


This item was submitted to Loughborough University as an MPhil thesis by the author and is made available in the Institutional Repository (<https://dspace.lboro.ac.uk/>) under the following Creative Commons Licence conditions.




creative
commons
C O M M O N S D E E D


Attribution-NonCommercial-NoDerivs 2.5


You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

 **BY:** **Attribution.** You must attribute the work in the manner specified by the author or licensor.


 **Noncommercial.** You may not use this work for commercial purposes.

 **No Derivative Works.** You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#) 

For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

LOUGHBOROUGH
UNIVERSITY OF TECHNOLOGY
LIBRARY

AUTHOR/FILING TITLE

DEMETRIOU, I

ACCESSION/COPY NO

120110/02

VOL. NO.

CLASS MARK

F

-i. 1003

LOAN COPY

F

012 0110 02



GENERALIZED PRECONDITIONING STRATEGIES

BY

IOANNIS CONSTANTINE DEMETRIOU, Diploma Mathematics

A Master's Thesis

Submitted in Partial Fulfilment of the Requirements

for the Award of Master of Philosophy

of the Loughborough University of Technology

September, 1980

Supervisor: PROFESSOR D.J. EVANS, Ph.D., D.Sc.

Department of Computer Studies.

Loughborough University of Technology Library	
DATE	Mar 81
Class	
Acc. No	120110/02

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.

I.C. DEMETRIOU

To my parents,

Constantinos and Catherin Demetriou

CONTENTS

	<u>PAGE</u>
PREFACE	1
ACKNOWLEDGEMENTS	vii
<u>Chapter 1:</u> INTRODUCTION	1
<u>Chapter 2:</u> BASIC ITERATIVE METHODS	
2.1 On the Iterative Method	7
2.2 On Convergence	9
2.3 On the Rate of Convergence	11
2.4 A Test Method	12
2.5 Basic Iterative Methods	14
2.6 Convergence Properties	17
2.7 On the Notion of Preconditioning	23
<u>Chapter 3:</u> THE PRECONDITIONING BY DIRECT FACTORIZATION METHOD	27
3.1 On a Conditioning Matrix	28
3.2 Generation of the Preconditioning Scheme	30
3.3 An Heuristic Method for the Matrix β	31
3.4 On the Eigenvalues of the Matrix $\tilde{U}L=UL+\beta\beta^T$	33
3.5 The Preconditioning by Direct Factorization Method (PDF Method)	35
3.6 Interval Estimation on the Bound of $P(B_\omega)$	46
3.7 On the Rate of Convergence	50
3.8 A Note on the Dirichlet Problem for the One-Dimensional Poisson Equation	52
3.9 A Certain Conditioning Matrix When A Possesses Property A	54

	<u>PAGE</u>
<u>Chapter 4:</u> EXPERIMENTAL RESULTS	61
4.1 Optima Parameters	62
4.2 Estimated Parameters	69
<u>Chapter 5:</u> THE PRECONDITIONING BY DIRECT FACTORIZATION, SEMI-ITERATIVE METHOD	82
5.1 The PDF-SI Method	83
5.2 Numerical Results	86
EPILOGUE	94
Appendix A: MATRIX THEORY PRELIMINARIES	97
Appendix B: PROOF OF THEOREM (2.7.12)	100
Appendix C: ARITHMETIC OPERATION COUNT	105
Appendix D: A COMPUTER PROGRAM	111
REFERENCES	114

Ἀρχιμήδης ... ἀπὸ τοῦ τῆς ἐπιστήμης τῶν μηχανῶν

PREFACE

*"And what is actual is actual only for one time
and only for one place"*

Ash-Wednesday, 1930

T.S. ELLIOT

Over the past decade Professor David J. Evans [1968] has suggested the use of "Preconditioning" in iterative methods for solving large, sparse systems of linear equations, which arise from the finite difference approximations to the partial differential equations. Since then, certain aspects on preconditioning have appeared in the literature and a whole new theory constructed. The versatility of the preconditioning concept is shown by the stimulating exploration of new numerical algorithms and methods of their realization.

The aim of this thesis is to emphasise in the theory we use and develop together with the practice we state. This study led to a new form of preconditioning, which has not yet appeared in the literature. Specifically, we consider the conditioning matrix factorized into two rectangular matrices⁽¹⁾, so as to develop a new preconditioned iterative method and its related properties as well. It requires the selection of two parameters to be applied, a preconditioning parameter at its optimal value and an acceleration parameter in such a fashion that a simultaneous

⁽¹⁾*To elucidate the difference between this conditioning matrix and the related ones used in other preconditioned schemes, we note that the up to date preconditioning techniques used a factorization into square matrices.*

displacement method is applicable.

Further, the method is accelerated by the classical semi-iterative technique. Our first aim is to develop the theoretical foundation of the new preconditioning concept. Our second aim is to present sufficient numerical details in the practical application of the theory to the numerical solution of certain elliptic partial differential equations.

Chapter 1: This is an introductory chapter stating the origin of the problem. We consider the discrete generalized Dirichlet problem obtained by applying the five-point difference approximation to a continuous generalized Dirichlet problem.

Chapter 2: Some basic iterative methods are introduced in a test of comparison from the Jacobi method to the Symmetric Successive Overrelaxation and the Preconditioned Simultaneous Displacement. Theorems on convergence are introduced and the rate of convergence is also defined for all methods included in the chapter. Furthermore, we give some attention in the Preconditioned Simultaneous Displacement method as it has been developed by Evans and others. However, we emphasise more to the preconditioning concept itself, since that is the original source of our study.

Chapter 3: With the goal of having all the theoretical analysis of our method self-contained we have devoted Chapter 3 to the development of the new preconditioning scheme we recommend, namely the Preconditioning by Direct Factorization method (PDF method).

A new conditioning matrix M is introduced which, in a sense, is close to the original matrix A . Matrix M is factorized into two rectangular matrices, i.e.,

$$M = \begin{bmatrix} A \\ U, \beta \end{bmatrix} \begin{bmatrix} A \\ U, \beta \end{bmatrix}^T$$

where the main square part A_U and A_U^T of each factor of M , arises by applying a backward and a forward process on the net, respectively. The semantic difference with the previous conditioning schemes is the existence of the matrix β , which is constructed by a simple heuristic method which is applied under the consideration that the matrix M has to be written as a sum of matrices having at least one non-zero element in any of their columns and rows. An analysis is then performed on the interaction of the heuristic part of M , to its eigenvalues, in order to have sufficient conditions in the latter seeking a bound for the spectral radius of the iteration matrix (or equivalently on the P-condition number of the preconditioned matrix) and hence the rate of convergence of the method.

The method can be seen as a fractional-step method which has three steps, a backward, a forward and a direct step of a simple Gaussian type elimination process. Following a technique due to Habetler and Wachpress [1961] we represent the eigenvalues of the preconditioned matrix $B_\omega = M^{-1}A$ in terms of certain inner products. A bound on the smallest eigenvalue of B_ω is then given concerning these inner products. By Wilkinson [1965] we have sufficient conditions in seeking bounds for the largest eigenvalue of the aforementioned matrix. The preconditioning parameter ω is chosen on the basis of a-priori information about the spectra of the operator involved in the algorithm. Moreover, we state a necessary and sufficient condition for the convergence of the method, in Theorem (3.5.54).

Next, in Theorem (3.6.4) we are concerned with the determination of the estimated parameters ω_1 and $P(B_{\omega_1})$ using a theorem given by Young [1977], suitably modified by Evans and Missirlis [1980]. Thus, Theorem (3.6.4) provides a theoretical foundation for the estimated parameters we use in a later chapter for our method, whereas an a-priori evaluation for the

rate of convergence is established. In Section 3.7 we see that the rate of convergence depends upon the bound on $P(B_\omega)$, by Theorem (3.6.4), and asymptotic results are given comparing our method with the SOR, SSOR and PSD method. Our theoretical expectations will be verified for the problems presented in Chapter 4.

In Section 3.8 we describe the simplest investigation of our method for the one dimensional Poisson equation.

Finally, a particular case is investigated when the original matrix is point two-cyclic (possesses Property A) and where a certain pre-conditioning scheme is used.

Chapter 4: We present in this chapter sufficient numerical details in the practical application of the theory developed in Chapter 3 to the numerical solution of six standard problems. The numerical results obtained with the optima and estimated parameters indicate that by applying the Preconditioning by Direct Factorization method, the number of iterations required for convergence varies approximately as h^{-1} , where h is the net mesh size.

The optima values of our method were found by a golden section search and the estimated values obtained by applying Theorem (3.6.4). Table (4.1.T3) portrays the results obtained from applying the PDF method with optima parameters. The number of iterations of the PDF method with optima parameters is almost the same number as the iterations required by the PSD method with optima parameters also. However, the treatment for obtaining the parameters requires the same number of iterations necessary for solving a problem itself, or even more.

In Table (4.2.T2) we can see that the SOR method requires a number of iterations which varies between 20-150% more than the required number of the PDF method with estimated parameters. Table (4.2.T2) again indicates

that the Symmetric SOR (SSOR) method requires 25-150% more iterations asymptotically as compared to the PDF method. Also, the PSD method required 8-115% more iterations than the PDF method, with all methods but the SOR considered with estimated parameters. It should be noticed that the PDF method requires about 30-40% more work per iteration, than the tested methods.

Chapter 5: In Chapter 5 the parameters involved are chosen at each step in such a way that the error vector approaches zero uniformly from the initial approximations as fast as possible. That is the acceleration of the PDF method by semi-iteration, which yields the PDF-SI method.

Acceleration by semi-iteration is possible since the eigenvalues of the iteration matrix of the PDF method are real and bounded in a certain region of the real axis. The parameters involved in the SI algorithm are again $P(B_\omega)$ together with a parametric set obtained by Chebyshev analysis.

Numerical results obtained in that chapter indicated that by applying the PDF-SI method with optima and estimated parameters to the six problems of Chapter 2, required an $O(h^{-\frac{1}{2}})$ number of iterations. The $O(h^{\frac{1}{2}})$ convergence was obtained even in cases with certain discontinuities amongst the coefficients of the initial differential equation. It is noticed here that the result of the proposition (4.2.8) of Chapter 4 establishes the improvement of an order of magnitude, of the PDF-SI method over PDF method. In Tables (5.2.T1) and (5.2.T2) we present the number of iterations required to solve the six problems of Chapter 4 by the PDF-SI method.

Appendices A-C: This part of the Thesis includes details which have not been covered in the main part since they are either considered well-known or trivial.

Appendix A: Appendix A includes preliminary results on matrix theory.

Appendix B: In this appendix we cite a theorem for a bound on the P-condition number of the Preconditioned Simultaneous Displacement method (Evans and Missirlis [1980]), by applying the conditioning matrix with reverse order than given in Evans and Missirlis [1980].

Appendix C: A detailed analysis is given for the arithmetic operations necessary to execute our algorithm by the Niethammer's [1964] scheme, in order to have a saving of operations. A saving of 20 percent is realized when Niethammer's scheme is applied to the PDF method. An operations count is given when we use the vector correction process for our algorithm.

Notation and Terminology: As a guide to the reader we state a word on references: "Theorem 3.5.54" refers to the Theorem 3.5.54 of Section 5 of Chapter 3. References are given in the form "Evans [1968]" which refer to a paper (or book) by Evans published in 1968.

Finally we mention that all the matrices used are real, except if otherwise mentioned. $\mu(A)$ denotes an eigenvalue of the matrix A , $M(A)$ denotes the eigenvalue of maximum algebraic value, $m(A)$ denotes the eigenvalue of minimum algebraic value, whereas the spectral radius of A is denoted by $S(A)$. The usual terminology in the text, defines A as a positive definite matrix to be the symmetric matrix A where $\langle x, Ax \rangle > 0$ for all $x \neq 0$. In addition, the definition implies that A is non-singular, has positive diagonal elements, has eigenvalues real and positive and has a complete system of eigenvectors.

ACKNOWLEDGEMENTS

Μέμνησο τοῦ ... Σόλωνος

I am grateful to Professor David J. Evans, who suggested and supervised this research.

I also wish to thank the staff of the Computer Centre at the University of Technology, Loughborough.

Sincere thanks are due to Miss Judith Briers, who, with much attention, typed the manuscript.

For continued encouragement and support, I thank my parents.

CHAPTER 1

INTRODUCTION

It is no paradox to say that in our most theoretical words we may be nearest to our most practical applications.

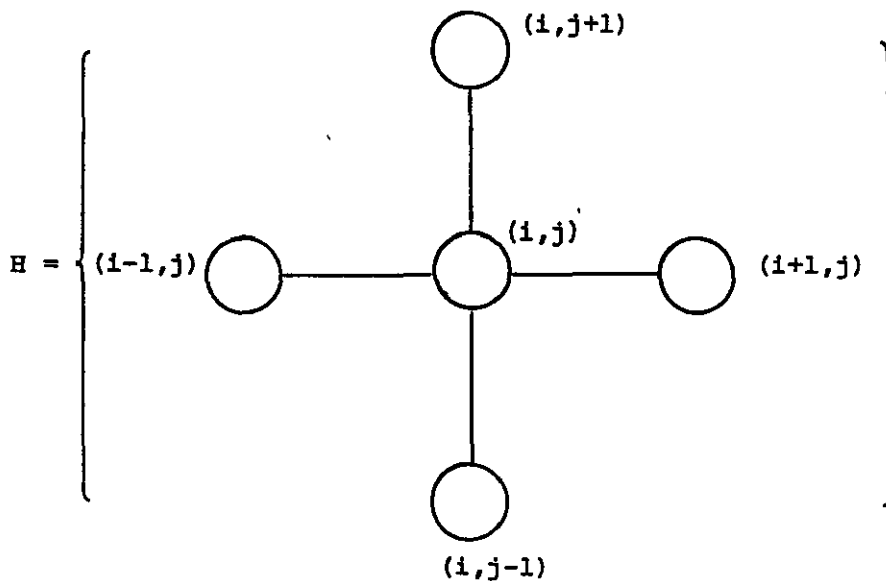
A.N. WHITEHEAD

Linear partial differential equations with given boundary conditions play a particularly important role in problems of applied mathematics. The numerical solution of such problems is in close relation to the solution of large scale systems of ordinary linear algebraic equations, characterized by a sparse matrix. The partial derivatives of a function can be conceived as limits of difference coefficients. The given partial differential equation can thus be replaced by a finite difference equation with an error which can be made as small as we wish at every point of the given domain. This does not guarantee that the solution of the algebraic system thus obtained will of necessity converge to the exact solution of the originally given partial differential equation, since the local errors committed may accumulate, resulting in a finite error.

However, since the development of large-scale computers have formed a basis for algorithmic constructions and extensive mathematical experiments on large scale systems, it seems desirable to exploit the close analogy between linear differential equations and linear algebraic equations, to the maximum degree.

Almost any method for solving a partial differential equation numerically reduces ultimately to the computations of discrete data. Moreover, tabulation of the solution over a finite lattice of mesh points appears as perhaps the most natural way in which to describe it in terms of numbers. Thus, a convenient and quite general procedure for calculating numerical solutions of boundary value problems is to approximate them by corresponding problems for finite difference equations.

In this investigation the partial differential equation with two independent variables is first transformed into a difference equation with the help of the "stencil":



Consider the general linear partial differential equation

$$L_{\sigma}[u] = \sum_{j,k=1}^n a_{jk} \frac{\partial^2 u}{\partial x_j \partial x_k} + \sum_{j=1}^n b_j \frac{\partial u}{\partial x_j} + cu = 0, \quad (1.1)$$

where the coefficients a_{jk} , b_j and c are suitably differentiable functions of the independent variables x_1, x_2, \dots, x_n and where it may be assumed that

$$a_{jk} = a_{kj} . \quad (1.2)$$

We shall suppose that (1.1) is of *elliptic type* in some n-dimensional region D, which means that the quadratic form

$$Q = \sum_{j,k=1}^n a_{jk} \lambda_j \lambda_k \quad (1.3)$$

is positive definite there, provided the signs are adjusted to make $a_{11} > 0$.

The *first boundary value problem* or the *Dirichlet problem* demands a solution u of (1.1) in D which takes as prescribed values

$$u = f \quad (1.4)$$

on the boundary ∂D of that region. Under various hypotheses about the geometry of ∂D and the behaviour of the coefficients a_{jk} , b_j and c , the most essential of which states that $c \leq 0$, it is possible to establish the existence, uniqueness and continuous dependence on boundary data of the solution of Dirichlet's problem.⁽¹⁾

We develop briefly the discretization of the self-adjoint differential equation

$$L[u] = \frac{\partial}{\partial x} \left(A \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(C \frac{\partial u}{\partial y} \right) + Fu = G , \quad (1.5)$$

involving Dirichlet boundary conditions.

In the square $D = \{D: 0 \leq x \leq 1, 0 \leq y \leq 1\}$ we seek a solution of the equation (1.5), satisfying the boundary conditions

$$u = g(x,y) , \quad (1.6)$$

where $A(x,y) > 0$, $C(x,y) > 0$ and $F(x,y) \leq 0$ in $D + \partial D$ and where the continuous

(1)

Even for the Laplace equation

$$\Delta u = \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} = 0$$

which has constant coefficients, it is not an easy matter to solve the Dirichlet problem in a region D of arbitrary shape. The solution can be obtained in closed form only for special choices of D, such as a sphere or a cube.

function $g(x,y)$ is defined on ∂D .

We now construct a difference approximation of our problem. Let D_h be the totality of points

$$x_i = ih, \quad y_j = jh \quad (1.7)$$

where $h = \frac{1}{N}$, $1 \leq i \leq N-1$, $1 \leq j \leq N-1$, $i, j, N \in \mathbb{N}$. Denote the points of intersection (x_i, y_j) by (x, y) and call them the mesh points of the lattice.

The positive number h is known as the mesh size of the lattice.

The set of mesh points for which one of $(i \pm 1, j)$, or $(i, j \pm 1)$ is not in D_h is denoted by ∂D_h .

For $u(x, y)$ for which $(x, y) \in D_h$, we get a system of finite difference equations

$$\begin{aligned} L_h[u] &= \frac{1}{h^2} \{ A(x+\frac{h}{2}, y) [u(x+h, y) - u(x, y)] - A(x-\frac{h}{2}, y) [u(x, y) - u(x-h, y)] \\ &\quad + C(x, y+\frac{h}{2}) [u(x, y+h) - u(x, y-h)] + C(x, y-\frac{h}{2}) [u(x, y) - u(x, y-h)] \} + Fu(x, y) \\ &= G(x, y). \end{aligned} \quad (1.8)$$

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

$$\begin{aligned} u(x, y) &= b_1(x, y)u(x+h, y) + b_2(x, y)u(x, y+h) \\ &\quad + b_3(x, y)u(x-h, y) + b_4(x, y)u(x, y-h) + \tau(x, y) \end{aligned} \quad (1.9)$$

where

$$\begin{aligned} b_1(x, y) &= \frac{A(x+\frac{h}{2}, y)}{S(x, y)}, & b_2(x, y) &= \frac{C(x, y+\frac{h}{2})}{S(x, y)} \\ b_3(x, y) &= \frac{A(x-\frac{h}{2}, y)}{S(x, y)}, & b_4(x, y) &= \frac{C(x, y-\frac{h}{2})}{S(x, y)} \end{aligned} \quad (1.10)$$

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

and where

$$S(x, y) = A(x+\frac{h}{2}, y) + A(x-\frac{h}{2}, y) + C(x, y+\frac{h}{2}) + C(x, y-\frac{h}{2}) - h^2 F(x, y). \quad (1.11)$$

The difference equation (1.9) contains the values of the solution at five lattice points and the approximation is of order $\theta(2h^2)$.

With each mesh point of the lattice we associated a specific difference equation (1.9). Therefore what we actually have to find is the solution of a system of $(N-1)(N-1)$ simultaneous linear equations (1.9) in $(N-1)(N-1)$ unknowns, where $(N-1)(N-1)$ stands for the total number of interior mesh points of the lattice exhausting D_h . Thus we obtain the matrix equation

$$Au = b \quad (1.12)$$

where the inhomogeneous term is derived from the boundary values we assign to $u(x,y)$.

The matrix A is real and symmetric. The difference operator can be chosen so that the matrix A is diagonally dominant and positive definite with the elements of $\text{diag}(A)$ positive and every other element of A non-positive.

Since ⁽¹⁾ any system of simultaneous linear equations can either be solved in a unique way for every choice of the inhomogeneous terms or it must possess non-trivial solutions in the homogeneous case, we can prove that a solution of the discrete Dirichlet problem always exists merely by verifying that when the data are made to vanish, the only answer available is the obvious one $u(x,y)=0$. On the other hand, an appeal to the maximum principle ⁽²⁾ assures us of the validity of the uniqueness assertion. The numerical answer has to converge toward the exact solution if the mesh size h approaches zero while at the same time the number l of decimal places retained in the calculations goes fast enough to infinity. As far as the more specific dependent of the

⁽¹⁾ *Fredholm alternative.*

⁽²⁾ *The maximum principle, states that the solution of certain partial differential equations of the elliptic type never achieve a strong relative maximum or minimum in the interior of their domain of definition.*

error on h and l is concerned, it corresponds directly to the kind of inhomogeneous terms that appear in (1.9). For a fixed choice of l the expected size of the round-off error is proportional to the number $(N-1)(N-1)$ of difference equations.

CHAPTER 2

BASIC ITERATIVE METHODS

*There are nine and sixty ways of constructing
tribal lays, and every single one of them is
right.* W

RUDYARD KIPLING

2.1 ON THE ITERATIVE METHOD

Many boundary value problems for partial differential equations can only be solved practically with the aid of difference methods. Whenever high precision is required, we are led to large systems of equations, often with thousands of unknowns. These systems of equations can be solved practically only by iterative methods. A well-studied topic is linear boundary value problems that occur with elliptic difference equations, where large systems of linear equations consequently result. Quite often in these systems, the terms on the principal diagonal "dominate" and an iterative method is recommended.

Replacing the continuous problem by an associated discrete one, may lead to a linear system

$$Au = b , \tag{2.1.1}$$

where A is a square (often sparse) matrix, b is a known and u is the unknown vector.

Methods of solutions for a general computational problem fall into the direct and iterative procedures. Iterative methods can be

programmed to take advantage of the zero elements in A . In elliptic partial differential equations the matrix resulting from the equation may be very large and sparse depending conversely on the mesh size of the lattice.

To solve the non-singular equation (2.1.1) by iteration we require a sequence $u^{(n)}$ so defined that

$$u^{(n)} \rightarrow A^{-1}b \text{ as } n \rightarrow \infty,$$

where $A^{-1}b$ is the exact solution. If $u^{(n)}$ is a function of $A, b, u^{(n-1)}, \dots, u^{(n-s)}$ we say that s is the degree of the iteration. In the case $s=1$, we could write

$$u^{(n)} = F(A, b, u^{(n-1)}) , \quad n=1, 2, \dots .$$

If F is independent of n , the iteration is said to be stationary and if F is linear in $u^{(n-1)}$, the iteration is termed linear.

The most general linear iteration is

$$u^{(n+1)} = Hu^{(n)} + k , \quad n=0, 1, \dots$$

where H is a matrix depending upon A and b and k a vector.

For a non-singular A we obtain, as a consistency condition, between (2.1.1) and (2.1.3),

$$k = (I-H)A^{-1}b .$$

The non-singularity, if it exists, of $I-H$ implies a reciprocal consistency condition

$$b = A(I-H)^{-1}k .$$

If both of the two consistency conditions are valid then the method is completely consistent, which means that the only solution of (2.1.3) is the solution $A^{-1}b$ of (2.1.1).

2.2 ON CONVERGENCE

The matrix H is called the iteration matrix for (2.1.3) and it is easy to see that if we split⁽¹⁾ A into

$$A = M - N, \quad M \text{ non-singular} \quad (2.2.1)$$

then for $H = M^{-1}N$ and $k = M^{-1}b$, $u = Hu + k$ if and only if $Au = b$.

If we subtract $u = Hu + k$ from (2.1.3), we obtain the error equation

$$u^{(n+1)} - u = H(u^{(n)} - u) = \dots = H^{n-1}(u^{(0)} - u). \quad (2.2.2)$$

Hence the sequence $u^{(0)}, u^{(1)}, \dots, u^{(n)}, \dots$ converges to u for each $u^{(0)}$ if and only if $\lim_{n \rightarrow \infty} H^n = 0$, that is, if and only if $S(H) < 1$,

by considering the Jordan form for H . We have, thus, proved the basic convergence lemma for (2.1.3).

Lemma (2.2.3): Let $A = M - N$ with A and M non-singular. Then for $H = M^{-1}N$ and $k = M^{-1}b$, the iterative method (2.1.3) converges to the solution $u = A^{-1}b$ of (2.1.1) for each $u^{(0)}$ if and only if $S(H) < 1$.

When the matrix A of (2.2.1) is symmetric⁽²⁾ the following theorem holds.

Theorem (2.2.4): If A is symmetric⁽²⁾ and A and $M^T + N$ are positive definite on some eigenset of H , then $S(H) < 1$. Conversely, if $\langle x, Mx \rangle > 0$ for all x in some eigenset E of H and $S(H) < 1$, then A and $M^T + N$ are positive definite on E .

As a result we have the following useful theorem:

(1) The splitting $A = M - N$ with A and M non-singular is called a regular splitting if $M^{-1} \geq 0$ and $N \geq 0$. If $M^{-1} \geq 0$ and $M^{-1}N \geq 0$ then it is called a weak regular splitting, where the symbol " \geq " is used with the sense of non-negativity for matrices.

(2) More generally A Hermitian.

Theorem (2.2.5): Assume that A is symmetric and that $M^T + N$ is positive definite. Then $S(H) < 1$ if and only if A is positive definite, or in another version,

Theorem (2.2.6): If A is a positive definite matrix and if (2.1.3) is completely consistent with (2.1.1), then $S(H) < 1$ if

$$M + M^T - A \tag{2.2.7}$$

is positive definite.

Moreover, we have

$$\|H\|_{A^{-1/2}} < 1 .$$

Conversely, if (2.2.8) holds, then M is positive definite.

2.3 ON THE RATE OF CONVERGENCE

Even if a method converges, it may converge too slowly to be of practical value. Therefore, it is essential to determine the effectiveness of each method. To accomplish this we must consider both the work required per iteration and the number of iterations necessary for convergence.

Definition (2.3.1): For a matrix H assume that $S(H) < 1$ and let $u = Hu + k$.

Then for

$$\alpha = \sup \left\{ \lim_{h \rightarrow \infty} \left| |u^{(n)} - u| \right| : u^{(0)} \text{ in a real space} \right\} \quad (2.3.2)$$

the number

$$R_{\infty}(H) = -\ln \alpha \quad (2.3.3)$$

is called the asymptotic rate of convergence of the iteration (2.1.3).

Since α defined by (2.3.2) satisfies $\alpha = S(H)$ the asymptotic rate of convergence of (2.1.3) is

$$R_{\infty}(H) = -\ln S(H). \quad (2.3.4)$$

The number of iterations required to reduce the size of the initial error, $u^{(0)} - u$, by a factor ζ is approximately determined by the equation

$$S(H)^n = \zeta. \quad (2.3.5)$$

Solving for n we get

$$n \doteq [-\ln S(H)]^{-1} \ln \zeta^{-1}. \quad (2.3.6)$$

We define the quantity,

$$RR(H) = [-\ln S(H)]^{-1} \quad (2.3.7)$$

as the reciprocal rate of convergence of the method (2.1.3). By (2.3.6) the number of iterations required for convergence is approximately proportional to the reciprocal rate of convergence.

2.4 A TEST METHOD

In Young [1977], is defined the method given by

$$u^{(n+1)} = \bar{\rho}(Bu^{(n)} + c) + (1 - \bar{\rho})u^{(n)}, \quad n=0,1,\dots \quad (2.4.1)$$

as the benchmark method. The method is often too slow to be of practical use. With a suitable $\bar{\rho}$, the method has the advantage of converging for any positive definite matrix. This method is useful for the purpose of comparison with other methods.

The construction of the benchmark method is summarized in the following steps. Consider (2.1.1) with A symmetric and positive definite, as well as the splitting $A=D-DB$, where $D=\text{diag}(A)$. We can rewrite (2.1.1) in the form

$$u = Bu + c \quad (2.4.2)$$

where $B = I - D^{-1}A$ (2.4.3)

and $c = D^{-1}b$, (2.4.4)

where I is the identity matrix.

Next step is the simultaneous over-relaxation method, defined by

$$u^{(n+1)} = \rho(Bu^{(n)} + c) + (1 - \rho)u^{(n)}, \quad n=0,1,\dots \quad (2.4.5)$$

where ρ is a real parameter.

