^2-2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2}. \tag{3.4.73a}$$

Proof

Let us first assume that the real parts of the eigenvalues of Λ_ω are positive, then we recall from (3.4.14) that

$$\omega < 2/\mu^2 \tag{3.4.74}$$

and that the eigenvalue of Λ_ω with the maximum real part is given by the expression

$$\Gamma(\omega,\mu^2) = \text{Re}\left\{\frac{2-\omega\mu^2+\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right\} . \tag{3.4.75}$$

Thus, the range of τ for the ESOR method to converge is the following (see (3.4.9))

$$0<\tau<2/\max_{\mu^2} \Gamma(\omega,\mu^2). \tag{3.4.76}$$

Since in this case we have that the eigenvalue relationship (3.4.1) is satisfied, we obtain again (3.4.12) and therefore (3.4.65) holds.

From (3.4.64) we see that

$$\underline{\mu}^2 \leq \mu^2 \leq \bar{\mu}^2 \tag{3.4.77}$$

which implies that in order to examine the position of μ_0^2 (see (3.4.18)) with respect to $\underline{\mu}^2$ and $\bar{\mu}^2$ we have to distinguish the following cases: (i) $0<\underline{\mu}^2\leq\bar{\mu}^2<\mu_0^2$, (ii) $0<\underline{\mu}^2\leq\mu_0^2\leq\bar{\mu}^2$ and (iii) $0\leq\mu_0^2<\underline{\mu}^2\leq\bar{\mu}^2$.

Case (i): $0<\underline{\mu}^2\leq\bar{\mu}^2<\mu_0^2$

In this case we have that $\omega^2\bar{\mu}^2-4(\omega-1)<0$ hence $\omega>1$ and $\omega^2\mu^2-4(\omega-1)<0$ for all μ^2 satisfying (3.4.77). In addition, we have from (3.4.75) that

$$\Gamma(\omega,\mu^2) = \frac{2-\omega\mu^2}{2} \tag{3.4.78}$$

thus

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega,\mu^2) = \frac{2-\omega\underline{\mu}^2}{2} = A \tag{3.4.79}$$

where the range for the parameter ω is

$$1<\omega\leq 2<2/\bar{\mu}^2. \tag{3.4.80}$$

Case (ii): $0<\underline{\mu}^2\leq\mu_0^2\leq\bar{\mu}^2$

In order to examine this case we distinguish two subcases according

to which (a) $\mu_0^2 \leq \mu^2$ and (b) $\mu^2 \leq \mu_0^2$.

Subcase (a): $\mu_0^2 \leq \mu^2$

For this subcase we have $\omega^2\mu^2-4(\omega-1) \geq 0$ thus (3.4.75) yields

$$\Gamma(\omega,\mu^2) = \frac{2-\omega\mu^2+\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2} \tag{3.4.81}$$

which is an increasing function of μ^2 since

$$\text{sign}\left(\frac{\partial\Gamma(\omega,\mu^2)}{\partial\mu^2}\right) = \text{sign}[(\sqrt{\omega^2\mu^2-4(\omega-1)}-\omega\mu)^2]. \tag{3.4.82}$$

Consequently,

$$\max_{\underline{\mu}^2 \leq \mu^2 \leq \bar{\mu}^2} \Gamma(\omega,\mu^2) = \Gamma(\omega,\bar{\mu}^2) = \frac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2} \tag{3.4.83}$$

where we easily verify that

$$\Gamma(\omega,\bar{\mu}^2) < \Gamma(\omega,1) = 2-\omega = B \tag{3.4.84}$$

Subcase (b): $\mu^2 \leq \mu_0^2$

Evidently, for this case we have $\omega^2\mu^2-4(\omega-1) \leq 0$ which has already been examined in Case (i).

Case (iii): $0 \leq \mu_0^2 < \underline{\mu}^2 \leq \bar{\mu}^2$

For this case we immediately find that $\omega^2\mu^2-4(\omega-1) \geq 0$ for all μ^2 which satisfy (3.4.77) and therefore we obtain the results of subcase (a).

On the other hand, we let

$$[\omega_b'(\underline{\mu})]^2\underline{\mu}^2-4(\omega_b'(\underline{\mu})-1) = 0 \tag{3.4.85}$$

since it is easily verified that

$$1<\omega_b'(\underline{\mu})<2<2/\bar{\mu}^2<\omega_b''(\underline{\mu}). \tag{3.4.86}$$

where

$$\omega_b'(\underline{\mu}) = \frac{2}{1+\sqrt{1-\underline{\mu}^2}} \tag{3.4.87}$$

and

$$\omega_b''(\underline{\mu}) = \frac{2}{1-\sqrt{1-\underline{\mu}^2}}. \tag{3.4.88}$$

Similarly, we let

$$[\omega_b'(\bar{\mu})]^2\bar{\mu}^2-4(\omega_b'(\bar{\mu})-1) = 0 \tag{3.4.89}$$

since

$$1<\omega_b'(\bar{\mu})<2<2/\bar{\mu}^2<\omega_b''(\bar{\mu}) \tag{3.4.90}$$

where

$$\omega_b'(\bar{\mu}) = \frac{2}{1+\sqrt{1-\bar{\mu}^2}} \tag{3.4.91}$$

and

$$\omega_b''(\bar{\mu}) = \frac{2}{1-\sqrt{1-\bar{\mu}^2}} \tag{3.4.92}$$

By combining (3.4.86) and (3.4.90) we find

$$1<\omega_b'(\underline{\mu})\leq\omega_b'(\bar{\mu})<2<2/\bar{\mu}^2<\omega_b''(\bar{\mu})\leq\omega_b''(\underline{\mu}) \tag{3.4.93}$$

From Case (i) we find that if $\omega_b'(\bar{\mu})\leq\omega$, then $\max_{\mu^2} \Gamma(\omega,\mu^2)$ is given by (3.4.79), whereas from Case (iii) if $\omega\leq\omega_b'(\underline{\mu})$, then $\max_{\mu^2} \Gamma(\omega,\mu^2)$ is obtained by (3.4.83). Since for ω in the range $\omega_b'(\underline{\mu})\leq\omega\leq\omega_b'(\bar{\mu})$ we have that $\max_{\mu^2} \Gamma(\omega,\mu^2)$ is either expressed by (3.4.79) or (3.4.83) we have to examine the sign of the quantity A-B (see (3.4.79) and (3.4.84)).

It can be easily seen that

$$\text{sign}(B-A) = \text{sign}(\hat{\omega}-\omega) \tag{3.4.94}$$

where $\hat{\omega}$ is given by (3.4.69).

Thus, for ω in the range

$$\omega_b'(\underline{\mu})\leq\omega\leq\omega_b'(\bar{\mu}) \tag{3.4.95}$$

we have

$$\max_{\mu^2} \Gamma(\omega,\mu^2) = \begin{cases} B & \text{, if } \omega<\hat{\omega} \\ A & \text{, if } \omega>\hat{\omega}. \end{cases} \tag{3.4.96}$$

On the other hand, we have that

$$\hat{\omega}\leq\omega_b'(\bar{\mu}) \tag{3.4.97}$$

if the relationship (3.4.67) is satisfied. This implies that if (3.4.67) holds, then $\max_{\mu^2} \Gamma(\omega,\mu^2)$ is given by (3.4.68), otherwise it is given by (3.4.70).

Let us consider the case where $2<2/\underline{\mu}^2<\omega<\infty$. By following a similar

analysis as previously and noting that $\hat{\omega}<2$, the validity of (II) can be easily verified and the proof of the theorem is now complete.

Lemma 3.4.7

Under the hypothesis of Theorem 3.4.6 and if $\bar{\mu}=S(B)<1$, then for $0\leq\omega\leq 2$

$$\min_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \gamma(\omega,\mu^2) = \gamma(\omega,\bar{\mu}^2) \tag{3.4.98}$$

where

$$\gamma(\omega,\mu^2) = \mathrm{Re}\left\{\frac{|2-\omega\mu^2|-\sqrt{\mu^2[\omega^2\mu^2-4(\omega-1)]}}{2}\right\}. \tag{3.4.99}$$

Moreover, for any ω in the range $0\leq\omega\leq 2$, we have

$$\gamma(\omega,\bar{\mu}^2) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2} & \text{,if } 0\leq\omega\leq\omega_b'(\bar{\mu}) \\ \dfrac{2-\omega\bar{\mu}^2}{2} & \text{,if } \omega_b'(\bar{\mu})\leq\omega\leq 2, \end{cases} \tag{3.4.100}$$

whereas for any ω in the range $2/\underline{\mu}^2<\omega<\infty$

$$\gamma(\omega,\bar{\mu}^2) = \begin{cases} \dfrac{\omega\underline{\mu}^2-2}{2} & \text{,if } 2/\underline{\mu}^2<\omega\leq\omega_b''(\underline{\mu}) \\ \dfrac{\omega\bar{\mu}^2-2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2} & \text{,if } \omega_b''(\underline{\mu})\leq\omega<\infty \end{cases} \tag{3.4.101}$$

where $\omega_b'(\bar{\mu})$ and $\omega_b''(\underline{\mu})$ are given by (3.4.71) and (3.4.88), respectively.

Proof

Similar to the one followed in Lemma 3.4.5.

Theorem 3.4.8

Let A be a consistently ordered matrix with nonvanishing diagonal elements such that the matrix $B=I-D^{-1}A$ has real eigenvalues μ_i, $i=1,2,\ldots,N$ with

$$\underline{\mu} = \min_i |\mu_i| \neq 0 , \quad \bar{\mu} = \max_i |\mu_i| \tag{3.4.102}$$

and such that $\bar{\mu}=S(B)<1$.

(i) For any ω in the range $0\leq\omega\leq 2$, we have that:

if

$$\sqrt{1-\bar{\mu}^2}<1-\underline{\mu}^2 , \tag{3.4.103}$$

then

$$k(\Lambda_\omega) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}} & \text{, if } 0\leq\omega\leq\hat{\omega} \\ \dfrac{2-\omega\underline{\mu}^2}{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}} & \text{, if } \hat{\omega}\leq\omega\leq\omega_b'(\bar{\mu}) \\ \dfrac{2-\omega\underline{\mu}^2}{2-\omega\bar{\mu}^2} & \text{, if } \omega_b'(\bar{\mu})\leq\omega\leq 2 \end{cases} \tag{3.4.104}$$

otherwise

$$k(\Lambda_\omega) = \begin{cases} \dfrac{2-\omega\bar{\mu}^2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{2-\omega\bar{\mu}^2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}} & \text{, if } 0\leq\omega\leq\omega_b'(\bar{\mu}) \\ \dfrac{2-\omega\underline{\mu}^2}{2-\omega\bar{\mu}^2} & \text{, if } \omega_b'(\bar{\mu})\leq\omega\leq 2 \end{cases} \tag{3.4.105}$$

where $\hat{\omega}$ and $\omega_b'(\bar{\mu})$ are defined by (3.4.69) and (3.4.71), respectively.

Moreover, $k(\Lambda_\omega)$ is a strictly decreasing function of ω for $0<\omega<\omega_b'(\bar{\mu})$ and if we let

$$\omega = \omega_b'(\bar{\mu}) = \omega_b', \tag{3.4.106}$$

then $k(\Lambda_\omega)$ is minimised and its corresponding value is given by

$$k(\Lambda_{\omega_b'}) = \frac{1}{\sqrt{1-\bar{\mu}^2}}\left[1-\frac{\underline{\mu}^2}{1+\sqrt{1-\bar{\mu}^2}}\right] = \frac{\omega_b'}{2-\omega_b'}\left[1-\frac{\underline{\mu}^2\omega_b'}{2}\right] \tag{3.4.107}$$

On the other hand, if we also let

$$\tau = \tau_0 = \frac{4\omega_b'}{4-\underline{\mu}^2[\omega_b']^2}, \tag{3.4.108}$$

then $S(L_{\tau,\omega})$ attains its minimum value which is given by the expression

$$S(L_{\tau_0,\omega_b'}) = 1+\tau_0-2\tau_0/\omega_b'. \tag{3.4.109}$$

(ii) For any ω in the range $2/\underline{\mu}^2<\omega<\infty$, if $2/\underline{\mu}^2<\omega_b''(\bar{\mu})$, then

$$k(\Lambda_\omega) = \begin{cases} \dfrac{\omega\bar{\mu}^2-2}{\omega\underline{\mu}^2-2} & \text{, if } \dfrac{2}{\underline{\mu}^2}<\omega\leq\omega_b''(\bar{\mu}) \\ \dfrac{\omega\bar{\mu}^2-2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{\omega\underline{\mu}^2-2} & \text{, if } \omega_b''(\bar{\mu})\leq\omega\leq\omega_b''(\underline{\mu}) \\ \dfrac{\omega\bar{\mu}^2-2+\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}}{\omega\underline{\mu}^2-2-\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2-4(\omega-1)]}} & \text{, if } \omega_b''(\underline{\mu})\leq\omega<\infty \end{cases} \tag{3.4.110}$$

otherwise

$$k(\Lambda_\omega) = \begin{cases} \dfrac{\omega\underline{\mu}^2 - 2 + \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}{\omega\underline{\mu}^2 - 2}, & \text{if } 2/\underline{\mu}^2 < \omega \le \omega_b''(\underline{\mu}) \\ \dfrac{\omega\bar{\mu}^2 - 2 + \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}{\omega\bar{\mu}^2 - 2 - \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}, & \text{if } \omega_b''(\underline{\mu}) \le \omega < \infty \end{cases} \tag{3.4.111}$$

where $\omega_b''(\underline{\mu})$ is defined by (3.4.88). Moreover, $k(\Lambda_\omega)$ is minimised if we let

$$\omega = \omega_b''(\bar{\mu}) = \omega_b''$$

and its corresponding value is given by the expression

$$k(\Lambda_{\omega_b''}) = (\omega_b'' - 2) / [\omega_b''(\omega_b''\underline{\mu}^2/2 - 1)].$$

Further, if we let $\tau = \tau_o = -\hat{\tau}_o$ where

$$\hat{\tau}_o = 4\omega_b'' / (\omega_b''^2\underline{\mu}^2 - 4) \tag{3.4.112}$$

then

$$S(\Lambda_{\hat{\tau}_o, \omega_b''}) = \hat{\tau}_o(1 - 2/\omega_b'') - 1. \tag{3.4.113}$$

Proof

From Theorem 3.4.6 and Lemma 3.4.7 we can easily verify (3.4.104), (3.4.105) and (3.4.110). Evidently, $k(\Lambda_\omega) > k(\Lambda_{\omega_b'(\bar{\mu})})$ if $\omega_b'(\bar{\mu}) < \omega < 2$.

Next, we seek to show that if $0 < \omega < \omega_b'(\bar{\mu})$, then $k(\Lambda_\omega)$ is a decreasing function of ω.

Since

$$\text{sign}\left(\frac{\partial}{\partial\omega}\left[\frac{2 - \omega\bar{\mu}^2 + \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}{2 - \omega\bar{\mu}^2 - \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}\right]\right) = \text{sign}(\bar{\mu} - 1)$$

and

$$\text{sign}\left(\frac{\partial}{\partial\omega}\left[\frac{2 - \omega\underline{\mu}^2}{2 - \omega\bar{\mu}^2 - \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}\right]\right)$$

$$= \text{sign}\left(-\underline{\mu}^2(2 - \omega\bar{\mu}^2 - \sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}) + \bar{\mu}^2(2 - \omega\underline{\mu}^2)\frac{\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]} + (\omega\bar{\mu}^2 - 2)}{\sqrt{\bar{\mu}^2[\omega^2\bar{\mu}^2 - 4(\omega-1)]}}\right)$$

where $2 - \omega\bar{\mu}^2 > \bar{\mu}\sqrt{\omega^2\bar{\mu}^2 - 4(\omega-1)}$

it follows from (3.4.104) and (3.4.105) that for $0 < \omega < \omega_b'(\bar{\mu})$, $k(\Lambda_\omega)$ is a decreasing function of ω and is minimised when ω takes the value given by (3.4.106). From either (3.4.104) or (3.4.105) we see that for this value of ω, $k(\Lambda_\omega)$ is given by

$$k(\Lambda_{\omega_b'}) = \frac{2-\underline{\mu}^2\omega_b'}{2-\bar{\mu}^2\omega_b'} \tag{3.4.114}$$

which by (3.4.91) yields (3.4.107). Evidently, for $\omega=\omega_b'(\bar{\mu})$ either (3.4.68) or (3.4.70) yields

$$\max_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \Gamma(\omega_b',\mu^2) = \frac{2-\omega_b'\underline{\mu}^2}{2} \tag{3.4.115}$$

whereas by (3.4.100) we have

$$\min_{\underline{\mu}^2\leq\mu^2\leq\bar{\mu}^2} \gamma(\omega_b',\mu^2) = \frac{2-\omega_b'\bar{\mu}^2}{2} \tag{3.4.116}$$

thus by (3.4.33) the optimum value of τ is easily verified to be given by (3.4.108). Finally, by combining (3.4.34) and (3.4.107) we obtain (3.4.109) which is the minimum value of $S(L_{\tau,\omega})$.

Similarly, we can prove the second part (ii) of the theorem, thus completing its proof.

From the above theorem we have the following corollary (see Hadjidimos [1978]).

Corollary 3.4.9

Under the hypotheses of Theorem 3.4.8 and if

$$0<\underline{\mu} = \bar{\mu} = \mu<1, \tag{3.4.117}$$

then

$$S(L_{\tau_0,\omega_b'}) = 0 \tag{3.4.118}$$

where

$$\omega_b' = \frac{2}{1+\sqrt{1-\mu^2}} \quad \text{and} \quad \tau_0 = \frac{1}{\sqrt{1-\mu^2}}\ . \tag{3.4.119}$$

Also we see that for $\underline{\mu}=0$ we obtain Theorem 3.4.4 as a special case of Theorem 3.4.8.

From Corollary 3.4.9 we have that under the special condition (3.4.117) one can obtain an exceptionally fast rate of convergence by applying the ESOR method. However, it can be easily verified that under the same condition, one can obtain an analogous relationship to (3.4.118) for the EGS method (i.e. when $\omega=1$).

On the other hand, Corollary 3.4.9 shows how in the ESOR method one can exploit the spectrum of the eigenvalues of the matrix B to achieve the best possible results, whereas such a possibility is precluded in the SOR. We note that the superiority of the ESOR method depends strongly upon $\underline{\mu}$. This can be easily seen since if we assume $\underline{\mu}\to 0$, then from (3.4.108) and (3.4.109) we have that $R(L_{\tau_0,\omega_b'})\to R(L_{\omega_b',\omega_b'})$, whereas if $\underline{\mu}\to\bar{\mu}$, then from Corollary 3.4.9 we see that $S(L_{\tau_0,\omega_b'})\to 0$. Finally, a comparison of the ESOR method and the SOR showing the dependence on $\underline{\mu}$, is given by the following theorem.

Theorem 3.4.10

Under the hypotheses of Theorem 3.4.8 and for fixed $\underline{\mu}$ we have

$$\lim_{\bar{\mu}\to 1^-} \frac{R(L_{\tau_0,\omega_b'})}{R(L_{\omega_b',\omega_b'})} = \frac{1}{1-\underline{\mu}^2} . \tag{3.4.120}$$

Proof

Since in this case $k(\Lambda_{\omega_b'})>>1$ for fixed $\underline{\mu}$, then we have $R(L_{\tau_0,\omega_b'})\sim\frac{2}{k(\Lambda_{\omega_b'})}$ and therefore from (3.4.107) we obtain

$$\lim_{\bar{\mu}\to 1^-} \frac{R(L_{\tau_0,\omega_b'})}{R(L_{\omega_b',\omega_b'})} = \lim_{\bar{\mu}\to 1^-} \frac{1}{1-\dfrac{\underline{\mu}^2}{1+\sqrt{1-\bar{\mu}^2}}} = \frac{1}{1-\underline{\mu}^2} . \tag{3.4.121}$$

As it was seen earlier, the advantages of the ESOR iterative procedure depends upon the value of $\underline{\mu}$. The determination of $\underline{\mu}$ is the added work in the ESOR method as compared with SOR and it may incur some extra computational effort. The need for knowing $\underline{\mu}$ is very strong especially in the more general cases where the matrix A is not consistently ordered and the ESOR theory is expected to hold approximately. Apart from the iterative procedures which can be considered in a similar way as for the determination of $\underline{\mu}$, (see Young [1954], Young and Shaw [1955], Hageman and Kellogg [1968]) another approach is to use the a priori exact and approximate methods.

4.4 THE PRECONDITIONED JACOBI METHOD (PJ METHOD)

In Section 4.2 it was noted that GS,SOR,SSOR and EMA iterative procedures are not the appropriate methods which can produce the maximum rate of convergence using the corresponding conditioning matrix R. A result of this observation was to develop the new iterative schemes (3.1) and (3.34) which were proved (under certain conditions) to be superior over their corresponding "counterparts" (i.e., GS and SOR respectively).

In this section we will attempt to follow a similar approach as in Section 4.3 in order to construct and study iterative schemes which use the more general form of conditioning matrix given by (2.4),(2.9) with $\omega_1=\omega$, $\omega_2=\omega$ and employing the same splitting of A as given by (2.10). We therefore consider the iterative schemes associated with conditioning matrix

$$R = (D-\omega C_L)D^{-1}(D-\omega C_U). \tag{4.1}$$

(SSOR and EMA are iterative methods which possess the above conditioning matrix).

Before we start defining any iterative process using the particular form of R given above, we can obtain a crude idea as to how effective this conditioning matrix is going to be, by comparing it with the conditioning matrix

$$R_1 = D(I-\omega L). \tag{4.2}$$

An alternative form of R, given by (4.1), is the following

$$R = D[I-\omega(L+U)+\omega^2 LU] \tag{4.3}$$

where we see that its effectiveness depends strongly upon the product LU since the remaining part of the right hand side in (4.3) is a good approximation to the matrix

$$A = D[I-(L+U)]. \tag{4.4}$$

Consequently, by comparing the conditioning matrices R_1 and R we conjecture that if any norm of LU (e.g. $\|LU\|_\infty$) is sufficiently small, then the conditioning matrix R may produce slightly better improvement on the "condition" of the preconditioned system than R_1. In other words, we

expect that under certain conditions, the iterative method which is associated with the conditioning matrix R, to possess slightly better rate of convergence than the SOR method.

The associated iterative scheme with R is of the form (2.7) and is given by

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{4.5}$$

where ω,τ are real parameters and their role will be considered later.

We will commence our study of the above scheme by considering first the case where $\tau=1$. Thus, we will concentrate our attention on the iterative process defined by

$$u^{(n+1)} = u^{(n)} + (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{4.6}$$

which is the Jacobi version of (4.1) and will be referred to as the Preconditioned Jacobi method (PJ method). If we consider vector corrections (see (2.8)), then the PJ method can be written as a two-level fractional method given by

$$\left.\begin{aligned} \zeta^{(n+\frac{1}{2})} &= \omega L\zeta^{(n+\frac{1}{2})} + r^{(n)} \\ \zeta^{(n+1)} &= \omega U\zeta^{(n+1)} + \zeta^{(n+\frac{1}{2})} \\ \text{and} \quad u^{(n+1)} &= u^{(n)} + \zeta^{(n+1)} \end{aligned}\right\} \tag{4.7}$$

where

$$r^{(n)} = D^{-1}(b-Au^{(n)}). \tag{4.8}$$

Finally, a more compact form can be obtained from (4.6) to yield

$$u^{(n+1)} = \mathcal{H}_\omega u^{(n)} + \eta_\omega \tag{4.9}$$

where

$$\mathcal{H}_\omega = I-(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \tag{4.10}$$

and

$$\eta_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}c. \tag{4.11}$$

From (4.10) and (3-2.39) we see that the PJ method and SSOR have similar forms and therefore it is expected that the amount of work involved is approximately the same (see Appendix A). It should be noted however that the PJ method defined by (4.7) is a modified version of (1.10), where in the former it is not required to use (1.4) after the criterion of convergence is satisfied, thus reducing the involved computational work.

4.5 CONVERGENCE OF THE PJ METHOD

From (4.6) we see that the preconditioned matrix of the PJ method is

$$B_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A. \tag{5.1}$$

If we assume that A is a real symmetric and positive definite matrix, then B_ω is similar to the matrix

$$\begin{aligned}\bar{B}_\omega &= D^{-\frac{1}{2}}(D-\omega C_U)B_\omega(D-\omega C_U)^{-1}D^{\frac{1}{2}} \\ &= D^{\frac{1}{2}}(D-\omega C_L)^{-1}A(D-\omega C_U)^{-1}D^{\frac{1}{2}} \\ &= [D^{\frac{1}{2}}(D-\omega C_L)^{-1}]A[D^{\frac{1}{2}}(D-\omega C_L)^{-1}]^T.\end{aligned} \tag{5.2}$$

The last expression of $\bar{B}_\omega$ implies that $\bar{B}_\omega$ is obtained from A by a congruence transformation since the matrix $D^{\frac{1}{2}}(D-\omega C_L)^{-1}$ is non-singular. Furthermore, by Theorem 2-2.4 we have that $\bar{B}_\omega$ is a positive definite matrix which implies that B_ω is similar to a positive definite matrix. From this observation we have that if γ_i and λ_i are the eigenvalues of $\mathcal{H}_\omega$ and B_ω respectively, then they are real and are related through the relationship

$$\gamma_i = 1-\lambda_i \tag{5.3}$$

where

$$\lambda_i > 0. \tag{5.4}$$

Consequently, by Theorem 3-3.1 and (5.3) the PJ method converges if and only if

$$0<\lambda_i<2. \tag{5.5}$$

Theorem 5.1

Let A by a symmetric matrix with positive diagonal elements, then

$$S(\mathcal{H}_\omega)<1 \tag{5.6}$$

if and only if A is positive definite and

$$\omega_\alpha<\omega<\omega_f \tag{5.7}$$

where $\omega_\alpha = 1-\sqrt{2}/2$ and $\omega_f = 1+\sqrt{2}/2$.

Proof

From (5.1) and (3-2.39) it follows that

$$\mathcal{L}_\omega = I-\omega(2-\omega)B_\omega \tag{5.8}$$

and therefore we have the eigenvalue relationship

$$\omega(2-\omega)\lambda = 1-\nu \tag{5.9}$$

where λ,ν are the eigenvalues of B_ω and $\mathcal{L}_\omega$, respectively.

If we now assume that A is a positive definite matrix and (5.7) holds, then from Theorem 3-5.6 we have that $\nu\in[0,1)$ which by (5.9) implies that

$$\lambda\in(0, \frac{1}{\omega(2-\omega)}]. \tag{5.10}$$

Moreover, from (5.7) we have

$$\frac{1}{\omega(2-\omega)} < 2, \tag{5.11}$$

hence (5.5) holds and the PJ method converges.

Suppose now that $S(\mathcal{H}_\omega)<1$, then we have that (5.5) holds.

If $\lambda>0$, then by (5.2) we have that A is positive definite matrix. But by Theorem 3-5.6 $0\leq\nu<1$ if $0<\omega<2$, hence by (5.9) $\lambda\in(0, \frac{1}{\omega(2-\omega)}]$. If now $\lambda<2$ which implies (5.11), we have that (5.7) is satisfied and the proof of the theorem is complete.

4.6 DETERMINATION OF GOOD BOUNDS ON $\lambda(B_\omega)$ AND $\Lambda(B_\omega)$

From the previous section, it is clear that in order to study the PJ method and to determine a good estimate of ω near the optimum we need to determine $S(\mathcal{H}_\omega)$ and then study its behaviour with respect to the preconditioning parameter ω. From (5.3) it follows that $S(\mathcal{H}_\omega)$ is given by

$$S(\mathcal{H}_\omega) = \max\{|1-\lambda(B_\omega)|, |1-\Lambda(B_\omega)|\} \tag{6.1}$$

where $\lambda(B_\omega)$ and $\Lambda(B_\omega)$ are the minimum and maximum eigenvalues of B_ω, respectively. The determination of $\lambda(B_\omega)$ and $\Lambda(B_\omega)$ is therefore essential for our analysis. By following a similar analysis of Habetler and Wachspress [1961], we will attempt to find the eigenvalues of B_ω in terms of certain inner products.

Let us assume that λ is an eigenvalue of B_ω and v an associated eigenvector, then it follows that

$$B_\omega v = \lambda v \tag{6.2}$$

which on substitution of B_ω from (5.1) becomes

$$(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}Av = \lambda v \tag{6.3}$$

or

$$Av = \lambda D(I-\omega L)(I-\omega U)v. \tag{6.4}$$

Furthermore, by taking inner products of both sides with respect to v, (6.4) yields

$$(v,Av) = \lambda(v,D(I-\omega L)(I-\omega U)v) \tag{6.5}$$

which can be solved for λ to give the expression

$$\lambda = \frac{(v,Av)}{(v,D(I-\omega L)(I-\omega U)v)}. \tag{6.6}$$

We can now expand the numerator and denominator in (6.6) to obtain

$$\lambda = \frac{(v,Dv)-(v,DBv)}{(v,Dv)-\omega(v,DBv)+\omega^2(v,DLUv)} \tag{6.7}$$

and if we divide both parts of the ratio by $(v,Dv)\neq 0$, then we have the final representation of λ which is given by the expression

$$\lambda = \frac{1-\hat{a}(v)}{1-\omega\hat{a}(v)+\omega^2\hat{\beta}(v)} \tag{6.8}$$

where

$$\hat{a}(v) = \frac{(v,DBv)}{(v,Dv)} \tag{6.9}$$

and

$$\hat{\beta}(v) = \frac{(v,DLUv)}{(v,Dv)} .$$

From the above expression of λ we see that it would be possible to determine the largest and the smallest eigenvalues of B_ω if we happened to know their associated eigenvectors, respectively. We therefore have to rely on bounds for $\hat{a}(v),\hat{\beta}(v)$ to yield reasonable bounds for $\lambda(B_\omega)$ and $\Lambda(B_\omega)$. Since B_ω is similar to a symmetric matrix, by Theorem 2-1.5, we find that for any $v\neq 0$ we have

$$\lambda(B_\omega) \leq \frac{1-\hat{a}(v)}{1-\omega\hat{a}(v)+\omega^2\hat{\beta}(v)} \leq \Lambda(B_\omega). \tag{6.10}$$

Lemma 6.1

If the eigenvalues μ of B lie in the range

$$m(B) = m \leq \mu \leq M = M(B), \tag{6.11}$$

then the quantities $\hat{a}(v)$ and $\hat{\beta}(v)$ are bounded as follows

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

$$\text{and} \qquad 0 \leq \hat{\beta}(v) \leq S(LU). \tag{6.12}$$

Proof

If we first consider $\hat{a}(v)$, then from (6.9) we have

$$\hat{a}(v) = \frac{(v,DBv)}{(v,Dv)} = \frac{(D^{\frac{1}{2}}v,(D^{\frac{1}{2}}BD^{-\frac{1}{2}})D^{\frac{1}{2}}v)}{(D^{\frac{1}{2}}v,D^{\frac{1}{2}}v)} = \frac{(w,\tilde{B}w)}{(w,w)} \tag{6.13}$$

where $w=D^{\frac{1}{2}}v$ and $\tilde{B}=D^{\frac{1}{2}}BD^{-\frac{1}{2}}$.

Thus, $\hat{a}(v)$ is a Rayleigh quotient with respect to $\tilde{B}$ which is similar to B and by applying Theorem 2-1.5 the first part of (6.12) follows. Similarly, from (6.9) we have

$$\hat{\beta}(v) = \frac{(v,DLUv)}{(v,Dv)} = \frac{(D^{\frac{1}{2}}v,(D^{\frac{1}{2}}LUD^{-\frac{1}{2}})D^{\frac{1}{2}}v)}{(D^{\frac{1}{2}}v,D^{\frac{1}{2}}v)} = \frac{(w,\tilde{L}\tilde{U}w)}{(w,w)} \tag{6.14}$$

where again $w=D^{\frac{1}{2}}v$ and $\tilde{L}=\tilde{U}^T=D^{\frac{1}{2}}LD^{-\frac{1}{2}}$. Hence $\hat{\beta}(v)$ is a Rayleigh quotient with respect to the symmetric and positive definite matrix $\tilde{L}\tilde{U}$ and the proof of the lemma is complete.

Since we assume A to be a positive definite matrix, B is similar to $\tilde{B}=D^{\frac{1}{2}}BD^{-\frac{1}{2}}=I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ hence

$$M(\tilde{B}) = M(B) = M < 1$$

and from (3-6.3) we finally have

$$m \lesssim 0 \lesssim M < 1. \tag{6.15}$$

Next, we note that $\tilde{B}=\tilde{L}+\tilde{U}$ is symmetric and similar to B, hence

$$S(B) = S(\tilde{B}) = S(\tilde{L}+\tilde{U}) = \|\tilde{L}+\tilde{U}\| \lesssim \|\tilde{L}\|+\|\tilde{U}\| = \|\tilde{L}\|+\|\tilde{L}^T\| =$$

$$= 2\|\tilde{L}\| = 2\sqrt{S(\tilde{L}\tilde{U})} = 2\sqrt{S(LU)} \lesssim 2\sqrt{\bar{\beta}} \tag{6.16}$$

where $$S(LU) \lesssim \bar{\beta}. \tag{6.17}$$

Moreover, from (6.16) it follows

$$-m \lesssim 2\sqrt{\bar{\beta}}$$

$$M \lesssim 2\sqrt{\bar{\beta}} \tag{6.18}$$

which implies that if the bounds -m and M exceed $2\sqrt{\bar{\beta}}$ we replace M by $2\sqrt{\bar{\beta}}$ or m by $-2\sqrt{\bar{\beta}}$. Finally, from the above analysis it is readily seen that the following inequalities hold

$$-2\sqrt{\bar{\beta}} \lesssim m \lesssim 0 \lesssim M \lesssim \min(1, 2\sqrt{\bar{\beta}}). \tag{6.19}$$

We are now in a position to determine upper and lower bounds for $\lambda(B_\omega)$ and $\Lambda(B_\omega)$. Although a crude upper bound for $\Lambda(B_\omega)$ was found in Section 4.5 based on the properties of the matrix ξ_ω, nevertheless we present an independent approach to the same problem.

Theorem 6.2

If A is a positive definite matrix, then

$$\Lambda(B_\omega) \lesssim \frac{1}{\omega(2-\omega)} \tag{6.20}$$

where $0<\omega<2$.

Proof

It can be easily verified that

$$D(I-\omega L)(I-\omega U) = \omega(2-\omega)A + D[(1-\omega)I+\omega L][(1-\omega)I+\omega U] \tag{6.21}$$

which when substituted in (6.6) yields

$$\lambda = \frac{(v,Av)}{(v,\omega(2-\omega)Av)+(v,D[(1-\omega)I+\omega L][(1-\omega)I+\omega U]v)} . \tag{6.22}$$

However, from Theorem 2-2.2 we have

$$(v,D[(1-\omega)I+\omega L][(1-\omega)I+\omega U]v) = (D^{\frac{1}{2}}v,D^{\frac{1}{2}}[(1-\omega)I+\omega L][(1-\omega)I+\omega U]D^{-\frac{1}{2}}D^{\frac{1}{2}}v)\geq 0 \tag{6.23}$$

and therefore (6.22) yields

$$\lambda \leq \frac{(v,Av)}{(v,\omega(2-\omega)Av)} \tag{6.24}$$

which is valid for all the eigenvalues of B_ω. On the other hand, it is known that B_ω is a positive definite matrix and therefore by (6.24) we obtain (6.20) provided $0<\omega<2$, hence the proof of the theorem is complete.

The following theorem gives a lower bound for $\lambda(B_\omega)$ obtained by studying the behaviour of the expression (6.8) with respect to $\hat{a}(v)$ and $\hat{\beta}(v)$.

Theorem 6.3

Let $\bar{\beta}$,m and M be numbers such that

$$\begin{aligned} &-2\sqrt{\bar{\beta}}\leq m\leq m(B), \\ &M(B)\leq M\leq \min(1,2\sqrt{\bar{\beta}}) \\ \text{and}\quad &S(LU)\leq\bar{\beta}. \end{aligned} \tag{6.25}$$

Then, a lower bound on $\lambda(B_\omega)$ is given by

$$\lambda(B_\omega) \geq \begin{cases} \dfrac{1-M}{1-\omega M+\omega^2\bar{\beta}} = \phi_1(\omega), \text{if } \bar{\beta}\geq\frac{1}{4} \text{ or if } \bar{\beta}\leq\frac{1}{4} \text{ and } \omega\leq\omega^* \\ \dfrac{1-m}{1-\omega m+\omega^2\bar{\beta}} = \phi_2(\omega), \text{if } \bar{\beta}<\frac{1}{4} \text{ and } \omega>\omega^*. \end{cases} \tag{6.26}$$

where for $\bar{\beta}<\frac{1}{4}$ we define ω^* by

$$\omega^* = \frac{2}{1+\sqrt{1-4\bar{\beta}}} . \tag{6.27}$$

Proof

Let us consider the eigenvalue λ of B_ω given by (6.8) as a function of the variables $\omega,\hat{a},\hat{\beta}$, then we will attempt to find a lower bound of this expression by studying its behaviour with respect to $\hat{a},\hat{\beta}$, hence we have

the following problem to solve

$$\lambda(B_\omega) \geq \min_{\hat{a},\hat{\beta}} \lambda(\omega,\hat{a},\hat{\beta}) = \min_{\hat{a},\hat{\beta}} \left\{ \frac{1-\hat{a}}{1-\omega\hat{a}+\omega^2\hat{\beta}} \right\} . \tag{6.28}$$

It can be easily verified, for fixed ω and $\hat{a}$, that

$$\operatorname{sign}\left[\frac{\partial}{\partial\hat{\beta}} \lambda(\omega,\hat{a},\hat{\beta}) \right] < 0 \tag{6.29}$$

since $\omega>0$ and $\hat{a}<1$.

Thus $\lambda(\omega,\hat{a},\hat{\beta})$ is a decreasing function of $\hat{\beta}$ and from (6.25),(6.28) and (6.12) it follows that our problem reduces to the following

$$\lambda(B_\omega) \geq \min_{\hat{a}} \lambda(\omega,\hat{a},\bar{\beta}). \tag{6.30}$$

Further, in this case for fixed ω, we have

$$\operatorname{sign}\left[\frac{\partial}{\partial\hat{a}} \lambda(\omega,\hat{a},\bar{\beta}) \right] = \operatorname{sign}(-\omega^2\bar{\beta}+\omega-1) \tag{6.31}$$

hence we can easily construct Table 6.1 which verifies (6.26). Thus the proof of the theorem is now complete.

$\bar{\beta}$-Domain	ω-Domain	$\omega^2\bar{\beta}-\omega-1$	$\lambda(B_\omega)$ Bound
$\bar{\beta}\geq\frac{1}{4}$	$0<\omega<2$	≤ 0	$\phi_1(\omega)=\lambda(\omega,M,\bar{\beta})$
$0\leq\bar{\beta}<\frac{1}{4}$	$0<\omega<\omega^*$	>0	$\phi_1(\omega)=\lambda(\omega,M,\bar{\beta})$
	$\omega=\omega^*$	$=0$	$\phi_1(\omega)=\lambda(\omega,\hat{a},\bar{\beta})$
	$\omega^*<\omega<2$	<0	$\phi_2(\omega)=\lambda(\omega,m,\bar{\beta})$

TABLE 6.1

BEHAVIOUR OF $\lambda(\omega,\hat{a},\bar{\beta})$ AS A FUNCTION OF $\hat{a}$

4.7 DETERMINATION OF $S(\mathcal{H}_{\omega_1})$ and ω_1

From the analysis of the previous section and (6.1) we see that

$$S(\mathcal{H}_\omega) = \max\{\nu_{\max}(\mathcal{H}_\omega), \ |\nu_{\min}(\mathcal{H}_\omega)|\} \tag{7.1}$$

where $\nu_{\max}(\mathcal{H}_\omega)$, $\nu_{\min}(\mathcal{H}_\omega)$ are the maximum and minimum eigenvalues of $\mathcal{H}_\omega$, respectively. It is evident that

$$\begin{aligned} \nu_{\max}(\mathcal{H}_\omega) &= 1-\lambda(\mathcal{B}_\omega) \\ \text{and} \qquad |\nu_{\min}(\mathcal{H}_\omega)| &= \frac{1}{\omega(2-\omega)} - 1. \end{aligned} \tag{7.2}$$

Let us first examine the behaviour of

$$\nu(\omega,\hat{a},\bar{\beta}) = 1-\lambda(\omega,\hat{a},\bar{\beta}) \tag{7.3}$$

with respect to ω in the range $(0,2)$.

Using the notation of the previous section, we have

$$\operatorname{sign}\left\{\frac{\partial}{\partial\omega}\nu(\omega,\hat{a},\bar{\beta})\right\} = \operatorname{sign}(2\omega\bar{\beta}-\hat{a}), \tag{7.4}$$

which simply means that we have to examine the behaviour of the function $\gamma(\omega,\hat{a},\bar{\beta})=2\omega\bar{\beta}-\hat{a}$. This is summarised in the following table.

ω-Domain	$\gamma(\omega,\hat{a},\bar{\beta})$	$\nu(\omega,\hat{a},\bar{\beta})$
$\omega > \frac{\hat{a}}{2\bar{\beta}}$	>0	Increasing
$\omega = \frac{\hat{a}}{2\bar{\beta}}$	$=0$	Stationary
$\omega < \frac{\hat{a}}{2\bar{\beta}}$	<0	Decreasing

TABLE 7.1

BEHAVIOUR OF $\nu(\omega,\hat{a},\bar{\beta})$ AS A FUNCTION OF ω

Using Table 7.1 we can determine the behaviour of the functions

$$\begin{aligned} \theta_1(\omega) &= \nu(\omega,M,\bar{\beta}), \\ \text{and} \qquad \theta_2(\omega) &= \nu(\omega,m,\bar{\beta}) \end{aligned} \tag{7.5}$$

with respect to ω. Hence we have that

$$\omega_M = \frac{M}{2\bar{\beta}} \tag{7.6}$$

is the critical point of $\theta_1(\omega)$. Since now $\gamma(\omega,m,\bar{\beta})>0$ for any non-negative

value of ω, $\theta_2(\omega)$ is an increasing function in the interval (0,2). Finally, from the previous analysis Table 7.2 and Figure 7.1 can be established.

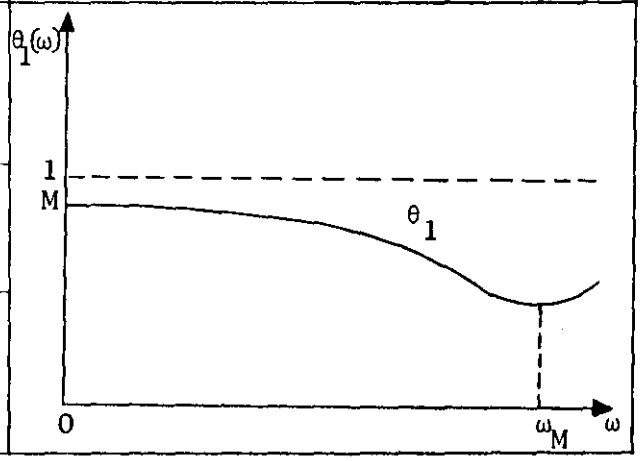
ω-Domain	$\gamma(\omega,M,\bar{\beta})$	$\theta_1(\omega)$	Graph
$0<\omega<\omega_M$	<0	Decreasing	
$\omega=\omega_M$	$=0$	Stationary	
$\omega_M<\omega$	>0	Increasing	

TABLE 7.2

BEHAVIOUR OF $\theta_1(\omega)$ AS A FUNCTION OF ω

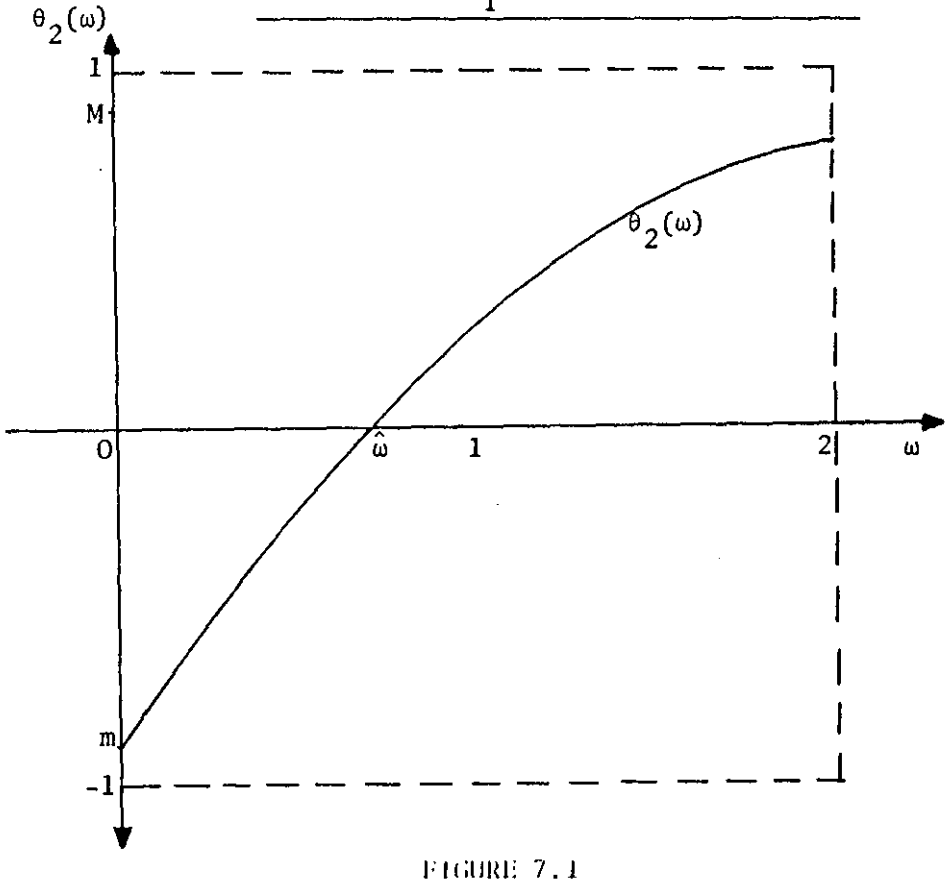


FIGURE 7.1

BEHAVIOUR OF $\theta_2(\omega)$ AS A FUNCTION OF ω

We can easily verify that $\theta_2(\omega)=0$ or $g(\omega)=\omega^2\bar{\beta}-\omega m+m=0$ when $\omega=\hat{\omega}=2/(1-\sqrt{1-\frac{4\bar{\beta}}{m}})$, also since $\text{sign}(g(0)g(1))\overset{<0}{}$ and $\frac{m}{\bar{\beta}}<0$, we have $0<\hat{\omega}<1$.

From the previous analysis, we can summarise our results by considering the following cases,

$$\begin{aligned}&\text{Case I} &&: \bar{\beta}\geq\tfrac{1}{4}\ ,\ 0<\omega<2, \theta_1(\omega)\geq\theta_2(\omega)\\ &\text{Case II} &&: 0\leq\bar{\beta}<\tfrac{1}{4},\ 0<\omega\leq\omega^*, \theta_1(\omega)\geq\theta_2(\omega)\\ &\text{and Case III} &&: 0\leq\bar{\beta}<\tfrac{1}{4},\ 0<\omega^*\leq\omega, \theta_1(\omega)\leq\theta_2(\omega)\end{aligned} \tag{7.7}$$

Moreover the relation between $\theta_1(\omega)$ and $\theta_2(\omega)$ can be easily seen from Table 7.3

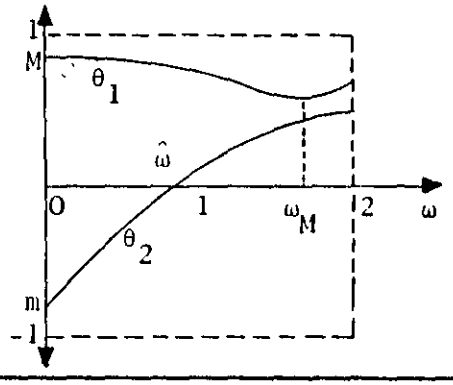
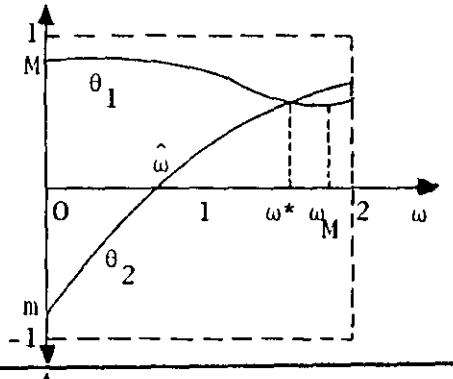
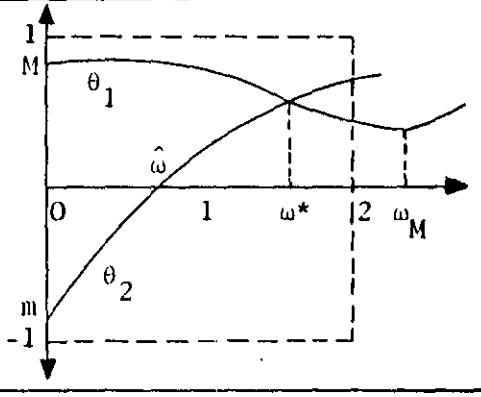
$\bar{\beta}$-Domain	ω-Domain	$\nu_{\max}(\mathcal{H}_\omega)$ Bound	Graph
$\bar{\beta}\geq\frac{1}{4}$ $M\leq 4\bar{\beta}$	$0<\omega<2$	θ_1	
$0\leq\bar{\beta}<\frac{1}{4}$ $M\leq 4\bar{\beta}$	$0<\omega\leq\omega^*$	θ_1	
$0\leq\bar{\beta}<\frac{1}{4}$ $M\geq 4\bar{\beta}$	$\omega^*\leq\omega<2$	θ_2	

TABLE 7.3

RELATION BETWEEN $\theta_1(\omega)$ AND $\theta_2(\omega)$

Up to now we have studied the behaviour of $\nu_{max}(\mathcal{H}_\omega)$, but in order to determine $S(\mathcal{H}_\omega)$ we have to find the relation of $\nu_{max}(\mathcal{H}_\omega)$ and $|\nu_{min}(\mathcal{H}_\omega)|$. This is shown in Figure 7.2.

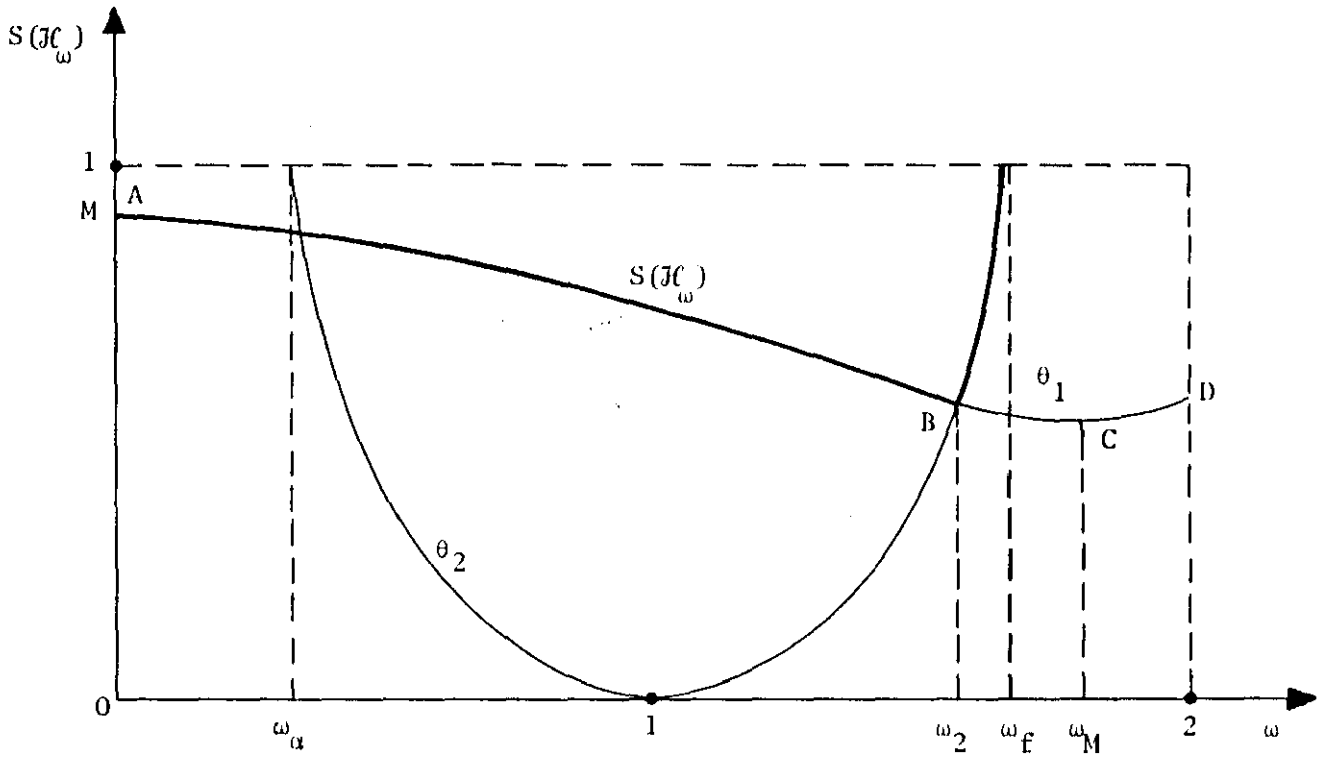


FIGURE 7.2 (case I)

BEHAVIOUR OF $S(\mathcal{H}_\omega)$ AS A FUNCTION OF ω

If we consider Case I, then from Table 7.3 it is clear that $\theta_1(\omega)$ dominates over $\theta_2(\omega)$ for all the values of $\omega \in (0,2)$. In this case we can determine a good value of ω which minimises the bound on $S(\mathcal{H}_\omega)$ from the equation

$$\nu_{max}(\mathcal{H}_\omega) = |\nu_{min}(\mathcal{H}_\omega)|. \tag{7.8}$$

This equation by using (7.2) and (7.5) can be written to yield

$$1 - \frac{1-M}{1-\omega M+\omega^2\bar{\beta}} = \frac{1}{\omega(2-\omega)} - 1 \tag{7.9}$$

or equivalently

$$p(\omega) = 2\bar{\beta}\omega^4 - 2(M+2\bar{\beta})\omega^3 + (1+5M+\bar{\beta})\omega^2 - (3M+2)\omega + 1 = 0. \tag{7.10}$$

By Déscartes rule we can find that equation (7.10) has either i) no positive roots or ii) at least two positive roots. Since now it can be easily verified that $sign(p(0)p(1)<0$ and $sign(p(1)p(2))<0$ we conclude that equation (7.10) has two positive roots ω_1,ω_2 such that $0<\omega_1<1$ and $1<\omega_2<2$. Clearly, from the above analysis we have justified Figure 7.2, also we have shown the existence of a unique value of the preconditioning parameter $\omega=\omega_2$ which minimises $S(\mathcal{H}_\omega)$ and lies in the interval (1,2). This value can be determined by using known methods (i.e. Newton Raphson, Bairstow) to solve numerically (7.10) and consequently to obtain the corresponding bound on $S(\mathcal{H}_\omega)$ from

$$S(\mathcal{H}_{\omega_2}) = \frac{1}{\omega_2(2-\omega_2)} - 1. \tag{7.11}$$

In the remainder of the cases, we have to distinguish whether i) $\omega_2<\omega^*$ or ii) $\omega^*<\omega_2$. Since this cannot be done unless we solve (7.10), we impose the restriction that if

$$\omega_f \lesssim \omega^* \tag{7.12}$$

or from (6.27) if

$$\bar{\beta} \gtrsim \frac{\omega_f - 1}{\omega_f^2} = 0.2426, \tag{7.13}$$

then ω_2 is a good choice of ω.

Summarising our results we have that

$$\omega_1 = \begin{cases} \omega_2 \text{ if } \bar{\beta}\gtrsim 0.2426 \text{ or if } \omega_2<\omega^* \\ \omega^* \text{ if } \omega^*\lesssim\omega_2 \end{cases} \tag{7.14}$$

whereas the corresponding bound on $S(\mathcal{H}_\omega)$ is given by

$$S(\mathcal{H}_{\omega_1}) \lesssim \begin{cases} \frac{1}{\omega_2(2-\omega_2)} - 1, & \text{if } \bar{\beta}\gtrsim 0.2426 \text{ or if } \omega_2<\omega^* \\ \frac{\omega^*-1}{\omega^*}, & \text{if } \bar{\beta}<0.2426 \text{ and if } \omega^*\lesssim\omega_2. \end{cases} \tag{7.15}$$

By a simple comparison of $S(\mathcal{L}_{\omega_1})$ (see (3-6.49)) and $S(\mathcal{H}_{\omega_1})$ we observe that under certain conditions the PJ method may attain slightly better rate of convergence than the SSOR method. However, this is a limited case and

it is expected to happen for large values of the mesh size h (small values of N). Finally, Figure 7.3 illustrates the behaviour of $S(\mathcal{H}_\omega)$ when the restrictions of the second part of (7.15) hold.

