THE UNIVERSITY OF MANITOBA

# SOME FINITE AND INFINITE MATRICES, THEIR COMPUTATIONS AND APPLICATIONS

BY

**CHUANXIANG JI**

DISSERTATION
PRESENTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

WINNIPEG, MANITOBA
JANUARY 1998

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

# SOME FINITE AND INFINITE MATRICES, THEIR COMPUTATIONS AND APPLICATIONS

BY

CHUANXIANG JU

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

## Abstract

In this thesis, we consider finite and infinite matrices in linear equations with different structures which arise mainly in the solution of some elliptic partial differential equations in two dimensions. In many of the cases, the solutions lead to infinite systems of linear equations associated with matrices of special structures like diagonal dominance, tridiagonal or having a new sign distribution. The regions considered are either doubly connected or semi infinite. We also consider the theory of finite and infinite tridiagonal matrices, improving some well-known classical results. Nonsingularity criteria are given for matrices with a new sign distribution, which occurs in a conformal mapping problem and viscous fluid flow problem. For the semi infinite region which is bounded on the top by a sloping sinusoidal curve, a theoretical solution in terms of infinite matrices is given leading to numerical evaluation and development of the software. The above problems occur in transmission of electricity in coaxial cables, groundwater flow, conformal mapping , recurrence relations for Bessels functions etc. We also give an error estimate for a finite element method for solution of Laplace's equation resulting in double integrals for physical quantities in applications. The thesis is mainly concerned with using estimates for solving infinite and finite systems with easily computable and meaningful error estimates. The problem in groundwater flow in an infinite region arose from a problem suggested by industry.

## ACKNOWLEDGEMENT

I would like to thank my advisor, Dr. P.N. Shivakumar, with my appreciation and gratitude for his guidance, encouragement and financial support while I was working towards this degree.

I dedicate this thesis to my father, Dr. Xiaohong Ji, who passed away in 1992 when I entered this program, who had always been my source of confidence and motivation, and to my mother Fuzhen Zong, who has given me her deep love and excellent education. I also dedicate my work to my wife, Xuesi Jin, whose understanding and support have made this thesis possible.

# Contents

**Contents** i

**List of Figures** iii

**List of Tables** iv

**1 Introduction** 1

**2 Conformal Mapping of Doubly Connected Regions** 5

2.1 Introduction . . . . . 7

2.2 Formulation of the problem . . . . . 8

2.3 Conformal mapping . . . . . 10

2.4 The boundary value problem . . . . . 11

2.5 Rate of flow . . . . . 13

2.6 Approximation and error analysis . . . . . 15

2.7 Numerical results . . . . . 16

2.8 Conclusions and comments . . . . . 22

**3 Evaluation of a Double Integral over a Doubly Connected Region** 24

3.1 Introduction . . . . . 24

3.2 Formulation of the problem . . . . . 26
3.3 A finite element method for the double integral . . . . . 27
3.4 Numerical results and conclusions . . . . . 32

**4 Non-singularity of Matrices of Certain Sign Distributions 39**
4.1 Introduction . . . . . 39
4.2 Main Theorem . . . . . 41
4.3 Proof . . . . . 43
4.4 Applications . . . . . 49

**5 Upper and Lower Bounds for the Inverse Elements of Finite and Infinite Tridiagonal Matrices 55**
5.1 Introduction . . . . . 55
5.2 Finite tridiagonal matrices . . . . . 56
5.3 Infinite tridiagonal matrices . . . . . 64
5.4 Infinite block tridiagonal matrices . . . . . 69
5.5 Applications . . . . . 71

**6 An Elliptic Boundary Value Problem Defined on an Infinite Domain 80**
6.1 Introduction . . . . . 81
6.2 Mathematical model . . . . . 82
6.3 Formal solution . . . . . 84
6.4 Numerical approximation and error estimation . . . . . 90
6.5 Numerical results . . . . . 92

6.6 Conclusions and future work . . . . . 94

# List of Figures


6.1 Semi-infinite domain . . . . . 83

6.2 Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.0$. $L = 80.000$. . . . . 94

6.3 Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.00235$. $L = 80.000$. . . . . 95

6.4 Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.0235$. $L = 80.000$. . . . . 96

# List of Tables


2.1 Rate of flow $R_e$ vs $R_c$=2.098121 . . . . . 18

2.2 Rate of flow $R_e$ vs $R_c$=1.337667 . . . . . 19

2.3 Rate of flow $R_e$ vs $R_c$=0.955380 . . . . . 20

2.4 Rate of flow $R_e$ vs $R_c$=0.395791 . . . . . 21

2.5 Rate of flow $R_e$ vs $R_c$=0.057245 . . . . . 21

3.1 Rates of flow $R_e$'s with $\partial\Omega_1$ : $x^2 + y^2 = 1$ . . . . . 34

3.2 Rates of flow $R_e$'s with $\partial\Omega_1$ : $\frac{x^2}{(5/4)^2} + \frac{y^2}{(4/5)^2} = 1$ . . . . . 35

3.3 Rates of flow $R_e$'s with $\partial\Omega_1$ : $\frac{x^2}{(3/2)^2} + \frac{y^2}{(2/3)^2} = 1$ . . . . . 36

5.1 Solutions for truncated system . . . . . 75

5.2 Actual errors . . . . . 76

5.3 Values for $y_k$, m=20 . . . . . 79

6.1 Approximation and actual error on top boundary . . . . . 93

# Chapter 1

# Introduction

The motivation for this thesis is to discuss the solution of some mathematical and physical problems whose solutions lead to linear systems of equations. In many such problems, infinite matrices occur and the solution is often approached by truncation of the infinite matrix or by considering the infinite matrix as an operator. The latter approach, while giving in some cases qualitative results such as existence, uniqueness and even justification of truncation, is of very limited help in obtaining explicit error bounds for approximate solutions or for computation. The algebra of finite matrices is often extended to treat the analysis of infinite matrices. This thesis is concerned with a variety of structures for the matrices which have arisen in many physical problems and in classical analysis [30]. Infinite matrices have a very interesting history and the excellent review by Bernkopf [31] traces the role of infinite matrices in the development of operator theory and integral equations. Due to the difficulties in treating infinite matrices, not much progress was achieved in the literature involving computations except in the framework of Operator Theory. Hilbert used infinite quadratic forms to solve Fredholm integral equations, while in 1929, John von Neumann demonstrated

that an abstract approach was powerful and preferable to using infinite matrices as a tool for the study of Operator Theory.

This thesis continues the work of Shivakumar and his collaborators over the last two decades. The problems dealt with in this thesis are of a classical nature and do not have many recent references. As an example, conformal mapping of specific doubly connected regions has been a long standing problem in classical analysis. Another example of a different classical problem not dealt with in the literature is the solution of an elliptic equation in an infinite region. This problem was suggested by industry. Diagonal dominance has been the main motivation in [33] - [37]. The topics include linear algebraic systems with matrices having structures like diagonal dominance, tridiagonal etc. The problems discussed include differential equations, infinite systems of first order differential equations, iteration techniques etc. The applications have included Mathieu equations, Bessel equations, conformal mapping of doubly connected regions etc.

In Chapter 2, we consider the problem of the conformal mapping of a doubly connected region which is equivalent to solving a Poisson's equation in $w(x, y)$ with $w(x, y)$ vanishing on the two bounding curves [41]. The problem also represents the velocity along the axis of a slow and steady viscous fluid flowing between two pipes. A practical application could be that of simultaneous flow of gas and oil in a situation where both gas and oil are found in one location. A matter of practical importance is the rate of flow of the fluid between the pipes and to maximize this flow by varying the eccentricity of the circles. Here we give a proof of our results based on analysis and computation. These problems occur in the transmission of electricity in coaxial cables and in many other applications. The

result we establish is that the rate of flow is not a maximum in the case of concentric circles and, in fact, the opposite is true. We use conformal mapping as a tool to map the given doubly connected region to that of a region bounded by concentric circles.

In Chapter 3, we continue the work of Chapter 2 and give a finite element method and an error estimate for evaluating double integrals over a smooth domain. The results are used to compare rates of flow of a viscous incompressible fluid in a pipe-in-pipe system discussed earlier. These numerical results confirm an earlier conjecture that the domain yielding the least flow is the case of concentric circles.

In Chapter 4, we consider the flow described in Chapter 2 for the case of a region bounded by an ellipse and a circle by adopting a technique of mapping functions used in [35]. The resulting infinite matrix has a certain sign distribution and has only partial diagonal dominance in its elements. For such matrices, we give a set of sufficient conditions to ensure that the finite truncated matrix becomes nonsingular. The criterion developed for nonsingularity is easily verified.

We give easily computable upper and lower bounds for the inverse elements of finite diagonally dominant tridiagonal matrices [42] in Chapter 5. We also improve the well-known upper bounds due to Ostrowski. The results are extended to infinite systems. The theory is used to evaluate Bessel functions and Mathieu functions by using their recurrence relations.

In the final chapter, a ground water flow problem is discussed. The problem reduces to solving an elliptic equation defined in a semivertical infinite region of finite width. The top boundary is a sloping sinusoidal curve. A mathematical

analysis leading to numerical computation is given [45]. The problem is reduced to an infinite system of linear equations by using the method of separation of variables and construction of a Grammian matrix. Truncation (though not justified) yields an approximate solution that gives the best approximation on the top boundary. This problem arose in the discussion of contaminated groundwater flows and was suggested by Atomic Energy of Canada Ltd.

Finally, the thesis is an attempt to deal with some difficult problems of applied mathematics by developing meaningful, easily computable solutions with error bounds. The techniques used are mostly based on known and derived estimates concerning the given matrix and its inverse.

# Chapter 2

# Conformal Mapping of Doubly Connected Regions

The solution of a large number of problems in modern technology such as leakage of a gas in a graphite brick of gas cooled nuclear reactor [1]. analysis of stresses in solid propellant rocket grains [5]. simultaneous flow of oil and gas in concentric pipes [35] hinges critically on conformally mapping a doubly connected region onto a circular annulus. Only a few specific regions have been studied and only approximate solutions have been given. Hockney [1] considers the region where the inner boundary is a circle and the outer boundary is a square. He obtains a series solution for Laplace's equation in the region and gives an approximate solution of the problem by replacing the outer square by a circle of equal area and solving the resulting one dimension radial problem: Laura [2] considers the region with circular external boundary and an internal boundary which consists of several axes of symmetry: Narodetskii and Sherman [3] discuss the mapping of a region bounded by an ellipse and a circle: Symm [4] considers the numerical mapping of a bounded doubly connected domain onto an annulus. He describes a technique of mapping a general ring-shaped domain onto an annulus based on

the use of integral equations and illustrates this method with several numerical examples. Fornberg [6] studies a numerical method for conformal mapping of doubly connected regions with discretized boundaries and develops an iterative approach for computation.

More recent work includes the papers by Menke [7], [8] and the paper by Wegmann [9]. Menke studies conformal mapping of a doubly connected region bounded by the unit circle and an analytic Jordan curve. He approximates numerically the conformal mapping of an annulus onto a doubly connected region bounded by two concentric squares. Wegmann gives an iterative method for the numerical conformal mapping of a circular annulus onto a doubly connected region with smooth boundary. Papamichael [10] introduces a singular function that reflects the singular behaviour of the conformal mapping of a doubly connected region onto an annulus and demonstrates the method by several numerical examples which include the conformal mapping of rings of different shapes.

Most of the methods use integrals of the Cauchy type and then use truncation procedures to get numerical results. Although some estimations of accuracy are included, checking the numerical results using theoretical considerations is far from satisfactory. In this chapter, we provide proof of our results by numerical work and its analysis. In the following sections, we will describe a method of reducing the conformal mapping problem to a problem of solving an infinite system of linear algebraic equations.

## 2.1 Introduction

The mapping functions of the following form

$$\omega(z) = e^{[\log z + o(z)]}, \qquad z = x + iy = re^{i\theta}. \tag{2.1}$$

are widely studied[4]. On a doubly connected region bounded by two disjoint smooth curves $C_0$ and $C_1$, there is a mapping which is unique except for an arbitrary rotation and which maps the region $D + \partial D$ onto the annulus $0 < a \leq |\omega| \leq b < \infty$, where the ratio $b/a$ is unique and $o(z)$ is regular in $D$. We will assume that the origin in the $z$-plane is not included in the doubly connected region, and the function $o(z)$ has the following series expansion

$$o(z) = \sum_{n=-\infty}^{\infty} c_n z^n.$$

Hence for all $z \in C_0$, we need

$$\log(z\bar{z}) + o(z) + \overline{o(z)} = \log b^2.$$

and for all $z \in C_1$, we need

$$\log(z\bar{z}) + o(z) + \overline{o(z)} = \log a^2.$$

Without loss of generality, we will assume $b$ to be unity.

If a mapping $z = f_1(\zeta)$ maps conformally the simply connected region enclosed

by $C_1$, to a disk of radius 1 in the $\zeta$-plane, then we have

$$\log[f_1(\zeta)\overline{f_1(\zeta)} + o(f_1(\zeta)) + \overline{o(f_1(\zeta))}] - \log a^2 = 0.$$

on $\zeta\bar{\zeta} = 1$.

Using a Laurent series expansion for $o(z)$, we can derive a set of infinite linear equations for the coefficients $c_n$ and $a$. Similarly, if the mapping $z = f_1(\zeta')$ maps conformally the simply connected region enclosed by $C_0$, to a disk of radius 1 in the $\zeta'$-plane, we get another set of equations for $c_n$. Combining the two sets of equations for the $c_n$, the existence and uniqueness of the mapping function depends on the existence and uniqueness of the solution of the infinite system for the $c_n$from the idea that and $a$.

We will apply the above method to our application problem in the following sections.

## 2.2 Formulation of the problem

We consider the problem that arises when two fluids are transported with one fluid inside a pipe of cross-section $E$ bounded by $C_2$ and the other flowing in an annular domain $D$ in the $xy$ plane bounded internally by $C_2$ and externally by $C_1$. The flow velocity $w(x, y)$ satisfies the Poisson's equation

$$w_{xx} + w_{yy} = -\frac{P}{\mu} \quad \text{in} \quad D. \quad P, \mu \quad \text{being positive constants.} \tag{2.2}$$

and the boundary condition

$$w = 0 \quad \text{on} \quad C_1. \tag{2.3}$$

and

$$w = 0 \quad \text{on} \quad C_2. \tag{2.4}$$

In the above, $P$ represents the pressure gradient of the flow and $\mu$ the viscosity of the fluid. The flow is assumed to be slow and steady and the fluid is assumed to be incompressible and viscous. We will be concerned with the rate of flow

$$R = \int_D \int w(x, y) dx dy. \tag{2.5}$$

for curves $C_1$ and $C_2$ of given included area. In [35] the following cases were discussed:

(a) $C_1$ and $C_2$ being concentric circles.

(b) $C_2$ being a circle and $C_1$ being an ellipse.

(c) $C_1$ and $C_2$ being confocal ellipses.

Denoting the respective rates of flow by $R_a$, $R_b$, $R_c$, numerical evidence was presented which suggested that $R_c < R_b < R_a$. In all the three cases, the area included by $C_1$ and $C_2$ respectively were held constant. In the following sections, we will prove that when $C_1$ and $C_2$ are both circles, $R$ has a lower value for all the cases in which $C_1$ and $C_2$ are concentric. To compute $R$, we first seek the

solution of (2.2) - (2.4) by using a conformal map which maps $D$ in the $xy$ plane onto a circular annulus in the $\xi\eta$ plane.