Evidently, the iteration matrix of (2.4.5) is

$$H = B_{\rho} = \rho B + (1 - \rho)I. \quad (2.4.6)$$

The eigenvalues of B are real and less than unity, and let $m(B)$ and $M(B)$ be real numbers such that for each eigenvalue μ of B

$$m(B) \leq \mu \leq M(B), \quad (2.4.7)$$

where $m(B) \leq 0 \leq M(B)$.

By (2.4.5) we have

$$S(B_{\rho}) = \max_{\mu \in [m(B), M(B)]} |\rho\mu + 1 - \rho|, \quad (2.4.8)$$

which is minimized with respect to ρ if

$$\rho = \bar{\rho} = \frac{2}{2-M(B)-m(B)} \quad (2.4.9)$$

with the corresponding value of $S(B_{\bar{\rho}})$ given by

$$S(B_{\bar{\rho}}) = \frac{M(B)-m(B)}{2-M(B)-m(B)} \quad (2.4.10)$$

Consider now the matrix \hat{A} , where

$$\hat{A} = D^{-1/2} A D^{-1/2} . \quad (2.4.11)$$

Since $m(B)=1-m(\hat{A})$ and $M(B)=1-M(\hat{A})$, the spectral radius of the benchmark method is given by

$$S(B_{\bar{\rho}}) = \frac{M(\hat{A})-m(\hat{A})}{M(\hat{A})+m(\hat{A})} \quad (2.4.12)$$

or

$$S(B_{\bar{\rho}}) = \frac{P(\hat{A})-1}{P(\hat{A})+1} , \quad (2.4.13)$$

where $P(\hat{A})$ the P-condition number⁽¹⁾ of the matrix \hat{A} .

The reciprocal rate of convergence is

$$RR(B_{\bar{\rho}}) = \left[-\ln \frac{P(\hat{A})-1}{P(\hat{A})+1} \right]^{-1} \approx 2P(\hat{A}) \quad (2.4.14)$$

for large $P(\hat{A})$, i.e. the reciprocal rate of convergence is approximately twice the P-condition number of the matrix \hat{A} .

⁽¹⁾ The spectral condition number.

2.5 BASIC ITERATIVE METHODS

We describe in this section some basic iterative formulae for solving the linear system (2.1.1). We assume that the coefficient matrix A for (2.1.1) is non-singular with all non-zero diagonal entries.

Consider the splitting

$$A = D - C_L - C_U \quad (2.5.1)$$

where $D = \text{diag}(A)$ and $-C_L, -C_U$ are the strictly lower and strictly upper triangular parts of A .

Now (2.1.1) can be rewritten

$$(D - C_L - C_U)u = b, \quad (2.5.2)$$

or

$$Du = (C_L + C_U)u + b. \quad (2.5.3)$$

Then clearly,

$$u = D^{-1}(C_L + C_U)u + D^{-1}b \quad (2.5.4)$$

or

$$u = Bu + c \quad (2.5.5)$$

where

$$B = D^{-1}(C_L + C_U) \quad (2.5.6)$$

$$= L + U$$

$$c = D^{-1}b \quad (2.5.7)$$

and where

$$L = D^{-1}C_L, \quad U = D^{-1}C_U. \quad (2.5.8)$$

Assume that the n^{th} approximation $u^{(n)}$ to the solution $u = A^{-1}b$ has been computed.

Then, the Jacobi method (J method) is given by

$$u^{(n+1)} = Bu^{(n)} + c, \quad n=0,1,\dots \quad (2.5.9)$$

or

$$u^{(n+1)} = (L+U)u^{(n)} + c, \quad n=0,1,\dots \quad (2.5.10)$$

Since $I - B = D^{-1}A$ we have that $(I - B)A^{-1}b = c$ and the method is completely consistent (vd Section (2.1)).

Related to the J method is the Jacobi overrelaxation method (JOR) (vd Section (2.4)) which is given by,

$$u^{(n+1)} = B_{\omega} u^{(n)} + \omega c, \quad n=0,1,\dots \quad (2.5.11)$$

where $B_{\omega} = \omega B + (1-\omega)I$ (2.5.12)

is the iteration matrix of the method and ω a real parameter⁽¹⁾.

If $\omega \neq 0$, the method is completely consistent and if $\omega=1$ we have the J method.

A method closely related to the J method may also be derived from the observation of the intermediate use of the improved values. That is, the latest estimates of the components of $u^{(n+1)}$ are employed immediately upon becoming available. This results in the iterative method

$$u^{(n+1)} = L u^{(n)} + (I-L)^{-1} c, \quad n=0,1,\dots \quad (2.5.13)$$

where $L = (I-L)^{-1} U$ (2.5.14)

is the iteration matrix of the method. We call this iterative scheme the Gauss-Seidel method (GS method).

The GS method can be modified using a relaxation parameter ω (if $\omega > 1$ we are "over-correcting" while if $\omega < 1$ we are "under-correcting") implying the successive overrelaxation (SOR) method,

$$u^{(n+1)} = \omega(Lu^{(n+1)} + Uu^{(n)} + c) + (1-\omega)u^{(n)}, \quad n=0,1,\dots \quad (2.5.15)$$

or $u^{(n+1)} = L_{\omega} u^{(n)} + (I-\omega L)^{-1} \omega c$, (2.5.16)

where $L_{\omega} = (I-\omega L)^{-1} (\omega U + (1-\omega)I)$. (2.5.17)

For $\omega \neq 0$ the SOR method is completely consistent. If $\omega=1$ the SOR method reduces to the GS method.

The symmetric successive overrelaxation method (SSOR method) can be considered as two half-iterations. The first half iteration is the same as the SOR method, while the second half iteration is the SOR method with the equations taken in reverse order. Consequently, the SSOR

⁽¹⁾ By the use of the relaxation parameter ω , in certain instances, we can maximize the asymptotic rate of convergence for the resulting process.

is determined by

$$u^{(n+\frac{1}{2})} = L_{\omega} u^{(n)} + (I - \omega L)^{-1} \omega c \quad (2.5.18)$$

and

$$u^{(n+1)} = U_{\omega} u^{(n+\frac{1}{2})} + (I - \omega U)^{-1} \omega c \quad n=0,1,\dots \quad (2.5.19)$$

where

$$\left. \begin{aligned} L_{\omega} &= (I - \omega L)^{-1} (\omega U + (1 - \omega) I) \\ U_{\omega} &= (I - \omega U)^{-1} (\omega L + (1 - \omega) I) \end{aligned} \right\} \quad (2.5.20)$$

Eliminating $u^{(n+\frac{1}{2})}$ in (2.5.18) and (2.5.19) we get

$$u^{(n+1)} = G_{\omega} u^{(n)} + k_{\omega}, \quad (2.5.21)$$

where

$$\begin{aligned} G_{\omega} &= (I - \omega U)^{-1} (\omega L + (1 - \omega) I) (I - \omega L)^{-1} (\omega U + (1 - \omega) I) \\ &= I - \omega(2 - \omega) (I - \omega U)^{-1} (I - \omega L)^{-1} D^{-1} A \\ &= U_{\omega} L_{\omega} \end{aligned} \quad (2.5.22)$$

with

$$k_{\omega} = \omega(2 - \omega) (I - \omega U)^{-1} (I - \omega L)^{-1} c. \quad (2.5.23)$$

The matrix $I - G_{\omega}$ is non-singular, if $0 < \omega < 2$ and if A is non-singular.

The SSOR is also a completely consistent method.

2.6 CONVERGENCE PROPERTIES

The matrix $A=(a_{i,j})$ emanating from the elliptic equation (1.5) with Dirichlet boundary conditions, has (as we have mentioned in Chapter 1) the following properties:

$$(i) \quad a_{i,i} > 0, \quad a_{i,j} \leq 0, \quad i \neq j$$

$$(ii) \quad a_{i,i} \geq \sum_{\substack{j \\ i \neq j}} |a_{i,j}|$$

$$a_{i,i} > \sum_{\substack{j \\ i \neq j}} |a_{i,j}| \quad \text{for some } i$$

and

$$(iii) \quad A \text{ is irreducible.}$$

Under these conditions the J method converges, whereas the relative magnitudes of the spectral radii of the iteration matrices associated with the J method and GS method can be listed in one of the following ways:

$$(i) \quad S(L) = S(B) = 0$$

$$(ii) \quad 0 < S(L) < S(B) < 1$$

$$(iii) \quad 1 = S(L) = S(B)$$

$$(iv) \quad 1 < S(B) < S(L).$$

Thus, if these methods both converge the GS converges faster than the J method.

If A is a positive definite matrix then the GS always converges, without further restriction on A .

The theorems we cited in Section (2.2), are now applied to these basic methods of Section (2.5). Thus, we have,

Theorem (2.6.1): Let A be a positive definite matrix and let $D=\text{diag}(A)$.

Then,

- (i) $\|B\|_{A^{-1/2}} < 1$ if $2D-A$ is positive definite
- (ii) $\|B_\omega\|_{A^{-1/2}} < 1$ if $2\omega^{-1}D-A$ is positive definite
- (iii) $\|L\|_{A^{-1/2}} < 1$
- (iv) $\|L_\omega\|_{A^{-1/2}} < 1$, if $0 < \omega < 2$.

Moreover by the famous Ostrowski-Reich theorem for the SOR method when A is symmetric we cite

Corollary (2.6.2): Let A be a real matrix with the decomposition (2.5.1) and assume that A is symmetric and that D is positive definite. Then, the SOR method converges for all $0 < \omega < 2$ if and only if A is positive definite.

Proof: Here we set

$$M = \omega^{-1}(D - \omega C_L), \quad N = \omega^{-1}[(1-\omega)D + \omega C_U].$$

Then since

$$M^T + N = \omega^{-1}(2-\omega)D$$

is positive definite for all $0 < \omega < 2$, the corollary follows immediately from Theorem (2.2.5).

Finally, we mention a convergence theorem for the SSOR method.

Theorem (2.6.3): Let A be a symmetric matrix with positive diagonal elements. For any real ω the eigenvalues of G_ω are real and non-negative. Moreover, if

$$0 < \omega < 2 \tag{2.6.4}$$

and if A is positive definite then

$$\|G_\omega\|_{A^{-1/2}} = S(G_\omega) = \|L_\omega\|_{A^{-1/2}}^2 < 1 \tag{2.6.5}$$

Conversely, if $S(G_\omega) < 1$, then $0 < \omega < 2$ and A is positive definite.

A proof for this theorem can be found in e.g. Young [1971].

We next give some comparison results of the convergence rate of the SOR and the SSOR method.

The introduction of the parameter ω into the GS method is not done so as to force convergence but, rather, to enhance the rate of convergence. We can, in certain instances, determine a ω_b , $0 < \omega_b < 2$ such that

$$R_\omega(L_{\omega_b}) \geq R_\omega(L_\omega), \quad 0 < \omega < 2. \quad (2.6.6)$$

The real parameter ω_b is called the "optimum SOR relaxation parameter" since it maximizes the asymptotic convergence rate of the SOR method. This optimum choice of ω which can allow the SOR method to converge faster by an "order of magnitude" than the benchmark method (vd. Section (2.4)), provided that the matrix A is consistently ordered.

If A is positive definite and consistently ordered, the optimum choice of ω , in the sense of minimizing $S(L_\omega)$ is given by

$$\omega_b = \frac{2}{1 + \sqrt{1 - S(B)^2}} \quad (2.6.7)$$

and the corresponding value of $S(L_{\omega_b})$ is

$$S(L_{\omega_b}) = \omega_b - 1 = \left(\frac{S(B)}{1 + \sqrt{1 - S(B)^2}} \right)^2 \quad (2.6.8)$$

When $S(B) \rightarrow 1$ it can be shown that

$$\frac{RR(L_{\omega_b})}{[RR(B_{\bar{\rho}})]^{\frac{1}{2}}} \doteq \frac{1}{2\sqrt{2}} \quad (2.6.9)$$

Thus, we have an order of magnitude improvement over the benchmark method.⁽¹⁾

⁽¹⁾ By (2.4.9) since $M(B) = -m(B) < 1$, $\bar{\rho} = 1$ and $B_{\bar{\rho}} = B$

By Evans [1973] we have that (2.6.8) can be expressed as

$$S(L_{\omega_b}) = \frac{1}{\left(1 + \frac{2}{\sqrt{P(\hat{A})}}\right)^2} \quad (2.6.10)$$

where $P(\hat{A})$ is the P-condition number of the matrix \hat{A} .⁽¹⁾

Since now,

$$R_{\infty}(L_{\omega_b}) = -\ln S(L_{\omega_b}) \quad (2.6.11)$$

we have

$$R_{\infty}(L_{\omega_b}) = \frac{4}{\sqrt{P(\hat{A})}} \quad (2.6.12)$$

or

$$RR(L_{\omega_b}) = \frac{\sqrt{P(\hat{A})}}{4} \quad (2.6.13)$$

In Young [1977] there is the following theorem for the SOR method assuming a positive definite matrix A.

Theorem (2.6.14): Let $\bar{\beta}, M$ and m be numbers such that

$$\begin{aligned} m(B) &\geq m \geq -2\sqrt{\bar{\beta}} \\ M(B) &\leq M \leq 2\sqrt{\bar{\beta}} \\ M &< 1 \\ S(LU) &\leq \bar{\beta}. \end{aligned} \quad (2.6.15)$$

Then

$$S(G_{\omega}) \leq \begin{cases} 1 - \omega(2 - \omega) \frac{1 - M}{1 - \omega M + \omega \bar{\beta}}, & \text{if } \bar{\beta} \geq \frac{1}{4} \text{ or if } \bar{\beta} < \frac{1}{4} \\ & \text{and } \omega \leq \omega^* \\ 1 - \omega(2 - \omega) \frac{1 - m}{1 - \omega m + \omega \bar{\beta}}, & \text{if } \bar{\beta} < \frac{1}{4} \text{ and } \omega > \omega^*. \end{cases} \quad (2.6.16)$$

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

$$\omega^* = \frac{2}{1 + \sqrt{1 - 4\bar{\beta}}} \quad (2.6.17)$$

Moreover, the bound (2.6.16) is minimized if we let

(1) The matrix \hat{A} has been defined in (2.4.11).

$$\omega_1 = \begin{cases} \frac{2}{1+\sqrt{1-2M+4\bar{\beta}}}, & \text{if } M \leq 4\bar{\beta} \\ \omega^* & \text{if } M > 4\bar{\beta}. \end{cases} \quad (2.6.18)$$

The corresponding value of $S(G_{\omega_1})$ is given by

$$S(G_{\omega_1}) \leq \begin{cases} \frac{1 - \frac{1-M}{\sqrt{1-2M+4\bar{\beta}}}}{1 + \frac{1-M}{\sqrt{1-2M+4\bar{\beta}}}}, & \text{if } M \leq 4\bar{\beta} \\ \frac{1-\sqrt{1-4\bar{\beta}}}{1+\sqrt{1-4\bar{\beta}}} = \omega^*-1, & \text{if } M > 4\bar{\beta}. \end{cases} \quad (2.6.19)$$

Young [1977] refers to the value of ω_1 given by (2.6.18) as a "good" value of ω , which is not necessarily the true optimum value in the sense of minimizing $S(G_{\omega})$. A proof of this theorem can be found in Benokratis [1974] or in Young [1977].

Habetler and Wachspress [1961] presupposing the continuity⁽¹⁾ of the eigenvectors of G_{ω} with respect to ω , developed a formula for finding the optimum ω .

Evans and Forrington [1963] derived an iterative procedure for determining the optimum ω . Since then many adaptive schemes are based on the Evans and Forrington procedure.

A modification to the above found (2.6.19) is given by

$$S(G_{\omega_1}) \leq \begin{cases} \frac{1-\sqrt{1-M}}{1+\sqrt{1-M}}, & \text{if } \bar{\beta} \leq \frac{M}{4} \\ \frac{1-\sqrt{\frac{1-M}{2}}}{1+\sqrt{\frac{1-M}{2}}}, & \text{if } \frac{M}{4} < \bar{\beta} \leq \frac{1}{4} \\ \frac{1-\gamma\sqrt{\frac{1-M}{2}}}{1+\gamma\sqrt{\frac{1-M}{2}}}, & \text{if } \bar{\beta} > \frac{1}{4} \end{cases} \quad (2.6.20)$$

⁽¹⁾ Which is not in general true.

where

$$\gamma = \left[1 + \frac{2(\bar{\beta} - \frac{1}{4})}{1-M} \right]^{-1/2} . \quad (2.6.21)$$

Comparing the bounds of $RR(G_{\omega_1})$ with $RR(B_{\rho})$ this result in

$$\frac{RR(G_{\omega_1})}{\sqrt{RR(B_{\rho})}} = \begin{cases} \frac{1}{\sqrt{2}} & , \text{ if } \bar{\beta} \leq \frac{M}{4} \\ 1 & , \text{ if } \frac{M}{4} < \bar{\beta} \leq \frac{1}{4} \\ \gamma^{-1} & , \text{ if } \bar{\beta} > \frac{1}{4} . \end{cases} \quad (2.6.22)$$

2.7 ON THE NOTION OF PRECONDITIONING

We have seen that the asymptotic rate of convergence of the aforementioned iterative methods for positive definite matrices depends inversely on the P-condition number of the coefficient matrix. Thus, a sensible approach to develop for a new technique of accelerating convergence, is to attempt to minimize that condition number, as Evans [1968] initially proposed.

The concept of a minimum P-condition number is an important one. The original system (2.1.1) is theoretically equivalent (Evans [1974]) to the system

$$PAQy = c \quad (2.7.1)$$

with $c=Pb$, $Qy=u$ and P, Q non-singular matrices. Thus, it was the question as to the existence of matrices P and Q for which

$$P(PAQ) < P(A) \quad (2.7.2)$$

Evans dealt with.

Evans [1968] considered P and Q as modified forms of the triangular components of \hat{A} (vd. Section (2.4)). Therefore (2.1.1) is transformed to

$$(I-\omega L)^{-1} \hat{A} (I-\omega U)^{-1} z = b_{\omega} \quad (2.7.3)$$

where $b_{\omega} = (I-\omega L)^{-1} b$ and $z = (I-\omega U)y$.

In the context of that theory ω is allowed to play the role of a preconditioning parameter in a range $0 < \omega < W$, where a minimum value of the P-condition number of the left hand side of (2.7.3) is obtained.

Further, the matrix

$$C = D[(I-\omega L)(I-\omega U)] \quad (2.7.4)$$

was defined as the conditioning matrix of the preconditioning transformation and the matrix,

$$C_{\omega} = C^{-1}A \quad (2.7.5)$$

as the preconditioned matrix of the iterative method.

Since then, certain aspects on preconditioning have appeared and a whole new theory constructed.

Recently, Evans [1980] generalized the preconditioning concept and defined C to be factorable into easily inverted factors in such a way that C^{-1} is an approximate inverse to A . Thus (2.1.1) is transformed in the preconditioned form,

$$C^{-1}Au = C^{-1}b \quad (2.7.6)$$

and the general preconditioned iterative scheme is then defined

$$u^{(n+1)} = u^{(n)} + \tau C^{-1}(b - Au^{(n)}), \quad n=0,1,\dots \quad (2.7.7)$$

where τ is a real parameter which is consistent with (2.7.6) iff C is non-singular and $\tau \neq 0$.

For the purpose of comparison with our method we cite one of the preconditioned schemes, the Preconditioned Simultaneous Displacement method (PSD method) which is given (Evans and Missirlis [1980]) by

$$u^{(n+1)} = D_{\tau,\omega} u^{(n)} + \delta_{\tau,\omega} \quad (2.7.8)$$

where

$$D_{\tau,\omega} = I - \tau(I - \omega U)^{-1}(I - \omega L)^{-1}D^{-1}A \quad (2.7.9)$$

and

$$\delta_{\tau,\omega} = \tau(I - \omega U)^{-1}(I - \omega L)^{-1}D^{-1}b. \quad (2.7.10)$$

The PSD method is convergent (Evans and Missirlis [1980]) if and only if $0 < \omega < 2$ and $0 < \tau < 2\omega(2-\omega) \leq 2$.

The rate of convergence of the PSD method is approximately $O(h)$ and is competitive with the SOR method in certain cases, but the PSD method requires more work per iteration than one SOR iteration.

Moreover, "good" values for the involved parameters are determined in terms of bounds on the eigenvalues of the preconditioned matrix $C = C^{-1}A$ by a theorem, presented here for convenience.

Theorem (2.7.11): Let $\bar{\beta}, M$ and m be numbers such that

$$\begin{aligned} m(B) &\geq m > -2\sqrt{\bar{\beta}} \\ M(B) &\leq M \leq 2\sqrt{\bar{\beta}} \\ M &< 1 \\ S(LU) &\leq \bar{\beta} \end{aligned} \quad (2.7.12)$$

Then, an upper bound on $P(C_\omega) = P(C^{-1}A)$ is given by

$$P(C_\omega) \leq \begin{cases} \frac{1 - \omega M + \omega^2 \bar{\beta}}{(1-M)\omega(2-\omega)} = p_M, & \text{if } \bar{\beta} \geq \frac{1}{4} \text{ or if } \bar{\beta} < \frac{1}{4} \text{ and } \omega \leq \omega^* \\ \frac{1 - \omega m + \omega^2 \bar{\beta}}{(1-m)\omega(2-\omega)} = p_m, & \text{if } \bar{\beta} < \frac{1}{4} \text{ and } \omega > \omega^* \end{cases} \quad (2.7.13)$$

where for $\bar{\beta} < \frac{1}{4}$, ω^* is defined by

$$\omega^* = \frac{2}{1 + \sqrt{1 - 4\bar{\beta}}} \quad (2.7.14)$$

Moreover, the bound on $P(C_\omega)$ is minimized if we let $\omega = \omega_1$ where

$$\omega_1 = \begin{cases} \frac{2}{1 + \sqrt{1 - 2M + 4\bar{\beta}}} = \omega_M, & \text{if } M \leq 4\bar{\beta} \\ \frac{2}{1 + \sqrt{1 - 4\bar{\beta}}} = \omega^*, & \text{if } M > 4\bar{\beta} \end{cases} \quad (2.7.15)$$

With respect to the benchmark method (vd Section (2.4)), we have the asymptotic result, and its corresponding value is given by,

$$P(C_{\omega_1}) \leq \begin{cases} \frac{1}{2} \left(1 + \frac{\sqrt{1 - 2M + 4\bar{\beta}}}{1 - M} \right) = \frac{1}{2} \frac{2 - M\omega}{\omega_M - M\omega_M}, & \text{if } M \leq 4\bar{\beta} \\ \frac{1}{2} \left(1 + \frac{1}{\sqrt{1 - 4\bar{\beta}}} \right) = \frac{1}{2 - \omega^*}, & \text{if } M > 4\bar{\beta}. \end{cases} \quad (2.7.16)$$

Theorem (2.7.11) is a modification on Theorem (2.6.14), exploiting the fact that the SSOR method and the PSD method both possess identical P-condition numbers. Since that theorem is a valuable tool to our analysis, a proof is stated in Appendix B.

With respect to the benchmark method (vd Section (2.4)) we have the asymptotic result

$$\frac{RR(D_{\tau_1, \omega_1})}{\sqrt{RR(B_{\bar{\rho}})}} \leq \begin{cases} \frac{1}{2\sqrt{2}} & , \quad \text{if } \bar{\beta} \leq \frac{M}{4} \\ \frac{1}{2} & , \quad \text{if } \frac{M}{4} < \bar{\beta} < \frac{1}{4} \\ \frac{1}{2} \gamma^{-1} & , \quad \text{if } \frac{1}{4} < \bar{\beta} \end{cases} \quad (2.7.17)$$

where we have used the notation of (2.6.15-21), whereas for the SSOR method asymptotically the following result is obtained

$$RR(D_{\tau_1, \omega_1}) \sim \frac{RR(G_{\omega_1})}{2} \quad (2.7.18)$$

namely, the number of iterations of SSOR is asymptotically twice the number of iterations of the PSD, for achieving the same level of accuracy.

CHAPTER 3

THE PRECONDITIONING BY DIRECT FACTORIZATION

ITERATIVE METHOD

Any attempt to improve...fundamental methods must clearly apply some form of preconditioning to the original equations in order to minimize the P-condition number and hence increase the rate of convergence.

D.J. EVANS

Let us begin with problem (2.1.1)

$$Au = b$$

where A is a positive definite matrix.

Equation (2.1.1) is usually solved using the iterative process

$$u^{(n+1)} = u^{(n)} + \tau(b - Au^{(n)}), \quad n=0,1,\dots$$

The parameter τ is chosen to maximize the asymptotic rate of convergence of the iterative process. This usually leads to a very slow convergence rate.

The attempt to accelerate the procedure has resulted in the generalized preconditioned iterative method (Evans [1968-80]), (2.7.8)

$$u^{(n+1)} = u^{(n)} + \tau M^{-1}(b - Au^{(n)}), \quad n=0,1,\dots$$

where M is a given positive definite matrix (conditioning matrix), the form of which is to be determined.

Note that the problem is solved in one iteration if we choose $M/\tau=A$. Taking into account that

$$\tau M^{-1}A = I$$

is the identity matrix, we conclude that for arbitrary $u^{(n)}$

$$u^{(n+1)} = A^{-1}b.$$

This is a formal expression for the exact solution to the problem. Even though the method looks impressive, its realization requires computation of the inverse A^{-1} , a task whose difficulty equals that of solving the original problem. For that reason the iterative process above is not constructive. It suggests, however, various other approaches for choosing the matrix M/τ , in a sense, would be close to the matrix A and easily solvable, so it is appropriate to extend the discussion on a certain conditioning matrix.

3.1 ON A CONDITIONING MATRIX

Let A be the matrix, obtained from the original problem of the elliptic partial differential equation.

Assuming the splitting

$$A = D - C_L - C_U \quad (3.1.1)$$

where C_L and C_U are strictly the lower and upper triangular parts of A and $D = \text{diag}(A)$, let us take the conditioning matrix M in the following form

$$M = D \begin{bmatrix} I - \omega U & \omega \beta \\ \omega \beta^T & \end{bmatrix}, \quad (3.1.2) \quad (1)$$

where $L = D^{-1} C_L$, $U = D^{-1} C_U$, β is a matrix to be explained later, and ω is a real parameter which can be chosen on the basis of a-priori information about the spectra of the operators involved in the algorithm.

It can be easily seen that M can further be written in the form

$$M = D [(I - \omega U) (I - \omega L) + \omega^2 \beta \beta^T]. \quad (3.1.3)$$

Assume that A is a positive definite matrix with positive diagonal elements and that $\beta \beta^T$ is a positive semidefinite matrix, then

(1) For our convenience let us introduce the notation

$$M = (I - \omega \tilde{U}) (I - \omega \tilde{L}). \quad (3.1.5)$$

M is similar⁽¹⁾ to the positive definite matrix M, where

$$\bar{M} = D^{\frac{1}{2}} [(I - \omega U) (I - \omega L) + \omega^2 \beta \beta^T] D^{\frac{1}{2}}. \quad (3.1.4)$$

It is not an easy matter to relate the closeness of M to A, but it can be conveniently handled by the well-known Wielandt-Hoffman theorem (vd Appendix (A.15)) a lower bound on their distance.

Also, by pre-assuming convergence to the iterative scheme concerning the iteration matrix $(I - M^{-1}A)$, an upper bound is given on their distance with respect to the L^2 -norm.⁽²⁾ Thus, we have

$$\sum (a_i - m_i)^2 \leq \|A - M\|_L^2 < \sum m_i^2 \quad (3.1.6)$$

where a_i and m_i are the eigenvalues of A and M respectively, arranged in non-increasing order.

The idea of "expanding" the matrices, from their original square form to a rectangular one, has been applied previously by Evans [1972] in a direct method for solving tridiagonal systems occurring in the solution of certain elliptic partial difference equations, as well as⁽³⁾ by Evans and Hadjidimos [1979] in a factorization method for the solution of constant quindagonal linear systems.

However, such a scheme of conditioning matrix (or otherwise-named of a similar role matrix) we consider, does not seem to be known in the literature of iterative methods.

(1) See Appendix (A.14). It is then evident that M is a non-singular matrix.

(2) The L^2 -norm of a positive definite matrix expresses its spectral radius.

(3) Or, at least that is the available work we have.