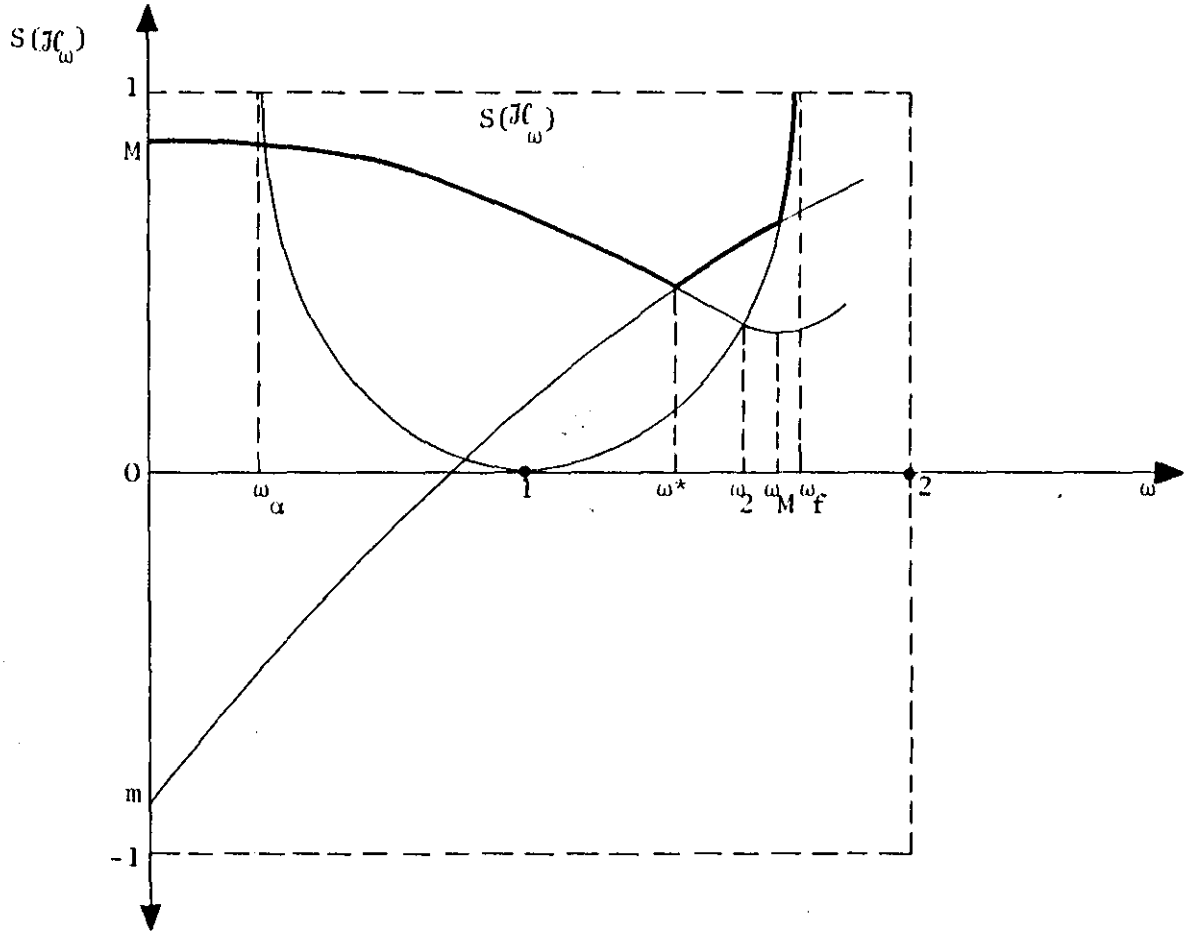


FIGURE 7.3

BEHAVIOUR OF $S(\mathcal{H}_\omega)$ AS A FUNCTION OF ω WHEN $\bar{\beta}<0.2426$ AND $\omega^*<\omega_2$

4.8 COMPUTATIONAL RESULTS

First, it was observed that the estimated upper bound on $\Lambda(B_\omega)$ which is given by (6.20) is a very good approximation for $1 \lesssim \omega < 2$. This can be easily seen if one carries out a comparison of this bound and $\Lambda(B_\omega)$ determined by the power method for the above range of ω. However, this does not seem to be the case, especially when ω is very close to zero. As a result of this we have Table 8.1 which shows the behaviour of $S(\mathcal{H}_\omega)$ for $\omega \in [0,2]$ in the Laplace problem. From this table we observe that $S(\mathcal{H}_\omega)$ is less than unity for $\omega = 0.0,\ 0.1,\ 0.2$ which is not expected because of Theorem 5.1. On the other hand, $S(\mathcal{H}_\omega)$ is approximated satisfactorily by $\nu_{max}(\mathcal{H}_\omega)$ given by (7.2) for $0 \lesssim \omega \lesssim \omega_1$ which is consistent with the observation that for $\omega = 0$ the PJ method coincides with the J method. Consequently the interval of the preconditioning parameter ω such as the PJ method to be convergent is the following

$$0 \lesssim \omega < 1 + \frac{\sqrt{2}}{2} \ . \tag{8.1}$$

Moreover, this justifies the way we indicated the behaviour of $S(\mathcal{H}_\omega)$ in Figures 7.2 and 7.3

ω \ h^{-1}	10	20	40
0.0	0.9510	0.9876	0.9965
0.1	0.9460	0.9863	0.9962
0.2	0.9402	0.9848	0.9959
0.3	0.9335	0.9830	0.9955
0.4	0.9255	0.9808	0.9949
0.5	0.9161	0.9783	0.9943
0.6	0.9048	0.9751	0.9935
0.7	0.8912	0.9713	0.9926
0.8	0.8745	0.9665	0.9913
0.9	0.8539	0.9604	0.9898
1.0	0.8282	0.9524	0.9877
1.1	0.7955	0.9419	0.9849
1.2	0.7537	0.9275	0.9810
1.3	0.6998	0.9273	0.9753
1.4	0.6307	0.8778	0.9667
1.5	0.5452	0.8327	0.9528
1.6	(-)0.5621	0.7610	0.9281
1.7	(-)0.9518	(-)0.9601	(-)0.9606
1.8	(-)1.6638	(-)1.7758	(-)1.7769
1.9	(-)3.0395	(-)4.0581	(-)4.2586
2.0	(-)5.8515	(-)13.9677	(-)30.9734

TABLE 8.1

BEHAVIOUR OF $S(\mathcal{H}_\omega)$ FOR THE LAPLACE EQUATION

As can be seen from Table 8.1, as $h\to 0$ the value ω_1 tends to ω_f, whereas the minus signs in parentheses indicate that $S(\mathcal{H}_\omega)$ is represented by $|\nu_{min}(\mathcal{H}_\omega)|$. Also we note that for large h (i.e. $h^{-1}=10$), $S(\mathcal{H}_\omega)$ is insensitive for $\omega\in[0,1]$ and then as ω increases from 1 to 2, $S(\mathcal{H}_\omega)$ decreases slightly more rapidly until ω gets close to $\omega_{opt}=1.5$ the optimum value at which point the decrease is very rapid. As ω increases further, $S(\mathcal{H}_\omega)$ increases faster to a value of unity when $\omega=\omega_f$ and then as ω approaches 2, $S(\mathcal{H}_\omega)$ takes values greater than unity which can become very large indeed. In the case where $h\to 0$ (i.e. $h^{-1}=40$) we observe that $S(\mathcal{H}_\omega)$ is very insensitive for $\omega\in[0,\omega_f]$ and then as $\omega\to 2$, increases very rapidly. Of course, this behaviour was expected, since as $h\to 0$ then $M\to 1^-$ and from our theoretical analysis we can observe (see Figure 7.2) that the part AB moves upwards and the point C moves to the right. Hence, in general we expect the PJ method to have a very slow rate of convergence.

In order to test the theoretical results obtained above a number of numerical experiments were carried out involving the generalised Dirichlet problem on the unit square with the differential equation

$$\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(C\frac{\partial u}{\partial y}\right) = 0. \tag{8.2}$$

Various choices of the coefficients $A(x,y)$ and $C(x,y)$ were used, as indicated in Table 8.2. The optimum preconditioning parameters ω_0 and $S(\mathcal{H}_{\omega_0})$ were determined as follows. The spectral radius $S(\mathcal{H}_\omega)$ was calculated by using the power method which is given by

$$\left.\begin{aligned} w^{(n+1)} &= \mathcal{H}_\omega w^{(n)} \\ \nu^{(n)} &= \frac{(w^{(n)},w^{(n+1)})}{(w^{(n)},w^{(n)})} \quad , \; n\geq 0 \end{aligned}\right\} \tag{8.3}$$

for $w^{(0)}\neq 0$. It is known that $w^{(n)}$ is an approximation to the normalised eigenvector associated with $S(\mathcal{H}_\omega)$ and that $\nu^{(n)}$ converges to $S(\mathcal{H}_\omega)$ as n tends to infinity (see e.g. Gourlay and Watson [1973]).

Furthermore, we assume that $w^{(0)}$ has a non-zero component in the direction of the dominant eigenvector. Then we can apply the Fibonacci

search technique (e.g. see Zahradnik [1971]) to obtain ω_0 and $S(\mathcal{H}_{\omega_0})$. The PJ iterative scheme was then applied with boundary values taken to be zero on all sides of the square. As starting vector $u^{(0)}$ was used the vector with all its components equal to unity in each case and the procedure was terminated when the inequality $\|u^{(n)}\|_\infty < 10^{-6}$ was satisfied. The number of iterations of the numerical experiments together with the optimum values ω_0 and $S(\mathcal{H}_{\omega_0})$ are presented in Table 8.2.

Problem	Coefficients	h^{-1}	ω_0	$S(\mathcal{H}_{\omega_0})$	n_{PJ}
1	$A=C=1$	20	1.6456	0.7147	43
		40	1.6859	0.8883	121
		60	1.6967	0.9435	247
2	$A=C=e^{10(x+y)}$	20	1.5370	0.4511	20
		40	1.6439	0.7082	49
		60	1.6555	0.8496	105
3	$A=(1+2x^2+y^2)^{-1}$	20	1.6471	0.7204	44
	$C=(1+x^2+2y^2)^{-1}$	40	1.6865	0.8913	124
		60	1.6970	0.9453	254
4	$A=C=\begin{cases}1+x, & 0\leq x\leq\frac{1}{2}\\ 2-x, & \frac{1}{2}<x\leq 1\end{cases}$	20	1.6453	0.7134	43
		40	1.6858	0.8878	120
		60	1.6967	0.9433	245
5	$A=1+4\left\|x-\frac{1}{2}\right\|^2$	20	1.6483	0.7252	44
	$C=\begin{cases}1, & 0\leq x<\frac{1}{2}\\ 9, & \frac{1}{2}\leq x\leq 1\end{cases}$	40	1.6857	0.8876	119
		60	1.6964	0.9420	239
6	$A=1+\sin\frac{\pi(x+y)}{2}$	20	1.5499	0.4703	21
	$C=e^{10(x+y)}$	40	1.6466	0.7187	48
		60	1.6722	0.8397	90

TABLE 8.2

OPTIMUM PARAMETERS ω_0 AND $S(\mathcal{H}_{\omega_0})$

4.9 THE PRECONDITIONED SIMULTANEOUS DISPLACEMENT METHOD (PSD METHOD)

As was seen in the previous sections, the PJ method has a less favourable rate of convergence although it requires approximately twice as much work as the SOR. Again as it will be shown later this is due to the fact that we did not let the parameter τ in (4.5) take its optimum value. Let us therefore consider the iterative scheme

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{9.1}$$

as defined by (4.5) where τ,ω are real parameters.

The iterative process (9.1) is the Simultaneous Displacement version of (4.6) and it will be referred to as the Preconditioned Simultaneous Displacement method (PSD method). At this point we are able to state that if we let τ and ω take their optimum values τ_0 and ω_0, respectively then the PSD method will be superior to any other iterative scheme which uses the same conditioning matrix R given by (4.1). We therefore expect that the PSD method will produce in general a better rate of convergence than the PJ and SSOR methods. Further, we also expect the iterative process (9.1) to meet our expectations as regards our earlier conjecture, (see Section 4.4) to produce a rate of convergence which in some cases might be better than SOR. Evidently, (9.1) can be written in a more "computable" form similar to (4.7) as

$$\left.\begin{aligned} \zeta^{(n+\frac{1}{2})} &= \omega L\zeta^{(n+\frac{1}{2})} + r^{(n)} \\ \zeta^{(n+1)} &= \omega U\zeta^{(n+1)} + \zeta^{(n+\frac{1}{2})} \\ u^{(n+1)} &= u^{(n)} + \tau\zeta^{(n+1)} \end{aligned}\right\} \tag{9.2}$$

where

$$r^{(n)} = D^{-1}(b-Au^{(n)}). \tag{9.3}$$

From (9.1) we also have

$$u^{(n+1)} = D_{\tau,\omega}u^{(n)} + \delta \tag{9.4}$$

where

$$D_{\tau,\omega} = I-\tau(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \tag{9.5}$$

and

$$\delta = \tau(I-\omega U)^{-1}(I-\omega L)^{-1}c. \tag{9.6}$$

Moreover, $I-D_{\tau,\omega}$ is non-singular if

$$\tau\neq 0 \tag{9.7}$$

and if A is non-singular.

4.10 CONVERGENCE OF THE PSD METHOD

Theorem 10.1

Let A be a symmetric matrix with positive diagonal elements. For any real ω and τ the eigenvalues of $D_{\tau,\omega}$ are real. Moreover,

$$S(D_{\tau,\omega})<1 \tag{10.1}$$

if and only if

$$0<\omega<2, \tag{10.2}$$

$$0<\tau<2\omega(2-\omega) \tag{10.3}$$

and A is positive definite.

Proof

From (9.5) and since $D^{\frac{1}{2}}$ exists we have that $D_{\tau,\omega}$ is similar to the matrix

$$\begin{aligned}\bar{D}_{\tau,\omega} &= D^{-\frac{1}{2}}(D-\omega C_U)D_{\tau,\omega}(D-\omega C_U)^{-1}D^{\frac{1}{2}} \\ &= I-\tau D^{\frac{1}{2}}(D-\omega C_L)^{-1}A(D-\omega C_U)^{-1}D^{\frac{1}{2}} \\ &= I-\tau[D^{\frac{1}{2}}(D-\omega C_L)^{-1}]A[D^{\frac{1}{2}}(D-\omega C_L)^{-1}]^T\end{aligned} \tag{10.4}$$

If A is a symmetric matrix, then the second matrix of the right hand side of (10.4) is symmetric as well. In addition, if A is positive definite, then by (5.2) we have

$$\bar{D}_{\tau,\omega} = I-\tau\bar{B}_{\omega} \tag{10.5}$$

where $\bar{B}_{\omega}$ is positive definite and similar to B_{ω}. As was proved in Section 4.5, B_{ω} is positive definite if and only if A is positive definite. Furthermore, if d and λ are the eigenvalues of $D_{\tau,\omega}$ and B_{ω}, respectively, then from (9.5) we have the following eigenvalue relationship

$$d = 1-\tau\lambda \tag{10.6}$$

where $\lambda>0$.

Evidently, $S(D_{\tau,\omega})<1$ if and only if

$$-1<1-\tau\lambda<1 \tag{10.7}$$

which by Theorem 6.2 gives (10.2) and (10.3) thus completing the proof of the theorem.

4.11 CHOICE OF τ AND ω FOR THE PSD METHOD

We now study the problem of determining good estimates for τ, the preconditioning parameter ω and the spectral radius of $D_{\tau,\omega}$. Our primary concern is the case where A does not possess the form (2-7.1).

Theorem 11.1

Let $\bar{\beta}$, m and M be numbers such that

$$-2\sqrt{\bar{\beta}} \leq m \leq m(B)$$
$$M(B) \leq M \leq \min(1, 2\sqrt{\bar{\beta}}) \tag{11.1}$$

and $$S(LU) \leq \bar{\beta}.$$

Then, a lower bound on the P-condition number of B_ω, $P(B_\omega)$ is given by

$$P(B_\omega) \gtrsim \begin{cases} \dfrac{1-\omega M+\omega^2\bar{\beta}}{\omega(2-\omega)(1-M)} = \phi_1(\omega), & \text{if } \bar{\beta} \geq \dfrac{1}{4} \text{ or if } \bar{\beta} < \dfrac{1}{4} \text{ and } \omega \leq \omega^* \\ \dfrac{1-\omega m+\omega^2\bar{\beta}}{\omega(2-\omega)(1-m)} = \phi_2(\omega), & \text{if } \bar{\beta} < \dfrac{1}{4} \text{ and } \omega > \omega^* \end{cases} \tag{11.2}$$

where for $\bar{\beta} < \frac{1}{4}$ we define ω^* by

$$\omega^* = \frac{2}{1+\sqrt{1-4\bar{\beta}}}. \tag{11.3}$$

Moreover, the bound (11.2) is minimised if we let

$$\omega_1 = \begin{cases} \dfrac{2}{1+\sqrt{1-2M+4\bar{\beta}}} = \omega_M, & \text{if } M \leq 4\bar{\beta} \\ \dfrac{2}{1+\sqrt{1-4\bar{\beta}}} = \omega^*, & \text{if } M \geq 4\bar{\beta} \end{cases} \tag{11.4}$$

and the corresponding value of $P(B_{\omega_1})$ is given by

$$P(B_{\omega_1}) \lesssim \begin{cases} \dfrac{1}{2}\left(1+\dfrac{\sqrt{1-2M+4\bar{\beta}}}{1-M}\right) = \dfrac{1}{2}\left(\dfrac{2-\omega_M M}{(1-M)\omega_M}\right), & \text{if } M \leq 4\bar{\beta} \\ \dfrac{1+\sqrt{1-4\bar{\beta}}}{2\sqrt{1-4\bar{\beta}}} = \dfrac{1}{2-\omega^*}, & \text{if } M \geq 4\bar{\beta}. \end{cases} \tag{11.5}$$

Proof

The validity of (11.2) can be easily seen by Theorems 6.2 and 6.3. Thus, we will be concerned with the behaviour of the bound (11.2) as a function of ω.

By letting

$$p(\omega,\hat{a},\hat{\beta}) = \frac{1}{\omega(2-\omega)\lambda(\omega,\hat{a},\hat{\beta})} \tag{11.6}$$

then from (6.28) we have

$$p(\omega,\hat{a},\hat{\beta}) = \frac{1-\omega\hat{a}+\omega^2\hat{\beta}}{\omega(2-\omega)(1-\hat{a})} \tag{11.7}$$

hence

$$p(\omega,M,\bar{\beta}) = \phi_1(\omega) \quad \text{and} \quad p(\omega,m,\bar{\beta}) = \phi_2(\omega) \tag{11.8}$$

From (11.7) it follows that

$$\text{sign}\left(\frac{\partial}{\partial\omega}p(\omega,\hat{a},\hat{\beta})\right) = \text{sign}(\omega^2(2\hat{\beta}-\hat{a})+2\omega-2) \tag{11.9}$$

and therefore the critical points of $\phi_1(\omega)$ and $\phi_2(\omega)$ in the interval $(0,2)$ are

$$\omega_M = \frac{2}{1+\sqrt{1-2M+4\bar{\beta}}} \tag{11.10}$$

and

$$\omega_m = \frac{2}{1+\sqrt{1-2m+4\bar{\beta}}} \tag{11.11}$$

respectively.

We can therefore establish Tables 11.1 and 11.2.

Domain	$p(\omega,M,\bar{\beta})$	$\phi_1(\omega)$
$0<\omega<\omega_M$	>0	Decreasing
$\omega=\omega_M$	$=0$	Stationary
$\omega_M<\omega<2$	<0	Increasing

TABLE 11.1

BEHAVIOUR OF $\phi_1=p(\omega,M,\bar{\beta})$ AS A FUNCTION OF ω

Domain	$p(\omega,m,\bar{\beta})$	$\phi_2(\omega)$
$0<\omega<\omega_m$	>0	Decreasing
$\omega=\omega_m$	$=0$	Stationary
$\omega_m<\omega<2$	<0	Increasing

TABLE 11.2

BEHAVIOUR OF $\phi_2(\omega)=p(\omega,m,\bar{\beta})$ AS A FUNCTION OF ω

It can be easily seen that

$$\omega_m \leq \omega_M \tag{11.12}$$

since m<M. Moreover, if $\bar{\beta} \leq \frac{1}{4}$ then $\omega^* \geq 1$ and from (11.11) we see that $\omega_m \leq 1$. Thus, we obtain

$$\omega_m < \omega^*. \tag{11.13}$$

Furthermore, we have

$$\text{sign}\left(\frac{\partial}{\partial \hat{a}} p(\omega, \hat{a}, \bar{\beta})\right) = \text{sign}(\omega^2 \bar{\beta} - \omega + 1) \tag{11.14}$$

and therefore we construct Table 11.3.

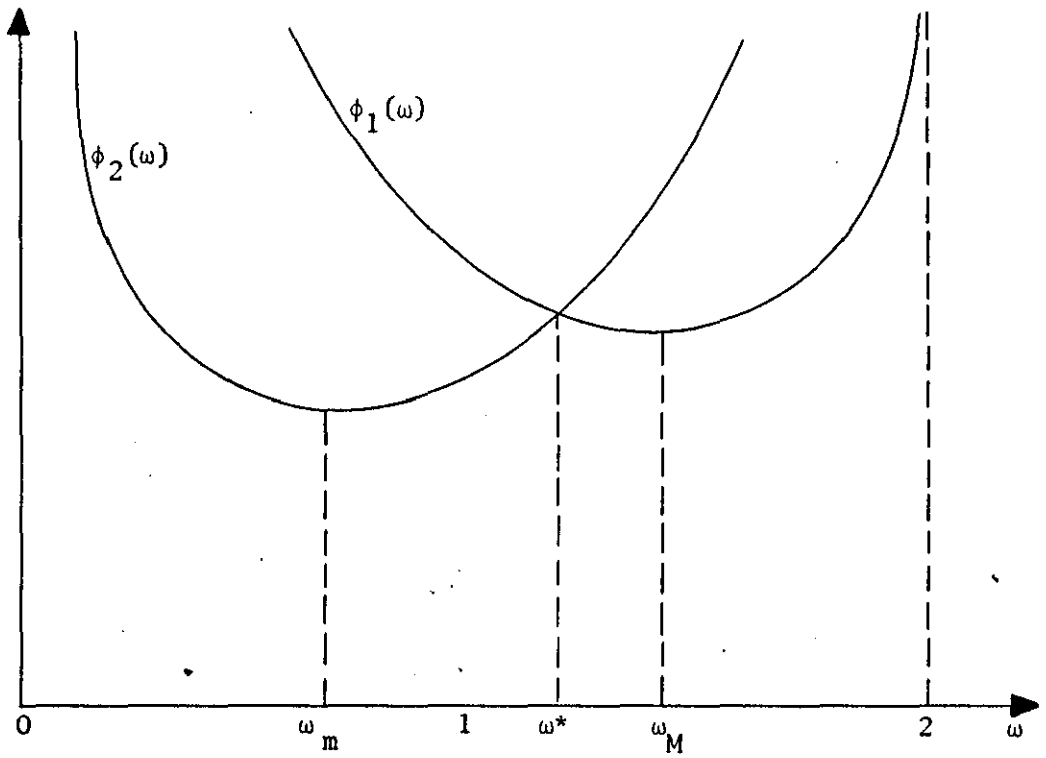
$\omega^2\bar{\beta}-\omega+1$	Relation	Bound on $P(B_\omega)$
≥ 0	$\phi_1(\omega) \geq \phi_2(\omega)$	$\phi_1(\omega)$
$=0$	$\phi_1(\omega) = \phi_2(\omega)$	$\phi_1(\omega)$ or $\phi_2(\omega)$
≤ 0	$\phi_1(\omega) \leq \phi_2(\omega)$	$\phi_2(\omega)$

TABLE 11.3

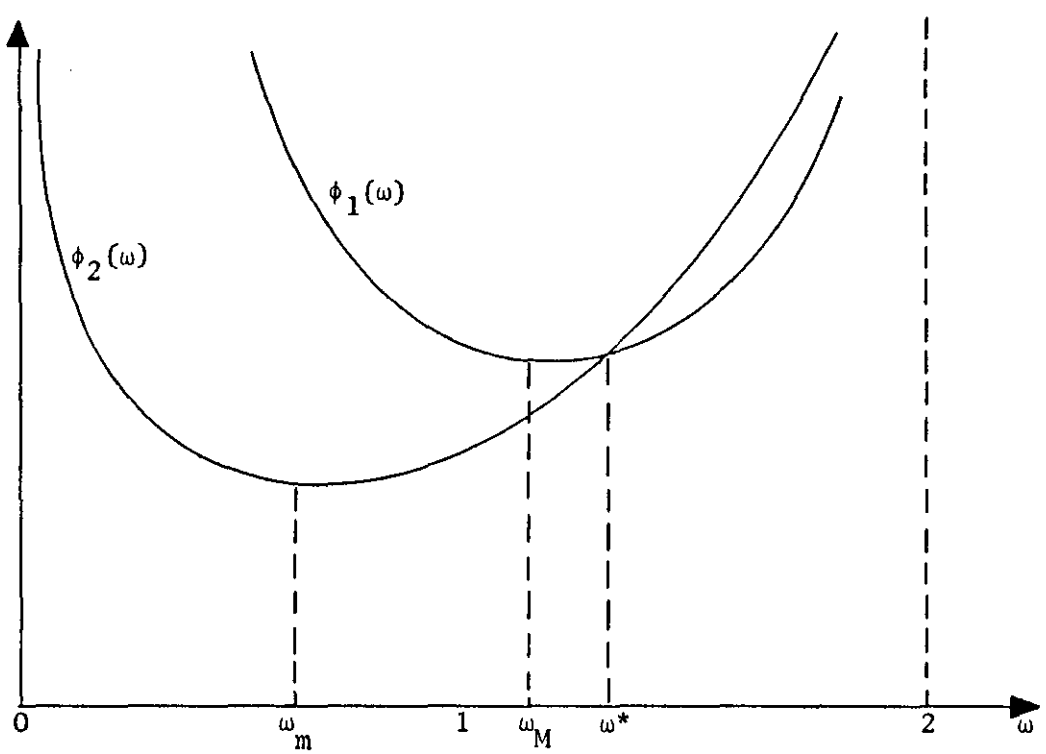
RELATION BETWEEN $\phi_1(\omega)$ and $\phi_2(\omega)$

Evidently, a similar table to Table 6.1 holds for $p(\omega, \hat{a}, \bar{\beta})$ as well, hence by combining also the properties of Table 11.3 we clearly see that for $\bar{\beta} \geq \frac{1}{4}$, then $\phi_1(\omega) \geq \phi_2(\omega)$ for all ω, hence from Table 11.1, $\omega_1 = \omega_M$. If $\bar{\beta} \leq \frac{1}{4}$, then we consider two cases:

Case I: $\omega_m \leq \omega^* \leq \omega_M$ and Case II: $\omega_m \leq \omega_M \leq \omega^*$ (see Figure 11.1). Clearly in Case I $\omega_1 = \omega^*$ while $\omega_1 = \omega_M$ in Case II. If now $\bar{\beta} \geq \frac{1}{4}$ and M<1, then these conditions imply $M \leq 4\bar{\beta}$. Also, in Case II we have $\bar{\beta} < \frac{1}{4}$ and $\omega_M \leq \omega^*$ which imply $M \leq 4\bar{\beta}$ as well. Finally, in Case I, $\bar{\beta} < \frac{1}{4}$ and $\omega_M \geq \omega^*$ thus $M \geq 4\bar{\beta}$. Thus a "good" estimation of ω namely ω_1 is given by (11.4) whereas the corresponding bounds on $P(B_{\omega_1})$ are found by direct substitution in (11.2), hence the proof of the theorem is complete.



Case I: $\omega_m \leq \omega^* \leq \omega_M$



Case II: $\omega_m \leq \omega_M \leq \omega^*$

FIGURE 11.1

It should be noted that the approach for determining the estimated values ω_1 and $P(B_{\omega_1})$ is similar to the one followed by Young [1974] and Axelsson [1974] for obtaining good bounds on $S(\mathfrak{G}_\omega)$.

Moreover, from Theorem 3-6.14 we conclude that the two estimates ω_1^{SSOR} and ω_1^{PSD} are identical (see (3-6.48) and (11.4)). This was expected since SSOR and PSD can be easily seen that they both possess identical P-condition numbers, thus

$$P(B_\omega) = \frac{1-\lambda(\mathfrak{G}_\omega)}{1-\Lambda(\mathfrak{G}_\omega)} \tag{11.15}$$

where $\lambda(\mathfrak{G}_\omega)$ and $\Lambda(\mathfrak{G}_\omega)$ are the minimum and maximum eigenvalues of $\mathfrak{G}_\omega$. But $\lambda(\mathfrak{G}_\omega)$ is approximated by zero under the restriction that A is positive definite, hence from (11.15) it follows

$$P(B_\omega) \sim \frac{1}{1-S(\mathfrak{G}_\omega)} \tag{11.16}$$

which implies that the optimum values $\omega_0^{PSD}, \omega_0^{SSOR}$ of $P(B_\omega)$ and $S(\mathfrak{G}_\omega)$, respectively are very close, whereas the estimated ω_1^{PSD} and ω_1^{SSOR} are identical.

From (11.5) we can modify the bound on $P(B_{\omega_1})$ to yield

$$P(B_{\omega_1}) \lesssim \begin{cases} \frac{1}{2}\left(1+\frac{1}{\sqrt{1-M}}\right), & \text{if } \bar{\beta} \lesssim \frac{M}{4} \\ \frac{1}{2}\left(1+\sqrt{\frac{2}{1-M}}\right), & \text{if } \frac{M}{4} \lesssim \bar{\beta} \lesssim \frac{1}{4} \\ \frac{1}{2}\left(1+\gamma^{-1}\sqrt{\frac{2}{1-M}}\right), & \text{if } \bar{\beta} \gtrsim \frac{1}{4} \end{cases} \tag{11.17}$$

where

$$\gamma = \left(1 + \frac{2(\bar{\beta}-\frac{1}{4})}{1-M}\right)^{-\frac{1}{2}} \tag{11.18}$$

The first two parts of (11.17) can be easily verified from (11.5) if we recall that $p(\omega,\hat{a},\hat{\beta})$ is an increasing function of $\hat{\beta}$, whereas the last part is obtained if we rewrite the first part of (11.5) to yield successively

$$\frac{\sqrt{1-2M+4\bar{\beta}}}{1-M} = \sqrt{\frac{1-2M+4\bar{\beta}}{(1-M)^2}} = \sqrt{\frac{1-2M+4\bar{\beta}}{2(1-M)}}\sqrt{\frac{2}{1-M}}$$
$$= \gamma^{-1}\sqrt{\frac{2}{1-M}}$$

where

$$\gamma^{-1} = \sqrt{\frac{1-2M+4\bar{\beta}}{2(1-M)}} = \sqrt{1+\frac{2(\bar{\beta}-1/4)}{1-M}} \ .$$

Furthermore the determination of the value $\tau=\tau_1$ can be achieved by using the relationships (6.20) and (11.5) thus we can easily verify that

$$\tau_1 = \frac{2\omega_1(2-\omega_1)}{1+1/P(B_{\omega_1})} \ . \tag{11.19}$$

Finally, from (11.17) we can derive the spectral radius of D_{τ_1,ω_1} since it can be expressed in terms of $P(B_{\omega_1})$.

From (11.19) we note that for $P(B_{\omega_1})>>1$, τ_1 tends to become equal to $2\omega_1(2-\omega_1)$ which according to our previous analysis, implies that the PSD method is expected to produce better rate of convergence than SSOR since in the latter we always have $\tau_1=\omega_1(2-\omega_1)$.

4.12 COMPARISON OF RECIPROCAL RATES OF CONVERGENCE

Let us now compare our bounds on $RR(D_{\tau_1,\omega_1})$ with $RR(B_{\bar{\omega}})$. By using the relationship

$$RR(D_{\tau_1,\omega_1}) \sim \tfrac{1}{2}P(B_{\omega_1}) \tag{12.1}$$

we can easily obtain bounds on $RR(D_{\tau_1,\omega_1})$ from (11.17) for the different cases. Also, by considering (3-6.39) we can obtain bounds on $RR(B_{\bar{\omega}})$ for the general case (when A is a Stieltjes matrix). Consequently, we are now able to construct Table 12.1.

Range of $\bar{\beta}$	Asymptotic Bounds on $RR(D_{\tau_1,\omega_1})/\sqrt{RR(B_{\bar{\omega}})}$	
	General Case	Property A
$\bar{\beta}\leq\frac{M}{4}$	$\frac{1}{2\sqrt{2}}$	$\frac{1}{4}$
$\frac{M}{4}\leq\bar{\beta}\leq\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2\sqrt{2}}$
$\bar{\beta}>\frac{1}{4}$	$\frac{1}{2^{\gamma-1}}$	$\frac{1}{2\sqrt{2}^{\gamma-1}}$

TABLE 12.1

By a simple comparison of Tables 3-6.1 and 12.1 we see that we obtain our main result between the asymptotic bounds on $RR(D_{\tau_1,\omega_1})$ and $RR(\mathcal{L}_{\omega_1})$

$$RR(D_{\tau_1,\omega_1}) \sim \frac{RR(\mathcal{L}_{\omega_1})}{2}. \tag{12.2}$$

Evidently, from (12.2) it follows that the number of iterations of the PSD method is asymptotically half the number of iterations of the SSOR for both methods to achieve the same level of accuracy. (For a comparison of the work involved see Appendix A). This result clearly justifies our earlier conjectures concerned with the superiority of the PSD method over SSOR (see Section 4.9).

Another observation is that the improvement by an order of magnitude of the rate of convergence of PSD over the JOR is retained in the general case as well. Furthermore, by comparing the best possible bound on $RR(L_{\omega_b})$

with $RR(D_{\tau_1,\omega_1})$ in the case where the matrix A is consistently ordered we obtain Table 12.2 (see (3-6.26))

Range of $\bar{\beta}$	Asymptotic Bounds on $RR(D_{\tau_1,\omega_1})/RR(L_{\omega_b})$
$\bar{\beta}\leq\frac{M}{4}$	$\frac{1}{\sqrt{2}}$
$\frac{M}{4}\leq\bar{\beta}\leq\frac{1}{4}$	1
$\bar{\beta}>\frac{1}{4}$	γ^{-1}

TABLE 12.2

PROPERTY A

From Table 12.2 if we compare the PSD with the SOR (without taking into account the computational work involved) then for $\bar{\beta}\leq\frac{M}{4}$ we have an improvement of approximately $\sqrt{2}$ of the rate of convergence of the former over the latter method, whereas for $\frac{M}{4}\leq\beta\leq\frac{1}{4}$ we expect to obtain asymptotically identical results. However for $\bar{\beta}>\frac{1}{4}$ the results depend strongly upon γ^{-1} and are useful if this quantity is not very large (see Section 5.5). Clearly Table 12.2 justifies our early comparison on the effectiveness of the conditioning matrices R,R_1 (see Section 4.4) establishing therefore the credibility of the used criteria. In addition, the construction of the PSD method by using the preconditioning techniques, its superiority over SSOR and in certain cases over SOR, confirm further the idea of how one should use these techniques in order to associate the most effective iterative scheme with a given conditioning matrix and on the other hand, the strong need to study the effectiveness of the different forms of conditioning matrices associated with the various splittings of the coefficient matrix A.

4.13 COMPUTATIONAL RESULTS

Before we verify our theoretical analysis by presenting various numerical experiments we consider the application of the above results to the following model problem. Given a continuous function $g(x,y)$ defined on the boundary S of the unit square $R: 0 \leq x \leq 1$, $0 \leq y \leq 1$ find a function $U(x,y)$ continuous in the closed square and satisfying in the interior R the Laplace equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0. \tag{13.1}$$

On the boundary, we require that

$$U(x,y) = g(x,y). \tag{13.2}$$

By considering the five-point discrete analogue with the natural ordering it is easy to show (see Young [1971], Chapter 4) that

$$S(B) = \cos\pi h \tag{13.3}$$

where h is the mesh size.

Moreover, the matrix A can be shown to possess Property A and to be consistently ordered (see Chapters 1 and 2). It is easy to show also (see, for instance, Ehrlich [1963], Appendix B) that

$$S(LU) = \frac{1}{4}\cos^2\frac{\pi h}{2(1-h)}. \tag{13.4}$$

From (13.4) we can let $\bar{\beta}$ be given by

$$\bar{\beta} = \frac{1}{4}\cos^2\frac{\pi h}{2}. \tag{13.5}$$

Evidently, by (13.4) and (13.5) we have

$$S(LU) \leq \bar{\beta}.$$

We now determine a good value of ω using (11.4) with $M=\cos\pi h$ and with $\bar{\beta}$ given by (13.5). We note that

$$2\sqrt{\bar{\beta}} = \cos\frac{\pi h}{2} \geq 4\bar{\beta} = \cos^2\frac{\pi h}{2} \geq \cos\pi h = M \tag{13.6}$$

and hence we can apply Theorem 11.1 to obtain

$$\omega_1 = \frac{2}{1+\sqrt{1-2\cos\pi h+\cos^2\frac{\pi h}{2}}} = \frac{2}{1+\sqrt{3}\sin\frac{\pi h}{2}} \tag{13.7}$$

and

$$P(B_{\omega_1}) \leq \frac{1}{2}\left(1+\frac{\sqrt{3}\sin\frac{\pi h}{2}}{1-\cos\pi h}\right) = \frac{1}{2}\left(1+\frac{\sqrt{3}}{2\sin\frac{\pi h}{2}}\right). \tag{13.8}$$

Therefore, for sufficiently small h we have

$$RR(D_{\tau_1,\omega_1}) \sim \frac{P(B_{\omega_1})}{2} \sim \frac{1}{4}(1+\frac{\sqrt{3}}{\pi}h^{-1}). \tag{13.9}$$

Similarly, by using (3-6.49) we can find

$$S(\mathcal{L}_{\omega_1}) \leq \frac{1-\frac{2}{\sqrt{3}}\sin\frac{\pi h}{2}}{1+\frac{2}{\sqrt{3}}\sin\frac{\pi h}{2}} \sim 1-\frac{2\pi h}{\sqrt{3}}, \tag{13.10}$$

therefore for small h we have

$$RR(\mathcal{L}_{\omega_1}) \sim \frac{\sqrt{3}}{2\pi}h^{-1}. \tag{13.11}$$

By comparing the PSD with SSOR we obtain the following result

$$\frac{RR(D_{\tau_1,\omega_1})}{RR(\mathcal{L}_{\omega_1})} \sim \frac{1}{2}(1+\frac{\pi h}{\sqrt{3}}). \tag{13.12}$$

The values of this ratio for h=1/20, 1/40, 1/60 and 1/80 are illustrated in the following tabulation

h	$RR(D_{\tau_1,\omega_1})/RR(\mathcal{L}_{\omega_1})$
1/20	0.545
1/40	0.523
1/60	0.515
1/80	0.511

Consequently, for the above model problem and as the mesh size h tends to zero the number of iterations of PSD tends to become half the number of iterations of SSOR.

For the SOR method, since A is consistently ordered, it follows from (3-6.22) that the optimum value of ω is given by

$$\omega_b = \frac{2}{1+\sin\pi h} \tag{13.13}$$

whereas by (3-6.23) we obtain

$$S(L_{\omega_b}) = \frac{1-\sin\pi h}{1+\sin\pi h} \sim 1-2\pi h \tag{13.14}$$

for small h. Finally, from (13.14) it follows that

$$RR(L_{\omega_b}) \sim \frac{1}{2\pi} h^{-1}. \tag{13.15}$$

By comparing now the PSD with the SOR we obtain

$$\frac{RR(D_{\tau_1,\omega_1})}{RR(L_{\omega_b})} \sim \frac{\sqrt{3}}{2}\left(1+\frac{\pi h}{\sqrt{3}}\right) \tag{13.16}$$

thus as $h \to 0$ the limiting value of the ratio in (13.16) is

$$\frac{RR(D_{\tau_1,\omega_1})}{RR(L_{\omega_b})} \sim \frac{\sqrt{3}}{2} = 0.866. \tag{13.17}$$

which implies that for the model problem the rate of convergence of the PSD method is even better than SOR. On the other hand, if one also compares the two methods in terms of required computational work (see Appendix A), then it seems interesting to investigate further the possibility of using the PSD method with Niethammer's scheme.

In order to test our theoretical results obtained so far, the same numerical experiments, as described in Section 4.8, were carried out. For purposes of comparison, we considered the application of SOR, SSOR and PSD methods with optimum and estimated parameters to the derived systems of equations corresponding to the problems presented in Table 8.2.

The quantities $\lambda(B_{\omega_0})$ and $S(\mathcal{G}_{\omega_0})$ presented in Table 13.1, were computed by using the power method combined with the Fibonacci search technique (e.g., see Gottfried and Weisman [1973] and Zahradnik [1971]), whereas $\Lambda(B_{\omega_0})$ was computed by (6.20). The value of ω_b based on the true value of $S(B)$, as determined by the power method, was used. Also, in Table 13.1 we present the number of iterations required to satisfy the convergence criterion $\|u^{(n)}\|_\infty \leq 10^{-6}$ for SOR, SSOR and PSD methods with optimum parameters. Furthermore, in Table 13.2 we present the estimated parameters τ_1, ω_1 and $P(B_{\omega_1})$ computed by (11.4),(11.19) and (11.5), respectively, where $\bar{\beta}$ is given by (B.6) (see Appendix B) and M is given in Young [1971a] (see also (5-5.27)). In Table 13.2 we also give the number of iterations for the SSOR

Problem	h^{-1}	ω_0	$S(\mathcal{E}_{\omega_0})$	$\lambda(B_{\omega_0})$	$\Lambda(B_{\omega_0})$	$P(B_{\omega_0})$	τ_0	OPTIMUM SSOR	OPTIMUM PSD	h^{-1}	ω_b	SOR
1	20	1.7641	0.8099	0.4568	2.4030	5.2604	0.6993	66	37	20	1.7295	61
	40	1.8750	0.9008	0.4233	4.2667	10.0806	0.4264	134	71	40	1.8547	121
	60	1.9157	0.9343	0.4068	6.1922	15.2207	0.3031	201	107	80	1.9237	253
2	20	1.5888	0.5876	0.6313	1.5307	2.4248	0.9251	24	17	20	1.5527	50
	40	1.7668	0.7663	0.5672	2.4271	4.2790	0.6679	48	30	40	1.7460	99
	60	1.8386	0.8386	0.5439	3.3698	6.1958	0.5110	71	44	80	1.8902	217
3	20	1.7652	0.8140	0.4488	2.4127	5.3763	0.6989	68	38	20	1.7326	60
	40	1.8756	0.9031	0.4153	4.2859	10.3200	0.4254	137	72	40	1.8564	121
	60	1.9163	0.9343	0.4096	6.2346	15.2207	0.3010	205	107	80	1.9247	252
4	20	1.7624	0.8088	0.4566	2.3881	5.2301	0.7031	66	37	20	1.7385	59
	40	1.8748	0.9002	0.4252	4.2603	10.0200	0.4268	133	70	40	1.8599	119
	60	1.9143	0.9324	0.4121	6.0955	14.7929	0.3073	200	104	80	1.9260	225
5	20	1.7479	0.8281	0.3901	2.2694	5.8173	0.7520	74	41	20	1.7233	60
	40	1.8665	0.9105	0.3592	4.0132	11.1732	0.4574	149	79	40	1.8515	118
	60	1.9093	0.9395	0.3494	5.7746	16.5289	0.3266	224	117	80	1.9191	274
6	20	1.6097	0.6035	0.6311	1.5917	2.5221	0.8998	28	17	20	1.5528	41
	40	1.7820	0.7855	0.5779	2.5742	4.4543	0.6345	57	32	40	1.7448	81
	60	1.8490	0.8438	0.5593	3.5817	6.4020	0.4829	85	47	80	1.8907	176

TABLE 13.1

Problem	h^{-1}	$\bar{\beta}$	$2\sqrt{\beta}$ *	M	ω_1	τ_1	$P(B_{\omega_1})$	ESTIMATED	
								SSOR	PSD
1	20	0.2500	1.0000	0.9877	1.7287	0.8188	6.8727	68	48
	40	0.2500	1.0000	0.9969	1.8544	0.5021	13.2357	138	93
	60	0.2500	1.0000	0.9986	1.9005	0.3598	19.6008	207	137
2	20	0.2350	0.9695	1.0000	1.6065	0.9073	2.5415	28	18
	40	0.2461	0.9922	1.0000	1.7788	0.6444	4.5208	45	33
	60	0.2483	0.9965	1.0000	1.8465	0.4914	6.5139	67	47
3	20	0.2505	1.0009	0.9969	1.8355	0.5661	14.9905	73	101
	40	0.2501	1.0002	0.9992	1.9142	0.3177	29.5540	145	195
	60	0.2501	1.0001	0.9997	1.9420	0.2203	44.1015	218	-
4	20	0.2500	1.0001	0.9918	1.7717	0.7223	8.3322	66	59
	40	0.2500	1.0000	0.9979	1.8790	0.4282	16.1660	133	114
	60	0.2500	1.0000	0.9991	1.9176	0.3034	23.9999	200	168
5	20	0.2500	0.9999	0.9978	1.8756	0.4379	15.2395	97	65
	40	0.2500	1.0000	0.9994	1.9359	0.2402	30.0138	193	100
	60	0.2500	1.0000	0.9998	1.9568	0.1654	44.7804	288	144
6	20	0.2416	0.9831	1.0000	1.6903	0.7994	3.2293	36	23
	40	0.2483	0.9966	1.0000	1.8475	0.4889	6.5567	82	47
	60	0.2493	0.9986	1.0000	1.8997	0.3463	9.9697	129	71

TABLE 13.2

In Problems 2 and 6 in the determination of ω_1 and $P(B_{\omega_1})$, the value $2\sqrt{\bar{\beta}}$ was used instead of M.

and PSD method using the estimated parameters.

To determine the rate of convergence which has been attained by applying the SSOR, SOR and PSD methods with optimum and estimated parameters, we plot the logarithm of the number of iterations versus $\log h^{-1}$ for problems 1,2 and 6 in Figures 13.1 and 13.2.

From Table 13.1 we see that our theoretical expectations are verified since in all cases we have a substantial reduction in the number of iterations of the PSD method as compared with SSOR and SOR. It is readily verified that in all the considered problems we have i) a reduction of at least 39% (except in problem 2 where the reduction is at least 29%) of the number of iterations of the PSD over SSOR and ii) a reduction of at least 32% of the number of iterations of the PSD over SOR. On the other hand if we consider the SSOR and the PSD methods with estimated parameters, then from Table 13.2 we observe that for problems 1,2,5 and 6 we have a reduction of at least 27% in the number of iterations, whereas for problem 4 this percentage is somewhat less. For problem 3, where $\bar{\beta}>1/4$, (see Table 12.1) the convergence of the PSD method is erratic. It is conjectured that this is due to the crude bounds used for the quantities S(B) and S(LU) since such a phenomenon does not exist in the case where optimum parameters are used. Furthermore, for all problems (again problem 3 is an exception) we have a reduction of at least 25% for h=1/60, using PSD with estimated parameters as compared to SOR.

Finally, from Figures 13.1 and 13.2 we confirm our theoretical results by noting that the rate of convergence is approximately O(h).

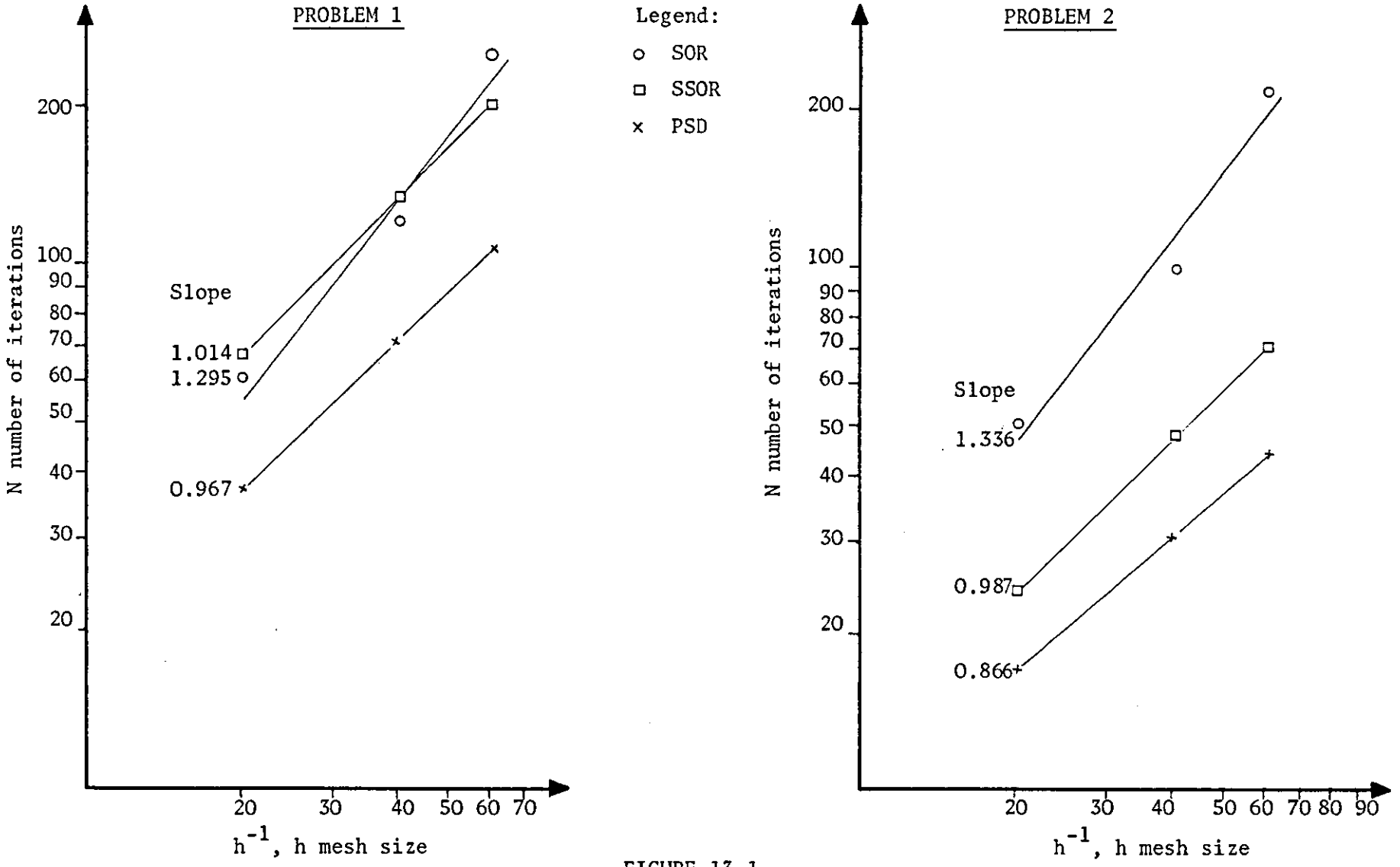


FIGURE 13.1

DETERMINATION OF THE RATE OF CONVERGENCE ATTAINED FOR PROBLEMS 1,2 AND 6 USING PSD AND SSOR WITH OPTIMUM PARAMETERS. THE SLOPE α INDICATES $O(h^{\alpha})$ RATE OF CONVERGENCE

PROBLEM 6

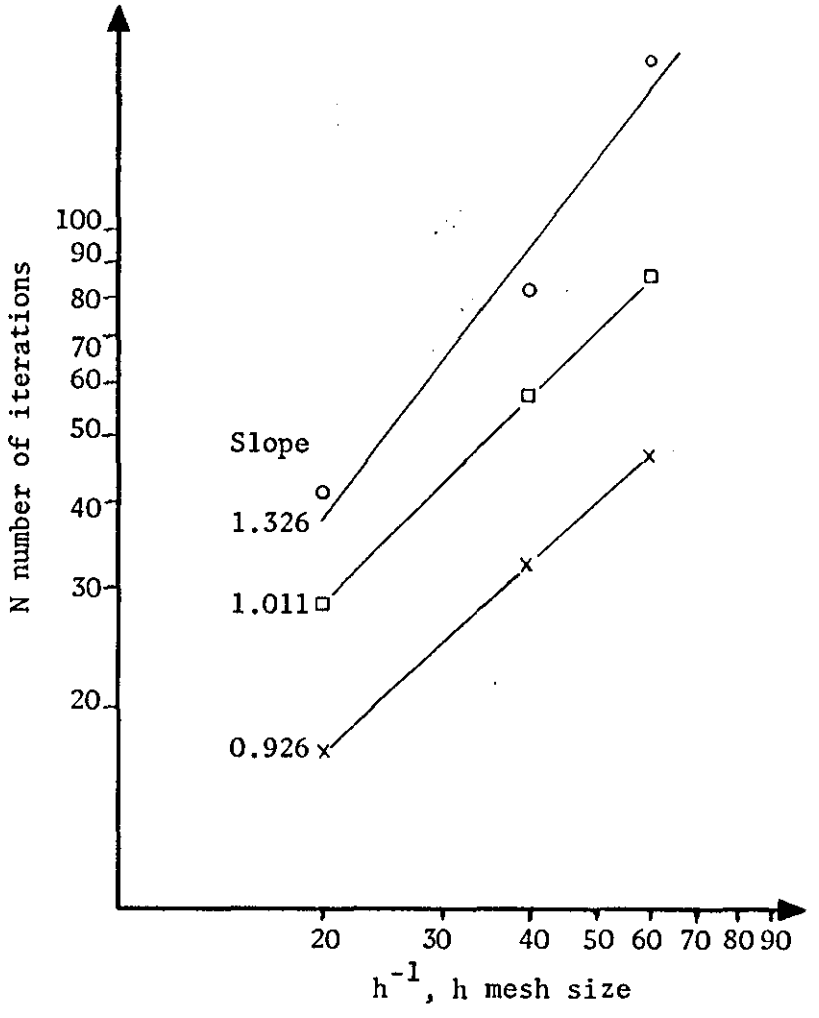


FIGURE 13.1 (CONTINUED)

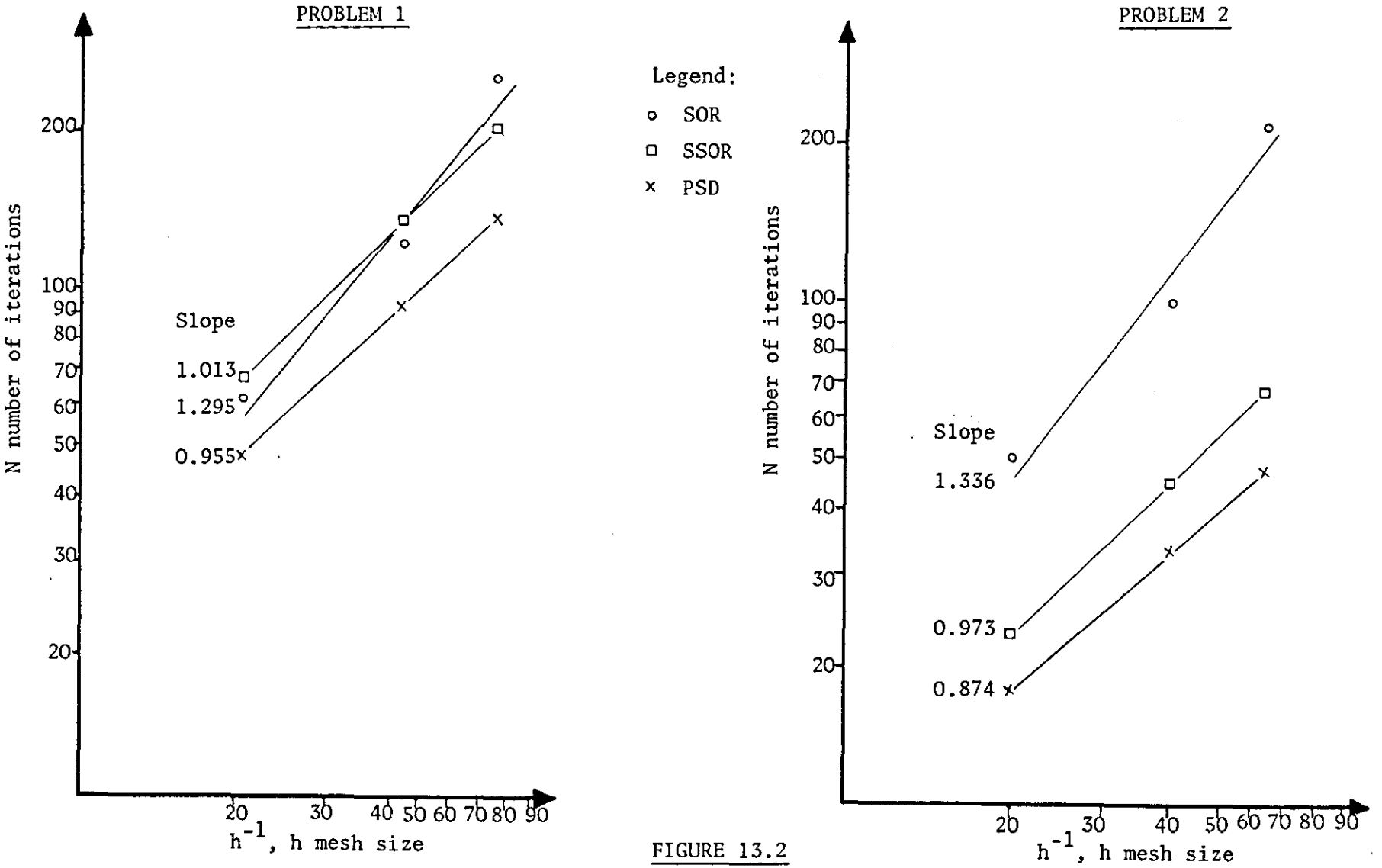


FIGURE 13.2

DETERMINATION OF RATE OF CONVERGENCE ATTAINED FOR PROBLEMS 1,2 AND 6 USING PSD AND SSOR WITH ESTIMATED PARAMETERS

PROBLEM 6

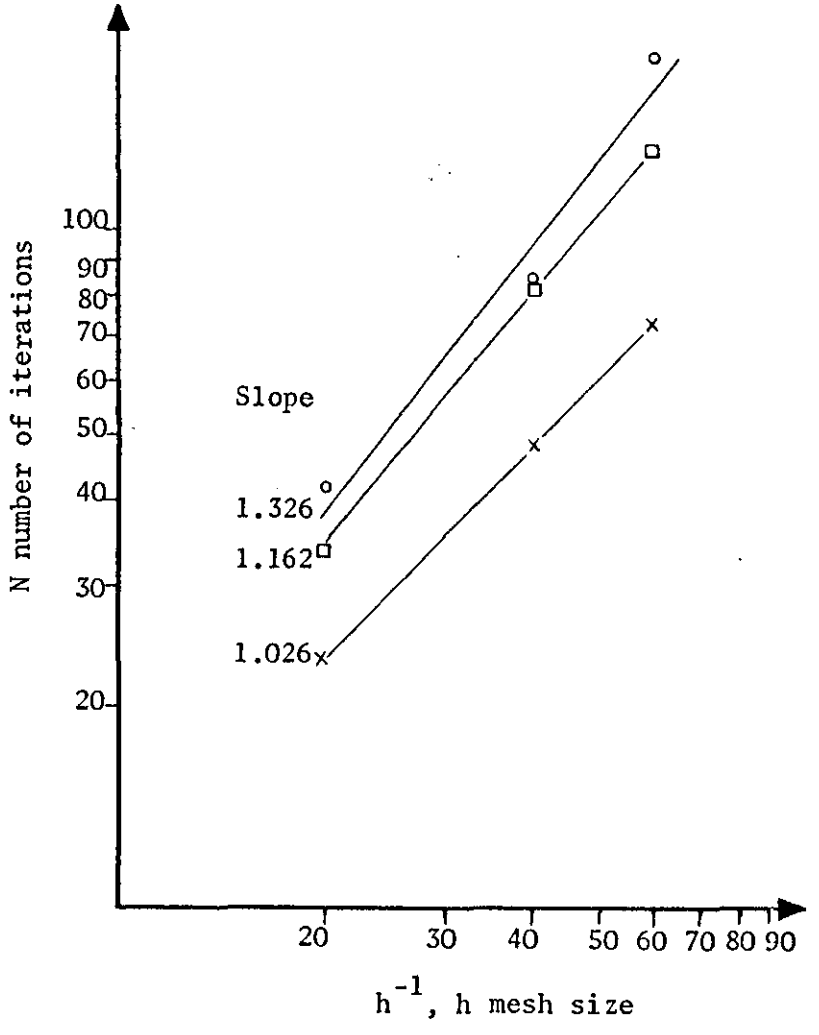


FIGURE 13.2 (CONTINUED)

4.14 THE UNSYMMETRIC PJ METHOD

In this section we consider the conditioning matrix R to have the general form

$$R = (D-\omega_1 C_L)D^{-1}(D-\omega_2 C_U) \tag{14.1}$$

which is obtained from (2.4) and (2.9) by assuming that A has the splitting

$$A = D-C_L-C_U. \tag{14.2}$$

As can be seen the preconditioned matrix is given by

$$P(B_{\omega_1,\omega_2}) = (I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}D^{-1}A \tag{14.3}$$

which is not a symmetric matrix.