## 2.3 Conformal mapping

For regions bounded by two eccentric circles enclosing a ring space [32], we take the mapping function

$$z = \frac{c}{1-\zeta}, \quad z = x + iy, \quad \zeta = \xi + i\eta, \quad c \quad \text{real}. \tag{2.6}$$

with $z'(\zeta) \neq 0$. For the transformation to be conformal, the ring space excludes the critical point $\zeta = 1$. The mapping (2.6) in cartesian coordinates takes the form

$$(x^2 + y^2)\zeta\bar{\zeta} = x^2 + y^2 - 2cx + c^2.$$

showing that the concentric circles $|\zeta| = \rho, \quad \rho = \rho_1, \rho_2, \quad \rho_1 < \rho_2$ transform onto the eccentric circles

$$(x-h)^2 + y^2 = a^2; \quad (x-k)^2 + y^2 = b^2, \quad a < b. \tag{2.7}$$

in the z-plane where $\rho_1 = \frac{a}{h}, \ \rho_2 = \frac{b}{k}$. Here $c = h - \frac{a^2}{h} = k - \frac{b^2}{k}$ and if $d$ is the distance between the two centers, $d = k - h = \frac{b^2}{k} - \frac{a^2}{h}$ and $\rho_2^2 - \rho_1^2 = \frac{cd}{hk}$. Note that $h = k$ implies $a = b$. We can only prescribe three quantities among $h, k, a, b$. If we fix $a, b$ then $h, k$ will have to satisfy the compatibility condition:

$$k - h = \frac{b^2}{k} - \frac{a^2}{h^2}. \tag{2.8}$$

## 2.4 The boundary value problem

Using the complex variables $z = x + iy$, $\bar{z} = x - iy$, (2.2) becomes

$$\frac{\partial^2 w}{\partial z \partial \bar{z}} = -\frac{P}{4\mu}$$

which on integrating gives, for real $w$

$$w = -\frac{P}{4\mu} z\bar{z} + \omega(z) + \overline{\omega(z)}. \tag{2.9}$$

The complex potential $\omega(z)$ is of the form

$$\omega(z) = B \ln z + \sum_{-\infty}^{\infty} b_n z^n. \tag{2.10}$$

In the $\zeta$-plane, we still use the notation $\omega(\zeta)$ for convenience, i.e. $\omega(\zeta) = \omega(z(\zeta))$ which leads to

$$w = -\frac{Pc^2}{4\mu} \frac{1}{(1-\zeta)(1-\bar{\zeta})} + \omega(\zeta) + \overline{\omega(\zeta)}. \tag{2.11}$$

where

$$\omega(\zeta) = A \ln \zeta + \sum_{-\infty}^{\infty} a_n \zeta^n. \tag{2.12}$$

and the boundary conditions (2.3), (2.4) reduce to

$$w = 0 \quad \text{on} \quad |\zeta| = \rho, \quad \rho = \rho_1, \rho_2, \quad \rho_1 < \rho_2 < 1. \tag{2.13}$$

Substituting (2.12) in (2.11) and using $\zeta\bar{\zeta} = \rho^2$, we get

$$\begin{aligned} w &= -\frac{Pc^2}{4\mu}\left(\sum_0^\infty \zeta^n\right)\left(\sum_0^\infty \left(\frac{\rho^2}{\zeta}\right)^n\right) \\ &\quad + A\ln\rho^2 + \sum_{-\infty}^{\infty} a_n\zeta^n + \sum_{-\infty}^{\infty} a_n\frac{\rho^{2n}}{\zeta^n} \\ &= 2a_0 + A\ln\rho^2 - \frac{Pc^2}{4\mu}\frac{1}{1-\rho^2} \\ &\quad + \sum_{n=1}^{\infty}\left[a_n + \frac{a_{-n}}{\rho^{2n}} - \frac{Pc^2}{4\mu}\frac{1}{1-\rho^2}\right]\left[\zeta^n + \frac{\rho^{2n}}{\zeta^n}\right]. \end{aligned}$$

Now applying the boundary conditions given by (2.13), we get

$$A\ln\rho^2 + 2a_0 - \frac{Pc^2}{4\mu}\frac{1}{1-\rho^2} = 0. \quad \text{on} \quad \rho = \rho_1, \rho_2, \quad \rho_1 < \rho_2.$$

and

$$a_n\rho^{2n} + a_{-n} - \frac{Pc^2}{4\mu}\frac{\rho^{2n}}{1-\rho^2} = 0. \quad \text{on} \quad \rho = \rho_1, \rho_2, \quad \rho_1 < \rho_2.$$

Solving uniquely for $A$ and $a_n$'s, we obtain

$$\begin{aligned} A &= \frac{Pc^2}{8\mu}\left[\frac{1}{1-\rho_2^2} - \frac{1}{1-\rho_1^2}\right] / \ln\left(\frac{\rho_2}{\rho_1}\right), \\ a_0 &= \frac{Pc^2}{16\mu}\left[\frac{\ln\rho_2}{1-\rho_2^2} - \frac{\ln\rho_1}{1-\rho_1^2}\right] / \ln\left(\frac{\rho_2}{\rho_1}\right). \end{aligned}$$

(2.14)

$$\begin{aligned} a_n &= \frac{Pc^2}{4\mu}\left[\frac{\rho_2^{2n}}{1-\rho_2^2} - \frac{\rho_1^{2n}}{1-\rho_1^2}\right] / \left(\rho_2^{2n} - \rho_1^{2n}\right), \quad n = 1, 2, 3, \cdots \\ a_{-n} &= \frac{Pc^2}{4\mu}\left[\frac{1}{1-\rho_2^2} - \frac{1}{1-\rho_1^2}\right] / \left(\rho_2^{2n} - \rho_1^{2n}\right), \quad n = 1, 2, 3, \cdots. \end{aligned}$$

The solution to the boundary value problem (2.2)-(2.4) is given by (2.11). (2.12) and (2.14). It can be shown that the resulting series in (2.12) is convergent in the domain under consideration.

## 2.5 Rate of flow

On using the complex form of the Green's theorem

$$\int_D \int \frac{\partial F}{\partial \bar{z}} dS = \frac{1}{2i} \int_{C_2 - C_1} F dz.$$

we have

$$\begin{aligned} R &= \int_D \int w dS \\ &= \frac{1}{2i} \int_{C_2 - C_1} z\bar{z} \left( \frac{P}{8\mu} \bar{z} - \omega'(z) \right) dz \\ &= \frac{1}{2i} \int_{\Gamma_2 - \Gamma_1} \frac{Pc^2}{8\mu} \frac{1}{(1-\zeta)(1-\bar{\zeta})} z'(\zeta) d\zeta \\ &- \frac{1}{2i} \int_{\Gamma_2 - \Gamma_1} \frac{c^2}{(1-\zeta)(1-\bar{\zeta})} \omega'(\zeta) d\zeta \end{aligned}$$

where $\Gamma_2$ and $\Gamma_1$ and respectively the circles $|\zeta| = \rho_2$ and $|\zeta| = \rho_1$.

After substitutions and some simplifications. we obtain

$$\begin{aligned} R &= \frac{Pc^4}{8\mu} \frac{1}{2i} \int_{\Gamma_2 - \Gamma_1} \frac{1}{(1-\zeta)^3 (1 - \frac{\rho^2}{\zeta})^2} d\zeta \\ &- \frac{c^2}{2i} \int_{\Gamma_2 - \Gamma_1} \frac{1}{(1-\zeta)(1-\frac{\rho^2}{\zeta})} \left\{ \frac{A}{\zeta} + \sum_{n=1}^{\infty} n a_n \zeta^{-n-1} - \sum_{n=1}^{\infty} n a_n \zeta^{-(n+1)} \right\} \end{aligned}$$

$$= \frac{P\pi c^4}{8\mu}\frac{\rho^2(2+\rho^2)}{(1-\rho^2)^4}\Big|_{\rho=\rho_1}^{\rho_2} - \pi c^2\left[\frac{A}{1-\rho^2} + \frac{1}{1-\rho^2}\sum_{n=1}^{\infty}[n(a_n\rho^{2n} - a_{-n})]\right]_{\rho=\rho_1}^{\rho_2}$$

which yields after some calculations

$$R = F + MS \tag{2.15}$$

where

$$F = \frac{P\pi c^4}{8\mu}\left\{\frac{\rho_2^4}{(1-\rho_2^2)^4} - \frac{\rho_1^4}{(1-\rho_1^2)^4} + \frac{(\rho_2^2-\rho_1^2)^2}{(1-\rho_1^2)^2(1-\rho_2^2)^2\ln(\rho_1/\rho_2)}\right\} \tag{2.16}$$

and

$$S = \sum_{n=1}^{\infty}\frac{n\rho_1^{2n}}{1-(\rho_1/\rho_2)^{2n}} \tag{2.17}$$

where

$$M = \frac{P\pi c^4}{2\mu}\frac{(\rho_2^2-\rho_1^2)^2}{(1-\rho_1^2)^2(1-\rho_2^2)^2}. \tag{2.18}$$

In fact, we can rewrite $S$ as

$$\begin{aligned} S &= \sum_{n=0}^{\infty} n\rho_1^{2n}\sum_{k=0}^{\infty}\left(\frac{\rho_1}{\rho_2}\right)^{2nk} \quad \text{since} \quad \frac{\rho_1}{\rho_2} < 1 \\ &= \sum_{k=0}^{\infty}\frac{\rho_1^{2k+2}}{\rho_2^{2k}}\sum_{n=0}^{\infty} n\left(\frac{\rho_1^{2k+2}}{\rho_2^{2k}}\right)^{n-1} \\ &= \sum_{k=0}^{\infty}\rho_1^2\left(\frac{\rho_1}{\rho_2}\right)^{2k}\left[1-\rho_1^2\left(\frac{\rho_1}{\rho_2}\right)^{2k}\right]^{-2}. \end{aligned} \tag{2.19}$$

## 2.6 Approximation and error analysis

Denoting by $S_N$ the sum of the first $N$ terms of $S$, we can now rewrite (2.15) as

$$R = F + M\, S_N + M\, E_N.$$

Here

$$E_N = \sum_{k=N}^{\infty} \rho_1^2 \left(\frac{\rho_1}{\rho_2}\right)^{2k} \left[1 - \rho_1^2 \left(\frac{\rho_1}{\rho_2}\right)^{2k}\right]^{-2}$$

represents the truncation error yielding an approximation $R_N$ for $R$ where $R_N = F + M\, S_N$. Now letting

$$r = \frac{\rho_1^2}{\rho_2^2} < 1 \quad \text{and} \quad \alpha = \rho_1^2,$$

we get

$$\begin{aligned} E_N &= \sum_{k=N}^{\infty} \frac{\alpha r^k}{(1 - \alpha r^k)^2} \\ &< \frac{\alpha}{(1 - \alpha r^N)^2} \sum_{k=N}^{\infty} r^k \\ &= \frac{\alpha}{(1 - \alpha r^N)^2} \frac{r^N}{1 - r}. \end{aligned}$$

We now proceed to find $N$ such that the truncation error in the evaluation of $R$ is less than $\varepsilon$, a prescribed number. Setting

$$\left(\frac{M}{1 - r}\right) \frac{\alpha r^N}{(1 - \alpha r^N)^2} < \varepsilon$$

we get

$$(\alpha r^N)^2 - (2+p)(\alpha r^N) + 1 > 0$$

where

$$p = \frac{M}{\varepsilon(1-r)}.$$

Noting that $\alpha r^N < 1$, we obtain

$$\begin{aligned} r^N &> \frac{2}{\alpha(2+p+\sqrt{(2+p)^2-4})} \\ &> \frac{1}{\alpha(2+p)} \end{aligned}$$

giving

$$N > \left[\ln\left(\frac{\rho_2^2}{\rho_1^2}\right)\right]^{-1} \left\{\ln\rho_1^2 + \ln\left[2 + \frac{4c^4\rho_2^2(\rho_2^2-\rho_1^2)}{(1-\rho_1^2)^2(1-\rho_2^2)}\frac{1}{\varepsilon}\right]\right\}. \tag{2.20}$$

If we choose $N$ given by (2.20), we will have

$$R_N - \varepsilon < R < R_N + \varepsilon.$$

## 2.7 Numerical results

For comparison purposes, we consider the following two cases:

(a) *Concentric Circles.*

The region $D$ is bounded by the two concentric circles

$$x^2 + y^2 = a^2, \qquad x^2 + y^2 = b^2, \qquad 0 < a < b.$$

It can be easily verified that

$$w(z) = A \ln z + E$$

where

$$A = \frac{P}{4\mu} \frac{(b^2 - a^2)}{\ln(b^2/a^2)}.$$

$$E = \frac{P}{8\mu} \left\{ \frac{a^2 \ln b - b^2 \ln a^2}{\ln b^2/a^2} \right\}.$$

when substituted in (2.9) satisfies (2.2)-(2.4). Further, the rate of flow $R_C$ per unit time per unit cross-section is given by

$$\frac{8\mu}{P\pi} R_c = b^4 - a^4 - \frac{(b^2 - a^2)^2}{\ln \frac{b}{a}}. \tag{2.21}$$

(b) *Eccentric Circles.*

The region $D$ is bounded by the eccentric circles (2.7) and the rate of flow $R$ per unit time per unit cross-section is given by (2.15). The series in (2.19) is truncated to $N$ terms where $N$ is determined by (2.20), thus assuring the error to be less than $\varepsilon$. In all calculations $\varepsilon = 10^{-6}$ was used. Also, $P/8\mu$ was taken to be 1 for all the calculations.

In both the above cases, the area of flow and the sum of the perimeters of the boundaries are held constant. In Tables I - V, we give the behaviour of $R_e$ as the inner boundary moves away from the concentric case. $h$ and $k$ are chosen such that (2.8) is satisfied. We find that $R_e$ increases as the $|h - k|$ increases and

for the same eccentricity. $R_e$ increases as the cross-section area increases. $a$ was chosen to be 1.0 for the calculations.

Table 2.1: Rate of flow $R_e$ vs $R_c$=2.098121

a = 1.000, b = 0.050, area = 3.134

| h-k | h | k | $R_e$ | N |
|---|---|---|---|---|
| 0.010000 | 99.749959 | 99.739959 | 2.098296 | 1 |
| 0.060000 | 16.624848 | 16.564848 | 2.104425 | 1 |
| 0.110000 | 9.067902 | 8.957902 | 2.119244 | 2 |
| 0.160000 | 6.233963 | 6.073963 | 2.141784 | 2 |
| 0.210000 | 4.749449 | 4.539449 | 2.172815 | 2 |
| 0.260000 | 3.835839 | 3.575839 | 2.211598 | 2 |
| 0.310000 | 3.216882 | 2.906882 | 2.257676 | 2 |
| 0.360000 | 2.769796 | 2.409796 | 2.310489 | 2 |
| 0.410000 | 2.431690 | 2.021690 | 2.396373 | 2 |
| 0.460000 | 2.167014 | 1.707014 | 2.433556 | 2 |
| 0.510000 | 1.954151 | 1.444151 | 2.502152 | 2 |
| 0.560000 | 1.779199 | 1.219199 | 2.574153 | 2 |
| 0.610000 | 1.632802 | 1.022802 | 2.648427 | 2 |
| 0.660000 | 1.508417 | 0.848417 | 2.723702 | 2 |
| 0.710000 | 1.401313 | 0.691313 | 2.798397 | 3 |
| 0.760000 | 1.307937 | 0.547937 | 2.871129 | 3 |
| 0.810000 | 1.225464 | 0.415464 | 2.939994 | 3 |
| 0.860000 | 1.151302 | 0.291302 | 3.002909 | 3 |
| 0.910000 | 1.081584 | 0.171584 | 3.057576 | 4 |

Table 2.2: Rate of flow $R_e$ vs $R_c$=1.337667

a = 1.000. b = 0.200. area = 3.016

| $\|h-k\|$ | h | k | $R_e$ | N |
|---|---|---|---|---|
| 0.010000 | 95.999583 | 95.989583 | 1.337857 | 2 |
| 0.060000 | 15.997490 | 15.937490 | 1.346150 | 3 |
| 0.110000 | 8.722628 | 8.612628 | 1.366210 | 3 |
| 0.160000 | 5.993143 | 5.833143 | 1.397838 | 4 |
| 0.210000 | 4.562238 | 4.352238 | 1.440719 | 4 |
| 0.260000 | 3.680614 | 3.420614 | 1.494418 | 4 |
| 0.310000 | 3.082346 | 2.772346 | 1.558385 | 4 |
| 0.360000 | 2.649193 | 2.289193 | 1.631946 | 4 |
| 0.410000 | 2.320527 | 1.910527 | 1.714312 | 5 |
| 0.460000 | 2.061988 | 1.601988 | 1.804572 | 5 |
| 0.510000 | 1.852559 | 1.342559 | 1.901696 | 5 |
| 0.560000 | 1.678524 | 1.118524 | 2.004535 | 5 |
| 0.610000 | 1.530307 | 0.920307 | 2.111826 | 6 |
| 0.660000 | 1.400530 | 0.740530 | 2.222197 | 7 |
| 0.710000 | 1.282208 | 0.572208 | 2.334175 | 8 |
| 0.760000 | 1.164196 | 0.404196 | 2.446191 | 12 |

Table 2.3: Rate of flow $R_e$ vs $R_c$=0.955380

a = 1.000. b = 0.300. area = 2.859

| $\|h-k\|$ | h | k | $R_e$ | N |
|---|---|---|---|---|
| 0.010000 | 90.999011 | 90.989011 | 0.955586 | 3 |
| 0.060000 | 15.160707 | 15.100707 | 0.964233 | 4 |
| 0.110000 | 8.261687 | 8.151687 | 0.985154 | 5 |
| 0.160000 | 5.671170 | 5.511170 | 1.018156 | 5 |
| 0.210000 | 4.311390 | 4.101390 | 1.062930 | 5 |
| 0.260000 | 3.471980 | 3.211980 | 1.119054 | 6 |
| 0.310000 | 2.900745 | 2.590745 | 1.185996 | 6 |
| 0.360000 | 2.485433 | 2.125433 | 1.263111 | 6 |
| 0.410000 | 2.168327 | 1.758327 | 1.349648 | 7 |
| 0.460000 | 1.916468 | 1.456468 | 1.444751 | 7 |
| 0.510000 | 1.709268 | 1.199268 | 1.547465 | 8 |
| 0.560000 | 1.532450 | 0.972450 | 1.656741 | 9 |
| 0.610000 | 1.374003 | 0.764003 | 1.771440 | 11 |
| 0.660000 | 1.217293 | 0.557293 | 1.890348 | 16 |

Table 2.4: Rate of flow $R_e$ vs $R_c$=0.395791

a = 1.000, b = 0.50, area = 2.356

| $\|h-k\|$ | h | k | $R_e$ | N |
|---|---|---|---|---|
| 0.010000 | 74.996653 | 74.986653 | 0.396014 | 7 |
| 0.060000 | 12.479870 | 12.419870 | 0.403840 | 8 |
| 0.110000 | 6.780704 | 6.670704 | 0.422783 | 9 |
| 0.160000 | 4.631591 | 4.471591 | 0.452693 | 10 |
| 0.210000 | 3.495333 | 3.285333 | 0.493331 | 11 |
| 0.260000 | 2.785630 | 2.525630 | 0.544372 | 12 |
| 0.310000 | 2.293302 | 1.983302 | 0.605412 | 14 |
| 0.360000 | 1.923428 | 1.563428 | 0.675965 | 16 |
| 0.410000 | 1.623202 | 1.213202 | 0.755475 | 22 |
| 0.460000 | 1.349322 | 0.889322 | 0.843315 | 31 |

Table 2.5: Rate of flow $R_e$ vs $R_c$=0.057245

a = 1.000, b = 0.750, area = 1.374

| $\|h-k\|$ | h | k | $R_e$ | N |
|---|---|---|---|---|
| 0.010000 | 43.737136 | 43.727136 | 0.057484 | 13 |
| 0.060000 | 7.213029 | 7.153029 | 0.062247 | 20 |
| 0.110000 | 3.825896 | 3.715896 | 0.073784 | 24 |
| 0.160000 | 2.493300 | 2.333300 | 0.092024 | 30 |
| 0.210000 | 1.707777 | 1.497777 | 0.116849 | 43 |

From the above five tables, we have established that the rate of flow in the concentric case is less than the rate of flow in all the eccentric cases considered. Further, as the inner circle moves away from the concentric position, the rate of flow increases.