3.2 GENERATION OF THE PRECONDITIONING SCHEME

Evans [1968] pointed out that "any attempt to improve... fundamental methods must clearly apply some form of preconditioning to the original equations, in order to minimize the P-condition number and hence increase the rate of convergence". Let us then describe the transformation of A, by means of the new conditioning matrix (3.1.2), into the preconditioned form. By premultiplying the original system (2.1.1) by (3.1.5) we obtain

$$M^{-1}Au = M^{-1}b \quad (3.2.1)$$

$$\text{or} \quad [(I-\omega\tilde{U})(I-\omega\tilde{L})]^{-1}Au = [(I-\omega\tilde{U})(I-\omega\tilde{L})]^{-1}b \quad (3.2.2)$$

assuming that $\det M \neq 0$.

Let

$$y = [(I-\omega\tilde{U})(I-\omega\tilde{L})]^{-1}Au \quad (3.2.3)$$

Calculate $z=Au$, starting with a guess value of the vector u . Hence

$$z = (I-\omega\tilde{U})(I-\omega\tilde{L})y. \quad (3.2.4)$$

Let v be an intermediate vector given by

$$v = (I-\omega\tilde{L})y. \quad (3.2.5)$$

Now we have to solve the system of equations

$$\left. \begin{array}{l} \text{(a)} \quad z = (I-\omega\tilde{U})y \\ \text{(b)} \quad v = (I-\omega\tilde{L})y. \end{array} \right\} \quad (3.2.6)$$

which may be written in the form

$$\left. \begin{array}{l} \text{(a)} \quad z = [I-\omega\tilde{U}; \beta] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ \text{(b)} \quad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} I-\omega\tilde{L} \\ \beta^T \end{bmatrix} y \end{array} \right\} \quad (3.2.7)^{(1)}$$

Thus, the initial system (3.2.3) can be replaced by the coupled

⁽¹⁾ For convenience where $\omega\beta$ we have put, simply β .

system (3.2.7) (a) and (b) where we ascertain that is underdetermined by k , say, and overdetermined by k , respectively, and hence (3.2.7) (a) and (3.2.7) (b) can only be solved in a coupled manner, assuming a similar partitioned form for $v = [v_1, v_2]^T$.

We can therefore, write (3.2.7) in the form

$$\left. \begin{array}{l} \text{(a) } z = (I - \omega U)v_1 + \beta v_2 \\ \text{(b) } \left. \begin{array}{l} v_1 = (I - \omega L)y \\ v_2 = \beta^T y. \end{array} \right\} \end{array} \right\} \quad (3.2.8)$$

By substituting (3.2.8) (b) in (3.2.8) (a), we derive the following result for y

$$z = (I - \omega U) [(I - \omega L)y] + \beta (\beta^T y) \quad (3.2.9)$$

and finally

$$y = [I + (I - \omega L)^{-1} (I - \omega U)^{-1} \beta \beta^T]^{-1} (I - \omega L)^{-1} (I - \omega U)^{-1} z. \quad (3.2.10)$$

In a similar manner we can show for b , the right part of (2.1.1) that

$$\bar{b} = [I + (I - \omega L)^{-1} (I - \omega U)^{-1} \beta \beta^T]^{-1} (I - \omega L)^{-1} (I - \omega U)^{-1} b. \quad (3.2.11)$$

Since it is up to us to define β we shall see later that for its sparseness the above result can be obtained in an algorithmic form consisting of a simple Gaussian elimination.

That approach for the determination of y and b is similar to the one followed by Evans and Hadjidimos [1979] for the factorization of special symmetric non-periodic quindagonal matrices.

3.3 AN HEURISTIC METHOD FOR THE MATRIX β

Considering the five-point approximation difference formula to an elliptic partial differential equation as we have developed in Chapter 1 we derive a system of (2.1.1) from where the matrix A is of order $[(N-1) \times (N-1)]^2$.

We attempt now to determine an heuristic method for the matrix β or for "completing the rank" of the matrix UL , with the proper precautions to avoid any instability arising from an arbitrary definition of that matrix.

Our requirement with respect to the conditioning matrix A is that it can be written as the sum of matrices having at least one non-zero element in any of their columns and rows.

The scheme (3.1.3) of M implies

$$M = D[I - \omega(L+U) + \omega^2 UL + \omega^2 \beta \beta^T] \quad (3.3.1)$$

where the last column and last row of the matrix UL are the zero vector and its transpose respectively.

Thus we propose the following principles,

- H.I If the elements of the matrix U follow a recursive property or formula, then β is chosen to be the $[(N-1) \times (N-1)] \times (N-1)$ sparse matrix with elements formulated by the above recursive property of the formula and have been situated towards the direction of the band of UL .
- H.II If no relation or property appears in the elements of the matrix U , then β is chosen to be the $[(N-1) \times (N-1)] \times (N-1)$ sparse matrix with elements \sqrt{b} which have been situated towards the direction of the band of U , where

$$0 < b \leq \bar{b} \quad , \quad \bar{b} \geq S(UL) \quad . \quad (3.3.2)$$

Moreover, independently of H.I and H.II, we choose

- H.III β to be the $[(N-1) \times (N-1)] \times 1$ sparse matrix where the $[(N-1) \times (N-1), 1]$ position has the non-zero element $\sqrt{b} > 0$ and (3.3.2) holds.

From the above situation we can see that the matrix $\beta\beta^T$ is very sparse with only a few non-zero elements (that gives an advantage to the arithmetic operation count of our algorithm) as well as that the matrix⁽¹⁾

$$\tilde{UL} = UL + \beta\beta^T \quad (3.3.3)$$

has at least one element different than zero in every of its columns and its rows.

If we partition the matrix A by placing all mesh points of successive horizontal mesh lines segments into successive sets, with the five-point approximation formula again, then the block form of A is tridiagonal (for such a derivation see for example Varga [1962]). Then, the matrix UL is block diagonal with its last diagonal block element equal to zero. In such a case, we propose again H.I and H.II where for H.II we have, $\beta\beta^T = b^{(\pi)} I$, $0 < b^{(\pi)} \leq \bar{b}^{(\pi)}$ and $\bar{b}^{(\pi)} \geq S(U^{(\pi)} L^{(\pi)})$, and (π) refers to the partition of the matrix A.

3.4 ON THE EIGENVALUES OF THE MATRIX $\tilde{UL} = UL + \beta\beta^T$

We attempt now to establish a relation between the eigenvalues of the matrices \tilde{UL} , UL and $\beta\beta^T$ and thus we consider the changing of the eigenvalues of UL by adding $\beta\beta^T$ to it, where $\beta\beta^T$ has been investigated in Section 3.

Let a_i, b_i and γ_i be the eigenvalues of UL, $\beta\beta^T$ and \tilde{UL} respectively, where all three sets are arranged in non-increasing order. Since $\tilde{UL} = UL + \beta\beta^T$, by the minimax theorem (vd for example Wilkinson [1965]) we have

$$\gamma_s = \min_{\langle p_i, x \rangle = 0} \max_{\|x\|=1} (\langle x, \tilde{UL}x \rangle), \quad i=1, 2, \dots, s-1; \quad s < (N-1)^2 \quad (3.4.1)$$

⁽¹⁾ Notice that the nullity of UL is at least one.

where the p_i 's are any s non-null vectors, in a subspace of the matrix UL .

If R is a real unitary matrix such that

$$R^T ULR = \text{diag}(a_i) \quad (3.4.2)$$

and since (3.4.1) holds for any p_i , then if $p_i = Re_i$, where e_i is the i^{th} unit base vector of $R^{s \times s}$, $i \leq s$, we have from (3.4.1) that

$$\begin{aligned} \gamma_s &\leq \max_{\|x\|=1} (\langle x, ULx \rangle + \langle x, \beta\beta^T x \rangle) \\ &\max_{\|x\|=1} \left(\sum_{i \geq s} a_i y_i^2 + \langle x, \beta\beta^T x \rangle \right) \end{aligned} \quad (3.4.3)$$

or

$$\gamma_s \leq a_s + M(\beta\beta^T), \text{ for any } x. \quad (3.4.4)$$

Similarly we can show that

$$\gamma_s \geq a_s + m(\beta\beta^T) \quad (3.4.5)$$

or since $m(\beta\beta^T) = 0$,

$$\gamma_s \geq a_s. \quad (3.4.6)$$

Thus, we have

$$a_s \leq \gamma_s \leq a_s + M(\beta\beta^T). \quad (3.4.7)$$

Relations (3.4.7) imply that when $\beta\beta^T$, is added to UL all its eigenvalues are changed by an amount which lies between the smallest and the largest of the eigenvalues of $\beta\beta^T$.

NOTICE: It is evident that in the case H.III of Section 3 no change in the eigenvalues of the matrix UL is observed by adding the matrix $\beta\beta^T$ to it.

3.5 THE PRECONDITIONING BY DIRECT FACTORIZATION METHOD (PDF METHOD)

In this section we shall establish an iterative method for approximating the solution $A^{-1}b$ of (2.1.1), where the matrix A has the properties stated in Chapter 1.

The method relies on the spectral characteristics of the operators involved. A brief description may be given as follows: We construct an iterative process with the iteration matrix depending on a set of two parameters. These parameters will be taken identical for all steps and chosen by minimizing the spectral radius of the iteration matrix or in other words by minimizing the P-condition number of the preconditioned matrix. Therefore, we use an a-priori information about the spectra of the corresponding matrices. The choice of these parameters is an integral part of the optimization of our algorithm. As a rule, the main problem is in finding the corresponding spectral bounds.

Let us therefore consider obtaining the solution of (2.1.1) for u and b belonging to a finite dimensional inner product space $R^{v \times v}$, and A a positive definite matrix on the same space. Let M be a positive definite matrix on $R^{v \times v}$ whose inverse can be discovered in an easy manner.

In the orthogonal space $R^{v \times v}$, A and M are bounded and compact, and the compactness of the unit sphere in $R^{v \times v}$ implies the existence of positive constants λ_1 and λ_n such that

$$\lambda_n \langle u, Mu \rangle \leq \langle u, Au \rangle \leq \lambda_1 \langle u, Mu \rangle, \quad (3.5.1)$$

for all non-zero u in $R^{v \times v}$.

As mentioned previously our primary and central problem is to choose M , in such a fashion that λ_n is close to λ_1 and therefore $\frac{\lambda_1}{\lambda_n}$ is a finite number.

We then define as our *conditioning matrix* M , the matrix

$$M = D[I - \omega(L+U) + \omega^2(UL + be_v e_v^T)] \quad (3.5.2)$$

where e_v the v^{th} unit vector of $R^{v \times v}$, b a constant by means of H.III of Section 3.3, $D = \text{diag}(A)$ with positive elements and DL and DU , the strictly lower and strictly upper triangular parts of the matrix A .

We define the *Preconditioning by Direct Factorization* stationary iterative scheme (Evans [1980]) as follows:

$$\begin{aligned} Mu^{(n+1)} &= Mu^{(n)} - \tau(Au^{(n)} - b), \quad n=0,1,\dots \\ \text{or} \quad u^{(n+1)} &= (I - \tau M^{-1}A)u^{(n)} + \tau M^{-1}b, \quad n=0,1,\dots \end{aligned} \quad (3.5.3)$$

where τ is a positive acceleration parameter to be chosen later so that the error $A^{-1}b - u^{(n)}$ go to zero in some norm, and where the solution $A^{-1}b$ of (2.1.1) is clearly a fixed point of (3.5.3). From its construction, the *Preconditioning by Direct Factorization* method (PDF method) is a completely consistent method.

We note that the above scheme can be seen as a fractional-step scheme (Marchuk [1975]). Namely, if we choose

$$\left. \begin{aligned} \xi^{(n+1/3)} &= \omega U \xi^{(n+1/3)} + r^{(n)} \\ \xi^{(n+2/3)} &= \omega L \xi^{(n+2/3)} + \xi^{(n+1/3)} \\ (I+F)\xi^{(n+1)} &= \xi^{(n+2/3)} \\ u^{(n+1)} &= u^{(n)} - \tau \xi^{(n+1)} \end{aligned} \right\} \quad (3.5.4)$$

where

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

and ⁽¹⁾

$$F = \omega^2 b (I - \omega L)^{-1} (I - \omega U)^{-1} e_v$$

and where the third step is a simple Gaussian elimination, given in Appendix C.

The *preconditioned matrix* B_ω is then defined (Evans [1968]) by

⁽¹⁾ The computational work for F is executed in similar fractional steps.

$$B_{\omega} = M^{-1}A. \quad (3.5.5)$$

Since M is a positive definite matrix (vd Section (3.5)), then

B_{ω} is similar to the matrix

$$\begin{aligned} \bar{B}_{\omega} &= M^{\frac{1}{2}} B_{\omega} M^{-\frac{1}{2}} \\ &= M^{\frac{1}{2}} (M^{-1}A) M^{-\frac{1}{2}} \\ &= M^{-\frac{1}{2}} A M^{-\frac{1}{2}} \end{aligned} \quad (3.5.6)$$

As we can ascertain from Theorem (A.7) and Definition (A.3) of Appendix A, since A is positive definite then \bar{B}_{ω} is positive definite which implies that B_{ω} is similar to a positive definite matrix.

Lemma (3.5.7) ⁽¹⁾: Let M, A be positive definite matrices in $R^{v \times v}$. Then, for any vector $x \neq 0$, we have

$$0 < m(B_{\omega}) \leq \frac{\langle x, Ax \rangle}{\langle x, Mx \rangle} \leq M(B_{\omega}), \quad (3.5.8)$$

where $m(B_{\omega})$ and $M(B_{\omega})$ denote the eigenvalues of minimum and maximum algebraic value respectively of matrix B_{ω} .

Furthermore, if $y \neq 0$ and $B_{\omega}y = \lambda y$ then

$$\lambda = \frac{\langle y, Ay \rangle}{\langle y, My \rangle}. \quad (3.5.9)$$

Proof

For a proof to this well-known theorem, see for example, Diamond [1972] or Young [1977]. ■

Lemma (3.5.7) states a sufficient condition for obtaining bounds on the P-condition number of the matrix B_{ω} , i.e. $P(B_{\omega})$, which we shall deal with later.

Since, from Lemma (3.5.7) the eigenvalues of B_{ω} are between the minimum and the maximum of

(1) In the case where A is symmetric and (3.5.1) is satisfied with constants λ_1 and λ_n then (Gunn [1964])

$$\|B_{\omega}\|_{L_2} \leq \max(|\lambda_1|, |\lambda_n|) \text{ or } \|M^{-\frac{1}{2}} A M^{-\frac{1}{2}}\|_{L_2} \leq \max(|\lambda_1|, |\lambda_n|).$$

$$\frac{\langle x, M^{-\frac{1}{2}} A M^{-\frac{1}{2}} x \rangle}{\langle x, x \rangle} \quad (3.5.10)$$

if we let $z = M^{-\frac{1}{2}} x$ then (3.5.10) becomes

$$\frac{\langle z, Az \rangle}{\langle z, Mz \rangle} \quad (3.5.11)$$

Thus, we shall investigate the behaviour of the ratio (3.5.11) in terms of inner products as introduced by Habetler and Wachspress [1961].

Let then μ be an eigenvalue of the matrix B_{ω} and v an associated eigenvector. Thus

$$B_{\omega} v = \mu v \quad (3.5.12)$$

and by (3.6.2) and (3.6.5) we have

$$Av = \mu D [I - \omega(U+L) + \omega^2 \tilde{U}L] v \quad (3.5.13)$$

where $\tilde{U}L = UL + b e_{\frac{v}{v}} e_{\frac{v}{v}}^T$.

By taking inner products of both sides with respect to v , then in relation to μ we get the expression

$$\mu = \frac{\langle v, Av \rangle}{\langle v, D [I - \omega(U+L) + \omega^2 \tilde{U}L] v \rangle} \quad (3.5.14)$$

Expanding numerator and denominator in (3.5.14) we obtain

$$\mu = \frac{\langle v, Dv \rangle - \langle v, DBv \rangle}{\langle v, Dv \rangle - \omega \langle v, DBv \rangle + \omega^2 \langle v, D\tilde{U}L v \rangle} \quad (3.5.15)$$

We now divide both parts of μ by the non-zero quantity $\langle v, Dv \rangle$. Then,

(3.5.15) becomes

$$\mu = \frac{1 - a(v)}{1 - \omega a(v) + \omega^2 b(v)} \quad (3.5.16)$$

where

$$a(v) = \frac{\langle v, DBv \rangle}{\langle v, Dv \rangle} \quad (3.5.17)$$

and

$$b(v) = \frac{\langle v, D\tilde{U}L v \rangle}{\langle v, Dv \rangle} \quad (3.5.18)$$

Now B is similar to the symmetric matrix $D^{\frac{1}{2}}BD^{-\frac{1}{2}}$, therefore

$$a(v) = \frac{\langle v, DBv \rangle}{\langle v, Dv \rangle} = \frac{\langle D^{\frac{1}{2}}v, (D^{\frac{1}{2}}BD^{-\frac{1}{2}})D^{\frac{1}{2}}v \rangle}{\langle D^{\frac{1}{2}}v, D^{\frac{1}{2}}v \rangle} \quad (3.5.19)$$

is a Rayleigh quotient with respect to $D^{\frac{1}{2}}BD^{-\frac{1}{2}}$, thus we have

$$m(B) \leq a(v) \leq M(B), \quad \text{for any } v \neq 0. \quad (3.5.20)$$

Similarly $\tilde{U}L$ is similar to the positive definite matrix $D^{\frac{1}{2}}\tilde{U}LD^{-\frac{1}{2}} = D^{\frac{1}{2}}(UL)D^{-\frac{1}{2}} + D^{\frac{1}{2}}(be \frac{e^T}{v} \frac{v}{v})D^{-\frac{1}{2}}$, therefore

$$b(v) = \frac{\langle v, \tilde{U}Lv \rangle}{\langle v, Dv \rangle} = \frac{\langle D^{\frac{1}{2}}v, (D^{\frac{1}{2}}\tilde{U}LD^{-\frac{1}{2}})D^{\frac{1}{2}}v \rangle}{\langle D^{\frac{1}{2}}v, D^{\frac{1}{2}}v \rangle} \quad (3.5.21)$$

is a Rayleigh quotient as well with respect to $D^{\frac{1}{2}}\tilde{U}LD^{-\frac{1}{2}}$, thus

$$0 \leq b(v) \leq S(\tilde{U}L) \quad (3.5.22)$$

We have thus seen that the quantities $a(v)$ and $b(v)$ are bounded by the extreme eigenvalues of the matrices B and $\tilde{U}L$ respectively.

Since now $\text{trace}(B) = 0$, it follows that

$$m(B) = m \leq 0 \leq M = M(B) \quad (3.5.23)$$

Moreover, since B is similar to $D^{\frac{1}{2}}BD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ and since $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$

is positive definite (for A is positive definite) we have

$$M(B) = M < 1. \quad (3.5.24)$$

Therefore

$$m \leq 0 \leq M < 1. \quad (3.5.25)$$

Furthermore, since $S(B) = S(D^{\frac{1}{2}}BD^{-\frac{1}{2}})$ and $S(\tilde{U}L) = S(D^{\frac{1}{2}}\tilde{U}LD^{-\frac{1}{2}})$, where

$S(\tilde{U}L) = S(UL)$ we have

$$\begin{aligned} S(B) &= S(D^{\frac{1}{2}}(L+U)D^{-\frac{1}{2}}) = \\ &= \left\| \left| D^{\frac{1}{2}}LD^{-\frac{1}{2}} + D^{\frac{1}{2}}UD^{-\frac{1}{2}} \right| \right\| \\ &\leq 2 \left\| \left| D^{\frac{1}{2}}UD^{-\frac{1}{2}} \right| \right\| \\ &= 2\sqrt{S(D^{\frac{1}{2}}\tilde{U}LD^{-\frac{1}{2}})} \\ &= 2\sqrt{S(UL)}. \end{aligned} \quad (3.5.26)$$

If \bar{b} is an upper bound for $S(UL)$ then

$$S(\tilde{U}L) \leq \bar{b} \quad (3.5.27)$$

and

$$S(B) \leq 2\sqrt{b} \quad (3.5.28)$$

It follows then,

$$\begin{aligned} -m &\leq 2\sqrt{b} \\ M &\leq 2\sqrt{b} \end{aligned} \quad (3.5.29)$$

and, if the bounds M or $-m$ exceed $2\sqrt{b}$ we replace M by $2\sqrt{b}$ or m by $-2\sqrt{b}$.

By the above analysis, we are now in a position to seek a lower bound on $m(B_\omega)$, the smallest eigenvalue of B_ω , involving only known quantities determined a-priori without the need to preassume continuity on ω , of any eigenvector $v=v(\omega)$ of B_ω .

This is a trivial problem which many authors have dealt with, Benokraitis [1974] and Young [1971] determining an upper bound for the spectral radius of SSOR method and Evans and Missirlis [1980] who gave a similar proof to them seeking a lower bound to the minimum eigenvalue of his preconditioned matrix. Therefore, following the previous authors we obtain Lemma (3.5.30), in which we attempt to find the minimal of the function (3.5.16) with respect to $a(v)$ and $b(v)$ by taking successive lower bounds in any order, given the analysis concerning the bounds on B and UL . Thus,

Lemma (3.5.30): If $-2\sqrt{b} \leq m \leq M(B)$

$$M(B) \leq M \leq \min(1, 2\sqrt{b})$$

$$S(\tilde{UL}) \leq \bar{b}$$

then a lower bound on $m(B_\omega)$ may be given by

$$m(B_\omega) \geq \begin{cases} \frac{1-M}{1-\omega M + \omega \frac{2^-}{b}}, & \text{if } \bar{b} > \frac{1}{4} \text{ or if } b < \frac{1}{4} \text{ and } \omega \leq \omega^* \\ \frac{1-m}{1-\omega m + \omega \frac{2^-}{b}}, & \text{if } \bar{b} < \frac{1}{4} \text{ and } \omega > \omega^*, \end{cases} \quad (3.5.31)$$

where for $\bar{b} < \frac{1}{4}$ we define ω^* by

$$\omega^* = \frac{2}{1 + \sqrt{1 - 4b}} \quad (3.5.32)$$

Our aim now is to estimate an upper bound on the largest eigenvalue $M(B_\omega)$, of the matrix B_ω . Thus, we merely state the following discussion on the lower bound of the Rayleigh quotient of the conditioning matrix M .

The matrix M given by (3.5.2) can be written in the form

$$M = \omega(2-\omega)A + D(C+E) \quad (3.5.33)$$

where

$$C = [(1-\omega)I + \omega U] [(1-\omega)I + \omega L] \quad (3.5.34)$$

and where $E = \omega^2 b \frac{e}{v} \frac{e^T}{v}$. It is evident that C and E are positive semi-definite matrices and exact symmetry is preserved in the matrix

$$C' = C + E. \quad (3.5.35)$$

Thus, we can seek relations between the eigenvalues of C and E or between the eigenvalues of C' , by applying a classical technique due to Wilkinson [1965].

Since E is rank unity, if we partition C in a similar form to E we have

$$C = \begin{bmatrix} C_1 & | & C \\ \hline C^T & | & \bar{C} \end{bmatrix} \quad (3.5.36)$$

where C_1 is the first minor matrix of C . Then, there is a real unitary matrix P of order $v-1$ such that

$$P^T C_1 P = \text{diag}(C_1) \quad (3.5.37)$$

and if we define R by the relation

$$R = \begin{bmatrix} P & | & O \\ \hline O & | & 1 \end{bmatrix} \quad (3.5.38)$$

then R is a real unitary and $R^T C' R$ implies

$$R^T (C+E) R = \begin{bmatrix} \text{diag}(C_1) & \ell \\ \hline \ell^T & \bar{C} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \hline 0 & \omega^2 b \end{bmatrix} \quad (3.5.39)$$

where $\ell = P^T c$ and $\omega^2 b$ is the unique non-zero eigenvalue of E .

The eigenvalues of C and C' are therefore those of

$$\begin{bmatrix} \text{diag}(C_1) & \ell \\ \hline \ell^T & \bar{C} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \text{diag}(C_1) & \ell \\ \hline \ell^T & \bar{C} + \omega^2 b \end{bmatrix} \quad (3.5.40)$$

and if we denote these by λ_i and λ'_i in decreasing order, then they satisfy the relation (vd Wilkinson [1965], p.97)

$$\left. \begin{aligned} \lambda'_i &= \lambda_i + \rho_i \omega^2 b \\ 0 < \rho_i &\leq 1, \quad \sum \rho_i = 1 \end{aligned} \right\} \quad (3.5.41)$$

Hence all the eigenvalues of C after the addition of E have been shifted by an amount which lies between zero and $\omega^2 b$, the unique eigenvalue of E .

Since C' is a symmetric matrix, for any vector $v \neq 0$ its Rayleigh quotient satisfies the relations

$$\min_{v \neq 0} \frac{\langle v, C' v \rangle}{\langle v, v \rangle} \leq \frac{\langle v, C' v \rangle}{\langle v, v \rangle} \leq \max_{v \neq 0} \frac{\langle v, C' v \rangle}{\langle v, v \rangle} \quad (3.5.42)$$

or in view of (3.5.41), relations (3.5.42) are written as

$$m(C) + \omega^2 \rho_1 b \leq \frac{\langle v, C' v \rangle}{\langle v, v \rangle} \leq M(C') \quad (3.5.43)$$

or since $m(C) = 0$ (for C is positive semi-definite), we have

$$\omega^2 \rho_i b \leq \frac{\langle v, DC'v \rangle}{\langle v, Dv \rangle} \leq M(C') \quad (3.5.44)$$

We can now obtain a bound on $M(B_\omega)$ given that (3.5.44) holds for any $v \neq 0$.

Lemma (3.5.45): Let A be a positive definite matrix and $\omega \in (1, \omega_b)$. Then,

$$M(B_\omega) \leq \frac{1}{\omega(2-\omega) + \frac{\omega^2 b}{2} \rho} \quad (3.5.46)$$

where $\rho \in (0, 1]$ and

$$\omega_b = \frac{2}{1 - \frac{b}{2} \rho}$$

Proof

As it has been shown in Lemma (3.5.7) the largest eigenvalue of B_ω is the maximum of (3.5.10). Let μ be an eigenvalue of B_ω and v an associated eigenvector. Thus,

$$\text{where } \langle v, v \rangle = 1, \text{ or} \quad \mu = \langle v, B_\omega v \rangle \quad (3.5.47)$$

$$\mu = \frac{\langle v, Av \rangle}{\langle v, Mv \rangle} \quad (3.5.48)$$

and by (3.5.33) and (3.5.41), we have

$$\mu \leq \frac{\langle v, Av \rangle}{\omega(2-\omega) \langle v, Av \rangle + \omega^2 \rho_i b} \quad (3.5.49)$$

Since now the right part of (3.5.49) is an increasing function of $\langle v, Av \rangle \in [1-M(B), 1-m(B)]$, then the maximum of that quantity with respect to $\langle v, Av \rangle$ occurs when $\langle v, Av \rangle = 1-m(B)$ and therefore

$$\mu \leq \frac{1}{\omega(2-\omega) + \omega^2 \rho \frac{b}{1-m(B)}} \quad (3.5.50)$$

where ρ has been chosen from the set $\{\rho_i : \sum \rho_i = 1, 0 < \rho_i \leq 1\}$, and $-m(B) \rightarrow 1^-$. Thus (3.5.46) is valid.

Notation: When the mesh size h of the net tends to zero then $-m(B) \rightarrow 1^-$, and (3.5.50) can be written

$$M(B_\omega) \leq \frac{1}{\omega(2-\omega) + \omega^2 \rho \frac{\bar{b}}{2}} \quad (3.5.51)$$

In our later arithmetical examples we consider ρ varying approximately as h^2 . Hence, without loss of generality we use \bar{b} instead of b in formula (3.5.51), given that $b \leq \bar{b}$.

With regard to ω_b , the supremum of ω ,

$$\omega_b(\bar{b}; m(B)) = \frac{2}{1 - \frac{\rho \bar{b}}{1 - m(B)}} \quad (3.5.52)$$

since it is an increasing function of \bar{b} , it is limited in the interval

$$\omega_b(b; m(B)) \in \begin{cases} [2, \frac{16}{7}\rho) & , \text{ if } 0 < \bar{b} < \frac{1}{4} \\ [\frac{16}{7}\rho, 4\rho) & , \text{ if } 1 > \bar{b} > \frac{1}{4} . \end{cases} \quad (3.5.53)$$

where $-m(B) \rightarrow 1^-$ and $\rho \in (0, 1]$.

We now state a necessary and sufficient condition for the iterative scheme (3.5.3) to be convergent.