By using the above defined conditioning matrix $P(B_{\omega_1,\omega_2})$ we form first the following iterative scheme (see (2.7) for $\tau=1$)

$$u^{(n+1)} = u^{(n)} + (I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}D^{-1}(b-Au^{(n)}) \tag{14.4}$$

which defines the unsymmetric PJ method (UPJ method). The UPJ method can also be written as

$$u^{(n+1)} = Q_{\omega_1,\omega_2}u^{(n)} + q_{\omega_1,\omega_2} \tag{14.5}$$

where
$$Q_{\omega_1,\omega_2} = I-(I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}D^{-1}A \tag{14.6}$$

and
$$q_{\omega_1,\omega_2} = (I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}c. \tag{14.7}$$

We will only concentrate on the case when A is of order N and has the form

$$A = I-B = \begin{pmatrix} I_1 & -U^* \\ -L^* & I_2 \end{pmatrix} \tag{14.8}$$

where U^* is an $m\times r$ submatrix, L^* is an $r\times m$ submatrix, I_1, I_2 are $m\times m$ and $r\times r$ identity submatrices, respectively and $m+r=N$.

Theroem 14.1

If A is a real symmetric matrix of the form (14.8), then

$$S(Q_{\omega_1,\omega_2})<1 \tag{14.9}$$

if and only if

$$-\frac{1-\bar{\mu}^2}{2\bar{\mu}^2} < \hat{\omega} < \frac{1+\bar{\mu}^2}{\bar{\mu}^2} \tag{14.10}$$

and $\bar{\mu} < 1$

where $\bar{\mu}=S(B)$ and

$$\hat{\omega} = \omega_1 + \omega_2 - \omega_1\omega_2. \tag{14.11}$$

Proof

It can be easily verified that if A has the form (14.8), then Q_{ω_1,ω_2} has the form

$$Q_{\omega_1,\omega_2} = \begin{bmatrix} 0 & (1-\omega_2)U^* \\ (1-\omega_1)L^* & (\omega_1+\omega_2-\omega_1\omega_2)L^*U^* \end{bmatrix} \tag{14.12}$$

If λ is an eigenvalue of Q_{ω_1,ω_2} and $y=\begin{bmatrix} b \\ d \end{bmatrix}$ is the corresponding eigenvector, the partitions of y corresponding to the partitions of A in (14.8), then we have

$$Q_{\omega_1,\omega_2} y = \lambda y \tag{14.13}$$

or from (14.12)

$$\begin{bmatrix} 0 & (1-\omega_2)U^* \\ (1-\omega_1)L^* & (\omega_1+\omega_2-\omega_1\omega_2)L^*U^* \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} = \lambda \begin{bmatrix} b \\ d \end{bmatrix} \tag{14.14}$$

which simplifies to the following system of equations

$$\begin{aligned} (1-\omega_2)U^*d &= \lambda b \\ (1-\omega_1)L^*b + (\omega_1+\omega_2-\omega_1\omega_2)L^*U^*d &= \lambda d. \end{aligned} \tag{14.15}$$

Eliminating b from (14.15) we have

$$[(1-\omega_1)(1-\omega_2)+\lambda(\omega_1+\omega_2-\omega_1\omega_2)]L^*U^*d = \lambda^2 d. \tag{14.16}$$

It is easily shown that the non-zero eigenvalues of B occur in pairs $\pm\mu_i$ $(i=1,2,\ldots,M)$, where M is less than or equal to the number of rows in L* or U*. Furthermore, the eigenvalues of L*U* are

$$0 \leq \mu_i^2 \leq \bar{\mu}^2 = S(B)^2 \tag{14.17}$$

hence by (14.16) we have the following eigenvalue relationship

$$[(1-\omega_1)(1-\omega_2)+\lambda(\omega_1+\omega_2-\omega_1\omega_2)]\mu^2 = \lambda^2 \tag{14.18}$$

which can be written as

$$\lambda^2-\lambda\hat{\omega}\mu^2+\mu^2(\hat{\omega}-1) = 0 \tag{14.19}$$

where $\hat{\omega}$ is given by (14.11).

Consequently, the theorem is a result of Theorem 3.4, since (14.19) is similar to (3.26).

Moreover, as a result of Theorem 3.5 we have

Theorem 14.2

If A is a real symmetric matrix of the form (14.8) and $\bar{\mu}=S(B)<1$, then

$$\bar{S}(Q_{\omega_1,\omega_2}) = S(Q_{\omega_1,\omega_2}) = \bar{\mu}(\omega_b-1)^{\frac{1}{2}} \tag{14.20}$$

where

$$\hat{\omega} = \omega_1 +\omega_2 -\omega_1\omega_2 = \omega_b \tag{14.21}$$

and

$$\omega_b = \frac{2}{1+\sqrt{1-\bar{\mu}^2}} . \tag{14.22}$$

Thus, since twice as much work is required per iteration using the UPJ method as with the SOR method and since the rate of convergence is no better (see (3.33)), the UPJ method is of academic interest, at least when A has the form (14.8). In addition, we note that when $\omega_1=\omega_2=\omega$, then we have the PJ method and

$$\hat{\omega} = \omega(2-\omega) \leq 1$$

with equality holding at $\omega=\bar{\omega}=1$ which is the best value of ω since $\omega_b \geq 1$ (see Theorem 14.2). For $\omega=1$, equation (14.19) reduces to

$$\lambda = \mu^2 \tag{14.23}$$

and the eigenvalues of the optimised PJ are identical with the eigenvalues of the GS method. Consequently, the rate of convergence of the PJ method is affected by different consistent orderings. We also note that under the same conditions, the eigenvalues of SSOR are given by (14.23) (see D'Sylva and Miles [1963]), thus PJ and SSOR are identical at the optimum stage when A has the form (14.8).

4.15 THE UNSYMMETRIC PSD METHOD (UPSD METHOD)

Evidently the unsymmetric PSD method (UPSD method) is defined by

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}D^{-1}(b-Au^{(n)}) \tag{15.1}$$

or

$$u^{(n+1)} = G_{\omega_1,\omega_2}u^{(n)}+g_{\omega_1,\omega_2} \tag{15.2}$$

where

$$G_{\omega_1,\omega_2} = I-\tau(I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}D^{-1}A \tag{15.3}$$

and

$$g_{\omega_1,\omega_2} = \tau(I-\omega_2 U)^{-1}(I-\omega_1 L)^{-1}c. \tag{15.4}$$

If λ is an eigenvalue of $P(B_{\omega_1,\omega_2})$ given by (14.3) and μ is an eigenvalue of B, then working similarly as in the previous section we have the following eigenvalue relationship when A has the form (14.8)

$$\lambda^2-\lambda(2-\hat{\omega}\mu^2)+1-\mu^2 = 0 \tag{15.5}$$

where $\hat{\omega}$ is given by (14.11).

Therefore, from the analysis of the subsection 4.3.4 we have as a result the following theorems.

Theorem 15.1

If A is a real symmetric matrix of the form (14.8), then

$$S(G_{\omega_1,\omega_2}) < 1 \tag{15.6}$$

if and only if

$$\bar{\mu} = S(B) < 1 \tag{15.7}$$

and the parameters τ and $\hat{\omega}$ lie in either of the following ranges:

$$\left.\begin{array}{llll} \text{for } \hat{\omega}\geq 0 & 0<\tau<1 & \text{and} & 0\leq\hat{\omega}<1 \\ \text{or} & 1\leq\tau<2 & \text{and} & 1\leq\hat{\omega}\leq 2 \end{array}\right\} \tag{15.8}$$

for $\hat{\omega}\leq 0$, the ranges of τ remain the same but the corresponding ranges of $\hat{\omega}$ are the following:

$$\left.\begin{array}{l} -1<\hat{\omega}\leq 0 \\ -2\leq\hat{\omega}\leq -1 \end{array}\right\} \tag{15.9}$$

Theorem 15.2

If A is a real symmetric matrix of the form (14.8) and $\bar{\mu}=S(B)<1$, then

$$\hat{\omega} = \omega_1 + \omega_2 - \omega_1\omega_2 = \omega_b = \tau_0. \tag{15.10}$$

and

$$\bar{S}(G_{\omega_1,\omega_2}) = S(G_{\omega_1,\omega_2}) = \omega_b - 1 \tag{15.11}$$

where ω_b is given by (14.22).

Thus, since twice as much work is required per iteration using the UPSD method as with SOR, the UPSD method would appear to be mainly of academic interest, at least when A has the form (14.8). If we now let $\omega_1=\omega_2=\omega$, then we have the PSD method and

$$\hat{\omega} = \omega(2-\omega) \leq 1 \tag{15.12}$$

with equality holding at $\omega=\bar{\omega}=1$ which is the best value of ω since $\omega_b \geq 1$. For $\omega=1$ equation (15.5) reduces to

$$\lambda^2-\lambda(2-\mu^2) + 1-\mu^2 = 0. \tag{15.13}$$

Thus we have

$$\lambda_{max} = 1$$

and

$$\lambda_{min} = 1-\bar{\mu}^2 \tag{15.14}$$

which implies that for convergence the parameter τ must lie in the range

$$0<\tau<2. \tag{15.15}$$

Finally,

$$P(B_{\hat{\omega}}) = \frac{1}{1-\bar{\mu}^2} \tag{15.16}$$

and the optimum value for τ is given by the formula

$$\tau_0 = \frac{2}{2-\bar{\mu}^2} \tag{15.17}$$

Therefore, the spectral radius of the PSD method is given by the expression

$$S(G_{\hat{\omega}=1}) = \frac{\bar{\mu}^2}{2-\bar{\mu}^2}. \tag{15.18}$$

Since it is known (D'Sylva and Miles [1963]) that under the same conditions the eigenvalues of SSOR are identical with those of the GS method, then from Theorem 3.3 it follows that:

Theorem 15.3

Under the hypotheses of Theorem 15.2 we have

$$\lim_{\bar{\mu}\to 1^-} \frac{R(G_{\hat{\omega}=1})}{R(\mathcal{L}_{\hat{\omega}=1})} = 2. \tag{15.19}$$

Thus the asymptotic improvement of the PSD method by a factor of 2 over SSOR is still retained and in the case where A has the form (14.8).

CHAPTER 5

BLOCK PRECONDITIONED ITERATIVE METHODS - ACCELERATED TECHNIQUES

SECTION A

BLOCK PRECONDITIONED ITERATIVE METHODS

5.1 INTRODUCTION

In the previous chapter we considered various iterative schemes where we determined each component of $u^{(n)}$ explicitly, i.e., by using already computed approximate values of the other unknowns. As is known these schemes are called point methods in order to be distinguished from the group iterative methods. In the latest methods, we first assign the equation to groups and then we solve the group of equations for the corresponding unknowns u_i, treating the other values of u_j as known (implicit methods).

A special case of a grouping is a partitioning where for some integers $n_1,n_2,\ldots,n_q$ such that $1\leq n_1<n_2<\ldots<n_q=N$ 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 partitionings are known as block methods. The theory of block methods is well known (Southwell [1946], Geiringer [1949]).

Arms, Gates and Zondek [1956] first generalised SOR to block method and Friedman [1957] analysed its convergence rate. In addition, Varga [1960] showed that the rate of convergence of the two-line SOR method with optimum ω is approximately twice that of point SOR, whereas Parter [1961] showed that the k-line SOR method with optimum ω converges approximately $(2k)^{\frac{1}{2}}$ as fast as point SOR. Finally, Ehrlich [1963,1964] considered the line SSOR for the five-point discrete Dirichlet problem and was able to show that the convergence is faster than the point SSOR method.

In the first part of this chapter we will extend the preconditioning techniques so that to show, in an analogous way to Chapter 4, how we can construct and develop the corresponding group methods of the previously considered iterative procedures. We therefore commence our consideration by presenting a brief review of some basic concepts concerned with the

definition and convergence of the group methods.

Definition 1.1

An ordered grouping π 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 π denote the ordered grouping defined by $R_k=\{k\}$, $k=1,2,\ldots,N$.

Given a matrix A and an ordered grouping π 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 (see Chapter 2).

Given a matrix A and an ordered grouping π, with q groups, we define the $q\times q$ matrix $Z=(z_{r,s})$ by

$$z_{r,s} = \begin{cases} 0, & \text{if } A_{r,s}=0 \\ 1, & \text{if } A_{r,s}\neq 0. \end{cases} \tag{1.1}$$

Definition 1.2

The matrix A has Property $A^{(\pi)}$ if Z has Property A.

Definition 1.3

The matrix A is a π-consistently ordered matrix (π-CO-matrix) if Z is consistently ordered.

Definition 1.4

A matrix A is a generalised π-consistently ordered matrix (π-GCO-matrix) if[†]

$$\Delta = \det(\alpha C_L^{(\pi)}+\alpha^{-1}C_U^{(\pi)}-kD^{(\pi)})$$

where

$$A = D^{(\pi)}-C_L^{(\pi)}-C_U^{(\pi)}, \tag{1.2}$$

is independent of α for all $\alpha\neq 0$ and for all k.

Here $D^{(\pi)}$ is the matrix formed from A by replacing with zeros 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

---

[†]*We will use the notation* $B^{(\pi)}$ *to denote the group form of the matrix* B.

that i and j belong to different groups and such that the group containing i comes after and before, respectively the group containing j.

Theorem 1.1 (Arms, Gates and Zondek [1956])

If A is a π-GCO-matrix such that $D^{(\pi)}$ is non-singular, then the conclusions of Theorem 3-6.4 are valid if we replace B by $B^{(\pi)}$ and L_ω by $L_\omega^{(\pi)}$.

From the above analysis we note that the definition of the group methods is based on the splitting (1.2). Thus following the analysis of the preconditioning techniques we can regard (1.2) 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

$$R = D^{(\pi)} \tag{1.3}$$

for any ordered grouping π, then we define the group SD method (using (4-2.2) and (1.3)) by

$$u^{(n+1)} = u^{(n)} + \tau(D^{(\pi)})^{-1}(b-Au^{(n)}) \tag{1.4}$$

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 π, since if $D^{(\pi)}=A$, then we solve our system immediately. On the other hand, the invertion of $D^{(\pi)}$ by using direct methods (Cuthill and Varga [1959]) is a limit to the above observation.

Further, by letting the conditioning matrix have the form

$$R = D^{(\pi)}(I-L^{(\pi)}) \tag{1.5}$$

we define the group EGS method by

$$u^{(n+1)} = u^{(n)}+\tau(I-L^{(\pi)})^{-1}(D^{(\pi)})^{-1}(b-Au^{(n)}) \tag{1.6}$$

where

$$L^{(\pi)} = (D^{(\pi)})^{-1}C_L^{(\pi)} \quad \text{and} \quad U^{(\pi)} = (D^{(\pi)})^{-1}C_U^{(\pi)}. \tag{1.7}$$

Finally, we can also define the group ESOR method, by letting the conditioning matrix have the form

$$R = D^{(\pi)}(I-\omega L^{(\pi)}) \tag{1.8}$$

hence

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}(b-Au^{(n)}) \tag{1.9}$$

where τ,ω are real parameters and their role is familiar to us from the previous chapter. A more compact form of the group ESOR method is given by

$$u^{(n+1)} = L_{\tau,\omega}^{(\pi)}u^{(n)} + \tau(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}b \tag{1.10}$$

where

$$\begin{aligned} L_{\tau,\omega}^{(\pi)} &= I-\tau(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}A \\ &= (I-\omega L^{(\pi)})^{-1}[(1-\tau)I+(\tau-\omega)L^{(\pi)}+\tau U^{(\pi)}]. \end{aligned} \tag{1.11}$$

For actual computation with the group ESOR method we solve the system

$$(D^{(\pi)}-\omega C_L^{(\pi)})u^{(n+1)} - [(1-\tau)D^{(\pi)}+(\tau-\omega)C_L^{(\pi)}+\tau C_U^{(\pi)}]u^{(n)} = \tau b \tag{1.12}$$

for $u^{(n+1)}$.

If the system Au=b is written in the form

$$\sum_{s=1}^{q} A_{r,s}U_s = B_r \ , \ r=1,2,\ldots,q \tag{1.13}$$

where the matrices $A_{r,s}$ have been defined earlier and if we define U_s and B_r similarly, then (1.12) is equivalent to solving

$$\begin{aligned} A_{r,r}U_r^{(n+1)} - \omega\sum_{s=1}^{r-1}A_{r,s}U_s^{(n+1)} - (1-\tau)A_{r,r}U_r^{(n)} + (\omega-\tau)\sum_{s=1}^{r-1}A_{r,s}U_s^{(n)} - \\ -\tau\sum_{s=r+1}^{q}A_{r,s}U_s^{(n)} = \tau B_r, \ r=1,2,\ldots,q \end{aligned} \tag{1.14}$$

successively for $U_1^{(n+1)},U_2^{(n+2)},\ldots,U_q^{(n+1)}$.

The conditions under which the previous schemes converge are similar to their point versions which have been thoroughly considered (see Chapters 3 and 4). However, the derivation of a relation between the eigenvalues of the preconditioned matrix $(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}A$ and $B^{(\pi)}=C_L^{(\pi)}+C_U^{(\pi)}$ similar to that obtained in Chapter 4 for $(I-\omega L)^{-1}D^{-1}A$ and B, is possible for π-GCO matrices.

We can therefore generalise in an anlogous way the results obtained in Theorem 4-3.4.8.

Theorem 1.2

Let A be a π-GCO matrix such that $D^{(\pi)}$ is non-singular. If $B^{(\pi)}$ has real eigenvalues $\mu_i^{(\pi)} | i=1(1)N$ with $\underline{\mu}^{(\pi)}=\min|\mu_i^{(\pi)}|$ and $\bar{\mu}^{(\pi)}=\max|\mu_i^{(\pi)}|$, such that $\bar{\mu}^{(\pi)}=S(B^{(\pi)})<1$ and if

$$\omega_b^{(\pi)} = \frac{2}{1+\sqrt{1-(\bar{\mu}^{(\pi)})^2}}, \qquad \tau_0^{(\pi)} = \frac{4\omega_b^{(\pi)}}{4-(\omega_b^{(\pi)})^2(\underline{\mu}^{(\pi)})^2}, \tag{1.15}$$

then for $\tau \neq \tau_0^{(\pi)}$ and $\omega \neq \omega_b^{(\pi)}$

$$\bar{S}(L_{\tau,\omega}^{(\pi)}) = S(L_{\tau,\omega}^{(\pi)}) > S(L_{\tau_0^{(\pi)},\omega_b^{(\pi)}}^{(\pi)}) \tag{1.16}$$

where

$$\bar{S}(L_{\tau_0^{(\pi)},\omega_b^{(\pi)}}^{(\pi)}) = S(L_{\tau_0^{(\pi)},\omega_b^{(\pi)}}^{(\pi)}) = 1-\tau_0^{(\pi)}-2\tau_0^{(\pi)}/\omega_b^{(\pi)}. \tag{1.17}$$

Evidently, if $\underline{\mu}^{(\pi)}=0$, then we have the well known results of the group SOR method.

5.2 GROUP PSD METHODS

For any ordered grouping π, we let the conditioning matrix have the form

$$R = D^{(\pi)}(I-\omega L^{(\pi)})(I-\omega U^{(\pi)}) \tag{2.1}$$

therefore we define the group PSD method by

$$u^{(n+1)} = u^{(n)} + \tau(I-\omega U^{(\pi)})^{-1}(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}(b-Au^{(n)}) \tag{2.2}$$

where τ,ω are some real parameters.

By using (4-2.8) we write the iterative scheme (2.1) in a computable form (another form can be produced by considering (A.10), see Appendix A)

$$\left.\begin{aligned} D^{(\pi)}\zeta^{(n+\frac{1}{2})} &= \omega C_L^{(\pi)}\zeta^{(n+\frac{1}{2})} + r^{(n)} \\ D^{(\pi)}\zeta^{(n+1)} &= \omega C_U^{(\pi)}\zeta^{(n+1)} + D^{(\pi)}\zeta^{(n+\frac{1}{2})} \\ u^{(n+1)} &= u^{(n)} + \tau\zeta^{(n+1)} \end{aligned}\right\} \tag{2.3}$$

where

$$r^{(n)} = b-Au^{(n)}. \tag{2.4}$$

For the analysis of the method, it is convenient to write (2.3) in the following form

$$u^{(n+1)} = D_{\tau,\omega}^{(\pi)}u^{(n)} + \delta^{(\pi)} \tag{2.5}$$

where

$$D_{\tau,\omega}^{(\pi)} = I-\tau(I-\omega U^{(\pi)})^{-1}(I-\omega L^{(\pi)})^{-1}(D^{(\pi)})^{-1}A \tag{2.6}$$

and

$$\delta^{(\pi)} = \tau(I-\omega U^{(\pi)})^{-1}(I-\omega L^{(\pi)})^{-1}b. \tag{2.7}$$

Evidently, by (2.6),(2.7) we see that the group PSD method is completely consistent if $D^{(\pi)}$ is non-singular and $\tau\neq 0$. Most of the analysis in Sections 4.10,4.11 can be applied to group PSD methods. Before we proceed in a more detailed analysis of the behaviour of $S(D_{\tau,\omega}^{(\pi)})$, we determine the spectral radius of the PSD method applied to a smaller system derived from Au=b.

Theorem 2.1

If A is a symmetric and positive definite matrix and if A_* is obtained from A by deleting certain rows and the corresponding columns of A, then

$$P(B_{*_\omega}) \leq P(B_\omega) \tag{2.8}$$

where

$$B_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A. \tag{2.9}$$

Proof

Let $\lambda(B_{*_\omega})$ and $\Lambda(B_{*_\omega})$ denote the smallest and largest eigenvalue of

$$B_{*_\omega} = (I_*-\omega U_*)^{-1}(I-\omega L_*)^{-1}D_*^{-1}A_*. \tag{2.10}$$

From (4-5.2) we have that B_{*_ω} is similar to the symmetric matrix

$$\bar{B}_{*_\omega} = D_*^{\frac{1}{2}}(D_*-\omega C_{L_*})^{-1}A_*(D_*-\omega C_{U_*})^{-1}D_*^{\frac{1}{2}}. \tag{2.11}$$

If now v_* is an eigenvector associated with $\lambda(B_{*_\omega})$, then

$$\lambda(B_{*_\omega}) = \frac{(v_*,B_{*_\omega}v_*)}{(v_*,v_*)} = \frac{(w_*,A_*w_*)}{(v_*,v_*)} \tag{2.12}$$

where

$$w_* = (D_*-\omega C_{U_*})^{-1}D_*^{\frac{1}{2}}v_* \neq 0 \tag{2.13}$$

or

$$D_*^{\frac{1}{2}}v_* = (D_*-\omega C_{U_*})w_*. \tag{2.14}$$

Next, we augment w_* with zero components (at the positions which were deleted from A to form A_*) to form w and define v such that

$$D^{\frac{1}{2}}v = (D-\omega C_U)w. \tag{2.15}$$

Evidently, from the definition of w and v we have

$$\lambda(B_{*_\omega}) = \frac{(w_*,A_*w_*)}{(v_*,v_*)} \geq \frac{(w,Aw)}{(v,v)} \tag{2.16}$$

since

$$(w_*,A_*w_*) = (w,Aw)$$

and the influence of the added rows and columns in A is annihilated by the zero components of w. Further, by the definition of w, the right hand side of (2.15) has identical components as the right hand side of (2.14) plus additional ones. Since $D^{\frac{1}{2}}$ is diagonal, the components of v are identical as the components of v_* plus additional ones. If $\lambda(B_\omega)$ denotes the smallest eigenvalue of B_ω, then we have (see Theorem 2-1.5)

$$\lambda(B_\omega) \leq \frac{(v,\bar{B}_\omega v)}{(v,v)} = \frac{(w,Aw)}{(v,v)} \leq \frac{(w_*,A_*w_*)}{(v_*,v_*)} = \lambda(B_{*_\omega}). \tag{2.17}$$

Similarly, we can prove

$$\Lambda(B_{*\omega}) \leq \Lambda(B_{\omega}) \tag{2.18}$$

where $\Lambda(B_{*\omega})$ and $\Lambda(B_{\omega})$ denote the largest eigenvalues of $B_{*\omega}$ and B_{ω}, respectively.

Hence, if A is positive definite, then (2.8) follows from (2.17) and (2.18) and the proof of the theorem is complete.

A similar result for the SSOR method has been proved by Ehrlich [1963]. Although Theorem 2.1 applies only for the point PSD method, the numerical results (see Table 4.1) indicate that the theorem is probably true for at least certain other partitions π.

The analysis for the determination of good estimates for τ, the preconditioning parameter ω and the spectral radius of $D_{\tau,\omega}^{(\pi)}$ is similar to the one developed for the point PSD method (see Section 4.11-4.12). Consequently we can easily derive the conclusion that the group PSD method produces a gain of approximately a factor of 2 in the rate of convergence as compared with the group SSOR method.

5.3 COMPARISON OF LINE PSD AND POINT PSD METHODS

As an example of a block method we will choose the partitioning by lines of mesh points (x,y) with y constant, where the ordering is with increasing y. In this case the system is partitioned such that all the equations with y constant are grouped together and solved simultaneously. In the literature, this partition is frequently referred to as "line" iteration (see e.g. Varga [1962]). Subsequently, this partition will be denoted by π_1 whereas π_0 will be used to denote the point form of the method.

By using the difference equation (1-2.7) we can exhibit the line PSD method (LPSD method) as follows (see (2.3))

$$a_0\bar{u}_{i,j}^{(n+\frac{1}{2})}-a_1\bar{u}_{i+1,j}^{(n+\frac{1}{2})}-a_3\bar{u}_{i-1,j}^{(n+\frac{1}{2})} = \omega a_4\bar{u}_{i,j-1}^{(n+\frac{1}{2})}+a_1u_{i+1,j}^{(n)}+a_2u_{i,j+1}^{(n)}+a_3u_{i-1,j}^{(n)}+ a_4\bar{u}_{i,j-1}^{(n)}-a_0u_{i,j}^{(n)}$$

$$a_0\bar{u}_{i,j}^{(n+1)}-a_1\bar{u}_{i+1,j}^{(n+1)}-a_3\bar{u}_{i-1,j}^{(n+1)} = \omega a_2\bar{u}_{i,j+1}^{(n+1)}+a_0\bar{u}_{i,j}^{(n+\frac{1}{2})}-a_1\bar{u}_{i+1,j}^{(n+\frac{1}{2})}-a_3\bar{u}_{i-1,j}^{(n+\frac{1}{2})}$$

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)}+\tau\bar{u}_{i,j}^{(n+1)}. \tag{3.1}$$

Further by (2.5) the LPSD method can be written in the matrix form

$$u^{(n+1)} = D_{\tau,\omega}^{(\pi_1)}u^{(n)}+\delta^{(\pi_1)} \tag{3.2}$$

where $D_{\tau,\omega}^{(\pi_1)}$ and $\delta^{(\pi_1)}$ are given by (2.6),(2.7), respectively with π replaced by π_1. For the estimation of the rate of convergence we consider the application of Theorem 4-11.1. Young [1971] has shown that

$$S(L^{(\pi_1)}U^{(\pi_1)}) \leq \frac{1}{4} \tag{3.3}$$

which implies that from Theorem 4-11.1 we obtain

$$\omega_1^{(\pi_1)} = \begin{cases} \dfrac{2}{1+\sqrt{1-M^{(\pi_1)}}}, & \text{if } \beta^{(\pi_1)} \leq \dfrac{M^{(\pi_1)}}{4} \\ \dfrac{2}{1+\sqrt{2(1-M^{(\pi_1)})}}, & \text{if } \dfrac{M^{(\pi_1)}}{4} \leq \beta^{(\pi_1)} \leq \dfrac{1}{4} \end{cases} \tag{3.4}$$

where the corresponding value of $P(B_{\omega_1^{(\pi_1)}}^{(\pi_1)})$ is given by

$$P(B_{\omega_1^{(\pi_1)}}^{(\pi_1)}) \leq \begin{cases} \frac{1}{2}\left[1 + \frac{1}{\sqrt{1-M^{(\pi_1)}}}\right], & \text{if } \beta^{(\pi_1)} \leq \frac{M^{(\pi_1)}}{4} \\ \frac{1}{2}\left[1 + \frac{2^{\frac{1}{2}}}{\sqrt{1-M^{(\pi_1)}}}\right], & \text{if } \frac{M^{(\pi_1)}}{4} \leq \beta^{(\pi_1)} \leq \frac{1}{4}. \end{cases} \tag{3.5}$$

In particular, if we consider these results to be applied directly to systems of linear equations arising from the five-point difference equations considered in Section 1-2.1 in the unit square, we have (see Varga [1962])

$$M^{(\pi_1)} = S(B^{(\pi_1)}) = \frac{\cos\pi h}{2-\cos\pi h} \sim 1-\pi^2h^2 \tag{3.6}$$

as compared with

$$M^{(\pi_0)} = S(B^{(\pi_0)}) = \cos\pi h \sim 1- \frac{\pi^2h^2}{2} \tag{3.7}$$

for sufficiently small h.

From (3.3) and (3.5) we note that generally in this case the spectral radius of $D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}}^{(\pi_1)}$ is given by

$$S(D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}}^{(\pi_1)}) = \frac{1 - \sqrt{\frac{1-M^{(\pi_1)}}{2}}}{1 + 3\sqrt{\frac{1-M^{(\pi_1)}}{2}}} \sim 1-2\sqrt{2}\sqrt{1-M^{(\pi_1)}} \tag{3.8}$$

which by (3.6) yields

$$S(D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}}^{(\pi_1)}) \sim 1-2\sqrt{2}\pi h. \tag{3.9}$$

Therefore the rate of convergence of the LPSD method for h sufficiently small, is approximately the same with the line SOR method (LSOR). But if we have the additional restriction $\beta^{(\pi_1)} \leq \frac{M^{(\pi_1)}}{4}$ we note that, as in the point version, LPSD has an improved rate of convergence over LSOR. However, the additional work involved in the LPSD method even if the reduction scheme (A.11) (see Appendix A) is applied, probably do not justify the gain in the convergence thus making the method less attractive than LSOR.

Since it is known that (see Section 4.13)

$$S(D_{\tau_1,\omega_1}^{(\pi_0)}) \sim 1-2\pi h \tag{3.10}$$

we have a similar result to the SOR method that

$$\frac{R(D_{\tau_1(\pi_1),\omega_1(\pi_1)}^{(\pi_1)})}{R(D_{\tau_1,\omega_1}^{(\pi_0)})} \sim \sqrt{2} \tag{3.11}$$

for sufficiently small h.

In other words, for small h, the line preconditioned simultaneous displacement iterative method in the unit square, or a subset thereof yields an increase of approximately 40% in the rate of convergence over the point preconditioned simultaneous displacement method. Also, another conclusion we have reached here is that the improvement (3.11) in the ratios of rates of convergence is a fixed factor, independent of the mesh h, in contrast to the alternating direction methods of Chapter 7.

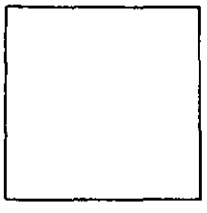
5.4 COMPUTATIONAL RESULTS

In order to test our theoretical results, the Laplace equation was solved in three different regions as shown in Figure 4.1. In each case, the unique solution was the vector $\bar{u}$ with all its components equal to zero while the initial guess was the vector $u^{(0)}$ with all its components equal to unity in the interior of the regions, with zero boundary values. The criterion used for convergence was again $\|u^{(n)}\|_\infty \leq 10^{-6}$.

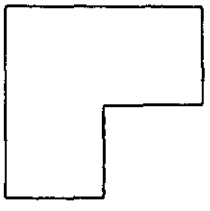
Although the problems considered here (and perhaps in other experiments in this thesis) were trivial only because of boundary conditions, the general behaviour of the iterative procedures could be expected to be typical of more complicated problems. The only change needed would be non-trivial boundary conditions.

The ordering considered in our experiments was the natural one as described in Section 3.6 for both the line and point PSD methods.

Region I: Unit square



Region II: Unit square with $\frac{1}{2}\times\frac{1}{2}$ square removed from one corner



Region III: Unit square with $\frac{4}{10}\times\frac{4}{10}$ square removed from center

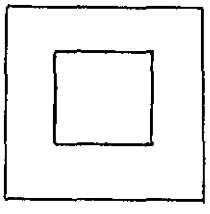


FIGURE 4.1

Region	h^{-1}	$\omega_0^{(\pi_1)}$	$\lambda(B_{\omega_0^{(\pi_1)}}^{(\pi_1)})$	$\Lambda(B_{\omega_0^{(\pi_1)}}^{(\pi_1)})$	$P(B_{\omega_0^{(\pi_1)}}^{(\pi_1)})$	$\tau_0^{(\pi_1)}$	$S(\xi_{\omega_0^{(\pi_1)}}^{(\pi_1)})$	LPSD	LSSOR	LPSD-SI†
I	20	1.7235	0.5667	2.0984	3.7026	0.7504	0.7299	26	45	14
	40	1.8487	0.5221	3.5752	6.8474	0.4881	0.8540	48	89	20
	60	1.8945	0.5006	5.0033	9.9941	0.3634	0.8999	71	134	25
II	20	1.5537	0.4699	1.4421	3.0693	1.0460	0.6742	22	36	12
	40	1.7496	0.4007	2.2826	5.6967	0.7454	0.8245	41	73	17
	60	1.7599	0.2638	2.3666	8.9720	0.7604	0.8885	65	121	23
III	20	1.3449	0.5172	1.1350	2.1945	1.2105	0.5443	15	21	9
	40	1.6267	0.3908	1.6468	4.2138	0.9816	0.7627	30	43	15
	60	1.7184	0.3569	2.0665	5.7907	0.8253	0.8273	41	64	17

TABLE 4.1

COMPARISON OF LPSD AND LSSOR $\|u^{(n)}\|_\infty \leq 10^{-6}$

†*For comparison reasons we also present the number of iterations using the semi-iterative line PSD method (SI-LPSD) (see (5.14))*

Table 4.1 contains the optimum values of the preconditioning parameter $\omega_0^{(\pi_1)}$, the acceleration parameter $\tau_0^{(\pi_1)}$, the maximum $\Lambda(B_{\omega_0^{(\pi_1)}}^{(\pi_1)})$, the minimum $\lambda(B_{\omega_0^{(\pi_1)}}^{(\pi_1)})$ eigenvalues and the P-condition number of the preconditioned matrix $B_{\omega_0^{(\pi_1)}}^{(\pi_1)}$, as well as the spectral radius of $\xi_{\omega_0^{(\pi_1)}}^{(\pi_1)}$. Also it contains the number of iterations of LPSD and LSSOR which were applied under the same conditions to solve the previously described problems for different values of the mesh size.

A study of Table 4.1 seems to imply that a monotonicity theorem for π_1 may be valid (and probably for any partition π). However, this remains to be proved. Also, one may notice immediately the confirmation of the fact that the LPSD method is asymptotically 2 times as effective as the LSSOR method in all the cases examined. Furthermore, although we predicted theoretically an improvement of about 40% in the rate of convergence of the LPSD over the point PSD (see (3.11)), the numerical results show (see Tables 4.1 and 4-13.1) that this gain is slightly greater for problem 1 in the unit square. In order to achieve this improvement in terms of overall computational effort one should carry out the method using a normalised block iteration scheme as described in Cuthill and Varga [1964].

From (3.9) we have that for line PSD and for any region we can find an $\omega_1^{(\pi_1)}$ such that the rate of convergence is O(h), hence one expects that the graph plot of log(N) versus $(\log h^{-1})$, where N is the number of iterations, to be a straight line with slope approximately unity. As previously, the slope α indicates $O(h^{\alpha})$ convergence rate. From Figure 4.2 we see that for all regions the rate of convergence is approximately O(h).

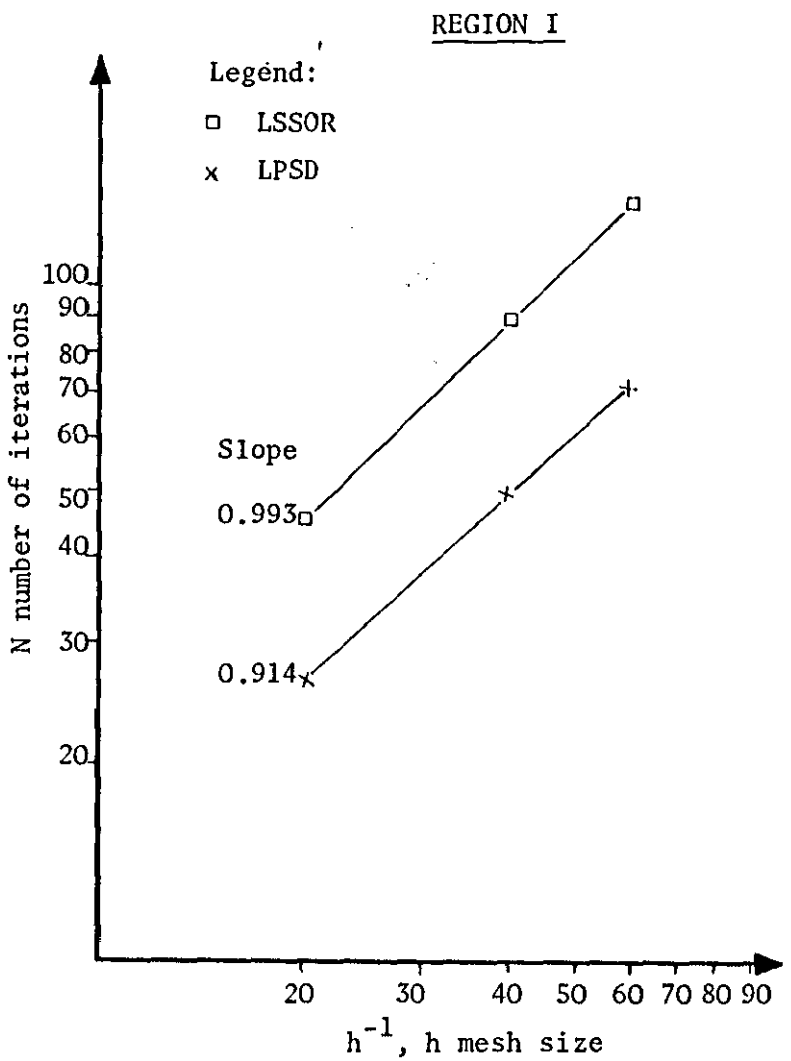
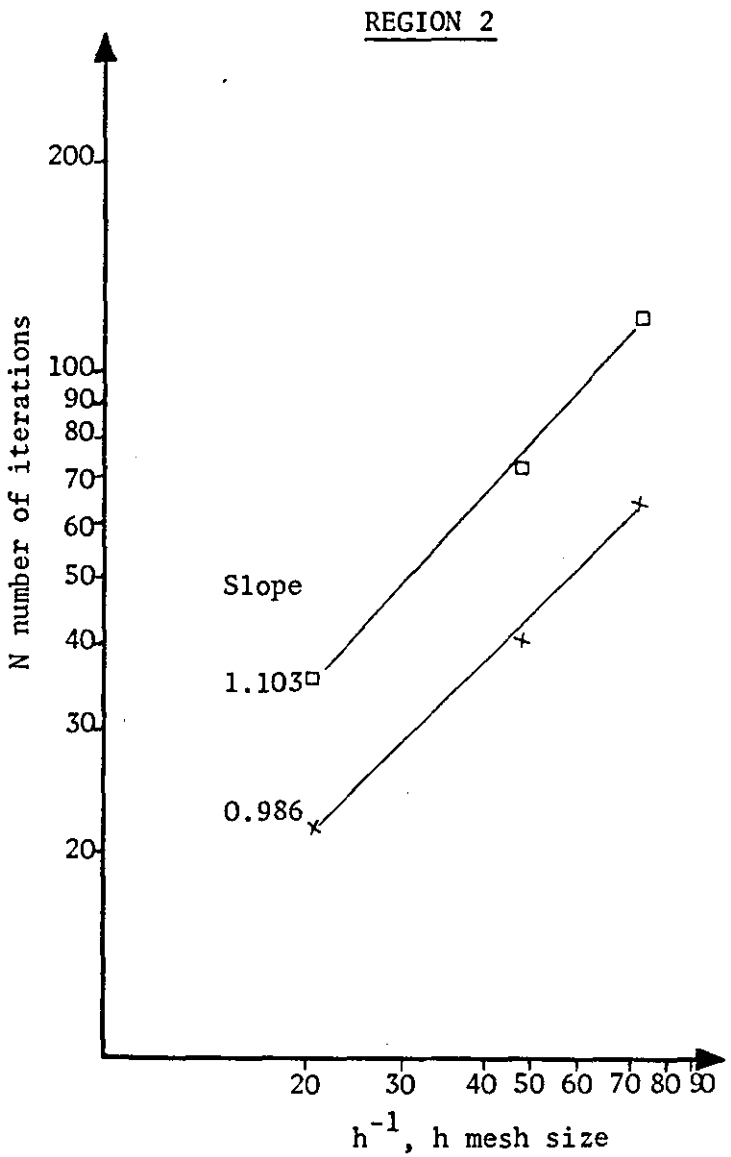


FIGURE 4.2

DETERMINATION OF RATE OF CONVERGENCE ATTAINED FOR REGIONS I, II AND III USING LPSD AND LSSOR WITH OPTIMUM PARAMETERS



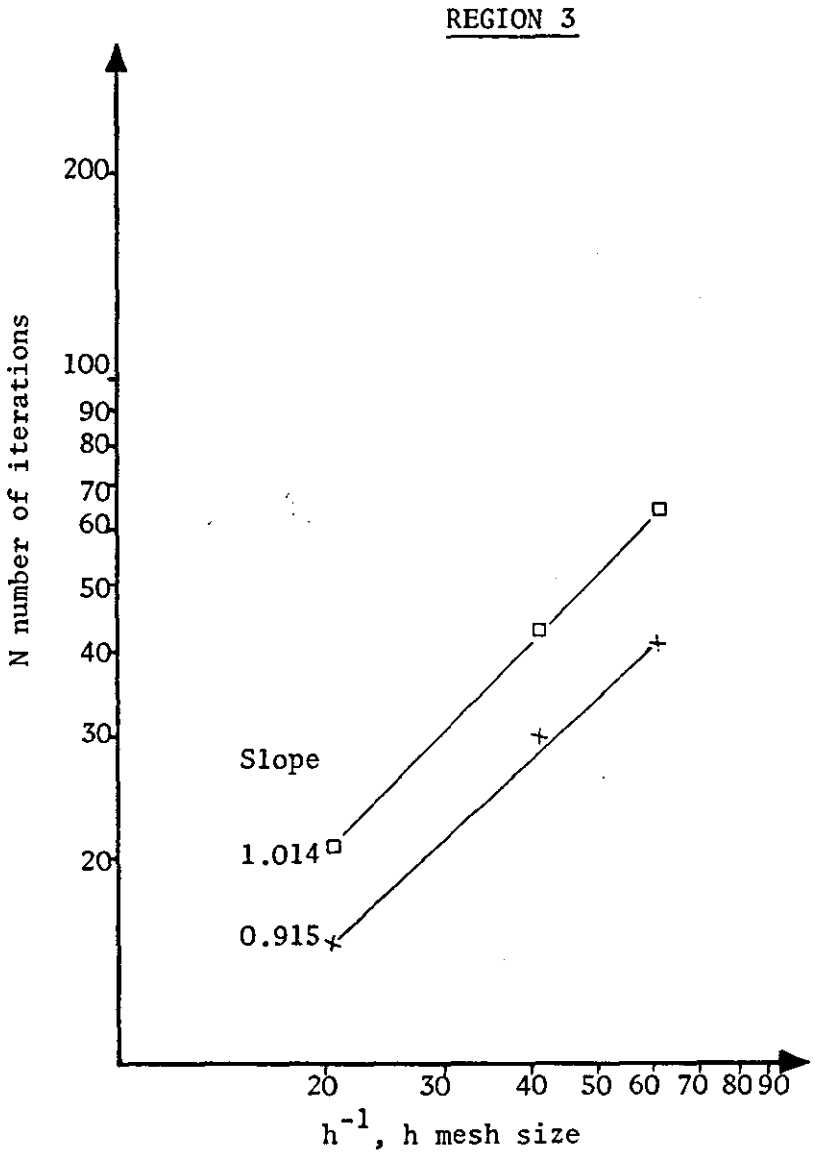


FIGURE 4.2 (CONTINUED)

SECTION B

ACCELERATED TECHNIQUES

5.5 PRECONDITIONED JACOBI-SEMI ITERATIVE METHOD (PJ-SI METHOD)

In Section 3.7 we showed how one can find a semi-iterative method (SI method) with respect to the linear stationary iterative process defined by

$$u^{(n+1)} = Gu^{(n)} + k \tag{5.1}$$

where the eigenvalues of G are real and lie in a certain interval. In the same section, we also considered the SI method as a two level acceleration procedure of (5.1).

The construction of the PJ method and its analysis in Sections 4.4 and 4.5 can be regarded as the first step of studying the behaviour and properties of a basic method of the form (5.1). On the other hand, the formulation and analysis of the PSD method is the second step which constitutes the first type of acceleration procedure similar to (3-7.25).

Next, we attempt to further accelerate the convergence of the PSD method by constructing the PJ-SI method.

As we have shown in Chapter 4, assuming the natural ordering of points and for a certain "good" choice of ω which depends on upper bounds of $S(B)$ and $S(LU)$, the rate of convergence of the PSD method is approximately $O(h)$. This rate of convergence is the same order of magnitude attained by the SOR method with optimum ω. Since a PSD iteration requires approximately twice the work involved in an SOR iteration, only with the Niethammer's scheme (A.11) we can consider PSD as being competitive with SOR in certain cases (see Section 4.12). The employment of this work-saving technique necessitates a more complicated program with greater storage requirements. On the other hand, the application of the PSD method with red-black ordering yields a convergence rate which differs by an order of magnitude from the natural ordering

(see Section 4.15).

Consequently, in order to establish the superiority over the SOR we need to consider the possibility of increasing the rate of convergence of the PSD method by an order of magnitude by means of semi-iteration (Varga [1957], Golub and Varga [1961]). This approach can be pursued for the PSD method since the eigenvalues of the iteration matrix $D_{\tau,\omega}$ are real, while this is precluded for SOR with optimum $\omega=\omega_b$ since the eigenvalues of L_{ω_b} are complex, although some progress has been made in accelerating SOR by semi-iteration for $\omega<\omega_b$ (Kincaid [1974]).

For the PJ method we recall that the iteration matrix is given by

$$\mathcal{H}_\omega = I-B_\omega \tag{5.2}$$

where

$$B_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A. \tag{5.3}$$

If A is positive definite, then as we have shown, all the eigenvalues of B_ω are positive and there exist positive numbers $\lambda(B_\omega)$ and $\Lambda(B_\omega)$ such that all eigenvalues λ of B_ω lie in the range

$$0<\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega). \tag{5.4}$$

Therefore, all the eigenvalues ν of $\mathcal{H}_\omega$ are real and lie in the range

$$\alpha = 1-\Lambda(B_\omega)\leq\nu\leq 1-\lambda(B_\omega) = \beta<1 \tag{5.5}$$

hence from (3-7.15) we have

$$z = \frac{P(B_\omega)+1}{P(B_\omega)-1}. \tag{5.6}$$

By (3-7.21) and (4-4.6) the formula for the optimum semi-iterative method based on PJ denoted by PJ-SI, is given by

$$\begin{aligned} u^{(n+1)} = \rho_{n+1}/(\lambda(B_\omega)+\Lambda(B_\omega))\{-[2(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \\ -(\lambda(B_\omega)+\Lambda(B_\omega))I]u^{(n)}+2(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}b\}+(1-\rho_{n+1})u^{(n-1)} \end{aligned} \tag{5.7}$$

or equivalently,

$$u^{(n+1)} = u^{(n-1)}+\rho_{n+1}(u^{(n)}-u^{(n-1)})+\rho_{n+1}\bar{\rho}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{5.8}$$

where

$$\bar{\rho} = \frac{2}{\lambda(B_\omega)+\Lambda(B_\omega)} \tag{5.9}$$

which by (4-11.19) becomes

$$\bar{\rho} = \frac{2\omega(2-\omega)}{1+1/P(B_\omega)} , \tag{5.10}$$

and

$$\begin{aligned} \rho_1 &= 1 , \\ \rho_2 &= \left(1 - \frac{\sigma^2}{2}\right)^{-1} , \\ \rho_{n+1} &= \left(1 - \frac{\sigma^2\rho_n}{4}\right)^{-1} , \quad n=2,3,\ldots \end{aligned} \tag{5.11}$$

where

$$\sigma = \frac{P(B_\omega)-1}{P(B_\omega)+1} . \tag{5.12}$$

From (5.8) we note that the PJ-SI method can also be written as

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)} + \rho_{n+1}[u^{(n)} + \bar{\rho}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)})] . \tag{5.13}$$

But the expression in the brackets is the PSD method since $\bar{\rho}=\tau_0$, thus a more compact form of the PJ-SI method is given by the following scheme

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)} + \rho_{n+1}(D_{\tau,\omega}u^{(n)} + \delta) . \tag{5.14}$$

From this observation we immediately conclude that (5.14) represents also the PSD-SI method. This can be more explicitly seen if we consider the range of the eigenvalues μ of $D_{\tau,\omega}$ which is

$$\alpha = 1-\tau\Lambda(B_\omega) \leq \mu \leq 1-\tau\lambda(B_\omega) = \beta < 1 \tag{5.15}$$

if $\tau>0$ and

$$\alpha = 1-\tau\Lambda(B_\omega) \geq \mu \geq 1-\tau\lambda(B_\omega) = \beta > 1 \tag{5.16}$$

if $\tau<0$. In either case, the SI method of Section 3.7 is applicable (see Young [1971]). It is easily verified that the formula for the PSD-SI method is independent of τ and it is identical to the one given by (5.14).

As can be seen from (5.14) the PJ-SI method is a linear non-stationary method of second degree. The improvement in convergence comes at the expense of requiring storage for one additional vector.

In order to determine the convergence of the PJ-SI method we have from (3-7.19) and (5.5) that

$$P_n(I-B_\omega) = T_n\left[\frac{2(I-B_\omega)-(2-\Lambda(B_\omega)-\lambda(B_\omega))}{\Lambda(B_\omega)-\lambda(B_\omega)}\right] \Big/ T_n\left[\frac{2-(2-\Lambda(B_\omega)-\lambda(B_\omega))}{\Lambda(B_\omega)-\lambda(B_\omega)}\right]$$

$$= T_n\left[\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2B_\omega}{\Lambda(B_\omega)-\lambda(B_\omega)}\right] \Big/ T_n\left[\frac{\Lambda(B_\omega)+\lambda(B_\omega)}{\Lambda(B_\omega)-\lambda(B_\omega)}\right] . \tag{5.17}$$

Thus the virtual spectral radius (see (3-7.27) and (3-7.28)) is given by

$$\bar{S}(P_n(\mathcal{H}_\omega)) = \frac{1}{T_n\left[\frac{P(B_\omega)+1}{P(B_\omega)-1}\right]} = \frac{2r^{n/2}}{1+r^n} = \frac{2\bar{r}^n}{1+\bar{r}^{2n}} \tag{5.18}$$

where by (3-7.29) and (5.12) we have

$$\bar{r} = r^{\frac{1}{2}} = \frac{\sigma}{1+\sqrt{1-\sigma^2}} = \frac{1-1/\sqrt{P(B_\omega)}}{1+1/\sqrt{P(B_\omega)}} . \tag{5.19}$$

Therefore, by (3-7.30) the average rate of convergence is defined by

$$R_n(P_n(\mathcal{H}_\omega)) = -\frac{1}{n}\log \bar{S}(P_n(\mathcal{H}_\omega)) = -\frac{1}{n}\log\frac{2\bar{r}^n}{1+\bar{r}^{2n}} \tag{5.20}$$

while as $n\to\infty$ we have by (3-7.31) that the rate of convergence for the PJ-SI method is given by

$$R_\infty(P_n(\mathcal{H}_\omega)) = -\frac{1}{2}\log r = -\log \bar{r}. \tag{5.21}$$

If now $P(B_\omega)>>1$, then from (5.19) we have

$$r = \left[\frac{1-1/\sqrt{P(B_\omega)}}{1+1/\sqrt{P(B_\omega)}}\right]^2 \sim 1 - \frac{4}{\sqrt{P(B_\omega)}} \tag{5.22}$$

hence

$$R_\infty(P_n(\mathcal{H}_\omega)) = -\frac{1}{2}\log r \sim \frac{2}{\sqrt{P(B_\omega)}} \sim \sqrt{2}\sqrt{R(D_{\tau,\omega})}. \tag{5.23}$$

The above result could also have been obtained immediately from (3-7.32).

If we express (5.23) in terms of reciprocal rates of convergence, then we obtain the following result

$$RR_\infty(P_n(\mathcal{H}_\omega)) \sim \frac{1}{\sqrt{2}}\sqrt{RR(D_{\tau,\omega})}. \tag{5.24}$$

By combining (5.24) and the results summarised in Table 4-12.1 we have the following comparisons between the PJ-SI and the JOR methods.

Range of $\bar{\beta}$	Asymptotic Bounds on $RR_\infty(P_n(\mathcal{H}_{\omega_1}))/(RR(B_{\bar{\omega}}))$	
	General Case	Property A
$\bar{\beta} \leq \frac{M}{4}$	$\frac{1}{2^{5/4}}$	$\frac{1}{2\sqrt{2}}$
$\frac{M}{4} \leq \bar{\beta} \leq \frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2^{5/4}}$
$\bar{\beta} > \frac{1}{4}$	$\frac{1}{2\sqrt{\gamma}}$	$\frac{1}{2^{5/4}\sqrt{\gamma}}$

TABLE 5.1

By (5.23) we see that the application of the semi-iterative techniques to the PJ method improves the rate of convergence by an order of magnitude over the PSD method. This is a substantial improvement and compares favourably with the frequently used SOR method as it (see Section 4.12) has approximately the same rate of convergence with the PSD method. A simple comparison between the asymptotic bounds on $RR_\infty(P_n(\mathcal{H}_{\omega_1}))$ and the best possible bound on $RR(L_{\omega_b})$ which is given by (3-6.26), when the matrix A is consistently ordered results in the construction of Table 5.2.