## 2.8 Conclusions and comments

The above results establish that the rate of flow is not a maximum in the case of concentric circles and. in fact. the opposite is true. For a given area of flow and fixed circumference lengths. the rate of flow increases as the inner circle moves away from the position of concentric circles. Again. for a given area of flow but variable circumference lengths. the rate of flow decreases as the total perimeter increases. This suggests a boundary layer effect on the flow.

The mathematical problem described above has a large number of applications. One example is the simultaneous transport of oil and gas from oil and gas fields in the Arctic or on an ocean floor where both oil and gas are present at the same source. In such a case oil flows between the pipes while gas flows in the inner pipe. Although a new technology for building a pipe-in-pipe system is needed. the laying of one system is cheaper than laying two different pipes. The present system of heating oil at intervals to keep it flowing can be avoided. since the heat in the inside pipe is conserved and used to its maximum advantage. Ecologically and environmentally. accidents are less harmful since a burst in sections of the inner pipe will not lead to a spill while a burst in the outer pipe will not interrupt the gas flow and the oil spill may be recoverable.

There are a large number of applications involving Poisson's equations in engineering. A notable one is simultaneous transmission of data in co-axial cables.

The problem dealt with in this chapter is part of the general problem of mapping doubly connected regions onto an annulus. A well-known existence theorem which will not give a method of construction states that a doubly connected re-

gion (with arbitrary bounding curves) can always be mapped onto an annulus with the ratio of the radii being unique [11].

# Chapter 3

# Evaluation of a Double Integral over a Doubly Connected Region

In this chapter, we give a finite element method and its error estimate for evaluating double integrals over a smooth domain. The results are used to compare rates of flow of a viscous incompressible fluid in a pipe-in-pipe system with different doubly connected cross sections. These numerical results confirm an earlier conjecture regarding the rates of flow.

## 3.1 Introduction

In a number of problems in physics and engineering, measurable physical quantities depend on the evaluation of a double integral over a given domain $\Omega$. The domain may be simply or multiply connected and the geometry of the domain may not consist necessarily of commonly known curves. The solutions of such problems are usually not feasible by analytical methods. In this chapter, we give a finite element method for evaluating double integrals with smooth boundary in

a plane. with an error analysis. We are interested in the double integral

$$R = \iint_{\Omega} u(x, y)dxdy \tag{3.1}$$

where $\Omega$ is a smooth domain. We are also interested in the finite element approximation of $R$

$$R_h = \iint_{\Omega} u_h(x, y)dxdy$$

where $u_h(x, y)$ is the standard finite element solution to the Dirichlet problem $u_{xx} + u_{yy} = -1$ in $\Omega$ with $u \equiv 0$ on the two boundary curves.

We prepare to apply the techniques to find the rate of flow of a steady state incompressible viscous fluid flow in a pipe-in-pipe configuration. The analysis leads to a two-dimensional Dirichlet problem. Shivakumar and Ji [42] discuss the case where the region of cross-section of the pipes is bounded by two eccentric circles. They provide a proof to show that the rate of flow per unit cross-section per unit time is a minimum in the concentric case with the area enclosed by the bounding curves held constant. [27] gives estimates for similar problems for multiply connected cross-sections arising in the determination of torsional rigidity of beams.

In our numerical experiments using the techniques of this chapter. we discuss various doubly connected regions bounded by (a) two circles. (b) two ellipses. (c) a circle and an ellipse. and (d) an ellipse and a circle. In each case. the area bounded by each pair of the curves is kept constant.

## 3.2 Formulation of the problem

We are concerned with the double integral

$$R = \iint_{\Omega} u(x, y)dxdy \tag{3.2}$$

where the domain, if simply connected, is bounded by a smooth curve $\partial\Omega$. In (3.2) $u(x, y)$ is the solution of the Dirichlet problem

$$u_{xx} + u_{yy} = -1 \qquad in\ \Omega, \tag{3.3}$$

$$u = 0 \qquad on \qquad \partial\Omega$$

If the domain $\Omega$ is doubly connected, and bounded by two curves $\partial\Omega_1$, $\partial\Omega_2$, then Dirichlet problem is given by (3.3) and

$$u = 0 \qquad on \qquad \partial\Omega_1, \partial\Omega_2. \tag{3.4}$$

In (3.2), $R$ is the rate of slow and steady flow of an incompressible viscous fluid in a pipe whose cross-section is given by $\Omega$.

For the doubly connected region, Shivakumar and Ji [42] discuss the case where $\partial\Omega_1$ and $\partial\Omega_2$ are two eccentric circles. They prove that the rate of flow is a minimum when the circles are in the concentric position. It is conjectured that, in general, the rate of flow is a minimum in the symmetric case when $\partial\Omega_1$ and $\partial\Omega_2$ bound a fixed area. We will give numerical values for $R$ in the following cases, using the finite element method:

(a) two circles, (b) two ellipses, (c) a circle and an ellipse, (d) an ellipse and a

circle.

## 3.3 A finite element method for the double integral

In order to evaluate the double integrals (3.2), we can solve equation (3.3) using the standard finite element method, and then calculate (3.2) by substituting $u_h$, the finite element solution.

To give a brief description of the standard finite element method(FEM)[28], we first consider a decomposition $\Gamma_h$ on domain $\Omega$ such that

$$\overline{\Omega} \cong \tilde{\Omega} = \bigcup_i \overline{e}_i \qquad e_i \bigcap e_j = \emptyset, \qquad i \neq j.$$

where the element $e_i$ can be a triangle, quadrilateral or their mappings. The diameter of $e_i$ is denoted by $h_i$, and the largest diameter

$$h = \max_i \{h_i\}.$$

The finite element space of order $k$, $k \geq 0$, is

$$V^h = \{v \in C(\tilde{\Omega});\ v|_e \in P_e^k,\ \forall e \in \Gamma_h\}.$$

where $P^k$ represents the polynomials of degree $k$, and

$$V_0^h = V^h \bigcap H_0^1(\tilde{\Omega}).$$

In this chapter, $H^1(\Omega)$ stands for the standard Sobolev space(see [29] for details).

$H_0^1(\Omega)$ is the closure of $C_0^\infty(\Omega)$ with the measure of $H^1(\Omega)$. The semi-norm

$$|\omega|_{s,\Omega} = \left( \iint_\Omega \sum_{|\alpha|=s} |\frac{\partial^{|\alpha|}\omega}{\partial x^{\alpha_1} \partial y^{\alpha_2}}|^2 dxdy \right)^{\frac{1}{2}}.$$

with $\alpha = (\alpha_1, \alpha_2)$. $|\alpha| = \alpha_1 + \alpha_2$. $\alpha_1, \alpha_2$ are non-negative integers. and

$$\|\omega\|_{s,\Omega} = \left( \sum_{l=0}^{s} |\omega|_{l,\Omega}^2 \right)^{\frac{1}{2}}.$$

The FEM for (3.3) is to find the finite element solution $u_h \in V_0^h$ satisfying given boundary conditions and

$$\iint_\Omega \nabla u_h \nabla v dxdy = \iint_\Omega v dxdy. \qquad \forall v \in V_h. \tag{3.5}$$

where $\nabla$ represents the gradient operator

$$\nabla v = (v_x, v_y).$$

We calculate the approximate double integral

$$\tilde{R}_h = \iint_{\tilde{\Omega}} u_h dxdy. \tag{3.6}$$

where $\tilde{\Omega} = \bigcup_i \bar{e}_i$ and $e_i$'s are the finite elements.

Note that $u_h$ is a polynomial of degree $k$ for every element $e$, and we can calculate $\tilde{R}_h$. If we use the quadrature with the accuracy of order $k+1$(for

instance, we can use Gaussian quadrature with one Gaussian point at the center of the element for the case $k = 1$), then we can get the exact value of $\tilde{R}_h$.

The error estimate for the above approximate double integral can be obtained as follows.

**Theorem 3.1** *Suppose $u$ is the solution of (3.3), $u_h \in V_0^h$ is its finite element approximation defined in (3.5), the double integral $R$ and its approximation $R_h$ is defined in (3.2) and (3.6). $\Omega$ is a smooth domain satisfying*

$$measure(\Omega - \tilde{\Omega}) \sim O(h^2).$$

*where $\tilde{\Omega} = \bigcup_i e_i$ is the mesh for domain $\Omega$, then*

$$|R - \tilde{R}_h| \sim O(h^2) \tag{3.7}$$

Proof: Suppose $\varphi$ is the solution of the auxiliary problem

$$\begin{cases} \varphi_{xx} + \varphi_{yy} = -1. & in\ \Omega. \\ \varphi = 0. & on\ \partial\Omega. \end{cases}$$

Then, by[28], for $k \geq 1$

$$|\varphi|_{k+1,\Omega} \leq C_1 |1|_{k-1,\Omega} \leq C_1 \sqrt{measure(\Omega)} = \overline{C}_1,$$

and there is $\varphi_I \in V_h$ such that

$$|\varphi - \varphi_I|_{1,\Omega} \leq C_2 h^k |\varphi|_{k+1,\Omega} \leq C_2 \overline{C}_1 h^k.$$

Note that $\varphi_I \in V_h \subset H_0^1(\Omega)$.

$$\iint_\Omega \nabla(u - u_h)\nabla\varphi_I dxdy = 0.$$

We consider

$$|R - \tilde{R}_h| \leq |R - R_h| + |R_h - \tilde{R}_h|.$$

where $R_h = \iint_\Omega u_h dxdy$.

By Schwarz inequality, we have

$$\begin{aligned}
|R - R_h| &= \left|\iint_\Omega (u - u_h)dxdy\right| \\
&= \left|\iint_\Omega (u - u_h)(-\varphi_{xx} - \varphi_{yy})dxdy\right| \\
&= \left|\iint_\Omega \nabla(u - u_h)\nabla\varphi dxdy\right| \qquad (3.8) \\
&= \left|\iint_\Omega \nabla(u - u_h)\nabla(\varphi - \varphi_I)dxdy\right| \\
&\leq \|u - u_h\|_1 \|\varphi - \varphi_I\|_1 \\
&\leq \overline{C}_1 C_2 h^k \|u - u_h\|_1.
\end{aligned}$$

It is well known[28] that

$$\|u - u_h\|_1 \leq \|u - u_I\|_1 \leq C_2 h^k \|u\|_{k-1}. \qquad (3.9)$$

Therefore,

$$|R - R_h| \sim O(h^{2k}). \quad k \geq 1. \qquad (3.10)$$

We also have

$$\begin{aligned} |R_h - \tilde{R}_h| &= \left| \iint_{\Omega - \tilde{\Omega}} u_h dx dy \right| \\ &\leq \iint_{\Omega - \tilde{\Omega}} |u_h| dx dy \\ &\leq measure\{\Omega - \tilde{\Omega}\} \|u_h\|_\infty \\ &\sim O(h^2). \end{aligned} \tag{3.11}$$

Hence, (3.7) follows from (3.10) and (3.11).

Remark: In this chapter, $R$ is assumed to be a double integral of the solution of a special Poisson equation. If $u$ in (3.2) is a solution of a general elliptic equation of second order

$$a(u, v) = (f, v). \qquad \forall v \in H_1^0(\Omega).$$

then Theorem 3.1 can be extended with a proof similar to the above argument. Here

$$a(u, v) = \iint_\Omega \left( \sum_{i,j=1}^{2} a_{ij} \partial_i u \partial_j v + \sum_{i=1}^{2} b_i \partial_i u v + \gamma u v \right) dx dy.$$

and

$$(f, v) = \iint_\Omega f v dx dy.$$

where $\partial_1 = \frac{\partial}{\partial x}$, $\partial_2 = \frac{\partial}{\partial y}$.

## 3.4 Numerical results and conclusions

For the Dirichlet problem

$$u_{xx} + u_{yy} = -1. \qquad in \quad \Omega$$

and

$$u = 0 \quad on \quad \partial\Omega_1. \quad \partial\Omega_2.$$

we evaluate the approximate double integral

$$\tilde{R}_h = \int\int_{\tilde{\Omega}} u_h dxdy$$

using the finite element solution $u_h$.

In the numerical experiments, we use triangular decomposition with $h = 0.05$ and piecewise linear finite element space(first order FE space, $k = 1$). The quadrature is chosen to be Gaussian quadrature with one Gaussian point at the barycenters of elements. By Theorem 3.1, $\tilde{R}_h$ is the approximation of $R$ with the accuracy of $O(h^2)$. The condition required by the theorem, $measure\{\Omega - \tilde{\Omega}\} \sim O(h^2)$, is satisfied in all the cases in which we apply the result in our computation. For example, using uniform triangle element mesh on a unit circle gives the difference of the order $O(h^2)$:

$$measure\{\Omega - \tilde{\Omega}\} = \pi - \pi\frac{h\sqrt{1 - h^2/4}}{\sin^{-1}(h/2)} \sim \pi h^2/4 \sim O(h^2).$$

It can be similarly verified for the regions in the following computation.

In the case of slow and steady incompressible viscous flow in a pipe-in-pipe, $\tilde{R}_h$ represents the approximate rate of flow per unit time per unit cross-section. We consider the region with a fixed area, $(1-\rho^2)\pi$, $\rho < 1$, bounded by

$$\partial\Omega_1 : \frac{x^2}{a^2} + \frac{y^2}{(1/a)^2} = 1. \qquad \partial\Omega_2 : \frac{(x-d)^2}{b^2} + \frac{y^2}{(1/b)^2} = \rho^2.$$

We calculate the rates of flow for the following 9 cases in three groups, as $a = 1, b = 1, \frac{5}{4}, \frac{3}{2}$; $a = \frac{5}{4}, b = 1, \frac{5}{4}, \frac{3}{2}$; $a = \frac{3}{2}, b = 1, \frac{5}{4}, \frac{3}{2}$, and $\rho = 0.2$.

Note that the areas enclosed between $\partial\Omega_1$ and $\partial\Omega_2$ in all the cases are kept constant, $0.96\pi$.