Theorem (3.5.54): If A is a positive definite matrix then the iterative scheme

$$u^{(n+1)} = (I - \tau M^{-1} A) u^{(n)} + \tau M^{-1} b, \quad n=0, 1, \dots$$

is convergent if and only if

$$0 < \tau < \frac{2}{M(B_\omega)} . \quad (3.5.55)$$

Proof

The iterative scheme (3.5.3) is convergent if and only if

$$|\langle v, (I - \tau M^{-1} A) v \rangle| < 1 \quad \text{for every non-zero vector } v. \quad (3.5.56)$$

Thus, on the unit sphere, since M and $B_\omega = M^{-1} A$ are positive definite matrices (vd Section (3.5) and (3.6)), equivalently from (3.5.56) we have

$$-1 < 1 - \tau \langle v, M^{-1} A v \rangle < 1 \quad (3.5.57)$$

or

$$0 < \tau < \frac{2}{M(B_\omega)} . \quad (3.5.58)$$

When we seek for estimated parameters the above theorem implies,

Proposition (3.5.58): If A is a positive definite matrix and $\omega \in (1, \omega_b)$ then (3.5.3) is convergent if and only if

$$0 < \tau < 2 \left[\omega(2-\omega) + \omega^2 \frac{\bar{b}}{2} \rho \right] \quad (3.5.59)$$

Proof:

By Lemma (3.5.45) and Theorem (3.5.54) the proof is evident. ■

We can maximize the rate of convergence of the method by choosing an optimal value for τ in its range. Thus we let τ to have the value

$$\tau_0 = \frac{2}{m(B_{\omega_0}) + M(B_{\omega_0})} \quad (3.5.60)$$

for the case of optima parameters, where ω_0 is the optimum point to be explained later, whereas

$$\tau_1 = \frac{2}{M(B_{\omega_1}) + M(B_{\omega_1})} \quad (3.5.61)$$

for the case of estimated ones, where ω_1 is an estimated point to be explained later also.

The acceleration parameter $\tau_{0,1}$ can now be given, by means of the P-condition number of B_{ω} , $P(B_{\omega})$, in the form

$$\tau_{0,1} = \frac{2/M(B_{\omega_{0,1}})}{1 + \frac{1}{P(B_{\omega_{0,1}})}} \quad (3.5.62)$$

where by the notation $\tau_{0,1}$ we mean either τ_0 or τ_1 .

3.6 INTERVAL ESTIMATION ON THE BOUND OF $P(B_\omega)$

From the analysis already developed we are now able to seek an interval in which an optimal value of a bound on the P-condition number of B_ω lies.

Since

$$P(B_\omega) = \frac{M(B_\omega)}{m(B_\omega)}, \quad (3.6.1)$$

from Lemmas (3.5.30) and (3.5.45) concerning $m(B_\omega)$ and $M(B_\omega)$ respectively, we have

$$P(B_\omega) \leq \begin{cases} \frac{1-\omega M + \omega^2 \bar{b}}{(1-M)[\omega(2-\omega) + \omega^2 \frac{\bar{b}}{2\rho}]}, & \text{if } \bar{b} > \frac{1}{4} \text{ or if } \bar{b} < \frac{1}{4} \text{ and } \omega \leq \omega^* \\ \frac{1-\omega m + \omega^2 \bar{b}}{(1-m)[\omega(2-\omega) + \omega^2 \frac{\bar{b}}{2\rho}]}, & \text{if } \bar{b} < \frac{1}{4} \text{ and } \omega > \omega^* \end{cases} \quad (3.6.2)$$

where for $\bar{b} < \frac{1}{4}$, ω^* is defined by

$$\omega^* = \frac{2}{1 + \sqrt{1 - 4\bar{b}}} \quad (3.6.3)$$

But $\frac{\bar{b}}{2\rho}$, $\rho \in (0, 1]$, is in general a small number in the interval $(0, \frac{\bar{b}}{2})$, so any attempt to minimize $P(B_\omega)$ with respect to ω , in the case of $\bar{b} > \frac{1}{4}$ implies implicit quantities, as optimal values, to be of practical use. On the other hand, the gain in the number of iterations is of little importance in comparison with the ones obtained with the estimated values from the following theorem (3.6.4). However, it is a simple formulae we need for the estimated parameters in order for it to be easily handled in an accelerating or an adaptive process. Theorem (3.6.4) proposes a less effective but more practical manner in seeking optimal values on $P(B_\omega)$ with respect to ω .

Theorem (3.5.4): Let \bar{b}, M and m be numbers such that

$$\begin{aligned}
 m(B) &\geq m \geq -2\sqrt{b} \\
 M(B) &\leq M \leq 2\sqrt{b} \\
 M &< 1 \\
 S(\tilde{U}L) &\leq \bar{b}.
 \end{aligned} \tag{3.6.5}$$

Then (3.6.2) is valid.

Moreover, a bound on $P(B_{\omega})$ may be given if we let

$$\omega_1 = \begin{cases} \frac{2}{1+\sqrt{1-2M+4\bar{b}}} = \omega_M, & \text{if } M \leq 4\bar{b} \\ \frac{2}{1+\sqrt{1-4\bar{b}}} = \omega^*, & \text{if } M \geq 4\bar{b} \end{cases} \tag{3.6.6}$$

and therefore the corresponding bound of $P(B_{\omega})$ is given by

$$P(B_{\omega_1}) \leq \frac{P(C_{\omega_1})}{1+kP(C_{\omega_1})} \tag{3.6.7}$$

where

$$P(C_{\omega_1}) \leq \begin{cases} \frac{1}{2} \left(1 + \frac{\sqrt{1-2M+4\bar{b}}}{1-M}\right) = \frac{1}{2} \frac{2-M\omega_M}{(1-M)\omega_M}, & \text{if } M \leq 4\bar{b} \\ \frac{1}{2} \left(1 + \frac{1}{\sqrt{1-4\bar{b}}}\right) = \frac{1}{2-\omega^*}, & \text{if } M \geq 4\bar{b}, \end{cases} \tag{3.6.8}$$

and where k is a positive constant lying in the interval

$$J = \begin{cases} \frac{(1-M)\omega_M^{2-\bar{b}}}{(2-\omega_M)(2-\omega_M)}, & \text{if } M \leq 4\bar{b} \\ \left(0, \frac{\omega^*\bar{b}}{2}\right), & \text{if } M \geq 4\bar{b} \end{cases} \tag{3.6.9}$$

Thus,

$$P(B_{\omega_1}) \leq \begin{cases} \frac{2-\omega_M}{(1-M)\omega_M} \frac{2-\omega_M}{4-\omega_M(2-\bar{b}\rho)}, & \text{if } M \leq 4\bar{b} \\ \frac{1}{2-\omega^*(1-\frac{\bar{b}}{2}\rho)}, & \text{if } M \geq 4\bar{b} \end{cases} \tag{3.6.10}$$

where $\rho \in (0, 1]$.

Proof

Now the boundary (3.6.2) on the P-condition number, $P(B_\omega)$ can be given in the form

$$P(B_\omega) \leq \frac{1}{\frac{(1-a)\omega(2-\omega)}{1-\omega a + \omega^2 \bar{b}} + \frac{1}{2} \frac{(1-a)\omega^2 \bar{b} \rho}{1-\omega a + \omega^2 \bar{b}}} \quad (3.6.12)$$

where

$$a = \begin{cases} M, & \text{if } \bar{b} \geq \frac{1}{4} \text{ or if } \bar{b} < \frac{1}{4} \text{ and } \omega \leq \omega^* \\ m, & \text{if } \bar{b} < \frac{1}{4} \text{ and } \omega > \omega^* \end{cases} \quad (3.6.13)$$

Let $F_1(\omega; a, \bar{b})$ and $F_2(\omega; a, \bar{b})$ be the first and the second terms respectively of the denominator of (3.6.12).

The function $F_2(\omega; a, \bar{b})$ is increasing with respect to ω , in the range (1,2), then in a subinterval of ω , say I, $F_2(\omega; a, \bar{b})$ has a minimum at $\omega = \omega_1$, where $\omega_1 = \inf_{\omega \in I} \omega$.

$$F_2(\omega_1; a, \bar{b}) = \min_{\omega \in I} F_2(\omega; a, \bar{b}) = k. \quad (3.6.14)$$

Furthermore, in order to find a minimum on the bound of $P(B_\omega)$ we have,

$$\begin{aligned} \min_{\omega} P(B_\omega) &\leq \frac{1}{\max_{\omega} (F_1 + F_2)} \\ &\leq \frac{1}{\max_{\omega} (F_1 + \min_{\omega \in I} F_2)} \\ &= \frac{1}{\max_{\omega} F_1 + \min_{\omega \in I} F_2} \\ &= \frac{1}{\max_{\omega} \left[\frac{1}{P(C_\omega)} \right] + \min_{\omega \in I} F_2} \\ &= \frac{P(C_{\omega_1})}{1+k P(C_{\omega_1})} \end{aligned} \quad (3.6.15)$$

where $P(C_\omega)$ is the minimum bound on the P-condition number of the matrix C_ω , by theorem (B.1) of Appendix B and where,

$$\omega_1 = \begin{cases} \frac{2}{1 + \sqrt{1 - 2M + 4\bar{b}}} = \omega_M & , \text{ if } M \leq 4\bar{b} \\ \frac{2}{1 + \sqrt{1 - 4\bar{b}}} = \omega^* & , \text{ if } M > 4\bar{b}. \end{cases} \quad (3.6.16)$$

From the analysis in the Appendix B concerning $P(C_\omega)$ the interval I may lie in the following interval

$$I \subseteq \begin{cases} (1, \omega^*] & , \text{ if } \bar{b} < \frac{1}{4} \text{ and } \omega_1 = \omega_M \\ [\omega^*, 2) & , \text{ if } \bar{b} < \frac{1}{4} \text{ and } \omega_1 = \omega^* \\ (1, 2) & , \text{ if } \bar{b} > \frac{1}{4}. \end{cases} \quad (3.6.17)$$

Thus, a realistic choice of I is

$$I = \begin{cases} (\omega_M, \omega^*] & , \text{ if } M \leq 4\bar{b} \\ [\omega^*, 2) & , \text{ if } M > 4\bar{b}. \end{cases} \quad (3.6.18)$$

By (3.6.14) we have that

$$k = \begin{cases} \frac{(1-M)\omega_M^2 \bar{b} \rho}{(2-\omega_M)(2-\omega_M)} & , \text{ if } M \leq 4\bar{b} \\ \frac{\omega^* \bar{b} \rho}{2} & , \text{ if } M > 4\bar{b} \end{cases} \quad (3.6.19)$$

and (3.6.9) and (3.6.10) are valid since $\rho \in (0, 1]$.

We can modify the bound on $P(B_{\omega_1})$ given by (3.6.10) to yield

$$P(B_{\omega_1}) \leq \begin{cases} \frac{\frac{1}{2} (1 + \frac{1}{\sqrt{1-M}})}{1 + \frac{M}{8\sqrt{1-M}} \rho} & , \text{ if } \bar{b} \leq \frac{M}{4} \\ \frac{\frac{1}{2} (1 + \sqrt{\frac{2}{1-M}})}{1 + \frac{1}{8\sqrt{2}(1-M)} \rho} & , \text{ if } \frac{M}{4} < \bar{b} \leq \frac{1}{4} \\ \frac{\frac{1}{2} (1 + \gamma^{-1} \sqrt{\frac{2}{1-M}})}{1 + \frac{\bar{b}\gamma}{4} \rho} & , \text{ if } \bar{b} > \frac{1}{4} \end{cases} \quad (3.6.20)$$

where $\rho \in (0, 1]$ and

$$\gamma = \left(1 + \frac{2(\bar{b}-\frac{1}{4})}{1-M} \right)^{-\frac{1}{2}} \quad (3.6.21)$$

As $M \rightarrow 1^-$ the bound on $P(B_{\omega_1})$ is a number of the interval

$$\left[\begin{array}{ll} \left(4, \frac{1}{2} \left(1 + \frac{1}{\sqrt{1-M}} \right) \right) & , \bar{b} \leq \frac{M}{4} \\ \left(8, \frac{1}{2} \left(1 + \sqrt{\frac{2}{1-M}} \right) \right) & , \text{if } \frac{M}{4} < \bar{b} \leq \frac{1}{4} \\ \left(\frac{\frac{1}{2} \left(1 + \sqrt{\frac{1}{\bar{b}-\frac{1}{4}}} \right)}{1 + \frac{\bar{b}\gamma}{4}} , \frac{1}{2} \left(1 + \sqrt{\frac{1}{\bar{b}-\frac{1}{4}}} \right) \right) & , \text{if } \bar{b} > \frac{1}{4} . \end{array} \right. \quad (3.6.22)$$

From Theorems (2.6.14) and (2.7.11) we conclude that the SSOR, PSD and PDF method possess the same estimated values of ω_1 whereas it is important to remember the reverse order of the equations of the PDF method. This influences the upper bound \bar{b} of UL as confirmed by table (4.2.T1).

3.7 ON THE RATE OF CONVERGENCE

For the PDF method with τ in the form (3.5.62) the spectral radius of the iteration matrix $H_{\omega_{0,1}}$ can be conveniently written as

$$S(H_{\omega_{0,1}}) = 1 - \frac{2}{1 + P(B_{\omega_{0,1}})} \quad (3.7.1)$$

whereby $\omega_{0,1}$ we mean either ω_0 or ω_1 .

If $P(B_{\omega_{0,1}}) \gg 1$, then

$$S(H_{\omega_{0,1}}) \doteq 1 - \frac{2}{P(B_{\omega_{0,1}})} \quad (3.7.2)$$

hence the asymptotic rate of convergence is given by

$$R_{\infty}(H_{\omega_{0,1}}) \doteq \frac{2}{P(B_{\omega_{0,1}})} \quad (3.7.3)$$

and the reciprocal rate of convergence by

$$RR(H_{\omega_{0,1}}) \doteq \frac{P(B_{\omega_{0,1}})}{2} \quad (3.7.4)$$

The a-priori knowledge of the eigenvalues of B_{ω} implies a dependence upon ρ of the estimated rate of convergence. Thus, comparing the bounds on $RR(H_{\omega_1})$ with $RR(B_{\rho})$ by (2.4.14) and taking into account that $\rho \in (0,1]$ we have

$$\frac{RR(H_{\omega_1})}{\sqrt{RR(B_{\rho})}} \leq \begin{cases} \frac{1}{2\sqrt{2}} (1-k_1) & , \text{ if } \bar{b} \leq \frac{M}{4} \\ \frac{1}{2} (1-k_2) & , \text{ if } \frac{M}{4} < \bar{b} \leq \frac{1}{4} \\ \frac{1}{2} \gamma^{-1} (1-k_3) & , \text{ if } \bar{b} > \frac{1}{4} \end{cases} \quad (3.7.5)$$

where $k_1, k_2, k_3 \in (0,1)$ depending on M, ρ and \bar{b} .

With respect now to the PSD method by comparing (2.7.18) and (3.7.5) we obtain that

$$RR(H_{\omega_1}) \leq RR(D_{\tau_1, \omega_1}) \quad (3.7.6)$$

or

$$\frac{RR(H_{\omega_1})}{RR(D_{\tau_1, \omega_1})} \doteq 1-k, \quad k \in (0,1) \quad (3.7.7)$$

However, we are not able to specify exactly an a-priori asymptotic number to the ratio (3.7.7) since it is not possible to have an a-priori knowledge of ρ ⁽¹⁾, but the numerical results of Chapter 4 justify our assertion, of the existence of k .

⁽¹⁾ A rough approach gives as ρ one of the elements of the set $\{c, ch, ch^2\}$ where h is the net mesh size.

Since now

$$RR(L_{\omega_b}) \leq \sqrt{RR(B_{\rho})} \quad (3.7.8)$$

where L_{ω_b} is the iteration matrix of the SOR method we establish that

$$\frac{RR(H_{\omega_1})}{RR(L_{\omega_b})} = \begin{cases} \frac{1}{\sqrt{2}}(1-k_1) & , \text{ if } \bar{b} < \frac{M}{4} \\ (1-k_2) & , \text{ if } \frac{M}{4} < \bar{b} < \frac{1}{4} \\ \gamma^{-1}(1-k_3) & , \text{ if } \bar{b} > \frac{1}{4} \end{cases} \quad (3.7.9)$$

where $k_i \in (0,1)$ and A is consistently ordered matrix.

Moreover, asymptotically for the SSOR method we have by (2.6.22)

that

$$\frac{RR(H_{\omega_1})}{RR(G_{\omega_1})} = \begin{cases} \frac{1}{2}(1-k_1) & , \text{ if } \bar{b} < \frac{1}{4} \\ \frac{1}{2}(1-k_2) & , \text{ if } \frac{M}{4} < \bar{b} < \frac{1}{4} \\ \frac{1}{2}(1-k_3) & , \text{ if } \bar{b} > \frac{1}{4} \end{cases} \quad (3.7.10)$$

where $k_i \in (0,1)$.

3.8 A NOTE ON THE DIRICHLET PROBLEM FOR THE ONE-DIMENSIONAL POISSON EQUATION

We will consider a simple but typical problem of Mathematical Physics and use it to illustrate the effectiveness of our method.

To begin with, let us consider the problem

$$-\frac{d^2 u}{dx^2} = F \quad (0 < x < 1) \quad (3.8.1)$$

$$u(0) = a, \quad u(1) = b,$$

where F represents the source term and a and b are given constants.

The difference analogue of (3.8.1) with the second order approximation, can be written in the matrix form,

$$Au = g \quad (3.8.2)$$

where in general

$$A = \left\{ \begin{array}{c} \dots, -\alpha, 1, -\alpha, \dots \\ \dots \end{array} \right\} \quad (3.8.3)$$

and where $\frac{1}{2} < \alpha < \frac{1}{2}$.

The preconditioned system

$$B_{\omega} u = g_{\omega} \quad (3.8.4)$$

by means of M having the form (3.1.2) has the optimistic p -condition number which can be easily ascertained from, (3.8.5)

Proposition (3.8.5): If A has the form (3.8.3) then

$$P(B_{\omega}) \leq \begin{cases} 1 & , \text{ if } \alpha^2 \leq \frac{1}{4} \\ k P(A) & , \text{ if } \alpha^2 > \frac{1}{4} \end{cases} \quad (3.8.6)$$

where k is a positive constant such that $k < 1$ and where $B_{\omega} = M^{-1}A$ is the preconditioned matrix of the PDF method.

Proof:

Let μ_1, μ_2 be two eigenvalues of B_{ω} and let v_1, v_2 be two eigenvectors associated with them respectively. Then, if we suppose that

$$\mu_1 = \mu_2 \quad \text{while} \quad \langle v_1, Bv_1 \rangle = a(v_1) \neq a(v_2) = \langle v_2, Bv_2 \rangle \quad (3.8.7)$$

by means of the relation (3.6.16) we have

$$\frac{1-a(v_1)}{1-\omega a(v_1)+\omega^2 \alpha^2} = \frac{1-a(v_2)}{1-\omega a(v_2)+\alpha^2 \omega^2} \quad (3.8.8)$$

which equivalently implies

$$[a(v_2)-a(v_1)] (\omega^2 \alpha^2 - \omega + 1) = 0 \quad (3.8.9)$$

or

$$\omega^* = \frac{2}{1+\sqrt{1-4\alpha^2}} \quad , \text{ if } \alpha^2 \leq \frac{1}{4} \quad (3.8.10)$$

where $\omega^* \in (0, 2)$. Therefore (3.8.8) is valid for (3.8.10) and subsequently $P(B_{\omega^*}) = 1$.

On the other hand, if $\alpha^2 \geq \frac{1}{4}$ since (3.6.16) with $b(v) = \alpha^2$ is a decreasing function with respect to $a(v)$, then any eigenvalue μ of B_ω lies on

$$\left[\frac{1-M}{1-\omega M + \omega^2 \alpha^2}, \frac{1-m}{1-\omega m + \omega^2 \alpha^2} \right] \tag{3.8.11}$$

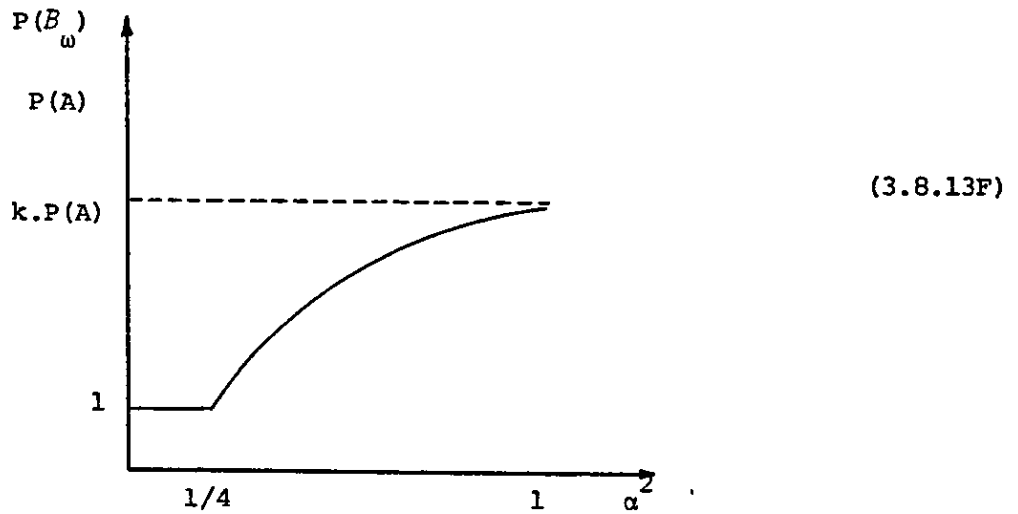
and hence the P-condition number of B_ω is minimized at $\omega_M = \frac{1}{\sqrt{\alpha^2}}$.
Therefore,

$$P(B_{\omega_M}) \leq P(A) \frac{2\sqrt{\alpha^2} - M}{2\sqrt{\alpha^2} - m} \tag{3.8.12}$$

where the quantity $\frac{2\sqrt{\alpha^2} - M}{2\sqrt{\alpha^2} - m}$ is less than one.

We have therefore, proved (3.8.6). ■

A schematic typical representation of the behaviour of the P-condition number of B_ω , with respect to α^2 , is given in (3.8.13F).



3.9 A CERTAIN CONDITIONING MATRIX WHEN A POSSESSES PROPERTY A

A particular case with another approach to the conditioning matrix M , is now investigated where we assume that σ_1 ordering has been applied which leads A to form (D.15). Actually, the finite difference equations can be reordered by choosing all the points in which $(i+j)$ is even, and then all the points in which $(i+j)$ is odd so that the coefficient matrix A has the simple partitioned form,

$$A = \begin{bmatrix} I & -U^* \\ -L^* & I \end{bmatrix} \quad (3.9.1)$$

where $L^*=U^*$.

Consider the conditioning matrix

$$M_2 = (I-\omega\tilde{U})(I-\omega\tilde{L}) \quad (3.9.2)$$

where

$$I-\omega\tilde{U} = \begin{bmatrix} I & -\omega U^* \\ & I & -\omega U^* \end{bmatrix} \quad (3.9.3)$$

and

$$I-\omega\tilde{L} = \begin{bmatrix} I & \\ -\omega L^* & I \\ & & -\omega L^* \end{bmatrix} \quad (3.9.4)$$

Therefore, M_2 has the form

$$M_2 = \begin{bmatrix} I+\omega^2 U^* L^* & -\omega U^* \\ -\omega L^* & I+\omega^2 U^* L^* \end{bmatrix}. \quad (3.9.5)$$

As we have seen the preconditioned matrix $B_\omega^{(2)}$ is given by

$$B_\omega^{(2)} = M_2^{-1} A. \quad (3.9.6)$$

whereas the iterative scheme has the form (3.6.3).

Lemma (3.9.7): Let A have the form (3.9.1). If μ is an eigenvalue

of the preconditioned matrix $B_\omega^{(2)}$ such that

$$\mu \neq \frac{\omega-1}{\omega} + \sqrt{\left(\frac{\omega-1}{\omega}\right)^2 + \frac{1}{\omega^2}} \quad (3.9.8)$$

and if

$$\delta [\mu^2 \omega^2 + 2\omega\mu(1-\omega) - 1] = (\mu-1)^2 \quad (3.9.9)$$

then δ is an eigenvalue of B^2 .

Proof:

Let us assume that A has the form, $A=I-L-U$, where

$$L = \begin{bmatrix} 0 & 0 \\ L^* & 0 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 0 & U^* \\ 0 & 0 \end{bmatrix} \quad (3.9.10)$$

Let v be an associated eigenvector to the eigenvalue μ of $B_{\omega}^{(2)}$.

Then

$$B_{\omega}^{(2)} v = \mu v \quad (3.9.11)$$

or

$$Av = \mu M_2 v \quad (3.9.12)$$

or

$$Av = \mu [I - \omega(U+L) + \omega^2 UL + \omega^2 UL] v \quad (3.9.13)$$

which implies

$$(\omega\mu - 1)Bv - 2\omega^2 \mu ULv = (\mu - 1)v. \quad (3.9.14)$$

Since A has the form (3.9.1) then,

$$B = \begin{bmatrix} 0 & U^* \\ L^* & 0 \end{bmatrix} \quad (3.9.15)$$

and

$$UL = \begin{bmatrix} U^*L^* & 0 \\ 0 & 0 \end{bmatrix} \quad (3.9.16)$$

Thus (3.9.14) becomes

$$\begin{bmatrix} -2\omega^2 \mu U^*L^* & (\omega\mu - 1)U^* \\ (\omega\mu - 1)L^* & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = (\mu - 1) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

where we assume that $v = [v_1, v_2]^T$ is similarly partitioned to A .

Furthermore,

$$\left. \begin{aligned} -2\omega^2 \mu U^*L^*v_1 + (\omega\mu - 1)U^*v_2 &= (\mu - 1)v_1 & (a) \\ (\omega\mu - 1)L^*v_1 + &= (\mu - 1)v_2 & (b) \end{aligned} \right\} \quad (3.9.17)$$

Multiplying (a) by $(\mu - 1)$ and (b) by $(\omega\mu - 1)U^*$ and substituting the first into the second we have

$$[\omega\mu(2\omega - 2 - \omega\mu) + 1]U^*L^*v_1 = (\mu - 1)^2 v_1. \quad (3.9.18)$$

Similarly,

$$[\omega\mu(2\omega - 2 - \omega\mu) + 1]L^*U^*v_2 = (\mu - 1)^2 v_2. \quad (3.9.19)$$

Since from (3.9.8), the relations (3.9.18) and (3.9.19) become

$$\left. \begin{aligned} U^*L^*v_1 &= \frac{(\mu-1)^2}{\omega\mu(2\omega-2-\omega\mu)+1} v_1 \\ L^*U^*v_2 &= \frac{(\mu-1)^2}{\omega\mu(2\omega-2-\omega\mu)+1} v_2 \end{aligned} \right\} \quad (3.9.20)$$

Since now

$$B^2 v = \begin{bmatrix} U^*L^* & 0 \\ 0 & L^*U^* \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{(\mu-1)^2}{\omega\mu(2\omega-2-\omega\mu)+1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (3.9.21)$$

we conclude that v is an eigenvector of B^2 and (3.9.9) is valid. ■

Following closely to a proof of Kahan [1958] concerning the eigenvalues of the SSOR iteration matrix and those of B^2 we set,

Proposition (3.9.22): Let A have the form (3.9.1). If η is an eigenvalue of the iteration matrix $H^{(2)}$ of (3.6.3) such that

$$\eta \neq \frac{1}{\omega} + \sqrt{\frac{1}{\omega^2} + \left(\frac{\omega-1}{\omega}\right)^2} \quad (3.9.23)$$

and if

$$\eta^2 = [(\omega-1)^2 + \omega\eta(2-\omega\eta)]\delta, \quad (3.9.24)$$

then δ is an eigenvalue of B^2 .

Proof:

The proof is similar to the one followed in Lemma (3.9.7). ■

We now proceed to the following theorem concerning the minimization of the P-condition number of the matrix $B_\omega^{(2)}$.

Theorem (3.9.25): If A possesses property A and σ_1 ordering then the P-condition number of the matrix $B_\omega^{(2)}$ is minimized if we let

$$\omega_1 = \begin{cases} \frac{1}{M(B)} = \omega_M, & \text{if } M(B) \geq \frac{1}{2} \\ \frac{1}{1 + \sqrt{1 - 4M(B)^2}} = \omega^*, & \text{if } M(B) \leq \frac{1}{2} \end{cases} \quad (3.9.26)$$

and the corresponding value of $P(B_{\omega_1}^{(2)})$ is given by

$$P(B_{\omega_1}^{(2)}) \leq \begin{cases} \frac{1}{3}P(A), & \text{if } \omega_1 = \omega_M \\ 1, & \text{if } \omega_1 = \omega^* \end{cases} \quad (3.9.27)$$

Proof:

Let A have the form $A=I-L-U$, where (3.9.10) is valid for the matrices L and U. Let μ be an eigenvalue of $B_{\omega}^{(2)}$ and v an associated eigenvector.