Range of $\bar{\beta}$	Asymptotic Bounds on $RR_\infty(P_n(\mathcal{H}_{\omega_1}))/\sqrt{RR(L_{\omega_b})}$
$\bar{\beta} \leq \frac{M}{4}$	$\frac{1}{2^{3/4}}$
$\frac{M}{4} \leq \bar{\beta} \leq \frac{1}{4}$	$\frac{1}{\sqrt{2}}$
$\bar{\beta} > \frac{1}{4}$	$\frac{1}{\sqrt{2\gamma}}$

TABLE 5.2

PROPERTY OF PJ-SI WITH SOR WHEN A HAS PROPERTY A

From Tables 5.2 and 5.1 we clearly see that for $0 \leq \bar{\beta} \leq \frac{1}{4}$ we have substantial improvements of the rate of convergence for the PJ-SI method over SOR, while for the case where $\bar{\beta} > \frac{1}{4}$ the gain on the convergence depends strongly upon the quantity γ which is given by the expression (see (4-11.18))

$$\gamma = \left(1 + \frac{2(\bar{\beta}-1/4)}{1-M}\right)^{-\frac{1}{2}}. \tag{5.25}$$

Young [1974] proved that

$$S(LU) \leq \frac{1}{4} + O(h^2) \tag{5.26}$$

for the discrete generalised Dirichlet problem assuming the mesh size h to be small and that the coefficients A(x,y) and C(x,y) in (1-2.3) belong to class $C^{(2)}$ in $RU\partial R$*. This result is significant because it establishes an order of magnitude improvement of the PJ-SI method over the SOR and PSD methods.

Young [1971,1971a] has also shown for the generalised Dirichlet problem that

$$S(B) \leq \frac{2(\bar{A}+\bar{C})}{2(\bar{A}+\bar{C})+h^2(-\underline{F})}\left\{1 - \frac{2\underline{A}\sin^2\frac{\pi}{2I}+2\underline{C}\sin^2\frac{\pi}{2J}}{\frac{1}{2}(\bar{A}+\underline{A})+\frac{1}{2}(\bar{C}+\underline{C})+\frac{1}{2}(\bar{A}-\underline{A})\cos\frac{\pi}{I}+\frac{1}{2}(\bar{C}-\underline{C})\cos\frac{\pi}{J}}\right\} \tag{5.27}$$

where the region R is included in an $Ih \times Jh$ rectangle for some positive integers I and J and where

$$\underline{A} \leq A(x,y) \leq \bar{A}, \quad \underline{C} \leq C(x,y) \leq \bar{C}, \quad (-\underline{F}) \leq -F(x,y)$$

in $RU\partial R$.

We note that (5.27) implies that

$$S(B) \leq M = 1 - ch^2 + O(h^4) \tag{5.28}$$

for some constant $c > 0$. By (5.26) and letting $\bar{\beta} = 1/4 + O(h^2)$, it follows that

$$\frac{\bar{\beta}-1/4}{1-M} = \frac{O(h^2)}{ch^2+O(h^4)} \tag{5.29}$$

and

$$\lim_{h\to 0} \frac{\bar{\beta}-1/4}{1-M} = \xi_0 > 0$$

---

*We note that Ehrlich [1963,1964] showed that $S(LU) \leq \frac{1}{4}$ for the model problem whereas Phien [1972] showed that this condition holds for the equation $(y^{-1}U_x)_x + (y^{-1}U_y)_y = 0$ as well.

hence

$$\gamma > \xi_1 > 0 \tag{5.30}$$

where

$$\xi_1 = (1+2\xi_0)^{-\frac{1}{2}}. \tag{5.31}$$

Therefore, the quantity γ^{-1}, where γ is given by (5.25) is bounded away from zero as $h \to 0$. On the other hand, for $\bar{\beta} > \frac{1}{4}$ we recall from (4-11.17) that

$$P(B_{\omega_1}) \leq \frac{1}{2}(1+\gamma^{-1}\sqrt{\frac{2}{1-M}}) \tag{5.32}$$

and using (5.28) we obtain

$$P(B_{\omega_1}) \sim \frac{1}{2}\left(1+\frac{\sqrt{2/c}}{\gamma h}\right) = \frac{1}{2}(1+\frac{1}{\xi_2 h}) \tag{5.33}$$

where

$$\xi_2 = \gamma\sqrt{2/c}\,.$$

Evidently, from (5.33) it follows that

$$R(D_{\tau_1,\omega_1}) \sim 2/P(B_{\omega_1}) \sim 4/(1+(\xi_2 h)^{-1}) \sim O(h) \tag{5.34}$$

and finally by (5.25) we obtain the expected result

$$R_\infty(P_n(\mathcal{H}_{\omega_1}) \sim \sqrt{2}\sqrt{R(D_{\tau_1,\omega_1})} \sim O(h^{\frac{1}{2}}). \tag{5.35}$$

Consequently, for the PJ-SI method to yield an order-of-magnitude improvement on the convergence rate over PSD and SOR it is sufficient that $\bar{\beta}$-1/4 be of the same order-of-magnitude as 1-M. This condition has been shown to hold for the generalised Dirichlet problem under the condition that A(x,y) and C(x,y) are in the class $C^{(2)}$ in the region of consideration. However, of we consider the self-adjoint equation

$$\frac{\partial}{\partial x}(A\frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(C\frac{\partial U}{\partial y}) = G \tag{5.36}$$

where $|A_x|$ and $|C_y|$ are bounded in the domain of consideration, then the application of the PJ-SI method to the corresponding difference equation yields a rate of convergence of $O(h^{\frac{3}{4}})$. This is proved if we determine S(LU) as follows. We note that

$$S(LU) \leq \|LU\|_\infty \leq \|L\|_\infty \|U\|_\infty \tag{5.37}$$

which implies that we have to estimate $\|U\|_\infty$. But the sum of the elements

of the matrix U in the row corresponding to the point (x,y), by (1-2.7) and (1-2.8) is

$$\left(\frac{1}{A(x,y)+C(x,y)+O(h^2)}\right)\left[A(x+\frac{h}{2},y)+C(x,y+\frac{h}{2})\right]$$

$$= \frac{1}{2} + \frac{h}{4}\left(\frac{A_x+C_y}{A+C}\right) + O(h^2) \tag{5.38}$$

thus for h sufficiently small we have

$$\|U\|_\infty \leq \frac{1}{2} + \xi h \tag{5.39}$$

and by (5.37) we obtain

$$S(LU) \leq \frac{1}{4} + \xi' h \tag{5.40}$$

for some constants ξ and ξ'.

In this case, by (4-11.17) we have

$$P(B_{\omega_1}) \leq \frac{1}{2}\left(1+\frac{\sqrt{1-2M+4\bar{\beta}}}{1-M}\right) \sim \frac{1}{2}(1+\xi'' h^{-3/2}) \tag{5.41}$$

and therefore by applying the PSD method we have

$$R(\mathcal{D}_{\tau_1,\omega_1}) \sim 2/P(B_{\omega_1}) \sim O(h^{3/2}). \tag{5.42}$$

Finally, by using semi-iteration we obtain by (5.35) the following result

$$R_\infty(P_n(\mathcal{H}_{\omega_1})) \sim O(h^{\frac{3}{4}}) \tag{5.43}$$

which indicates again an order of magnitude improvement in the convergence rate.

5.6 PRECONDITIONED JACOBI-VARIABLE EXTRAPOLATION METHOD (PJ-VE METHOD)

We recall from Section 4.9 that the PSD method is defined by the iterative scheme

$$u^{(n+1)} = \tau(\mathcal{H}_\omega u^{(n)} + \gamma_\omega) + (1-\tau)u^{(n)}. \tag{6.1}$$

By comparing (6.1) and (3-7.25) we see that if we allow τ to vary in each iteration we immediately have the variable extrapolation (see Section 3.8) version of the PJ method which is

$$u^{(n+1)} = \theta_{n+1}(\mathcal{H}_\omega u^{(n)} + \gamma_\omega) + (1-\theta_{n+1})u^{(n)}. \tag{6.2}$$

The iterative procedure (6.2) defines the PJ-varaible extrapolation method (PJ-VE method). We note that the first expression in brackets of the right hand side in (6.2) is the PJ method thus by using (4-4.6), the expression in (6.2) becomes

$$u^{(n+1)} = \theta_{n+1}[u^{(n)} + (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)})] + (1-\theta_{n+1})u^{(n)} \tag{6.3}$$

which can be simplified to yield the iterative process

$$u^{(n+1)} = u^{(n)} + \theta_{n+1}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}). \tag{6.4}$$

The iteration parameters θ_{n+1} can be determined by using (3-8.5) and (5.5). Consequently, the parameters θ_k for the variable extrapolation as applied to the PJ method are given by

$$\theta_k = \frac{2}{(\lambda(B_\omega)-\Lambda(B_\omega))\cos\frac{(2k-1)\pi}{2m} + (\lambda(B_\omega)+\Lambda(B_\omega))}, \quad k=1,2,\ldots,m \tag{6.5}$$

or by

$$\theta_k = \frac{\omega(2-\omega)}{1/P(B_\omega)\cos^2\frac{(2k-1)\pi}{4m} + \sin^2\frac{(2k-1)\pi}{4m}}, \quad k=1,2,\ldots,m. \tag{6.6}$$

Using the optimum θ_k as given by (6.6) we can see by (3-8.6) that the virtual spectral radius of the PJ-VE method depends upon the P-condition number of B_ω, whereas we do not expect the virtual rate of convergence to be as effective as the PJ-SI method since care must be taken with the values of m (see Young [1954a].

5.7 SECOND DEGREE-PRECONDITIONED JACOBI METHOD (SD-PJ METHOD)

Instead of using the non-stationary methods as described in the previous section one can obtain almost as rapid convergence using the stationary second degree version of the PJ method (see Section 3.9). The second degree PJ method can be easily obtained, if we let $\rho_1=1$ as for the SI-PJ method, but for $n\geq 2$ we let $\rho_n=\hat{\omega}_0$ where $\hat{\omega}_0$ is given by (3-9.15). Evidently, $\hat{\omega}_0$ is the limit of the sequence $\rho_1,\rho_2,\ldots$ as defined in the SI-PJ method.

If now A is positive definite and if $G=\mathcal{H}_\omega$, corresponding to the PJ method, then from (5.5) we have

$$\sigma = \frac{P(B_\omega)-1}{P(B_\omega)+1} . \tag{7.1}$$

Therefore

$$\hat{\omega}_0 = \frac{2}{1+\sqrt{1-\sigma^2}} = 1 + \left(\frac{\sqrt{P(B_\omega)}-1}{\sqrt{P(B_\omega)}+1}\right)^2 \tag{7.2}$$

and by (3-9.17) we have the result

$$S(M) = (\hat{\omega}_0-1)^{\frac{1}{2}} = \frac{\sqrt{P(B_\omega)}-1}{\sqrt{P(B_\omega)}+1} . \tag{7.3}$$

If $P(B_\omega)$ is very large, as is frequently the case, then the second degree PJ method converges much faster than the SOR method. After the determination of $\hat{\omega}_0$ we can easily find ξ_0 and η_0 by (3-9.14) and (3-9.16), respectively, hence the SD-PJ method with A positive definite is defined by (see (3-9.2))

$$u^{(n+1)} = u^{(n)}+(\hat{\omega}_0-1)(u^{(n)}-u^{(n-1)})+\tau_0\hat{\omega}_0(\mathcal{H}_\omega u^{(n)}+\gamma_\omega-u^{(n)}) \tag{7.4}$$

where $\hat{\omega}_0$ is given by (7.2) and

$$\tau_0 = \frac{2\omega(2-\omega)}{1+1/P(B_\omega)} . \tag{7.5}$$

We simplify (7.4) to obtain successively

$$u^{(n+1)} = u^{(n)}+(\hat{\omega}_0-1)(u^{(n)}-u^{(n-1)})+\tau_0\hat{\omega}_0(\gamma_\omega-B_\omega u^{(n)})$$

or

$$u^{(n+1)} = \hat{\omega}_0[u^{(n)}+\tau_0(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)})]+(1-\hat{\omega}_0)u^{(n-1)} \quad (7.6)$$

which can be written in the more compact form

$$u^{(n+1)} = \hat{\omega}_0(D_{\tau_0,\omega}u^{(n)}+\delta) + (1-\hat{\omega}_0)u^{(n-1)} \quad (7.7)$$

where the expression in the first brackets in the right hand side can be easily recognised to be the PSD method.

For a direct comparison with the PJ-SI method, we specify the iterant $u^{(1)}$ to be given by a PSD iteration, thus we finally define the SD-PJ method by (7.7), where

$$\hat{\omega}_0 = \begin{cases} 1 & , \text{ if } n=0 \\ \dfrac{2}{1+\sqrt{1-\sigma^2}}, & \text{if } n\geq 1 \end{cases} \quad (7.8)$$

or more analytically

$$\begin{aligned} u^{(1)} &= D_{\tau_0,\omega}u^{(0)}+\delta, \\ u^{(n+1)} &= \hat{\omega}_0(D_{\tau_0,\omega}u^{(n)}+\delta) + (1-\hat{\omega}_0)u^{(n-1)}. \end{aligned} \quad (7.9)$$

In this case we are able to determine the virtual spectral radius of the SD-PJ method and it can be proved (see Young [1971] p.490-491) that

$$\bar{S}(Q_n(D_{\tau_0,\omega})) = \frac{2\hat{r}^{n/2}}{1+\hat{r}}\left[1+\left(\frac{n-1}{2}\right)(1-\hat{r})\right] \quad (7.10)$$

where

$$\hat{r} = \hat{\omega}_0-1 \quad (7.11)$$

and the polynomials $Q_n(D_{\tau_0,\omega})$ satisfy the recurrence relation

$$\begin{aligned} Q_0(D_{\tau_0,\omega}) &= I, \qquad Q_1(D_{\tau_0,\omega}) = D_{\tau_0,\omega} \\ Q_{n+1}(D_{\tau_0,\omega}) &= \hat{\omega}_0 D_{\tau_0,\omega}Q_n(D_{\tau_0,\omega})+(1-\hat{\omega}_0)Q_{n-1}(D_{\tau_0,\omega}). \end{aligned} \quad (7.12)$$

By recalling (5.19), the virtual spectral radius for the PJ-SI method is given by

$$\bar{S}(P_n(\mathcal{H}_\omega)) = \frac{2\hat{r}^{n/2}}{1+\hat{r}^n} \quad (7.13)$$

thus by the theory of Chebyshev polynomials we have that

$$\bar{S}(P_n(\mathcal{H}_\omega)) \leq \bar{S}(Q_n(D_{\tau_0,\omega})) \tag{7.14}$$

which implies that the PJ-SI method converges faster than the SD-PJ method.

5.8 GENERALISED CONJUGATE GRADIENT METHOD

In Section 3.10 we showed how the CG method can be regarded as an acceleration procedure analogous to the SI method. In this section we will consider the above idea in more detail with particular reference to the basic iterative method (5.1) of a certain form. This will help us to apply the CG method to the PJ method in order to produce a powerful iterative scheme (see Section 5.9).

Let us consider the basic iterative method of the form

$$u^{(n+1)} = Gu^{(n)}+k \tag{8.1}$$

where

$$k = (I-G)A^{-1}b. \tag{8.2}$$

Further, we make the assumption that the iteration matrix G has the form

$$G = I-R^{-1}A \tag{8.3}$$

where the conditioning matrix R is the product of a matrix times its transpose i.e.,

$$R = QQ^{T} \tag{8.4}$$

and Q is a non-singular matrix. Evidently, the matrix G has real eigenvalues since it is similar to the symmetric matrix

$$\tilde{G} = Q^{T}G(Q^{T})^{-1} = I-Q^{-1}A(Q^{T})^{-1}. \tag{8.5}$$

Next, we will develop a version of the conjugate gradient procedure with respect to the basic method (8.1) in a way similar to the one followed for the SI method.

Let us consider the original form of the preconditioned system (as was first introduced by Evans [1968]),

$$[Q^{-1}A(Q^{T})^{-1}](Q^{T}u) = Q^{-1}b. \tag{8.6}$$

System (8.6) can be written in the following compact form

$$\hat{A}\hat{u} = \hat{b} \tag{8.7}$$

where

$$\hat{A} = Q^{-1}A(Q^{T})^{-1}, \tag{8.8}$$

$$\hat{u} = Q^{T}u \tag{8.9}$$

and

$$\hat{b} = Q^{-1}b. \tag{8.10}$$

It can be readily seen that $\hat{A}$ is symmetric and

$$\hat{A} = Q^T(I-G)(Q^{-1})^T. \tag{8.11}$$

We now consider the application of the CG method to the preconditioned system (8.7). If we use the non-stationary second degree version as developed in Section 3.10, then by (3-10.37) we have the iterative scheme

$$\hat{u}^{(n+1)} = \rho_{n+1}(\hat{u}^{(n)}+\gamma_{n+1}\hat{r}^{(n)}) + (1-\rho_{n+1})\hat{u}^{(n-1)} \tag{8.12}$$

where

$$\hat{r}^{(n)} = \hat{b}-\hat{A}u^{(n)} = Q^{-1}r^{(n)} \tag{8.13}$$

and

$$r^{(n)} = b-Au^{(n)}. \tag{8.14}$$

Using the relationships (8.9) and (8.13) we rewrite (8.12) to yield

$$u^{(n+1)} = \rho_{n+1}(u^{(n)}+\gamma_{n+1}(Q^T)^{-1}Q^{-1}r^{(n)}) + (1-\rho_{n+1})u^{(n-1)} \tag{8.15}$$

which by noting that

$$(Q^T)^{-1}Q^{-1}r^{(n)} = (Q^T)^{-1}Q^{-1}(b-Au^{(n)}) = Gu^{(n)}+k-u^{(n)} \tag{8.16}$$

can be written in the following compact form

$$u^{(n+1)} = \rho_{n+1}[\gamma_{n+1}(Gu^{(n)}+k)+(1-\gamma_{n+1})u^{(n)}]+(1-\rho_{n+1})u^{(n-1)}. \tag{8.17}$$

By turning our attention to the expressions for the parameters ρ_{n+1} and γ_{n+1} we have from (3-10.39) that

$$\gamma_{n+1} = \frac{(\hat{r}^{(n)},\hat{r}^{(n)})}{(\hat{r}^{(n)},\hat{A}\hat{r}^{(n)})} = \frac{(\hat{r}^{(n)},\hat{r}^{(n)})}{(\hat{r}^{(n)},Q^{-1}A(Q^{-1})^T\hat{r}^{(n)})}$$

$$= \frac{(\hat{r}^{(n)},\hat{r}^{(n)})}{((Q^{-1})^T\hat{r}^{(n)},A(Q^{-1})^T\hat{r}^{(n)})} = \frac{(\hat{r}^{(n)},\hat{r}^{(n)})}{(\hat{s}^{(n)},A\hat{s}^{(n)})}. \tag{8.18}$$

where

$$\hat{s}^{(n)} = (Q^{-1})^T\hat{r}^{(n)} = (Q^{-1})^TQ^{-1}r^{(n)} = Gu^{(n)}+k-u^{(n)}. \tag{8.19}$$

Finally, from (3-10.38) we have the following expression for ρ_{n+1}

$$\rho_{n+1} = \left[1-\frac{\gamma_{n+1}}{\gamma_n}\frac{(\hat{r}^{(n)},\hat{r}^{(n)})}{(\hat{r}^{(n-1)},\hat{r}^{(n-1)})}\frac{1}{\rho_n}\right]^{-1} \quad n=1,2,\ldots . \tag{8.20}$$

Evidently, the relationships (8.17), (8.18) and (8.20) define the CG method with respect to the iterative scheme (8.1).

5.9 PRECONDITIONED JACOBI-CONJUGATE GRADIENT METHOD (PJ-CG METHOD)

In this section, we present an alternative acceleration procedure of the PJ method which results in a more effective iterative scheme (as far as the rate of convergence is concerned) than the PJ-SI method developed in Section 5.5.

The combination of the PJ method with the application of CG method has also been considered by Evans [1973a] whereas similar accelerated schemes have been developed by Axelsson [1974] and Young [1975] for the SSOR method.

Let us assume, without loss of generality (see Young [1971] p.112), that the matrix A has the splitting

$$A = I-L-U \tag{9.1}$$

where

$$L = U^T \tag{9.2}$$

and L,U are strictly lower triangular and strictly upper triangular matrices, respectively.

We note that if we let Q be the matrix

$$Q = I-\omega L \ , \tag{9.3}$$

the PJ method can take the form (8.1) where the matrix R is given by (8.4).

From (9.3) and (8.13) we immediately have

$$\hat{r}^{(n)} = Q^{-1}r^{(n)} = (I-\omega L)^{-1}r^{(n)} \tag{9.4}$$

whereas from (8.19) we obtain the following expression for $\hat{s}^{(n)}$

$$\hat{s}^{(n)} = (I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)} . \tag{9.5}$$

We recall from Chapter 4 that the PJ method is given by the iterative process

$$u^{(n+1)} = u^{(n)}+(I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)} \tag{9.6}$$

and has the form (8.1). If we substitute (9.6) in (8.17) we obtain

$$u^{(n+1)} = \rho_{n+1}[\gamma_{n+1}(u^{(n)}+(I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)})+(1-\gamma_{n+1})u^{(n)}] +(1-\rho_{n+1})u^{(n-1)} \tag{9.7}$$

or

$$u^{(n+1)} = \rho_{n+1}(u^{(n)}+\gamma_{n+1}(I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)})+(1-\rho_{n+1})u^{(n-1)} \tag{9.8}$$

which defines the PJ-CG method.

The parameters ρ_{n+1} and γ_{n+1} are determined by using (8.18) and (8.20), hence we have

$$\gamma_{n+1} = \frac{((I-\omega L)^{-1}r^{(n)},(I-\omega L)^{-1}r^{(n)})}{((I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)},A(I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)})}$$

$$= \frac{(r^{(n)},\tilde{s}^{(n)})}{(\tilde{s}^{(n)},A\tilde{s}^{(n)})} \tag{9.9}$$

where

$$\tilde{s}^{(n)} = (I-\omega U)^{-1}(I-\omega L)^{-1}r^{(n)} \tag{9.10}$$

and in a similar way we find

$$\rho_{n+1} = \left[1 - \frac{\gamma_{n+1}}{\gamma_n}\,\frac{(r^{(n)},r^{(n)})}{(r^{(n-1)},r^{(n-1)})}\,\frac{1}{\rho_n}\right]^{-1}. \tag{9.11}$$

Summarising our results we have that the PJ-CG method is defined by

$$u^{(n+1)} = u^{(n-1)} + \rho_{n+1}(u^{(n)}-u^{(n-1)}) + \rho_{n+1}\gamma_{n+1}(I-\omega U)^{-1}(I-\omega L)^{-1}(b-Au^{(n)}) \tag{9.12}$$

where $\rho_1=1$ and ρ_{n+1} is given by (9.11) whereas γ_{n+1} can be obtained from (9.9).

In order to examine the average rate of convergence of the PJ-CG method we recall from (3-10.40) and (3-10.42) that the average rate of convergence is expected to be better than the PJ-SI method in the sense of minimising the $A^{\frac{1}{2}}$-norm of the error vector. Consequently, we expect that the number of iterations will behave like $O(h^{-\frac{1}{2}})$ in the PJ-CG method as well.

5.10 COMPARISONS AND COMPUTATIONAL RESULTS

In the previous sections we have developed various accelerated procedures based on the PJ method which resulted in an order of magnitude improvement of the rate of convergence as compared with the SOR method (and the PSD method). These comparisons were based on the fact that the preconditioning parameter ω takes its optimum value ω_0. However, in practice we use a value of ω which is near its optimum ω_0. This value is given by (4-11.4) and it requires the determination of the quantities $\bar{\beta}$ and M.

Since for a given linear system we may have some difficulties to estimate the quantity $\bar{\beta}$, we consider the effectiveness of the PJ-SI method with $\omega=1$.

Next, we prove a lemma which establishes the effectiveness of the PJ-SI method even with $\omega=1$.

Lemma 10.1

If A is a positive definite L-matrix then

$$P(B_1) \leq P(A). \tag{10.1}$$

Proof

If A is a positive definite L-matrix, then it follows (see (3- 6.38)) that $M(B) \geq -m(B)$. Furthermore, from (4-11.2) we have the following expression for $P(B_1)$

$$P(B_{\omega=1}) = P(B_1) \leq \frac{1-S(B)+S(LU)}{1-S(B)} . \tag{10.2}$$

Since now

$$B^2 = (L+U)^2 = LU+UL+L^2+U^2 \tag{10.3}$$

we have

$$LU = B^2-UL-L^2-U^2 \leq B^2 \tag{10.4}$$

and from Theorem 2-1.3 it follows that

$$S(LU) \leq S(B)^2. \tag{10.5}$$

If we combine (10.2) and (10.5) we have the alternative bound

$$P(B_1) \leq \frac{1-S(B)+S(B)^2}{1-S(B)} . \tag{10.6}$$

On the other hand, we have

$$1-S(B)+S(B)^2 = 1-S(B)(1-S(B)) \leq 1$$

hence (10.6) yields

$$P(B_1) \leq \frac{1}{1-S(B)} \leq P(A) \tag{10.7}$$

and the proof of the lemma is complete.

From the above lemma and from the fact that the effectiveness of the SI method (see 3-7.32) depends upon the P-condition number of the preconditioned matrix, we conclude that if A is a positive definite L-matrix, then the PJ-SI method with $\omega=1$ is at least as effective as the Jacobi-SI method.

If we further assume that the matrix A has Property A, then the rate of convergence for the PSD method with $\omega=1$ is given by

$$R(D_{\tau_0,1}) \sim 2/P(B_1) \tag{10.8}$$

which using (10.7) and (3-2.14) yields

$$\begin{aligned} R(D_{\tau_0,1}) &\sim 2(1-S(B)) \sim 2(-\log S(B)) \\ &= 2(-\log S(B_\omega)) = 2R(B_\omega). \end{aligned} \tag{10.9}$$

Consequently, from (5.24) the rate of convergence for the PJ-SI method with $\omega=1$ is

$$R_\infty(P_n(\mathcal{H}_1)) \sim \sqrt{2}\sqrt{R(D_{\tau_0,1})} \sim 2\sqrt{R(B_\omega)}. \tag{10.10}$$

On the other hand, the rate of convergence for the J-SI method is

$$R_\infty(P_n(B)) \sim \sqrt{2}\sqrt{R(B)} = \sqrt{2}\sqrt{R(B_\omega)} \tag{10.11}$$

hence

$$R_\infty(P_n(\mathcal{H}_1)) \sim \sqrt{2}\, R_\infty(P_n(B)) \tag{10.12}$$

for sufficiently small h.

Further, if A is consistently ordered, then it is known (see (3-6.32)) that

$$R(L_{\omega_b}) \sim 2\sqrt{2}\sqrt{R(B_\omega)} \tag{10.13}$$

hence by combining (10.13) and (10.10) we find

$$\frac{R_\infty(P_n(\mathcal{H}_1))}{R(L_{\omega_b})} \sim \frac{\sqrt{2}}{2} \tag{10.14}$$

which implies that if A is consistently ordered, then the PJ-SI method with $\omega=1$ is asymptotically $\sqrt{2}/2$ times as effective as the SOR method. From the above analysis, we conclude that if A is a positive define L-matrix, then it would seem appropriate to use the PJ-SI method as opposed to the J-SI method, whereas if A is consistently ordered, as opposed to the SOR method. Thus, even with $\omega=1$ and even taking into account the extra work per iteration, the PJ-SI method is nearly as effective as the other methods.

Let us now return to the case where the preconditioning parameter takes the value given by (4-11.4) and consider the application of the previous results to the model problem as described in Section 4.13. Specifically we will consider the application of the SI techniques to the J and GS methods and compare their rates of convergence with PJ-SI, SOR and other iterative procedures as well.

We recall from (4-13.9) that for the model problem we have

$$R(D_{\tau_1,\omega_1}) \sim 4/(1+\frac{\sqrt{3}}{\pi}h^{-1}) \sim \frac{4\pi}{\sqrt{3}}\,h \tag{10.15}$$

for sufficiently small h.

Thus from (5.24) we have that the reciprocal rate of convergence for the PJ-SI method is

$$RR_\infty(P_n(\mathcal{H}_{\omega_1})) \sim \frac{3^{\frac{1}{4}}}{2\sqrt{2\pi}}\,h^{-\frac{1}{2}} \tag{10.16}$$

which is better than the value of $RR(L_{\omega_b})$ by an order-of-magnitude. In fact, from (4-13.15) we have

$$\rho = \frac{RR(L_{\omega_b})}{RR_\infty(P_n(\mathcal{H}_{\omega_1}))} \sim \sqrt{\frac{2}{\sqrt{3\pi}}}\,h^{-\frac{1}{2}} = 0.606h^{-\frac{1}{2}}. \tag{10.17}$$

If we now compute the values of ρ for h=1/20, 1/40, 1/60, 1/80, then we have the corresponding values of the ratio of the asymptotic bounds on the reciprocal rates of convergence of the SOR method over the PJ-SI method presented in Table 10.1. From this table, we observe that the PJ-SI represents a substantial saving over the SOR method even if one

counts each PJ-SI iteration as two full SOR iterations. We also note that the factor of saving increases as the mesh size h decreases so we expect further increases as $h \to 0$.

h^{-1}	ρ	$\rho/2$
20	2.71	1.36
40	3.83	1.92
60	4.69	2.35
80	5.42	2.71

TABLE 10.1

We recall again from (4-13.3) that for the JOR method we have

$$RR(B_{\bar{\omega}}) = \frac{1}{-\log \cos \pi h} \sim \frac{2}{\pi^2} h^{-2} \tag{10.18}$$

for sufficiently small h. Thus by (3-7.32) the reciprocal asymptotic rate of convergence of the J-SI method is

$$RR_\infty(P_n(B)) \sim \frac{1}{\sqrt{2}} \sqrt{RR(B_{\bar{\omega}})} \sim \frac{1}{\pi} h^{-1}. \tag{10.19}$$

On the other hand, it is known (see Young [1971], Golub and Varga [1961]) that the Cyclic Chebyshev Semi-Iterative method (CCSI method) has twice as fast the rate of convergence of the J-SI and therefore from (10.19) we obtain

$$RR_\infty(CCSI) \sim \frac{1}{2\pi} h^{-1} \tag{10.20}$$

which is the same as the value of $RR(\mathcal{L}_{\omega_b})$ (see (4-13.15)).

Moreover, for the GS method, since A is consistently ordered, we have (see Chapter 4) that

$$S(\mathcal{L}) = S(B)^2 = \cos^2 \pi h \tag{10.21}$$

thus we obtain

$$RR(\mathcal{L}) \sim \frac{1}{\pi^2} h^{-2}. \tag{10.22}$$

Finally for the GS-SI method we have the following result

$$RR_\infty(P_n(\mathcal{L})) \sim \frac{1}{2\sqrt{2}} \sqrt{RR(B_{\bar{\omega}})} \sim \frac{1}{2\pi} h^{-1} \tag{10.23}$$

which implies that the convergence rate of the GS-SI method is approximately twice as fast as the J-SI method. However, for stability reasons (see Young

[1971]), before using the GS-SI method, one should first permute the rows and corresponding columns of A so as to obtain the form (2-7.1). Evidently, for the model problem this corresponds to the relabelling of the interior mesh points to correspond to the "red-black" ordering (see Chapter 2).

Let us now consider the effectiveness of the SSOR-SI method (see Habelter and Wachspress [1961], Sheldon [1955], Young [1974, 1971]). From the simple observation that the rates of convergence of the accelerated iterative schemes considered so far, depend also upon the P-condition number of the (preconditioned) coefficient matrix, we expect the SSOR-SI method to possess approximately the same rate of convergence as the PJ-SI method. This observation is concluded from the fact that the two methods have conditioning matrix which differ only by a scalar factor. This can be more explicitly seen if we consider the first step of acceleration (see (3-7.25)) applied to the SSOR method, hence we have

$$u^{(n+1)} = (I-\hat{\tau}\omega(2-\omega)B_\omega)u^{(n)}+\hat{\tau}\omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}c \tag{10.24}$$

or

$$u^{(n+1)} = (I-\tau B_\omega)u^{(n)}+\tau(I-\omega U)^{-1}(I-\omega L)^{-1}c \tag{10.25}$$

where

$$\tilde{\tau} = \hat{\tau}\omega(2-\omega). \tag{10.26}$$

Evidently, in order for the rate of convergence of (10.25) to be maximised $\hat{\tau}$ will take that optimum value so that the optimum value $\tilde{\tau}_0$ of $\tilde{\tau}$ to become identical with τ_0. In other words, at the optimum stage the iterative scheme (10.25) is identical with the PSD method, which if accelerated using the semi-iterative techniques becomes identical with the PJ-SI method. Consequently the optimum SI method based on SSOR is identical to the optimum SI method based on the PJ method. This conclusion can also be extended to include all the previously considered accelerated techniques based on the PJ method.

Finally, we consider the application of the SI method based on the LPSD. From (5.24) we have that the rate of convergence for the LPSD-SI

method is given by (see (3.11))

$$R_\infty(P_n(D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}})) \sim \sqrt{2}\sqrt{R(D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}})} \sim 2^{\frac{1}{4}}\sqrt{2}\sqrt{R(D_{\tau_1,\omega_1})} \tag{10.27}$$

hence

$$\frac{RR_\infty(P_n(D_{\tau_1,\omega_1})}{RR_\infty(P_n(D_{\tau_1^{(\pi_1)},\omega_1^{(\pi_1)}}))} \sim 2^{\frac{1}{4}} \tag{10.28}$$

for sufficiently small h. This result implies that for the unit square, or a subset thereof, there is a gain of approximately a factor of 1.2 in using the LPSD method with semi-iteration as compared with point PSD with semi-iteration. However, in order to achieve this relatively small improvement for the former scheme in terms of overall computational effort one should carry out the method using a normalised block iteration scheme (see Cuthill and Varga [1962]).

Summarising our results, we have the following asymptotic expressions for the reciprocal convergence rates of the various methods considered for the model problem

Method	Reciprocal Convergence Rate
J-SI	$\frac{1}{\pi} h^{-1}$
CCSI	$\frac{1}{2\pi} h^{-1}$
GS-SI	$\frac{1}{2\pi} h^{-1}$
SOR	$\frac{1}{2\pi} h^{-1}$
SSOR	$\frac{\sqrt{3}}{2\pi} h^{-1}$
PSD	$\frac{\sqrt{3}}{4\pi} h^{-1}$
LSSOR	$\frac{\sqrt{3}}{2^{3/2}\pi} h^{-1}$
LPSD	$\frac{\sqrt{3}}{2^{5/2}\pi} h^{-1}$
SSOR-SI	$\frac{3^{\frac{1}{4}}}{2^{3/2}\sqrt{\pi}} h^{-\frac{1}{2}}$
PJ-SI	$\frac{3^{\frac{1}{4}}}{2^{3/2}\sqrt{\pi}} h^{-\frac{1}{2}}$
LPSD-SI	$\frac{3^{\frac{1}{4}}}{2^{7/2}\sqrt{\pi}} h^{-\frac{1}{2}}$

FIGURE 10.1

From the above figure we notice that for the PJ-SI (or SSOR-SI) and LPSD-SI methods the number of iterations varies like $O(h^{-\frac{1}{2}})$. From our previous analysis we expect that this would also happen for the SD-PJ, PJ-VE and PJ-CG methods.

In order to obtain some information about the above methods in practice, we considered again the six problems as described in Section 4.8 under the same boundary conditions, starting vector and convergence criterion.

Next, the PJ method was accelerated both by variable extrapolation (PJ-VE) and by semi-iteration (PJ-SI). In each case, the optimum parameters were used whereas for the PJ-VE method the value of m was determined as the smallest integer such that

$$\left[-\frac{1}{m}\log\frac{2r^{m/2}}{1+r^{m}}\right]^{-1} \lesssim 1.25\ \frac{1}{(-\frac{1}{2}\log r)}\ . \tag{10.29}$$

This guarantees that the reciprocal rate of convergence does not exceed 125% of the reciprocal rate of convergence of the corresponding semi-iterative method.

In Table 10.2 we present the number of iterations of the two aforementioned iterative schemes. From this table we note that the number of iterations required for convergence using the PJ-SI and PJ-VE methods behaves approximately as $h^{-\frac{1}{2}}$, even though the coefficients A(x,y) and C(x,y) are not necessarily in class $C^{(2)}$ (see problem 5). However, it should be noted here that for a higher degree of discontinuity the behaviour is expected to be $h^{-\frac{3}{4}}$, as was shown in Section 5.5, under the assumption that $|A_x|$ and $|C_y|$ are bounded in the region under consideration. This is somewhat better than O(h) convergence of SOR.

As is shown (see Appendix A), the number of operations required per iteration using the PJ-SI method is approximately twice that required using the SOR method. This should be considered in comparing the PJ-SI with the SOR method. However, if we use the PJ-VE method, then we can

Problem	h^{-1}	ω_0	$\lambda(B_{\omega_0})$	$\lambda(B_{\omega_0})$	τ_0	PJ-SI	PJ-VE	m_0
1	20	1.7641	0.4568	2.4030	0.6993	17	20	4
	40	1.8750	0.4233	4.2667	0.4264	24	30	6
	60	1.9157	0.4068	6.1922	0.3031	30	35	7
2	20	1.5888	0.6313	1.5307	0.9251	12	12	3
	40	1.7668	0.5672	2.4271	0.6679	17	19	4
	60	1.8386	0.5439	3.3698	0.5110	21	25	5
3	20	1.7652	0.4488	2.4127	0.6989	17	20	4
	40	1.8756	0.4153	4.2859	0.4254	24	30	6
	60	1.9163	0.4096	6.2346	0.3010	29	35	7
4	20	1.7624	0.4566	2.3881	0.7031	17	20	4
	40	1.8748	0.4252	4.2603	0.4268	24	30	6
	60	1.9143	0.4121	6.0955	0.3073	29	35	7
5	20	1.7479	0.3901	2.2694	0.7520	18	20	4
	40	1.8665	0.3592	4.0132	0.4574	25	30	6
	60	1.9093	0.3494	5.7746	0.3266	31	35	7
6	20	1.6097	0.6311	1.5917	0.8998	11	12	3
	40	1.7820	0.5779	2.5742	0.6343	17	20	4
	60	1.8490	0.5595	3.5817	0.4829	19	25	5

TABLE 10.2

A COMPARISON OF PJ-SI AND PJ-VE METHODS WITH OPTIMUM PARAMETERS

h^{-1}	PJ-SI	PJ-CG	LPSD-SI	PJ-VE	m	SOR	PSD	SSOR
20	17	14	14	20	4	61	37	66
40	24	20	20	30	6	121	71	134
60	30	25	25	35	7	253	107	201

TABLE 10.3

NUMBER OF ITERATIONS FOR THE MODEL PROBLEM

reduce the work involved to be approximately equal (see(A.11)) to the work in SOR, thus making this iterative procedure more attractive than the former accelerated versions of the PJ method.

Finally, in Table 10.3 we present the number of iterations of the various procedures considered so far for solving the model problem using the same starting vector and convergence criterion.

CHAPTER 6

THE ADAPTIVE ALGORITHM

6.1 INTRODUCTION

In the last two chapters we were concerned with the construction of various iterative schemes and their comparison with respect to rates of convergence and related computational work. A lot of emphasis was also dedicated to the theoretical determination of the involved parameters to attain optimal rates of convergence. Finally, the formulation of the accelerated versions of the PJ method and especially the PJ-SI, PJ-VE, SD-PJ and PJ-CG procedures were also considered and it was shown that they form a variety of different algorithms with rapid rates of convergence. These latest schemes together with PSD, SOR and ADI-methods (see Chapter 7) can give an answer to the question as to which method should be used to solve systems of the form (3-1.1). However, the problem as to how the iteration parameters should be chosen so that the anticipated rapid rate of convergence will be attained, still remains since the number of iterations required to obtain the optimum parameters may exceed the number of iterations necessary to solve (3-1.1) itself. This is also very closely related to the problem of how one should decide when the iteration process should be terminated. A step towards the solution of the above problem was the work by Diamond [1971] and later on by Hageman [1972], Young [1974a] and Benokraitis [1974] who have considered a number of techniques for accelerating various iterative methods by Chebyshev acceleration whereas Ikebe et al [1973] considered an adaptive scheme which does not depend upon estimating the spectral radius of the iteration matrix. These methods adaptively update the required acceleration parameters and improve the approximate solution at the same time. The goal of the adaptive schemes is to attain convergence in only a few more iterations than would be required if the best possible values of the iteration parameters were used from the outset.

In this chapter we will develop an adaptive algorithm to accelerate

the PJ method which does not require any knowledge of the eigenvalues of B_ω. In particular, we will consider the PJ-SI method with the parameters being improved during the course of the iterations and we will show that this algorithm under certain conditions performs better than the PJ-SI method with estimated parameters.

6.2 SOME CONSIDERATIONS FOR CHOOSING THE OPTIMUM PARAMETERS

As we have seen in the previous chapter, in order to apply any accelerated version of the PJ method we need to have the optimum parameters ω_0 and $P(B_{\omega_0})$ or $S(D_{\tau_0,\omega_0})$ (although in the PJ-CG method we do not require $P(B_{\omega_0})$ in the actual iteration, its computation is essential for the determination of ω_0). A simple technique for determining these quantities is the selection of various values of τ and ω and the computation of the corresponding $S(D_{\tau,\omega})$ using the power method (see e.g. Gourlay and Watson [1973]). Thus, the triple $(\omega_0,\tau_0,S(D_{\tau_0,\omega_0}))$ which produces the smallest spectral radius $S(D_{\tau,\omega})$ can be chosen as the optimum parameter set. The most sophisticated scheme for this approach is to use an optimization method (e.g. Fibonacci search technique or golden section) for the appropriate selection of τ and ω. In Table 2.1 we present the optimum parameters ω_0,τ_0 and $S(D_{\tau_0,\omega_0})$ obtained by the power method for the problems considered in Section 4.8 (see Table 4-8.2).

Problem	h^{-1}	ω_0	τ_0	$S(D_{\tau_0,\omega_0})$	$P(B_{\omega_0})$
1	20	1.7641	0.6993	0.6805	5.2604
	40	1.8750	0.4264	0.8195	10.0806
	60	1.9157	0.3031	0.8767	15.2207
2	20	1.5888	0.9251	0.4160	2.4248
	40	1.7668	0.6659	0.6211	4.2790
	60	1.8386	0.5110	0.7221	6.1958
3	20	1.7652	0.6989	0.6863	5.3763
	40	1.8756	0.4254	0.8233	10.3200
	60	1.9163	0.3010	0.8767	15.2207
4	20	1.7624	0.7031	0.6790	5.2301
	40	1.8748	0.4268	0.8185	10.0200
	60	1.9143	0.3073	0.8734	14.7929
5	20	1.7479	0.7520	0.7066	5.8173
	40	1.8665	0.4574	0.8357	11.1732
	60	1.9093	0.3266	0.8859	16.5289
6	20	1.6097	0.8998	0.4322	2.5221
	40	1.7820	0.6343	0.6333	4.4543
	60	1.8490	0.4829	0.7298	6.4020

TABLE 2.1

OPTIMUM PARAMETERS ω_0,τ_0 AND $S(D_{\tau_0,\omega_0})$ OBTAINED BY POWER METHOD

Evidently, the aforementioned technique of obtaining the optimum parameters is impractical, we are therefore bound to consider other approaches. The same problem was also encountered for the determination of ω_0 in the SSOR method and the first step towards its solution was the work by Habelter and Wachspress [1961] who determined an implicit formula for the optimum parameters ω_0 and $S(\xi_{\omega_0})$. Later, Evans and Forrington [1963] modified the determination of the optimum parameters by devising an iterative scheme which although successful for the model problem does not guarantee to produce the optimum parameters for a wider class of problems.

Here, we present an algorithm for the determination of ω_0, τ_0 and $S(D_{\tau_0,\omega_0})$ which is based on the analysis presented in Chapter 4 and is similar to an algorithm for the determination of ω_0 and $S(\xi_{\omega_0})$ (Benokraitis [1974]).

Algorithm 6.1

1. Choose convergence tolerances $\varepsilon_1, \varepsilon_2$ and initial values of ω, τ and $v \neq 0$.
2. Iterate with the power method to obtain $S(D_{\tau,\omega})$ and a vector v such that

$$\begin{aligned} D_{\tau,\omega} v &= S(D_{\tau,\omega}) v \\ (v, Dv) &= 1. \end{aligned} \tag{2.1}$$

3. Compute

$$\begin{aligned} a &= (v, DBv) \\ \text{and} \qquad \beta &= (v, DLUv). \end{aligned} \tag{2.2}$$

4. Compute

$$\omega' = \begin{cases} \dfrac{2}{1+\sqrt{1-2a+4\beta}} & , \text{ if } a \leq 4\beta \\ \dfrac{2}{1+\sqrt{1-4\beta}} = \omega^* & , \text{ if } a > 4\beta, \end{cases} \tag{2.3}$$

$$P' = \begin{cases} \dfrac{1}{2}\left(1+\dfrac{\sqrt{1-2a+4\beta}}{1-a}\right) & , \text{ if } a \leq 4\beta \\ \dfrac{1-\sqrt{1-4\beta}}{2\sqrt{1-4\beta}} = \dfrac{1}{2-\omega^*} & , \text{ if } a > 4\beta, \end{cases} \tag{2.4}$$

$$\tau' = \frac{2\omega'(2-\omega')}{1+1/P'} \quad \text{and} \quad S' = \frac{P'-1}{P'+1}. \tag{2.5}$$

5. Terminate the process if

$$\begin{aligned} |\omega-\omega'| &< \varepsilon_1 \\ |P(B_\omega)-P'| &< \varepsilon_2 \end{aligned} \tag{2.6}$$

and choose

$$\begin{aligned} &\omega_0 = \omega', \quad \tau_0 = \tau' \\ &P(B_{\omega_0}) = P'. \end{aligned} \tag{2.7}$$

Otherwise set $\omega=\omega'$, $\tau=\tau'$ and go to step 2.

However, we note that the deficiency of the power method approach is still retained since the number of iterations required to obtain the optimum parameters by applying Algorithm 6.1 can be of the same order as the number of iterations required to find the approximate solution of (3-1.1). We are therefore motivated to consider a comparison of the PJ-SI procedure using estimated and optimum parameters since the former are obtained relatively easily.

6.3 STOPPING PROCEDURES

In this section we consider various criteria which will be used for terminating the procedures for the adaptive determination of the parameters ω and $P(B_\omega)$. The analysis of the stopping procedures is similar to the one developed by Young [1974a], Benokraitis [1974] and Cullen [1974] modified slightly to suit our purposes.

Let $\bar{u}=(I-G)^{-1}k$ be the exact solution of (3-1.3) and hence of (3-1.1). We recall from (3-3.3) that the error vector at the n^{th} stage of the PJ-SI iteration is defined by

$$\varepsilon^{(n)} = u^{(n)}-\bar{u}. \tag{3.1}$$

Here we will accept $u^{(n)}$ as an adequate approximation to the exact solution $\bar{u}$ provided the following inequality is satisfied by the relative error

$$\frac{\|\varepsilon^{(n)}\|_{A^{\frac{1}{2}}}}{\|\varepsilon^{(0)}\|_{A^{\frac{1}{2}}}} = \frac{\|u^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}}}{\|\bar{u}\|_{A^{\frac{1}{2}}}} < \zeta \quad , \tag{3.2}$$

where we assume that $u^{(0)}=0$ and ζ is some small tolerance (e.g. $\zeta=10^{-6}$). It can be observed that we cannot use the convergence criterion (3.2) directly, for $\bar{u}$ is not available at the outset. Thus, we have to consider various upper bounds on the relative error defined in (3.2) in order to avoid this difficulty.

In particular, for the PJ-SI we have from (5-5.14), (5-5.17) that

$$u^{(n)} = P_n(\mathcal{H}_\omega)u^{(0)} + k_n \tag{3.3}$$

and by consistency, we obtain the following

$$\bar{u} = P_n(\mathcal{H}_\omega)\bar{u} + k_n. \tag{3.4}$$

From (3.3) and (3.4) we have

$$u^{(n)}-\bar{u} = P_n(\mathcal{H}_\omega)(u^{(0)}-\bar{u}) \tag{3.5}$$

hence (3.2) can be modified to yield

$$\frac{\|u^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}}}{\|\bar{u}\|_{A^{\frac{1}{2}}}} \leq \frac{\|P_n(\mathcal{H}_\omega)\|_{A^{\frac{1}{2}}}\|u^{(0)}-\bar{u}\|_{A^{\frac{1}{2}}}}{\|\bar{u}\|_{A^{\frac{1}{2}}}} = \|P_n(\mathcal{H}_\omega)\|_{A^{\frac{1}{2}}} \tag{3.6}$$

since $u^{(0)}=0$.

We therefore define by (3.2) and (3.6) stopping "Procedure I" for PJ-SI as

$$K_I = \frac{2\bar{r}^n}{1+\bar{r}^{2n}} \leq \zeta \tag{3.7}$$

since

$$\|P_n(\mathcal{H}_\omega)\|_{A^{\frac{1}{2}}} = \|A^{\frac{1}{2}}P_n(\mathcal{H}_\omega)A^{-\frac{1}{2}}\| = \|P_n(A^{\frac{1}{2}}\mathcal{H}_\omega A^{-\frac{1}{2}})\| = S(P_n(A^{\frac{1}{2}}\mathcal{H}_\omega A^{-\frac{1}{2}}))$$

$$= S(A^{\frac{1}{2}}P_n(\mathcal{H}_\omega)A^{-\frac{1}{2}}) = S(P_n(\mathcal{H}_\omega)) \tag{3.8}$$

where $\bar{r}$ is given by (5-5.19). This procedure is an a priori criterion since we can determine in advance how many iterations are required such that (3.2) is satisfied.

Next, we consider an alternative stopping procedure which may, in favourable cases, lead us to terminate the iteration process sooner than Procedure I (and in some cases later).

The pseudo-residual vector (or incremental vector see Diamond [1971]) is defined by (see Young [1974a]) the expression

$$\delta^{(n)} = Gu^{(n)}+k-u^{(n)}. \tag{3.9}$$

Since $\bar{u}=G\bar{u}+k$, we have

$$\delta^{(n)} = Gu^{(n)}+(\bar{u}-G\bar{u})-u^{(n)} = (G-I)(u^{(n)}-\bar{u}) \tag{3.10}$$

and by (3.1) we obtain

$$\delta^{(n)} = (G-I)\varepsilon^{(n)} \tag{3.11}$$

which indicates that the error vector $\varepsilon^{(n)}$ can be expressed in terms of the pseudo-residual vector $\delta^{(n)}$. By combining (3.2) and (3.11) we find the result

$$\frac{\|u^{(n)}-\bar{u}\|_{A^{\frac{1}{2}}}}{\|\bar{u}\|_{A^{\frac{1}{2}}}} \leq \frac{\|(G-I)^{-1}\|_{A^{\frac{1}{2}}}\|\delta^{(n)}\|_{A^{\frac{1}{2}}}}{\|\bar{u}\|_{A^{\frac{1}{2}}}} \leq \zeta . \tag{3.12}$$

Depending on how we approximate $\|\bar{u}\|_{A^{\frac{1}{2}}}$, we obtain various stopping criteria. Here, we first define "Procedure II" if we substitute $\|\bar{u}\|_{A^{\frac{1}{2}}}$ from the expression

$$\|\delta^{(0)}\|_{A^{\frac{1}{2}}} \leq \|G-I\|_{A^{\frac{1}{2}}} \|u^{(0)}-\bar{u}\|_{A^{\frac{1}{2}}} = \|G-I\|_{A^{\frac{1}{2}}} \|\bar{u}\|_{A^{\frac{1}{2}}}. \tag{3.13}$$

Thus (3.12) holds, if the following inequality is satisfied

$$\|(G-I)^{-1}\|_{A^{\frac{1}{2}}} \|G-I\|_{A^{\frac{1}{2}}} \frac{\|\delta^{(n)}\|_{A^{\frac{1}{2}}}}{\|\delta^{(0)}\|_{A^{\frac{1}{2}}}} \leq \zeta \tag{3.14}$$

which can also be written as

$$K_{II} = k(G-I) \frac{\|\delta^{(n)}\|_{A^{\frac{1}{2}}}}{\|\delta^{(0)}\|_{A^{\frac{1}{2}}}} \leq \zeta \tag{3.15}$$

where $k(G-I)$ denotes the spectral condition number of $(G-I)$.

If we replace $\|\bar{u}\|_{A^{\frac{1}{2}}}$ by $\|u^{(n)}\|_{A^{\frac{1}{2}}}$ in (3.13), then we can approximate $\|\delta^{(0)}\|_{A^{\frac{1}{2}}}$ by $\|u^{(n)}\|_{A^{\frac{1}{2}}}$ if $\|G-I\|_{A^{\frac{1}{2}}} = S(G-I)<1$, hence (3.15) becomes

$$K_{III} = k(G-I) \frac{\|\delta^{(n)}\|_{A^{\frac{1}{2}}}}{\|u^{(n)}\|_{A^{\frac{1}{2}}}} \leq \zeta \tag{3.16}$$

which defines "Procedure III".

We note that in order to apply the tests (3.15) and (3.16) we need a bound on the spectral condition number of G-I. However, the effect of inaccuracies on these bounds as far as the convergence testing is concerned is much less than the effect on the rapidity of convergence. Therefore, we can often use any crude bound which may be available without a substantial alteration on the number of iterations.

6.4 COMPUTATIONAL PROCEDURES AND NUMERICAL RESULTS

In this section, we compare the effectiveness of the estimated parameters with the optimum ones by applying the PJ-SI and PJ-VE methods (for solving the self-adjoint equation (5-5.36) for the different expressions of the coefficients A(x,y) and C(x,y) (see Table 4-8.2). Furthermore, we use the stopping procedures introduced in the previous section for terminating the iterations. Before we present any numerical results we summarise the procedures for applying the PJ-SI and PJ-VE methods with estimated parameters for solving the linear system corresponding to (1-2.6).

Algorithm 6.2 (PJ-SI method)

1. As a starting vector we choose $u^{(0)}$ such that

$$\|u^{(0)}-\bar{u}\|_{A^{\frac{1}{2}}} \leq \|\bar{u}\|_{A^{\frac{1}{2}}}. \tag{4.1}$$

   The choice $u^{(0)}=0$ will suffice.

2. Compute M=-m by (5-5.27).

   This involves the determination of an Ih×Jh rectangle containing R+∂R.

3. Compute $\bar{\beta}$ using the expression (see Appendix B)

$$\bar{\beta} = \max_{(x,y)\in Rh} \{\beta_3(x,y)[\beta_1(x-h,y)+\beta_2(x-h,y)]+\beta_4(x,y)[\beta_1(x,y-h)+\beta_2(x,y-h)]\} \tag{4.2}$$

4. Adjust M if necessary.

   If $M>2\sqrt{\bar{\beta}}$, replace M by $2\sqrt{\bar{\beta}}$.

5. Compute ω_1 and $P(B_{\omega_1})$ by (4-11.4),(4-11.5), respectively.

6. Iterate using the PJ-SI method

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)}+\rho_{n+1}[u^{(n)}+\bar{\rho}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)})] \tag{4.3}$$

   The values of $\bar{\rho},\rho_1,\rho_2,\ldots$ are given by the following expressions

$$\bar{\rho} = \frac{2\omega_1(2-\omega_1)}{1+1/P(B_{\omega_1})}, \tag{4.4}$$

$$\left.\begin{aligned} \rho_1 &= 1 , \\ \rho_2 &= \left(1 - \frac{\sigma^2}{2}\right)^{-1} \\ \rho_{n+1} &= \left(1 - \frac{\sigma^2}{4}\rho_n\right)^{-1} \quad , \ n=2,3,\ldots \end{aligned}\right\} \tag{4.5}$$

where

$$\sigma = \frac{P(B_{\omega_1})-1}{P(B_{\omega_1})+1} . \tag{4.6}$$

7. Terminate the process after n iterations where n satisfies the inequality

$$K_I \leq \zeta. \tag{4.7}$$

The alternative procedure of accelerating the PJ method is the PJ-VE and is represented by the same steps as in Algorithm 6.2 but instead of step 6 and 7 we have (Algorithm 6.3):

6. Choose m_1 as the smallest integer such that

$$\left[-\frac{1}{m_1}\log\frac{2r^{m_1/2}}{1+r^{m_1}}\right]^{-1} \leq 1.25\left(-\frac{1}{2}\log r\right)^{-1}. \tag{4.8}$$

7. Iterate using the PJ-VE method defined by

$$u^{(n+1)} = u^{(n)} + \theta_{n+1}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{4.9}$$

where

$$\theta_k = \frac{\omega_1(2-\omega_1)}{1/P(B_{\omega_1})\cos^2\frac{(2k-1)\pi}{4m_1} + \sin^2\frac{(2k-1)\pi}{4m_1}} , \ k=1,2,\ldots,m_1 \tag{4.10}$$

8. Terminate the process after tm_1 iterations where

$$\left(\frac{2r^{m_1/2}}{1+r^{m_1}}\right)^t \leq \zeta . \tag{4.11}$$

In order to test the efficiency of the Algorithms 6.2 and 6.3 we considered their application to the six problems (see Section 4.8) and compared their results with the PJ-SI and PJ-VE with optimum parameters.

Here the boundary values were taken to be zero on all sides of the unit square except for the side y=0, where they were taken to be unity. The natural ordering was used and $u^{(0)}=0$ was taken as a starting vector, whereas we let $\zeta=10^{-6}$.