Table 3.1: Rates of flow $R_e$'s with $\partial\Omega_1 : \quad x^2 + y^2 = 1$

| $\partial\Omega_2$ | $(x-d)^2 + y^2 = 0.2^2$ | $\frac{(x-d)^2}{(5/4)^2} + \frac{y^2}{(4/5)^2} = 0.2^2$ | $\frac{(x-d)^2}{(3/2)^2} + \frac{y^2}{(2/3)^2} = 0.2^2$ |
|---|---|---|---|
| d | $R_{1a}$ | $R_{1b}$ | $R_{1c}$ |
| 0.0000 | 0.16601048 | 0.16421933 | 0.16025561 |
| 0.0500 | 0.16674468 | 0.16493076 | 0.16094677 |
| 0.1000 | 0.16893996 | 0.16705783 | 0.16301312 |
| 0.1500 | 0.17257436 | 0.17057884 | 0.16643316 |
| 0.2000 | 0.17761125 | 0.17545760 | 0.17117111 |
| 0.2500 | 0.18399926 | 0.18164343 | 0.11717692 |
| 0.3000 | 0.19167224 | 0.18907110 | 0.18438624 |
| 0.3500 | 0.20054919 | 0.19766094 | 0.19272067 |
| 0.4000 | 0.21053419 | 0.20731875 | 0.20208782 |
| 0.4500 | 0.22151642 | 0.21793610 | 0.21238177 |
| 0.5000 | 0.23337013 | 0.22939046 | 0.22348365 |
| 0.5500 | 0.24595493 | 0.24154589 | 0.23526268 |
| 0.6000 | 0.25911599 | 0.25425380 | 0.24757783 |
| 0.6500 | 0.27268466 | 0.26735440 | 0.26028022 |

Table 3.2: Rates of flow $R_e$'s with $\partial\Omega_1$ : $\frac{x^2}{(5/4)^2} + \frac{y^2}{(4/5)^2} = 1$

| $\partial\Omega_2$ | $(x-d)^2 + y^2 = 0.2^2$ | $\frac{(x-d)^2}{(5/4)^2} + \frac{y^2}{(4/5)^2} = 0.2^2$ | $\frac{(x-d)^2}{(3/2)^2} + \frac{y^2}{(2/3)^2} = 0.2^2$ |
|---|---|---|---|
| d | $R_{2a}$ | $R_{2b}$ | $R_{2c}$ |
| 0.0000 | 0.16243616 | 0.15879773 | 0.15335999 |
| 0.0500 | 0.16289040 | 0.15924697 | 0.15380700 |
| 0.1000 | 0.16424981 | 0.16059133 | 0.15514500 |
| 0.1500 | 0.16650426 | 0.16282079 | 0.15736300 |
| 0.2000 | 0.16963698 | 0.16591857 | 0.16044500 |
| 0.2500 | 0.17362440 | 0.16986127 | 0.16436701 |
| 0.3000 | 0.17843620 | 0.17461875 | 0.16909900 |
| 0.3500 | 0.18403541 | 0.18015414 | 0.17460400 |
| 0.4000 | 0.19037817 | 0.18642393 | 0.18083800 |
| 0.4500 | 0.19741397 | 0.19337787 | 0.18775199 |
| 0.5000 | 0.20508543 | 0.20095906 | 0.19528700 |
| 0.5500 | 0.21332847 | 0.20910391 | 0.20338200 |
| 0.6000 | 0.22072257 | 0.21774232 | 0.21196499 |
| 0.6500 | 0.23123926 | 0.22679769 | 0.22096001 |

Table 3.3: Rates of flow $R_e$'s with $\partial\Omega_1$ : $\frac{x^2}{(3/2)^2} + \frac{y^2}{(2/3)^2} = 1$

| $\partial\Omega_2$ | $(x-d)^2 + y^2 = 0.2^2$ | $\frac{(x-d)^2}{(5/4)^2} + \frac{y^2}{(4/5)^2} = 0.2^2$ | $\frac{(x-d)^2}{(3/2)^2} + \frac{y^2}{(2/3)^2} = 0.2^2$ |
|---|---|---|---|
| d | $R_{3a}$ | $R_{3b}$ | $R_{3c}$ |
| 0.0000 | 0.15169299 | 0.14767602 | 0.14240605 |
| 0.0500 | 0.15193447 | 0.14791816 | 0.14265074 |
| 0.1000 | 0.15265770 | 0.14864333 | 0.14338353 |
| 0.1500 | 0.15385883 | 0.14984769 | 0.14460056 |
| 0.2000 | 0.15553147 | 0.15152489 | 0.14629541 |
| 0.2500 | 0.15766667 | 0.15366602 | 0.14845908 |
| 0.3000 | 0.16025296 | 0.15625958 | 0.15107997 |
| 0.3500 | 0.16327621 | 0.15929151 | 0.15414388 |
| 0.4000 | 0.16671972 | 0.16274512 | 0.15763398 |
| 0.4500 | 0.17056416 | 0.16660117 | 0.16153079 |
| 0.5000 | 0.17478755 | 0.17083767 | 0.16581215 |
| 0.5500 | 0.17936523 | 0.17543077 | 0.17045322 |
| 0.6000 | 0.18426988 | 0.18035106 | 0.17542642 |
| 0.6500 | 0.18947136 | 0.18557061 | 0.18070145 |

From the above tables, we first observe that in all the cases, the rate of flow per unit cross-section per unit time, evaluated by the double integral, increases as the eccentricity of the annulus increases, and attains its minimum in the concentric case (when $d = 0$). This conclusion agrees well with the results in [27].

We can also notice that for a fixed eccentricity $d$ and a fixed compression constant $\mu_1$ for $\partial\Omega_1$, the outer ellipse($\mu = 1$ for circles), the rate of flow decreases as the compression constant of $\partial\Omega_2$, $\mu_2$, increases. We can also get exactly the same information if we switch the roles of $\partial\Omega_1$ and $\partial\Omega_2$ by looking at the columns of the three tables for a fixed $d$. To summarize, we may conclude the following.

- i). For a fixed eccentricity $d$ and a fixed ellipse(outer or inner), the rate of flow decreases as the compression constant of the other ellipse increases. It attains a maximum value when the latter one is a circle ($\mu = 1$).

- ii). For any fixed eccentricity $d$, the rate of flow decreases as the sum of the compression constants of the two ellipses, $\mu_1 + \mu_2$ increases. For a fixed value of $\mu_1 + \mu_2$, $\mu_1$ plays more dominant role in the two, e.g., if $\mu_1 + \mu_2 = \mu_1^* + \mu_2^* =$ a fixed number, then the corresponding rates of flow, $R_h > R_h^*$ if $\mu_1 > \mu_1^*$, or else $R_h < R_h^*$.

- iii). The area enclosed by $\partial\Omega_1$ and $\partial\Omega_2$ is $0.96\pi$. We now consider the rate of flow over a simply connected region $D : |z| \leq c$, where $c = \sqrt{0.96}$. The solution of the Dirichlet problem in $D$ can easily be obtained:

$$w(r, \theta) = -\frac{1}{4}(r^2 - c^2), \quad r \leq c.$$

and the rate of flow:

$$R_D = \frac{1}{8}\pi c^4 = 0.36191147.$$

The numerical results in Table 1. column $R_{1a}$. show that the rates of flow over the cross section bounded by two eccentric circles increase as the eccentricity $d$ increases. but they are bounded above by $R_D = 0.36191147$.

# Chapter 4

# Non-singularity of Matrices of Certain Sign Distributions

## 4.1 Introduction

Nonsingularity of matrices plays a key role in the solution of linear systems, matrix computation and numerical analysis. A large variety of problems arising in computational mechanics, fluid dynamics and material engineering, modelled by using difference equations or finite element methods, demand the matrices be non-singular for the numerical approaches to be convergent.

Two typical criteria for a non-singularity test are (i) non-vanishing determinant: and (ii) diagonal dominance. Some disadvantages are well known: criterion (i) costs too much computing time and (ii) is too strict for most application problems to fit. Since a large number of matrices resulting from physical models have certain structures or sign distributions, consideration of non-singularity related to sign distributions becomes useful and effective.

Nonsingularity related to M-matrices and positive matrices, two classes of matrices with fixed sign distributions, was first studied by M. Fiedler[14] , K. Fan

and A.S. Householder[13]. J. Drew and C.R. Johnson[12] consider Hessenberg and Hadamard matrices. In recent years, sign-non-singular matrices have been extensively explored. A matrix $\mathbf{B}$ is called a sign-non-singular matrix if its entries are among $\{1, 0, -1\}$ and any other matrix $\mathbf{A}$ with the same sign distribution as $\mathbf{B}$'s is non-singular. If, in addition, the sign distribution of the inverse of $\mathbf{A}$ is the same as $\mathbf{B}$'s for all $\mathbf{A}$, then $\mathbf{B}$ is called a strong sign-non-singular matrix. Although the sign-non-singular matrices have received considerable attention, most of the results remain theoretical and specific sign distributions are barely studied thoroughly for practical purposes, and few computable conditions are given on non-singularity of matrices of certain sign distributions.

In this chapter, we impose easily computable sufficient conditions for matrices of the following two different sign distributions:

$$\begin{pmatrix} + & + & + & + & + & \cdots \\ - & + & + & + & + & \cdots \\ + & - & + & + & + & \\ - & + & - & + & + & \\ + & - & + & - & + & \\ \vdots & \vdots & & & & \ddots \end{pmatrix} \qquad \begin{pmatrix} + & + & + & + & + & \cdots \\ + & + & + & + & + & \cdots \\ - & + & + & + & + & \\ + & - & + & + & + & \\ - & + & - & + & + & \\ \vdots & \vdots & & & & \ddots \end{pmatrix}$$

Sign Distribution 1(SD1) Sign Distribution 2(SD2)

The matrices in this chapter are assumed to be square matrices of arbitrary but fixed size, and contain no zero entries.

This work is mainly motivated by the problems of viscous flow in pipes whose cross-sections are doubly connected regions[35] in which the velocity of the fluid in the direction of the axis of the pipe satisfies Poisson's equation with homogeneous boundary conditions. The solution of the problem can be expressed as a truncated infinite series which can be found by solving a linear system whose coefficient matrix has SD1.

## 4.2 Main Theorem

The sign distributions 1 and 2 can be formulated for the matrix $\mathbf{A} = (a_{ij})_{m\times m}$ by

SD1: For $i \le j$. $a_{ij} > 0$: For $i > j$. $(-1)^{i-j}a_{ij} > 0$.
SD2: For $i \le j$. $a_{ij} > 0$: For $i > j$. $(-1)^{i-j+1}a_{ij} > 0$.

where $i, j = 1, 2, \ldots, m$. We will prove that, under certain stated conditions, a matrix with either of the above sign distributions is non-singular.

For convenience, we first define for the matrices of both cases SD1 and SD2, the following quantities:

$$\begin{cases} \nu_{ij} &= a_{ij} - \sum_{k=j+1}^{m} a_{ik}, \ for\ i \le j. \\ \mu_{ij} &= a_{ij} - \sum_{k=1}^{i-1} a_{kj}, \ for\ i \le j. \\ \omega_{ij}^{l} &= \min\left\{|a_{ij}| - \left|\sum_{k=i+1}^{l} a_{kj}\right|, \ \left|\sum_{k=i+1}^{l} a_{kj}\right| - |a_{l+1\,j}|\right\}, \ for\ j < i < l < m. \\ \delta_{ij}^{l} &= \left|\sum_{k=j+1}^{l} a_{ik}\right| - |a_{ij}|, \ for\ j < l \le i. \end{cases}$$

(4.1)

**Theorem 4.1 (Sign Distributions 1 and 2)** *Let* $\mathbf{A} = (a_{ij})_{m\times m}$, $m > 0$, *be a real matrix of Sign Distribution 1 or 2 satisfying*

$$\begin{cases} For\ i \le j. & (A1)\ \nu_{ij} > 0 & (A2)\ \mu_{ij} > \sum_{k=j+1}^{m} \mu_{ik} > 0; \\ For\ i > j. & (A3)\ \omega_{ij}^{l} > 0\ (j < i < l) & (A4)\ \delta_{ij}^{l} > \sum_{k=i+1}^{m} \delta_{kj}^{l} > 0\ (j < l \le i). \end{cases}$$

(4.2)

*Then* $\mathbf{A}$ *is non-singular.*

We give below an example of a matrix $\mathbf{A}$ which satisfies the conditions in Theorem 4.1.

Let $\mathbf{A} = (a_{ij})_{m\times m}$ be given by

$$a_{ij} = \begin{cases} \frac{1}{2^{j-1}} & i \leq j. \\ \left(-\frac{1}{2}\right)^{i-j} & i > j. \end{cases}$$

i.e.

$$\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2^2} & \frac{1}{2^3} & \cdots & \frac{1}{2^{m-1}} \\ -\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2^2} & \cdots & \\ \frac{1}{2^2} & -\frac{1}{2} & 1 & \frac{1}{2} & & \\ \vdots & \vdots & & \ddots & & \\ \left(-\frac{1}{2}\right)^{m-1} & & & & -\frac{1}{2} & 1 \end{pmatrix}$$

It is easy to verify that $\mathbf{A}$ satisfies $(A1)-(A4)$ and SD1. in Theorem 4.1. hence it is non-singular. although it is not diagonally dominant.

Remarks:

1. In Theorem 4.1. conditions $(A1)$ and $(A3)$ show a decreasing absolute value for upper trianglular entries of $\mathbf{A}$ along the horizontal direction and of lower trianglular entries along the vertical downward direction respectively: conditions $(A2)$ and $(A4)$ may be considered as second order distribution properties. which also reflect a similar trend along horizontal and vertical directions. Basically. conditions $(A1)-(A4)$ describe a scattering distribution along the horizontal and vertical directions for any element of $\mathbf{A}$.

2. All the conditions that appear in Theorem 4.1 are easily computable because they involve only additions and logical operations.

## 4.3 Proof

To formulate the problem, we denote by $\hat{\mathbf{A}} = (\hat{a}_{ij})_{n\times n}$ the matrix obtained by eliminating the last row of $\mathbf{A} = (a_{ij})_{(n+1)\times(n+1)}$, in the sense that

$$\begin{pmatrix} \hat{\mathbf{A}} & V_{n\times 1} \\ O_{1\times n} & a_{n+1\,n+1} \end{pmatrix} = \mathbf{A}\,\mathbf{L}^{(n+1)}.$$

where $\mathbf{L}^{(n+1)}$ is a $(n+1)\times(n+1)$ non-singular lower triangular matrix due to Gaussian transformation. $O^t_{1\times n}$ is a zero row vector and $V_{n\times 1}$ is a column vector. The elements of $\hat{\mathbf{A}}$ are given by:

$$a^{(1)}_{ij} = a_{ij} - \frac{a_{i\,n+1}\,a_{n+1\,j}}{a_{n+1\,n+1}}, \qquad i,j = 1,2,\ldots,n.$$

As a consequence of elimination process, $\hat{\mathbf{A}}$ and $\mathbf{A}$ are both singular or both non-singular. Correspondingly, the quantities defined by (4.1) for $\mathbf{A}$ can be written for $\hat{\mathbf{A}}$ with a little modification as

$$\begin{cases} \hat{\nu}_{ij} &= \hat{a}_{ij} - \sum_{k=j+1}^{n} \hat{a}_{ik},\ for\ i \le j, \\ \hat{\mu}_{ij} &= \hat{a}_{ij} - \sum_{k=1}^{i-1} \hat{a}_{kj},\ for\ i \le j; \\ \hat{\omega}^{l}_{ij} &= \min\left\{|\hat{a}_{ij}| - \left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right|,\ \left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right| - |\hat{a}_{l+1\,j}|\right\},\ for\ j < i < l. \\ \hat{\delta}^{l}_{ij} &= \left|\sum_{k=j+1}^{l} \hat{a}_{ik}\right| - |\hat{a}_{ij}|,\ for\ j < l \le i. \end{cases}$$

(4.3)

We give a proof for Theorem 4.1 for the case of SD1 only, to illustrate the method based on mathematical induction. The proof for the case of SD2 can be similarly shown by parallel arguments.

**Proof of Theorem 4.1:** (For the case of SD1)

(a) It is trivial that Theorem 4.1 is true for the cases of $m = 1$ and $m = 2$.

(b) We now assume that Theorem 4.1 holds for $m = n$, i.e., any $n \times n$ matrix $\mathbf{A}^{(n)}$ that has SD1 and satisfies conditions $(A1) - (A4)$ is non-singular.

(c) For $m = n + 1$, we suppose that the matrix $\mathbf{A}^{(n+1)} = (a_{ij})_{(n+1)\times(n+1)}$ has SD1 and satisfies conditions $(A1) - (A4)$. We will next show that $\mathbf{A}^{(n+1)}$ can be reduced to an $n$ by $n$ matrix that also has SD1 and satifies conditions $(A1) - (A4)$, using a non-singular transformation. This allows us to conclude that $\mathbf{A}^{(n+1)}$ is nonsingular and completes the proof by induction.