Since now, $B_{\omega}^{(2)}v = \mu v$, we have that

$$(A - \mu M_2)v = 0 \quad (3.9.28)$$

or assuming in v a partition similar to that of A we get from (3.9.5) and (3.9.28)

$$\begin{bmatrix} (1-\mu)I - \mu\omega^2 U^*L^* & (\mu\omega - 1)U^* \\ (\mu\omega - 1)L^* & (1-\mu)I - \mu\omega^2 U^*L^* \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0 \quad (3.9.29)$$

which simplifies to

$$\left. \begin{aligned} [(1-\mu)I - \mu\omega^2 U^*L^*]v_1 + (\mu\omega - 1)U^*v_2 &= 0 \\ (\mu\omega - 1)L^*v_1 + [(1-\mu)I - \mu\omega^2 U^*L^*]v_2 &= 0 \end{aligned} \right\} \quad (3.9.30)$$

Eliminating v_1 from the equations of (3.9.30), this results in

$$(1-\mu\omega)^2 U^*L^*v_2 - [(1-\mu)^2 I + \mu^2 \omega^4 (U^*L^*)^2 - 2(1-\mu)\mu\omega^2 U^*L^*]v_2 = 0 \quad (3.9.31)$$

Since the non-zero eigenvalues of $(L+U)$ occur in pairs

$\pm b_i$ ($i=1,2,\dots,r$) where r is less than or equal to the number of rows in L^* or U^* , the eigenvalues of U^*L^* are precisely b_i^2 ($i=1,2,\dots,r$) or zero.

Therefore, since $v_2 \neq 0$, we must have

$$(1-\mu\omega)^2 b_i^2 - [(1-\mu)^2 I + \mu^2 \omega^4 b_i^4 - 2(1-\mu)\mu\omega^2 b_i^2] = 0, \quad (i=1,2,\dots,r) \quad (3.9.32)$$

Now, the P-condition number of the matrix $B_{\omega}^{(2)}$ is given by the expression

$$P(B_{\omega}^{(2)}) = \frac{M(B_{\omega})}{m(B_{\omega})} \quad (3.9.33)$$

and is obtained by taking $i=1$ in both choices of solution to (3.9.32) or in its transformed form

$$[(1+b_1) - \mu(\omega^2 b_1^2 + \omega b_1 + 1)] [(b_1 - 1) + \mu(\omega^2 b_1^2 - \omega b_1 + 1)] = 0 \quad (3.9.34)$$

Thus,

$$\begin{aligned} P(B_{\omega}^{(2)}) &= \left(\frac{1+b_1}{\omega^2 b_1^2 + \omega b_1 + 1} \right) \left(\frac{\omega^2 b_1^2 - \omega b_1 + 1}{1-b_1} \right) \\ &= \left(\frac{1+b_1}{1-b_1} \right) \left(\frac{\omega^2 b_1^2 - \omega b_1 + 1}{\omega^2 b_1^2 + \omega b_1 + 1} \right). \end{aligned} \quad (3.9.35)$$

Now, if $b_1 \geq \frac{1}{2}$, $P(B_{\omega})$ receives its minimum value with respect to ω at the point $\omega_M = \frac{1}{b_1}$, thus,

$$P(B_{\omega_M}^{(2)}) = \frac{1}{3} P(A) \quad (3.9.36)$$

where $P(A) = \frac{1+b_1}{1-b_1}$ is the P-condition number of A.

On the other hand, if $b_1 \leq \frac{1}{2}$, a minimum occurs at the point

$$\omega^* = \frac{2}{1 + \sqrt{1 - 4b_1^2}}. \quad \text{In that case } P(B_{\omega^*}^{(2)}) = 1.$$

Finally, we conclude that (3.9.26) and (3.9.27) are valid, where $M(B) = b_1$. ■

Note that the above approach for determining a bound on $P(B_{\omega}^{(2)})$ is similar to that followed by Evans [1968] concerning the classical scheme of preconditioning methods, when A has property A.

The asymptotic rate of convergence when A possesses property A and $M = M_2$ for $P(B_{\omega}^{(2)}) \gg 1$ is

$$R_{\infty}(H_{\omega_M}^{(2)}) \doteq \frac{2}{P(B_{\omega_M}^{(2)})} \quad (3.9.37)$$

or

$$R_{\infty}(H_{\omega_M}^{(2)}) \doteq \frac{6}{P(A)} \quad (3.9.38)$$

whereas the reciprocal rate of convergence is given by

$$RR(H_{\omega_M}^{(2)}) \doteq \frac{P(A)}{6} \quad (3.9.39)$$

In the case of Property A we have

$$R_{\infty}(B_{\rho^-}) = R_{\infty}(B)$$

and

$$R_{\infty}(B_{\rho^-}) \doteq \frac{1}{2P(A)} \quad .$$

Thus

$$R_{\infty}(H_{\omega_M}^{(2)}) \doteq 12 R_{\infty}(B_{\rho^-}).$$

This improvement is really only of academic interest and does not convey the use of this form of preconditioning for practical problems.

CHAPTER 4

EXPERIMENTAL RESULTS

The sequence of values of the (solution) vector after the successive cycles of the iteration must eventually either reach a fixed terminal state or enter a periodic phase...either of these two terminations of the iteration is possible but have no information as to which occur more often.

G.E. FORSYTHE

Six distinct problems will be studied in this section, in order to test the theoretical results obtained in Chapter 3. All the numerical experiments were carried out involving the Dirichlet problem with the differential equation

$$\frac{\partial}{\partial x} \left(A \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(C \frac{\partial u}{\partial y} \right) = 0 \quad (4.0.1)$$

on the unit square

$$\{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1\} \quad (4.0.2)$$

with zero boundary conditions.

The number of iterations required to solve the systems of equations using SOR, SSOR and PSD methods are compared with the number of iterations required when the same problems are solved with the method developed in this thesis, i.e. the PDF method.

4.1 OPTIMA PARAMETERS

We discuss now, the determination of the optima parameters ω_0 and $P(B_{\omega_0})$, where by the optimum ω we mean the value of ω which minimizes the P-condition number of B_{ω} .

At first we deal with the searching of optima parameters for the iterative scheme

$$u^{(n+1)} = u^{(n)} - M^{-1} A u^{(n)} + M^{-1} b, \quad n=0,1,\dots \quad (4.1.1)$$

evidently without the acceleration of parameter τ of (3.5.3).

That method as we can easily verify by Theorem (2.2.5) is convergent for any ω in the range, $1 - \frac{\sqrt{2}}{2} < \omega < 1 + \frac{\sqrt{2}}{2}$, where we assume that A is positive definite and M possesses form (3.5.2). By the optimum ω we mean the value of ω which minimizes the spectral radius $S(H_{\omega})$ of

$$H_{\omega} = I - M^{-1} A \quad (4.1.2)$$

The optimum $S(H_{\omega})$ is the largest eigenvalue of H_{ω} for $\omega = \omega_0$, i.e.

$$S(H_{\omega_0}) = \min_{\omega} \max_{x \neq 0} \frac{\langle x, H_{\omega} x \rangle}{\langle x, x \rangle} \quad (4.1.3)$$

Since $S(H_{\omega})$ is an unimodal function of ω where $\omega \in (1 - \frac{\sqrt{2}}{2}, 1 + \frac{\sqrt{2}}{2})$, then using the method of golden section search (vd for example, Himmelbau [1972]) with the power method (vd for example, Wilkinson [1965]),

$$\left. \begin{aligned} y^{(m+1)} &= H_{\omega} z^{(m)}, \quad m=0,1,\dots \\ z^{(m+1)} &= \frac{y^{(m+1)}}{\max(y^{(m+1)})} \end{aligned} \right\} \quad (4.1.4)$$

where $z^{(0)}$ an arbitrary vector ⁽¹⁾ and $\max(x)$ denotes the element of maximum modulus of the vector x , we were able to determine the optimum value of the spectral radius $S(H_{\omega})$, of H_{ω} , at the point $\omega = \omega_0$.

⁽¹⁾ In our case, $z^{(0)} = (1, 1, \dots, 1)$.

Clearly, we have, provided that the first component of the initial vector $z^{(0)}$ is different from zero, that

$$z^{(m)} \rightarrow \frac{x}{\max(x)} \quad (4.1.5)$$

and

$$\max(y^{(m)}) \rightarrow S(H_{\omega}) \quad (4.1.6)$$

where $S(H_{\omega})$ is the dominant eigenvalue of H_{ω} and $\frac{x}{\max(x)}$ the (computed) corresponding eigenvector. Hence, the above process provides simultaneously the spectral radius at its optimum point, i.e. $S(H_{\omega_0})$.

The scheme (4.1.1) then applied with a guess vector $u^{(0)} = (1, 1, \dots, 1)$ and the procedure was terminated when the inequality $\|u^{(n)}\|_{\infty} \leq 10^{-6}$ was satisfied.

We present the number of iterations n_I of six problems together with the optima values of ω and $S(H_{\omega})$ i.e. ω_0 and $S(H_{\omega_0})$ given that the terminated criterion for the golden section search was $(.618)^{n_G^{-1}} \leq 10^{-3}$, namely the a-priori number of functional evaluations to reduce the initial interval of ω to 10^{-3} , was $n_G = 10$.

The results of the numerical experiments are given in Table (4.1.T1).

TABLE (4.1.T1)

OPTIMA PARAMETERS ω_0 AND $S(H_{\omega_0})$ OBTAINED BY THE POWER METHOD

COMBINED WITH A GOLDEN SECTION SEARCH

PROBLEM	COEFFICIENTS	h^{-1}	ω_0	$S(H_{\omega_0})$	n_I
I	$A=C=1$	20	1.6465	.7032	42
		40	1.6854	.8806	121
		60	1.6969	.9432	246
II	$A=C=e^{10(x+y)}$	20	1.5383	.4508	19
		40	1.6437	.7051	48
		60	1.6549	.8491	105
III	$A = \frac{1}{1+2x^2+y^2}$ $C = \frac{1}{1+x^2+2y^2}$	20	1.6478	.7109	43
		40	1.6861	.8911	123
		60	1.6964	.9454	254
IV	$A=C = \begin{cases} 1+x, & 0 \leq x \leq \frac{1}{2} \\ 2-x, & \frac{1}{2} \leq x \leq 1 \end{cases}$	20	1.6459	.7212	46
		40	1.6455	.8903	126
		60	1.6963	.9451	248
V	$A = 1 + 4 x - \frac{1}{2} ^2$ $C = \begin{cases} 1, & 0 \leq x \leq \frac{1}{2} \\ 9, & \frac{1}{2} \leq x \leq 1 \end{cases}$	20	1.6493	.7254	44
		40	1.6850	.8878	120
		60	1.6956	.9429	246
VI	$A = 1 + \sin \frac{\pi(x+y)}{2}$ $C = e^{10(x+y)}$	20	1.5600	.4687	20
		40	1.6459	.7098	48
		60	1.6714	.8400	90

We now discuss the accelerated version of (4.1.1), namely the PDF iterative technique, in the sense of searching for the optima parameters.

Evans [1968] has shown that the P-condition number of the classical preconditioned matrix is an unimodal function of ω . Since our preconditioned matrix, B_ω , preserves the same property then by a simultaneous process using the method of golden section search⁽¹⁾, again, with the power method⁽²⁾ applied twice in B_ω to determine at first $M(B_\omega)$, the largest eigenvalue of B_ω and secondly $m(B_\omega)$, the smallest one of B_ω at their optimum point $\omega=\omega_0$ we were able to determine the optimum $P(B_\omega)$, in the sense of its minimization with respect to ω , i.e.

$$P(B_{\omega_0}) = \min_{\omega} \frac{\max_{x \neq 0} \frac{\langle x, B_\omega x \rangle}{\langle x, x \rangle}}{\min_{x \neq 0} \frac{\langle x, B_\omega x \rangle}{\langle x, x \rangle}} \quad (4.1.7)$$

Since the golden section search loses accuracy in determining the optima ω_0 and $P(B_\omega)$ in its final steps these optima are obtained by using some trial values of ω between the last of the search.

The scheme (3.6.3) is then applied with a guess vector $u^{(0)} = (1, 1, \dots, 1)$ and the procedure was terminated when the inequality $\|u^{(n)}\|_\infty \leq 10^{-6}$ was satisfied. Our experiments were carried out involving (4.0.1) on (4.0.2) with zero boundary conditions, whereas the coefficients, $A(x, y)$ and $C(x, y)$ used, can be found in Table (4.1.T1).

The results of the numerical experiments are given in Table (4.1.T3) where τ_0 has been computed from (3.5.60) using the optima values, $m(B_{\omega_0})$ and

(1) The terminated criterion was, again, $(.618)^{n_G-1} \leq 10^{-3}$.

(2) The power method was applied in similar fashion to (4.1.4).

$M(B_{\omega_0})$ of $m(B_{\omega_0})$ and $M(B_{\omega_0})$ respectively and n_{IO} indicates the number of iterations of each problem. Furthermore, SOR, SSOR and PSD have been computed with their optima values. Table (4.1.T2) includes the optima values of the three last methods, given that ω_b for the SOR has been computed by formula (2.6.7), while the SSOR and PSD method are implied with the same optima parameters (vd Missirlis [1978]). Analytically, ω_0 , $S(G_{\omega_0})$, $P(C_{\omega_0})$ and τ_0 are given in Missirlis [1978] and $S(B), \omega_b$ given in Young [1977]. Besides the values of $M(B_{\omega_0})$, $m(B_{\omega_0})$ the number of power iterations, n_p necessary to evaluate these values are tabulated.

As we see in Table (4.1.T3) the number of iterations of the PDF method with optima parameters is almost the same with the iterations required by the PSD with optima parameters, thus PSD preserves its superiority since less work is required per iteration than PDF. Evenmore, PDF, PSD, SSOR with optima parameters require a considerable computing effort for the approach of these parameters, which means that the generation of $\omega_0, P(B_{\omega_0}), P(C_{\omega_0}), S(G_{\omega_0})$ requires the same number of iterations necessary for solving a finite difference problem itself, or even more. We also note that the number of iterations required to satisfy the convergence tolerance 10^{-6} using the aforementioned methods varies approximately as h^{-1} .

TABLE (4.1.T2)

OPTIMA PARAMETERS, USED FOR THE SSOR, PSD AND SOR METHOD

Prob	h^{-1}	ω_0	SSOR	PSD		SOR		
			$S(G_{\omega_0})$	$P(C_{\omega_0})$	τ_0	h^{-1}	$S(B)$	ω_b
I	20	1.7641	.8099	5.2604	.6993	20	.9877	1.7295
	40	1.8750	.9008	10.0806	.4264	40	.9969	1.8547
	60	1.9157	.9343	15.2207	.3031	80	.9992	1.9237
II	20	1.5888	.5876	2.4248	.9251	20	.9576	1.5527
	40	1.7668	.7663	4.2790	.6679	40	.9894	1.7460
	60	1.8386	.8386	6.1958	.5110	80	.9983	1.8902
III	20	1.7652	.8140	5.3763	.6989	20	.9880	1.7326
	40	1.8756	.9031	10.3200	.4254	40	.9970	1.8564
	60	1.9163	.9343	15.2207	.3010	80	.9992	1.9242
IV	20	1.7624	.8088	5.2301	.7031	20	.9882	1.7385
	40	1.8748	.9002	10.0200	.4268	40	.9972	1.8599
	60	1.9143	.9324	14.7929	.3073	80	.9993	1.9260
V	20	1.7479	.8281	5.8173	.7520	20	.9870	1.7233
	40	1.8665	.9105	11.1732	.4574	40	.9968	1.8515
	60	1.9093	.9395	16.5289	.3266	80	.9991	1.9191
VI	20	1.6097	.6065	2.5221	.8998	20	.9576	1.5528
	40	1.7820	.7855	4.4543	.6345	40	.9982	1.7448
	60	1.8490	.8438	6.4020	.4829	80	.9983	1.8907

TABLE (4.1.T3)

OPTIMA PARAMETERS ω_0 AND $P(B_{\omega_0})$ OBTAINED BY THE POWER METHOD COMBINED WITH
A GOLDEN SECTION SEARCH

PROBLEM	h^{-1}	ω_0	$M(B_{\omega_0})$		$m(B_{\omega_0})$		τ_0	$P(B_{\omega_0})$	n_{IO} OPTIMUM			h^{-1}	SOR
			n_p	n_p	PDF	SSOR			PSD				
I	20	1.7642	2.3892	30	.4569	55	.7027	5.2291	37	66	37	20	61
	40	1.8741	4.2312	58	.4185	51	.4301	10.1104	71	134	71	40	121
	60	1.9155	6.1721	73	.4134	54	.3037	14.9301	105	201	107	80	253
II	20	1.5878	1.5263	89	.6331	101	.9262	2.4108	18	24	17	20	50
	40	1.7661	2.4179	92	.5695	85	.6695	4.2449	35	48	30	40	99
	60	1.8377	3.3477	96	.5479	82	.5134	6.1100	51	71	44	80	217
III	20	1.7667	2.4112	39	.4587	40	.6969	5.2566	37	68	38	20	60
	40	1.8760	4.2902	68	.4248	40	.4242	10.0993	71	137	72	40	121
	60	1.9158	6.1934	39	.4149	39	.3027	14.9274	105	205	107	80	252
IV	20	1.7643	2.3903	30	.4249	59	.7104	5.6256	33	66	37	20	59
	40	1.8748	4.2532	56	.3887	49	.4308	10.9421	64	133	70	40	119
	60	1.9143	6.0919	69	.3704	31	.3108	16.4468	102	200	104	80	225
V	20	1.7483	2.2696	27	.3866	121	.7529	5.8706	41	74	41	20	60
	40	1.8662	4.0051	42	.3548	125	.4587	11.2883	79	149	79	40	118
	60	1.9090	5.7568	41	.3444	122	.3278	16.7154	115	224	117	80	274
VI	20	1.6065	1.5820	6	.6198	2	.9083	2.5520	17	28	17	20	41
	40	1.7790	2.5437	7	.5578	2	.6448	4.5602	31	57	32	40	81
	60	1.8464	3.5051	7	.5404	2	.4944	4.4861	47	85	47	80	176

4.2 ESTIMATED PARAMETERS

In order to test the theoretical results obtained in Section(3.7) the problems of Section(4.1) were considered, whereas the numerical experiments were carried out involving estimated parameters. We denote the bound of $P(B_{\omega_1})$ by $P(B_{\omega_1})$ itself. The parameters ω_1 and $P(B_{\omega_1})$ will be referred to as the estimated preconditioning parameters.

Before we can compute ω_1 and $P(B_{\omega_1})$, we must find an upper bound M for $S(B)$ and an upper bound b for $S(UL)$.

Young [1971a] has shown for the Dirichlet problem (4.0.1) that

$$S(B) \leq 1 - \frac{2\underline{A}\sin^2 \frac{\pi}{2I} + 2\underline{C}\sin^2 \frac{\pi}{2I}}{\frac{1}{2}(\overline{A}+\underline{A}) + \frac{1}{2}(\overline{C}+\underline{C}) + \frac{1}{2}(\overline{A}-\underline{A})\cos\frac{\pi}{I} + \frac{1}{2}(\overline{C}-\underline{C})\cos\frac{\pi}{I}} \quad (4.2.1)$$

= M

where the region D_h is included in an $(Ih \times Ih)$ rectangle for a positive I and

where

$$\underline{A} \leq A(x,y) \leq \overline{A}, \quad \underline{C} \leq C(x,y) \leq \overline{C} \quad (4.2.2)$$

in $D_h' + \partial D_h$.

The bound b for $S(UL)$ is determined by a similar process as in Benokraitis [1974] who determined a bound on $S(LU)$. Thus, in order to investigate how the matrix UL operates on a vector defined on the net, let the system be

$$ULu = v. \quad (4.2.3)$$

Let w be an intermediate vector such that

$$Lu = w. \quad (4.2.4)$$

Thus we have

$$w(x,y) = b_3(x,y)u(x-h,y) + b_4(x,y)u(x,y-h) \quad (4.2.5)$$

and

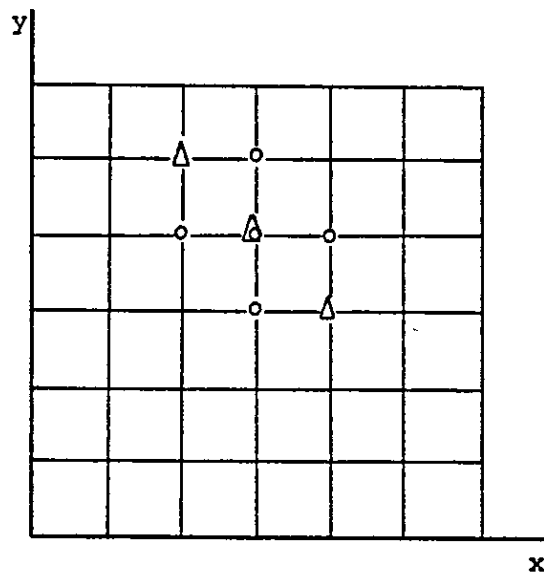
$$\begin{aligned} v(x,y) &= b_1(x,y)w(x+h,y) + b_2(x,y)w(x,y+h) \\ &= b_1(x,y) [b_3(x+h,y)u(x,y) + b_4(x+h,y)u(x+h,y-h)] \\ &\quad + b_2(x,y) [b_3(x,y+h)u(x-h,x+h) + b_4(x,y+h)u(x,y)] \end{aligned}$$

$$\begin{aligned}
&= [b_1(x,y)b_3(x+h,y)+b_2(x,y)b_4(x,y+h)]u(x,y) \\
&\quad + [b_1(x,y)b_4(x+h,y)]u(x+h,y-h) \\
&\quad + [b_2(x,y)b_3(x,y+h)]u(x-h,y+h) \\
&= \gamma_0 u(x,y) + \gamma_1 u(x+h,y-h) + \gamma_2 u(x-h,y+h) \quad (4.2.6)
\end{aligned}$$

for each $(x,y) \in D_h$.

Thus, the operator UL only involves values of $u(x,y)$ at the diagonal points (x,y) , $(x+h,y-h)$ and $(x-h,y+h)$.

Geometrically with notation (4.2.6) it is convenient to indicate the net points involved in the equation (4.2.6) of the operator UL by means of a diagram as given in (4.2.F1) as well as the net points involved in the equation (1.9).



(4.2.F1)

O, represents point involved in (1.9)

Δ, represents point involved in (4.2.6)

A bound for the largest eigenvalue of UL can be obtained from

$$\begin{aligned}
S(UL) &\leq \|UL\|_{\infty} = \max_{(x,y) \in D_h} (\gamma_0 + \gamma_1 + \gamma_2) \\
&= \max_{(x,y) \in D_h} [b_1(x,y) [b_3(x+h,y) + b_4(x+h,y)] \\
&\quad + b_2(x,y) [b_3(x,y+h) + b_4(x,y+h)]] \\
&= \bar{b} \quad (4.2.7)
\end{aligned}$$

Proposition (4.2.8): If $A(x,y)$ and $C(x,y) \in C^{(2)}(D+\partial D)$ then

$$S(UL) \leq \frac{1}{4} + O(h^2), \quad \text{as } h \rightarrow 0. \quad (4.2.9)$$

Proof:

From Chapter 1 we have (1.9) where (1.10) and (1.11) are valid.

Furthermore, (1.11) can be written as

$$S(x,y) = 2[A(x,y)+C(x,y)] + O(h^2) \quad (4.2.10)$$

We seek to determine a bound on $\|UL\|_\infty$ by obtaining a bound on $\gamma_0 + \gamma_1 + \gamma_2$.

From (4.2.7) we have

$$\begin{aligned} S_1 &= b_1(x,y) [b_3(x+h,y) + b_4(x+h,y)] \\ &\quad \frac{A(x+\frac{h}{2},y) [A(x+\frac{h}{2},y) + C(x+h,y-\frac{h}{2})]}{S(x,y)S(x+h,y)} \\ &= \frac{A(x+\frac{h}{2}) [A(x+\frac{h}{2},y) + C(x+h,y-\frac{h}{2})]}{4[A(x,y) + C(x,y) + O(h^2)] [A(x+h,y) + C(x+h,y) + O(h^2)]} \\ &= \frac{A(x+\frac{h}{2}) [A(x+\frac{h}{2},y) + C(x+h,y-\frac{h}{2})]}{4[A(x,y) + C(x,y)] [A(x+h,y) + C(x+h,y)]} + O(h^2) \quad (4.2.11) \end{aligned}$$

Moreover,

$$\begin{aligned} &A(x+\frac{h}{2},y) [A(x+\frac{h}{2},y) + C(x+h,y-\frac{h}{2})] \\ &= A(x,y) [A(x+h,y) + C(x+h,y)] \\ &\quad + (A+\frac{h}{2}Ax+O(h^2)) [A+\frac{h}{2}Ax+C+hCx-\frac{h}{2}Cy+O(h^2)] \\ &\quad - A[A+hAx+c+hCx+O(h^2)], \\ \Delta_1 &= A(x+\frac{h}{2},y) [A(x+\frac{h}{2},y) + C(x+h,y-\frac{h}{2})] - A(x,y) [A(x+h,y) + C(x+h,y)] \\ &= (A+\frac{h}{2}Ax+O(h^2)) [A+\frac{h}{2}Ax+C+hCx-\frac{h}{2}Cy+O(h^2)] - A[A+hAx+C+hCx+O(h^2)] \\ &= \dots \\ &= h \left[\frac{1}{2}ACy + \frac{1}{2}AxC \right] + O(h^2) \\ &= \frac{h}{2} [AxC - ACy]. \end{aligned}$$

Thus,

$$\begin{aligned}
 S_1 &= \frac{A(x,y)}{A[A(x,y)+C(x,y)]} + \frac{h}{2} \frac{Ax-C-Acy}{4[A(x,y)+C(x,y)][A(x+h,y)+C(x+h,y)]} + O(h^2) \\
 &= \frac{A}{4[A+C]} + \frac{h}{8} \frac{Ax-C-Acy}{[A+C]^2} + O(h^2) .
 \end{aligned}$$

Similarly,

$$S_2 = \frac{C}{4[A+C]} - \frac{h}{8} \frac{Ax-C-Acy}{[A+C]^2} + O(h^2) .$$

Hence
$$S_1 + S_2 = \frac{1}{4} + O(h^2)$$

Therefore $\|UL\|_{\infty} \leq \frac{1}{4} + O(h^2)$ and (4.2.9) is valid. ■

The proof of the above proposition is based on Young [1977], who obtained a bound on $S(LU)$.

The result $S(UL) \leq \frac{1}{4} + O(h)$ is significant because, it establishes an order of magnitude improvement of the PDF-SI method (vd Chapter 5) over PDF, SOR, SSOR and PSD.

For the purpose of comparison we include in Table (4.2.T1) the bounds on $S(UL)$ computed by (4.2.7), as well as the bounds on $S(LU)$ given in Missirlis [1978] and Young [1977].

TABLE (4.2.T1)

PROBLEM	h^{-1}	$\bar{b}=\bar{b}$ (UL)	\bar{b} (LU)
I	20	.2500	.2500
	40	.2500	.2500
	60	.2500	.2500
II	20	.2350	.2350
	40	.2461	.2461
	60	.2482	.2483
III	20	.2499	.2506
	40	.2499	.2502
	60	.2500	.2501
IV	20	.2521	.2511
	40	.2501	.2505
	60	.2500	.2500
V	20	.2499	.2499
	40	.2500	.2499
	60	.2500	.2500
VI	20	.2366	.2360
	40	.2469	.2468
	60	.2486	.2493

By (3.6.20) given that $M=1-ch^2+O(h^4)$ we have that $P(B_{\omega_1})=c_1h^{-1}$. Thus by (3.7.3) we obtain the expected result for the asymptotic convergence rate of the PDF method

$$R_{\infty}(H_{\omega_1}) \doteq O(h) \quad (4.2.12)$$

where h the net mesh size. The numerical results of Table (4.2.T2), however indicate that we can also attain $O(h)$ convergence for the PDF method with estimated parameters. In this way, we see that for the SOR, SSOR, PSD and PDF method the number of iterations varies as h^{-1} .