In Table 4.1 we present the number of iterations of PJ-SI and PJ-VE with optimum and estimated parameters, required to satisfy stopping Procedure I. On the other hand, under the column headings I,II,III in Tables 4.2 and 4.3 we present the number of iterations required to satisfy stopping procedures I,II and III using the PJ-SI method with optimum and estimated parameters, respectively. For the PJ-VE method, the number of iterations was tm_0 (or tm_1), where m_0 (or m_1) is determined by (4.8) and is given by the relationship

$$\left(\frac{2r^{m_0/2}}{1+r^{m_0}}\right)^t \leq \zeta. \tag{4.12}$$

From Table 4.1 we verify again that for the problems considered the number of iterations required to satisfy stopping Procedure I using the PJ-SI and PJ-VE method varies approximately as $h^{-\frac{1}{2}}$ even though the coefficients A(x,y) and C(x,y) are not necessarily in the class $C^{(2)}$. A comparison of the results obtained in Table 5-10.2 with the ones obtained in Table 4.1 shows that although ideal conditions were used (see Section 5.10) the behaviour of the number of iterations is typical for other boundary conditions as well.

Also, from Table 4.1 we note that the results using the estimated parameters were reasonably good in comparison with the results based on the optimum parameters for both the PJ-SI and PJ-VE methods. Only in the third and fifth cases were there substantial differences which indicates that in such cases it would appear worthwhile to attempt to improve the parameters ω and $P(B_\omega)$ adaptively. The same situation is true when we use Procedures II and III as stopping criteria. These also provide a suitable indication when (3.2) is satisfied for PJ-SI (see Tables

Problem	h^{-1}	$\bar{\beta}$	$2\sqrt{\bar{\beta}}$	M	ω_1	ω_0	$P(B_{\omega_1})$	$P(B_{\omega_0})$	PJ-SI		PJ-VE			
									n_{opt}	n_{est}	n_{opt}	m_0	n_{est}	m_1
1	20	0.2500	1.0000	0.9877	1.7287	1.7641	6.8727	5.2604	16	19	20	4	25	5
	40	0.2500	1.0000	0.9969	1.8544	1.8750	13.2357	10.0806	23	26	30	6	35	7
	60	0.2500	1.0000	0.9986	1.9005	1.9157	19.6008	15.2207	28	32	35	7	40	8
2	20	0.2350	0.9695	1.0000	1.6065	1.5888	2.5415	2.4248	10	10	12	3	12	3
	40	0.2461	0.9922	1.0000	1.7788	1.7668	4.5208	4.2790	14	15	16	4	20	4
	60	0.2483	0.9965	1.0000	1.8465	1.8386	6.5139	6.1958	18	18	20	5	25	5
3	20	0.2505	1.0009	0.9969	1.8355	1.7652	14.9905	5.3763	16	29	20	4	35	7
	40	0.2501	1.0002	0.9992	1.9142	1.8756	29.5540	10.3200	23	39	30	6	50	10
	60	0.2501	1.0001	0.9997	1.9420	1.9163	44.1015	15.2207	28	48	35	7	66	11
4	20	0.2500	1.0001	0.9918	1.7717	1.7624	8.3322	5.2301	16	21	20	4	25	5
	40	0.2500	1.0000	0.9979	1.8790	1.8748	16.1660	10.0200	23	29	30	6	35	7
	60	0.2500	1.0000	0.9991	1.9176	1.9143	23.9999	14.7929	28	36	35	7	45	9
5	20	0.2500	0.9999	0.9978	1.8756	1.7479	15.2395	5.8173	17	28	20	4	35	7
	40	0.2500	1.0000	0.9994	1.9359	1.8665	30.0138	11.1732	24	40	30	6	50	10
	60	0.2500	1.0000	0.9998	1.9568	1.9093	44.7804	16.5289	29	49	35	7	66	11
6	20	0.2416	0.9831	1.0000	1.6903	1.6097	3.2293	2.5221	10	12	12	3	15	3
	40	0.2483	0.9966	1.0000	1.8475	1.7820	6.5567	4.4543	15	18	20	4	25	5
	60	0.2493	0.9986	1.0000	1.8997	1.8490	9.9697	6.4020	18	23	25	5	30	6

TABLE 4.1

Problem	h^{-1}	Optimum Parameter PJ-SI				
		ω_0	$P(B_{\omega_0})$	I	II	III
1	20	1.7641	5.2604	16	17	18
	40	1.8750	10.0806	23	26	28
	60	1.9157	15.2207	28	32	36
2	20	1.5888	2.4248	10	12	11
	40	1.7668	4.2790	14	16	17
	60	1.8386	6.1958	18	20	22
3	20	1.7652	5.3763	16	17	18
	40	1.8756	10.3200	23	26	28
	60	1.9163	15.2207	28	32	35
4	20	1.7624	5.2301	16	17	18
	40	1.8748	10.0200	23	26	28
	60	1.9143	14.7929	28	32	35
5	20	1.7479	5.8173	17	18	19
	40	1.8665	11.1732	24	27	29
	60	1.9093	16.5289	29	35	38
6	20	1.6097	2.5221	10	11	11
	40	1.7820	4.4543	15	16	17
	60	1.8490	6.4020	18	20	22

TABLE 4.2

NUMBER OF ITERATIONS REQUIRED TO SATISFY STOPPING CRITERIA I, II AND III USING PJ-SI WITH OPTIMUM PARAMETERS FOR THE SIX PROBLEMS

Problem	h^{-1}	Estimated Parameters PJ-SI				
		ω_1	$P(B_{\omega_1})$	I	II	III
1	20	1.7287	6.8727	19	21	21
	40	1.8544	13.2357	26	30	32
	60	1.9005	19.6008	32	37	41
2	20	1.6065	2.5415	10	12	12
	40	1.7788	4.5208	15	17	18
	60	1.8465	6.5139	18	21	23
3	20	1.8355	14.9905	28	32	33
	40	1.9142	29.5540	39	47	51
	60	1.9420	44.1015	48	59	66
4	20	1.7717	8.3322	21	23	24
	40	1.8790	16.1660	29	34	37
	60	1.9176	23.9999	36	43	47
5	20	1.8756	15.2395	28	33	34
	40	1.9359	30.0138	40	49	54
	60	1.9568	44.7804	49	63	69
6	20	1.6903	3.2293	12	12	13
	40	1.8475	6.5567	18	17	18
	60	1.8997	9.9697	26	21	28

TABLE 4.3

NUMBER OF ITERATIONS REQUIRED TO SATISFY STOPPING CRITERIA I, II AND III USING PJ-SI WITH ESTIMATED PARAMETERS FOR THE SIX PROBLEMS

4.2 and 4.3). The above stopping procedures will be proved to be suitable for the adaptive acceleration of the PJ method. Finally, if we also consider the amount of work involved to determine the optimum parameters (see Algorithm 6.1) we are motivated by these observations to seek adaptive or dynamic procedures which approximate the parameters ω_0 and $P(B_{\omega_0})$ and at the same time, obtain the solution of Au=b. Here, it should be mentioned that Benokraitis [1974] considered similar procedures for the SSOR-SI method involving the simultaneous determination of both ω and $S(\xi_\omega)$.

6.5 THE THEORETICAL BASIS FOR THE ADAPTIVE DETERMINATION OF PARAMETERS

Our aim in the remainder of this chapter will be to develop an efficient procedure which will use the PJ-SI method for solving elliptic difference equations of the form (1-2.6).

From Section 4.9 we recall that the preconditioning parameter is optimum for that value of ω for which the P-condition number of B_ω is minimised. We also have seen that for the largest and smallest eigenvalue of B_ω we can let, respectively

$$\Lambda(B_\omega) = \frac{1}{\omega(2-\omega)} \quad , \tag{5.1}$$

$$\lambda(B_\omega) = \frac{1-\omega a+\omega^2\beta}{1-a} = \phi(\omega,v) \tag{5.2}$$

where

$$a = \frac{(v,DBv)}{(v,Dv)} \quad , \tag{5.3}$$

$$\beta = \frac{(v,DLUv)}{(v,Dv)} \tag{5.4}$$

and

$$B_\omega v = \lambda(B_\omega)v. \tag{5.5}$$

Therefore, the P-condition number of B_ω is given by the expression

$$P(B_\omega) = \frac{1-\omega a+\omega^2\beta}{\omega(2-\omega)(1-a)} \tag{5.6}$$

where a and β are given by (5.3),(5.4), respectively. From (5.6) and (5.2) we see that finding the optimum parameters ω_0, $P(B_{\omega_0})$ depends upon the availability of an eigenvector corresponding to the smallest eigenvalue of B_ω. Evidently, if we happened to know this eigenvector, then we would be able to determine $\lambda(B_\omega)$ from the formula

$$\lambda(B_\omega) = \frac{(v,B_\omega v)}{(v,v)} \tag{5.7}$$

and then compute $P(B_\omega)$ from the expression

$$P(B_\omega) = [\omega(2-\omega)\lambda(B_\omega)]^{-1}.$$

It is therefore clear that the determination of v such that (5.5) is satisfied, is essential for obtaining the optimum parameters. Thus we are motivated by this observation to seek for a vector which is

automatically calculated in the practical implementation of the PJ-SI method and can be made to approach an eigenvector of B_ω corresponding to $\lambda(B_\omega)$. As a result our task will be to compute an approximation to the quantities a and β defined by (5.3) and (5.4), respectively thus computing by (5.6) approximations (using the analysis of Section 4-4.9) to the optimum parameters ω_0 and $P(B_{\omega_0})$.

We commence our analysis by proving the following theorem

Theorem 5.1

Let A be a positive definite matrix, then for any vector $v\neq 0$ the representation $\phi(\omega,v)$ given by (5.2) is a Rayleigh quotient with respect to the vector $w=D^{\frac{1}{2}}(I-\omega U)v$ and the positive definite matrix

$$\overline{B}_\omega = D^{\frac{1}{2}}(I-\omega U)B_\omega(I-\omega U)^{-1}D^{-\frac{1}{2}}, \tag{5.8}$$

that is

$$\phi(\omega,v) = \frac{(w,\overline{B}_\omega w)}{(w,w)} . \tag{5.9}$$

Furthermore,

$$\lambda(B_\omega) = \lambda(\overline{B}_\omega) \leq \phi(\omega,v) . \tag{5.10}$$

Proof

We first show that $\overline{B}_\omega$ is positive definite. We recall from (4-5.1) that

$$B_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \tag{5.11}$$

hence

$$\begin{aligned}\overline{B}_\omega &= D^{\frac{1}{2}}(I-\omega U)B_\omega(I-\omega U)^{-1}D^{-\frac{1}{2}} \\ &= D^{\frac{1}{2}}(I-\omega L)^{-1}D^{-1}A(I-\omega U)^{-1}D^{-\frac{1}{2}} \\ &= (I-\omega\tilde{L})^{-1}D^{-\frac{1}{2}}AD^{-\frac{1}{2}}(I-\omega\tilde{U})^{-1} \\ &= [(I-\omega\tilde{L})^{-1}D^{-\frac{1}{2}}]A[(I-\omega\tilde{L})^{-1}D^{-\frac{1}{2}}]^T \end{aligned} \tag{5.12}$$

where

$$\tilde{L} = D^{\frac{1}{2}}LD^{-\frac{1}{2}} \quad \text{and} \quad \tilde{U} = D^{\frac{1}{2}}UD^{-\frac{1}{2}} \tag{5.13}$$

since $\tilde{L}^T=\tilde{U}$. From (5.12) and Theorem 2-2.4 it follows that $\overline{B}_\omega$ is positive definite.

From (5.11) we also have

$$(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A = B_\omega$$

or

$$D^{\frac{1}{2}}(I-\omega L)^{-1}D^{-1}A(I-\omega U)^{-1}D^{-\frac{1}{2}} = \bar{B}_\omega$$

thus

$$\begin{aligned} A &= D(I-\omega L)D^{-\frac{1}{2}}\bar{B}_\omega D^{\frac{1}{2}}(I-\omega U) \\ &= D^{\frac{1}{2}}(I-\omega\tilde{L})\bar{B}_\omega(I-\omega\tilde{U})D^{\frac{1}{2}}. \end{aligned} \tag{5.14}$$

If we take inner products of both sides of the last equation with respect to $v\neq 0$ we have

$$\begin{aligned} (v,Av) &= (v,D^{\frac{1}{2}}(I-\omega\tilde{L})\bar{B}_\omega(I-\omega\tilde{U})D^{\frac{1}{2}}) \\ &= ((I-\omega\tilde{U})D^{\frac{1}{2}}v,\bar{B}_\omega(I-\omega\tilde{U})D^{\frac{1}{2}}v) \\ &= (w,\bar{B}_\omega w) \end{aligned} \tag{5.15}$$

where $w = (I-\omega\tilde{U})D^{\frac{1}{2}}v.$

Dividing by (w,w) both sides of (5.15) we obtain

$$\frac{(v,Av)}{(w,w)} = \frac{(w,\bar{B}_\omega w)}{(w,w)}. \tag{5.16}$$

Expanding the inner product (w,w) we have successively

$$\begin{aligned} (w,w) &= ((I-\omega\tilde{U})D^{\frac{1}{2}}v,(I-\omega\tilde{U})D^{\frac{1}{2}}v) \\ &= (v,D^{\frac{1}{2}}(I-\omega\tilde{L})(I-\omega\tilde{U})D^{\frac{1}{2}}v) \\ &= (v,D(I-\omega L)(I-\omega U)v) \\ &= (v,(D-\omega DB+\omega^2 DLU)v) \\ &= (v,Dv)-\omega(v,DBv)+\omega^2(v,DLUv). \end{aligned} \tag{5.17}$$

Since we also have that

$$(v,Av) = (v,(D-DB)v) = (v,Dv)-(v,DBv) \tag{5.18}$$

then by using (5.17), the left hand side of (5.16) yields

$$\begin{aligned} \frac{(v,Av)}{(w,w)} &= \frac{(v,Dv)-(v,DBv)}{(v,Dv)-\omega(v,DBv)+\omega^2(v,DLUv)} \\ &= \frac{1-a}{1-\omega a+\omega^2\beta} = \phi(\omega,v) \end{aligned} \tag{5.19}$$

thus (5.16) becomes

$$\phi(\omega,v) = \frac{(w,\bar{B}_\omega w)}{(w,w)}. \tag{5.20}$$

Finally, since $\bar{B}_\omega$ is similar to B_ω, then by Theorem 2-1.5 we have (5.10) and the proof of the theorem is complete.

Corollary 5.2

Under the hypotheses of Theorem 5.1 any eigenvalue of B_ω can be represented by (5.2).

From the above analysis we see that the inequality (5.10) is satisfied for any non-zero vector v. On the other hand, we observe from (5.10) that the closer we approach an eigenvector corresponding to the smallest eigenvalue of B_ω, the better we will be able to determine $\lambda(B_\omega)$ from $\phi(\omega,v)$. It is evident now there is a strong need for finding this eigenvector. However, we have to devise another approach other than using the power method since as we have seen the power iterations require extensive computational effort and do not contribute directly to the solution of the system (3-1.1). The answer to this problem was given by Diamond [1971] for the general case. Here we have modified this approach to suit our purposes.

Next, we will first show that the pseudo-residual vector, as defined in (3.9), satisfies the relationship $\delta^{(n)}=P_n(\mathcal{H}_\omega)\delta^{(0)}$ and secondly that $\delta^{(n)}$ approaches the vector v which satisfies (5.5).

Theorem 5.3

The pseudo-residual vector

$$\delta^{(n)} = \mathcal{H}_\omega u^{(n)} + k - u^{(n)} \tag{5.21}$$

where $u^{(n)}$ is the latest PJ-SI iterate, satisfies the relationship

$$\delta^{(n)} = P_n(\mathcal{H}_\omega)\delta^{(0)}. \tag{5.22}$$

Proof

We recall from (3.10) that for the PJ-SI method we have

$$\delta^{(n)} = (\mathcal{H}_\omega - I)(u^{(n)} - \bar{u}) \tag{5.23}$$

where $u^{(n)}$ is the latest PJ-SI iterate. Alternatively, we can write the PJ-SI method in the form

$$u^{(n+1)} = P_n(\mathcal{H}_\omega)u^{(0)} + k_n \tag{5.24}$$

where

$$P_n(\mathcal{H}_\omega) = \frac{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2B_\omega}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)}{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)} . \tag{5.25}$$

Moreover, since (5.24) is consistent, we have the relationship

$$\bar{u} = P_n(\mathcal{H}_\omega)\bar{u} + k_n . \tag{5.26}$$

Subtracting (5.26) from (5.24) yields the result

$$u^{(n)}-\bar{u} = P_n(\mathcal{H}_\omega)(u^{(0)}-\bar{u}) . \tag{5.27}$$

Further, by combining (5.23) and (5.27) we obtain the following expression for $\delta^{(n)}$

$$\delta^{(n)} = (\mathcal{H}_\omega-I)P_n(\mathcal{H}_\omega)(u^{(0)}-\bar{u}) . \tag{5.28}$$

Letting n=0 in (5.23) we obtain

$$u^{(0)}-\bar{u} = (\mathcal{H}_\omega-I)^{-1}\delta^{(0)} \tag{5.29}$$

which on substitution in (5.28) yields (5.22) and the proof of the theorem is complete.

The next theorem will establish the fact that $\delta^{(n)}$ does converge to a multiple of the eigenvector corresponding to the smallest eigenvalue of B_ω.

Theorem 5.4

The pseudo-residual vector $\delta^{(n)}$ given by (5.21) approaches a multiple of the eigenvector associated with $\lambda(B_\omega)$ as $n\to\infty$.

Proof

Let v_k, k=1,2,...,N, form a complete set of eigenvectors of B_ω corresponding to the positive eigenvalues

$$\Lambda(B_\omega) = \lambda_1>\lambda_2\geq\ldots\geq\lambda_{N-1}>\lambda_N = \lambda(B_\omega) > 0. \tag{5.30}$$

Next, we express $\delta^{(0)}$ as a linear combination of v_k, thus

$$\delta^{(0)} = \sum_{k=1}^{N} d_k v_k \ , \ d_N\neq 0 \tag{5.31}$$

and let

$$\lambda_E(B_\omega) > \lambda_N \tag{5.32}$$

where $\lambda_E(B_\omega)$ is an estimate of $\lambda(B_\omega)$.

From Theorem 5.3 we have that the pseudo-residual vector defined by (5.21) satisfies the relationship

$$\delta^{(n)} = P_n(\mathcal{H}_\omega)\delta^{(0)} \tag{5.33}$$

which by using (5.31) and (5.25) yields successively

$$\delta^{(n)} = P_n(\mathcal{H}_\omega)\sum_{k=1}^{N} d_k v_k = \sum_{k=1}^{N} d_k P_n(\lambda_k) v_k$$

$$= P_n(\lambda_N)\left\{d_N v_N + \sum_{k=1}^{N-1} d_k \frac{P_n(\lambda_k)}{P_n(\lambda_N)} v_k\right\}$$

$$= P_n(\lambda_N)\left\{d_N v_N + \sum_{k=1}^{N-1} d_k \frac{T_n\left(\frac{\lambda_1+\lambda_E(B_\omega)-2\lambda_k}{\lambda_1-\lambda_E(B_\omega)}\right)}{T_n\left(\frac{\lambda_1+\lambda_E(B_\omega)-2\lambda_N}{\lambda_1-\lambda_E(B_\omega)}\right)} v_k\right\}$$

$$= P_n(\lambda_N)\left\{d_N v_N + \sum_{k=1}^{N-1} d_k \frac{T_n\left(1-\frac{2(\lambda_k-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)}{T_n\left(1-\frac{2(\lambda_N-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)} v_k\right\}. \tag{5.34}$$

Consequently, from (5.34) we have

$$\frac{\delta^{(n)}}{P_n(\lambda_N)} = d_N v_N + \sum_{\lambda_k \in [\lambda_E(B_\omega), \lambda_1]} d_k \frac{T_n\left(1-\frac{2(\lambda_k-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)}{T_n\left(1-\frac{2(\lambda_N-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)} v_k$$

$$+ \sum_{\substack{\lambda_k \notin [\lambda_E(B_\omega), \lambda_1] \\ \lambda_k \neq \lambda_N}} d_k \frac{T_n\left(1-\frac{2(\lambda_k-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)}{T_n\left(1-\frac{2(\lambda_N-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)}\right)} v_k . \tag{5.35}$$

Applying Theorem C1 (see Appendix C) to the terms of the first sum we see that as n increases these terms are decreasing at an optimal rate. Let us know concentrate on the terms of the second sum where we have that $\lambda_k \in [\lambda_N, \lambda_E(B_\omega)]$, i.e.,

$$\lambda_E(B_\omega) > \lambda_k > \lambda_N \ . \tag{5.36}$$

We note that by letting

$$\left.\begin{aligned} x &= 1 - \frac{2(\lambda_k-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)} \\ y &= 1 - \frac{2(\lambda_N-\lambda_E(B_\omega))}{\lambda_1-\lambda_E(B_\omega)} \end{aligned}\right\} \tag{5.37}$$

and using (5.36) we can easily find the following relationship between x and y

$$y > x > 1. \tag{5.38}$$

Furthermore, from (5.38) we have

$$\cosh^{-1}y = \log(y+\sqrt{y^2-1}) > \log(x+\sqrt{x^2-1}) = \cosh^{-1}x \tag{5.39}$$

or

$$\cosh^{-1}x - \cosh^{-1}y < 0. \tag{5.40}$$

On the other hand, from (C4) (see Appendix C) we use the following expression for $T_n(x)$ since $x>1$

$$T_n(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n}\right]$$

which by (5.39) can be written alternatively as

$$T_n(x) = \frac{1}{2}\left[e^{n\cosh^{-1}x} + e^{-n\cosh^{-1}x}\right] .$$

Finally, by using (5.40) we obtain

$$\lim_{n\to\infty} \frac{T_n(x)}{T_n(y)} = \lim_{n\to\infty} \frac{e^{n\cosh^{-1}x}+e^{-n\cosh^{-1}x}}{e^{n\cosh^{-1}y}+e^{-n\cosh^{-1}y}} = \lim_{n\to\infty} e^{n(\cosh^{-1}x-\cosh^{-1}y)} = 0$$

which indicates that the terms of the second sum in (5.35) vanish also as $n\to\infty$ and therefore the proof of the theorem is complete.

The above theorem establishes the theoretical basis for the use of the PJ-SI method to solve Au=b and simultaneously compute an approximate eigenvector associated with $\lambda(B_\omega)$. This can be more explicitly seen if we consider again the pseudo-residual vector

$$\delta^{(n)} = \mathcal{H}_\omega u^{(n)} + \gamma_\omega - u^{(n)} .$$

Then by (4-4.6) we have that

$$\delta^{(n)} = (I-\omega U)^{-1}(I-\omega L)^{-1}(b-Au^{(n)}) \tag{5.41}$$

where $u^{(n)}$ is the latest PJ-SI iteration. But, the next PJ-SI iteration $u^{(n+1)}$ is given by (see (5-5.14))

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)}+\rho_{n+1}(u^{(n)}+\bar{\rho}\delta^{(n)}) \tag{5.42}$$

which shows that the pseudo-residual vector is essentially obtained as a by-product of the application of the PJ-SI method. This last observation is the main advantage of determining the parameters ω and $P(B_\omega)$ adaptively since we obtain the fundamental eigenvector v by exploiting the iteration used to improve the accuracy of the approximate solution of Au=b.

Furthermore, it remains to be shown that any approximation to an eigenvector of B_ω yields a corresponding eigenvalue approximation. This is derived from a theorem in Wachspress [1966] (see also Diamond [1971]) and is presented here without proof.

Theorem 5.5

If A and B are positive definite matrices, then the eigenvectors x_k and the corresponding eigenvalues λ_k of the generalised eigenvalue problem

$$Ax_k = \lambda_k Bx_k \tag{5.43}$$

satisfy the following properties:

a) The eigenvalues λ_k are all positive, i.e., $\lambda_k>0$ for k=1,2,...,N.

b) The eigenvectors x_k of $B^{-1}A$ are orthogonal with respect to B, i.e., $(x_k,Bx_j)=0$ for $j\neq k$.

We note that if we apply the above theorem to A and $B\equiv D(I-\omega(L+U)+\omega^2LU)$, then we have that the eigenvectors v_k of B_ω are orthogonal with respect to $D(I-\omega(L+U)+\omega^2LU)$.

Finally, the next lemma defines the approximate eigenvalue $\lambda(B_\omega)$.

Lemma 5.6

Let λ_k k=1,2,...,N, be the eigenvalues of B_ω and v_k the corresponding

eigenvectors. If [†]

$$\mu = \frac{(y,Ay)}{(y,(I-\omega U)^{-1}(I-\omega L)^{-1}y)} \tag{5.44}$$

where y is approximately equal to v_N with errors ε_k in the direction v_k and $\varepsilon_N >> \varepsilon_k, k \neq N$, then $\mu \simeq \lambda_N$ with error of order $\left(\frac{\varepsilon_k}{\varepsilon_N}\right)^2$.

Proof

From the hypotheses of the lemma we have that

$$B_\omega v_j = \lambda_j v_j \tag{5.45}$$

also we have seen that

$$(v_k,(I-\omega U)^{-1}(I-\omega L)^{-1}v_j) = \delta_{j,k} \tag{5.46}$$

where $\delta_{j,k}$ is the Kronecker delta.

Next, we express y in terms of v_k, hence

$$y = \sum_{k=1}^{N} \varepsilon_k v_k$$

and if we substitute y in (5.44) it follows that

$$\mu = \frac{\left(\sum_{k=1}^{N} \varepsilon_k v_k,\ A \sum_{j=1}^{N} \varepsilon_j v_j\right)}{\left(\sum_{k=1}^{N} \varepsilon_k v_k,(I-\omega U)^{-1}(I-\omega L)^{-1} \sum_{j=1}^{N} \varepsilon_j v_j\right)}$$

$$= \frac{\left(\sum_{k=1}^{N} \varepsilon_k v_k,(I-\omega L)(I-\omega U)B_\omega \sum_{j=1}^{N} \varepsilon_j v_j\right)}{\left(\sum_{k=1}^{N} \varepsilon_k v_k,(I-\omega U)^{-1}(I-\omega L)^{-1} \sum_{j=1}^{N} \varepsilon_j v_j\right)} . \tag{5.47}$$

Combining now (5.45), (5.46) and (5.47) we obtain the result

$$\mu = \frac{\sum_{k=1}^{N} \varepsilon_k^2 \lambda_k}{\sum_{k=1}^{N} \varepsilon_k^2} = \frac{\lambda_N + \sum_{k=1}^{N-1} \varepsilon_k^2 \lambda_k/\varepsilon_N^2}{1 + \sum_{k=1}^{N-1} \varepsilon_k^2/\varepsilon_N^2} \tag{5.48}$$

which completes the proof of the lemma.

[†]*Here it is assumed that* A=I-L-U.

As a result of Theorem 5.4 and Lemma 5.6 we have that an approximation to $\lambda(B_\omega)$ can be calculated by the expression

$$\mu^{(n)} = \frac{(\delta^{(n)}, A\delta^{(n)})}{(\delta^{(n)}, (I-\omega U)^{-1}(I-\omega L)^{-1}\delta^{(n)})} . \tag{5.49}$$

6.6 THE ADAPTIVE ALGORITHM

In this section a precise definition of the algorithm which uses the PJ-SI and simultaneously improves the parameters ω and $P(B_\omega)$ is given.

From Theorem 5.1 which gives an upper bound $\phi(\omega,v)$ for the smallest eigenvalue $\lambda(B_\omega)$ we conclude that we can determine a lower bound for the P-condition number $P(B_\omega)$ from the relationship

$$p(\omega,v) \leq P(B_\omega) \tag{6.1}$$

where

$$p(\omega,v) = \frac{1}{\omega(2-\omega)\phi(\omega,v)} . \tag{6.2}$$

The lower bound $p(\omega,v)$ of $P(B_\omega)$ as defined in (6.2) indicates that it should be possible to approximate $P(B_\omega)$ by using the PJ-SI method. In order to obtain more information about the role of a and β given by (5.3) and (5.4), respectively, we examine the behaviour of $\phi(\omega,v)$ with respect to these quantities.

We recall from (5.2) that

$$\begin{aligned} \phi(\omega,v) &= \frac{1-\omega a+\omega^2\beta}{1-a} \\ &= \omega + \frac{1-\omega+\omega^2\beta}{1-a} \end{aligned} \tag{6.3}$$

and by a similar approach used to construct Table 4-11.3 we have Table 6.1,

β-Domain	ω-Domain	$\omega^2\beta-\omega+1$
$\beta \leq 1/4$	$0 \leq \omega \leq \omega^*$	≥ 0
	$\omega = \omega^*$	$=0$
	$0 \leq \omega^* \leq \omega$	≤ 0
$\beta > 1/4$	$0 \leq \omega < 2$	>0

TABLE 6.1

BEHAVIOUR OF $\omega^2\beta-\omega+1$ AS A FUNCTION OF ω

where

$$\omega^* = \frac{2}{1+\sqrt{1-4\beta}} \tag{6.4}$$

From (6.2),(6.3) and a cursory examination of Table 6.1 reveals that for $\beta \leq 1/4$ we have i) if $\omega \leq \omega^*$, then $p(\omega,v)$ is maximised when a 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 a is minimised. Finally, if $\beta>1/4$, then $p(\omega,v)$ is maximised when a is maximised for $0 \leq \omega < 2$. Furthermore, if we maximise $\phi(\omega,v)$ then we have an approximation to $P(B_\omega)$ which can be minimised with respect to ω.

Next, we consider some practical aspects of the application of the adaptive algorithm. From the above analysis we see that the quantity a is either maximised or minimised depending upon the range in which the pre-conditioning parameter ω lies. Since now a depends upon the vector v, then it is clear that we have to find two alternative forms of v such that the quantity a always maximises $p(\omega,v)$ independently of the position of ω with respect to ω^*. Let us therefore assume that we have an approximation or an initial guess $v=v^{(1)} \neq 0$ to the eigenvector of $\lambda(B_\omega)$. If A is an L-matrix the quantity a is maximised for $v \geq 0$ (i.e., all the components of v are non-negative). Thus, for $\beta \leq 1/4$ and $\omega \leq \omega^*$ or $\beta>1/4$, where $p(\omega,v)$ is maximised if a is maximised,we may always let v have the form

$$v^{(1)} = (v_1^{(1)}, v_2^{(1)}, v_3^{(1)}, \ldots, v_N^{(1)})^T$$

where $v_i^{(1)} \geq 0$. This choice of v gives

$$a_1 = \frac{(v^{(1)}, DBv^{(1)})}{(v^{(1)}, Dv^{(1)})} \geq 0 \qquad (6.5)$$

if A is an L-matrix.

On the other hand, it is required that the quantity a be minimised (if $\omega^* \leq \omega$ and $\beta \leq 1/4$) which can be achieved if one chooses v to have the alternative form

$$v^{(2)} = (v_1^{(2)}, v_2^{(2)}, v_3^{(2)}, \ldots, v_N^{(2)})^T$$

where

$$v_k^{(2)} = \begin{cases} v_k^{(1)}, & \text{on even points} \\ -v_k^{(1)}, & \text{on odd points.} \end{cases}$$

The above choice of v gives

$$a_2 = \frac{(v^{(2)}, DBv^{(2)})}{(v^{(2)}, Dv^{(2)})} \leq 0 \tag{6.6}$$

which tends to maximise $p(\omega,v)$ if $\omega^* \leq \omega$ and $\beta \leq 1/4$.

Finally, we see that if A is an L-matrix with Property A, then

$$\beta_1 = \frac{(v^{(1)}, DLUv^{(1)})}{(v^{(1)}, Dv^{(1)})} = \frac{(v^{(2)}, DLUv^{(2)})}{(v^{(2)}, Dv^{(2)})} = \beta_2. \tag{6.7}$$

Consequently, a lower bound on $P(B_\omega)$ is given by the following expression

$$P(B_\omega) \geq \begin{cases} \dfrac{1-M}{1-\omega M+\omega^2\beta}, & \text{if } \omega \leq \omega^* \\ \dfrac{1-m}{1-\omega m+\omega^2\beta}, & \text{if } \omega \geq \omega^* \end{cases} \tag{6.8}$$

where

$$\left.\begin{aligned} M &= a_1 \geq 0, \\ m &= a_2 \leq 0 \\ \text{and} \quad \beta &= \beta_1 = \beta_2. \end{aligned}\right\} \tag{6.9}$$

We continue to adhere to the analysis of Section 4.11 concerned with the estimation of good parameters ω and $P(B_\omega)$. Thus, from Theorem 4-11.1 we have that a good choice of the preconditioning parameter ω in the sense of minimising the bound (6.8) is given by (see (4-11.4)) the formula

$$\omega_1 = \begin{cases} \dfrac{2}{1+\sqrt{1-2M+4\beta}} = \omega_M, & \text{if } M \leq 4\beta \\ \dfrac{2}{1+\sqrt{1-4\beta}} = \omega^*, & \text{if } M \geq 4\beta, \end{cases} \tag{6.10}$$

whereas the corresponding value of $P(B_\omega)$ is given by (see (4-11.5))

$$P(B_{\omega_1}) \geq \begin{cases} \dfrac{1}{2}\left(1 + \dfrac{\sqrt{1-2M+4\beta}}{1-M}\right) = \dfrac{1}{2}\left(\dfrac{2-M\omega_M}{(1-M)\omega_M}\right), & \text{if } M \leq 4\beta \\ \dfrac{1+\sqrt{1-4\beta}}{2\sqrt{1-4\beta}} = \dfrac{1}{2-\omega^*}, & \text{if } M \geq 4\beta. \end{cases} \tag{6.11}$$

Another approach which does not require the analysis of Section 4.11 for finding a good estimate to the preconditioning parameter is the use of a direct search technique, such as the Fibonacci method.

Following this approach we can determine an approximation to the optimum parameter ω by minimising

$$P_1(\omega) = \max\{p_1,p_2\} = \max_i \{p(\omega,v^{(i)})\}$$
$$= \max_i \left\{\frac{1-\omega a_i+\omega^2\beta_i}{\omega(2-\omega)(1-a_i)}\right\}, \quad i=1,2. \tag{6.12}$$

As soon as we determine a good estimate $\omega=\omega_1$ we can immediately obtain our first estimate of $P(B_\omega)$ by evaluating

$$P_E(B_{\omega_1}) = P_1(\omega_1) \tag{6.13}$$

and then we can apply the PJ-SI method using ω_1 and $P_E(B_{\omega_1})$. As we have seen, at the same time we can determine $v^{(3)}$ to be another estimate for the eigenvector v and we proceed by forming

$$v^{(4)} = (v_1^{(4)},v_2^{(4)},\ldots,v_N^{(4)})^T$$

where

$$v_k^{(4)} = \begin{cases} v_k^{(3)}, & \text{on even points} \\ -v_k^{(3)}, & \text{on odd points.} \end{cases}$$

At this stage we determine a good estimate $\omega=\omega_2$ by minimising

$$P_2(\omega) = \max\{p_1,p_2,p_3,p_4\}$$
$$= \max_i \{p_i\}, \quad i=1,2,3,4$$

and computing the corresponding estimate of $P(B_{\omega_2})$ by

$$P_E(B_{\omega_2}) = P_2(\omega_2). \tag{6.14}$$

It becomes clear after this analysis that if there are r available estimates of the eigenvector v, then we have 2r vectors $v^{(1)},v^{(2)},\ldots v^{(2r)}$, where for i odd, $v^{(i)}$ is an eigenvector approximation whereas for i even, $v^{(i)}$ is given by

$$v^{(i)} = (v_1^{(i)},v_2^{(i)},\ldots,v_N^{(i)})^T$$

where

$$v_k^{(i)} = \begin{cases} v_k^{(i-1)} & , \text{ on even points} \\ -v_k^{(i-1)} & , \text{ on odd points.} \end{cases}$$

The r^{th} estimate ω_r is determined by minimising the quantity

$$P_r(\omega) = \max_i \{p_i\} \ , \ i=1,2,\ldots,2r$$

and the corresponding estimate $P(B_{\omega_r})$ is given by

$$P_E(B_{\omega_r}) = P_r(\omega_r). \tag{6.15}$$

In addition, we note that since we have

$$p(\omega,v) \leq P(B_\omega) \tag{6.16}$$

it follows that the inequalities

$$P_E(B_{\omega_r}) \leq P(B_{\omega_r}) \leq P(B_{\omega_0}) \tag{6.17}$$

are valid where ω_0 is the optimum preconditioning parameter.

We will now present an adaptive algorithm which uses the PJ-SI method to solve the system Au=b and automatically improve the parameters $(\omega,P(B_\omega))$. The algorithm will use the PJ-SI iterative scheme and a sequence of parameters $(\omega_i,P(B_{\omega_i}))$ which converge to the optimum parameter set $(\omega_0,P(B_{\omega_0}))$. The theoretical basis of the algorithm has been developed in Theorems 5.3, 5.4 and Lemma 5.6.

Algorithm 6.4

1. Choose an initial approximation $u^{(0)}$ such that $\|\varepsilon^{(0)}\|_{A^{\frac{1}{2}}} \leq \|\bar{u}\|_{A^{\frac{1}{2}}}$ and choose a convergence tolerance ζ.

   Also, let

   $$v = (v_1,v_2,\ldots,v_N)^T = (1,1,\ldots,1)^T$$

   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

$$a_i = \frac{(v,DBv)}{(v,Dv)} , \qquad a_{i+1} = \frac{(\hat{v},DB\hat{v})}{(\hat{v},D\hat{v})}$$
$$\beta_i = \frac{(v,DLUv)}{(v,Dv)} , \qquad \beta_{i+1} = \frac{(\hat{v},DLU\hat{v})}{(\hat{v},D\hat{v})} \tag{6.18}$$

if they have not been previously computed.

3. Use a Fibonacci search technique to determine ω by minimising the function

$$P(\omega) = \max_i \left\{ \frac{1-\omega a_i+\omega^2\beta_i}{\omega(2-\omega)(1-a_i)} \right\} \quad \dagger \tag{6.19}$$

for all available pairs (a_i,β_i). Moreover, compute the corresponding value $P_E(B_\omega)$ from the expression

$$P_E(B_\omega) = P(\omega) . \tag{6.20}$$

4. Choose n_q to be the least integer n which satisfies the inequality

$$-\frac{1}{n} \log \frac{2\bar{r}^n}{1+\bar{r}^{2n}} \geq -0.9 \log \bar{r} \tag{6.21}$$

where (see (5-5.19))

$$\bar{r} = \frac{\sqrt{P_E(B_\omega)}-1}{\sqrt{P_E(B_\omega)}+1} . \tag{6.22}$$

5. Iterate n_q times with the PJ-SI method using the latest parameters ω and $P_E(B_\omega)$. After each iteration, check for convergence by computing the pseudo-residual vector

$$\delta^{(n)} = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}), \quad n \leq n_q \tag{6.23}$$

and testing by stopping Procedure III whether or not

$$\frac{P_E(B_\omega)\|\delta^{(n)}\|_{A^{\frac{1}{2}}}}{\|u^{(n)}\|_{A^{\frac{1}{2}}}} \leq \zeta . \tag{6.24}$$

If (6.24) is satisfied terminate the algorithm, otherwise continue to the next step.

---

†*For the unimodality of the function* $P(\omega)$ *see Appendix D.*

6. In this step we test whether we should update the parameters $\omega, P_E(B_\omega)$ or not. From the previous step we have obtained the pseudo-residual vector $\delta^{(n_q)}$, thus we now compute

$$a = \frac{(\delta^{(n_q)}, DB\delta^{(n_q)})}{(\delta^{(n_q)}, D\delta^{(n_q)})}$$
$$\beta = \frac{(\delta^{(n_q)}, DLU\delta^{(n_q)})}{(\delta^{(n_q)}, D\delta^{(n_q)})} \qquad (6.25)$$

and

$$p = \frac{1-\omega a+\omega^2\beta}{\omega(2-\omega)(1-a)} \, . \qquad (6.26)$$

If the following inequality is satisfied

$$p \leq P_E(B_\omega) \, , \qquad (6.27)$$

then go to step 5 and note that the next PJ-SI iteration can be computed from (see (5.42))

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)} + \rho_{n+1}(u^{(n)} + \bar{\rho}\delta^{(n)}) \qquad (6.28)$$

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 $\delta^{(n_q)}$ in (6.23), apply a PJ-SI iteration using (6.28) with the old parameters ω and $P_E(B_\omega)$. Furthermore, let

$$v = \delta^{(n_q)} = \left(\delta_1^{(n_q)}, \delta_2^{(n_q)}, \ldots, \delta_N^{(n_q)}\right)^T$$

and

$$\hat{v} = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_N)^T$$

where

$$\hat{v}_k = \begin{cases} \delta_k^{(n_q)} & \text{, on even points} \\ -\delta_k^{(n_q)} & \text{, on odd points.} \end{cases}$$

Then, set i=i+2 and go to step 2 to compute new quantities for a and β in order to update ω and $P_E(B_\omega)$. Evidently, in step 2, a_i and β_i have already been computed by (6.25) in step 6.

Next we proceed to make various comments and justify some points in the above algorithm. First, we note that if the matrix A is an L-matrix with Property A, then

$$a_{i+1} = -a_i \quad \text{and} \quad \beta_{i+1} = \beta_i \ . \tag{6.29}$$

which implies that in step 2 we only need to compute a_i and β_i. The use of the Fibonacci search technique for the minimisation of $P(\omega)$ does not require the knowledge of the analysis of $P(B_\omega)$ (see Section 4.11) and it could be regarded as a hint for the use of such techniques for more general and complex problems where the mathematical analysis may be laborious.

In step 4 we have chosen n_q such that the average rate of convergence after n_q iterations is 90% of the asymptotic average rate of convergence.

Furthermore, in order to justify step 6 and in particular the criterion (6.27) as to whether or not we should modify the parameters, we reason as follows.

We have seen that

$$P_n(\mathcal{H}_\omega) = \frac{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2B_\omega}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)}{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)} \tag{6.30}$$

and that the virtual spectral radius of PJ-SI iterative procedure is given by

$$\bar{S}(P_n(\mathcal{H}_\omega)) = \max_{\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda)| \tag{6.31}$$

where

$$P_n(\lambda) = \frac{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2\lambda}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)}{T_n\left(\frac{\Lambda(B_\omega)+\lambda(B_\omega)-2}{\Lambda(B_\omega)-\lambda(B_\omega)}\right)} \ . \tag{6.32}$$

Let us assume that $\underline{R}$ is an estimate of $\lambda(B_\omega)$ and define

$$P_n(\lambda,\underline{R}) = \frac{T_n\left(\frac{\Lambda(B_\omega)+\underline{R}-2\lambda}{\Lambda(B_\omega)-\underline{R}}\right)}{T_n\left(\frac{\Lambda(B_\omega)+\underline{R}-2}{\Lambda(B_\omega)-\underline{R}}\right)} \ , \tag{6.33}$$

then by the minimax theorem of Chebyshev polynomials (see Appendix C)

we have

$$\max_{\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda)| \leq \max_{\underline{R}\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda,\underline{R})|. \tag{6.34}$$

In addition, if $\phi(\omega,v)\geq\lambda_E(B_\omega)\geq\lambda(B_\omega)$ or in other words if

$$p \leq P_E(B_\omega) \leq P(B_\omega),$$

then we also have

$$\max_{\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda)| \leq \max_{\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda,\underline{R})| \leq \max_{\lambda(B_\omega)\leq\lambda\leq\Lambda(B_\omega)} |P_n(\lambda,p)|. \tag{6.35}$$

Consequently, if $p\leq P_E(B_\omega)$, the parameters ω and $P_E(B_\omega)$ are not changed since the rate of convergence is not likely to be improved. However, we cannot be certain that a change of parameters in this case would not improve the rate of convergence.

6.7 NUMERICAL RESULTS

In order to examine the performance of the Algorithm 6.4 as compared with the Algorithms 6.1 and 6.2 we solved the six problems considered in Section 4.8 by applying the PJ-SI iterative procedure which determines adaptively the parameters ω and $P(B_\omega)$ (Algorithm 6.4). As a starting vector we used the one which has all its components equal to zero, whereas the convergence tolerance was taken $\zeta=10^{-6}$. In all cases the natural ordering was used. In Table 7.1 we present the number of iterations required to satisfy stopping Procedure III for the PJ-SI method with optimum, estimated and adaptive parameters. The subscripts on the number of iterations given for the adaptive algorithm refer to the number of parameter changes which were necessary to attain convergence. Under the headings n_A/n_0 and n_E/n_0 we give the ratio of the number of iterations, n(adaptive)/n(optimum) and n(estimated)/n(optimum), respectively where n(adaptive) is the effective number of iterations taking into account the additional work required for updating the parameters. This is done if we convert the additional operations that are performed each time the parameters are changed into the corresponding number of iterations. It can be seen that four parameter changes are approximately equivalent to three PJ-SI iterations of work performed. Consequently, from this information we can work out n_A for each particular case, e.g. if three parameter changes are required, then the number of iterations for the adaptive procedure should effectively be increased by approximately 2+1/4 iterations. Since for the determination of the optimum and estimated parameters a considerable preprocessing is required, their number of iterations should also be increased. However, even if we do not take into account this additional work we can safely state that the adaptive algorithm performs better than the estimated one in half of the problems considered (see Table 7.1).

Furthermore, from Table 7.1 we see that in general the adaptive algorithm requires effectively about 25% more iterations than the PJ-SI

Problem	h=1/20					h=1/40					h=1/60				
	n_0	n_A	n_A/n_0	n_E	n_E/n_0	n_0	n_A	n_A/n_0	n_E	n_E/n_0	n_0	n_A	n_A/n_0	n_E	n_E/n_0
1	18	19_3	1.18	21	1.17	28	29_4	1.14	32	1.14	36	39_3	1.16	41	1.14
2	11	12_2	1.23	12	1.09	17	18_3	1.19	18	1.06	22	24_4	1.23	23	1.05
3	18	19_3	1.18	33	1.83	28	31_3	1.19	51	1.82	35	40_3	1.21	66	1.89
4	18	19_3	1.18	24	1.33	28	29_4	1.14	37	1.32	35	39_5	1.18	47	1.34
5	19	21_3	1.22	34	1.79	29	31_4	1.17	54	1.86	38	41_4	1.16	69	1.81
6	11	12_1	1.16	13	1.18	17	18_2	1.15	18	1.06	22	26_2	1.25	28	1.27

TABLE 7.1

RESULTS OF APPLYING ALGORITHM 6.4 TO PROBLEMS 1-6 USING PROCEDURE III

with optimum parameter iterative scheme. Another observation is that for problems 2 and 6 the PJ-SI method with estimated parameters gives approximately the same results for the different mesh sizes as the PJ-SI method with optimum parameters. This happens because we have replaced M by $2\sqrt{\bar{\beta}}$ in determining the estimated parameters.

In Table 7.2 we can see a more detailed presentation as to how Algorithm 6.4 performs by including intermediate results of the various stages. Thus we can observe how the parameters are updated and how close agreement can be obtained with the optimum ω and $P(B_\omega)$. The number of cycles show how many times the η_q-iterations are repeated without changing the parameters. For comparison reasons we have also included the optimum parameters in parentheses.

From Table 7.2 we see that the maximum of four parameter changes are needed for each problem to be solved. In particular, for problem 6 a maximum of two changes were required for all the different mesh sizes. Furthermore, we observe that the parameters ω and $P(B_\omega)$ obtained adaptively were quite satisfactory approximations to the optimum parameters, especially for small mesh sizes in all the cases considered.

In Figure 7.1 we plot the logarithm of the number of iterations versus $\log h^{-1}$ for all the problems. This was carried out for the PJ-SI method using optimum, estimated and adaptive parameters. From this figure we see that the rate of convergence (of the three different approaches) is approximately $O(h^{\frac{1}{2}})$ convergence even when A(x,y) and C(x,y) do not belong to the class $C^{(2)}$.

From the above analysis of the obtained results we see that for problems 1,2 and 6 using PJ-SI with estimated parameters (determined by bounds on S(B) and S(LU) (see (4-11.4) and (4-11.5)) required fewer iterations than the adaptive scheme (see Table 7.1). We also observe that in these cases the coefficients A(x,y) and C(x,y) both belong to the class $C^{(2)}$ and

PROBLEM 1			
h^{-1}	20	40	60
n_1	5	6	7
cycles	1	1	1
ω_1	1.6228	1.7224	1.7688
$P_E(B_{\omega_1})$	2.7079	3.6425	4.3568
n_2	8	10	11
cycles	1	1	1
ω_2	1.7794	1.8753	1.9041
$P_E(B_{\omega_2})$	4.9675	8.1766	10.6848
n_3	8	11	14
cycles	1	1	2
ω_3	1.7428(1.7641)	1.8733	1.9178(1.9157)
$P_E(B_{\omega_3})$	5.1949(5.2604)	10.1336	14.9163(15.2207)
n_4		11	
cycles		1	
ω_4		1.8749(1.8750)	
$P_E(B_{\omega_4})$		10.1464(10.0806)	
PROBLEM 2			
n_1	4	6	7
cycles	2	1	1
ω_1	1.5050	1.6653	1.7337
$P_E(B_{\omega_1})$	2.0203	2.9881	3.7546
n_2	5	6	8
cycles	1	1	1
ω_2	1.5570(1.5888)	1.7162	1.7930
$P_E(B_{\omega_2})$	2.2575(2.4248)	3.5242	4.8307
n_3		7	8
cycles		1	1
ω_3		1.7454(1.7668)	1.8198
$P_E(B_{\omega_3})$		3.9271(4.2790)	5.5484
n_4			8
cycles			1
ω_4			1.8278(1.8386)
$P_E(B_{\omega_4})$			5.8083(6.1958)

TABLE 7.2

SUCCESSIVE PARAMETERS OBTAINED BY ALGORITHM 6.4

AND STOPPING PROCEDURE III

PROBLEM 3			
h^{-1}	20	40	60
n_1	5	6	7
cycles	1	1	1
ω_1	1.6237	1.7228	1.7690
$P_E(B_{\omega_1})$	2.7268	3.6563	4.3683
n_2	8	10	12
cycles	1	1	1
ω_2	1.7845	1.8791	1.9068
$P_E(B_{\omega_2})$	5.1868	8.5747	11.1830
n_3	8	11	14
cycles	1	2	2
ω_3	1.7403(1.7652)	1.8732(1.8756)	1.9190(1.9163)
$P_E(B_{\omega_3})$	5.2454(5.3763)	10.3795(10.3200)	15.3582(15.2207)
PROBLEM 4			
n_1	5	6	7
cycles	1	1	1
ω_1	1.6231	1.7225	1.7688
$P_E(B_{\omega_1})$	2.7062	3.6410	4.3556
n_2	8	10	11
cycles	1	1	1
ω_2	1.7784	1.8748	1.9039
$P_E(B_{\omega_2})$	4.9342	8.1558	10.6306
n_3	8	11	14
cycles	1	1	2
ω_3	1.7426(1.7624)	1.8728	1.9173(1.9143)
$P_E(B_{\omega_3})$	5.1690(5.2301)	10.0737	14.8158(14.7929)
n_4		11	
cycles		1	
ω_4		1.8749(1.8748)	
$P_E(B_{\omega_4})$		10.1030(10.0200)	

TABLE 7.2 (CONTINUED)

PROBLEM 5			
h^{-1}	20	40	60
n_1	6	7	7
cycles	1	1	1
ω_1	1.5642	1.6766	1.7294
$P_E(B_{\omega_1})$	2.9198	3.9643	4.7608
n_2	8	11	12
cycles	1	1	1
ω_2	1.7669	1.8692	1.8871
$P_E(B_{\omega_2})$	5.6765	9.8403	11.4930
n_3	8	12	14
cycles	1	1	1
ω_3	1.7593(1.7479)	1.8699	1.9157
$P_E(B_{\omega_3})$	5.7465(5.8173)	11.1733	16.3066
n_4		12	14
cycles		1	1
ω_4		1.8738(1.8665)	1.9142(1.9093)
$P_E(B_{\omega_4})$		11.1942(11.1732)	16.6335(16.5289)
PROBLEM 6			
n_1	5	7	8
cycles	3	2	2
ω_1	1.6065(1.6097)	1.6817	1.7178
$P_E(B_{\omega_1})$	2.5415(2.5221)	4.1905	5.3461
n_2		7	9
cycles		1	2
ω_2		1.7700(1.7820)	1.8438(1.8490)
$P_E(B_{\omega_2})$		4.3481(4.4543)	6.4037(6.4020)

TABLE 7.2 (CONTINUED)

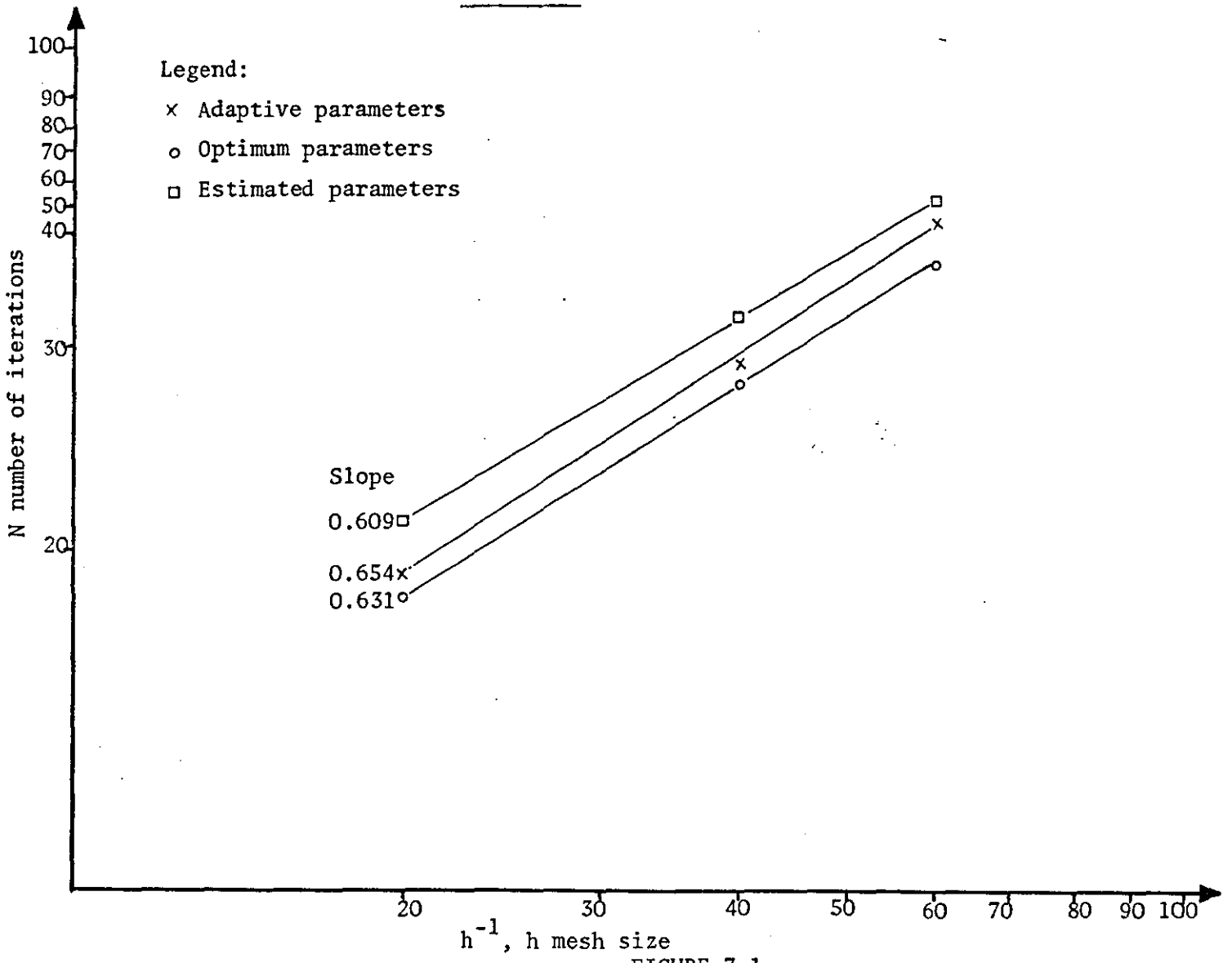


FIGURE 7.1

DETERMINATION OF RATE OF CONVERGENCE ATTAINED FOR THE SIX PROBLEMS USING ALGORITHM 6.4 AND PJ-SI WITH OPTIMUM AND ESTIMATED PARAMETERS