Suppose that $\mathbf{A}^{(n+1)} = (a_{ij})_{(n+1)\times(n+1)}$ has SD1 and satisfies the following conditions:

For $i \leq j$

$(A1)$ $\nu_{ij} > 0$

$(A2)$ $\mu_{ij} > \sum_{k=j+1}^{n+1} \mu_{ik} > 0.$

For $i > j$

$(A3)$ $\omega_{ij}^{l} > 0 \ \ (j < i < l)$

$(A4)$ $\delta_{ij}^{l} > \sum_{k=i+1}^{n+1} \delta_{kj}^{l} > 0 \ \ (j < l \leq i).$

Consider the matrix $\hat{\mathbf{A}}^{(n)} = (\hat{a}_{ij})_{n\times n}$, produced from $\mathbf{A}^{(n+1)}$ by Gaussian elimination as discussed above. In order to conclude that $\mathbf{A}^{(n+1)}$ is nonsingular, we need to show that $\hat{\mathbf{A}}^{(n)}$ preserves SD1 as given by $\hat{a}_{ij} > 0$ for $i \leq j$ and $(-1)^{i+j}\hat{a}_{ij} > 0$ for $i \geq j$, and in addition satisfies the following conditions:

For $i \leq j$

$(B1)$ $\hat{\nu}_{ij} > 0$

$(B2)$ $\hat{\mu}_{ij} > \sum_{k=j+1}^{n} \hat{\mu}_{ik} > 0.$

For $i > j$

$(B3)$ $\hat{\omega}_{ij}^{l} > 0 \ \ (j < i < l)$

$(B4)$ $\hat{\delta}_{ij}^{l} > \sum_{k=i+1}^{n} \hat{\delta}_{kj}^{l} > 0 \ \ (j < l \leq i).$

Recall that

$$\hat{a}_{ij} = a_{ij} - \frac{a_{i\,n+1}a_{n+1\,j}}{a_{n+1\,n+1}}.$$

To prove that SD1 is preserved by $\mathbf{A}^{(n)}$, we note that for $i \leq j \leq n$,

$$\begin{aligned} \hat{a}_{ij} &\geq a_{ij} - a_{i\,n+1}\left|\frac{a_{n+1\,j}}{a_{n+1\,n+1}}\right| \\ &> a_{ij} - a_{i\,n+1} \qquad \textit{by} \ \ (A4) \\ &> 0 \qquad \textit{by} \ \ (A1). \end{aligned}$$

Similarly, for $j < i \leq n$,

$$\begin{aligned} (-1)^{i+j}\hat{a}_{ij} &\geq (-1)^{i+j}a_{ij} - (-1)^{i+j}a_{n+1\,j}\frac{a_{i\,n+1}}{a_{n+1\,n+1}} \\ &\geq (-1)^{i+j}a_{ij} - (-1)^{i+j}a_{n+1\,j} \qquad \textit{by} \ \ (A2) \\ &> 0 \qquad \textit{by} \ \ (A3). \end{aligned}$$

Proof of (B1): (B1) follows from

$$\begin{aligned} \hat{\nu}_{ij} &= \hat{a}_{ij} - \sum_{k=j+1}^{n} \hat{a}_{ik} \\ &= a_{ij} - \sum_{k=j+1}^{n} a_{ik} - \frac{a_{i\,n+1}}{a_{n+1\,n+1}}\left(a_{n+1\,j} - \sum_{k=j+1}^{n} a_{n+1\,k}\right) \end{aligned}$$

$$> \quad a_{ij} - \sum_{k=j-1}^{n+1} a_{ik} \qquad by\ (A4)$$

$$= \quad \nu_{ij} > 0 \qquad by\ (A1).$$

Proof of (B3): It is easy to verify $(B3)$ if $l = j + 1$. From (4.3), we have

$$\hat{\omega}_{ij}^{l} = \min\left\{|\hat{a}_{ij}| - \left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right|, \left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right| - |\hat{a}_{l+1\,j}|\right\}.$$

Evaluating the two quantities in $\hat{\omega}_{ij}^{l}$ separately, we have

$$|\hat{a}_{ij}| - \left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right|$$

$$= (-1)^{i+j}\hat{a}_{ij} - (-1)^{i+j+1} \sum_{k=i+1}^{l} \hat{a}_{kj}$$

$$= (-1)^{i+j}a_{ij} - (-1)^{i+j+1} \sum_{k=i+1}^{l} a_{kj} - (-1)^{i+j}a_{n+1\,j}\frac{\sum_{k=i}^{l} a_{k\,n+1}}{a_{n-1\,n+1}}$$

$$> (-1)^{i+j} \sum_{k=i}^{l} a_{kj} - |a_{n+1\,j}| \qquad by\ (A2)$$

$$\geq (-1)^{i+j} \sum_{k=i}^{l} a_{kj} - |a_{l+1\,j}| \qquad by\ (A4)$$

$$\geq \omega_{ij}^{l} > 0 \qquad by\ (A3)$$

and

$$\left|\sum_{k=i+1}^{l} \hat{a}_{kj}\right| - |\hat{a}_{l+1\,j}|$$

$$
\begin{aligned}
&= (-1)^{i+j+1} \sum_{k=i+1}^{l} \hat{a}_{kj} - (-1)^{l+j+1} \hat{a}_{l+1\,j} \\
&= (-1)^{i+j+1} \sum_{k=i+1}^{l} a_{kj} - (-1)^{l+j+1} a_{l+1\,j} - \\
&a_{n+1\,j} \frac{(-1)^{i+j+1} \sum_{k=i+1}^{l} a_{k\,n+1} - (-1)^{l+j+1} a_{l+1\,n+1}}{a_{n+1\,n+1}} \\
&> (-1)^{i+j+1} \sum_{k=i+1}^{l+1} a_{kj} - |a_{n+1\,j}| \qquad \textit{by (A2)} \\
&> (-1)^{i+j+1} \sum_{k=i+1}^{l+1} a_{kj} - |a_{l-2\,j}| \qquad \textit{by (A4)} \\
&\geq \omega_{ij}^{l+1} > 0 \qquad \textit{by (A3).}
\end{aligned}
$$

Now combining the results, we have $\hat{\omega}_{ij}^{l} > 0$.

Proof of (B2): Now $\hat{\mu}_{ij} > 0$ since

$$
\begin{aligned}
\hat{\mu}_{ij} &= \hat{a}_{ij} - \sum_{k=1}^{i-1} \hat{a}_{kj} \\
&= a_{ij} - \sum_{k=1}^{i-1} a_{kj} - \frac{a_{n+1\,j}}{a_{n+1\,n+1}} \left( a_{i\,n+1} - \sum_{k=1}^{i-1} a_{k\,n+1} \right) \\
&> \mu_{ij} - \mu_{i\,n+1} \qquad \textit{by (A2), (A4)} \\
&> 0 \qquad \textit{by (A2).}
\end{aligned}
$$

Using the positivity of $\hat{\mu}_{ij}$, the full result follows from

$$\begin{aligned}
&\hat{\mu}_{ij} - \sum_{k=j+1}^{n} \hat{\mu}_{ik} \\
&= \left(\hat{a}_{ij} - \sum_{k=1}^{i-1} \hat{a}_{kj}\right) - \sum_{k=j+1}^{n} \left(\hat{a}_{ik} - \sum_{p=1}^{i-1} \hat{a}_{pk}\right) \\
&= \mu_{ij} - \sum_{k=j+1}^{n} \mu_{ik} - \mu_{i\,n+1} \cdot \frac{a_{n+1\,j} - \sum_{k=j+1}^{n} a_{n+1\,k}}{a_{n+1\,n+1}} \\
&> \mu_{ij} - \sum_{k=j+1}^{n+1} \mu_{ik} \qquad \textit{by (.42), (.44)} \\
&> 0 \qquad \textit{by (.42).}
\end{aligned}$$

Proof of (B4):

$$\begin{aligned}
\delta_{ij}^{l} &= \left|\sum_{k=j-1}^{l} \hat{a}_{ik}\right| - |\hat{a}_{ij}| = (-1)^{i+l} \sum_{k=j+1}^{l} \hat{a}_{ik} - (-1)^{i+j}\hat{a}_{ij} \\
&\geq (-1)^{i+l} \sum_{k=j+1}^{l} a_{ik} - (-1)^{i+j} a_{ij} \\
&\quad - \left|\frac{a_{i\,n+1}}{a_{n+1\,n+1}}\right| \left((-1)^{n+l+1} \sum_{k=j+1}^{l} a_{n+1\,k} - (-1)^{n+j+1} a_{n+1\,j}\right) \\
&> \delta_{ij}^{l} - \delta_{n+1\,j}^{l} \qquad \textit{by (.42), (.44)} \\
&> 0 \qquad \textit{by (.44).}
\end{aligned}$$

and,

$$\begin{aligned}
\hat{\delta}^l_{ij} - \sum_{k=i+1}^{n} \hat{\delta}^l_{kj} &= \left|\sum_{k=j+1}^{l} \hat{a}_{ik}\right| - |\hat{a}_{ij}| - \sum_{k=i+1}^{n} \left(\left|\sum_{p=j+1}^{l} \hat{a}_{kp}\right| - |\hat{a}_{kj}|\right) \\
&= \left((-1)^{i+l} \sum_{k=j+1}^{l} \hat{a}_{ik} - (-1)^{i+j}\hat{a}_{ij}\right) - \sum_{k=i+1}^{n} \left((-1)^{k+l} \sum_{p=j+1}^{l} \hat{a}_{kp} - (-1)^{k+j}\hat{a}_{kj}\right) \\
&= \delta^l_{ij} - \sum_{k=i+1}^{n} \delta^l_{kj} - \delta^l_{n+1\,j} \cdot \frac{\sum_{k=i+1}^{n}(-1)^{n+k+1}a_{k\,n+1} - (-1)^{n+i-1}a_{i\,n+1}}{a_{n+1\,n+1}} \\
&> \delta^l_{ij} - \sum_{k=i+1}^{n+1} \delta^l_{kj} \qquad \textit{by (42), (44)} \\
&> 0 \qquad \textit{by (44)}.
\end{aligned}$$

□

## 4.4 Applications

We now consider the application problem of viscous steady flow. From the optimal rate of flow point of view[44], we consider in this application, the region bounded by the outer circle: $\partial\Omega_1 : x^2 + y^2 = a^2$, and the inner ellipse $\partial\Omega_2 : \frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 1$, with $\alpha < \beta < a$(as in Figure 1), instead of the physical configuration used in [35] whose outer curve is an ellipse and inner curve is a circle.

The infinite series solution for the velocity of the flow in the pipe, which is geometrically convergent, is truncated, and the problem reduces to a linear system associated with a finite matrix.

For the new configuration, we derive the matrix using arguments and manip-

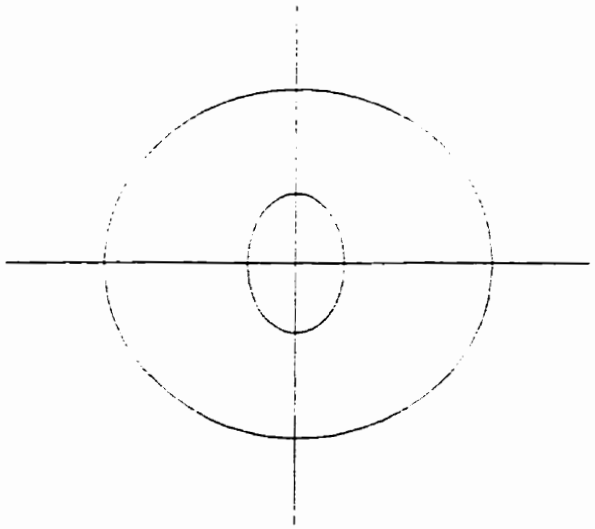


Figure 1. cross-section of pipe

ulations similar to those used to generate (4.15) and (4.16) in [35], and also let $b = 1$ without losing generality. The transpose of the resulting matrix is

$$a_{11} = \log K, \qquad a_{1j} = \frac{(\lambda/K^4)^{j-1}}{j-1}, \qquad a_{i1} = (-1)^{i-1}\lambda^{i-1}\binom{2(i-1)}{i-1}.$$

$$a_{ij} = \begin{cases} \left(\frac{\lambda}{K^4}\right)^{j-i}\binom{j+i-3}{j-i} & i < j. \\ 1 - \left(\frac{\lambda}{K^2}\right)^{2(i-1)} - K^{-4(i-1)} & i = j. \\ (-1)^{i-j}\lambda^{i-j}K^{-4(j-1)}(1+\lambda^{2(j-1)})\binom{2(i-j)}{i-j} & i > j. \end{cases} \tag{4.4}$$

$i, j = 1, 2, \ldots, N-1$, and the parameters

$$\lambda = \frac{\beta - \alpha}{\beta + \alpha}, \quad c = \frac{\alpha + \beta}{2}, \quad K = \frac{a}{c}. \tag{4.5}$$

We will next show that **A** satisfies the sign distribution as well as the conditions required in Theorem 4.1 for suitably chosen $K \geq 2$ and $\lambda < \frac{1}{2}$. We set $\alpha = 0.15$, $\beta = 0.25$ and $a = 1$ as an example. Correspondingly, we get $K = 5$, $\lambda = \frac{1}{4}$ and the truncation size is chosen to be $N \leq 500$. We only verify conditions (A1) and (A3). Conditions (A2) and (A4) can be verified in a similar manner.

It is evident that **A** satisfies SD1.

To verify (A1) of (4.1), we need to show, for $i \leq j$, that

$$\nu_{ij} = a_{ij} - \sum_{k=j+1}^{N} a_{ik} > 0. \qquad i \leq j,\ i, j = 1, 2, \ldots, N.$$

We first show that $\nu_{ii} > 0$. For convenience, we define

$$\rho = \frac{\lambda}{K^4}, \qquad q_l = \begin{pmatrix} l + 2i - 3 \\ l \end{pmatrix}.$$

For $i = 1$,

$$\begin{aligned} \nu_{11} &= a_{11} - \sum_{k=2}^{N} a_{1k} = \log K - \sum_{k=2}^{N} \frac{\rho^{k-1}}{k-1} \\ &> \log K - \rho \sum_{l=1}^{\infty} \frac{\rho^{l-1}}{l} = \log K - \frac{\rho}{(1-\rho)^2} > 0 \end{aligned}$$

For $i > 1$.

$$\begin{aligned}
\nu_{ii} &= a_{ii} - \sum_{k=i+1}^{N} a_{ik} \\
&= 1 - \rho^{i-1}\lambda^{i-1} - \rho^{i-1}/\lambda^{i-1} - \sum_{k=i+1}^{N} \rho^{k-i} \begin{pmatrix} k+i-3 \\ k-i \end{pmatrix} \\
&= 1 - \rho^{i-1}\left(\lambda^{i-1} + \frac{1}{\lambda^{i-1}}\right) - \sum_{l=1}^{N-i} \rho^l q_l
\end{aligned}$$

Notice that

$$\rho^{l+1} q_{l+1} = \left[\frac{l+2i-2}{l+1}\rho\right] \rho^l q_l < [N\rho]\rho^l q_l < \frac{1}{3}\rho^l q_l. \tag{4.6}$$

which implies that

$$\sum_{l=1}^{N-i} \rho^l q_l < \sum_{l=1}^{\infty} \frac{1}{3}\rho^l q_l < \sum_{l=1}^{\infty} \frac{1}{3^l} = \frac{1}{2}.$$

Also. $\rho^{i-1}\left(\lambda^{i-1} + \frac{1}{\lambda^{i-1}}\right) < \frac{2}{K^{4(i-1)}} < \frac{1}{4}$.

Hence.

$$\nu_{ii} > \frac{1}{2} - \rho^{i-1}\left(\lambda^{i-1} + \frac{1}{\lambda^{i-1}}\right) > 0.$$

We next show that $\nu_{ij} > 0$. for $i < j$.

$$\begin{aligned}
\nu_{ij} &= a_{ij} - \sum_{k=j+1}^{N} a_{ik} \\
&= \rho^{j-i} \begin{pmatrix} j+i-3 \\ j-i \end{pmatrix} - \sum_{k=i+1}^{N} \rho^{k-i} \begin{pmatrix} k+i-3 \\ k-i \end{pmatrix}
\end{aligned}$$

$$= \rho^{j-i}\left[\binom{j+i-3}{j-i} - \sum_{l=1}^{N-j}\rho^{l}\binom{l+j+i-3}{l+j-i}\right].$$

Using the argument similar to (4.6), we can show that

$$\sum_{l=1}^{N-j}\rho^{l}\binom{l+j+i-3}{l+j-i} < \sum_{l=1}^{N-j}\frac{1}{3^{l}}\binom{j+i-3}{j-i} < \frac{1}{2}\binom{j+i-3}{j-i}.$$

Therefore,

$$\nu_{ij} > \rho^{j-i}\left[\binom{j+i-3}{j-i} - \frac{1}{2}\binom{j+i-3}{j-i}\right] > 0.$$

To summarize, we have $\nu_{ij} > 0$, for $i \le j$.