We conclude this section by presenting in Table (4.2.T2) the estimated parameters ω_1 and $P(B_{\omega_1})$ for the same as in Table (4.1.T1) problems on the unit square. The upper bound M for $S(B)$ is given in Young [1977]; the upper bound \bar{b} for $S(UL)$ is computed by (4.2.7). Note that in Problems II and VI we would replace M by $2\sqrt{b}$ since $M > 2\sqrt{b}$. The estimated parameters ω_1, τ_1 and $P(B_{\omega_1})$ computed by (3.6.6), (3.5.61) and (3.6.10)⁽¹⁾ respectively. The same guess vector $u^{(0)}=1$ as previously was used as well as the same terminated criterion. The number of iterations required to carry out the solution is indicated by n_{IE} . We also present the number of iterations of the SSOR and PSD methods with estimated parameters (by Missirlis [1978] the estimated parameters of PSD are presented in Table (4.2.T3)).

As it was expected from the analysis in Section(3.7)of Chapter 3 a significant improvement of the convergence rate of the PDF method is observed in comparison to SOR and SSOR method. The results are fairly optimistic when compared with the ones of PSD. In Figure (4.2.F1-4) we plot the logarithm of the required number of iterations using PDF, PSD, SSOR and SOR versus $\ln h^{-1}$. The slope indicates the approximate $O(h)$ rate of convergence.

(1) In formula (3.6.10) we took ρ equal to h^2 , i.e. $\rho=h^2$, where h the net mesh size.

In Table (4.2.T2) it is shown that the SOR method requires at least 33%, 150%, 26%, 32%, 17% and 128% more iterations than the PDF method with estimated parameters, respectively to Problems I, II, III, IV, V and VI. However, it should be noted that the PDF method requires more work per iteration than the SOR method.

The SSOR method, asymptotically requires about 63%, 25%, 77%, 53%, 95% and 153% more iterations as compared to the PDF method (both the methods using estimated parameters), respectively to the six problems in increasing order. The numerical results which we obtained indicate that the PDF method is much more effective than the PSD method, in view of the Table (4.2.T2) and formula (3.7.7). In case of Problem II, the convergence of the PDF method was slower than the one of the PSD method, while in Case V we have almost the same results for both methods. Moreover, the PSD method requires asymptotically 8%, 28% and 39% more iterations than the PDF method for the Problems I, IV and VI, respectively. Even noting that in Case III when $h=1/20$ and $h=1/40$ the PSD requires 115% and 103% more iterations than the PDF method, whereas for the case of $h=1/60$ the convergence of the PSD is erratic.

TABLE (4.2.T2)

ESTIMATED PARAMETERS ω_1 AND $P(B_{\omega_1})$

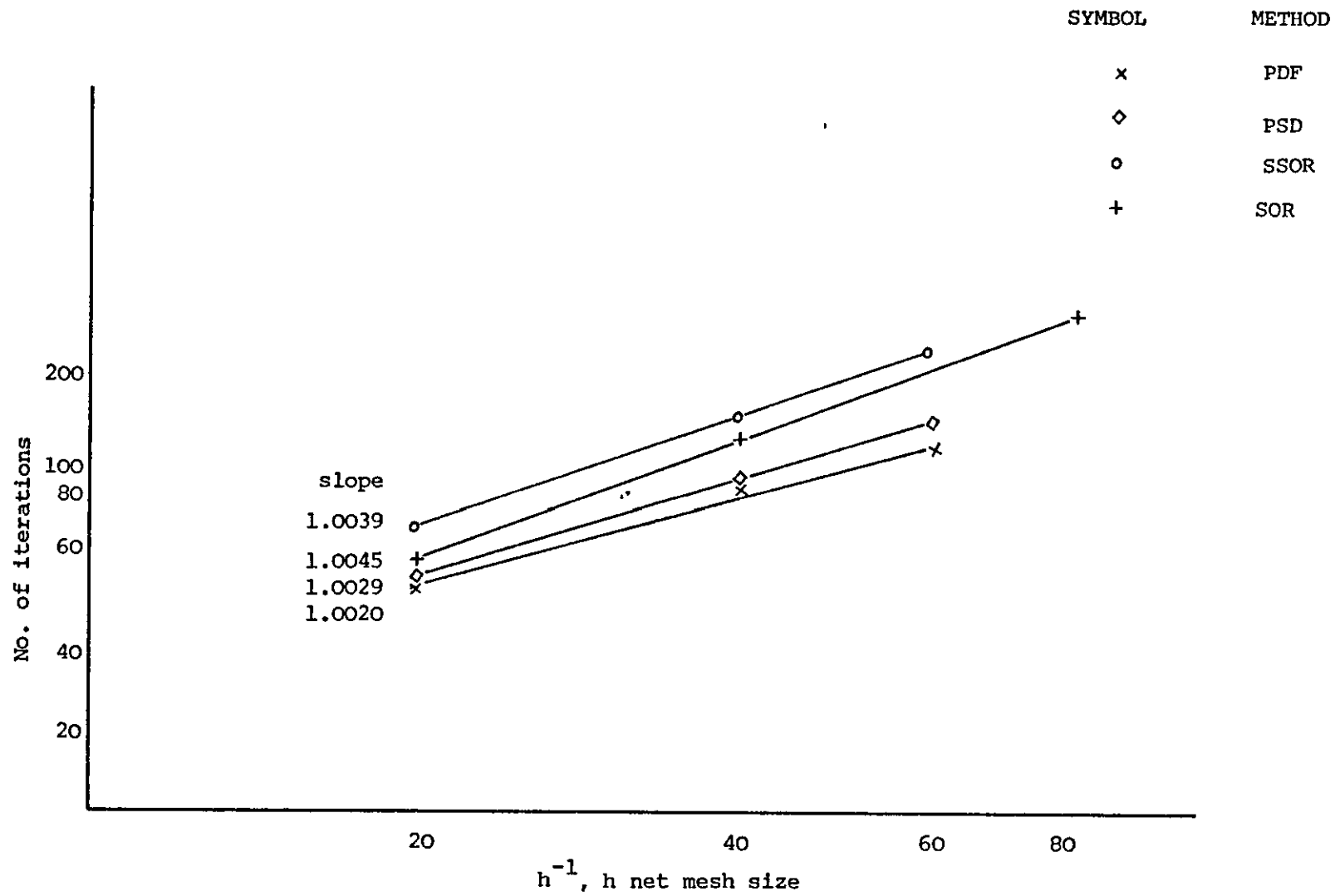
PROBLEM	h^{-1}	\bar{b}	$2\sqrt{b}$	M	ω_1	τ_1	$P(B_{\omega_1})$	$M(B_{\omega_1})$	n_{IE} ESTIMATED		
									PDF	SSOR	PSD
I	20	.2500	1.0000	.9877	1.7288	.8186	6.8760	2.1329	46	68	48
	40	.2500	1.0000	.9969	1.8544	.5018	13.1661	3.7037	91	138	93
	60	.2500	1.0000	.9986	1.8992	.3634	19.3686	5.2334	127	207	137
II	20	.2350	.9695	.9999	1.6065	.9094	2.5377	1.5779	20	28	18
	40	.2461	.9921	.9999	1.7779	.6469	4.4997	2.5293	37	45	33
	60	.2482	.9963	.9999	1.8436	.4993	6.3872	3.4633	54	67	47
III	20	.2499	.9997	.9967	1.8540	.5011	12.4429	3.6943	47	73	101
	40	.2499	.9992	.9992	1.9330	.2478	22.0863	7.7214	96	145	195
	60	.2500	1.0000	.9996	1.9422	.2156	23.4865	8.8976	123	218	-
IV	20	.2521	1.0041	.9914	1.7241	.8033	9.8038	2.1023	41	66	59
	40	.2501	1.0002	.9979	1.8790	.4282	15.8543	4.3935	90	133	114
	60	.2500	1.0000	.9990	1.9143	.3145	22.8430	6.0908	131	200	168
V	20	.2499	.9997	.9977	1.8782	.4282	14.5976	4.3713	60	93	65
	40	.2500	1.0000	.9994	1.9357	.2403	24.3183	8.0155	101	193	100
	60	.2500	1.0000	.9997	1.9521	.1826	36.9602	10.6643	148	288	144
VI	20	.2366	.9728	.9999	1.6240	.8883	2.6561	1.6356	18	36	23
	40	.2469	.9937	.9999	1.7996	.6012	4.9865	2.7709	35	82	47
	60	.2486	.9971	.9999	1.8607	.4552	7.1754	3.8563	51	129	71

TABLE (4.2.T3) (1)

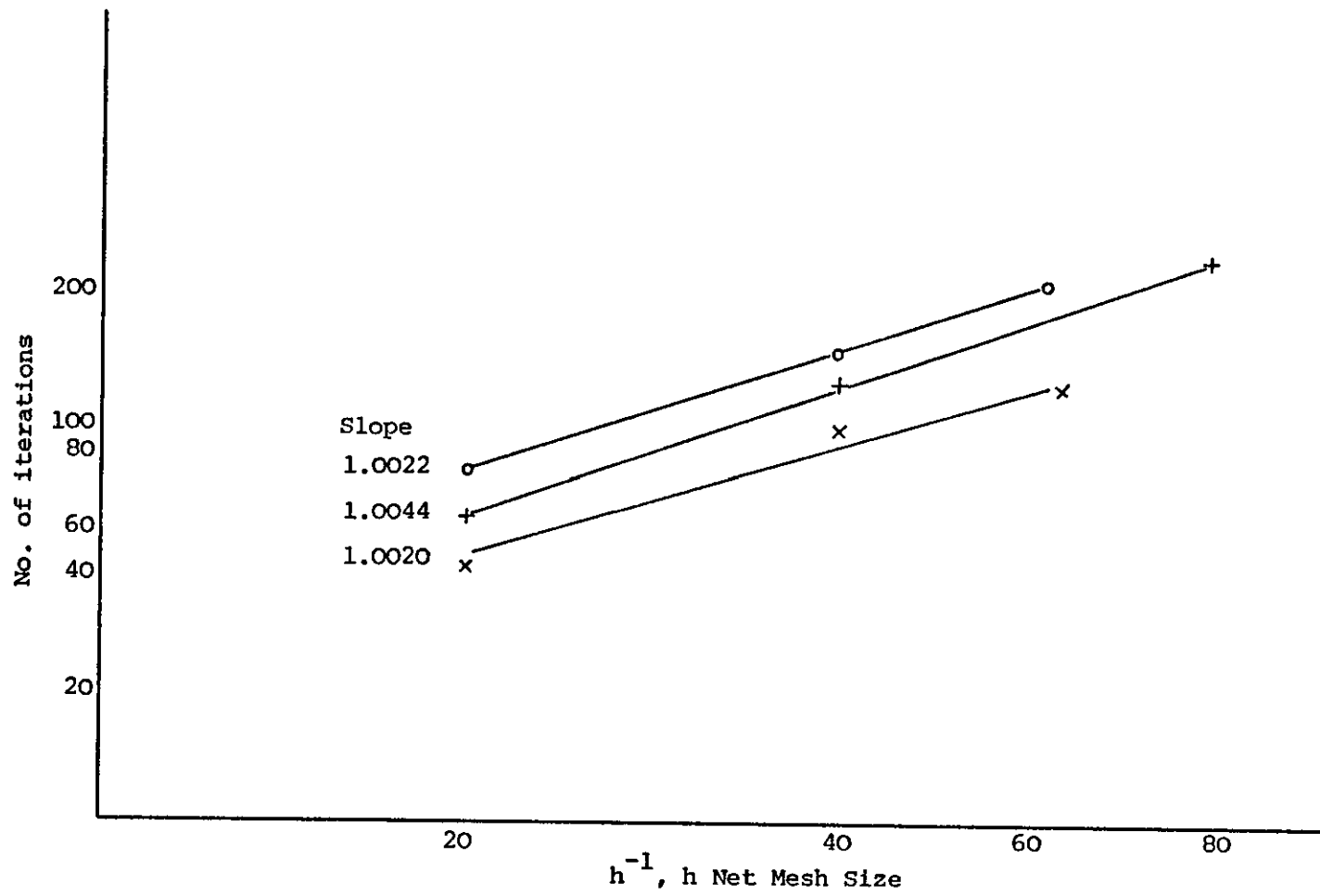
ESTIMATED PARAMETERS FOR THE PSD METHOD

PROBLEM	h^{-1}	ω_1	τ_1	P (PSD)
I	20	1.7287	.8188	6.8727
	40	1.8544	.5021	13.2357
	60	1.9005	.3598	19.6008
II	20	1.6065	.9073	2.5415
	40	1.7788	.6444	4.5208
	60	1.8465	.4914	6.5139
III	20	1.8355	.5661	14.9905
	40	1.9142	.3177	29.5540
	60	1.9420	.2203	44.1015
IV	20	1.7717	.7223	8.3322
	40	1.8790	.4282	16.1660
	60	1.9176	.3034	23.9999
V	20	1.8756	.4379	15.2395
	40	1.9359	.2402	30.0138
	60	1.9568	.1654	44.7804
VI	20	1.6903	.7994	3.2293
	40	1.8475	.4889	6.5567
	60	1.8997	.3463	9.9697

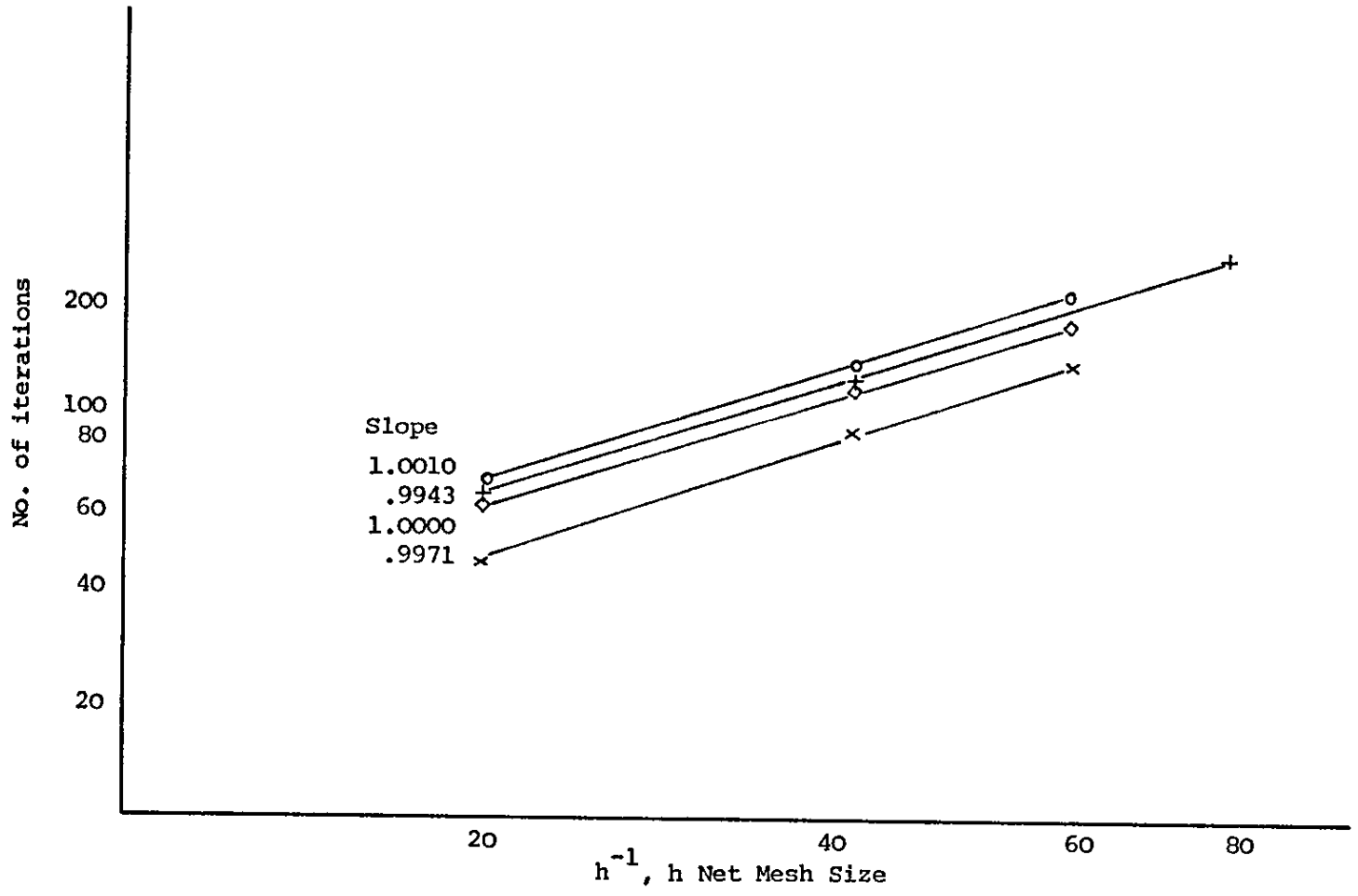
(1) *The results have been obtained by Missirlis [1978]*



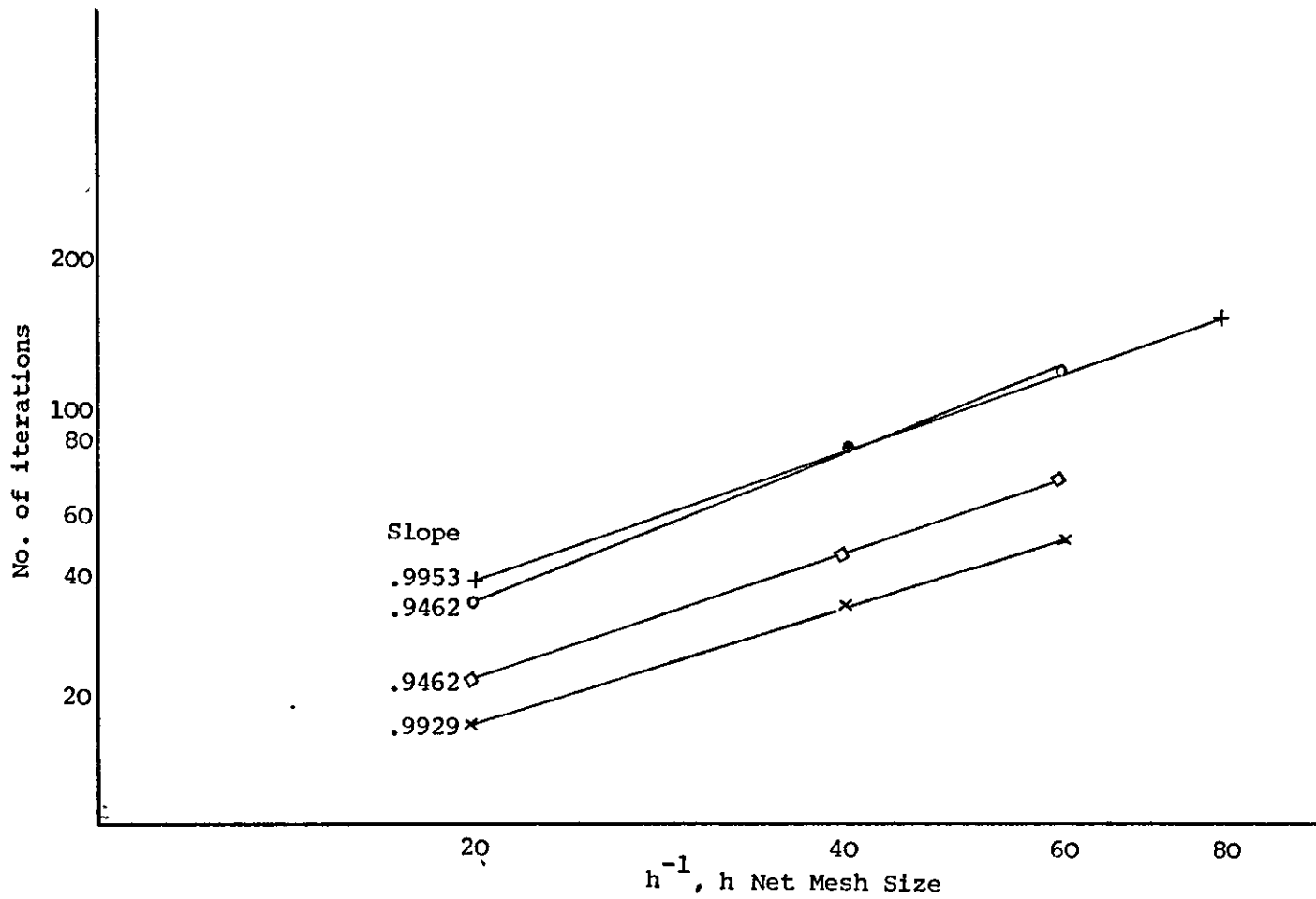
(4.2.F1): Problem 1. The Slope α indicates $O(h^\alpha)$ Rate of Convergence



(4.2.F2): Problem 3



(4.2.F3): Problem 4



(4.2.F4): Problem 6

CHAPTER 5

THE PRECONDITIONING BY DIRECT FACTORIZATION, SEMI-ITERATIVE METHOD

καὶ μετροῦμεν καὶ τό μέγεθος
τῆ κινήσει καὶ τὴν κίνησιν
τῷ μεγέθει

Ἀριστοτέλους φυσικῆς IV, 12, 29⁽¹⁾

As we noticed in Section 3.5 of Chapter 3, the PDF method relies on the spectral characteristics of the operator involved, and the parameters are considered identical for all steps of the algorithm. In that chapter the parameters involved are chosen at each step in such a way that the error vector approaches zero uniformly from the initial approximation as fast as possible. Evenmore, the method requires the transient information of the last two vectors.

⁽¹⁾ *That we do measure linear magnitude by movement and vice versa.*

Aristotle's Physics IV, 12, 29.

5.1 THE PSD-SI METHOD

Let us consider the completely consistent linear stationary iterative method defined by (2.1.3) where $(I-H)$ is non-singular and (2.1.5) is valid. We assume that the eigenvalues μ of H are real and lie in the interval

$$a \leq \mu \leq b < 1 \quad (5.1.1)$$

where a and b are real.

The convergence properties of (2.1.3) method can often be improved by the use of a semi-iterative method based on the (2.1.3).

Varga [1957] and Golub and Varga [1961] have shown that a gain by an order of magnitude is attained if one uses the linear non-stationary method of second degree

$$u^{(n+1)} = \rho_{n+1} [\bar{\rho} (Hu^{(n)} + k) + (1-\bar{\rho})u^{(n)}] + (1-\rho_{n+1})u^{(n-1)}, \quad n=0,1,\dots \quad (5.1.2)$$

where

$$\bar{\rho} = \frac{2}{2-(a+b)}, \quad (5.1.3)$$

$$\rho_1 = 1 \quad (5.1.4)$$

$$\rho_2 = \left(1 - \frac{\sigma^2}{2}\right)^{-1}$$

$$\rho_{n+1} = \left(1 - \frac{\sigma^2}{4} \rho_n\right)^{-1}, \quad n=2,3,\dots$$

and where

$$\sigma = \frac{b-a}{2-(a+b)} = S(H). \quad (5.1.5)$$

For the PDF method (vd Chapter 3) we have

$$\begin{aligned} H_\omega &= I - \tau M^{-1} A \\ &= I - \tau B_\omega, \end{aligned} \quad (5.1.6)$$

where $\tau > 0$ and A is a positive definite matrix.

Since B_ω is a positive definite matrix (vd (3.5.6)) there exists positive numbers $m(B_\omega)$, $M(B_\omega)$ such that

$$0 < m(B_\omega) \leq \mu(B_\omega) \leq M(B_\omega), \quad (5.1.7)$$

hence all the eigenvalues of H_ω are real and lie in the interval

$$a \leq \mu(H_\omega) \leq b < 1 \quad (5.1.8)$$

where $\mu(H_\omega) = \mu(I - \tau B_\omega)$.

Then from (5.1.2) the optimum semi-iterative method corresponding to the PDF method (3.5.3) is

$$u^{(n+1)} = u^{(n-1)} + \rho_{n+1} (u^{(n)} - u^{(n-1)}) + \rho_{n+1} \bar{\rho} M^{-1} (b - Au^{(n)}), \quad n=0,1,\dots \quad (5.1.9)$$

where $\bar{\rho}, \rho_1, \rho_2, \dots$ have been defined by (5.1.3), (5.1.4) and (5.1.5), and by means of the PDF method transform to

$$\bar{\rho} = \frac{2}{m(B_{\omega_0}) + M(B_{\omega_0})} \quad (5.1.10)$$

$$\rho_1 = 1$$

$$\rho_2 = \left(1 - \frac{\sigma^2}{2}\right)^{-1} \quad (5.1.11)$$

$$\rho_{n+1} = \left(1 - \frac{\sigma^2}{4} \rho_n\right)^{-1}, \quad n=2,3,\dots$$

and

$$\sigma = \frac{P(B_{\omega_0}) - 1}{P(B_{\omega_0}) + 1} \quad (5.1.12)$$

where in Chapter 3, $m(B_{\omega_0})$, $M(B_{\omega_0})$ and $P(B_{\omega_0})$ have been defined as the minimum, maximum eigenvalue of B_{ω_0} and their reciprocal ratio, respectively.

Below we will omit the subscript on ω and τ for simplicity.

By (5.1.2) and (3.5.3) the formula for the PDF-SI method is

$$u^{(n)} = P_n(H_\omega) u^{(0)} + k_n, \quad (5.1.13)$$

where

$$P_n(H_\omega) = T_n \left(\frac{2H_\omega - (b+a)I}{b-a} \right) / T_n \left(\frac{2-(b+a)}{b-a} \right) \quad (5.1.14)$$

$$k_n = (I - P_n(H_\omega))A^{-1}b \quad (5.1.15)$$

and T_n the Chebyshev polynomial of degree n .

By Missirlis and Evans [1980] we have that

$$P_n(H_\omega) = T_n \left(\frac{M(B_\omega) + m(B_\omega) - 2B_\omega}{M(B_\omega) - m(B_\omega)} \right) / T_n \left(\frac{M(B_\omega) + m(B_\omega)}{M(B_\omega) - m(B_\omega)} \right) \quad (5.1.16)$$

where we suppose the optimum value of any quantity involved in (5.1.16).

Young [1971] has shown that

$$S(P_n(H_\omega)) = \frac{2r^{h/2}}{1+r^h}, \quad (5.1.17)$$

where

$$r = \left(\frac{\sigma}{1 + \sqrt{1 - \sigma^2}} \right)^2 \quad (5.1.18)$$

$$= \left(\frac{\sqrt{P(B_\omega)} - 1}{\sqrt{P(B_\omega)} + 1} \right)^2,$$

hence as a measure of the rapidity of the convergence we take the asymptotic rate of convergence defined by

$$\begin{aligned} R_\infty(P_n(H_\omega)) &= \lim_{n \rightarrow \infty} \left(-\frac{1}{n} \ln S(P_n(H_\omega)) \right) \\ &= \lim_{n \rightarrow \infty} \left(-\frac{1}{n} \ln \frac{2r^{h/2}}{1+r^h} \right) \\ &= -\frac{1}{2} \ln r. \end{aligned} \quad (5.1.19)$$

Taking $M(B_\omega) \gg m(B_\omega)$, we obtain the asymptotic relation

$$r \doteq 1 - \frac{4}{\sqrt{P(B_\omega)}} \quad (5.1.20)$$

and therefore the asymptotic rate of convergence of the PDF-SI method is given by the formula

$$R_\infty(P_n(H_\omega)) \doteq \frac{2}{\sqrt{P(B_\omega)}} \quad (5.1.21)$$

whereas the reciprocal rate of convergence is given by

$$RR(P_n(H_\omega)) \doteq \frac{\sqrt{P(B_\omega)}}{2}, \quad (5.1.22)$$

which means that the use of the PDF-SI method results in an order of magnitude improvement in the method, as we can ascertain from the formula (3.7.4).

From (3.7.5) and (3.7.9) we see that we have a substantial improvement of the rate of convergence comparing the PDF-SI method, the benchmark method and the SOR method.

5.2 NUMERICAL RESULTS

In order to test the effectiveness of the PDF-SI method for solving elliptic difference equations, we carried out six numerical experiments similar to Table (4.1.T1). 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 starting vector $u^{(0)}$ was the vector with all its components equal to unity. The process terminated after n iterations where n satisfied the stopping procedure

$$\frac{2r^{h/2}}{1+r^h} \leq 10^{-6} \quad (5.2.1) \quad (1)$$

where r is given by (5.1.18).

The procedure was carried out with two classes of parameters, the optima and the ones estimated, by Chapter 4.

(1) When (5.2.1) is satisfied then $\frac{\|u^{(n)} - u\|_{A^2}}{\|u\|_{A^2}} \leq 10^{-6}$.

A detailed analysis on that can be found in Benokraitis [1974].