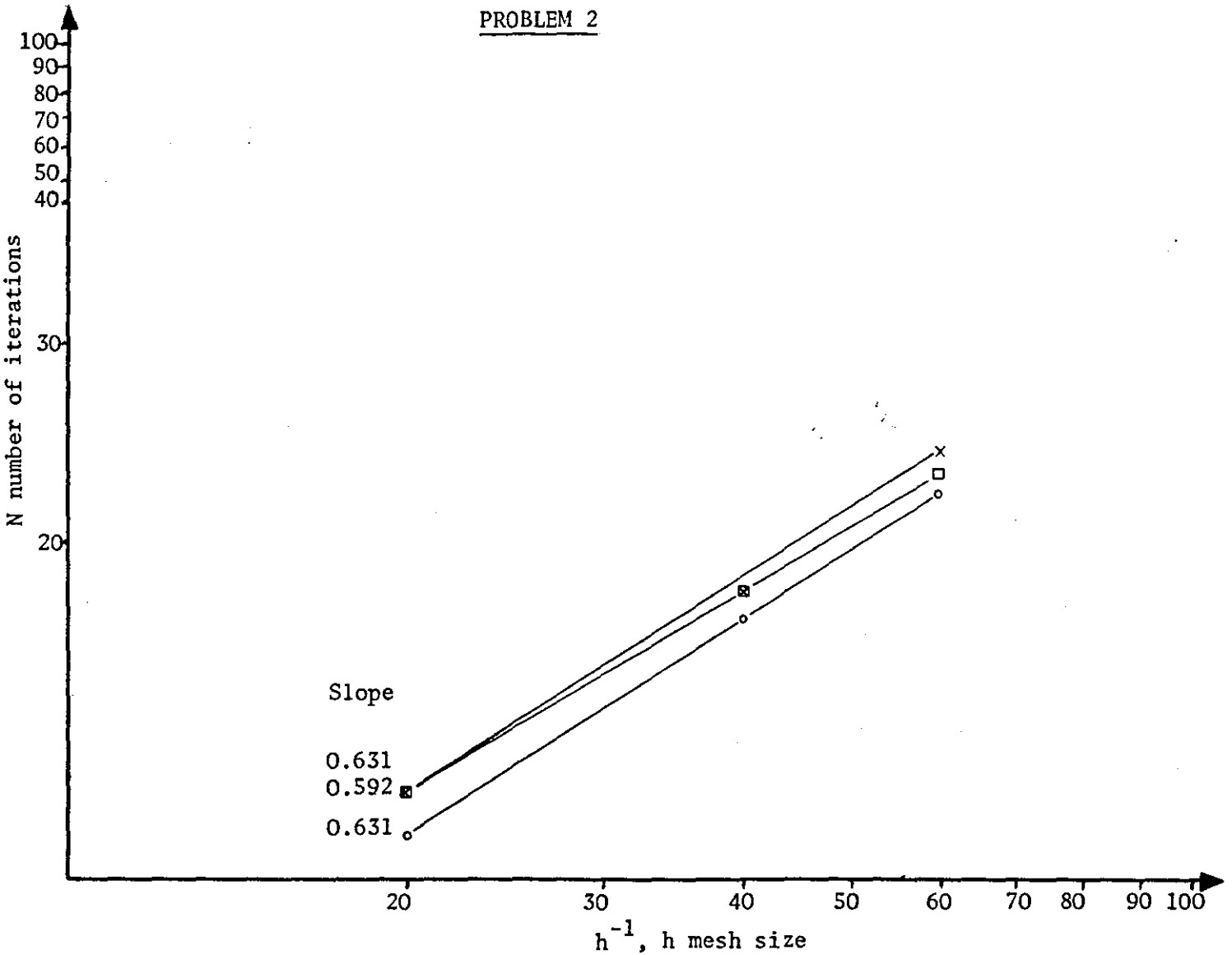


FIGURE 7.1 (CONTINUED)

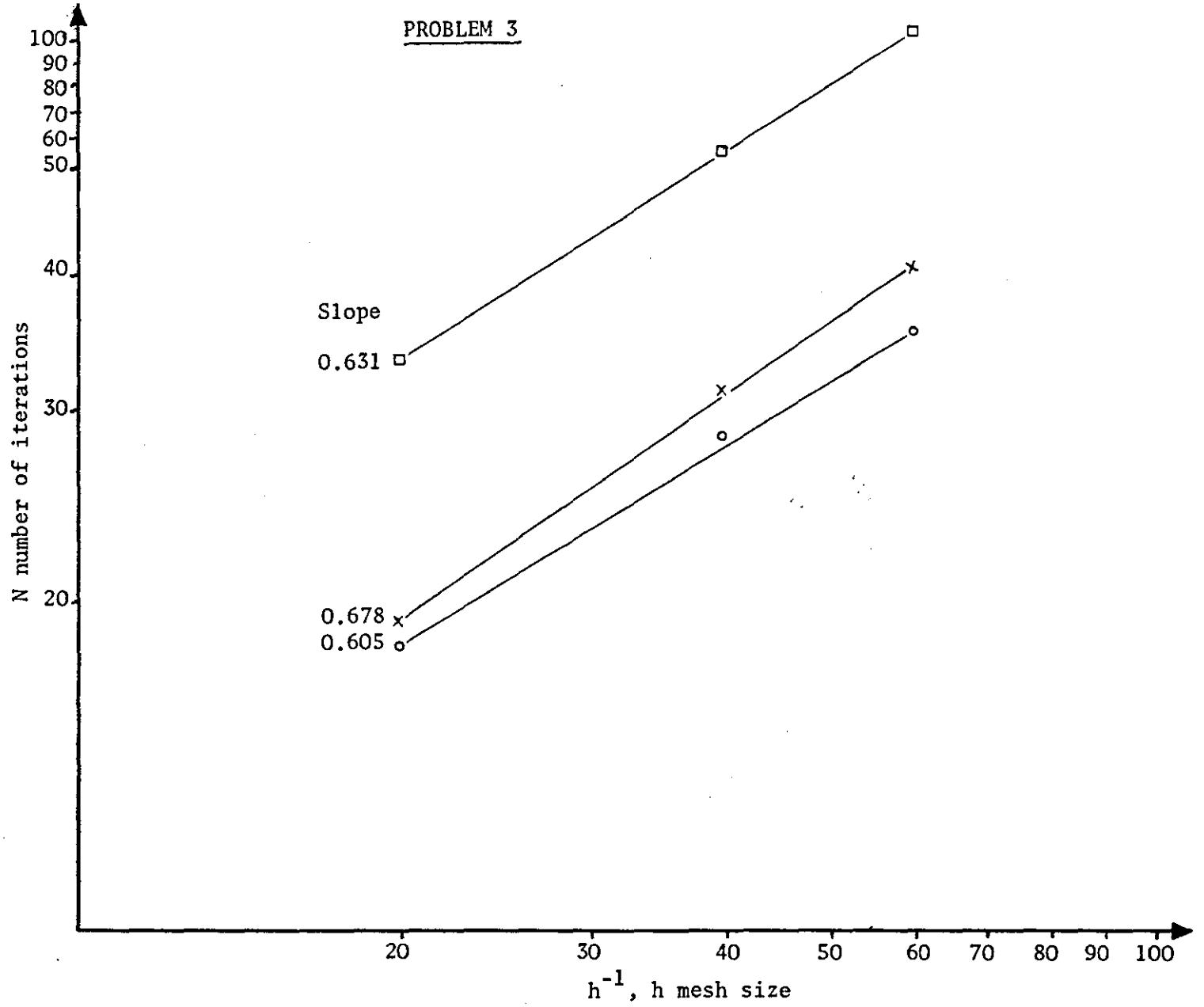


FIGURE 7.1 (CONTINUED)

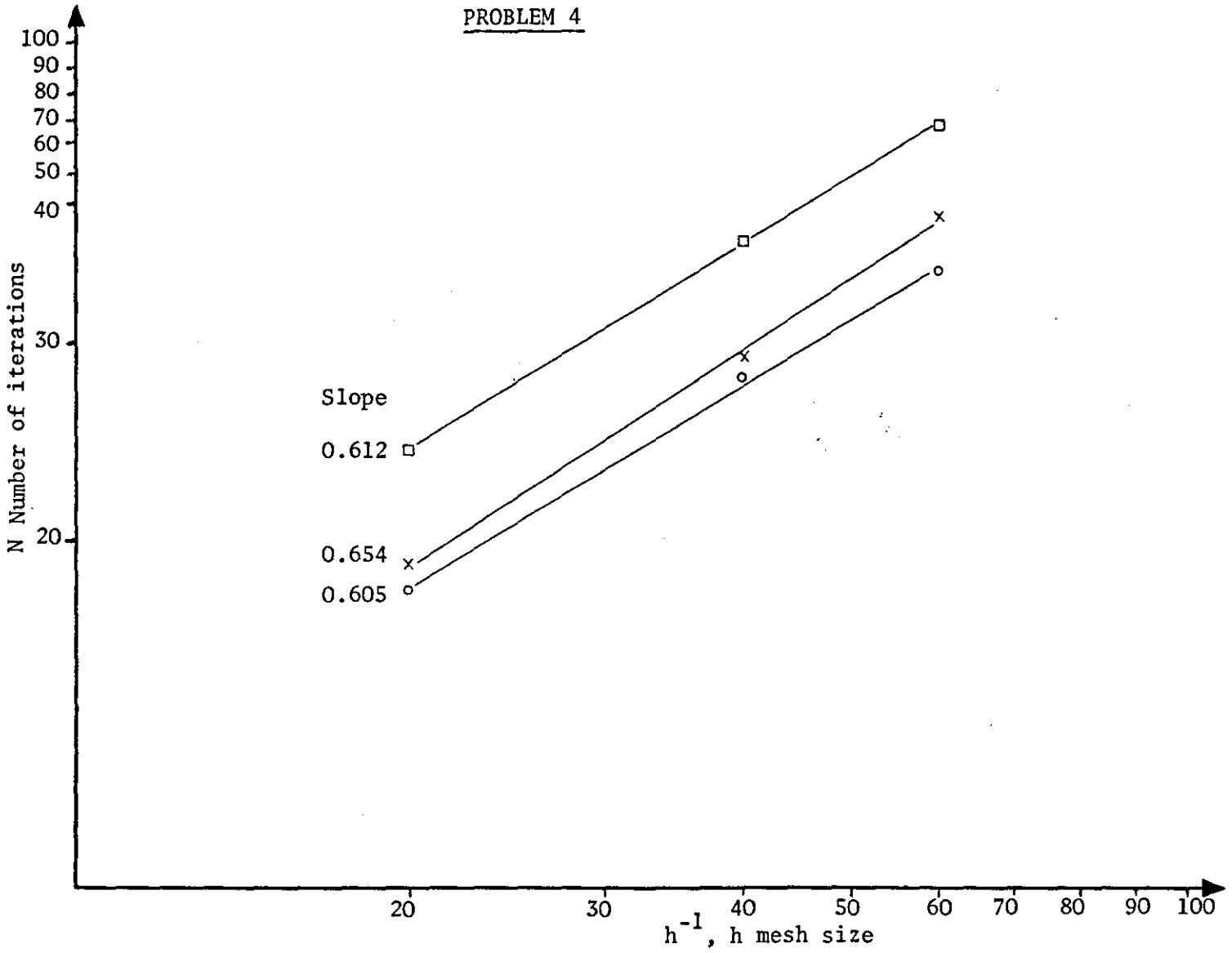


FIGURE 7.1 (CONTINUED)

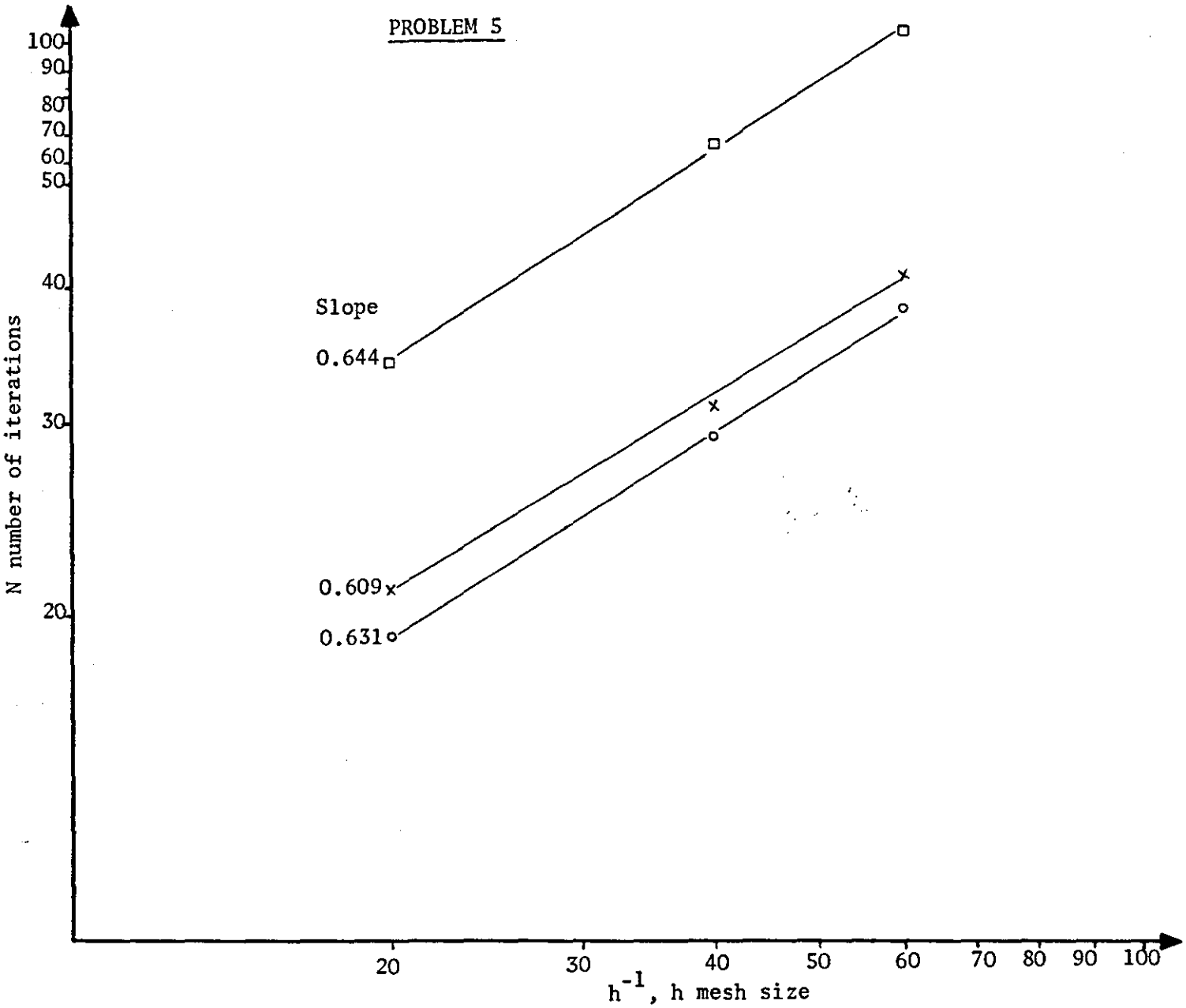


FIGURE 7.1 (CONTINUED)

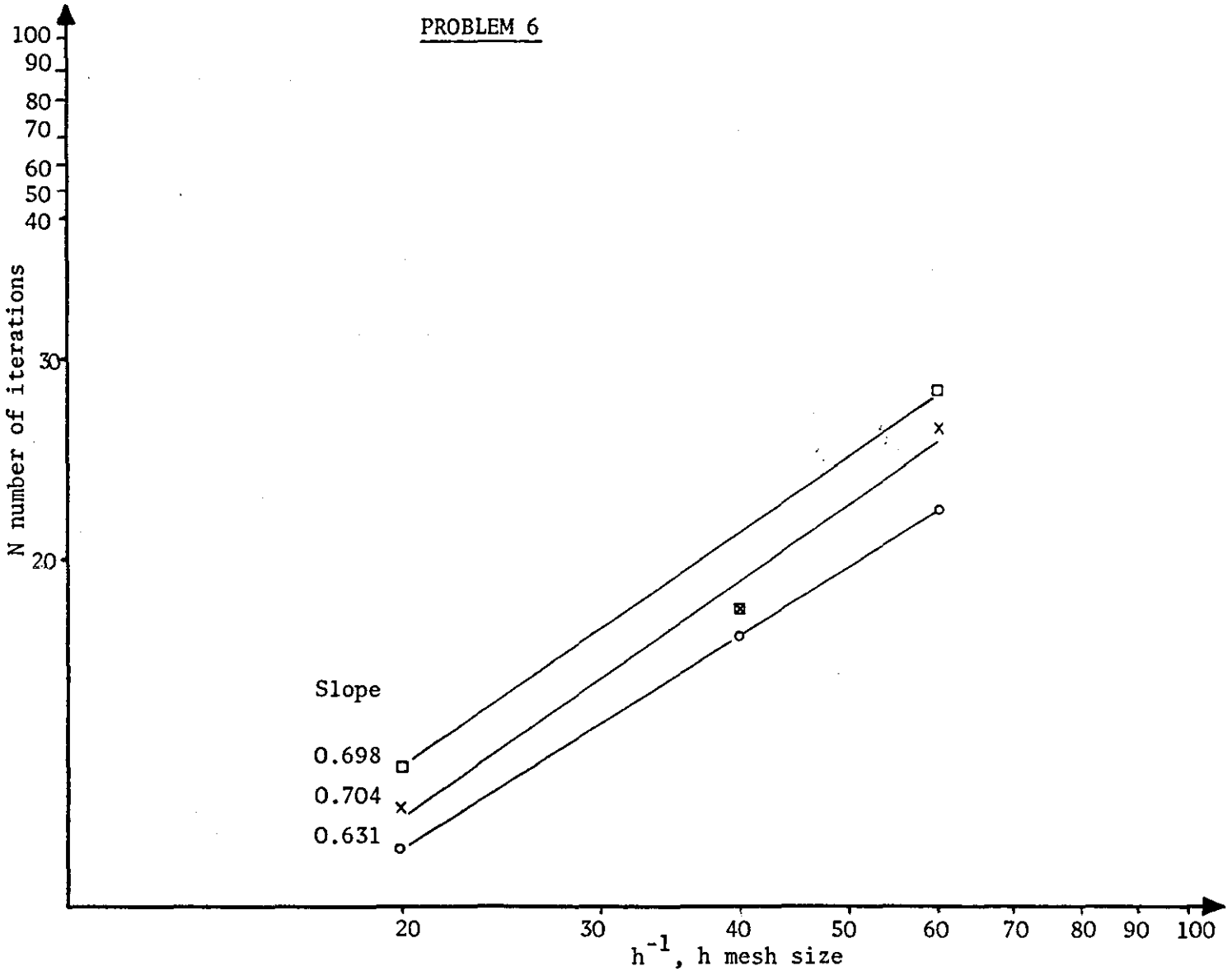


FIGURE 7.1 (CONTINUED)

that the bound $\bar{\beta}$ of S(LU) is less than or equal to 1/4 (see Table 4.1). On the other hand, for problems 3,4 and 5 the adaptive algorithm performed better than the estimated-parameter PJ-SI method. It is characteristic that in each of the problems 3-5 we either have $\bar{\beta} \geq 1/4$ or the coefficients A or C do not belong to class $C^{(2)}$.

Finally, we note that stopping Procedure III used in the adaptive algorithm was derived in Section 4.3 assuming that ω is fixed, hence we rely on the fact that $P(B_\omega)$ is a continuous function of ω for Algorithm 6.4 to yield acceptable results.

CHAPTER 7

ALTERNATING DIRECTION PRECONDITIONING TECHNIQUES FOR THE NUMERICAL SOLUTION OF THE ELLIPTIC SELF-ADJOINT SECOND ORDER AND BIHARMONIC EQUATIONS

7.1 INTRODUCTION

In Section 4.2 we exhibited a general idea of how one should proceed using the preconditioning technqiues in order to construct various iterative schemes for the solution of Au=b associated with the splitting of the coefficient matrix A. A result of this (when the matrix A had the splitting I-L-U) was to see the strong need for reconsidering the known iterative schemes such as to be consistent with the preconditioning approach. We therefore were able to produce new iterative methods (EGS, ESOR and PSD) which proved to be more effective than their known unextrapolated counterparts (GS, SOR and SSOR).

In this chapter we will attempt to follow a similar strategy (as the one in Chapter 4) by assuming a different well known splitting of A, the one on which the Alternating Direction Implicit (ADI) methods have been based (see Peaceman and Rachford [1955], Douglas [1955], Douglas and Rachford [1956]).

The ADI methods are somewhat similar to single line iterative methods with alternating directions. In order to see the ADI schemes as preconditioned methods we define the splitting of A by considering the discrete analogue of the self-adjoint partial differential equation

$$\frac{\partial}{\partial x}(A(x,y)\frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(C(x,y)\frac{\partial U}{\partial y}) + F(x,y)U = G \tag{1.1}$$

where A,C,F and G are such that $A>0$, $C>0$ and $F\leq 0$ for all $(x,y)\in R$ and R is the region under consideration. The five point finite difference analogue of (1.1) using a uniform mesh size h is given by

$$H_0[u](x,y)+V_0[u](x,y)+E_0[u](x,y) = -h^2G \tag{1.2}$$

where

$$H_0[u](x,y) = [A(x+\tfrac{1}{2}h,y)+A(x-\tfrac{1}{2}h,y)]u(x,y) - A(x+\tfrac{1}{2}h,y)u(x+h,y)-A(x-\tfrac{1}{2}h,y)u(x-h,y) \tag{1.3}$$

$$V_0[u](x,y) = [C(x,y+\tfrac{1}{2}h)+C(x,y-\tfrac{1}{2}h)]u(x,y) - C(x,y+\tfrac{1}{2}h)u(x,y+h)-C(x,y-\tfrac{1}{2}h)u(x,y-h) \tag{1.4}$$

$$E_0[u](x,y) = -h^2F(x,y)u(x,y) \tag{1.5}$$

From equation (1.1) we see that $H_0[u](x,y)$ and $V_0[u](x,y)$ correspond

to the discrete analogues of the terms $-h^2\frac{\partial}{\partial x}(A(x,y)\frac{\partial U}{\partial x})$ and $-h^2\frac{\partial}{\partial y}(C(x,y)\frac{\partial U}{\partial y})$, respectively. Evidently, the difference equation (1.2) can be written in the matrix form

$$Au = (H_0+V_0+E_0)u = b \tag{1.6}$$

where the matrices H_0,V_0 and E_0 correspond to the operators $H_0[u]$,$V_0[u]$ and $E_0[u]$, respectively. Moreover, by ordering the mesh points by rows, H_0 is tri-diagonal and V_0 can be made so by permution of its rows and columns whereas E_0 is a non-negative diagonal matrix. If we let

$$Au = (H+V)u = b \tag{1.7}$$

where

$$H = H_0+\tfrac{1}{2}E_0 \quad \text{and} \quad V = V_0+\tfrac{1}{2}E_0\ , \tag{1.8}$$

then it can be easily seen that (e.g. see Varga [1962]) H and V are real, symmetric, diagonally dominant matrices with positive diagonal entries and non-positive off-diagonal entries (see (1.3) and (1.4)).

Once we have defined the splitting of A (see (1.7) and (1.8)) we let R, the conditioning matrix, have the following general form (which is similar to (4-2.4) when A had the form (4-2.10))

$$R = (I+rH)(I+r'V) \tag{1.9}$$

where r,r' are real preconditioning parameters. The iterative scheme associated with the conditioning matrix defined by (1.9) is given by

$$u^{(n+1)} = u^{(n)} + \tau(I+r'V)^{-1}(I+rH)^{-1}(b-Au^{(n)}) \tag{1.10}$$

and will be referred to as the Modified Alternating Direction Preconditioning method (MADP method). We note that for the different values of the involved parameters in (1.10) we obtain the known ADI schemes presented in Table 1.1, where $\bar{a}$ and $\bar{b}$ are the minimum and maximum eigenvalues of the preconditioned matrix $R^{-1}A$, respectively, i.e.,

$$\bar{a} \leq \lambda(R^{-1}A) \leq \bar{b}.$$

From Table 1.1 we see that the ADI schemes can be regarded as "preconditioned methods" and as such the effectiveness of the conditioning matrix is not exploited in the DR-ADI and PR-ADI iterative methods since τ does not take

its optimum value (see Section 4.2), whereas in the EADI scheme we have $\tau_0'=r_{opt}\tau_0$ which indicates that the presence of r_{opt} does not produce any effect on the rate of convergence and therefore can be omitted. Thus we can immediately state that the EADI method degenerates into the ADP method at the optimum stage.

Preconditioning Parameters		τ	Conditioning Matrix	Iterative Method
r	r'	1	$(I+rH)(I+r'V)$	Douglas-Rachford ADI (DR-ADI)
r	r'	r+r'	"	Peaceman-Rachford ADI (PR-ADI)
r	r	$2r/(\bar{a}+\bar{b})$	$(I+rH)(I+rV)$	Extrapolated Alternating Direction Implicit method (EADI method) (see Hadjidimos [1975])
r	r	$2/(\bar{a}+\bar{b})$	"	Alternating Direction Preconditioning method (see Gane and Evans [1974]) (ADP method)

TABLE 1.1

Another question which emerges (see Section 4.3) is the study of the iterative scheme which is produced if we let r'=0 in (1.10) i.e., the iterative scheme

$$u^{(n+1)} = u^{(n)} + \tau(I+rH)^{-1}(b-Au^{(n)}). \tag{1.11}$$

However, before we proceed any further, we impose some additional conditions on the matrices H_0,V_0 which characterise the "commutative case" (see Birkhoff et al [1962]). In the remainder of this chapter we will assume that the matrices H_0,V_0 and E_0 of (1.6) satisfy the conditions

$$\left.\begin{aligned} &H_0V_0 = V_0H_0 \\ &E_0 = \sigma I, \text{ where } \sigma \text{ is a non-negative constant} \\ &H_0 \text{ and } V_0 \text{ are similar to non-negative diagonal matrices.} \end{aligned}\right\} \tag{1.12}$$

If these conditions are satisfied, then

$$\left.\begin{aligned} &HV = VH \\ \text{and} \quad &H \text{ and } V \text{ are similar to non-negative diagonal matrices.} \end{aligned}\right\} \tag{1.13}$$

Moreover, I+rH and I+rV are non-singular for any r>0 whereas as observed by Birkhoff et al [1962], one can obtain matrices H_0, V_0 and E_0 satisfying (1.12) from partial differential equations of the form

$$\frac{1}{E_2(x)}\ \frac{\partial}{\partial x}(E_1(x)\frac{\partial U}{\partial x}) + \frac{1}{F_1(y)}\ \frac{\partial}{\partial y}(F_2(y)\frac{\partial U}{\partial y})\ - kU = \frac{G(x,y)}{E_2(x)F_1(y)} \tag{1.14}$$

in the rectangle $R: 0 \leq x \leq L_x, \quad 0 \leq y \leq L_y,$ (1.15)

where the functions $E_1(x), E_2(x), F_1(y), F_2(y)$ are assumed to be continuous and positive in R, and $k \geq 0$. Evidently, (1.14) is a special case of (1.1) with $A(x,y)=E_1(x)F_1(y)$, $C(x,y)=E_2(x)F_2(y)$ and $F(x,y)=-kE_2(x)F_1(y)$.

7.2 SOME CONSIDERATIONS ON THE ITERATIVE SCHEME (1.11)

Let us consider the solution of the partial differential equation (1.14) defined in the rectangular region R given by (1.15) and the boundary condition

$$U(x,y) = g(x,y) , \qquad (x,y)\in\partial R \tag{2.1}$$

where $g(x,y)$ is a prescribed function on the boundary ∂R of R. A difference equation leading to matrices H_0, V_0 and E_0 satisfying the properties given by (1.12) is obtained as follows: First, we impose a uniform grid of mesh size h_x and h_y in the x- and y-directions, respectively, such that

$$N_a = \frac{L_a}{h_a} , \qquad a=x,y$$

where N_a is an integer. Next, we use the difference equation

$$H_0'[u](x,y)+V_0'[u](x,y)+E_0'[u](x,y) = -h_x h_y G(x,y)/(E_2(x)F_1(y)) \tag{2.2}$$

where

$$H_0'[u](x,y) = \left(\frac{h_y}{h_x}\right)\left\{\left[\frac{E_1(x+\frac{1}{2}h_x)+E_1(x-\frac{1}{2}h_x)}{E_2(x)}\right]u(x,y) - \frac{E_1(x+\frac{1}{2}h_x)}{E_2(x)}u(x+h_x,y) - \frac{E_1(x-\frac{1}{2}h_x)}{E_2(x)}u(x-h_x,y)\right\} \tag{2.3}$$

$$V_0'[u](x,y) = \left(\frac{h_y}{h_x}\right)\left\{\left[\frac{F_2(y+\frac{1}{2}h_y)+F_2(y-\frac{1}{2}h_y)}{F_1(y)}\right]u(x,y) - \frac{F_2(y+\frac{1}{2}h_y)}{F_1(y)}u(x,y+h_y) - \frac{F_2(y-\frac{1}{2}h_y)}{F_1(y)}u(x,y-h_y)\right\} \tag{2.4}$$

and

$$E_0'[u](x,y) = h_x h_y k u(x,y) . \tag{2.5}$$

By using the natural ordering, the difference equation (2.2) can be written in the matrix form (1.6) where H_0, V_0 and E_0 correspond now to the operators $H_0'[u], V_0'[u]$ and $E_0'[u]$, respectively.

As it was indicated earlier, we will continue our study on these preconditioned iterative schemes which are constructed by using the splitting (1.7) of A. As a first step in this section we will consider the iterative scheme (1.11) which from now on will be referred to as scheme (I). The motivation for the examination of scheme (I) can be justified by noting its similarity with the line ESOR method (see Chapter 5). Moreover, it would be desirable to obtain some information

about the effectiveness of the conditioning matrix

$$R = I+rH \tag{2.6}$$

as compared with the conditioning matrix given by (1.9) since the work involved in scheme (I) is considerably less than in the ADP method.

An alternative form of scheme (I) is given by

$$u^{(n+1)} = Q_{\tau,r}u^{(n)} + q \tag{2.7}$$

where

$$Q_{\tau,r} = I-\tau(I+rH)^{-1}A \tag{2.8}$$

and

$$q = \tau(I+rH)^{-1}b. \tag{2.9}$$

By Theorem 3-1.4 we note that scheme (I) is completely consistent if $\tau\neq 0$. Moreover, we obtain a more computable form if we write (2.7) as

$$(I+rH)u^{(n+1)} = [I+(r-\tau)H-\tau V]u^{(n)}+\tau b \tag{2.10}$$

where we now have to solve a tri-diagonal system which can be easily solved (see Cuthill and Varga [1959]) since the right hand side vectors are all known. Evidently, the preconditioned matrix of the iterative scheme (I) is given by

$$B_r = (I+rH)^{-1}A \tag{2.11}$$

where it can be seen that the matrix B_r is positive definite for all $r\geq 0$. For convergence, we prove the following theorem.

Theorem 2.1

If the matrices H,V defined by (2.3),(2.4) and (1.8) satisfy the conditions (1.13) and if their eigenvalues μ,ν respectively lie in the range

$$0<a\leq\mu,\nu\leq b, \tag{2.12}$$

then for $r\in[0,\infty)$ the iterative scheme (I) converges if and only if the parameters r and τ take values from the intervals I_r and I_τ, respectively defined as follows

$$I_r \equiv [0,1/b] \quad \text{and} \quad I_\tau \equiv (0,r+1/b) \tag{2.13}$$

or

$$I_r \equiv [1/b,\infty) \quad \text{and} \quad I_\tau \equiv (0,2(1+ra)/(a+b)). \tag{2.14}$$

Proof

From the hypotheses we have that H and V satisfy the conditions (1.13), hence there exist a set of linear independent vectors which are eigenvectors both of H and V.

Let v be any such vector and let

$$Hv = \mu v, \quad Vv = \nu v , \tag{2.15}$$

then the eigenvalues of B_r will be given by the expression

$$\lambda(\mu,\nu,r) = \frac{\mu+\nu}{1+r\mu} . \tag{2.16}$$

From (2.16) we see that if $r\geq 0$, then $\lambda>0$ and B_r is positive definite.

A sufficient and necessary condition for the iterative scheme (I) to converge is that the parameter τ to lie in the range

$$0<\tau<2/\max_{\mu,\nu}\{\lambda(\mu,\nu,r)\}. \tag{2.17}$$

From the above inequalities we see that we have to determine the largest eigenvalue of B_r with respect to μ,ν and for $r\geq 0$. We therefore study the behaviour of $\lambda(\mu,\nu,r)$ as a function of μ,ν.

Taking partial derivatives of $\lambda(\mu,\nu,r)$ with respect to μ and ν we obtain the following expressions

$$\operatorname{sign}\left(\frac{\partial\lambda}{\partial\mu}\right) = \operatorname{sign}(1-\nu r) \tag{2.18}$$

and

$$\operatorname{sign}\left(\frac{\partial\lambda}{\partial\nu}\right) = +1. \tag{2.19}$$

From (2.19) and for fixed $r\geq 0$ the continuous function $\lambda(\mu,\nu,r)$ is an increasing function of ν. We therefore conclude that if ν satisfies the inequalities (2.12), then

$$\max_{\mu,\nu}\{\lambda(\mu,\nu,r)\} \leq \max_{\mu}\{\lambda(\mu,b,r)\} \tag{2.20}$$

and from (2.18) we easily obtain the following expression for the largest eigenvalue of B_r,

$$\max\{\lambda(\mu,\nu,r)\} = \begin{cases} \lambda(b,b,r), & \text{if } 0\leq r\leq 1/b \\ \lambda(a,b,r), & \text{if } 1/b\leq r<\infty. \end{cases} \tag{2.21}$$

By combining now (2.17),(2.16) and (2.21) we can easily see that the proof of the theorem is complete.

For the determination of τ and r such as the iterative scheme (I) attains its maximum rate of convergence we have first to select r to minimise the function p(a,b,r) where

$$P(B_r) = \frac{\lambda_M}{\lambda_m} \equiv p(a,b,r) \tag{2.22}$$

$$\lambda_M = \max_{\mu,\nu}\{\lambda(\mu,\nu,r)\}, \quad \lambda_m = \min_{\mu,\nu}\{\lambda(\mu,\nu,r)\} \tag{2.23}$$

and secondly to calculate τ from the expression

$$\tau_O = \frac{2}{\lambda_M+\lambda_m} . \tag{2.24}$$

Theorem 2.2

Let H,V be the matrices defined by (2.3),(2.4) and (1.8) with eigenvalues μ,ν, respectively such that

$$0<a\leq\mu,\nu\leq b. \tag{2.25}$$

Then for any $r\in[0,\infty)$ the P-condition number of B_r is given by

$$P(B_r) = \begin{cases} \frac{b(1+ra)}{a(1+rb)} , & \text{if } 0\leq r\leq 1/b \\ \frac{a+b}{2a} , & \text{if } 1/b\leq r\leq 1/a \\ \frac{1+rb}{1+ra} , & \text{if } 1/a\leq r<\infty . \end{cases} \tag{2.26}$$

Moreover, $P(B_r)$ is minimised if we let r take values from the interval I_r where

$$I_r \equiv [1/b,1/a] \tag{2.27}$$

and its minimum value is given by

$$P(B_r) = \frac{a+b}{2a} . \tag{2.28}$$

Finally, if we let

$$\tau = \tau_O = \frac{2(1+ra)}{b+3a} , \tag{2.29}$$

then the spectral radius of $Q_{\tau,r}$ attains its minimum value which is given by the expression

$$S(Q_{\tau_O,r}) = \frac{b-a}{b+3a} . \tag{2.30}$$

Proof

From the relationships (2.18) and (2.19) we also find that

$$\min_{\mu,\nu}\{\lambda(\mu,\nu,r)\} = \begin{cases} \lambda(a,a,r), & \text{if } 0\leq r\leq 1/a \\ \lambda(b,a,r), & \text{if } 1/a\leq r<\infty. \end{cases} \tag{2.31}$$

By using (2.22),(2.23),(2.21) and (2.31) we determine the bound (2.26) for $P(B_r)$. In order to find the value of r such that $P(B_r)$ attains its minimum value we have to study the behaviour of the bound (2.26) as a function of r. From (2.26) and (2.22) by taking partial derivatives with respect to r we find that

$$\operatorname{sign}\left[\frac{\partial p}{\partial r}\right] = \begin{cases} \operatorname{sign}(a-b), & \text{if } 0\leq r\leq 1/b \\ 0, & \text{if } 1/b\leq r\leq 1/a \\ \operatorname{sign}(b-a), & \text{if } 1/a\leq r<\infty \end{cases}$$

which shows that the minimum value of $P(B_r)$ is given by (2.28) for all $r\in I_r$ where I_r is defined by (2.27). Moreover, we have from (2.24) that for the value of τ given by (2.29) the spectral radius of $Q_{\tau,r}$ (see (2.8)) is given by (2.30) since

$$S(Q_{\tau_0,r}) = \frac{P(B_r)-1}{P(B_r)+1} \tag{2.32}$$

and the proof of the theorem is complete.

A comparison of the effectiveness with respect to rates of convergence of the iterative scheme (I) when the involved parameters take the values which minimise $S(Q_{\tau,r})$ is provided by the following corollary.

Corollary 2.3

Under the hypotheses of Theorem 2.2 the iterative scheme (I) has asymptotically the same rate of convergence with the SD method (see (3-2.31)).

Proof

From (2.30) we have

$$S(Q_{\tau_0,r}) \leq \frac{P(A)-1}{P(A)+3} \tag{2.33}$$

where

$$P(A) = \frac{b}{a} \tag{2.34}$$

is the P-condition number of A. Evidently, by comparing $R(Q_{\tau_0,r})=-\log S(Q_{\tau_0,r})$ and $R(R_{\hat{\alpha}}) = \frac{P(A)-1}{P(A)+1}$ (see (3-2.32)) when $P(A)\gg 1$ we can clearly verify the validity of the corollary.

From Corollary 2.3 we conclude that the conditioning matrix given by (2.6) does not improve the P-condition number of the original system hence the iterative scheme (I) is only of academic interest. Another observation is that the conditioning matrix $R=I+rH$ is no better approximation to A than the conditioning matrix $R'=D^{(\pi_1)}$ (see Chapter 5). This can be regarded as an additional condition for the selection of the conditioning matrix (see Section 4.2).

From the above analysis we can conjecture that if two conditioning matrices are different but possess the same structure (e.g. they are tri-diagonal), then the associated iterative schemes which are produced using the preconditioning approach, will produce approximately the same rate of convergence for $h\to 0$.

7.3 THE MODIFIED ALTERNATING DIRECTION PRECONDITIONING METHOD (MADP METHOD)

Next, we continue our study of the same problem, as defined in the previous section, by considering the MADP iterative scheme (see (1.10)). That is, for the solution of our problem we consider the conditioning matrix to possess the form

$$R = (I+r_1H)(I+r_2V) \tag{3.1}$$

where r_1, r_2 are real preconditioning parameters.

Then, the MADP method is given by

$$u^{(n+1)} = u^{(n)} + \tau(I+r_2V)^{-1}(I+r_1H)^{-1}(b-Au^{(n)}) \tag{3.2}$$

where again the matrices H and V satisfy the condition (1.13). From (3.1) we have that

$$R = I+r_1H+r_2V+r_1r_2HV \tag{3.3}$$

which indicates that the effectiveness of R depends on the quantity r_1r_2HV. If we now assume that $r_1=r_2$, then we can immediately verify that the conditioning matrix $R'=I-\omega(L+U)+\omega^2LU$ is a better approximate to the matrix A than the conditioning matrix $R=I+r(H+V)+r^2HV$ (an easy way of verifying this is if we consider the molecules of R',R and compare which one is a better approximate to the molecule of the matrix A). We can therefore predict that the PSD method will have a slightly better rate of convergence than the ADP method. This result has been confirmed numerically (see Gane [1974]p.209) for the Laplace equation in the unit square.

Next, we see that the iterative scheme given by (3.2) is similar to the PSD method so we can either work with vector corrections and obtain a "computable" form similar to (4-9.2) or alternatively use a form similar to (A.10) (see Appendix A) which will allow us to save some computational effort. Following the latter suggestion we can write (3.2) in the form

$$\begin{aligned} &(I+r_1H)u^{(n+\frac{1}{2})} = [I+(r_1-\tau)H]u^{(n)}+\tau(b-Vu^{(n)}) \\ \text{and} \quad &(I+r_2V)u^{(n+1)} = u^{(n+\frac{1}{2})}+r_2Vu^{(n)} \ , \end{aligned} \tag{3.4}$$

where we observe that it is not necessary to recompute $Vu^{(n)}$ in the second

half iteration and therefore we can apply a scheme similar to (A.11) (see Appendix A) to considerably reduce the computational work.

An alternative form of the MADP method is given by

$$u^{(n+1)} = T_{\tau,r_1,r_2} u^{(n)} + t \tag{3.5}$$

where

$$T_{\tau,r_1,r_2} = I-\tau(I+r_2V)^{-1}(I+r_1H)^{-1}A \tag{3.6}$$

and

$$t = \tau(I+r_2V)^{-1}(I+r_1H)^{-1}b. \tag{3.7}$$

Moreover, we note that the MADP method is completely consistent (see Theorem 3-1.4) if $\tau\neq 0$ and $I+r_2V$, $I+r_1H$ are non-singular matrices. Evidently, the preconditioned matrix of the MADP scheme is given by the expression

$$B_{r_1,r_2} = (I+r_2V)^{-1}(I+r_1H)^{-1}A \tag{3.8}$$

and is positive definite for all $r_1,r_2\in(0,\infty)$.

Since now the matrices H and V satisfy the conditions (1.13) there exists a common basis of vectors for both matrices. From this observation and (3.8) we can easily find that the P-condition number of B_{r_1,r_2} is given by the expression

$$P(B_{r_1,r_2}) = \frac{\lambda_M}{\lambda_m} \tag{3.9}$$

where

$$\begin{aligned} \lambda_M &= \max_{\mu,\nu} \{\lambda(\mu,\nu,r_1,r_2)\}, \\ \lambda_m &= \min_{\mu,\nu} \{\lambda(\mu,\nu,r_1,r_2)\}, \end{aligned} \tag{3.10}$$

$$\lambda \equiv \lambda(\mu,\nu,r_1,r_2) = \frac{\mu+\nu}{(1+r_1\mu)(1+r_2\nu)} \tag{3.11}$$

and μ,ν are the eigenvalues of H,V, respectively which lie in the following ranges

$$0<a\leq\mu\leq b, \qquad 0<\alpha\leq\nu\leq\beta. \tag{3.12}$$

In order now to maximise the rate of convergence of the MADP method we will seek to select r_1 and r_2 to minimise $P(B_{r_1,r_2})$ given by (3.9) and on the other hand, to determine the optimum value for τ by the expression

$$\tau_0 = 2/(\lambda_m+\lambda_M) . \tag{3.13}$$

Finally, the MADP method converges for all $r_1,r_2\in(0,\infty)$ and $\tau\in(0,2/\lambda_M)$.

We distinguish two cases in our analysis i) the eigenvalue ranges of H and V are the same and ii) the eigenvalue ranges of H and V are different.

7.3.1 The case where the eigenvalue ranges of H and V are the same

In this case we prove the following theorem:

Theorem 3.1.1

Let H,V be the matrices as defined in Section 7.2 with real eigenvalues μ,ν, respectively such that

$$0<a\leq\mu,\nu\leq b. \tag{3.1.1}$$

Then, the P-condition number of B_{r_1,r_2} is given in Tables 3.1.1 and 3.1.3 for the different ranges of the preconditioning parameters $r_1,r_2\in(0,\infty)$. Moreover, $P(B_{r_1,r_2})$ is minimised if we let

$$r_1 = r_2 = r' = 1/(ab)^{\frac{1}{2}} \tag{3.1.2}$$

and its corresponding value is given by the expression

$$P(B_{r',r'}) = (a+b)r'/2. \tag{3.1.3}$$

On the other hand, of we also let

$$\tau = \tau_0 = 2r' \tag{3.1.4}$$

then, the spectral radius $S(T_{\tau,r_1,r_2})$ attains its minimum value which is given by the expression

$$S(T_{\tau_0,r',r'}) = \left(\frac{b^{\frac{1}{2}}-a^{\frac{1}{2}}}{b^{\frac{1}{2}}+a^{\frac{1}{2}}}\right)^2 . \tag{3.1.5}$$

Proof

As it can be seen from (3.10) we have to examine the behaviour of λ given by (3.11) as a function of μ,ν. Therefore by taking partial derivatives of the continuous function λ with respect to μ and ν we can easily obtain the following results

$$\begin{aligned} \text{sign}\left(\frac{\partial\lambda}{\partial\mu}\right) &= \text{sign}(1-\nu r_1) \\ \text{sign}\left(\frac{\partial\lambda}{\partial\nu}\right) &= \text{sign}(1-\mu r_2). \end{aligned} \tag{3.1.6}$$

From (3.1.6) we see that for fixed $r_1, r_2 > 0$ neither of the expressions $\frac{\partial \lambda}{\partial \mu}$, $\frac{\partial \lambda}{\partial \nu}$ changes sign as μ and ν vary in the interval (3.1.1). We therefore conclude that the possible extreme values of λ will occur at the points (a,a), (a,b), (b,a), and (b,b) (see Guittet [1967] Lemma 1). On the other hand, the values of the function λ at these points are the following

$$\begin{aligned} A &= \lambda(a,a,r_1,r_2), \qquad B = \lambda(a,b,r_1,r_2) \\ D &= \lambda(b,a,r_1,r_2), \qquad C = \lambda(b,b,r_1,r_2). \end{aligned} \tag{3.1.7}$$

Evidently, from (3.1.7) and for fixed r_1, r_2 we have that

$$\lambda_m = \min\{A,B,C,D\} \text{ and } \lambda_M = \max\{A,B,C,D\} \tag{3.1.8}$$

which indicates that we have to examine the relations of the quantities given by (3.1.7).

But we can easily obtain the following results

$$\left.\begin{aligned} \operatorname{sign}(A-B) &= \operatorname{sign}(r_2 - 1/a) \\ \operatorname{sign}(B-C) &= \operatorname{sign}(r_1 - 1/b) \\ \operatorname{sign}(A-D) &= \operatorname{sign}(r_1 - 1/a) \\ \text{and} \quad \operatorname{sign}(D-C) &= \operatorname{sign}(r_2 - 1/b). \end{aligned}\right\} \tag{3.1.9}$$

The above results suggest that for finding the order of the quantities A,B,C and D which will allow us to determine λ_m and λ_M from (3.1.8), we have to examine the relative positions of r_1 and r_2 with respect to the values $\frac{1}{a}$ and $\frac{1}{b}$. We therefore have to distinguish nine cases which are presented in Table 3.1.1 together with the values of λ_m, λ_M and $P(B_{r_1,r_2})$ for each case.

From Table 3.1.1 we see that we have determined $P(B_{r_1,r_2})$ thus we can now study its behaviour as a function of r_1 and r_2. By assuming that r_1 is kept fixed, then we obtain the results summarised in Table 3.1.2 for i=1, whereas if r_2 is fixed, we have the same results where now i=2.

r_1-Domain	r_2-Domain	λ_M	λ_m	$P(B_{r_1,r_2})$
$0<r_1\leq 1/b$	$0<r_2\leq 1/b$	C	A	C/A
	$1/b\leq r_2\leq 1/a$	D	A	D/A
	$1/a\leq r_2<\infty$	D	B	D/B
$1/b\leq r_1\leq 1/a$	$0<r_2\leq 1/b$	B	A	B/A
	$1/b\leq r_2\leq 1/a$	max{B,D}	min{A,C}	max{B,D}/min{A,C}
	$1/a\leq r_2<\infty$	D	C	D/C
$1/a\leq r_1<\infty$	$0<r_2\leq 1/b$	B	D	B/D
	$1/b\leq r_2\leq 1/a$	B	C	B/C
	$1/a\leq r_2<\infty$	A	C	A/C

TABLE 3.1.1

THE P-CONDITION NUMBER OF B_{r_1,r_2}

r_i-Domain†	r_j-Domain††	$\text{sign}(\partial P(B_{r_1,r_2})/\partial r_i)$	$P(B_{r_1,r_2})$
$0<r_i\leq 1/b$	$0<r_j\leq 1/b$	sign(a-b)	Decreasing
	$1/b\leq r_j\leq 1/a$	0	Stationary
	$1/a\leq r_j<\infty$	sign(b-a)	Increasing
$1/b\leq r_i\leq 1/a$	$0<r_j\leq 1/b$	sign(a-b)	Decreasing
	$1/b\leq r_j\leq 1/a$	-	-
	$1/a\leq r_j<\infty$	sign(b-a)	Increasing
$1/a\leq r_i<\infty$	$0<r_j\leq 1/b$	sign(a-b)	Decreasing
	$1/b\leq r_j\leq 1/a$	0	Stationary
	$1/a\leq r_j<\infty$	sign(b-a)	Increasing

TABLE 3.1.2

BEHAVIOUR OF $P(B_{r_1,r_2})$ AS A FUNCTION OF r_1,r_2

† $i=1,2$

†† $j=\{1,2\}-\{i\}$

From Table 3.1.2 we conclude that $P(B_{r_1,r_2})$ attains its minimum value when r_1 and r_2 lie in the following range

$$1/b \leq r_1, r_2 \leq 1/a. \tag{3.1.10}$$

But for this range of the preconditioning parameters, $P(B_{r_1,r_2})$ is given by the expression (see Table 3.1.1)

$$P(B_{r_1,r_2}) = \frac{\max\{B,D\}}{\min\{A,C\}}. \tag{3.1.11}$$

The order of B,D and A,C when r_1, r_2 lie in the range (3.1.10) is determined from the relationships

$$\text{sign}(B-D) = \text{sign}(r_1 - r_2) \tag{3.1.12}$$

and

$$\text{sign}(A-C) = \text{sign}(r_1 r_2 ab - 1). \tag{3.1.13}$$

Consequently, we have the following expressions for λ_M and λ_m

$$\lambda_M = \begin{cases} D, & \text{if } r_1 \leq r_2 \\ B, & \text{if } r_1 \geq r_2 \end{cases} \tag{3.1.14}$$

and

$$\lambda_m = \begin{cases} A, & \text{if } \frac{1}{b} \leq r_2 \leq \frac{1}{r_1 ab} \\ C, & \text{if } \frac{1}{r_1 ab} \leq r_2 \leq \frac{1}{a}. \end{cases} \tag{3.1.15}$$

Evidently, the quantity $\frac{1}{r_1 ab}$ belongs to the interval $[\frac{1}{b}, \frac{1}{a}]$ for all $r_1 \in [\frac{1}{b}, \frac{1}{a}]$. Moreover, we note that

$$\frac{1}{b} \leq r_1 \leq \frac{1}{r_1 ab} \leq \frac{1}{a}, \quad \text{if } r_1 \leq \frac{1}{\sqrt{ab}}$$

and

$$\frac{1}{b} \leq \frac{1}{r_1 ab} \leq r_1 \leq \frac{1}{a}, \quad \text{if } r_1 \geq \frac{1}{\sqrt{ab}}. \tag{3.1.16}$$

From the inequalities (3.1.16) it follows that we have to consider six cases which emerge for the different values of $r_2 \in [1/b, 1/a]$ (see Table 3.1.3) by keeping r_1 fixed. The results which are obtained after the examination of these cases are summarised in Table 3.1.3.

r_1-Domain	r_2-Domain	λ_M	λ_m	$P(B_{r_1,r_2})$	$\text{sign}(\partial P(B_{r_1,r_2})/\partial r_2)$
$1/b \leq r_1 \leq 1/(ab)^{\frac{1}{2}}$	$1/b \leq r_2 \leq r_1$	B	A	B/A	sign(a-b)
	$r_1 \leq r_2 \leq 1/(r_1 ab)$	D	A	D/A	0
	$1/(r_1 ab) \leq r_2 \leq 1/a$	D	C	D/C	sign(b-a)
$1/(ab)^{\frac{1}{2}} \leq r_1 \leq 1/a$	$1/b \leq r_2 \leq 1/(r_1 ab)$	B	A	B/A	sign(a-b)
	$1/(r_1 ab) \leq r_2 \leq r_1$	B	C	B/C	0
	$r_1 \leq r_2 \leq 1/a$	D	C	D/C	sign(b-a)

TABLE 3.1.3

BEHAVIOUR OF $P(B_{r_1,r_2})$ AS A FUNCTION OF r_2

In Table 3.1.3 we present the expressions by which $P(B_{r_1,r_2})$ is represented for the different values of $r_1, r_2 \in [1/b, 1/a]$ as well as the behaviour of $P(B_{r_1,r_2})$ as a function of r_2. Our main conclusion from Table 3.1.3 is that for fixed r_1, $P(B_{r_1,r_2})$ attains its minimum value for r_2 such that

$$r_2 \in [\min(r_1, \frac{1}{r_1 ab}), \quad \max(r_1, \frac{1}{r_1 ab})]. \tag{3.1.17}$$

On the other hand, by keeping r_2 fixed in the above interval we find that

i) if $r_1 \in [\frac{1}{b}, \frac{1}{\sqrt{ab}}]$, then $\text{sign}\left(\frac{\partial P(B_{r_1,r_2})}{\partial r_1}\right) = \text{sign}(a-b)$

ii) if $r_1 \in [\frac{1}{\sqrt{ab}}, \frac{1}{a}]$, then $\text{sign}\left(\frac{\partial P(B_{r_1,r_2})}{\partial r_1}\right) = \text{sign}(b-a)$

which indicate that $P(B_{r_1,r_2})$ is minimised when the preconditioning parameter r_1 becomes equal to the quantity

$$r_1 = r' = \frac{1}{\sqrt{ab}} \tag{3.1.18}$$

hence by (3.1.17) we have

$$r_2 = r_1 \tag{3.1.19}$$

and from (3.1.18),(3.1.19) we see that (3.1.2) holds.

In addition, from (3.1.14),(3.1.15),(3.1.7) and (3.11) we obtain the following expressions for the smallest and largest eigenvalue of $B_{r',r'}$

$$\lambda_M = B = D = \frac{(a+b)\sqrt{ab}}{(\sqrt{a}+\sqrt{b})^2}$$

and (3.1.20)

$$\lambda_m = A = C = \frac{2ab}{(\sqrt{a}+\sqrt{b})^2} .$$

By combining (3.9) and (3.1.20) we easily obtain the minimised value of the P-condition number of $B_{r',r'}$ which is given by (3.1.3). Finally, from (3.13) and (3.1.20) we obtain the optimum value of τ (see (3.1.4)). But for this optimum value τ_0 of τ the spectral radius of the iteration matrix is given by the formula

$$S(T_{\tau_0,r',r'}) = \frac{P(B_{r',r'})-1}{P(B_{r',r'})+1} \qquad (3.1.21)$$

which if combined with (3.1.3) and (3.1.2) gives (3.1.5) and the proof of the theorem is complete.

7.3.2 The case where the eigenvalue ranges of H and V may be different

In this case we prove the following theorem:

Theorem 3.2.1

Let H,V be the matrices as defined in Section 7.2 with eigenvalues μ,ν, respectively such that

$$0<a\leq\mu\leq b \quad \text{and} \quad 0<\alpha\leq\nu\leq\beta. \qquad (3.2.1)$$

Then, the P-condition number of B_{r_1,r_2} is minimised when the parameters r_1,r_2 take the values

$$r_1^* = \frac{1-\Sigma sc^{\frac{1}{2}}}{-t+\Sigma qc^{\frac{1}{2}}}, \quad r_2^* = \frac{1+\Sigma sc^{\frac{1}{2}}}{t+\Sigma qc^{\frac{1}{2}}}, \qquad (3.2.2)$$

where

$$c = \frac{1}{1+\theta+[\theta(2+\theta)]^{\frac{1}{2}}}, \qquad (3.2.3)$$

$$\theta = \frac{2(\beta-\alpha)(b-a)}{(a+\alpha)(b+\beta)}, \qquad (3.2.4)$$

$$\Sigma s = \frac{(\beta-\alpha)-(b-a)}{(b+\beta)-(a+\alpha)c} , \tag{3.2.5}$$

$$\Sigma q = \frac{(b+\beta)+(b-\beta)\Sigma s}{2} , \tag{3.2.6}$$

$$t = \frac{(b-\beta)+(b+\beta)\Sigma s}{2} , \tag{3.2.7}$$

and its corresponding value is given by

$$P(B_{r_1^*,r_2^*}) = (c^{\frac{1}{2}} + c^{-\frac{1}{2}})/2. \tag{3.2.8}$$

On the other hand, if we also let

$$\tau = \tau_0^* = r_1^* + r_2^* , \tag{3.2.9}$$

then the spectral radius of T_{τ,r_1,r_2} attains its minimum value which is given by the expression

$$S(T_{\tau_0^*,r_1^*,r_2^*}) = \left(\frac{1-c^{\frac{1}{2}}}{1+c^{\frac{1}{2}}}\right)^2 . \tag{3.2.10}$$

Proof

Under the hypotheses of the theorem we have that μ,ν lie in the different ranges given by (3.2.1). Since we have solved the problem for the case where μ,ν lie in the same range (see Section 7.3.1), we attempt to find a technique of transforming our present problem so that we return to the previous case of the "single range". This will prevent us repeating the laborious procedure (see proof of Theorem 3.1.1) of the more complex problem in the present case. The technique for achieving this is quite well known and is due to Wachspress and Jordan (see Wachspress [1966], Young [1971]).