To verify (A3) of (4.1), we need to show, for $j < i < l$, that

$$\omega_{ij}^{l} = \min\left\{|a_{ij}| - \left|\sum_{k=i+1}^{l} a_{kj}\right|, \quad \left|\sum_{k=i+1}^{l} a_{kj}\right| - |a_{l+1\,j}|\right\} > 0.$$

We first show that $|a_{ij}| - \left|\sum_{k=i+1}^{l} a_{kj}\right| > 0$. For convenience, we let $p_{ij} = (1+\lambda^{2(j-1)})\lambda^{i-j}K^{-4(j-1)}$ and $r_s = \binom{2(s+i-j)}{s+i-j}$.

$$\begin{aligned}
&|a_{ij}| - \left|\sum_{k=i+1}^{l} a_{kj}\right| \\
&= p_{ij}\binom{2(i-j)}{i-j} - \left|\sum_{k=i+1}^{l}(-1)^{k-j}p_{kj}\binom{2(k-j)}{k-j}\right| \\
&= p_{ij}\left[\binom{2(i-j)}{i-j} - \left|\sum_{k=i+1}^{l}(-\lambda)^{k-i}\binom{2(k-j)}{k-j}\right|\right]
\end{aligned}$$

$$= p_{ij}\left[\binom{2(i-j)}{i-j} - \left|\sum_{s=1}^{l-i}(-\lambda)^s \binom{2(s+i-j)}{s+i-j}\right|\right]$$

$$= p_{ij}\left[r_0 - \left|\sum_{s=1}^{l-i}(-\lambda)^s r_s\right|\right]$$

Note that

$$\lambda^{s+1} r_{s+1} = \left[\lambda\left(2 - \frac{1}{s+1+i-j}\right)\right]\lambda^s r_s.$$

and letting $f_s = \lambda\left(2 - \frac{1}{s+1+i-j}\right)$, we have

$$\left|\sum_{s=1}^{l-i}(-\lambda)^s r_s\right| = -\sum_{s=1}^{l-i}(-\lambda)^s r_s$$

$$= \lambda \cdot r_1 - \lambda^2 r_2 + \lambda^3 r_3 + \cdots - (-1)^{l-i}\lambda^{l-i} r_{l-i}$$

$$= \lambda r_1 \sum_{k=0}^{l-i-1}(-1)^k \prod_{s=1}^{k} f_s < \lambda r_1. \quad (f_{s+1} < f_s < \frac{1}{2}).$$

Hence, we have

$$|a_{ij}| - \left|\sum_{k=i+1}^{l} a_{kj}\right| > p_{ij}(r_0 - r_1\lambda) > p_{ij}(r_0 - f_1 r_0) = p_{ij} r_0 \frac{1}{2} > 0 \tag{4.7}$$

We can similarly show that $\left|\sum_{k=i+1}^{l} a_{kj}\right| - |a_{l+1\,j}| > 0$ and complete the verification of (4.3). Conditions (4.2) and (4.4) can be verified in the same way with slightly more complicated manipulations.

By the theory developed in this chapter, we conclude that the coefficient matrix defined in (4.4) is non-singular and a unique solution for the system $\mathbf{A}X = B$ is ensured.

# Chapter 5

# Upper and Lower Bounds for the Inverse Elements of Finite and Infinite Tridiagonal Matrices

## 5.1 Introduction

Tridiagonal matrices, finite or infinite occur in a large number of applications including the solution of boundary value problems by finite difference methods, cubic splines, data fitting, and three term difference equations and inverses of Toeplitz matrices and in the theory of continued fractions. Infinite systems occur in many areas including the solution of Mathieu's equations[39], three term recurrence relations for Bessel functions. For an algorithm to find the solution of a finite linear system or for Givens or Householder methods, see [16]. Estimates for upper bounds for the inverse elements of tridiagonal matrices arising in some boundary value problems are given by Mattheij[18]. Upper bounds for a special tridiagonal matrix is given by Kershaw[17], while a lower bound for the smallest singular value of a matrix is given by Varah[20]. Considerable work has been done in numerical treatment of tridiagonal matrices. Ostrowski[19] has given

upper bounds for the inverse elements of a diagonally dominant matrix.

In the following sections, we will give easily computable upper and lower bounds for the inverse elements and infinity norms for the inverse. The results improve Ostrowski's upper bounds as well as give new lower bounds. The results are extended to the infinite case and to block tridiagonal infinite systems. In later sections, we will apply the theory to a special matrix considered by Kershaw. We will also discuss the evaluation of Bessel functions and Mathieu functions by using their recurrence relations and numerical results are given.

## 5.2 Finite tridiagonal matrices

We will be concerned with finite and infinite tridiagonal matrices of the form

$$\mathbf{A} = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & a_n & b_n & c_n \\ & & & & \cdot & \cdot \end{pmatrix}.$$

denoted by $A = \{a_i, b_i, c_i\}$, where $b$'s are the diagonal elements and $a$'s and $c$'s are the off-diagonal elements. We will use the notation $A^{(r,s)}$, $s \geq r$ to represent the tridiagonal square submatrix of order $s - r + 1$ whose diagonal entry in the first row is $b_r$ and the diagonal entry in the last row is $b_s$.

We will now prove the following Lemma:

**Lemma 5.1** *For the tridiagonal $n \times n$ matrix $A = \{a_i, b_i, c_i\}$, the cofactors $A_{ij}$*

*of $A$ are given by*

$$(5.1) \qquad A_{ij} = \begin{cases} (-1)^{i+j} \left(\prod_{k=i+1}^{j} a_k\right) \det A^{(1,i-1)} \det A^{(j+1,n)}. & i \le j. \\ (-1)^{i+j} \left(\prod_{k=j}^{i-1} c_k\right) \det A^{(1,j-1)} \det A^{(i+1,n)}. & i > j. \end{cases}$$

*for $i, j = 2, 3, \cdots, n$. In the above,* $\det A^{(1,0)}$, $\det A^{(n+1,n)}$, *and* $\left(\prod_{k=i-1}^{i} p_k\right)$ *are each defined to be unity.*

*Proof:* We need only consider $1 \le i \le j \le n$ as the results for $1 \le j \le i \le n$ can be derived similarly. For $i = j = 1$ or $i = j = n$, the lemma is trivially true, while for $1 < i = j < n$, we can rewrite $A_{ii}$ in the form

$$A_{ii} = \det \begin{pmatrix} A^{(1,i-1)} & 0 \\ 0 & A^{(i+1,n)} \end{pmatrix}$$

from which the result follows. For $i = 1, j = n$,

$$\begin{aligned} A_{1n} &= (-1)^{n+1} \det \begin{pmatrix} a_2 & b_2 & c_2 & & \\ 0 & a_3 & b_3 & c_3 & \\ & \cdot & \cdot & \cdot & \\ & \cdot & \cdot & \cdot & \\ & & & & a_n \end{pmatrix} \\ &= (-1)^{n+1} \prod_{k=2}^{n} a_k. \end{aligned}$$

Similarly for $1 < i < n, j = n$, the theorem is true.

For $1 < i < j < n$.

$$A_{ij} = (-1)^{i+j} \det \left(\begin{array}{c|ccccc|c} A^{(1,i-1)} & c_{i-1} & & & & & \\ \hline 0 & a_{i-1} & b_{i+1} & c_{i+1} & & & \\ & 0 & a_{i-2} & b_{i+2} & c_{i+2} & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 0 & a_j & c_j & \\ \hline & & & & 0 & A^{(j-1,n)} & \end{array}\right)$$

$$= (-1)^{i+j} \det A^{(1,i-1)} \det A^{(j+1,n)} \prod_{k=i+1}^{j} a_k. \qquad \square$$

**Lemma 5.2** *[Ostrowski[19]] Let $B = (b_{ij})_{n\times n}$ be a strictly row diagonally dominant matrix and*

$$\mu_i = \frac{1}{|b_{ii}|} \sum_{j=1, j\neq i}^{n} |b_{ij}|. \quad 0 \leq \mu_i < 1. \quad i = 1, 2, \cdots, n.$$

*Then for $B^{-1} = \left(\frac{B_{ji}}{\det B}\right)$. the following hold:*

$$\frac{1}{|b_{jj}|(1+\mu_j)} \leq \left|\frac{B_{jj}}{\det B}\right| \leq \frac{1}{|b_{jj}|(1-\mu_j)} \tag{5.2}$$

*and*

$$|B_{ij}| \leq \mu_j |B_{ii}|. \tag{5.3}$$

Now we will prove the following theorem for tridiagonal matrices.

**Theorem 5.1** *Let $A$ be an $n \times n$ tridiagonal matrix, $a_i, b_i, c_i \neq 0$. and let $A$ be diagonally dominant in the sense $\mu_i(|b_i|) = |a_i| + |c_i|$, $i = 1, 2, \cdots, n$. $0 \leq \mu_i \leq 1$.*

*Then we have the following upper and lower bounds:*

$$\left(\prod_{k=i+1}^{j}\frac{|a_k|}{|b_k|(1+\mu_k)}\right)|A_{ii}| \le |A_{ij}| \le \left(\prod_{k=i+1}^{j}\mu_k\right)|A_{ii}|. \qquad i<j \tag{5.4}$$

$$\left|\frac{a_j}{a_i}\right|\left(\prod_{k=i}^{j-1}\frac{|a_k|}{|b_k|(1+\mu_k)}\right)|A_{jj}| \le |A_{ij}| \le \left|\frac{a_j}{a_i}\right|\left(\prod_{k=i}^{j-1}\mu_k\right)|A_{jj}|. \qquad i<j \tag{5.5}$$

$$\left(\prod_{k=j}^{i-1}\frac{|c_k|}{|b_k|(1+\mu_k)}\right)|A_{ii}| \le |A_{ij}| \le \left(\prod_{k=j}^{i-1}\mu_k\right)|A_{ii}|. \qquad i\ge j \tag{5.6}$$

$$\left|\frac{c_j}{c_i}\right|\left(\prod_{k=j+1}^{i}\frac{|c_k|}{|b_k|(1+\mu_k)}\right)|A_{jj}| \le |A_{ij}| \le \left|\frac{c_j}{c_i}\right|\left(\prod_{k=j+1}^{i}\mu_k\right)|A_{jj}|. \qquad i\ge j \tag{5.7}$$

*Proof:* We only show (5.4). The other results can be derived similarly.

For $i<j$, we have from Lemma 5.1.

$$\begin{aligned}\frac{A_{ij}}{A_{ii}} &= (-1)^{i+j}\left(\prod_{k=i+1}^{j}a_k\right)\frac{\det A^{(j+1,n)}}{\det A^{(i+1,n)}}\\ &= (-1)^{i+j}\left(\prod_{k=i+1}^{j}a_k\right)\prod_{p=i+1}^{j}\frac{\det A^{(p+1,n)}}{\det A^{(p,n)}}\end{aligned}$$

which gives on using (5.2),

$$\left(\prod_{k=i+1}^{j}a_k\right)\prod_{p=i+1}^{j}\frac{1}{|b_p|+|c_p|} \le \left|\frac{A_{ij}}{A_{ii}}\right| \le \left(\prod_{k=i+1}^{j}a_k\right)\prod_{p=i+1}^{j}\frac{1}{|b_p|-|c_p|}. \tag{5.8}$$

Now using $\mu_p|b_p| = |a_p| + |c_p|$ and

$$\frac{1}{|b_p|-|c_p|} = \frac{\mu_p}{|a_p|+|c_p|(1-\mu_p)} \le \frac{\mu_p}{|a_p|},$$

and

$$\frac{1}{|b_p| + |c_p|} > \frac{1}{|b_p|(1 + \mu_p)}.$$

(5.8) reduces to (5.4). □

Comparing the above results with Ostrowski's upper bound, we note that (5.4) and (5.6) lead to (5.3) for the tridiagonal case.

**Theorem 5.2** *For the matrix $A$ defined in Theorem 5.1, the following inequality holds for $i = 1, 2, \cdots, n$*

$$\frac{1}{|b_i| + |a_i|\mu_{i-1} + |c_i|\mu_{i+1}} \leq \left|\frac{A_{ii}}{\det A}\right| \leq \frac{1}{|b_i| - |a_i|\mu_{i-1} - |c_i|\mu_{i+1}}. \tag{5.9}$$

*where $\mu_0 = \mu_{n+1} = 0$.*

*Proof:* Expanding $\det A$ by the $i$th row, we have

$$a_i \frac{A_{i\,i-1}}{\det A} + b_i \frac{A_{ii}}{\det A} + c_i \frac{A_{i\,i+1}}{\det A} = 1, \quad i = 1, 2, \cdots, n.$$

where $A_{i0} = A_{i\,n+1} = 0$.

By taking absolute values and using (5.3), the above reduces to

$$\left|1 - b_i \frac{A_{ii}}{\det A}\right| \leq (|a_i|\mu_{i-1} + |c_i|\mu_{i+1}) \left|\frac{A_{ii}}{\det A}\right|$$

from which (5.9) follows. □

Combining Theorem 5.1 and Theorem 5.2, we immediately have

**Theorem 5.3** *Let* $A^{-1} = \left(\frac{A_{ji}}{\det A}\right)$ *be the inverse of matrix* $A$ *defined in Theorem 5.1. then*

$$(5.10) \quad \frac{\prod_{k=i+1}^{j} |a_k|}{\prod_{k=i}^{j} |b_k|(1+\mu_k)} \leq \left|\frac{A_{ij}}{\det A}\right| \leq \frac{\prod_{k=i+1}^{j} \mu_k}{|b_i| - |a_i|\mu_{i-1} - |c_i|\mu_{i+1}}. \qquad i < j$$

$$(5.11) \quad \frac{\prod_{k=j}^{i-1} |c_k|}{\prod_{k=j}^{i} |b_k|(1+\mu_k)} \leq \left|\frac{A_{ij}}{\det A}\right| \leq \frac{\prod_{k=j}^{i-1} \mu_k}{|b_i| - |a_i|\mu_{i-1} - |c_i|\mu_{i+1}}. \qquad i \geq j.$$

Based on Theorem 5.3. we will now establish some results for $\| A^{-1} \|_\infty$.

**Theorem 5.4** *Let* $A$ *be the matrix defined in Theorem 5.1. and define* $\mu = \sup_k\{\mu_k\}$. $\delta = \inf_k\{|b_k|\}$ *and* $\tau = \sup_k\{|b_k|\}$. *then*

$$\max\{\alpha, \tilde{\alpha}\} \leq \| A^{-1} \|_\infty \leq \beta.$$

*where* $\alpha^{-1} = \sup_k\{|b_k| + |a_k| + |c_k|\}$. $\tilde{\alpha}^{-1} = \frac{1}{2}\tau(1+\mu)$ *and* $\beta^{-1} = \delta(1-\mu)$.

*Proof:* It suffices to show that $(\alpha, \tilde{\alpha}) \leq \| A^{-1} \|_\infty \leq \beta$.

By definition of $\| \cdot \|_\infty$.

$$\| A^{-1} \|_\infty = \sup_i \left\{ \sum_{j=1}^{n} \left| \frac{A_{ji}}{\det A} \right| \right\}.$$

To show $\| A^{-1} \|_\infty \geq \alpha$, we have, from $\text{cond}(A) = \| A \|_\infty \cdot \| A^{-1} \|_\infty \geq 1$,

$$\| A^{-1} \|_\infty \geq \frac{1}{\| A \|_\infty} = \alpha.$$

To show $\| A^{-1} \|_\infty \geq \tilde{\alpha}$, let $\omega = \inf_k\{|a_k|, |c_k|\}$ and $\gamma = \frac{\omega}{\tau(1+\mu)}$. We now consider, for $n > 0$,

$$\sum_{j=1}^{n} \left| \frac{A_{ji}}{\det A} \right| = \sum_{j=1}^{i-1} \left| \frac{A_{ji}}{\det A} \right| + \sum_{j=i}^{n} \left| \frac{A_{ji}}{\det A} \right|. \tag{5.12}$$

For $i = 1$, we have, on applying (5.11),

$$\begin{aligned} \sum_{j=1}^{n} \left| \frac{A_{j1}}{\det A} \right| &\geq \frac{1}{|b_1|(1+\mu_1)} + \sum_{j=2}^{n} \frac{\prod_{k=1}^{j-1} |c_k|}{\prod_{k=1}^{j} |b_k(1+\mu_k)|} \\ &\geq \frac{1}{\tau(1+\mu)} + \sum_{j=2}^{n} \frac{\omega^{j-2}}{\tau^{j-1}(1+\mu)^{j-1}} \\ &\geq \frac{1}{\tau(1+\mu)} \left( 1 + \frac{1-\gamma^{n-1}}{1-\gamma} \right) \geq \frac{2}{\tau(1+\mu)}. \end{aligned}$$

Using same arguments, we can show that for $i = n$,

$$\sum_{j=1}^{n} \left| \frac{A_{jn}}{\det A} \right| \geq \frac{2}{\tau(1+\mu)}.$$

For $1 < i < n$,

$$\sum_{j=1}^{n} \left| \frac{A_{ji}}{\det A} \right| \geq \sum_{j=1}^{i-1} \frac{\prod_{k=j+1}^{i} |a_k|}{\prod_{k=j}^{i} |b_k|(1+\mu_k)} + \sum_{j=i+1}^{n} \frac{\prod_{k=i}^{j-1} |c_k|}{\prod_{k=i}^{j} |b_k|(1+\mu_k)}$$

$$\geq \frac{1}{|b_i|(1+\mu_i)} + \sum_{j=1}^{i-1} \frac{\omega^{i-j-1}}{\tau^{i-j}(1+\mu)^{i-j}} + \sum_{j=i+1}^{n} \frac{\omega^{j-i-1}}{\tau^{j-i}(1+\mu)^{j-i}}$$

$$= \frac{1}{\tau(1+\mu)} \left[ 1 + \sum_{j=1}^{i-1} \gamma^{i-j-1} + \sum_{j=i+1}^{n} \gamma^{j-i-1} \right] \geq \frac{3}{\tau(1+\mu)}.$$

Combine the above results. we get

$$\| A^{-1} \|_\infty \geq \frac{2}{\tau(1+\mu)} = \tilde{\alpha}.$$

To prove $\| A^{-1} \|_\infty \leq \beta$. we have. by Theorem 1 in [20].

$$\| A^{-1} \|_\infty \leq \frac{1}{\inf_k \{|b_k| - |a_k| - |c_k|\}}.$$

and hence

$$\| A^{-1} \|_\infty \leq \frac{1}{\delta(1-\mu)} = \beta.$$

Hence.

$$\max\{\alpha, \tilde{\alpha}\} \leq \| A^{-1} \|_\infty \leq \beta.$$

□

## 5.3 Infinite tridiagonal matrices

We now consider infinite, tridiagonal and diagonally dominant matrices of the form

$$\mathbf{A} = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & a_n & b_n & c_n \\ & & & & \cdot & \cdot \end{pmatrix}$$

and the infinite linear systems of algebraic equations associated with such matrices. Under some certain conditions, the infinite matrix $\mathbf{A}$ can be regarded as a linear operator on the $\ell_p$ space, and the existence and uniqueness of the solution in the $\ell_p$ for the above system can be established. A useful numerical approach for approximating the solution by using the solution for the truncated system, with an explicit error bound, is suggested. The results are comparable to the results for the general case discussed in [34] and [40]. Moreover, the results can also be extended to the infinite tridiagonal block systems that satisfy similar conditions.