TABLE (5.2.T1)

NUMBER OF ITERATIONS REQUIRED TO SATISFY STOPPING CRITERION
(5.2.1) USING PDF-SI WITH OPTIMA PARAMETERS

PROBLEM	h^{-1}	ω_0	$\bar{\rho} = \tau_0$	$P(B_{\omega_0})$	OPTIMA PARAMETERS PDF-SI
I	20	1.7642	.7027	5.2291	16
	40	1.8741	.4301	10.1104	23
	60	1.9155	.3037	14.9301	28
II	20	1.5878	.9262	2.4108	10
	40	1.7661	.6695	4.2449	14
	60	1.8377	.5134	6.1100	17
III	20	1.7667	.6969	5.2566	16
	40	1.8760	.4242	10.0993	23
	60	1.9158	.3027	14.9274	28
IV	20	1.7643	.7104	5.6256	17
	40	1.8748	.4308	10.9140	24
	60	1.9143	.3108	16.4468	29
V	20	1.7483	.7429	5.8706	17
	40	1.8662	.4587	11.2883	24
	60	1.9090	.3278	16.7154	30
VI	20	1.6065	.9083	2.5520	10
	40	1.7790	.6448	4.5602	15
	60	1.8464	.4944	6.4861	18

TABLE (5.2.T2)

NUMBER OF ITERATIONS REQUIRED TO SATISFY STOPPING CRITERION

(5.2.1) USING PDF-SI WITH ESTIMATED PARAMETERS

PROBLEM	h^{-1}	ω_1	$\bar{\rho}=\tau_1$	$P(B_{\omega_1})$	ESTIMATED PARAMETERS	
					PDF-SI	PSD-SI
I	20	1.7288	.8186	6.8760	19	19
	40	1.8544	.5018	13.1600	26	26
	60	1.8994	.3634	19.3686	32	32
II	20	1.6065	.9094	2.5377	10	10
	40	1.7779	.6469	4.4997	15	15
	60	1.8436	.4993	6.3872	18	18
III	20	1.8540	.5011	12.4429	25	29
	40	1.9330	.2478	22.0863	34	39
	60	1.9422	.2156	23.4865	36	48
IV	20	1.7241	.8033	9.8038	16	21
	40	1.8790	.4282	15.8543	29	29
	60	1.9143	.3145	22.8430	35	36
V	20	1.8782	.4282	14.5976	28	28
	40	1.9357	.2403	24.3183	36	40
	60	1.9521	.1826	36.9602	47	49
VI	20	1.6240	.8883	2.6561	11	12
	40	1.7996	.6012	4.9865	16	18
	60	1.8607	.4552	7.1754	19	23

By (5.1.19) and since $P(B_{\omega_1}) = c_1 h^{-1}$ we obtain the expected result

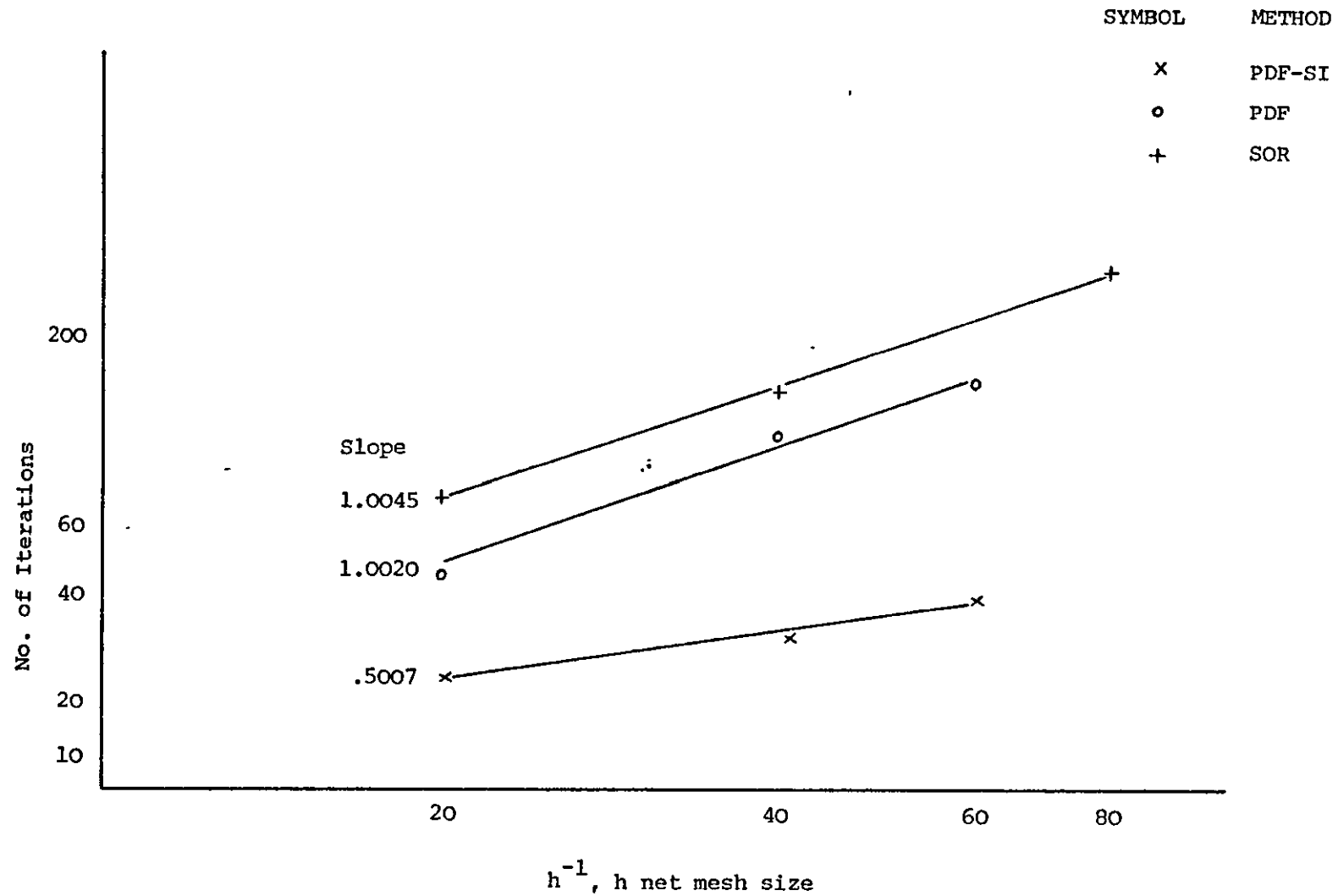
$$R_{\infty}(P_n(H_{\omega_1})) \doteq \frac{2}{\sqrt{c_1 h^{-1}}} = O(h^{\frac{1}{2}}) \quad (5.2.2)$$

The numerical results of Tables (5.2.T1) and (5.2.T2) indicate $O(h^{\frac{1}{2}})$ convergence even if the coefficient functions are not restricted to class $C^{(2)}$. This is an order of magnitude improvement of the PDF-SI over PDF. However, an $h^{-\frac{1}{2}}$ behaviour is attained by using SSOR-SI or PSD-SI. Also all three methods in the SI processes require approximately twice the work required by the SOR method.

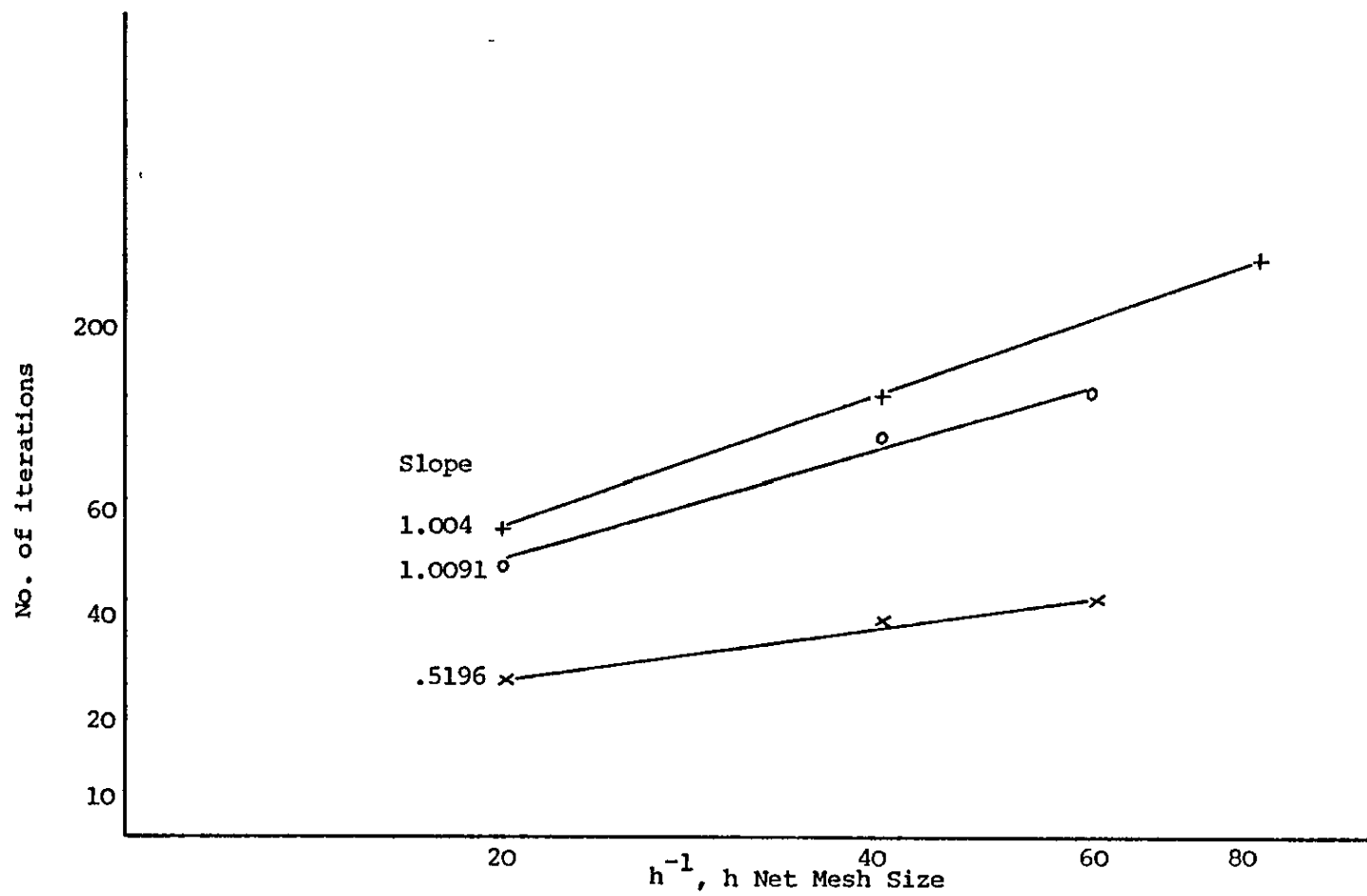
What we have to point out here is that the accuracy of the PDF-SI with estimated parameters is stronger than the one offered by the SSOR-SI and PSD-SI, since the estimated parameters of the PDF method are better posed than in the other two methods, namely PSD and SSOR (compare for example Table (4.2.T2)).

For the problems considered we obtain approximately $O(h^{\frac{1}{2}})$ convergence with the PDF-SI. As indicated in Table (5.2.T2), the PDF-SI method is applicable for problems with certain kinds of discontinuities, as shown by Problem IV and the one given by Case V. In cases like this, the expected analytical solution (if it is possible to be found) is a weak solution and this leads to the consideration of generalized functions. In the numerical solution it was suggested by Young [1971] that the coefficient of equation (1.5) to be bounded in a suitable chosen space.

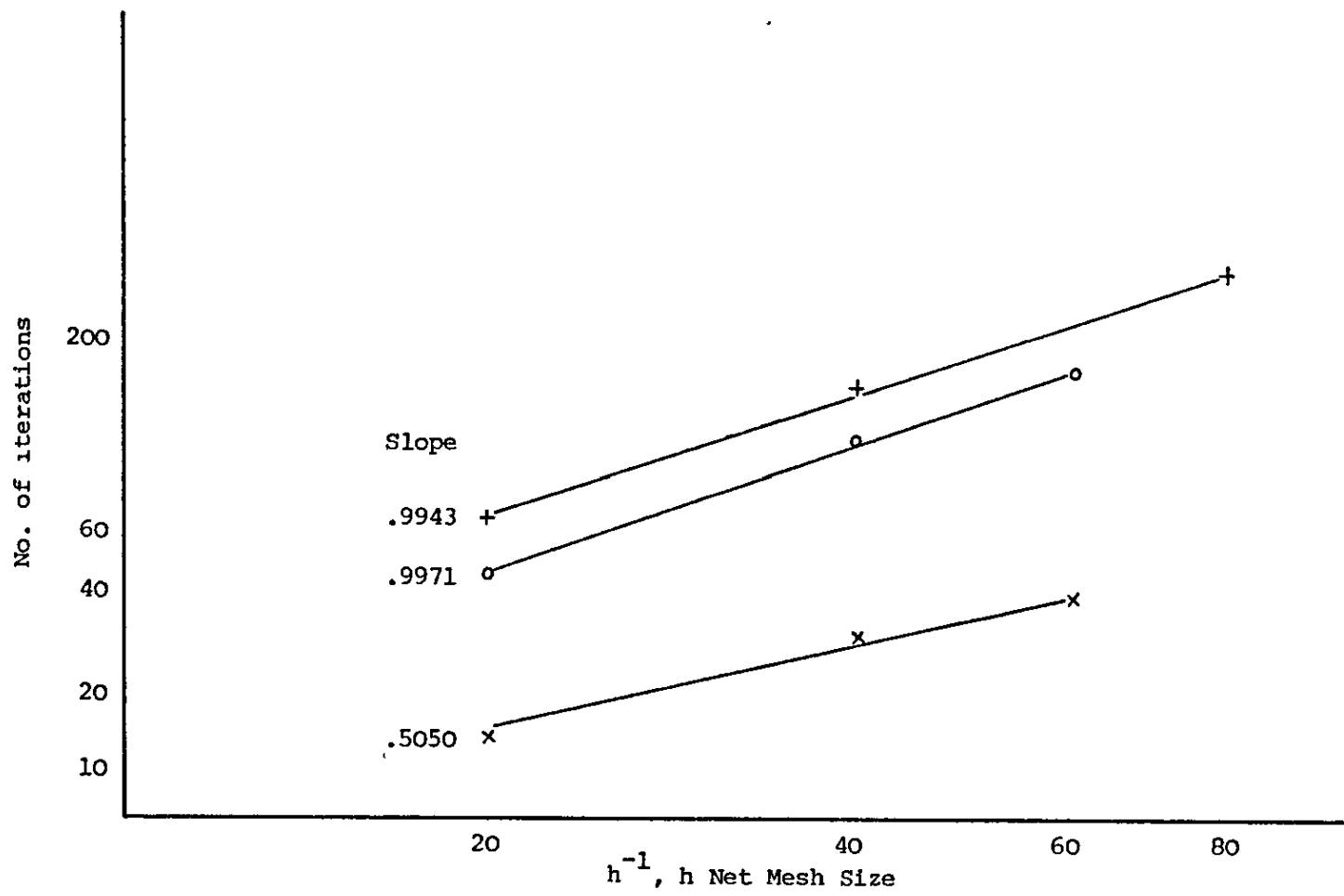
In Figures (5.2.F1-4) we plot the logarithm of the required number of iterations using PDF-SI versus $\ln h^{-1}$. The slope indicates the approximate order $O(h^{\frac{1}{2}})$ of the convergence rate.



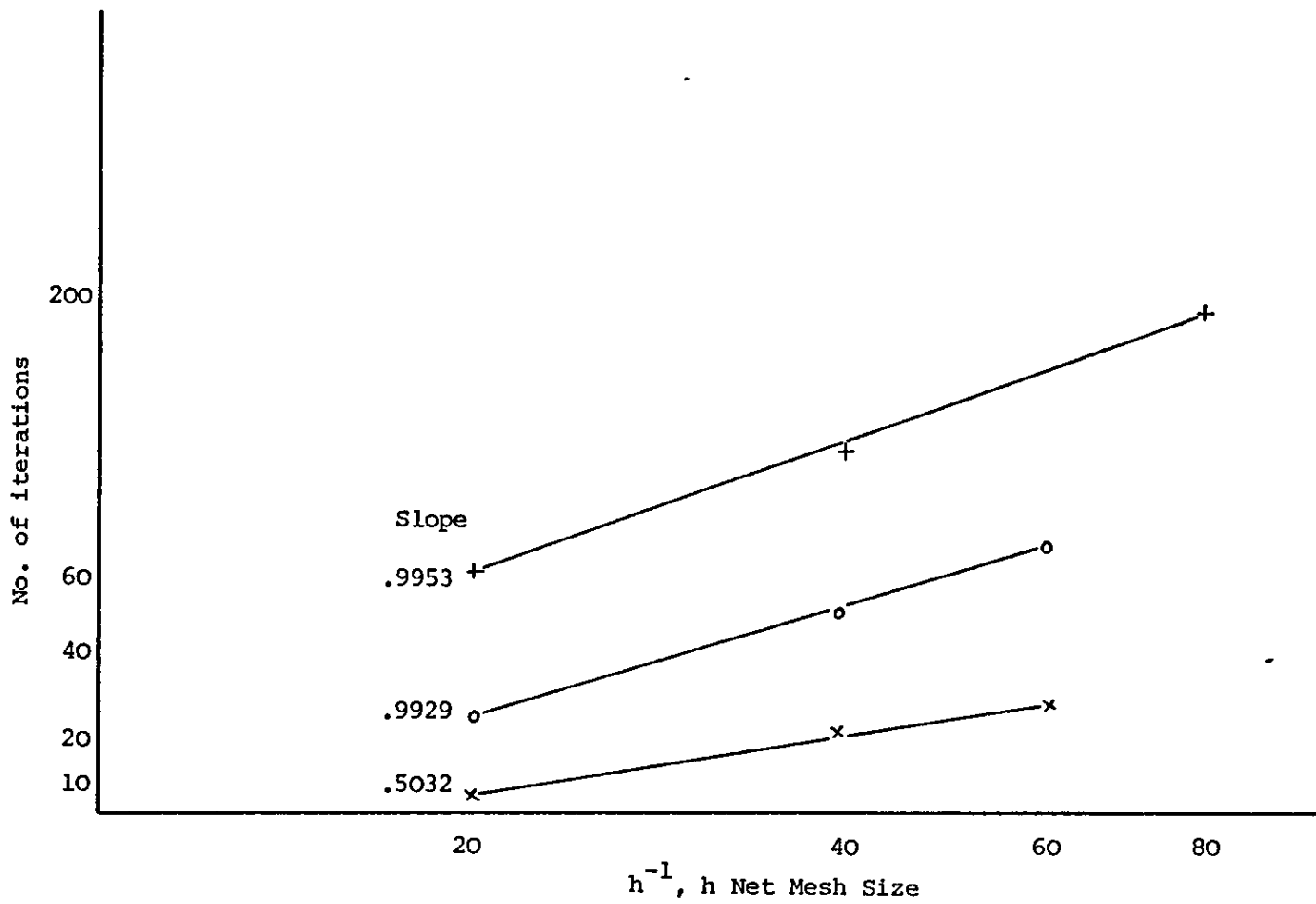
(5.2.F1): Problem 1. The Slope α indicates $O(h^\alpha)$ Rate of Convergence



(5.2.F2): Problem 3



(5.2.F3): Problem 4



(5.2.F4): Problem 6

EPILOGUE

Panoramix: Alors, Obelix, l'Helvétie, c'est comment?

Obelix: Plat.

Asterix Chez Les Helvètes

Gosciny

In this text, we have attempted to establish and describe an iterative method for solving large sparse systems of linear algebraic equations, based on the preconditioning concept. We have tried to present the theory making a detailed study on the formulation of the method. This investigation provided us with useful information on the validity of this particular type of preconditioning considered, on the spectral condition number and on the rate of convergence of the associated method.

The numerical experiments demonstrated the theoretical foundation of the Preconditioning by Direct Factorization method, even for certain kinds of discontinuities amongst the coefficients of the elliptic partial differential equations.

Since the iteration matrix of the PDF method has positive eigenvalues the semi-iterative technique applied resulted in a $O(h^{\frac{1}{2}})$ acceleration of the convergence rate, where h is the net mesh size.

Principally, we have been concerned with a certain conditioning matrix M of form (3.5.2). The field is open for research when M possess one of the forms stated in Section(3.3) of Chapter 3. Our aim was to reduce the original P-condition number (spectral condition number) of the

matrix A , by premultiplying A by M^{-1} . The resulting preconditioned matrix, $B_{\omega} = M^{-1}A$, turned the initial system (2.1.1) which was ill-conditioned to the well-conditioned system (3.2.1).

The PDF optimal procedure required the knowledge of the two parameters ω, τ at their optimal values, for carrying out the numerical solution of a partial differential equation efficiently. The theoretical foundation for the evaluation of ω was given in Theorem (3.6.4). Further, that theorem provided us with a bound on the P-condition number of the preconditioned matrix B_{ω} and hence a theoretical evaluation for the convergence rate of the PDF method.

The numerical results presented in Chapter 4 show a substantial improvement of the rate of convergence of the PDF method as compared with the SOR, SSOR and PSD method, where all the methods except SOR used estimated parameters. The percentage increase in the number of iterations required by the SOR, SSOR and PSD method over the PDF method, with estimated parameters, including the SOR method with optima parameters can be found in Tables (4.2.T2) and (4.1.T3) of Chapter 4. The PDF algorithm, nevertheless, required more operations than the aforementioned methods for the program to be executed. The Niethammer's scheme was applied to the PDF method in order to reduce the number of operations. A saving of only 20 percent was realized in the arithmetic operations count (vd Appendix C).

The PDF method accelerated by semi-iteration gave an order of magnitude improvement of the rate of convergence over the PDF method. Further research is required to investigate the accuracy offered by the PDF-SI method as compared to the SSOR-SI and the PSD-SI method. The acceleration by semi-iterative techniques shows that the PDF model can be

applied effectively to second degree and to non-stationary methods even in cases of differential equations containing discontinuous coefficients, as Problem V.

Our results permit further generalization to cases where the conditioning matrix M possesses one of the forms described in Section (3.3) of Chapter 3. However, the exposition of such cases of the conditioning matrix will occupy a larger amount of computer storage. Of particular interest, would be the investigation of a preconditioning method with such a matrix M as well as the problem of finding the optimal process (3.5.3) from the condition of minimizing the numerical work.

APPENDIX A

MATRIX THEORY PRELIMINARIES

In this Appendix we present some basic definitions and theorems on matrix theory, preassuming that the reader is familiar with it. The proofs can be found in any standard book of matrix theory.

Definition (A.1): If $x, y \in R^h$ then by $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ we call the inner product of x, y over R .

Definition (A.2): The L^2 -norm of a vector $x \in R^h$, is defined by

$$\|x\|_{L^2} = \sqrt{\langle x, x \rangle}$$

Definition (A.3): The matrices A and $A_1 = BAB^{-1}$ are said to be similar. The relation is symmetric for $A = B^{-1}A_1B$.

Theorem (A.4): If A is symmetric matrix then for any $x \in R^h$

$$\min_{x \neq 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \leq \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \leq \max_x \frac{\langle x, Ax \rangle}{\langle x, x \rangle}$$

Theorem (A.5): For any matrix A , the matrix AA^T is symmetric and non-negative definite. If A is non-singular then AA^T is positive definite.

Theorem (A.6): If A is positive definite matrix then there exists a unique positive definite matrix B such that

$$B^2 = A.$$

(The matrix B is denoted by $A^{\frac{1}{2}}$).

Theorem (A.7): If A is positive definite then for any non-singular matrix L the matrix $M=LAL^T$ is positive definite.

Definition (A.8): The L^2 -norm of a positive definite matrix A is defined by

$$\|A\|_{L^2} = \sup_{x \neq 0} \frac{\|Ax\|_{L^2}}{\|x\|_{L^2}} = \sup_{\|x\|_{L^2}=1} \langle Ax, Ax \rangle .$$

Definition (A.9): The M-norm of a matrix A is defined as

$$\|A\|_M = \|MAM^{-1}\| ,$$

for any non-singular matrix M and for any norm $\|\cdot\|$.

Theorem (A.10): For any matrix norm we have $S(A) \leq \|A\|$ where $A \in \mathbb{R}^{h \times h}$ and $S(A)$ denotes the spectral radius of A.

Theorem (A.11): If A and B are square matrices, then AB and BA have the same eigenvalues with the same multiplicities.

Definition (A.12): A matrix $A=(a_{ij})$ has weak diagonal dominance if

$$|a_{ii}| \geq \sum_{j, j \neq i} |a_{ij}|$$

and for some i,

$$|a_{ii}| > \sum_{j, j \neq i} |a_{ij}|$$

If for every i the second relation is valid, then A has strong diagonal dominance.

Definition (A.13): A $(n \times n)$ matrix A is reducible if there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

where B and C are square matrices, or if $n=1$ and $A=0$. Otherwise, A is irreducible.

Definition (A.14): A matrix A of order n has "Property A" if there exist two disjoint subsets S_1 and S_2 of W, the set of the first n positive integers, such that $S_1+S_2=W$ and such that if $i \neq j$ and if either $a_{ij} \neq 0$ or $a_{ji} \neq 0$, then $i \in S_1$ and $j \in S_2$ or else $i \in S_2$ and $j \in S_1$.

Simply by rearranging the rows and corresponding columns of A, A can obtain the form

$$\begin{bmatrix} D_1 & H \\ K & D_2 \end{bmatrix} \quad (\text{A.15})$$

where D_1 and D_2 are square diagonal matrices.

Wielandt-Hoffman Theorem (A.16): If $C=A+B$, where A, B and C are symmetric matrices having the eigenvalues a_i, b_i, γ_i respectively arranged in non-increasing order, then

$$\sum_1 (\gamma_i - a_i)^2 \leq \|B\|_{L^2}^2 = \sum_1 b_i^2 .$$

APPENDIX B

PROOF OF THEOREM (2.7.11)

In this Appendix a proof is stated in Theorem (2.7.11).

Theorem (2.7.11)'s exclamation states except that $S(UL) \leq \bar{b}$ is used instead of $S(LU) \leq \bar{b}$ in (2.7.12).

Proof:

Evidently (2.7.13) is derived from (3.6.1) by Lemma (3.5.30) and the relation $M(C_\omega) \leq 1/[\omega(2-\omega)]$.

Consider now the functions $p_M(\omega) = p_M(\omega; M, \bar{b})$ and $p_m(\omega) = p_m(\omega; M, \bar{b})$ as they have been defined.

In order to minimize $p_M(\omega)$ and $p_m(\omega)$ we first note that in the range (1,2) of ω they have a minimum at

$$\omega_M = \frac{2}{1 + \sqrt{1 - 2M + 4\bar{b}}}$$

and at

$$\omega_m = \frac{2}{1 + \sqrt{1 - 2m + 4\bar{b}}}$$

respectively, where $\omega_m \leq \omega_M$.

We now attempt to locate the functions $p_M(\omega)$ and $p_m(\omega)$ with respect to their critical and optimal values (in the sense of minimization), so we distinguish the following cases.

Case I.a.1: $\bar{b} < \frac{1}{4}$, $1 < \omega \leq \omega^*$ and $\frac{2}{3} < \omega_M \leq \omega^*$

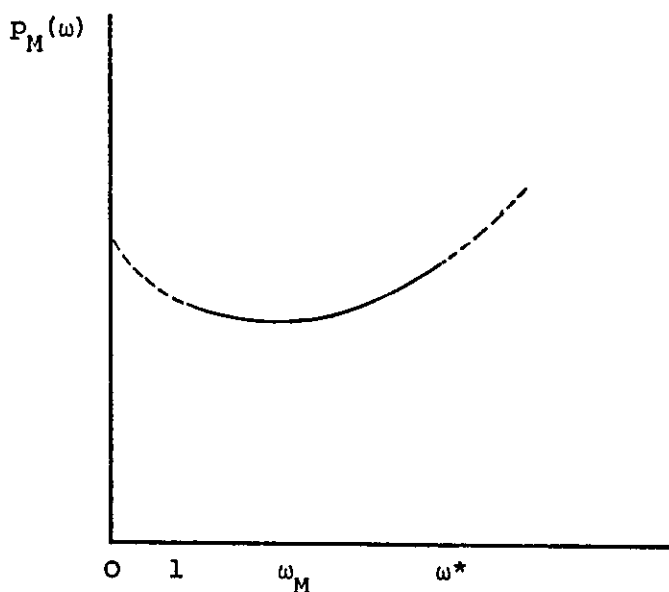
Since the critical point ω_M is in the range of ω , then

$$p_M(\omega_M) = \min_{\substack{\omega \in (1, \omega^*] \\ \omega_M \in (1, \omega^*]}} p_M(\omega)$$

Graphically $p_M(\omega)$ has the form (B.F1) as we can establish from Table (B.T1).