We commence our analysis by noting that the function λ defined by (3.11) can be written alternatively as

$$\lambda = \frac{\omega_1\omega_2}{\omega_1+\omega_2}\left[1 - \frac{(\mu-\omega_2)(\nu-\omega_1)}{(\mu+\omega_1)(\nu+\omega_2)}\right] \tag{3.2.11}$$

where

$$\omega_1 = \frac{1}{r_1} \quad \text{and} \quad \omega_2 = \frac{1}{r_2} . \tag{3.2.12}$$

We see that by expressing λ in the above form our interest is focused on the second term in the brackets. By adhering to the analysis of Wachspress and Jordan we seek to introduce new variables $\hat{\mu}$ and $\hat{\nu}$ such that

$$\mu = \frac{t+q\hat{\mu}}{1+s\hat{\mu}} , \qquad \nu = \frac{t'+q'\hat{\nu}}{1+s'\hat{\nu}} \tag{3.2.13}$$

so that for some $\hat{\omega}_1$ and $\hat{\omega}_2$ we have

$$\left(\frac{\mu-\omega_2}{\mu+\omega_1}\right)\left(\frac{\nu-\omega_1}{\nu+\omega_2}\right) = \left(\frac{\hat{\mu}-\hat{\omega}_2}{\hat{\mu}+\hat{\omega}_1}\right)\left(\frac{\hat{\nu}-\hat{\omega}_1}{\hat{\nu}+\hat{\omega}_2}\right) \tag{3.2.14}$$

and where $\hat{\mu}$ and $\hat{\nu}$ vary over the ranges

$$\sigma \lesssim \hat{\mu} \lesssim \Sigma , \quad \sigma \lesssim \hat{\nu} \lesssim \Sigma. \tag{3.2.15}$$

It can be shown that (for details see Young [1971] pp.511) by using certain conditions such that for (3.2.14) to hold, the relationships given by (3.2.13) become

$$\mu = \frac{t+q\hat{\mu}}{1+s\hat{\mu}} , \quad \nu = \frac{-t+q\hat{\nu}}{1-s\hat{\nu}} . \tag{3.2.16}$$

In order to determine t,q,s,σ and Σ we require that $\mu=a$ corresponds to $\hat{\mu}=\sigma$, $\mu=b$ corresponds to $\hat{\mu}=\Sigma$, $\nu=\alpha$ corresponds to $\hat{\nu}=\sigma$ and that $\hat{\nu}=\beta$ corresponds to $\hat{\nu}=\Sigma$, hence we have

$$a = \frac{t+q\sigma}{1+s\sigma} , \qquad b = \frac{t+q\Sigma}{1+s\Sigma} ,$$
$$\alpha = \frac{-t+q\sigma}{1-s\sigma}, \qquad \beta = \frac{-t+q\Sigma}{1-s\Sigma}. \tag{3.2.17}$$

After some algebraic manipulation we obtain the relationship

$$c+1/c = 2(1+\theta) \tag{3.2.18}$$

where

$$c = \sigma/\Sigma \tag{3.2.19}$$

and θ is given by (3.2.4). The quantities $\Sigma s, \Sigma q$ and t are also determined and are given by (3.2.5), (3.2.6) and (3.2.7), respectively. Therefore, we rewrite (3.2.16) to yield

$$\mu = \frac{t+(\Sigma q)(\hat{\mu}/\Sigma)}{1+(\Sigma s)(\hat{\mu}/\Sigma)} \qquad \text{and} \quad \nu = \frac{-t+(\Sigma q)(\hat{\nu}/\Sigma)}{1-(\Sigma s)(\hat{\nu}/\Sigma)} , \tag{3.2.20}$$

whereas by combining (3.2.13) and (3.2.14) we find

$$\omega_1 = \frac{-t+(\Sigma q)(\hat{\omega}_1/\Sigma)}{1-(\Sigma s)(\hat{\omega}_1/\Sigma)} \quad \text{and} \quad \omega_2 = \frac{t+(\Sigma q)(\hat{\omega}_2/\Sigma)}{1+(\Sigma s)(\hat{\omega}_2/\Sigma)} . \tag{3.2.21}$$

At this stage we note that we have transformed our problem to be identical with the one discussed in the previous section, where now instead of μ,ν we have the transformed variables $\hat{\mu},\hat{\nu}$, respectively possessing the same range given by (3.2.15). Evidently, from Theorem 3.1.1 and the relationships (3.2.12), (3.2.15) we see that the optimum parameters $\hat{\omega}_1$ and $\hat{\omega}_2$ for the transformed problem are

$$\hat{\omega}_1 = \hat{\omega}_2 = (\sigma\Sigma)^{\frac{1}{2}}. \tag{3.2.22}$$

Thus from (3.2.21),(3.2.22) and (3.2.19) we find that the optimum parameters for the given problem are given by the following expressions

$$\omega_1^* = \frac{-t+\Sigma qc^{\frac{1}{2}}}{1-\Sigma sc^{\frac{1}{2}}}, \qquad \omega_2^* = \frac{t+\Sigma qc^{\frac{1}{2}}}{1+\Sigma sc^{\frac{1}{2}}}. \tag{3.2.23}$$

Finally, from (3.2.12) and (3.2.23) we see that the optimum values for the preconditioning parameters are given by the expressions (3.2.2).

It is a trivial matter now from the above analysis and using the relationships (3.1.3),(3.1.4),(3.1.5) to show the validity of (3.2.8), (3.2.9) and (3.2.10), respectively. Thus, the proof of the theorem is complete.

Evidently, we can choose any positive value for R, e.g. R=1.

As we have shown (see Theorems 3.1.1 and 3.2.1) in the case where A is given by

$$A = H+V \tag{3.2.24}$$

then at the optimum stages the MADP method coincides with the Peaceman-Rachford ADI method (see Birkhoff et al [1962], Young [1971], Wachspress [1963]) when all the iteration parameters are kept fixed during the iterations. However, the advantage of the MADP method over the PR-ADI comes when more accurate finite difference equations are used (e.g. the nine-point difference formula).

It should be mentioned that similar results to Section 7.3.1 for the

case $r_1=r_2$ and for the EADI method have been obtained by Guittet [1967] whereas for the same case but with the eigenvalues of H and V lying in different ranges the optimum parameters have been found by Hadjidimos and Iordanidis [1974]. Moreover, for the ADP method (i.e. when $r_1=r_2$) Gane [1974] (see also Gane and Evans [1974]) found similar results to the ones given in Section 7.3.1. However, for the case where the eigenvalue ranges of the basic matrices involved are different, the optimum parameters were found (see Gane and Evans [1974]) under the assumption that $0<\alpha'\leq\mu,\nu\leq\beta'$, where $\alpha'=\min(a,\alpha)$, and $\beta'=\max(b,\beta)$.

7.4 APPLICATION OF THE ACCELERATED PROCEDURES TO THE MADP METHOD

In this section, we will briefly consider the application of the accelerated procedures developed in Chapter 3 to the Modified Alternating Direction method. In general, we will assume that $r_1 \neq r_2$ and that the matrix A has the form (3.2.24) with H and V defined as in Section 7.2. The properties of the basic matrices H,V guarantee the effectiveness of the acceleration techniques for the improvement (by an order of magnitude) in the rate of convergence of the MADP method since its iteration matrix (see (3.6)), will always have real eigenvalues. Furthermore, because of the similarity of the MADP and PSD iterative schemes we will follow closely the formulation of the corresponding accelerated iterative schemes as developed in Sections 5.5, 5.6, 5.7 and 5.8. Thus, we define the semi-iterative method based on MADP (denoted by MADP-SI) by (see (5-5.15))

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n+1)} + \rho_{n+1}(T_{\tau,r_1,r_2}u^{(n)}+t)^{\dagger} \tag{4.1}$$

where the second term in the brackets is the MADP method (see (3.5),(3.6) and (3.7)). The sequence of parameters is given by

$$\left.\begin{aligned} \rho_1 &= 1 \ , \\ \rho_2 &= \left(1-\frac{\sigma^2}{2}\right)^{-1} , \\ \rho_{n+1} &= \left(1-\frac{\sigma^2\rho_n}{4}\right)^{-1} , \quad n=2,3,\ldots \end{aligned}\right\} \tag{4.2}$$

where

$$\sigma = S(T_{\tau,r_1,r_2}) = \frac{P(B_{r_1,r_2})-1}{P(B_{r_1,r_2})+1} \ . \tag{4.3}$$

By expanding (4.1) we obtain a more explicit form which is given by

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n-1)}+\rho_{n+1}[u^{(n)}+\tau(I+r_2V)^{-1}(I+r_1H)^{-1}(b-Au^{(n)})] . \tag{4.4}$$

The virtual spectral radius of the iterative scheme (4.4) is given by (see (see 5-5.19)

---

†*It is assumed that the parameters* τ, r_1, r_2 *take their optimum values.*

$$\bar{S}(P_n(T_{\tau,r_1,r_2})) = \frac{2r^{n/2}}{1+r^n} \tag{4.5}$$

where

$$r^{\frac{1}{2}} = \frac{\sigma}{1+\sqrt{1-\sigma^2}} = \frac{1-1/\sqrt{P(B_{r_1,r_2})}}{1+1/\sqrt{P(B_{r_1,r_2})}} \quad . \tag{4.6}$$

Therefore, the rate of convergence is (see (5-5.24))

$$R_\infty(P_n(T_{\tau,r_1,r_2})) = -\frac{1}{2}\log r \sim 2/\sqrt{P(B_{r_1,r_2})} \quad . \tag{4.7}$$

In a similar manner we define the MADP-Variable Extrapolation method (MADP-VE method) by (see (5-6.4))

$$u^{(n+1)} = u^{(n)} + \theta_{n+1}(I+r_2V)^{-1}(I+r_1H)^{-1}(b-Au^{(n)}) \tag{4.8}$$

where the iteration parameters θ_n can be determined by the expression (see (5-6.5))

$$\theta_k = \frac{\tau_0}{1-\sigma\cos\frac{(2k-1)\pi}{2m}} \quad , \quad k=1,2,\ldots,m \tag{4.9}$$

and σ given by (4.3).

An alternative form of the MADP-VE method is given by the following two level iterative scheme (see (3.4))

$$\begin{aligned} (I+r_1H)u^{(n+\frac{1}{2})} &= [I+(r_1-\theta_{n+1})H]u^{(n)}+\theta_{n+1}(b-Vu^{(n)}) \\ \text{and} \qquad (I+r_2V)u^{(n+1)} &= u^{(n+\frac{1}{2})}+r_2Vu^{(n)}. \end{aligned} \tag{4.10}$$

The main advantage of the above scheme is that it is possible to apply the reduction scheme (A.11) thus saving some computational effort.

The second degree version of the ADP method has been introduced by Gane and Evans [1974]. Here, we define in an analogous manner the MADP-second degree method to be given by (see (5-7.6))

$$u^{(n+1)} = \hat{\omega}_0[u^{(n)}+\tau(I+r_2V)^{-1}(I+r_1H)^{-1}(b-Au^{(n)})]+(1-\hat{\omega}_0)u^{(n-1)} \tag{4.11}$$

where

$$\hat{\omega}_0 = \begin{cases} 1 & , \text{ if } n=0 \\ \dfrac{2}{1+\sqrt{1-\sigma^2}} & , \text{ if } n\geq 1 \end{cases} \tag{4.12}$$

and σ is given by (4.3).

As we have seen (see Chapter 5) the MADP-VE and the MADP-second degree methods are strong alternatives to the MADP-SI since their rate of convergence tends to be approximately the same as the latter method.

Finally, we can also define the Conjugate Gradient method with respect to the MADP scheme by (see (5-9.8))

$$u^{(n+1)} = \rho_{n+1}[u^{(n)}+\gamma_{n+1}(I+r_2V)^{-1}(I+r_1H)^{-1}r^{(n)}]+(1-\rho_{n+1})u^{(n-1)} \tag{4.13}$$

where

$$\left.\begin{aligned} \gamma_{n+1} &= \frac{(r^{(n)},\tilde{s}^{(n)})}{(\tilde{s}^{(n)},A\tilde{s}^{(n)})}, \\ \tilde{s}^{(n)} &= (I+r_2V)^{-1}(I+r_1H)^{-1}r^{(n)}, \\ \rho_{n+1} &= \left[1-\frac{\gamma_{n+1}}{\gamma_n}\frac{(r^{(n)},r^{(n)})}{(r^{(n-1)},r^{(n-1)})}\frac{1}{\rho_n}\right]^{-1} \end{aligned}\right\} \tag{4.14}$$

and

$$r^{(n)} = b-Au^{(n)}.$$

7.5 THE MODEL PROBLEM - COMPARISON OF RATES OF CONVERGENCE

Let us now consider the generalised Dirichlet problem involving the Laplace's equation (see Section 4.13) and try to compare the effectiveness of the methods considered in this chapter with already known ones. Evidently, we have $E_0=0$ in (1.6), hence

$$A = H_0+V_0 = H+V \ , \tag{5.1}$$

$$H = I \otimes U \quad \text{and} \quad V = U \otimes J \tag{5.2}$$

where I is the unit matrix of order N-1 and U is of the same order and given by

$$U = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} . \tag{5.3}$$

The symbol $\otimes$ denotes the tensor product (see Halmos [1958]) and has already been used in connection with ADI methods (see Keast and Mitchell [1967]). In the above we have assumed a uniform mesh size h=1/N. For this problem we see that the conditions of the "commutative case" hold. Furthermore, if μ,ν are the eigenvalues of H,V, respectively, then

$$\mu_i = 4\sin^2(\tfrac{i\pi}{2N}) \quad \text{and} \quad \nu_j = 4\sin^2(\tfrac{j\pi}{2N}), \ 1\leq i,j\leq N-1. \tag{5.4}$$

Hence, if $a\leq\mu_i\leq b$ and $\alpha\leq\nu_j\leq\beta$ with $1\leq i,j\leq N-1$, then

$$a = \alpha = 4\sin^2(\tfrac{\pi h}{2}) \quad \text{and} \quad b = \beta = 4\cos^2(\tfrac{\pi h}{2}). \tag{5.5}$$

From (5.5) and Theorem 3.1.1 we have that for our problem the value of r' is given by

$$r' = (ab)^{-\frac{1}{2}} = \frac{1}{4\sin(\frac{\pi h}{2})\cos(\frac{\pi h}{2})} \tag{5.6}$$

whereas the corresponding value of $B_{r',r'}$ becomes

$$P(B_{r',r'}) = 2r' = \frac{1}{\sin(\pi h)} \ . \tag{5.7}$$

On the other hand, by letting $\tau_0=2r'$ we find that the spectral radius of T_{τ,r_1,r_2} is given by the expression

$$S(T_{\tau_0,r',r'}) = \frac{1-\sin(\pi h)}{1+\sin(\pi h)} \tag{5.8}$$

which is identical with the spectral radius of the SOR method (see (4-13.14)). Since for this case we also have that the PR method is identical with the ADP (see discussion after the proof of Theorem 3.2.1) we obtain the following result

$$S(T_{\tau_0,r',r'}) = S(L_{\omega_b}). \tag{5.9}$$

Consequently, for the model problem and when H,V are defined as in Section 7.2 PR-ADI, SOR and MADP have identical rates of convergence at the optimum stage. It should be noted here that the above result has also been obtained by Gane [1974] for the ADP method. In addition, let us also examine the case where more accurate finite difference analogues are used to approximate the Laplace's equation. In particular, we consieer the nine point difference formula, then the totality of the difference equations produced yields the following splitting of the coefficient matrix

$$A = H+V-kHV \tag{5.10}$$

where k=1/6 and again H,V are the same matrices as defined in Section 7.2. By assuming the same conditioning matrix, we have that

$$P_k(B_{r_1,r_2}) = \frac{\lambda_M}{\lambda_m} \tag{5.11}$$

where λ_M,λ_m are the maximum and minimum bounds of

$$\lambda = \frac{\mu+\nu-k\mu\nu}{(1+r_1\mu)(1+r_2\nu)} \tag{5.12}$$

respectively. Since now the inequalities

$$(b-\mu)\nu + (b-\nu)\mu \gtrless 0$$

$$(\mu-a)\nu + (\nu-a)\mu \gtrless 0$$

always hold, we can easily obtain the following result

$$a(\mu+\nu)/2 \lesseqgtr \mu\nu \lesseqgtr b(\mu+\nu)/2. \tag{5.13}$$

By combining (5.13) and (5.12) we can bound λ as follows

$$(1-kb/2)\phi \lesssim \lambda \lesssim (1-ka/2)\phi \tag{5.14}$$

where

$$\phi = \frac{\mu+\nu}{(1+r_1\mu)(1+r_2\nu)} . \tag{5.15}$$

From (5.14) and (5.15) we see that the P-condition number of B_{r_1,r_2} is minimised for these values of r_1,r_2 for which the ratio ϕ_M/ϕ_m is minimised, where ϕ_M and ϕ_m are the largest and smallest bound of ϕ, respectively. But this problem is identical with the one studied in Section 7.3. Thus using the results of Theorem 3.1.1 and (5.14) we have that if we let

$$r_1 = r_2 = r' = \frac{1}{\sqrt{ab}} \tag{5.16}$$

then $P_k(B_{r_1,r_2})$ is minimised and its corresponding value is given by

$$P_k(B_{r',r'}) = k'P(B_{r',r'}) \tag{5.17}$$

where

$$k' = \frac{1-ka/2}{1-kb/2} . \tag{5.18}$$

Moreover, for

$$\tau_0 = \frac{(1+r'a)^2}{a(1-kb/2)(1+k'P(B_{r',r}))} \tag{5.19}$$

the spectral radius is also minimised and given by the expression

$$S_k(T_{\tau_0,r',r'}) = \frac{k'P(B_{r',r'})-1}{k'P(B_{r',r'})+1} . \tag{5.20}$$

Finally, the rate of convergence is given by

$$R_k(T_{\tau_0,r',r'}) \sim \frac{2}{k'P(B_{r',r'})} \tag{5.21}$$

where we can clearly see the effect of k.

As can be seen in this case $\tau_0 \neq 2r'$ which means that the MADP method does not coincide with the PR-ADI process. On the other hand, since for the PR-ADI the parameter τ does not take its optimum value (which is given by (5.19)) it follows that in this case MADP will have a slightly faster rate of convergence than the former method.

Let us now consider the application of the MADP-SI method for the solution of the present problem, then the rate of convergence is given by

$$R_{k,\infty}(P_n(T_{\tau_0},r',r')) \sim 2/\sqrt{k'P(B_{r',r'})}$$

thus for k=0 we obtain

$$R_{\infty}(P_n(T_{\tau_0},r',r')) \sim 2\sqrt{\sin(\pi h)} \sim 2\sqrt{\pi h} \qquad (5.22)$$

for sufficiently small h. The above expression for the rate of convergence serves also as an approximate bound for the rate of convergence of the other accelerated techniques applied to the MADP method (MADP-SD, MADP-VE and MADP-CG).

7.6 NUMERICAL RESULTS

For comparison reasons we consider again the Laplace equation

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0 \tag{6.1}$$

in the unit square with zero boundary values. The starting vector $u^{(0)}$ with all its components equal to unity is used, whereas the convergence criterion is $\|u^{(n)}\|_\infty \leq 10^{-6}$. For the solution of the above problem we approximated (6.1) by using the five and the nine point difference analogues (k=0 and k=1/6). The so produced system was solved by applying the MADP method and also its accelerated versions MADP-VE, MADP-SI and MADP-CG as they have been developed previously. For the case k=0, the optimum value of r' was computed from (5.6) and $\tau_0=2r'$ whereas for k=1/6 τ_0 was determined by (5.19). In Tables 6.1 and 6.2 we present the number of iterations required to solve the present problem with the iterative procedures mentioned above for the different mesh sizes $h^{-1}=20,30,40,60,80$. Under the column headings n_E we give the estimated number of iterations whereas under n_0 we have the observed number of iterations. The quantities $P(B_{r',r'}),S(T_{\tau_0,r',r'}),P_k(B_{r',r'})$ and $S_k(T_{\tau_0,r'r'})$ were computed from (5.7), (5.8),(5.17) and (5.20), respectively whereas $P(B_{r=0})=P(A)=b/a$ and $P_k(B_{r=0})=k'p(A)$ where k' is given by (5.18). The selection of m in the MADP-VE method is similar to the one developed in Section 5.10 (see (10.29)). In this example we cannot see the advantage of using more accurate difference approximations because the theoretical solution of Au=0 with A given by (5.10) is the zero solution for both k=0 and k=1/6 which is the same with the theoretical solution of (6.1).

Figure 6.1 shows graphs with logarithmic scales, of the observed number of iterations versus h^{-1} for the MADP, MADP-SI and for the MADP-CG methods for k=0 and k=1/6.

From Tables 6.1 and 6.2 we see that the number of iterations n_E for the considered methods agree closely with the observed values n_0. As a

h^{-1}	r'	τ_0	$P(B_{r=0})$	$P(B_{r',r'})$	$S(T_{\tau_0,r',r'})$	MADP		MADP-VE			MADP-SI		MADP-CG
						n_E	n_0	n_E	n_0	m	n_E	n_0	n_0
20	3.1962	6.3925	81.2238	6.3925	0.7295	44	46	20	24	5	18	19	15
30	4.7834	9.5668	182.5449	9.5668	0.8107	66	69	30	29	6	22	23	19
40	6.3727	12.7455	324.3945	12.7455	0.8545	88	91	35	34	7	26	27	22
60	9.5537	19.1073	729.6792	19.1073	0.9005	132	137	40	40	8	32	34	26
80	12.7357	25.4713	1297.0779	25.4713	0.9244	176	183	45	45	9	37	39	31

TABLE 6.1

NUMERICAL RESULTS FOR THE MADP METHOD WHEN k=0

h^{-1}	r'	τ_0	$P_k(B_{r=0})$	$P_k(B_{r',r'})$	$S_k(T_{\tau_0,r',r'})$	MADP		MADP-VE			MADP-SI		MADP-CG
						n_E	n_0	n_E	n_0	m	n_E	n_0	n_0
20	3.1962	6.7048	121.2127	9.5396	0.8102	66	50	30	29	6	22	23	17
30	4.7834	9.8859	273.1932	14.3174	0.8694	99	75	35	34	7	27	28	21
40	6.3727	13.0681	485.9673	19.0937	0.9005	132	100	40	40	8	32	32	24
60	9.5537	19.4335	1093.8941	28.6446	0.9325	198	150	50	50	10	39	40	30
80	12.7357	25.7993	1944.9919	38.1947	0.9490	264	200	55	55	11	45	46	34

TABLE 6.2

NUMERICAL RESULTS FOR THE MADP METHOD WHEN k=1/6

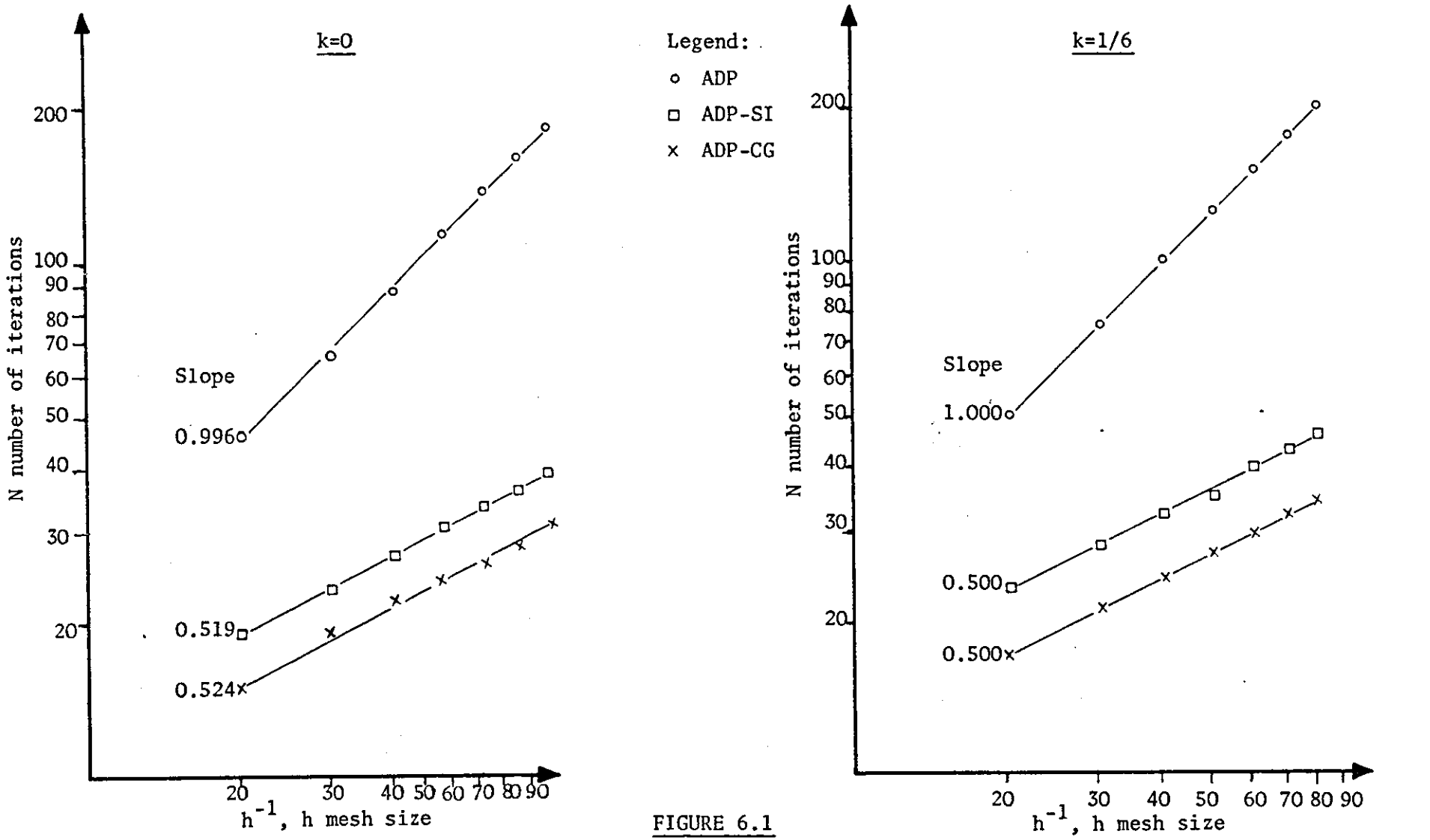


FIGURE 6.1

DETERMINATION OF RATE OF CONVERGENCE ATTAINED FOR THE MODEL PROBLEM USING ADP, ADP-SI AND ADP-CG WITH k=0 AND k=1/6

consequence of the agreement between the number of iterations predicted by the theory and the actual number, it follows that the MADP method is extremely effective. Furthermore, from Figure 6.1 we observe that the number of iterations of the MADP method varies approximately like $O(h^{-1})$, whereas for the other accelerated procedures like $O(h^{-\frac{1}{2}})$. This also confirms the theory developed in the previous sections. Finally, by comparing the PSD and the MADP methods (see Tables 6.1 and 4-13.1) for the model problem with optimum parameters we verify our earlier conjecture that the former scheme should produce slightly better rates of convergence than the latter. On the other hand, by comparing the accelerated versions of the above basic methods we see that they have approximately the same rates of convergence (see Table 6.1 and Table 5-10.3). It should be pointed out that the actual application of one complete iteration of the MADP iterative method requires more arithmetic work (even if the reduction scheme is applied) than the PSD and the SOR iterative methods which makes the former method less atrractive than the two latter methods. However, this should not obscure our final evaluation of the MADP methods since their real power is expected to be brought forth (as this is the case for the ADI methods) when a sequence of parameters $\{r_i\}$ is used instead of the fixed preconditioning parameters r_1 and r_2. On the other hand, the commutative property HV=VH which is the basic condition to the theoretical development of the MADP method will restrict their application to partial differential equations of the form (1.14) where the region under consideration is rectangular.

7.7 THE BIHARMONIC EQUATION

In this section we will consider the application of the MADP method for the numerical solution of the biharmonic equation

$$\frac{\partial^4 U}{\partial x^4} + 2\frac{\partial^4 U}{\partial x^2 \partial y^2} + \frac{\partial^4 U}{\partial y^4} = f(x,y) \tag{7.1}$$

for $(x,y)\in R$, where R is the rectangular region defined by (1.15). If $f(x,y)=0$ the biharmonic equation (7.1) together with appropriate boundary conditions governs the slow flow of a viscous fluid or the transverse displacement of the middle surface of a uniform elastic plate, where $f(x,y)$ is the transverse loading on the plate.

In particular, we consider the following boundary conditions in connection with the solution of (7.1)

$$\left.\begin{aligned} U(x,y) &= e(x,y) \\ \frac{\partial^2 U(x,y)}{\partial n^2} &= g(x,y) \end{aligned}\right\} \tag{7.2}$$

where $e(x,y)$ and $g(x,y)$ are prescribed functions on the boundary ∂R of R and $\frac{\partial}{\partial n}$ is the normal derivative to ∂R. By imposing a uniform grid of mesh sizes h_x and h_y in the x- and y-directions, respectively such that

$$N_a = \frac{L_a}{h_a} \quad , \quad a=x,y \tag{7.3}$$

where N_a is an integer, then the application of the thirteen point finite difference analogue approximating (7.1) yields the difference equation

$$\hat{H}_0[u](x,y)+\hat{V}_0[u](x,y)+\hat{E}_0[u](x,y) = h_x^2 h_y^2 f(x,y) \tag{7.4}$$

where

$$\begin{aligned} \hat{H}_0[u](x,y) = \left(\frac{h_y}{h_x}\right)^2 &[u(x+2h_x,y)-4u(x+h_x,y)+6u(x,y) \\ &-4u(x-h_x,y)+u(x-2h_x,y)], \end{aligned} \tag{7.5}$$

$$\begin{aligned} \hat{V}_0[u](x,y) = \left(\frac{h_x}{h_y}\right)^2 &[u(x,y+2h_y)-4u(x,y+h_y)+6u(x,y) \\ &-4u(x,y-h_y)+u(x,y-2h_y)] \end{aligned} \tag{7.6}$$

and

$$\hat{E}_0[u](x,y) = 2\{u(x-h_x,y-h_y)+u(x+h_x,y-h_y)+u(x-h_x,y+h_y)$$
$$+u(x+h_x,y+h_y)-2[u(x-h_x,y)+u(x+h_x,y)+u(x,y+h_y)$$
$$+u(x,y-h_y)]+4u(x,y)\}. \tag{7.7}$$

On the other hand, if we approximate the second normal derivatives such that (7.4) incorporates the boundary conditions, then by using the natural ordering of the grid points we arrive at a system of equations of the form

$$Au = (\hat{H}_0+\hat{V}_0+\hat{E}_0)u = b \tag{7.8}$$

where the matrices $\hat{H}_0,\hat{V}_0$ and $\hat{E}_0$ correspond to the operators $\hat{H}_0[u],\hat{V}_0[u]$ and $\hat{E}_0[u]$, respectively. More precisely, using tensor products we have that the coefficient matrix has the form

$$A = (H+V)^2 \tag{7.9}$$

where

$$H = I_x \otimes U_x\ , \quad V = U_y \otimes I_y \tag{7.10}$$

and

$$U_a = \frac{h_b}{h_a}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{7.11}$$

with a=x,y, $b\in\{x,y\}-\{a\}$ and I_a is the unit matrix of order N_b-1. It can be easily verified now that the matrices H,V are symmetric and commute, thus we can apply the MADP iterative method for the numerical solution of the matrix equation (7.8).

7.8 THE MADP METHOD FOR THE NUMERICAL SOLUTION OF THE BIHARMONIC EQUATION

From (7.9) we have that the form of A can be given more explicitly as

$$A = H^2+V^2+2HV \tag{8.1}$$

which indicates that if we consider the conditioning matrix to have the form

$$R = (I+r_1H^2)(I+r_2V^2) = I+r_1H^2+r_2V^2+r_1r_2(HV)^2 , \tag{8.2}$$

then by comparing (8.1) and (8.2) we see that R approximates the matrix A reasonably well. Consequently, if we use the matrix R given by (8.2) as the conditioning matrix, then the MADP method is defined by

$$u^{(n+1)} = u^{(n)} + \tau(I+r_2V^2)^{-1}(I+r_1H^2)^{-1}(b-Au^{(n)}) \tag{8.3}$$

where again we have that r_1, r_2 and τ are real parameters to be defined later. In order to compute the iterative scheme (8.3) we can either work with vector corrections or we can employ the following two-level form (similar to (3.4))

$$(I+r_1H^2)u^{(n+\frac{1}{2})} = [I+(r_1-\tau)H^2]u^{(n)}-\tau[V^2u^{(n)}+2HVu^{(n)}-b]$$

and

$$(I+r_2V^2)u^{(n+1)} = u^{(n+\frac{1}{2})} + r_2V^2u^{(n)} \tag{8.4}$$

where we see that it is not necessary to recompute $V^2u^{(n)}$ in the second half iteration, hence we can apply a reduction scheme similar to (A.11). Since now the matrix V^2 is quindiagonal this technique results in the saving of a considerable amount of computational effort.

From (8.3) we have that the iteration matrix is given by

$$\hat{T}_{\tau,r_1,r_2} = I-\tau(I+r_2V^2)^{-1}(I+r_1H^2)^{-1}A \tag{8.5}$$

whereas the preconditioned matrix

$$\hat{B}_{r_1,r_2} = (I+r_2V^2)^{-1}(I+r_1H^2)^{-1}A \tag{8.6}$$

is positive definite for all $r_1, r_2 \in (0,\infty)$.

Since H and V are pairwise commutative, we can easily find again that the eigenvalues of $\hat{B}_{r_1,r_2}$ are given by the expression

$$\lambda \equiv \lambda(\mu,\nu,r_1,r_2) = \frac{(\mu+\nu)^2}{(1+r_1\mu^2)(1+r_2\nu^2)} \tag{8.7}$$

where μ,ν denote the eigenvalues of H and V, respectively.

Furthermore, it is known that the eigenvalues μ and ν are given by

$$\begin{aligned} \mu &= (h_y/h_x)4\sin^2(i\pi/2N_x) \ , \text{ for } i=1,2,\ldots,N_x-1 \\ \nu &= (h_x/h_y)4\sin^2(j\pi/2N_y) \ , \text{ for } j=1,2,\ldots,N_y-1 \end{aligned} \tag{8.8}$$

and therefore they are bounded as follows

$$\begin{aligned} 0<a = (h_y/h_x)4\sin^2(\pi/2N_x) \leq \mu \leq (h_y/h_x)4\cos^2(\pi/2N_x) = b \\ 0<\alpha = (h_x/h_y)4\sin^2(\pi/2N_y) \leq \nu \leq (h_x/h_y)4\sin^2(\pi/2N_y) = \beta \end{aligned} \tag{8.9}$$

The determination of the involved parameters r_1,r_2 such that the rate of convergence of the iterative scheme (8.4) is maximised are obtained for those values of r_1 and r_2 for which the P-condition number of $\hat{B}_{r_1,r_2}$ which is given by the expression

$$P(\hat{B}_{r_1,r_2}) = \frac{\lambda_M}{\lambda_m} \tag{8.10}$$

where

$$\lambda_M = \max_{\mu,\nu} \lambda \quad \text{and} \quad \lambda_m = \min_{\mu,\nu} \lambda \tag{8.11}$$

is minimised, whereas the optimum value of τ is again

$$\tau_0 = \frac{2}{\lambda_M+\lambda_m} \ . \tag{8.12}$$

Next, we can proceed to develop a similar analysis for the determination of the above parameters as in Sections 7.3.1 and 7.3.2.

7.8.1 The case where the eigenvalue ranges of H and V are the same

In this case we prove the following theorem:

Theorem 8.1.1

Let H,V be the matrices defined by (7.10) and (7.11) with eigenvalues μ,ν respectively such that

$$0<a\leq\mu,\nu\leq b. \tag{8.1.1}$$

Then $P(\hat{B}_{r_1,r_2})$ is given in Tables 8.1.4 and 8.1.5 for the different ranges of the preconditioning parameters $r_1,r_2\in(0,\infty)$.

Moreover, $P(\hat{B}_{r_1,r_2})$ is minimised if we let

$$r_1 = r_2 = r' = 1/(ab) \tag{8.1.2}$$

and its corresponding value is given by the expression

$$P(\hat{B}_{r',r'}) = (a+b)^2 r'/4. \tag{8.1.3}$$

On the other hand, if we also let

$$\tau = \tau_0 = 2r'/(1+1/P(\hat{B}_{r',r'})) \ , \tag{8.1.4}$$

then the spectral radius $S(\hat{T}_{\tau,r_1,r_2})$ attains its minimum value which is given by the expression

$$S(\hat{T}_{\tau_0,r',r'}) = \frac{(b-a)^2}{(a+b)^2+4ab} \tag{8.1.5}$$

Proof

We notice that (8.7) can be rewritten as

$$\lambda(\mu,\nu,r_1,r_2) = g(\mu,\nu,r_1,r_2)+h(\mu,\nu,r_1,r_2) \tag{8.1.6}$$

where

$$g \equiv g(\mu,\nu,r_1,r_2) = \frac{\mu^2+\nu^2}{(1+r_1\mu^2)(1+r_2\nu^2)} \tag{8.1.7}$$

and

$$h \equiv h(\mu,\nu,r_1,r_2) = \frac{2\mu\nu}{(1+r_1\mu^2)(1+r_2\nu^2)} \ . \tag{8.1.8}$$

Evidently, $g,h>0$ for all $r_1,r_2\in(0,\infty)$ and μ,ν lying in the range given by (8.1.1).

From (8.1.6) we have that

$$\lambda_M = \max_{\mu,\nu}\{g+h\}\leq\max_{\mu,\nu}\{g\} + \max_{\mu,\nu}\{h\} \tag{8.1.9}$$

and similarly

$$\lambda_m = \min_{\mu,\nu}\{g+h\}\geq\min_{\mu,\nu}\{g\} + \min_{\mu,\nu}\{h\}. \tag{8.1.10}$$

The reason we expressed λ_M,λ_m by (8.1.9) and (8.1.10) is that in this way we have to study the behaviour of the functions g,h instead of

the behaviour of λ. But from (8.1.7) and (3.11) it follows that the behaviour of g can be summarised in Table 8.1.1 (which is similar to Table 3.1.1)

r_1-Domain	r_2-Domain	max{g}	min{g}
$0<r_1\leq 1/b^2$	$0<r_2\leq 1/b^2$	C'	A'
	$1/b^2\leq r_2\leq 1/a^2$	D'	A'
	$1/a^2\leq r_2<\infty$	D'	B'
$1/b^2\leq r_1\leq 1/a^2$	$0<r_2\leq 1/b^2$	B'	A'
	$1/b^2\leq r_2\leq 1/a^2$	max{B',D'}	min{A',C'}
	$1/a^2\leq r_2<\infty$	D'	C'
$1/a^2\leq r_1<\infty$	$0<r_2\leq 1/b^2$	B'	D'
	$1/b^2\leq r_2\leq 1/a^2$	B'	C'
	$1/a^2\leq r_2<\infty$	A'	C'

TABLE 8.1.1

THE FUNCTION $g(\mu,\nu,r_1,r_2)$

where

$$A' = g(a,a,r_1,r_2)\,, \qquad B' = g(a,b,r_1,r_2)\,,$$
$$C' = g(b,b,r_1,r_2)\,, \qquad D' = g(b,a,r_1,r_2). \tag{8.1.11}$$

We therefore have to study only the behaviour of the simpler function h instead of λ.

By taking partial derivatives of h with respect to μ and ν we obtain the following results

$$\text{sign}\left(\frac{\partial h}{\partial \mu}\right) = \text{sign}(1/\mu^2 - r_1)$$
$$\text{and} \qquad \text{sign}\left(\frac{\partial h}{\partial \nu}\right) = \text{sign}(1/\nu^2 - r_2). \tag{8.1.12}$$

From the above relationships we see that for fixed $r_1, r_2 > 0$ neither of the expressions $\frac{\partial h}{\partial \mu}$, $\frac{\partial h}{\partial \nu}$ changes sign as μ and ν vary in the interval (8.1.1). Consequently, the possible extreme values of h will occur at the points (a,a), (a,b), (b,a) and (b,b). On the other hand,

if we let

$$A = h(a,a,r_1,r_2), \qquad B = h(a,b,r_1,r_2),$$
$$C = h(b,b,r_1,r_2) \quad \text{and} \quad D = h(b,a,r_1,r_2), \tag{8.1.13}$$

then the order of the quantities A,B,C and D is determined by the following relationships

$$\text{sign}(A-B) = \text{sign}(D-C) = \text{sign}(r_2-1/(ab))$$
$$\text{and} \quad \text{sign}(A-D) = \text{sign}(B-C) = \text{sign}(r_1-1/(ab)). \tag{8.1.14}$$

In view of (8.1.14) we construct Table 8.1.2 which presents the maximum and minimum values of h with respect to μ,ν for the different values of r_1 and r_2 in the interval $(0,\infty)$.

r_1-Domain	r_2-Domain	max{h}	min{h}
$0<r_1\leq 1/(ab)$	$0<r_2\leq 1/(ab)$	C	A
	$1/(ab)\leq r_2<\infty$	D	B
$1/(ab)\leq r_1<\infty$	$0<r_2\leq 1/(ab)$	B	D
	$1/(ab)\leq r_2<\infty$	A	C

TABLE 8.1.2

THE FUNCTION $h(\mu,\nu,r_1,r_2)$

In order to form the function λ (using the relationships (8.1.9) and (8.1.10)) we note from Table 8.1.2 that we have to examine further the relative positions of r_1 and r_2 with respect to the value $1/(ab)$ in the study of the function $g(\mu,\nu,r_1,r_2)$. As a first step towards this direction we extend the case where $1/b^2\leq r_1\leq 1/a^2$ in Table 8.1.1 by constructing (in a similar manner to Table 3.1.3) Table 8.1.3 which can be properly modified to yield Table 8.1.4. Further, by taking also into consideration the position of r_2 with respect to the point $1/(ab)$ in the remaining cases in Table 8.1.1, we form Table 8.1.5.

r_1-Domain	r_2-Domain	max{g}	min{g}
$\frac{1}{b^2} \leq r_1 \leq \frac{1}{ab}$	$\frac{1}{b^2} \leq r_2 \leq r_1$	B'	A'
	$r_1 \leq r_2 \leq \frac{1}{r_1 (ab)^2}$	D'	A'
	$\frac{1}{r_1 (ab)^2} \leq r_2 \leq \frac{1}{a^2}$	D'	C'
$\frac{1}{ab} \leq r_1 \leq \frac{1}{a^2}$	$\frac{1}{b^2} \leq r_2 \leq \frac{1}{r_1 (ab)^2}$	B'	A'
	$\frac{1}{r_1 (ab)^2} \leq r_2 \leq r_1$	B'	C'
	$r_1 \leq r_2 \leq \frac{1}{a^2}$	D'	C'

TABLE 8.1.3

r_1-Domain	r_2-Domain	max{g}	max{h}	min{g}	min{h}	$P(\hat{B}_{r_1,r_2})$
$\frac{1}{b^2} \leq r_1 \leq \frac{1}{ab}$	$\frac{1}{b^2} \leq r_2 \leq r_1$	B'	C	A'	A	(B'+C)/(A'+A)
	$r_1 \leq r_2 \leq \frac{1}{ab}$	D'	C	A'	A	(D'+C)/(A'+A)
	$\frac{1}{ab} \leq r_2 \leq \frac{1}{r_1 (ab)^2}$	D'	D	A'	B	(D'+D)/(A'+B)
	$\frac{1}{r_1 (ab)^2} \leq r_2 \leq \frac{1}{a^2}$	D'	D	C'	B	(D'+D)/(C'+B)
$\frac{1}{ab} \leq r_1 \leq \frac{1}{a^2}$	$\frac{1}{b^2} \leq r_2 \leq \frac{1}{r_1 (ab)^2}$	B'	B	A'	D	(B'+B)/(A'+D)
	$\frac{1}{r_1 (ab)^2} \leq r_2 \leq \frac{1}{ab}$	B'	B	C'	D	(B'+B)/(C'+D)
	$\frac{1}{ab} \leq r_2 \leq r_1$	B'	A	C'	C	(B'+A)/(C'+C)
	$r_1 \leq r_2 \leq \frac{1}{a^2}$	D'	A	C'	C	(D'+A)/(C'+C)

TABLE 8.1.4

r_1-Domain	r_2-Domain	max{g}	max{h}	min{g}	min{h}	$P(\hat{B}_{r_1,r_2})$
$0<r_1\leq\frac{1}{b^2}$	$0<r_2\leq\frac{1}{b^2}$	C'	C	A'	A	(C'+C)/(A'+A)
	$\frac{1}{b^2}\leq r_2\leq\frac{1}{ab}$	D'	C	A'	A	(D'+C)/(A'+A)
	$\frac{1}{ab}\leq r_2\leq\frac{1}{a^2}$	D'	D	A'	B	(D'+D)/(A'+B)
	$\frac{1}{a^2}\leq r_2<\infty$	D'	D	B'	B	(D'+D)/(B'+B)
$\frac{1}{a^2}\leq r_1<\infty$	$0<r_2\leq\frac{1}{b^2}$	B'	B	D'	D	(B'+B)/(D'+D)
	$\frac{1}{b^2}\leq r_2\leq\frac{1}{ab}$	B'	B	C'	D	(B'+B)/(C'+D)
	$\frac{1}{ab}\leq r_2\leq\frac{1}{a^2}$	B'	A	C'	C	(B'+A)/(C'+C)
	$\frac{1}{a^2}\leq r_2<\infty$	A'	A	C'	C	(A'+A)/(C'+C)

TABLE 8.1.5

In Tables 8.1.4 and 8.1.5 we present the expressions of $P(\hat{B}_{r_1,r_2})$ for the different values of r_1 and r_2 in the interval $(0,\infty)$.

If one studies the behaviour of $\frac{\partial P(\hat{B}_{r_1,r_2})}{\partial r_2}$ (assuming r_1 is kept fixed) for all the cases in the latter two tables, then it can be easily verified that the minimum value of $P(\hat{B}_{r_1,r_2})$ is attained if

$$r_2 = \frac{1}{ab} . \tag{8.1.15}$$

Because of the symmetry of the problem we can work similarly for determining the optimum value of r_1 which obviously is identical with the value of r_2 given by (8.1.15), hence (8.1.2) follows. From Table 8.1.5 we have the following values for the smallest and largest eigenvalue of $\hat{B}_{r',r'}$

$$\lambda_M = D'+C = D'+D = B'+B = B'+A = ab$$

and

$$\lambda_m = A'+A = A'+B = C'+D = C'+C = \frac{4(ab)^2}{(a+b)^2}. \tag{8.1.16}$$

Thus from (8.1.16) and (8.10) we see that $P(\hat{B}_{r',r'})$ is given by (8.1.3) while from (8.12) the optimum value for τ is given by (8.1.4). But for this optimum value τ_0 of τ the spectral radius of the iteration matrix is given by the formula

$$S(\hat{T}_{\tau_0,r',r'}) = \frac{P(\hat{B}_{r',r'})-1}{P(\hat{B}_{r',r'})+1} \tag{8.1.17}$$

which by (8.1.3) gives (8.1.5) and the proof of the theorem is complete.

7.8.2 The case where the eigenvalue ranges of H and V may be different

In this case we prove the following theorem:

Theorem 8.2.1

Let H and V be the matrices defined by (7.10) and (7.11) with eigenvalues μ,ν, respectively such that

$$0<a\leq\mu\leq b \quad \text{and} \quad 0<\alpha\leq\nu\leq\beta\ . \tag{8.2.1}$$

Then the P-condition number of $\hat{B}_{r_1,r_2}$ is minimised if we let

$$r_1^* = \frac{1-\Sigma^2 sc^{\frac{1}{2}}}{-t+\Sigma^2 qc^{\frac{1}{2}}}, \quad r_2^* = \frac{1+\Sigma^2 sc^{\frac{1}{2}}}{t+\Sigma^2 qc^{\frac{1}{2}}} \tag{8.2.2}$$

where

$$c = \frac{1}{1+\theta+[\theta(2+\theta)]^{\frac{1}{2}}}, \tag{8.2.3}$$

$$\theta = \frac{2(\beta^2-\alpha^2)(b^2-a^2)}{(a^2+\alpha^2)(b^2+\beta^2)}, \tag{8.2.4}$$

$$\Sigma_s^2 = \frac{(\beta^2-\alpha^2)-(b^2-a^2)}{(b^2+\beta^2)-(a^2+\alpha^2)c}, \tag{8.2.5}$$

$$\Sigma_q^2 = \frac{(b^2+\beta^2)+(b^2-\beta^2)\Sigma_s^2}{2}, \tag{8.2.6}$$

$$t = \frac{(b^2-\beta^2)+(b^2+\beta^2)\Sigma_s^2}{2}, \tag{8.2.7}$$

and its corresponding value is given by

$$P(\hat{B}_{r_1^*,r_2^*}) = \frac{(1+c^{\frac{1}{2}})^2}{4c^{\frac{1}{2}}} . \tag{8.2.8}$$

On the other hand, if we also let

$$\tau = \tau_0^* = (r_1^*+r_2^*)/(1+1/P(\hat{B}_{r_1^*,r_2^*})), \tag{8.2.9}$$

then the spectral radius $S(\hat{T}_{\tau,r_1^*,r_2^*})$ attains its minimum value which is given by the expression

$$S(\hat{T}_{\tau_0^*,r_1^*,r_2^*}) = \frac{(1-c^{\frac{1}{2}})^2}{(1+c^{\frac{1}{2}})^2+4c^{\frac{1}{2}}} . \tag{8.2.10}$$

Proof

From the previous section is can be noticed that the value of the optimum parameters which minimise the P-condition number of the matrix $\hat{B}_{r_1,r_2}$ is identical with the one which minimises the ratio

$$G = \frac{\max\{g\}}{\min\{g\}} \tag{8.2.11}$$

Indeed, we observe that the function $g(\mu,\nu,r_1,r_2)$ is obtained from the function $\lambda(\mu,\nu,r_1,r_2)$ given by (3.11) with μ,ν being replaced by μ^2,ν^2, respectively. Consequently, from Theorem 3.1.1 we have that G is minimised if we let r_1,r_2 take the values given by (8.1.2) since $0<a^2\leq\mu^2,\nu^2\leq b^2$. On the other hand, the behaviour of $P(\hat{B}_{r_1,r_2})$ is not affected by the bilinear transformation (3.2.13), in the sense that it is the same between the original and the corresponding transformed intervals. Thus, if we transform our problem (using a similar analysis to Section 7.3.2) so that we return to the previous case of the "single range", then the optimum values of the corresponding transformed parameters r_1 and r_2 will still remain the same as the ones which minimise the transformed ratio G. In order words our problem is identical with the one tackled in Section 7.3.2, the only difference being that instead of having μ,ν in (3.2.11) here we have μ^2,ν^2. Thus, by adhering again to the analysis of Wachspress and Jordan we seek to introduce new variables

$\hat{\mu}^2,\hat{\nu}^2$ such that

$$\mu^2 = \frac{t+q\hat{\mu}^2}{1+s\hat{\mu}^2}, \qquad \nu^2 = \frac{t'+q'\hat{\nu}^2}{1+s\hat{\nu}^2} \tag{8.2.12}$$

so that for some $\hat{\omega}_1$ and $\hat{\omega}_2$ we have

$$\left(\frac{\mu^2-\omega_2}{\mu^2+\omega_1}\right)\left(\frac{\nu^2-\omega_1}{\nu^2+\omega_2}\right) = \left(\frac{\hat{\mu}^2-\hat{\omega}_2}{\hat{\mu}^2+\hat{\omega}_1}\right)\left(\frac{\hat{\nu}^2-\hat{\omega}_1}{\hat{\nu}^2+\hat{\omega}_2}\right) \tag{8.2.13}$$

where $\hat{\mu}^2$ and $\hat{\nu}^2$ vary over the ranges

$$\sigma^2 \leq \hat{\mu}^2 \leq \Sigma^2, \qquad \sigma^2 \leq \hat{\nu}^2 \leq \Sigma^2. \tag{8.2.14}$$

By following the same analysis as in Section 7.3.2 we can show the validity of Theorem 8.2.1.

By Theorem 8.1.1 we see that as the P-condition number of the preconditioned matrix $\hat{B}_{r',r'}$ increases, then the optimum value of τ tends to be equal to $2r'$. In other words, for sufficiently small mesh size the Peaceman-Rachford ADI scheme for the numerical solution of the biharmonic equation (i.e. $u^{(n+1)}=u^{(n)}+2r'(I+r'V^2)^{-1}(I+r'H^2)^{-1}(b-Au^{(n)})$) tends to attain the same rate of convergence as the MADP method. However, this will not be the case if more accurate difference analogues are used (e.g. 25-point difference formula).

It should be mentioned that similar results to Section 7.8.1 for the case where the eigenvalue ranges of H and V are the same, have been obtained by Gane [1974] whereas another approach for the same problem using the EADI method has been developed by Hadjidimos [1975]. However, for the case where the eigenvalue ranges of the basic matrices involved are different, the optimum parameters were found (see Gane and Evans [1974]) under the assumption that $0<\alpha' \leq \mu,\nu \leq \beta'$, where $\alpha'=\min(a,\alpha)$ and $\beta'=\max(b,\beta)$. In an analogous way to Section 7.4 we can easily define the accelerated procedures based on the iterative scheme (8.3) and obtain an order of magnitude improvement on the convergence rate.

7.9 RATES OF CONVERGENCE ON THE UNIT SQUARE

If we consider the solution of the biharmonic equation in the unit square with $h_x=h_y=h$, then by (8.9) we have

$$a = \alpha = 4\sin^2(\tfrac{\pi h}{2}) \quad \text{and} \quad b = \beta = 4\cos^2(\tfrac{\pi h}{2}) \tag{9.1}$$

hence from Theorem 8.1.1 we obtain successively

$$r' = \frac{1}{4\sin^2(\pi h)}, \tag{9.2}$$

$$P(\hat{B}_{r',r'}) = \frac{1}{\sin^2(\pi h)} \tag{9.3}$$

and

$$\tau_0 = \frac{1}{2\sin^2(\pi h)[1+\sin^2(\pi h)]}. \tag{9.4}$$

Finally, the spectral radius is given by the expression

$$S(\hat{T}_{\tau_0,r',r'}) = \frac{1-\sin^2(\pi h)}{1+\sin^2(\pi h)} \tag{9.5}$$

thus the rate of convergence of the iterative scheme (8.3) is

$$R(\hat{T}_{\tau_0,r',r'}) \sim 2\pi^2 h^2 \tag{9.6}$$

for sufficiently small h.

If on the other hand, we use the 25-point difference analogue to approximate the biharmonic equation, then the matrix A has the following splitting

$$A = (H+V-kHV)^2 \tag{9.7}$$

where k=1/6. Evidently, for k=0 the iterative scheme (8.3) is fourth order correct in h, while for k=1/6 it is eighth order correct in h. By following a similar approach as in Section 7.5 we can find that the eigenvalues of $\hat{B}_{r_1,r_2}$ are given by the expression

$$\lambda = \frac{(\mu+\nu-k\mu\nu)^2}{(1+r_1\mu^2)(1+r_2\nu^2)} \tag{9.8}$$

and can be bound as follows

$$(1-kb/2)^2\phi \leq \lambda \leq (1-ka/2)^2\phi \tag{9.9}$$

where

$$\phi = \frac{(\mu+\nu)^2}{(1+r_1\mu^2)(1+r_2\nu^2)} . \tag{9.10}$$

From the above we find again that if we let

$$r_1 = r_2 = r' = \frac{1}{ab} , \tag{9.11}$$

then $P_k(\hat{B}_{r_1,r_2})$ is minimised and its corresponding value is given by

$$P_k(\hat{B}_{r',r'}) = k''P(\hat{B}_{r',r'}) \tag{9.12}$$

where

$$k'' = \left(\frac{1-ka/2}{1-kb/2}\right)^2 . \tag{9.13}$$

Moreover, from (8.1.16) and (9.9) we find that if we let

$$\tau_0 = \frac{(a+b)^2 r'^2}{2(1-kb/2)^2[1+(k'')^2 P(\hat{B}_{r',r'})]} , \tag{9.14}$$

the spectral radius is also minimised and given by the expression

$$S_k(\hat{T}_{\tau_0,r',r'}) = \frac{k''P(\hat{B}_{r',r'})-1}{k''P(\hat{B}_{r',r'})+1} , \tag{9.15}$$

therefore the rate of convergence is

$$R_k(\hat{T}_{\tau_0,r',r'}) \sim \frac{2}{k''P(\hat{B}_{r',r'})} . \tag{9.16}$$

Finally, if we consider the application of the MADP-SI method for the solution of the present problem, then the rate of convergence is

$$R_{k,\infty}(P_n(\hat{T}_{\tau_0,r',r'})) \sim 2/\sqrt{k''P(\hat{B}_{r',r'})} \tag{9.17}$$

which for k=0, (9.3),(9.12) and (9.13) give the result

$$R_\infty(P_n(\hat{T}_{\tau_0,r',r'})) \sim 2\pi h \tag{9.18}$$

for sufficiently small h.