To formulate the problems in $\ell_\infty$, we first define $\mathbf{A}$ as an infinite, strictly row diagonally dominant tridiagonal matrix and

$$\mathbf{A} = \{a_i, b_i, c_i\}_{i=1,2,\cdots}, \quad D = (d_1, d_2, \cdots)^t \in \ell_\infty.$$

$$\mathbf{A}^{(n)} = \{a_i, b_i, c_i\}_{i=1,2,\cdots,n}, \quad D^{(n)} = (d_1, d_2, \cdots, d_n)^t. \tag{5.13}$$

where $a_1 = 0$ and $a_i, b_i, c_i \neq 0$, for $i = 1, 2, \cdots$.

A finite truncated system can be written out as $\mathbf{A}^{(n)} X^{(n)} = D^{(n)}$. For conve-

nience. we define an extended infinite truncated system $\tilde{\mathbf{A}}^{(n)}\tilde{X}^{(n)} = D$ with

$$(5.14)\quad \tilde{\mathbf{A}}^{(n)} = \begin{pmatrix} \mathbf{A}^{(n)} & 0 \\ 0 & \mathbf{B}^{(n+1)} \end{pmatrix}. \quad \tilde{X}^{(n)} = \left((X^{(n)})^t, \frac{d_{n+1}}{b_{n+1}}, \frac{d_{n+2}}{b_{n+1}}, \ldots\right)^t.$$

where $\mathbf{B}^{(n+1)}$ is the infinite diagonal matrix.

It is easy to see that the above two truncated systems are equivalent. We now rewrite the above systems in iterative form:

$$X = \mathbf{G}X + P. \quad \textit{and} \quad \tilde{X}^{(n)} = \tilde{\mathbf{G}}^{(n)}\tilde{X}^{(n)} + P.$$

The above leads to iteration formulas:

$$X^{(k+1)} = \mathbf{G}X^{(k)} + P. \quad \textit{and} \quad \tilde{X}^{(n,k+1)} = \tilde{\mathbf{G}}^{(n)}\tilde{X}^{(n,k)} + P. \quad X^{(0)} = \tilde{X}^{(n,0)} = P.$$

where

$$\mathbf{G} = \begin{pmatrix} 0 & -\frac{c_1}{b_1} & & & & \\ -\frac{a_2}{b_2} & 0 & -\frac{c_2}{b_2} & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & -\frac{a_n}{b_n} & 0 & -\frac{c_n}{b_n} \\ & & & & \cdot & \cdot \end{pmatrix} \quad \tilde{\mathbf{G}}^{(n)} = \begin{pmatrix} 0 & -\frac{c_1}{b_1} & & & & \\ -\frac{a_2}{b_2} & 0 & -\frac{c_2}{b_2} & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & -\frac{a_n}{b_n} & 0 & 0 \\ & & & & & O \end{pmatrix}$$

and

$$P = (\frac{d_1}{b_1}, \frac{d_2}{b_2}, \cdots)^t. \quad \tilde{X}^{(n)} = (x_1, x_2, \cdots, x_n, \frac{d_{n+1}}{b_{n+1}}, \cdots)^t$$

We now prove the following theorem.

**Theorem 5.5** *Let* $\mathbf{A}$, $D$ *and* $\tilde{\mathbf{A}}^{(n)}$ *be defined by (5.13) and (5.14). A satisfies: (H1)* $\mu_n \to 0$ *as* $n \to \infty$*: (H2)* $\delta \equiv \inf_k\{|b_n|\} > 0$. *Then there exists a unique* $X \in \ell_\infty$ *such that* $\mathbf{A}X = D$. *Moreover, if* $\tilde{X}^{(n)}$ *is a solution for the truncated system (3.2), then*

$$\| X - \tilde{X}^{(n)} \|_\infty \leq \frac{\sup_{l>n}\{\mu_l\} \| D \|_\infty}{\delta(1-\mu)^2} \to 0. \quad n \to \infty. \tag{5.15}$$

*where* $\mu \equiv sup_k\{\mu_k\}$.

*Proof:*

We first show that there exists a unique $X \in \ell_\infty$ such that $\mathbf{A}X = D$. For the linear transformation $Y = GX + P$, and $X_1, X_2 \in \ell_\infty$, let $Y_1 = \mathbf{G}X_1 + P$, $Y_2 = \mathbf{G}X_2 + P$, we have

$$\| Y_2 - Y_1 \|_\infty \leq \| \mathbf{G} \|_\infty \| X_2 - X_1 \|_\infty = \mu \| X_2 - X_1 \|_\infty. \quad (\mu < 1).$$

It turns out that $Y = \mathbf{G}X + P$ is a contraction mapping onto $\ell_\infty$. By the Schauder Fixed Point Theorem [21], there exists a unique $X \in \ell_\infty$ such that $X = \mathbf{G}X + P$.

We shall use $\| \cdot \|$ instead of $\| \cdot \|_\infty$.

We first notice, for any $k > 0$,

$$\| X - \tilde{X}^{(n)} \| \leq \| X - X^{(,k)} \| + \| X^{(,k)} - \tilde{X}^{(n,k)} \| + \| \tilde{X}^{(n)} - \tilde{X}^{(n,k)} \|, \tag{5.16}$$

and

$$\| X - X^{(,k)} \| \leq \| \mathbf{G} \|^k \| X - X^{(,0)} \| \leq \| \mathbf{G} \|^k \| X - P \| \leq \| \mathbf{G} \|^{k+1} \| X \|$$
$$\leq \| \mathbf{G} \|^{k+1} \| (\mathbf{I} - \mathbf{G})^{-1} \| \cdot \| P \| \leq \frac{\| \mathbf{G} \|^{k+1} \| P \|}{1 - \| \mathbf{G} \|}.$$

Recalling that $\| \mathbf{G} \| = \sup_n \{ \frac{|a_n|}{|b_n|} + \frac{|c_n|}{|b_n|} \} = \mu$ and $\| P \| \leq \delta^{-1} \| D \|$, we get

$$\| X - X^{(.k)} \| \leq \frac{\mu^{k+1} \| D \|}{\delta(1-\mu)}.$$

Similarly, we have

$$\| \tilde{X}^{(n)} - \tilde{X}^{(n.k)} \| \leq \frac{\mu^{k+1} \| D \|}{\delta(1-\mu)}.$$

We also have

$$\begin{aligned} \| X^{(.k)} - \tilde{X}^{(n.k)} \| &\leq \| \mathbf{G} X^{(.k)} - \tilde{\mathbf{G}}^{(n)} \tilde{X}^{(n.k)} \| \\ &\leq \| \mathbf{G} X^{(.k)} - \mathbf{G} \tilde{X}^{(n.k)} \| + \| \mathbf{G} - \tilde{\mathbf{G}}^{(n)} \| \cdot \| \tilde{X}^{(n.k)} \| \\ &\leq \| \mathbf{G} \| \cdot \| X^{(.k)} - \tilde{X}^{(n.k)} \| + \| \mathbf{G} - \tilde{\mathbf{G}}^{(n)} \| \cdot \| \tilde{X}^{(n.k)} \| . \end{aligned} \tag{5.17}$$

On using

$$\| \mathbf{G} - \tilde{\mathbf{G}}^{(n)} \| = \sup_{l>n} \{ \mu_l \}.$$

and

$$\begin{aligned} \| \tilde{X}^{(n.k)} \| &= \| \tilde{\mathbf{G}}^{(n)} \tilde{X}^{(n.k)} + P \| \leq \| P \| + \| \mathbf{G} \| \cdot \| \tilde{X}^{(n.k)} \| \\ &\leq \| P \| \left( \sum_{l=0}^{k} \| \mathbf{G} \|^l \right) \leq \frac{\| P \|}{1 - \| \mathbf{G} \|} \leq \frac{\| D \|}{\delta(1-\mu)}. \end{aligned}$$

(5.17) becomes

$$\| X^{(.k)} - \tilde{X}^{(n.k)} \| \leq \| \mathbf{G} \| \cdot \| X^{(.k)} - \tilde{X}^{(n.k)} \| + \frac{\sup_{l>n} \{ \mu_l \} \| D \|}{\delta(1-\mu)}$$

$$\leq \| \mathbf{G} \|^k \| X^{(.k)} - P \| + \frac{\sup_{l>n}\{\mu_l\} \| D \|}{\delta(1-\mu)} \sum_{l=0}^{k} \| \mathbf{G} \|^l$$

$$\leq \frac{\mu^{k+1} \| D \|}{\delta(1-\mu)} + \frac{\sup_{l>n}\{\mu_l\} \| D \|}{\delta(1-\mu)^2}.$$

So (5.16) reduces to

$$\| X - \tilde{X}^{(n)} \| \leq \frac{3\mu^{k+1} \| D \|}{\delta(1-\mu)} + \frac{\sup_{l>n}\{\mu_l\} \| D \|}{\delta(1-\mu)^2}.$$

By letting $k \to \infty$. we get

$$\| X - \tilde{X}^{(n)} \| \leq \frac{\sup_{l>n}\{\mu_l\} \| D \|}{\delta(1-\mu)^2} \to 0. \quad n \to \infty. \qquad \square$$

This proof gives an efficient way for estimating the solution for the infinite system (5.13) by using a solution for the truncated system (5.14). One can even use an iteration formula to execute this truncation. with considerable precision given by $O(\mu^{k+1})$.

**Corollary 5.1** *Let* $\mathbf{A}$ *and* $\mathbf{A}^{(n)}$ *be the matrices defined by (5.13). then for any fixed* $i, j$

$$\lim_{n\to\infty} \frac{\mathbf{A}_{ij}^{(n)}}{\det \mathbf{A}^{(n)}} = \frac{A_{ij}}{\det \mathbf{A}}.$$

*where the right hand side is defined as the inverse element of* $\mathbf{A}$.

To show that for any fixed $i, j \leq n$,

$$\lim_{n\to\infty} \frac{A_{ij}^{(n)}}{\det \mathbf{A}^{(n)}} = \frac{A_{ij}}{\det \mathbf{A}}.$$

we need only choose $D$ in (5.13), the given vector, to be the unit vector with the $i$th component $d_i = 1$, and $d_j = 0$, for $j \neq i$, $i = 1, 2, \cdots$. Then the result follows by Theorem 5.5.

## 5.4 Infinite block tridiagonal matrices

We now turn our attention to an infinite block tridiagonal system of the form

$$\mathbf{A}_b X = D. \quad \mathbf{A}_b = \begin{pmatrix} B_1 & C_1 & & & & \\ A_2 & B_2 & C_2 & & & \\ & \ddots & & & & \\ & & & A_n & B_n & C_n \\ & & & & & \ddots \end{pmatrix}. \tag{5.18}$$

where each element of $\mathbf{A}_b$ is a non-zero $m \times m$ matrix. $D = ((D_1)^t, (D_2)^t, \cdots, (D_n)^t, \cdots)^t$ $X = ((X_1)^t, (X_2)^t, \cdots, (X_n)^t, \cdots)^t$. $D_k$ and $X_k$ are $m \times 1$ vectors.

This part of the work is motivated by solving a block tridiagonal system arising in the problem of evaluating non-hierarchical networks which can be modelled as a strictly diagonally dominant infinite system with block tridiagonal structure as given in the above.

We define $\mu_n$ as the quantity for matrix $A$ in the usual sense and $\sigma_k$, $k =$

$1, 2, \cdots, m$ by

$$\sigma_k \| B_k^{-1} \|_\infty^{-1} = \| A_k \|_\infty + \| C_k \|_\infty .$$

We also define an extended infinite truncated block system:

$$\tilde{\mathbf{A}}_b^{(n)} \tilde{X}^{(n)} = D. \quad \tilde{\mathbf{A}}_b^{(n)} = \begin{pmatrix} B_1 & C_1 & & & & & \\ A_2 & B_2 & C_2 & & & & \\ & & & A_n & B_n & O & \\ & & & & O & B_{n+1} & O \\ & & & & & O & B_{n+2} & O \\ & & & & & & & \end{pmatrix}.$$

$$\tilde{X}^{(n)} = \left( (X_1)^t, (X_2)^t, \cdots, (X_n)^t, ((B_{n+1})^{-1} D_{n+1})^t, \cdots \right)^t . \tag{5.19}$$

where $B$'s are diagonal matrices and $O$ is the zero matrix

We first assume that $\mathbf{A}_b$ is a strictly row diagonally dominant matrix Therefore each of $B_k$'s is strictly row diagonally dominant and hence nonsingular We also assume that $\mathbf{A}_b$ satisfies the following conditions:

(B1) $\| B_k^{-1} \|_\infty^{-1} \geq \delta > 0$, for all k's

(B2) $\sigma_n \leq \sigma < 1$ and $\sigma_n \to 0$ as $n \to \infty$

Then we have the following results

**Theorem 5.6** *Let* $\mathbf{A}_b$ *and* $\tilde{\mathbf{A}}_b^{(n)}$ *be defined by (5.18) and (5.19), and* $D \in \ell_\infty$. *Then the system* $\mathbf{A}_b X = D$ *has a unique bounded solution. Further, if* $\tilde{X}_{(n)}$ *is a bounded solution of* $\tilde{\mathbf{A}}_b^{(n)} \tilde{X}^{(n)} = D$, *then*

$$\| X - \tilde{X}^{(n)} \|_\infty \leq \frac{\sup_{l>n} \{\sigma_l\} \| D \|_\infty}{\delta (1-\sigma)^2} \to 0, \qquad n \to \infty.$$

Remark: For the special case that all $B_k$'s happen to be tridiagonal matrices, $\| B_k^{-1} \|_\infty$ may be estimated by using the results in the previous sections.

A sketch of the proof for Theorem 5.6 parallels the proof of Theorem 3.1, with $\mu_k$ and $\mu$ replaced by $\sigma_k$ and $\sigma$ respectively, the notations standing for matrix elements replaced by block matrices and the absolute operator replaced by infinite norm.

## 5.5 Applications

We now give three examples for illustrating our results on finite and infinite matrices.