ω	0	ω_M	ω^*	2
$\text{sign}(p'_M(\omega))$	-	0	+	+
$p_M(\omega)$	↘		↗	

(B.T1)



(B.F1)

Case I.a.2: $\bar{b} < \frac{1}{4}$, $1 < \omega \leq \omega^*$ and $\omega_M \geq \omega^*$

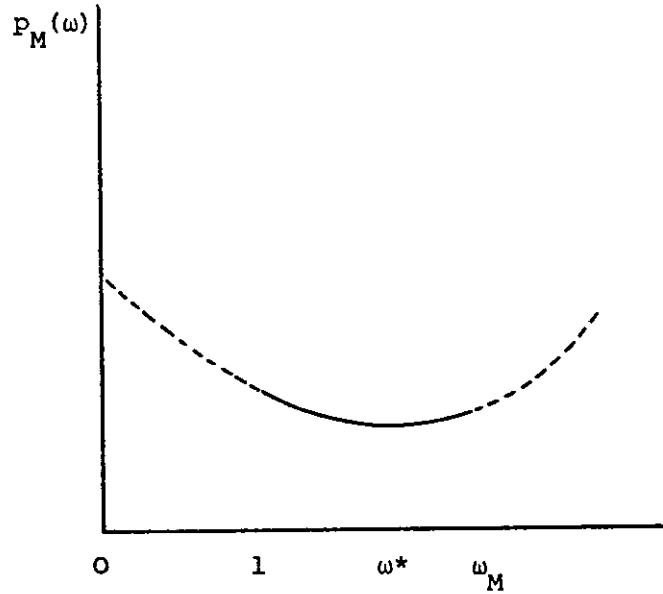
Since the critical point ω_M is out of the range of ω and $p_M(\omega)$ decreases on $(1, \omega^*]$, then

$$p_M(\omega^*) = \min_{\substack{\omega \in (1, \omega^*] \\ \omega_M \in (1, \omega^*)}} p_M(\omega)$$

Graphically $p_M(\omega)$ has the form (B.F2) as we can establish from Table (B.T2).

ω	0	ω^*	ω_M	2
$\text{sign}(p'_M(\omega))$	-	-	0	+
$p_M(\omega)$	↘		↘	

(B.T2)



(B.F2)

Case I.b: $\bar{b} < \frac{1}{4}$, $\omega^* \leq \omega < 2$.

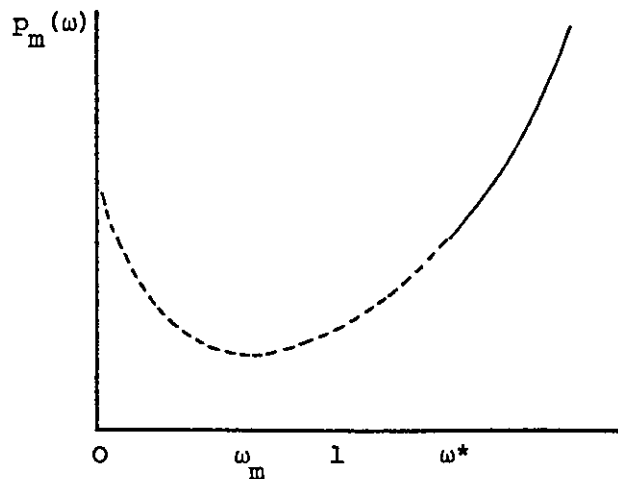
Since now, $\bar{b} < \frac{1}{4}$ and $\omega^* > 1$ then $\omega_M < 1 < \omega^*$. Therefore since $p_m(\omega)$ increases on $[\omega^*, 2)$, we have

$$p_m(\omega^*) = \min_{\omega \in [\omega^*, 2)} p_m(\omega)$$

Graphically $p_m(\omega)$ has the form (B.T3) as we can establish from Table (B.F3).

ω	0	ω_m	1	ω^*	2
$\text{sign}(p'_m(\omega))$	-		0	+	
$p_m(\omega)$	↘		↗	↗	

(B.T3)



(B.F3)

Respectively to the location of ω^* , ω_M and ω_m we ascertain the Cases I.a and I.b imply together that

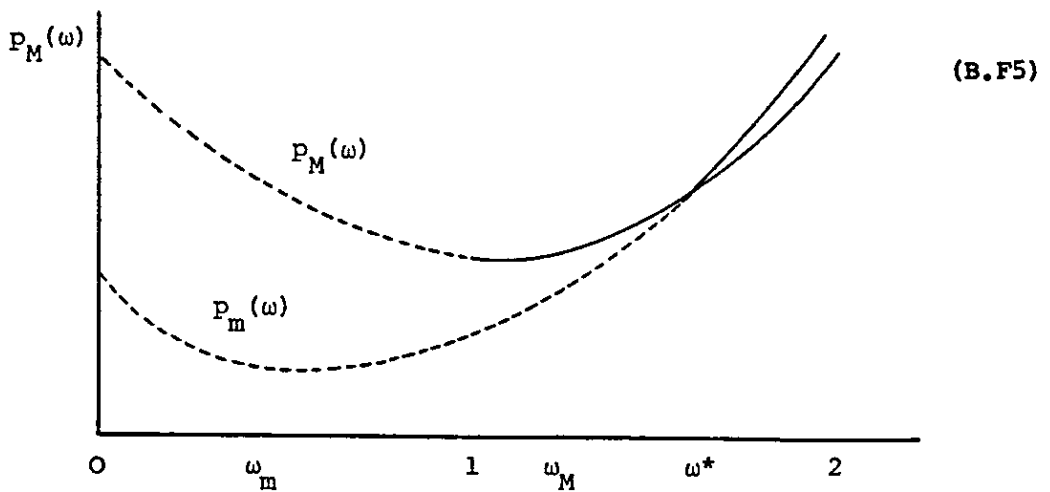
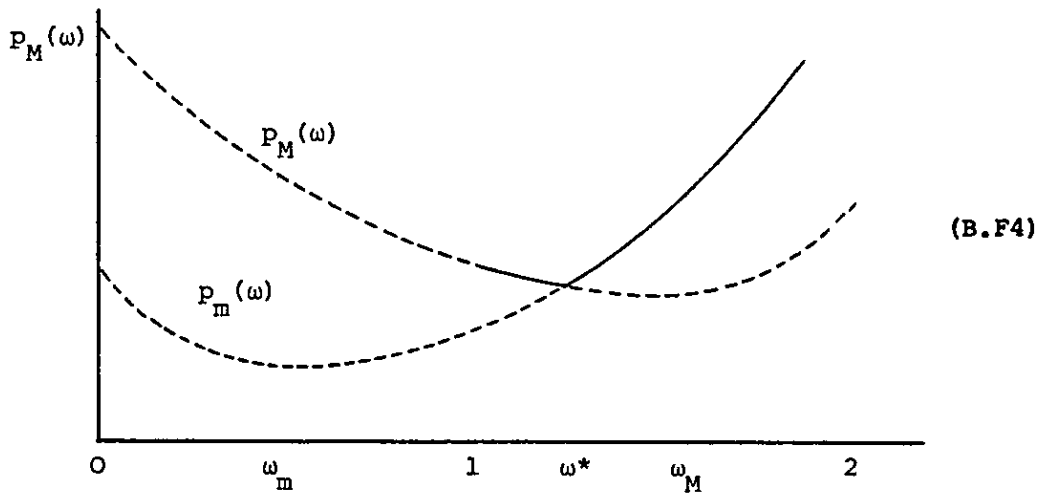
$$p_M(\omega_M) = \min_{\omega \in (1,2)} \{p_M(\omega), p_m(\omega)\} \\ \omega_m \leq \omega_M \leq \omega^*$$

when $\omega_m \leq \omega_M \leq \omega^*$ and that

$$p_M(\omega^*) = \min_{\omega \in (1,2)} \{p_M(\omega), p_m(\omega)\} \\ \omega_m \leq \omega^* \leq \omega_M$$

when $\omega_m \leq \omega^* \leq \omega_M$,

where graphically for $p_M(\omega)$ we have Figures (B.F4) and (B.F5).



Case II: $b > \frac{1}{4}$

Two subcases are observed which imply the same result. Hence always ω_M is in the range of ω . However, in Case II we have that

$$p_M(\omega_M) = \min_{\omega \in (1,2)} p_M(\omega) .$$

APPENDIX C

ARITHMETIC OPERATION COUNT

In this Appendix, we determine the number of arithmetic operations (ops) required by the PDF method to solve the problem

$$\frac{\partial}{\partial x} \left(A \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(C \frac{\partial u}{\partial y} \right) = 0, \quad (\text{C.1})$$

which is described in Chapter 1, for the unit square.

The discretised form of (C.1) is

$$u(x,y) = b_1(x,y)u(x+h,y) + b_2(x,y)u(x,y+h) + b_3(x,y)u(x-h,y) + b_4(x,y)u(x,y-h) \quad (\text{C.2})$$

where

$$\left. \begin{aligned} b_1(x,y) &= \frac{A(x+\frac{h}{2},y)}{S(x,y)}, & b_2(x,y) &= \frac{C(x,y+\frac{h}{2})}{S(x,y)} \\ b_3(x,y) &= \frac{A(x-\frac{h}{2},y)}{S(x,y)}, & b_4(x,y) &= \frac{C(x-\frac{h}{2},y)}{S(x,y)} \end{aligned} \right\} \quad (\text{C.3})$$

and

$$S(x,y) = A(x+\frac{h}{2},y) + A(x-\frac{h}{2},y) + C(x,y+\frac{h}{2}) + C(x-\frac{h}{2},y). \quad (\text{C.4})$$

We assume that the coefficients A and C for each mesh point are in storage and need only be computed once. Since the matrix A is symmetric then slightly more than a full array of size $(J-1) \times (J-1)$ is needed to store both coefficients A and C if $h=1/J$.

In the operation count we consider products as well as summation processes equally.

In order to compute the b_i 's in (C.2)

$$\left. \begin{array}{l} 4 \text{ divisions} \\ 3 \text{ additions} \end{array} \right\}$$

are required as can be seen from (C.3) and (C.4).

Let us now consider the PDF method defined by

$$u^{(n+1)} = u^{(n)} + \tau(I+F)^{-1}(I-\omega L)^{-1}(I-\omega U)^{-1}(b-Au^{(n)}) \quad (C.5)$$

$$\text{where } F = \omega^2 b(I-\omega L)^{-1}(I-\omega U)^{-1} e_{J-1} e_{J-1}^T \quad (C.6) \quad (1)$$

and $e_{J-1} = [0, 0, \dots, 0, 1]^T$ the $(J-1)$ -basic vector of $R^{(J-1) \times (J-1)}$.

An alternative form of (C.5) is the following scheme due to Niethammer [1964], which is a three step process.

$$\left. \begin{aligned} u^{(n+1/3)} &= (1-\tau)u^{(n)} + (\tau-\omega)Uu^{(n)} + \omega Uu^{(n+1/3)} + \tau(Lu^{(n)} + b) \\ u^{(n+1/3)}_{(J-1) \times (J-1)} &= \omega^2 b u^{(n)}_{(J-1) \times (J-1)} \\ u^{(n+2/3)} &= u^{(n+1/3)} + \omega Lu^{(n+2/2)} - \omega Lu^{(n)} \\ u^{(n+1)} &= (I+F)^{-1} u^{(n+2/3)} \end{aligned} \right\} \quad (C.7)$$

where the third step is a simple Gaussian elimination process, since the matrix $I+F$ is upper triangular with non-zero elements only on the diagonal and the last column.

Evidently we have

$$u^{(n+1)} = (I+F)^{-1}(I-\omega L)^{-1}(u^{(n+1/3)} - \omega Lu^{(n)})$$

and

$$u^{(n+1/3)} = (I-\omega U)^{-1} \{ (1-\tau)u^{(n)} + (\tau-\omega)Uu^{(n)} + \tau(Lu^{(n)} + b) + \omega^2 b e_{J-1} e_{J-1}^T u^{(n)} \}.$$

(1) Evidently F has only one non-zero column, i.e., the last one. Thus the form of F is relatively easy.

Eliminating $u^{(n+1/3)}$ we get the formula (3.5.4)

$$\begin{aligned} u^{(n+1)} &= M^{-1} [M - \tau A] u^{(n)} + \tau M^{-1} b \\ &= (I - \tau B_{\omega}^{-1} A) u^{(n)} + \tau M^{-1} b . \end{aligned}$$

By Niethammer's scheme we can reduce the computational work, taking advantage of the appearance of $Lu^{(n)}$ in both equations of (C.7). Thus, we can store $Lu^{(n)}$ after the first half iteration and use it for the second half one. Similarly at the end of the second half iteration we can store $Lu^{(n+2/3)}$ and use it in the next iteration step after the direct one. For each step after the first it is necessary to store only $Lu^{(n+1)}$ and not $Lu^{(n+1+2/3)}$.

Explicitly, it can be seen as indicated,

$$\left. \begin{aligned} u^{(n+1/3)} &= (1-\tau)u^{(n)} + (\tau-\omega)Uu^{(n)} + \omega Uu^{(n+1/3)} + \tau(Lu^{(n)} + b) \\ u^{(n+1/3)}_{(J-1) \times (J-1)} &= \omega^2 b u^{(n)}_{(J-1) \times (J-1)} \\ \text{save } Lu^{(n)} & \\ u^{(n+2/3)} &= u^{(n+1/3)} + \omega Lu^{(n+2/3)} - \omega Lu^{(n)} \\ u^{(n+1)} &= (I+F)^{-1} u^{(n+2/3)} \end{aligned} \right\}$$

$$\left. \begin{aligned} u^{(n+4/3)} &= (1-\tau)u^{(n+1)} + (\tau-\omega)Uu^{(n+1)} + \omega Uu^{(n+4/3)} + \tau(Lu^{(n+1)} + b) \\ u^{(n+4/3)}_{(J-1) \times (J-1)} &= \omega^2 b u^{(n)}_{(J-1) \times (J-1)} \\ \text{save } Lu^{(n+1)} & \\ u^{(n+5/8)} &= u^{(n+4/3)} + \omega Lu^{(n+5/8)} - \omega Lu^{(n+1)} \\ u^{(n+2)} &= (I+F)^{-1} u^{(n+5/8)} \end{aligned} \right\}$$

etc.
 \vdots
 \vdots

We proceed to determine the number of operations necessary to complete one PDF iteration. From (C.7) we have the following PDF computation for a particular point (x,y) ,

$$\begin{aligned}
 \text{(a)} \quad u^{(n+1/3)}(x,y) &= (1-\tau)u^{(n)}(x,y) + (\tau-\omega) \left[b_1(x,y)u^{(n)}(x+h,y) + b_2(x,y)u^{(n)}(x,y+h) \right] \\
 &\quad + \omega \left[b_1(x,y)u^{(n+1/3)}(x+h,y) + b_2(x,y)u^{(n+1/3)}(x,y+h) \right] \\
 &\quad + \tau \left[b_3(x,y)u^{(n)}(x-h,y) + b_4(x,y)u^{(n)}(x,y-h) \right] \\
 \text{(b)} \quad u^{(n+1/3)}(J-1,J-1) &= \omega^2 b u^{(n)}(J-1,J-1)
 \end{aligned} \tag{C.8}$$

$$\begin{aligned}
 u^{(n+2/3)}(x,y) &= u^{(n+1/3)}(x,y) + \omega \left[b_3(x,y)u^{(n+2/3)}(x-h,y) + b_4(x,y)u^{(n+2/3)}(x,y-h) \right] \\
 &\quad - \omega \left[b_3(x,y)u^{(n)}(x-h,y) + b_4(x,y)u^{(n)}(x,y-h) \right]
 \end{aligned} \tag{C.9}$$

$$\begin{aligned}
 u^{(n+1)}(J-1,J-1) &= u^{(n+2/3)}(J-1,J-1) / [1+F(J-1,J-1)] \\
 u^{(n+1)}(x,y) &= u^{(n+2/3)}(x,y) - F(x,y)u^{(n+1)}(J-1,J-1)
 \end{aligned} \tag{C.10}$$

The algorithm needs $8J^2$ operations to perform F , additional to the total number of operations involved to execute the program.

For a single point, from (C.8a) we have

10 multiplications

and 6 additions

not counting the operations involved to form $(1-\tau)$ and $(\tau-\omega)$ since these can be computed once and stored.

For the computation of (C.9)

6 multiplications

3 additions

are required and

1 subtraction

Surplus to that for the computation of (C.10)

1 subtraction

and 1 multiplication

is required.

Therefore, for one full iteration

$$(16+7)J^2 + (10+5)J^2 + 2J^2 \text{ ops} = 40J^2 \text{ ops}$$

is required.

Using Neithammer's scheme we have seen that it is not necessary to compute $Lu^{(n)}$ in the second half iteration. This means a saving of 8 operations, hence for one iteration applying the Niethammer's process

$$(40J^2 - 8J^2) \text{ ops} = 32J^2$$

operations are required.

In comparison one SOR iteration requires $17J^2$ operations, one SSOR Niethammer's iteration requires $26J^2$ operations for the first iteration and $18J^2$ operations after the first and one PSD Niethammer's iteration requires $30J^2$ operations for the first iteration with $22J^2$ operations for every iteration after the first. Thus, we conclude that the PDF method requires about 30%, 40% and 45% more work for an intermediate iteration compared to the SOR, SSOR and PSD methods respectively. This was expected for the direct step (C.10) and was included in the algorithm.

A less sophisticated scheme than the Niethammer's one is the fractional-step scheme making use of vector corrections. This scheme is described in (3.5.4) of Chapter 3. One complete PDF iteration of (3.5.4) requires $36J^2$ operations.

In a simple manner we see that the PDF-SI method given by formula (5.1.9) requires $39J^2$ operations since it is a second degree formula.

If we applied a Chebyshev acceleration technique we should have a smaller number of operations in the PDF version, but we expect the same number of iterations at the PSD method. That is because both methods possess almost identical P-condition numbers even in some cases of estimated parameters (vd. Tables (4.2.T2) and (4.2.T3)). However, better accuracy for SSOR and PSD methods is expected by using a similar version of the PDF.

APPENDIX D

In this Appendix we present a program which was employed for the solution of the six problems in Chapter 4, by the PDF method, either with optima or with estimated parameters. The same program can be used for the PDF-SI method if we replace the last step of the PDF algorithm by the non-stationary SI step.

For all numerical tests performed in this work, the ICL 1904S* computer was used.

```
'BEGIN' 'INTEGER' NI,N,J,I,K;
'REAL' A0,A1,A2,A3,A4,W,D1,T;
'PROCEDURE' CALCOEF(C0,C1,C2,C3,C4,K,L,H);
'VALUE' K,L,H;
'REAL' C0,C1,C2,C3,C4,H;
'INTEGER' K,L;
'BEGIN' C1_C2_C3_C4_1;
C0_C1+C2+C3+C4;
'END';
'PROCEDURE' NORMINF(M,A,N);
'VALUE' N; 'INTEGER' N; 'REAL' M;
'ARRAY' A;
'BEGIN' 'INTEGER' I,J;
M_ABS(A[1,1]);
'FOR' I_1 'STEP' 1 'UNTIL' N 'DO'
'FOR' J_1 'STEP' 1 'UNTIL' N 'DO'
'IF' ABS(A[I,J]) 'GT' M 'THEN'
M_ABS(A[I,J]);
'END';
SELECT INPUT(0); SELECT OUTPUT(0);
NEWLINE(3);
WRITETEXT('(%PROBLEM1%)');
'FOR' W 1.7288, 1.8544, 1.8994 'DO'
'BEGIN'
N_READ;
PRINT(N,2,0);
NEWLINE(3);
WRITETEXT('(%T%)');
T_READ;
PRINT(T,0,8);
NEWLINE(2);
```

```

'BEGIN'
'ARRAY' X,Y,Z,D,H,G,F,Z1,L[0:N,0:N];
'FOR' I_0 'STEP' 1 'UNTIL' N 'DO'
'FOR' J_0 'STEP' 1 'UNTIL' N 'DO'
Y[I,J]_Z[I,J]_D[I,J]_X[I,J]_
L[I,J]_H[I,J]_G[I,J]_F[I,J]_Z1[I,J]_0;
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
X[I,J]_1;
NI_0;
'FOR' J_N-1 'STEP' -1 'UNTIL' 1 'DO'
'FOR' I_N-1 'STEP' -1 'UNTIL' 1 'DO'
'BEGIN'
CALCOEF(A0,A1,A2,A3,A4,I,J,1/N);
H[N-1,N-1]_W*W/4;
G[I,J]_H[I,J]+W/A0*(A1*G[I+1,J]+A2*G[I,J+1]);
'END';
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'BEGIN'
CALCOEF(A0,A1,A2,A3,A4,I,J,1/N);
F[I,J]_G[I,J]+W/A0*(A3*F[I-1,J]+A4*F[I,J-1]);
'END';
LAB1:
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'BEGIN' CALCOEF(A0,A1,A2,A3,A4,I,J,1/N);
Z[I,J]_X[I,J]-1/A0*(A1*X[I+1,J]+A2*X[I,J+1]+
A3*X[I-1,J]+A4*X[I,J-1]);
'END';
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'

Z[I,J]_Z[I,J]-D[I,J];
'FOR' J_N-1 'STEP' -1 'UNTIL' 1 'DO'
'FOR' I_N-1 'STEP' -1 'UNTIL' 1 'DO'
'BEGIN' CALCOEF(A0,A1,A2,A3,A4,I,J,1/N);
Y[I,J]_Z[I,J]+W/A0*(A1*Y[I+1,J]+A2*Y[I,J+1]);
'END';
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'BEGIN' CALCOEF(A0,A1,A2,A3,A4,I,J,1/N);
Y[I,J]_Y[I,J]+W/A0*(A3*Y[I-1,J]+A4*Y[I,J-1]);
'END';
Z1[N-1,N-1]_Y[N-1,N-1]/(1+F[N-1,N-1]);
'FOR' J_N-2 'STEP' -1 'UNTIL' 1 'DO'
Z1[N-1,J]_Y[N-1,J]-F[N-1,J]*Z1[N-1,N-1];
'FOR' I_N-2 'STEP' -1 'UNTIL' 1 'DO'
'FOR' J_N-1 'STEP' -1 'UNTIL' 1 'DO'
Z1[I,J]_Y[I,J]-F[I,J]*Z1[N-1,N-1];
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
Y[I,J]_X[I,J]-T*Z1[I,J];
NI_NI+1;

```



```

K N-1;
NORMINF(D1,Y,K);
'IF' D1 'LE' &-6 'THEN' 'GOTO' LAB2;
'FOR' J_1 'STEP' 1 'UNTIL' N-1 'DO'
'FOR' I_1 'STEP' 1 'UNTIL' N-1 'DO'
'BEGIN' L[I,J]_X[I,J];
X[I,J]_Y[I,J];
'END';
'GOTO' LAB1;
LAB2:
'FOR' I_0 'STEP' 1 'UNTIL' N 'DO'
Y[I,0]_0;
NEWLINE(3);
WRITETEXT('(%W%)');PRINT(W,1,4);
NEWLINE(3);
WRITETEXT('(%NI%)');PRINT(NI,1,4);
NEWLINE(3);
'FOR' J_0 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'
NEW LINE(3);
'FOR' I_0 'STEP' 1 'UNTIL' N 'DO'
PRINT(Y[I,J],0,8);
'END';
'END';
'END';
'END';
FINISH

```

REFERENCES

- [1] BENOKRAITIS, V.J. (1974): "*On the Adaptive Acceleration of Symmetric Successive Overrelaxation*", Ph.D. Thesis, Univ. of Texas at Austin, U.S.A.
- [2] BERMAN, A. and R.J. PLEMMONS (1979): "*Non-negative Matrices in the Mathematical Sciences*", Academic Press, New York, San Francisco, London.
- [3] DIAMOND, M.A. (1972): "*An Economical Algorithm for the Solution of Elliptic Difference Equations Independent of User-Supplied Parameters*", Ph.D. Thesis, Univ. of Illinois at Urbana-Champaign, U.S.A.
- [4] EVANS, D.J. (1968): "*The Use of Preconditioning in Iterative Methods for Solving Linear Equations with Symmetric Positive Definite Matrices*", J.Inst.Math.Appl. 4, 295-314.
- [5] EVANS, D.J. (1973): "*Comparison of the Convergence Rates of Iterative Methods for Solving Linear Equations with Preconditioning*", Greek Mathematical Society, Carathéodory Symposium, 106-135.
- [6] EVANS, D.J. (1972): "*An Algorithm for the Solution of Certain Tridiagonal Systems of Linear Equations*", The Computer Journal, 15, 356-359.

- [7] EVANS, D.J. (1974): *"Iterative Sparse Matrix Algorithms"*, In *"Software in Numerical Mathematics"* (D.J. Evans, ed.), Academic Press, 1974, 49-83.
- [8] EVANS, D.J. (1980): *"On Preconditioned Iterative Methods for Elliptic Partial Differential Equations"*, Elliptic Problem Solvers Conference, Los Alamos Scientific Laboratory, July (1980), Academic Press
- [8a] EVANS, D.J. (1980): *Private Communication.*
- [9] EVANS, D.J. and C.V.D. FORRINGTON (1963): *"An Iterative Process for Optimizing Symmetric Overrelaxation"*, The Computer Journal, 6, 271-273.
- [10] EVANS, D.J. and A. HADJIDIMOS (1979): *"On the Factorization of Special Symmetric Periodic and Non-Periodic Quindiagonal Matrices"*, Computing, 259-266.
- [11] EVANS, D.J. and N.M. MISSIRLIS (1980): *"The Preconditioned Simultaneous Displacement Method"*, Math. and Comp. in Simulation (in press).
- [12] GOLUB, G.H. and R.S. VARGA (1961): *"Chebyshev semi-iterative methods, Successive Overrelaxation Iterative Methods and Second Order Richardson Iterative Methods"*, Numer.Math. Part I and II, 3, 147-168.
- [13] GUNN, J.E. (1964): *"The Solution of Elliptic Difference Equations by Semi-Explicit Iterative Techniques"*, J.SIAM Numer.Anal., Ser.B, 2, 24-45.

- [14] HABETLER, G.J. and E.L. WACHSPRESS (1961): "*Symmetric Successive Overrelaxation in Solving Diffusion Difference Equations*", Math.Comp. 15, 356-362.
- [15] HIMMELBAU, D.M. (1972): "*Applied Non-linear Programming*", McGraw-Hill.
- [16] KAHAN, W. (1958): "*Gauss-Seidel Methods of Solving Large Systems of Linear Equations*", Ph.D. Thesis, Univ. of Toronto, Canada.
- [17] MARCHUK, G.I. (1975): "*Methods of Numerical Mathematics*", (translated by J. Ruzicka from Russian), Springer-Verlag, New York, Heidelberg, Berlin.
- [18] MISSIRLIS, N.M. (1978): "*Preconditioned Iterative Methods for Solving Elliptic Partial Differential Equations*", Ph.D. Thesis, Univ. Of Technology, Loughborough, U.K.
- [19] MISSIRLIS, N.M. and D.J. EVANS (1980): "*On the Acceleration of the Preconditioned Simultaneous Displacement Method*", submitted to I.M.A. Jour.Num.Anal.
- [20] NIETHAMMER, W. (1964): "*Relaxation bei Komplexen Matrizen*", Math. Zeitsch, 86, 34-40.
- [21] VARGA, R.S. (1957): "*A Comparison of the Successive Overrelaxation Method and Semi-Iterative Methods Using Chebyshev Polynomials*", J.Soc.Indust.Appl.Math., 5, 39-46.
- [22] VARGA, R.S. (1962): "*Matrix Iterative Analysis*", Prentice-Hall, New Jersey.

- [23] WILKINSON, J.H. (1965): *"The Algebraic Eigenvalue Problem"*,
Clarendon Press, Oxford.
- [24] YOUNG, D.M. (1971): *"Iterative Solution of Large Linear Systems"*,
Academic Press, New York.
- [25] YOUNG, D.M. (1971a): *"A Bound for the Optimum Relaxation Factor for
the Successive Overrelaxation Method"*, Numer.Math., 16, 408-413.
- [26] YOUNG, D.M. (1972): *"Second-Degree Iterative Methods for the Solution
of Large Linear Systems"*, J. of Approximation Theory, 5,
137-148.
- [27] YOUNG, D.M. (1977): *"On the Accelerated SSOR Method for Solving
Large Linear Systems"*, Advances in Mathematics, 23, 215-271.