7.10 NUMERICAL RESULTS

In order to verify our theoretical results of the previous section we solved the biharmonic equation

$$\frac{\partial^4 U}{\partial x^4} + 2\frac{\partial^4 U}{\partial x^2 \partial y^2} + \frac{\partial^4 U}{\partial y^4} = 0, \quad (x,y)\in R, \tag{10.1}$$

where the region R was the unit square. The boundary conditions were as given in Figure 10.1. By applying the 13-point difference analogue we approximated 10.1 and the produced system was solved with the MADP method as defined by (8.4) and the MADP-SI method defined by

$$u^{(n+1)} = (1-\rho_{n+1})u^{(n+1)} + \rho_{n+1}(\hat{T}_{\tau_0,r',r'}u^{(n)}+\hat{t}) \tag{10.2}$$

where

$$\left.\begin{aligned} \rho_1 &= 1 \\ \rho_2 &= \left(1-\frac{\sigma^2}{2}\right)^{-1} \\ \rho_{n+1} &= \left(1-\frac{\sigma^2\rho_n}{4}\right)^{-1}, \quad n=2,3,\ldots \end{aligned}\right\} \tag{10.3}$$

and

$$\sigma = \frac{P(\hat{B}_{r',r'})-1}{P(\hat{B}_{r',r'})+1}.$$

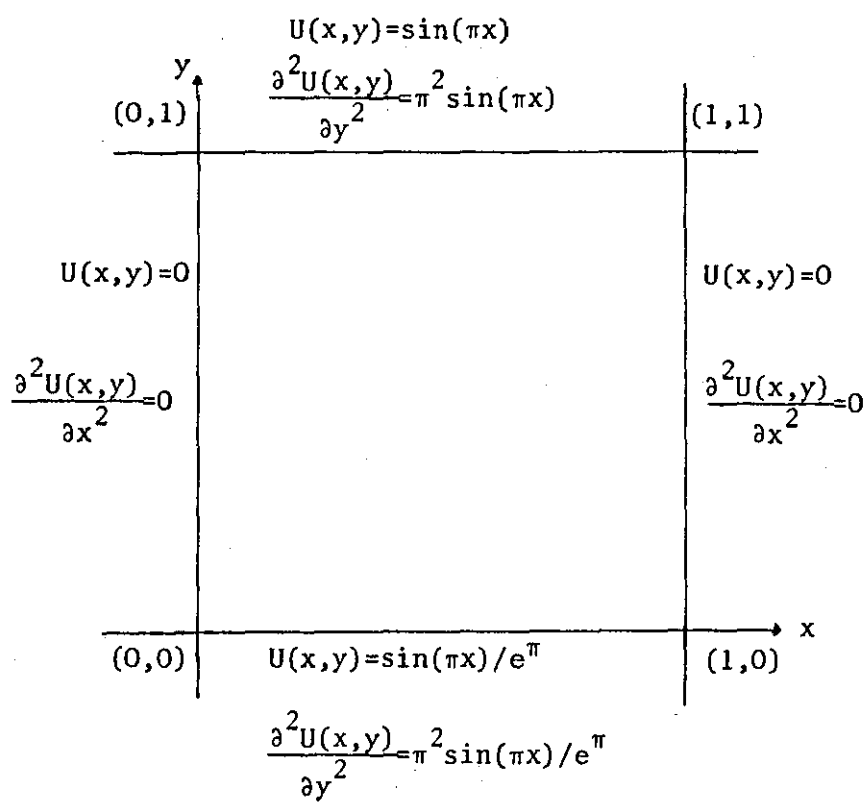


FIGURE 10.1

As starting vector $u^{(0)}$ we used the vector with all its components equal to unity while for convergence the following criterion was required to be satisfied

$$\max|u^{(n+1)}-u^{(n)}|\leq 10^{-6}.$$

In Table 10.1 we present the number of iterations required to solve the present problem with the iterative procedures mentioned above for the different mesh sizes shown. Furthermore, Figure 10.2 shows graphs with logarithmic scales of the observed number of iterations versus h^{-1} for the MADP and MADP-SI methods.

h^{-1}	r'	τ_0	$P(\hat{B}_{r',r'})$	$S(\hat{T}_{\tau_0,r',r'})$	MADP	MADP-SI
15	5.7834	11.0875	23.1335	0.9171	144	36
20	10.2159	19.9437	40.8635	0.9522	256	50
30	22.8808	45.2670	91.5231	0.9784	570	76
40	40.6119	80.7269	162.4476	0.9878	1012	103
50	63.4091	126.3202	253.6366	0.9921	-	129

TABLE 10.1

NUMERICAL RESULTS FOR THE BIHARMONIC EQUATION

From Table 10.1 we see that for the different mesh sizes shown the optimum value of τ_0 is close to $2r'$ which indicates that the rate of convergence of the MADP and the Peaceman-Rachford ADI method is approximately the same for the biharmonic equation. This observation could have been made earlier, when we found the expression for the optimum value of τ_0 (see (8.1.4)). On the other hand, Figure 10.2 verifies our expectations (see (9.6) and (9.18)) by showing that the number of iterations of the MADP method for the biharmonic equation (7.1) varies approximately like $O(h^{-2})$, whereas for the MADP-SI method like $O(h^{-1})$. Finally, a further improvement (perhaps by an order of magnitude) on the rate of convergence can be achieved by considering a sequence of parameters $\{r_i\}$ (see Conte and Dames [1958] and Hadjidimos [1969]).

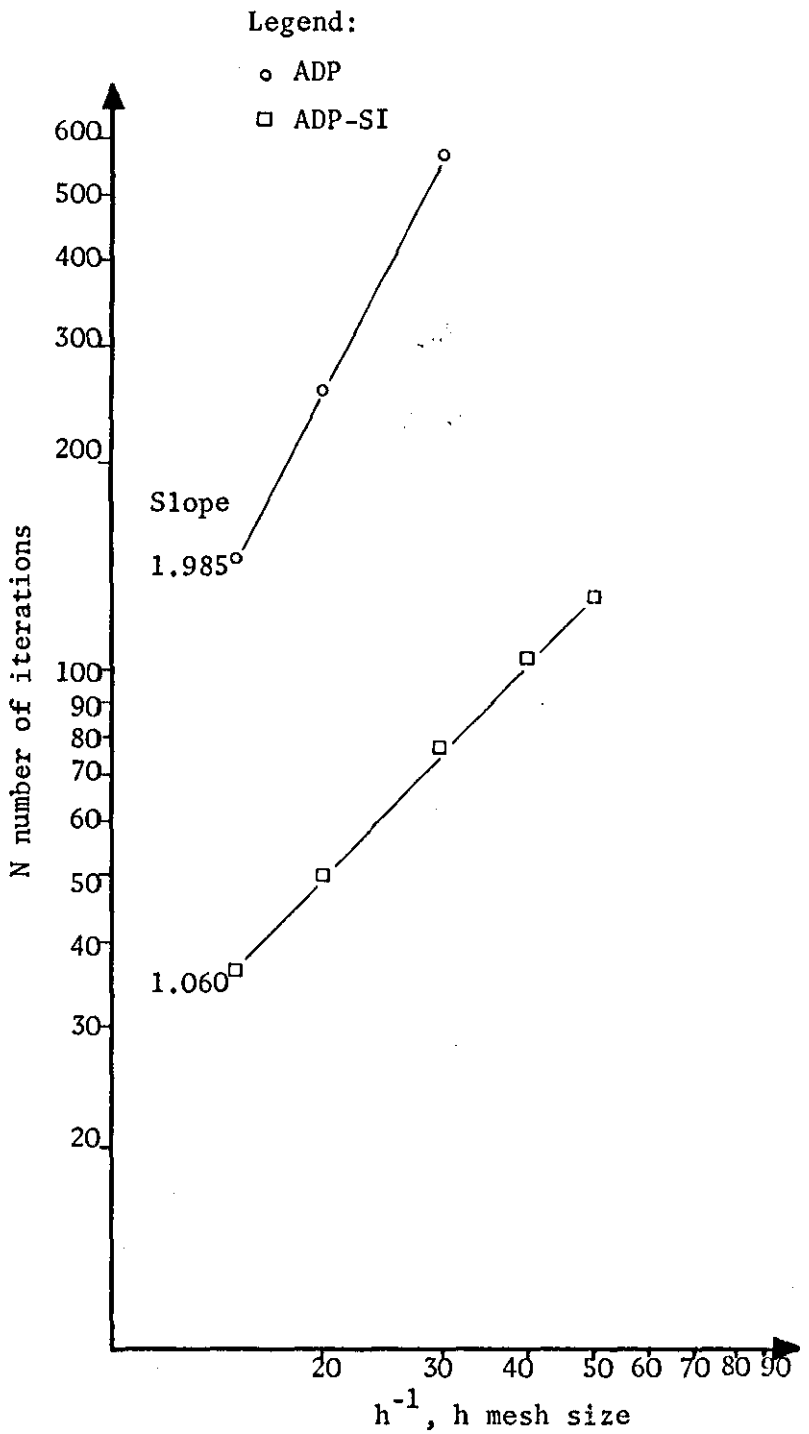


FIGURE 10.2

DETERMINATION OF RATE OF CONVERGENCE ATTAINED FOR THE BIHARMONIC EQUATION USING ADP AND ADP-SI METHODS

CHAPTER 8

SUMMARY AND CONCLUSIONS

Preconditioning techniques have been considered in the form of the "Preconditioned Simultaneous Displacement" (see Evans [1968]) and the "Alternating Direction Preconditioning" (see Gane and Evans [1974]) since their introduction. In this investigation we have shown that the preconditioning techniques can be generalised in such a way so that all known iterative schemes are special cases of a wider class of iterative methods. On the other hand, it is hoped that they will also provide a practical test and a guide line for the formulation of perhaps more efficient iterative procedures in the future. In fact, this is achieved from the experience which one obtains by attempting to explain under the "preconditioning" approach their origins and to establish a priori criteria as far as the efficiency of the basic iterative processes is concerned.

In this context we commenced our study by introducing in Chapter 4 new iterative schemes (the extrapolated versions of the GS and SOR) and also the related theory as well. For example, we have shown that the technique of extrapolating the GS method in order to obtain the SOR method can be regarded as a special case of a more general approach which yields an iterative scheme (ESOR method) with faster rate of convergence than the SOR. However, the rapidity on the rate of convergence of ESOR depends strongly upon the quantity $\underline{\mu}$ and therefore further research is needed in order to establish in what degree the required extra computational work affects the efficiency of the method as compared with SOR.

In the remainder of Chapter 4 we considered the two classical methods of the "Preconditioning theory", namely the Preconditioned Jacobi and the Preconditioned Simultaneous Displacement in a different form which does not necessitate any transformation, thus resulting in a reduction of the involved computational work as compared with the form they were first introduced.

The two aforementioned methods were studied in detail and their properties were clarified through the development of the theory concerning

their convergence and the choice of "good" values for the involved parameters. As a result of this investigation, we were able to determine a substantial improvement on the rate of convergence of the PSD method over SSOR which was also confirmed by our numerical results. Furthermore, the alternative form of the PSD method, presented in Appendix A, in combination with Niethammer's scheme may be regarded as an alternative procedure to SOR (see Tables 4-13.1 and 4-13.2) for problems with $\bar{\beta} \leq 1/4$. It is conjectured that $\bar{\beta} \leq 1/4$ may also be a necessary condition for other problems where the coefficients do not belong in class $C^{(2)}$. However, further research is clearly needed towards this direction before any firm conclusions are drawn. In the last two sections of the same chapter we considered a more general form of the PSD method, namely the unsymmetric PSD method in combination with the red-black ordering. As a conclusion from this study we have that although the aforementioned method has an identical spectral radius with SOR at the optimum stage (with red-black ordering) it requires twice the computational work. Although the UPSD method was not proved to be an efficient method with red-black ordering it will be interesting to investigate the possibility of using it with the natural ordering. A final result of Chapter 4 was that the application of the PSD method with red-black ordering yielded a rate of convergence which differed by an order of magnitude from that using the natural ordering. We therefore conclude that the PSD method should always be used in connection with the natural ordering.

In the first part of Chapter 5 we defined the line ESOR and the line PSD methods. Further, by adhering to the analysis of the point PSD method we were able to determine good estimates for the involved parameters in the LPSD in the sense that, at least for the model problem, the rate of convergence was approximately O(h). Also it was found that LPSD was $\sqrt{2}$ times faster than the point PSD. This result characterises the SOR method as well. Further study showed that, as in the point methods, the LPSD method is approximately 2 times faster than the line SSOR method. These

findings were also confirmed by our results obtained from numerical experiments. These results indicate that (see Table 5-4.1) although we do not have a monotonicity theorem for π_1, the LPSD method attains a convergence of about O(h) for subregions of the square. From the analysis and results of the LPSD method we conclude that this method possesses all the features which characterise the line methods and therefore it should be preferred over LSSOR since it also requires approximately the same storage and computational work.

In our attempt to further increase the rate of convergence of the PSD method, in the second part of Chapter 5 we consider various accelerating techniques which essentially prove that there exists a possibility of improving the rate of convergence of the PSD method by an order of magnitude. A principal result of this analysis is that if the coefficients A(x,y) and C(x,y) are in the class $C^{(2)}$ in R+∂R, then for h small we have

$$S(LU) \leq 1/4 + O(h^2)$$

which implies that the constant γ^{-1} appearing in (4-11.18) is bounded away from zero as h→0. The above condition guarantees that one indeed obtains an order of magnitude improvement in the rate of convergence of the PJ-SI method as compared with the J-SI,PSD and the SOR methods. Applying semi-iterative techniques to the PSD method, when A is a positive definite L-matrix, we proved that the PJ-SI method is asymptotically at least as good as the J-SI method. However, this comparison was based on the number of iterations and did not take into account the fact that each PJ-SI iteration requires about twice as much work as each Jacobi iteration. From the analysis and results presented in Chapter 5 one concludes that the accelerated versions of the PSD method offer a substantial saving as compared with the SOR method for many problems. From these procedures, the PJ-SI and the PJ-VE are very promising for the following reasons. The former method achieves a fast rate of convergence but requires approximately twice as much work as the SOR whereas the latter yields approximately the same rate of

convergence but the amount of computational work is substantially reduced (approximately the same as SOR) with the application of Niethammer's scheme (see (A.17)). However, in order to avoid the instability which may occur by using the PJ-VE method one should follow the suggestions of Lebedev and Finogenov [1971] for the choice of the iteration parameters.

In Chapter 6 we essentially considered the adaptive algorithms in order to further accelerate the PSD method. In conclusion we found that the PJ-SI method with either estimated or adaptively determined parameters yields faster rates of convergence than the SOR method (although in terms of the work required, SOR may still be preferable in certain cases). The analysis for the development of the adaptive algorithm based on the PJ-SI procedure can also be applied in conjunction with the PJ-VE method. Since for the latter method, there is a possibility of reducing the work involved, it would be interesting to develop an algorithm based on the PJ-VE method with Niethammer's scheme, which adaptively determines the involved parameters.

Finally in Chapter 7, following the suggestions which emerge from the preconditioning techniques, we considered another known splitting of the matrix A, as used in the Alternating Direction Implicit methods and determined the Modified Alternating Direction Preconditioning method. As a first step, we assumed that all the involved parameters were fixed and we developed the analysis for determining their optimum values in the general case where the matrices H and V had different eigenvalue ranges. For the numerical solution of the partial differential equation of the form (7-1.14) we conclude that in the case of using the five point difference analogue, the MADP method becomes identical with the Peaceman-Rachford ADI method at the optimum stage. However, if more accurate difference formulae are employed, then the former method is different from the latter and is expected to yield a better rate of convergence. It is hoped that this will become apparent when one considers the application of a sequence of parameters $\{r_i\}$. Finally, the MADP method for the biharmonic equation is

different from the one presented by Conte and Dames [1958] and it is reasonable to assume that if we follow their analysis of determining a sequence of parameters $\{r_i\}$, the MADP method will probably yield a slightly faster rate of convergence as well.

REFERENCES

AITKEN, A.C. [1950], *"Studies in practical mathematics V. On the iterative solution of a system of linear equations"*,
Proc.Roy.Soc. Edinburgh Sec. A63, 52-60.

ARMS, R.J., GATES, L.D. and ZONDEK, B. [1956], *"A method of block iteration"*,
J.Soc.Indust.Appl.Math. 4, 220-229.

AXELSSON, O. [1972], *"Generalised SSOR methods"*,
Report DD/72/8 CERN-Data Handling Division, Geneva.

AXELSSON, O. [1974], *"On preconditioning and convergence acceleration in sparse matrix problems"*,
CERN 70-74,Data Handling Division, Geneva.

BECKMAN, F.S. [1960], *"The solution of linear equations by the conjugate gradient method"*,
Chapter 4 of Ralston, A. and Wilf, H.S. [1960], "Mathematical Methods for Digital Computers", Vol.I, John Wiley and Sons, Inc., New York.

BELLMAN, R. [1960], *"Introduction to Matrix Analysis"*,
McGraw-Hill, New York.

BENOKRAITIS, V.J. [1974], *"On the adaptive acceleration of symmetric successive overrelaxation"*,
Ph.D. Thesis, University of Texas, Austin.

BIRKHOFF, G. and MACLANE, S. [1953], *"A Survey of Modern Algebra"*,
MacMillan, New York.

BIRKHOFF, G., VARGA, R.S. and YOUNG, D. [1962], *"Alternating direction implicit methods"*,
Advances in Computers 3, 189-273.

COLLATZ, L. [1960], *"The Numerical Treatment of Differential Equations"*,
3rd ed., Springer-Verlag, Berlin.

CONRAD, V. and WALLACH, Y. [1977], *"A faster SSOR algorithm"*,
Numer.Math., 27, 371-372.

CONTE, S.D. and DAMES, R.T. [1958], *"An alternating direction method for solving linear equations with symmetric positive definite matrices"*,
M.T.A.C. 12, 198-205.

COURANT, R. and HILBERT, D. [1962], *"Methods of Mathematical Physics"*,
Vol.II, Interscience Publishers, New York.

CULLEN, C.G. [1974], *"Dynamic parameter determination with the Jacobi semi-iterative method"*,
Report CNA-95, University of Texas, Austin, Texas.

CUTHILL, E.H. and VARGA, R.S. [1959], *"A method of normalised block iteration"*,
J.Assoc.Comput.Mach. 6, 236-244.

DANIEL, J.W. [1967], *"Convergence of the conjugate gradient method with computationally convenient modifications"*,
Numer.Math. 10, 125-131.

DIAMOND, M.A. [1971], *"An economical algorithm for the solution of finite difference equations independent of user-supplied parameters"*, Ph.D. Thesis, Department of Computer Science, University of Illinois.

DOUGLAS, J. [1955], *"On the numerical integration of* $\partial^2 u/\partial x^2+\partial^2 u/\partial y^2=\partial u/\partial t$ *by implicit methods"*, J.Soc.Indust.Appl.Math. 3, 42-65.

DOUGLAS, J. and RACHFORD, H. [1956], *"On the numerical solution of heat conduction problems in two and three space variables"*, Trans.Amer.Math.Soc. 82, 421-439.

D'SYLVA, E. and MILES, G.A. [1963], *"The SSOR iteration scheme for equations with* σ_1 *ordering"*, Comput.J. 6, 271-273.

EHRLICH, L.W. [1963], *"The block symmetric successive overrelaxation method"*, Ph.D. Thesis, University of Texas, Austin, Texas.

EHRLICH, L.W. [1964], *"The block symmetric successive overrelaxation method"*, J.Soc.Indust.Appl.Math. 12, 807-826.

EVANS, D.J. [1963], *"The extrapolated modified Aitken iteration method for solving elliptic difference equations"*, Computer J. 6, 193-201.

EVANS, D.J. [1964], *"The extrapolated modified Aitken iteration method applied to* σ_1*-ordered systems of linear equations"*, Computer J. 7, 137-140.

EVANS, D.J. [1968], *"The use of preconditioning in iterative methods for solving linear equations with symmetric positive definite matrices"*, J.Inst.Math.Applics. 4, 295-314.

EVANS, D.J. [1973], *"Comparison of the convergence rates of iterative methods for solving linear equations with preconditioning"*, Greek Mathematical Society, Caratheodory Symposium, 106-135.

EVANS, D.J. [1973a], *"The analysis and application of sparse matrix algorithms in the finite element method"*, in "Mathematics for Finite Elements and Applications", edited by J.R. Whiteman, Academic Press, London and New York, 427-447.

EVANS, D.J. [1974], *"Iterative sparse matrix algorithms"*, in "Software for Numerical Mathematics" edited by D.J. Evans, Academic Press, 49-83.

EVANS, D.J. and FORRINGTON, C.V.D. [1963], *"An iterative process for optimising symmetric overrelaxation"*, Computer J. 6, 271-273.

FADDEEV, D.K. and FADDEEVA, V.N. [1963], *"Computational Methods of Linear Algebra"*, Freeman, San Francisco, California.

FLANDERS, D. and SHORTLEY, G. [1950], *"Numerical determination of fundamental modes"*, J.Appl.Phys. 21, 1326-1332.

FORSYTHE, G.E. and WASOW, W.R. [1960], *"Finite difference methods for partial differential equations"*, Wiley, New York.

FRIEDMAN, B. [1957], *"The iterative solution of elliptic difference equations"*,
A.E.C. Research and Development Report NYO-7698, Institute of Mathematical Sciences, New York University, New York.

GANE, C.R. [1974], *"Computational techniques for the numerical solution of parabolic and elliptic differential equations"*,
Ph.D. Thesis, Loughborough University, Loughborough.

GANE, C.R. and EVANS D.J., [1974], *"Alternating direction preconditioning techniques"*,
Report No.18, Department of Computer Studies, Loughborough University, Loughborough.

GEIRINGER, H. [1949], *"On the solution of systems of linear equations by certain iterative methods"*,
Reissner Anniversary Volume, Contribution to Applied Mechanics, 365-393. Edwards, Ann Arbor, Michigan.

GEORGE, A. [1971], *"Computer implementation of the finite element method"*,
Ph.D. Thesis, Computer Studies Department, Stanford University, STAN-CS-71-208.

GOLUB, G.H. and VARGA, R.S. [1961], *"Chebyshev semi-iterative methods, successive overrealxation iterative methods and second order Richardson iterative methods"*,
Numer.Math., Parts I and II 3, 147-168.

GOTTFRIED, B.S. and WEISMAN, J. [1973], *"Introduction to optimisation theory"*,
Prentice-Hall, Inc.,New Jersey.

GOURLAY, A.R. and WATSON, G.A. [1973], *"Computational methods for matrix eigenproblems"*,
John Wiley and Sons, New York.

GUITTET, J. [1967], *"Une nouvelle methode de directions alternees a q variables"*,
J.Math.Anal.Appl. 17, 199-213.

HABELTER, G.J. and WACHSPRESS, E.L. [1961], *"Symmetric successive overrelaxation in solving diffusion difference equations"*,
Math.Comp. 15, 356-362.

HADJIDIMOS, A. [1975], *"On comparing optimum alternating direction preconditioning and extrapolated direction implicit schemes"*,
J.Math.Anal.Appl. 59, 573-586.

HADJIDIMOS, A. [1978], *"Accelerated overrelaxation method"*,
Math.Comp. 32, 149-157.

HADJIDIMOS, A. and IORDANIDIS, K. [1974], *"Solving Laplace's equation in a rectangle by alternating direction implicit methods"*,
J.Math.Appl. 48, 353-367.

HAGEMAN, L.A. [1972], *"The estimation of acceleration parameters for the Chebyshev polynomial and successive overrelaxation iteration methods"*,
Report WAPD-TM-1038, Bettis Atomic Power Laboratory, Westinghouse Electric Corp., Pittsburgh.

HAGEMAN, L.A. and KELLOGG, R.B. [1968], *"Estimating optimum overrelaxation parameters"*,
Math.Comp. 22, 60-68.

HALMOS, P.R. [1958], *"Finite dimensional vector spaces"*,
Van Nostrand, Princeton.

HATZOPOULOS, M. [1974], *"Preconditioning and other computational techniques for the direct solution of linear systems"*,
Ph.D. Thesis, Loughborough University, Loughborough.

HESTENES, M.R. and STIEFELL, E. [1952], *"Method of conjugate gradients for solving linear systems"*,
J.Res.Nat.Bur. Standards 49, 409-436.

HOUSHOLDER, A.S. [1964], *"The Theory of Matrices in Numerical Analysis"*,
Blaisdell, New York.

IKEBE, Y. BENOKRAITIS, V. and SULLIVAN J. [1973], *"Acceleration of stationary iterative processes by adaptive extrapolation"*,
Report CNA-74, Center for Numerical Analysis, University of Texas, Austin, Texas.

KAHAN, W. [1958], *"Gauss-Seidel methods of solving large systems of linear equations"*,
Ph.D. Thesis, University of Toronto, Toronto, Canada.

KEAST, P. and MITCHELL, A.R. [1967], *"Finite difference solution of the third boundary problem in elliptic and parabolic equations"*,
Numer.Math. 10, 67-75.

KIM, Y.J. [1973], *"An efficient iterative procedure for use with the finite element method"*,
Ph.D. Thesis, Department of Computer Science, University of Illinois, UIUCDCS-R-73-600.

KINCAID, D.R. [1974], *"On complex second-degree iterative methods"*, SIAM J.Numer.Anal. 11, 211-218.

LEVEDEV, V.I. and FINOGENOV, S.A. [1971], *"Ordering of the iterative parameters in the cyclical Chebyshev iterative method"*, Zh.Vychisl,Mat. Fiz.11, 155-170.

MARCHUK, G.I. [1971], *"On the theory of the splitting up method in the numerical solution of partial differential equations"*, Academic Press, New York.

MARKOFF, W. [1892], *"Uber Polynome die in einem gegeben intervale moglichst wenig von Null abweichen"*, Math.Ann. 77 [1916], 213-258 (translation and condensation by J. Grossman of Russian article published in 1892).

NIETHAMMER, W. [1964], *"Relaxation by kimplexen matrizen"*, Math.Zeitsch 86, 34-40.

PARTER, S.V. [1961], *"Multi-line iterative methods for the elliptic difference equations and fundamental frequencies"*, Numer.Math. 3, 305-319.

PEACEMAN, D.W. and RACHFORD, H. [1955], *"The numerical solution of parabolic and elliptic differential equations"*, J.Soc.Indust.Appl.Math. 3, 28-41.

PHIEN, T. [1972], *"An application of semi-iterative and second degree symmetric successive overrelaxation iterative methods"*, M.A. Thesis, University of Texas, Austin, Texas.

REID, J.K. [1971], *"On the method of conjugate gradients for the solution of large sparse systems of linear equations"*, in "Large Sparse Sets of Linear Equations", ed. by J.K. Reid, Academic Press, London, 231-254.

RICHARDSON, L.F. [1910], *"The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stresses in a mansory dam"*, Philos.Trans.Roy.Soc. London Ser.A210, 307-357.

RUTISHAUER, H. [1959], *"Theory of gradient methods"*, Chapter 2 of "Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-adjoint Boundary Value Problems" by M.Engeli, T. Ginsburg, H. Rutishauer and E. Stiefel. Birkhäuser, Baser, 1959.

SHELDON, J. [1955], *"On the numerical solution of elliptic difference equations"*, Math.Tables Aids Comput. 9, 101-112.

SOUTHWELL, R.V. [1946], *"Relaxation methods in Theoretical Physics"*, Oxford University Press, New York.

STEIN, P. and ROSENBERG, R. [1948], *"On the solution of linear simultaneous equations by iteration"*, J.London Math.Soc. 93, 111-118.

STIEFEL, E. [1952], *"Über einige methoden der Relaxationrechnung"*, Z.angew.Math.Phys. 3, 1-33.

STRANG, G. [1972], *"Approximation in the finite element method"*, Numer.Math. 19, 81-98.

VARGA, R.S. [1957], *"A comparison of the successive overelaxation method and semi-iterative method using Chebyshev polynomials"*, J.SIAM 5, 39-46.

VARGA, R.S. [1959], *"Orderings of the successive overrelaxation scheme"*, Pacific J.Math. 9, 925-939.

VARGA, R.S. [1960], *"Factorisation and normalised iterative methods"*, in "Boundary Problems in Differential Equations", (R.E. Langer ed.), 121-142, University of Wisconsin, Madison, Wisconsin.

VARGA, R.S. [1962], *"Matrix Iterative Analysis"*, Prentice-Hall, Englewood Cliffs, New Jersey.

WACHSPRESS, E.L. [1963], *"Extended application of alternating direction implicit iteration to model problem theory"*, J.Soc.Indust.Appl.Math. 11, 994-1016.

WACHSPRESS, E.L. [1966], *"Iterative solution of elliptic systems and applications to the neutron diffusion equations of reactor physics"*, Prentice-Hall, Englewood Cliffs, New Jersey.

YOUNG, D.M. [1954], *"Iterative methods for solving partial difference equations of elliptic type"*, Trans.Amer.Math.Soc. 76, 92-111.

YOUNG, D.M. [1954a], *"On Richardson's method for solving linear systems with positive definite matrices"*, J.Math.Phys. XXXII, 243-255.

YOUNG, D.M. [1956], *"On the solution of linear systems by iteration"*, Proc.Sixth.Symp. in Appl.Math.Amer.Math.Soc. 6, 283-298. McGraw-Hill, New York.

YOUNG, D.M. [1971], *"Iterative solution of large linear systems"*, Academic Press, New York and London.

YOUNG, D.M. [1971a], *"A bound for the optimum relaxation factor for the successive overrelaxation method"*, Numer.Math. 16, 408-413.

YOUNG, D.M. [1974], *"On the accelerated SSOR method for solving large linear systems"*, Report CNA-92, Center for Numerical Analysis, University of Texas, Austin, Texas.

YOUNG, D.M. [1974a], *"Iterative solution of linear and non-linear systems derived from elliptic partial differential equations"*, Report CNA-93, Center for Numerical Analysis, University of Texas, Austin, Texas.

YOUNG, D.M. [1975], *"Notes on conjugate gradient method"*, (unpublished).

YOUNG, D.M. and GREGORY, R.T., [1973], *"A survey of numerical mathematics"*, Vol.II, Addison-Wiley, Reading, Massachusetts.

YOUNG, D.M. and KINCAID, D.R. [1969], *"Norms of the successive overrelaxation method and related methods"*, Report TNN-94, Computation Center, University of Texas, Austin, Texas.

YOUNG, D.M. and SHAW, H. [1955], *"Ordvac solutions of* $\partial^2 u/\partial x^2+\partial^2 u/\partial y^2+(k/y)\partial u/\partial y=0$ *for boundary value problems and problems of mixed type"*,
Interim.Tech.Report No.14, Office of Ordnance Research Contract DA-36-034-ORD-1486, University of Maryland, College Park, Maryland.

YOUNG, D.M. and WARLICK, C.H. [1953], *"On the use of Richardson's method for the numerical solution of Laplace's equation on the ORDVAC"*,
Ballistic Research Labs. Memorandum Report No.707, Aberdeen Proving Ground, Maryland.

ZAHRADNIK, R.L. [1971], *"Theory and techniques of optimisation for practising engineers"*,
Barnes and Noble, Inc., New York.

ZIENKIEWICZ, O.C. [1971], *"The finite element method in engineering science"*,
McGraw-Hill, London.

ZLAMAL, M. [1968], *"On the finite element method"*,
Numer.Math. 12, 394-409.

APPENDIX A

ARITHMETIC OPERATION COUNT

In this appendix, we compare the number of arithmetic operations required for various methods to solve the problem

$$\frac{\partial}{\partial x}(A\frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(C\frac{\partial U}{\partial y}) = 0. \tag{A.1}$$

Since there is not a great difference between the time required to perform product and summation operations on present-day computers, we will consider product (multiplication and division) as well as summation (addition and subtraction) processes equally.

We recall from (1-2.7) that the discretised form of (A.1) is

$$u(x,y) = \beta_1(x,y)u(x+h,y)+\beta_2(x,y)u(x,y+h)+\beta_3(x,y)u(x-h,y)+\beta_4(x,y)u(x,y-h) \tag{A.2}$$

where

$$\left.\begin{aligned} \beta_1(x,y) &= \frac{A(x+\frac{h}{2},y)}{S(x,y)} \quad , \quad \beta_2(x,y) = \frac{C(x,y+\frac{h}{2})}{S(x,y)} \quad , \\ \beta_3(x,y) &= \frac{A(x-\frac{h}{2},y)}{S(x,y)} \quad , \quad \beta_4(x,y) = \frac{C(x,y-\frac{h}{2})}{S(x,y)} \quad , \end{aligned}\right\} \tag{A.3}$$

and

$$S(x,y) = A(x+\tfrac{h}{2},y)+A(x-\tfrac{h}{2},y)+C(x,y+\tfrac{h}{2})+C(x,y-\tfrac{h}{2}) \quad . \tag{A.4}$$

We assume that the coefficients A and C for each mesh point are in storage and need only be computed once. As a first step we proceed to determine the number of operations necessary to compute one SOR iteration. For a particular point (x,y) we have the following SOR computation

$$\begin{aligned} u^{(n+1)}(x,y) = \omega[&\beta_3(x,y)u^{(n+1)}(x-h,y)+\beta_4(x,y)u^{(n+1)}(x,y-h) \\ &+\beta_1(x,y)u^{(n)}(x+h,y)+\beta_2(x,y)u^{(n)}(x,y+h)]+(1-\omega)u^{(n)}(x,y). \end{aligned} \tag{A.5}$$

In order to compute the β_i's in (A.5) we have that

4 divisions

and

3 additions

are required as can be seen from (A.3) and (A.4).

Next, for a single point by applying SOR we have from (A.5) that

6 multiplications
and
4 additions

are required not counting the subtraction to form $(1-\omega)$, since $(1-\omega)$ can be computed once and stored rather than recomputed for each point. If now $h=1/J$, then one full SOR iteration traverses $(J-1)\times(J-1)\sim J^2$ points and therefore requires $17J^2$ operations.

Similarly working we will attempt to determine the number of operations needed to complete one PSD iteration. We recall from (4-9.2) that one full PSD iteration for a particular point (x,y) can be computed as

$$\zeta^{(n+\frac{1}{2})}(x,y) = b/S(x,y)-u^{(n)}(x,y)+\beta_1(x,y)u^{(n)}(x+h,y)+\beta_2(x,y)u^{(n)}(x,y+h)$$
$$+\beta_3(x,y)u^{(n)}(x-h,y)+\beta_4(x,y)u^{(n)}(x,y-h)+\omega[\beta_3(x-h,y)\zeta^{(n+\frac{1}{2})}(x-h,y)$$
$$+\beta_4(x,y)\zeta^{(n+\frac{1}{2})}(x,y-h)] \qquad (A.6)$$

$$\zeta^{(n+1)}(x,y) = \zeta^{(n+\frac{1}{2})}(x,y) +\omega[\beta_1(x,y)\zeta^{(n+1)}(x+h,y)+\beta_2(x,y)\zeta^{(n+1)}(x,y+h)] \qquad (A.7)$$

and

$$u^{(n+1)}(x,y) = u^{(n)}(x,y)+\tau\zeta^{(n+1)}(x,y). \qquad (A.8)$$

In order to compute $\zeta^{(n+\frac{1}{2})}(x,y)$ for a single point we have from (A.6) that

1 subtraction
6 additions
and
1 division
7 multiplications

are required. To compute $S(x,y)$ and the β_i's it is required

4 divisions
and
3 additions.

Moreover, to compute $\zeta^{(n+1)}(x,y)$ we have from (A.7) that

2 additions
and
3 multiplications

are required, whereas for the computation of β_1 and β_2

3 additions

and

2 divisions

are needed. Finally, to compute $u^{(n+1)}$ we have from (A.8) that

1 addition

and

1 multiplication

is required. Thus one full PSD iteration of the form (A.6-A.8) requires $(15+7+5+5+2)J^2=34J^2$ operations which is exactly twice the number of operations of one SOR iteration and equal to the number of operations of one SSOR iteration. Referring to (5-5.14) after the PSD iteration has been completed

and 1 addition

2 multiplications

are still required to obtain a PJ-SI iteration. Thus $(34+3)J^2=37J^2$ operations are needed to complete one PJ-SI iteration.

Even though one PSD iteration requires twice the number of operations of an SOR iteration there is a way to reduce the computational work by providing storage space for an extra N-vector. This can be accomplished by following a technique which is due to Niethammer [1964].

First, let us consider the PSD method defined by

$$u^{(n+1)} = u^{(n)}+\tau(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}). \tag{A.9}$$

Up to now we have seen a computable form of PSD in terms of vector corrections (see (4-9.2)). Next, we will present another form of PSD in terms of intermediate vector approximations similar to (3-2.33), (3-2.34) of SSOR. Let us consider the iterative process

$$u^{(n+\frac{1}{2})} = (1-\tau)u^{(n)}+\omega Lu^{(n+\frac{1}{2})}+(\tau-\omega)Lu^{(n)}+\tau(Uu^{(n)}+c)$$

and

$$u^{(n+1)} = u^{(n+\frac{1}{2})}+\omega Uu^{(n+1)}-\omega Uu^{(n)}\ , \tag{A.10}$$

then we can readily see that it is an alternative form of the PSD method since by eliminating $u^{(n+\frac{1}{2})}$ in (A.10) we obtain (A.9).

Consequently, (A.10) is another form of PSD which does not use vector corrections as in (4-9.2) and is more familiar to us as this form is similar to the other point methods. By expressing the PSD method using

(A.10) we exploit the fact that it is not necessary to recompute $Uu^{(n)}$ in the second half iteration and therefore this vector can be saved at each half iteration.

From this observation we see that we can apply Niethammer's process in order to reduce the amount of work for each PSD iteration. This can be seen by explicitly exhibiting two full PSD iterations given by (A.10) as follows

$$\left.\begin{array}{l}
\left\{\begin{array}{l}
u^{(n+\frac{1}{2})} = (1-\tau)u^{(n)}+\omega Lu^{(n+\frac{1}{2})}+(\tau-\omega)Lu^{(n)}+\tau(Uu^{(n)}+c) \\
\qquad\qquad\qquad \text{save } Uu^{(n)} \\
u^{(n+1)} = u^{(n+\frac{1}{2})}+\omega Uu^{(n+1)}-\omega Uu^{(n)} \\
\qquad\qquad\qquad \text{save } Uu^{(n+1)}
\end{array}\right. \\
\left\{\begin{array}{l}
u^{(n+3/2)} = (1-\tau)u^{(n+1)}+\omega Lu^{(n+3/2)}+(\tau-\omega)Lu^{(n+1)}+\tau(Uu^{(n+1)}+c) \\
u^{(n+2)} = u^{(n+3/2)}+\omega Uu^{(n+2)}-\omega Uu^{(n+1)} \\
\qquad\qquad\qquad \text{save } Uu^{(n+2)}
\end{array}\right. \\
\quad \vdots
\end{array}\right\} \tag{A.11}$$

If we now consider one full PSD iteration without using the above reduction scheme, then from (A.10) we have

$$\begin{aligned}
u^{(n+\frac{1}{2})}(x,y) &= (1-\tau)u^{(n)}(x,y)+\omega[\beta_3(x,y)u^{(n+\frac{1}{2})}(x-h,y)+\beta_4(x,y)u^{(n+\frac{1}{2})}(x,y-h)] \\
&+(\tau-\omega)[\beta_3(x,y)u^{(n)}(x-h,y)+\beta_4(x,y)u^{(n)}(x,y-h)]+\tau[\beta_1(x,y)u^{(n)}(x+h,y) \\
&+\beta_2(x,y)u^{(n)}(x,y+h)]
\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}
u^{(n+1)}(x,y) &= u^{(n+\frac{1}{2})}(x,y)+\omega[\beta_1(x,y)u^{(n+1)}(x+h,y)+\beta_2(x,y)u^{(n+1)}(x,y+h)] \\
&-\omega[\beta_1(x,y)u^{(n)}(x+h,y)+\beta_2(x,y)u^{(n)}(x,y+h)].
\end{aligned} \tag{A.13}$$

Thus for a single point, from (A.12) we have

10 multiplications

and

6 additions

not counting the operations involved to form $(1-\tau)$ and $(\tau-\omega)$ since they can be computed once and stored.

In addition, for the computation of (A.13) it is required

6 multiplications

3 additions

and

1 subtraction.

Thus we immediately determine that one full PSD iteration of the form (A.10) requires $(16+7)+(10+5)J^2=38J^2$ operations. Finally $(38+3)J^2=41J^2$ operations are needed to complete one PJ-SI iteration.

As we have seen using Niethammer's scheme with PSD it is not necessary to recompute $Uu^{(n)}$ in the second half iteration. It can be easily seen that this is a saving of 8 operations. Thus, for one PSD iteration and applying Niethammer's process it is required $(38-8)J^2=30J^2$ operations compared to $17J^2$ operations for SOR. Admittedly, this is but a modest saving. However the advantage comes when more iterations are computed with Niethammer's approach. Let us consider the PSD iteration given by (A.10) then (see (A.11))

Computing $u^{(n+\frac{1}{2})}$ requires $23J^2$ operations; by storing $Uu^{(n)}$ we save 8 operations in the next half-iteration.

Computing $u^{(n+1)}$ requires $(15-8)J^2=7J^2$ operations; by storing $Uu^{(n+1)}$ we save 8 operations in the next half-iteration.

Computing $u^{(n+3/2)}$ requires $(23-8)J^2=15J^2$ operations.

Computing $u^{(n+2)}$ requires $7J^2$ operations; by storing $Uu^{(n+2)}$ we save 8 operations in the next half-iteration.

⋮

Therefore we see that each PSD iteration past the first requires just

$(15+7)J^2=22J^2$ operations, which is $5J^2$ more than necessary for an SOR iteration. However, the first PSD iteration always requires $(23+7)J^2=30J^2$ operations.

In Table A.1 we summarise the results obtained so far for h=1/J, where under the column headings A.6-A.8 we include the number of operations required for the computation of the PSD method given by (A.6), (A.7) and (A.8).

Method	Number of Operations for n Iterations		
	With Niethammer's Scheme	Without Niethammer's Scheme	A.6-A.8
SOR	-	$17nJ^2$	-
PSD	$(22(n-1)+30)J^2$	$38nJ^2$	$34nJ^2$
SSOR†	$(18(n-1)+26)J^2$	$34nJ^2$	-
PJ-SI	-	$41nJ^2$	$37nJ^2$
SSOR-SI	-	$39nJ^2$	-

TABLE A.1

OPERATION COUNT FOR SOR, PSD, SSOR, SSOR-SI AND PJ-SI WITH h=1/J

From Table A.1 we observe that by expressing PSD in the form (A.10) the amount of operations is increased (without Niethammer's scheme) as compared with the number of operations required by the form (A.6)-(A.8). However, this form of PSD method enables us to apply Niethammer's approach and reduce the computational effort such as to be competitive with the work involved in SOR. Unfortunately, this is not the case for the PJ-SI method. This can be seen if we attempt to write (5-5.8) in a similar form to (A.10). Indeed, if we write (5-5.8) in a two-level iterative form involving intermediate vector approximations, we may end up with the following iterative

†*See Benokraitis [1974].*

scheme

$$u^{(n+\frac{1}{2})} = \omega Lu^{(n+\frac{1}{2})} + (I-\omega L)\tilde{u}^{(n)} + \bar{\rho}\rho_{n+1}[Lu^{(n)} + Uu^{(n)} - u^{(n)} + c]$$

and (A.14)

$$u^{(n+1)} = u^{(n+\frac{1}{2})} + \omega Uu^{(n+1)} - \omega\rho_{n+1}Uu^{(n)} - \omega(1-\rho_{n+1})Uu^{(n-1)}$$

where

$$\tilde{u}^{(n)} = u^{(n-1)} + \rho_{n+1}(u^{(n)} - u^{(n-1)}). \tag{A.15}$$

If we eliminate $u^{(n+\frac{1}{2})}$ in (A.14), we can readily see that (5-5.8) is obtained. On the other hand, the amount of computational work has now been increased considerably and even though we can apply Niethammer's scheme the number of operations for n iterations is greater than $41nJ^2$. Thus it is preferable to use (5-5.14) combined with (4-9.2) and (4-9.3) for the computation of the PJ-SI method rather than using (A.14)-(A.15). This difficulty is expected to be present for the SD-PJ and PJ-CG iterative procedures since they possess similar form with the PJ-SI method. However, the advantage of the Neithammer's scheme can be exploited in the PJ-VE method since its form is similar to the PSD method (see (5-6.4)). We recall from Section 5.6 that the PJ-VE has been defined by

$$u^{(n+1)} = u^{(n)} + \theta_{n+1}(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}(b-Au^{(n)}) \tag{A.16}$$

which can be written alternatively as

$$u^{(n+\frac{1}{2})} = (1-\theta_{n+1})u^{(n)} + \omega Lu^{(n+\frac{1}{2})} + (\theta_{n+1}-\omega)Lu^{(n)} + \theta_{n+1}(Uu^{(n)} + c)$$

and (A.17)

$$u^{(n+1)} = u^{(n+\frac{1}{2})} + \omega Uu^{(n+1)} - \omega Uu^{(n)}.$$

Consequently, it can be easily verified that if we apply Niethammer's scheme to (A.17), then the number of operations of the PJ-VE procedure is identical with the number of operations in the PSD method (see Table A.1).

APPENDIX B

DETERMINATION OF A BOUND ON S(LU)

In this appendix we show how one can determine the bound $\overline{\beta}$ on S(LU).

By Theorem 2-3.1 we have that

$$S(LU) \leq \|LU\|_\infty \tag{B.1}$$

so we seek to determine the quantity $\|LU\|_\infty$.

We note that equation (A.2) corresponds to the following computational stencil given in Figure B.1

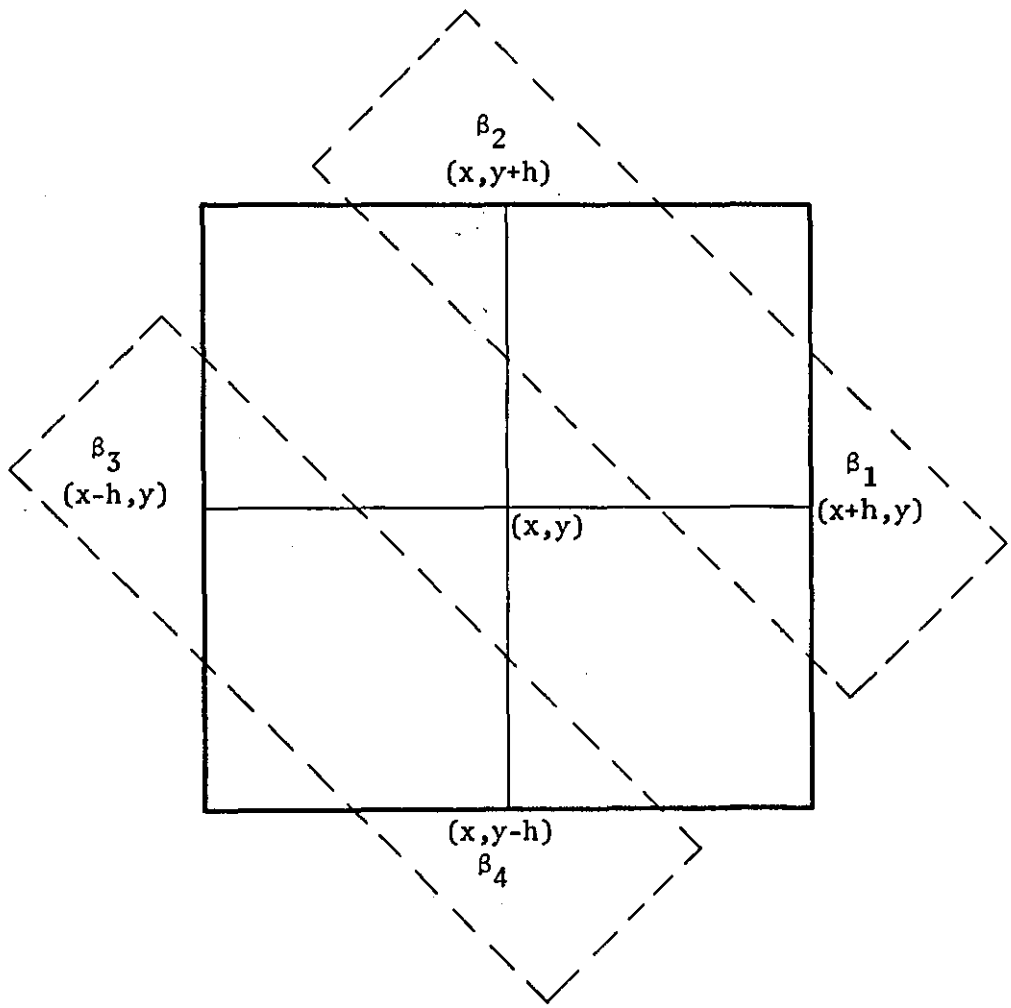


FIGURE B.1

The part of the stencil enclosed by dotted lines corresponds to the operators L and U. In order to see how LU operates on the function u(x,y) we consider two stages

$$v(x,y) = Uu(x,y) \tag{B.2}$$

and

$$w(x,y) = Lv(x,y) = LUu(x,y). \tag{B.3}$$

From (B.2) and Figure B.1 we have

$$v(x,y) = \beta_1(x,y)u(x+h,y)+\beta_2(x,y)u(x,y+h). \tag{B.4}$$

But from (B.3) and (B.4) we obtain successively the result

$$
\begin{aligned}
w(x,y) &= \beta_3(x,y)v(x-h,y)+\beta_4(x,y)v(x,y-h) \\
&= \beta_3(x,y)[\beta_1(x-h,y)u(x,y)+\beta_2(x-h,y)u(x-h,y+h)] \\
&+ \beta_4(x,y)[\beta_1(x,y-h)u(x+h,y-h)+\beta_2(x,y-h)u(x,y)] \\
&= [\beta_3(x,y)\beta_1(x-h,y)+\beta_4(x,y)\beta_2(x,y-h)]u(x,y) \\
&+ [\beta_3(x,y)\beta_2(x-h,y)]u(x-h,y+h)+[\beta_4(x,y)\beta_1(x,y-h)]u(x+h,y-h) \\
&= \gamma_0 u(x,y)+\gamma_1 u(x-h,y+h)+\gamma_2 u(x+h,y-h). \qquad \text{(B.5)}
\end{aligned}
$$

Therefore, the operational stencil for LU can be represented by Figure B.2 illustrating that LU operates only on values of u(x,y) at the diagonal points (x,y), (x-h,y+h) and (x+h,y-h).

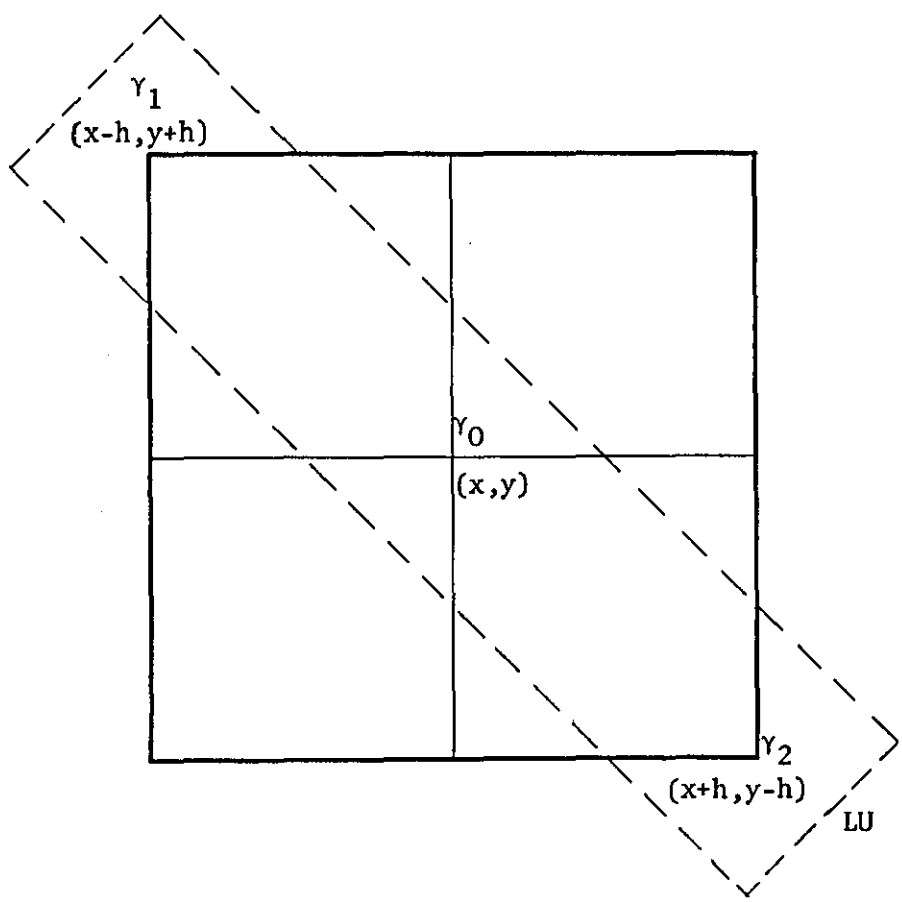


FIGURE B.2

Finally by (B.1),(B.5) and (2-3.7) we obtain the required bound $\bar{\beta}$ as follows

$$S(LU) \leq \|LU\|_\infty = \max_{(x,y)\in R_h} (\gamma_0+\gamma_1+\gamma_2)$$

$$= \max_{(x,y)\in R_h} \{\beta_3(x,y)[\beta_1(x-h,y)+\beta_2(x-h,y)]$$

$$+\beta_4(x,y)[\beta_2(x,y-h)+\beta_1(x,y-h)]\} = \bar{\beta}. \qquad (B.6)$$

APPENDIX C

CHEBYSHEV MINIMAX THEOREM

In this appendix we present an important theorem which concerns the Chebyshev polynomials.

Theorem C.1 (Markhoff [1916], Flanders and Shortley [1950])

Let $P_n(G)$ be a real polynomial of degree n in the matrix G such that the set of all eigenvalues λ of G satisfy the inequality

$$a < \lambda < b < 1. \tag{C.1}$$

Moreover, for each $n \geq 0$, let S_n be the set of all real polynomials $Q_n(\lambda)$ of degree n such that $Q_n(1)=1$. Then the polynomials $P_n(\lambda) \in S_n$ which minimises the quantity

$$\max_{a \leq \lambda \leq b} |P_n(\lambda)| \tag{C.2}$$

is unique and is given in terms of Chebyshev polynomials by

$$P_n(\lambda) = \frac{T_n\left(\frac{2\lambda-(b+a)}{b-a}\right)}{T_n\left(\frac{b+a}{b-a}\right)} \tag{C.3}$$

where $T_n(x)$ is the Chebyshev polynomial of degree n given by

$$T_n(x) = \begin{cases} \cos(n\cos^{-1}x) \\ \cosh(n\cosh^{-1}x) \\ \frac{1}{2}[(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n}], \ n \geq 0. \end{cases} \tag{C.4}$$

APPENDIX D

UNIMODALITY OF THE FUNCTION $P(\omega)$

Definition D.1

A function f(x) is said to be unimodal on [a,b] if it decreases monotonically to its minimum, after which it increases monotonically (Zahradnik [1971]).

Let us now consider the function

$$P(\omega) = \max_i \left\{ \frac{1-\omega a_i+\omega^2\beta_i}{\omega(2-\omega)(1-a_i)} \right\} = \max_i \{p(\omega,v^{(i)})\} \tag{D.1}$$

where

$$a_i = \frac{(v^{(i)},DBv^{(i)})}{(v^{(i)},Dv^{(i)})}, \qquad \beta_i = \frac{(v^{(i)},DLUv^{(i)})}{(v^{(i)},Dv^{(i)})}, \tag{D.2}$$

for the given vector $v^{(i)}$ and the pair (a_i,β_i).

We seek to show that $P(\omega)$ is unimodal, that is, according to Definition D.1, $P(\omega)$ decreases monotonically to its minimum $P(\omega_0)$, after which it increases monotonically.

From (D.1) we have that

$$\operatorname{sign}\left[\frac{\partial}{\partial\omega}p(\omega,v^{(i)})\right] = \operatorname{sign}(\omega^2(2\beta_i-a_i)-2(1-\omega)) \tag{D.3}$$

where $a_i<1$. If we let ω_i denote the value of $\omega\in(0,2)$ such that

$$\omega_i^2(2\beta_i-a_i)-2(1-\omega_i) = 0, \tag{D.4}$$

then we see that as ω varies from 0 to ω_i, $p(\omega,v^{(i)})$ decreases until $\omega=\omega_i$ and then increases as ω varies from ω_i to 2. Thus each $p(\omega,v^{(i)})$ is unimodal for $\omega\in(0,2)$. It remains to show that $P(\omega)$, defined by (D.1), first decreases when $\omega<\omega_0$ and then increases when $\omega>\omega_0$.

Let $P(\hat{\omega}_0)$ be a relative minimum of $P(\omega)$, where $\hat{\omega}_0\in(0,2)$, then at least one curve $p(\omega,v^{(i)})$ which passes through the point $(\hat{\omega}_0,P(\hat{\omega}_0))$ must not decrease for $\omega>\hat{\omega}_0$, otherwise $P(\hat{\omega}_0)$ is not a relative minimum. The curve $p(\omega,v^{(i)})$ increases as ω varies from $\hat{\omega}_0$ to 2. If $p(\omega,v^{(i)})$ is maximum in the interval $[\hat{\omega}_0,2)$, then $P(\omega)$ increases for $\omega>\hat{\omega}_0$. If $p(\omega,v^{(i)})$ is not maximum in the internal $[\hat{\omega}_0,2)$, then there are other functions

$p(\omega,v^{(i)})$ which pass through $(\hat{\omega}_0,P(\hat{\omega}_0))$ and are increasing in $[\hat{\omega}_0,2)$. In any case, $P(\omega)$ is increasing for $\omega>\hat{\omega}_0$.

Similarly, there exists at least one function $p(\omega,v^{(k)})$ which passes through $(\hat{\omega}_0,P(\hat{\omega}_0))$ and is decreasing in $(0,\hat{\omega}_0]$. The curve $p(\omega,v^{(k)})$ decreases as ω varies from 0 until $\omega=\hat{\omega}_0$. If $p(\omega,v^{(k)})$ is maximum in the interval $(0,\hat{\omega}_0]$, then $P(\omega)$ decreases in $(0,\hat{\omega}_0]$. If $p(\omega,v^{(k)})$ is not maximum in $(0,\hat{\omega}_0]$, then there are other functions $p(\omega,v^{(\ell)})$ which pass through $(\hat{\omega}_0,P(\hat{\omega}_0))$ and are decreasing in $(0,\hat{\omega}_0]$. Thus $P(\omega)$ is decreasing in the interval $(0,\hat{\omega}_0]$ and the relative minimum $P(\hat{\omega}_0)=P(\omega_0)$ is an absolute minimum. We therefore conclude that $P(\omega)$ is unimodal.