Example 1: (Kershaw[17]) Consider matrix $\mathbf{A}$ of the form

$$\mathbf{A} = \begin{pmatrix} \lambda_1 & 1-\alpha_1 & 0 & \cdots & 0 & 0 \\ \alpha_2 & \lambda_2 & \alpha_2 & \cdots & 0 & 0 \\ 0 & \alpha_3 & \lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{n-1} & 1-\alpha_{n-1} \\ 0 & 0 & 0 & \cdots & \alpha_n & \lambda_n \end{pmatrix}$$

where $0 < \alpha_i < 1$ and $\lambda_i > 1$, for $i = 1, 2, ..., n$. Upper bounds $K_i$ for the inverse elements of $\mathbf{A}$ are given by Kershaw[17], which can be written as follows with our notation:

$$0 < \left| \frac{A_{ij}}{\det \mathbf{A}} \right| < \frac{\nu_i \prod_{l=l_1}^{l_2} \lambda_l^{-1}}{\nu_i - 1} \equiv K_i, \quad i \neq j, \tag{5.20}$$

where $l_1 = \min\{i, j\}$, $l_2 = \max\{i, j\}$ and $\nu_s = \min\{\lambda_{s-1}\lambda_s, \lambda_s\lambda_{s+1}\}$ with $\nu_1 = \lambda_1\lambda_2$ and $\nu_n = \lambda_{n-1}\lambda_n$.

Using Theorems 5.2 and 5.3, we can easily get other upper bounds denoted by $S_i$ for $\frac{A_{ij}}{\det \mathbf{A}}$, which can be shown to be better than Kershaw's bounds. For $i \leq j$, noting that $|b_l| = \lambda_l$, $\mu_l = \frac{1}{\lambda_l}$, $a_l = \alpha_l$ and $c_l = 1 - \alpha_l$, we have

$$\begin{aligned}\left|\frac{A_{ij}}{\det \mathbf{A}}\right| \leq S_i &= \left[\prod_{l=i+1}^{j} \mu_l\right] \frac{1}{\lambda_i - \alpha_i \mu_{i-1} - (1-\alpha_i)\mu_{i+1}} \\ &= \left[\prod_{l=i}^{j} \lambda_l^{-1}\right] \frac{1}{1 - \frac{\alpha_i}{\lambda_{i-1}\lambda_i} - \frac{1-\alpha_i}{\lambda_i \lambda_{i+1}}} \\ &\leq \left[\prod_{l=i}^{j} \lambda_l^{-1}\right] \frac{1}{1 - \frac{\alpha_i}{\nu_i} - \frac{1-\alpha_i}{\nu_i}} = \left[\prod_{l=i}^{j} \lambda_l^{-1}\right] \frac{\nu_i}{\nu_i - 1} = K_i.\end{aligned}$$

giving $S_i \leq K_i$.

It can similarly be shown that result holds for $i \geq j$.

From the above we can see Kershaw's upper bound is improved.

Regarding the lower bound for Kershaw's matrix, we start with (5.6) and (5.8), for $i \leq j$ and we obtain

$$\begin{aligned}\left|\frac{A_{ij}}{\det \mathbf{A}}\right| &\geq \prod_{k=i+1}^{j} \frac{|a_k|}{|b_k|(1+\mu_k)} \cdot \frac{1}{|b_i| + |a_i|\mu_{i-1} + |c_i|\mu_{i+1}} \\ &\geq \prod_{k=i+1}^{j} \frac{\alpha_k}{\lambda_k(1+\lambda_k^{-1})} \cdot \frac{1}{\lambda_i(1 + \frac{\alpha_i}{\lambda_{i-1}\lambda_i} + \frac{1-\alpha_i}{\lambda_i \lambda_{i+1}})}\end{aligned}$$

$$\geq \lambda_i^{-1} \prod_{k=i+1}^{j} \frac{\alpha_k}{1+\lambda_k} \cdot \frac{\nu_i}{1+\nu_i}.$$

and similarly, for $i \geq j$

$$\left| \frac{A_{ij}}{\det \mathbf{A}} \right| \geq \lambda_i^{-1} \prod_{k=j}^{i-1} \frac{1-\alpha_k}{1+\lambda_k} \cdot \frac{\nu_i}{1+\nu_i}.$$

Hence we get the lower bounds

$$\left| \frac{A_{ij}}{\det \mathbf{A}} \right| \geq \begin{cases} \lambda_i^{-1} \prod_{k=i+1}^{j} \frac{\alpha_k}{1+\lambda_k} \cdot \frac{\nu_i}{1+\nu_i}. & i \leq j. \\ \lambda_i^{-1} \prod_{k=j}^{i-1} \frac{1-\alpha_k}{1+\lambda_k} \cdot \frac{\nu_i}{1+\nu_i}. & i \geq j. \end{cases}$$

Example 2: The Bessel functions, $J_n(x), n = 0, 1, 2, \cdots$, satisfy the following well-known recurrence relation

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x). \qquad n = 0, 1, 2, \cdots.$$

To find values of $J_n(x)$ at the chosen point $x = L, 0 < |L| < 2$, one can reduce this problem to the solution for an infinite system upon introducing $x_n = J_n(x)$, so that

$$x_{n+1} = \frac{2n}{L} x_n - x_{n-1}, \qquad n = 1, 2, \cdots. \tag{5.21}$$

where $x_0 = J_0(L)$ is assumed to be given.

If $|L| \geq 2$, then the matrix is not strictly diagonally dominant. We can still manipulate the system by eliminating the first $[L]$ rows of the matrix so that the

matrix can be transformed to a strictly diagonally dominant matrix. For instance, $L = 4$, the system

$$\begin{bmatrix} \frac{1}{2} & -1 & & & & \\ -1 & 1 & -1 & & & \\ & -1 & \frac{3}{2} & -1 & & \\ & & -1 & 2 & -1 & \\ & & & -1 & \frac{5}{2} & -1 \\ & & & & \ddots & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix},$$

which after truncating, can be reduced to

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & \frac{5}{2} & -1 & & \\ & -1 & 3 & -1 & \\ & & -1 & \frac{7}{2} & \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} x_4 \\ x_5 \\ x_6 \\ x_7 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_3 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix},$$

which is strictly diagonally dominant, and $x_1$, $x_2$ ,$x_3$ can be easily obtained once the infinite system is estimated in terms of $x_3$. For convenience, we only discuss the case when $|L| < 2$ for convenience.

In matrix notation, the above system can be written as $\mathbf{A}X = D$, where $\mathbf{A}$ is the resulting tridiagonal infinite matrix with $a_n = c_n = -1$ as its off diagonal entries and $b_n = \frac{2n}{L}$ as diagonal entries; $X = (x_1, x_2, \cdots)^t$ and $D = (x_0, 0, 0, \cdots)^t$.

It is easy to check that all the conditions required by Theorem 5.5 are satisfied by this example, even though it does not satisfy the conditions required to apply results in [34].

As an example, we choose $L = 1$ and apply the results from Theorem 5.5. This

enables us to use the solution of a truncated system $\mathbf{A}^{(n)}X^{(n)} = D^{(n)}$ to estimate $X$, the solution of the infinite system. We can estimate as many values of $J_n(1)$ as we wish. For illustrating the truncation method, we let $n = 4, 8, 12$ respectively. The numerical results are listed in the following tables.

Table 5.1: Solutions for truncated system

| k | Truncated solutions | | | Absolute sol. $J_k(1.0)$ |
|---|---|---|---|---|
| | n=4 | n=8 | n=12 | |
| 1 | 0.4400497878E+00 | 0.4400505858E+00 | 0.4400505858E+00 | 0.4400505858E+00 |
| 2 | 0.1149018890E+00 | 0.1149034849E+00 | 0.1149034849E+00 | 0.1149034849E+00 |
| 3 | 0.1955776835E-01 | 0.1956335398E-01 | 0.1956335398E-01 | 0.1956335398E-01 |
| 4 | 0.2444721043E-02 | 0.2476638964E-02 | 0.2476638964E-02 | 0.2476638964E-02 |
| 5 | - | 0.2497577300E-03 | 0.2497577302E-03 | 0.2497577302E-03 |
| 6 | - | 0.2093833601E-04 | 0.2093833800E-04 | 0.2093833800E-04 |
| 7 | - | 0.1502302135E-05 | 0.1502325818E-05 | 0.1502325818E-05 |
| 8 | - | 0.9389388347E-07 | 0.9422344173E-07 | 0.9422344173E-07 |
| 9 | - | - | 0.5249250180E-08 | 0.5249250180E-08 |
| 10 | - | - | 0.2630615105E-09 | 0.2630615124E-09 |
| 11 | - | - | 0.1198003084E-10 | 0.1198006746E-10 |
| 12 | - | - | 0.4991679517E-12 | 0.4999718180E-12 |

Table 5.2: Actual errors

| k | Actual errors: $\lvert x_k - J_k(1.0)\rvert$ | | |
|---|---|---|---|
| | n=4 | n=8 | n=12 |
| 1 | 0.7979480199E-06 | 0.6441113105E-15 | 0.1256154054E-25 |
| 2 | 0.1595896040E-05 | 0.1288222621E-14 | 0.2512308107E-25 |
| 3 | 0.5585636139E-05 | 0.4508779174E-14 | 0.8793078378E-25 |
| 4 | 0.3191792080E-04 | 0.2576445242E-13 | 0.5024616216E-24 |
| 5 | - | 0.2016068402E-12 | 0.3931762189E-23 |
| 6 | - | 0.1990303949E-11 | 0.3881516027E-22 |
| 7 | - | 0.2368204055E-10 | 0.4618501610E-21 |
| 8 | - | 0.3295582638E-09 | 0.6427087094E-20 |
| 9 | - | - | 0.1023715433E-18 |
| 10 | - | - | 0.1836260693E-17 |
| 11 | - | - | 0.3662284232E-16 |
| 12 | - | - | 0.8038662703E-15 |

Example 3: Mathieu functions are encountered in physical problems involving elliptical boundaries. The wave equation in elliptical coordinates, when using the method of separation of variables, can be reduced to the Mathieu equation given by

$$y'' + (\lambda - 2q\cos 2x)y = 0. \tag{5.22}$$

where $q$ is given and $\lambda$ is the eigenvalue parameter.

The equation (5.22) is a nonsingular Sturm-Liouville problem and has real distinct eigenvalues clustering at $\infty$. The eigenvalues $\lambda_k$, $k = 1, 2, \cdots$, can be estimated by various techniques. In [39], a simple but powerful method gives upper and lower bounds for eigenvalues. We will be concerned here with eigenfunctions

subject to $y'(0) = y'(\frac{\pi}{2}) = 0$, which are usually denoted by $ce_{2n}(x, q)$. We assume

$$y(x) = \frac{y_0}{2} + \sum_{n=1}^{\infty} y_n \cos 2nx. \tag{5.23}$$

which, on substituting in (5.22), gives for arbitrary integers $p \geq 1$

$$\begin{aligned} \frac{\lambda}{2} y_0 - q\, y_1 &= 0. \\ y_{n-1} + \frac{4n^2 - \lambda}{q} y_n + y_{n+1} &= 0. \qquad n = 1, 2, \ldots, p. \\ y_{n-1} + \frac{4n^2 - \lambda}{q} y_n + y_{n-1} &= 0. \quad n = p+1, p+2, \ldots \infty. \end{aligned} \tag{5.24}$$

the first equation serving as a normalizing relation for a given $\lambda$.

The computation of Mathieu function (5.23) reduces to solving the following infinite tridiagonal linear system with given $q$ and known $\lambda$:

$$\begin{bmatrix} \frac{4-\lambda}{q} & 1 & & & & \\ 1 & \frac{4\cdot 2^2-\lambda}{q} & 1 & & & \\ & & \ddots & & & \\ & & 1 & \frac{4\cdot n^2-\lambda}{q} & 1 & \\ & & & & \ddots & \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ \vdots \end{bmatrix} = \begin{bmatrix} y_0 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix}.$$

For a given $q$ and a known value of $\lambda$, it is not difficult to see that the system (5.24), for a suitable choice of $p$, gives an infinite diagonally dominant system where the infinite matrix satisfies all the conditions of Theorem 5.5. Also from Section 3,

$$\mu_n = \frac{|a_n| + |c_n|}{|b_n|} = \frac{2q}{4n^2 - \lambda} \sim \frac{q}{2} n^{-2}. \tag{5.25}$$

From (5.15) we get an error bound.

$$\| Y - \tilde{Y}^{(m)} \|_\infty \leq \frac{\sup_{l>m}\{\mu_l\} \| D \|_\infty}{\delta(1-\mu)^2}.$$

where $Y = (y_{p+1}, y_{p+2}, \cdots\cdots, y_{p+m})$ and $\tilde{Y}^{(m)}$ is the solution for the $m \times m$ truncated system of (5.24).

It is easy to check that the error bound given in (5.15) is $O(n^{-2})$. To illustrate the theory with numerical work, we let $q = 1$, and the corresponding eigenvalue $\lambda = -0.45513860$ [34] and normalize the solution with $y_0 = -7.5$. The truncation size is chosen to be $m = 20$. From an asymptotic analysis[22], we have, for large $n$

$$y_n = \left(\frac{-e^2 q}{4n^2}\right)^n \left[\frac{1}{n} + O(n^{-2})\right]. \tag{5.26}$$

The following table lists the eigenfunctions calculated from the $m \times m$ truncated system of (5.24) and the asymptotic approximations from (5.26).

Table 5.3: Values for $y_k$, m=20

| k | Truncated solutions | Asymptotic approximations |
|---|---|---|
| 1 | -0.1713596846E+01 | -0.1849600000E+01 |
| 2 | 0.1043114538E+00 | 0.1069068800E+00 |
| 3 | -0.2862583347E-02 | -0.2893241375E-02 |
| 4 | 0.4441888755E-04 | 0.4464484762E-04 |
| 5 | -0.4422068347E-06 | -0.4433217470E-06 |
| 6 | 0.3061313166E-08 | 0.3065495235E-08 |
| 7 | -0.1558306875E-10 | -0.1559818680E-10 |
| 8 | 0.6076406233E-13 | 0.6082648984E-13 |
| 9 | -0.1872817619E-15 | -0.1875394644E-15 |
| 10 | 0.4676746761E-18 | 0.4685739788E-18 |

# Chapter 6

# An Elliptic Boundary Value Problem Defined on an Infinite Domain

In this chapter we give a mathematical analysis with numerical computation for a groundwater flow problem described by an elliptic equation of the form

$$\nabla \cdot \left( e^{dz} \nabla \phi(x, z) \right) = 0, \quad d \geq 0$$

in a semi-infinite vertical region bounded on top by a sloping sinusoidal curve, under given boundary conditions. $\phi(x, z)$ represents the hydraulic head and $e^{dz}$ represents the relative hydraulic conductivity (or permeability). We reduce the problem to an infinite system of linear equations using the method of separation of variables and construction of a Grammian matrix. Truncation of this system yields an approximate solution that gives the best match on the top boundary. Computational results for some typical parameters are presented.

## 6.1 Introduction

We consider the problem of analyzing the motion of groundwater in a small drainage basin. If a cross-section of this basin is taken which is normal to the regional topographic trend and parallel to the regional hydraulic gradient, this results in a two-dimensional system in $(x, z)$ coordinates, with $x$ representing the horizontal coordinate, and $z$ represents the elevation. The velocity potential, $\phi(x, z)$, satisfies the equation $\nabla \cdot (K(z)\nabla\phi(x, z)) = 0$, where $K(z)$ is the hydraulic conductivity. For reference, see the papers by Tóth [25],[26].

There is a wide variety of concepts and modelling approaches to groundwater flow problems in the literature (see, for example, [24]). In order to produce a mathematical solution, the region under consideration has been taken to be finite, the hydraulic conductivity, $K(z)$, a constant or changing only in discrete regions, and the boundaries and boundary values approximated. Moreover, it has been observed that the usual approximation techniques of obtaining the solution including finite differences, finite elements, and perturbation techniques do not give completely satisfactory numerical results for the flow. Tóth [25], [26] has given analytical solutions for the boundary value problem for Laplace's equation representing a steady-state flow in a finite vertical, two-dimensional, saturated, homogeneous, isotropic region bounded on top by a sloping sinusoidal curve, which represents the watertable. However, he approximates the problem by replacing the semi-infinite region with a finite rectangle, and projecting the given boundary values onto the top of this rectangle. He then solves a reconstructed problem on this rectangle. This assumes that the solution has the same value on the top of

the rectangle as it did on the given boundary, which of course, is not accurate: this approach gives only a rough approximation, and then, only if both the angle of the sloping watertable and the amplitude of the sinusoidal curve are very small.

In this application, we are concerned with finding the hydraulic head, $o$, in a non-homogeneous porous medium. The region considered is bounded between two vertical impermeable boundaries, bounded on top by a sloping sinusoidal curve and unbounded in depth. A mathematical analysis is developed which reduces the problem to solving an infinite system of linear equations. There are many ways of producing such an infinite system. Our method yields the Grammian matrix which is positive definite, and the truncation of this system yields an approximate solution that provides the best match with the given values on the top sloping sinusoidal boundary. Graphs of the equi-potential lines for $o$ . (i.e., the curves $o = c_1$), and their corresponding orthogonal trajectories, the streamlines, (given by curves $\psi = c_2$) are given. There has been a scarcity of work in groundwater flow problems involving infinite regions, particularly with complicated boundaries. In fact, for the problem under consideration, finite difference methods and finite element methods gave completely different descriptions of the flows.

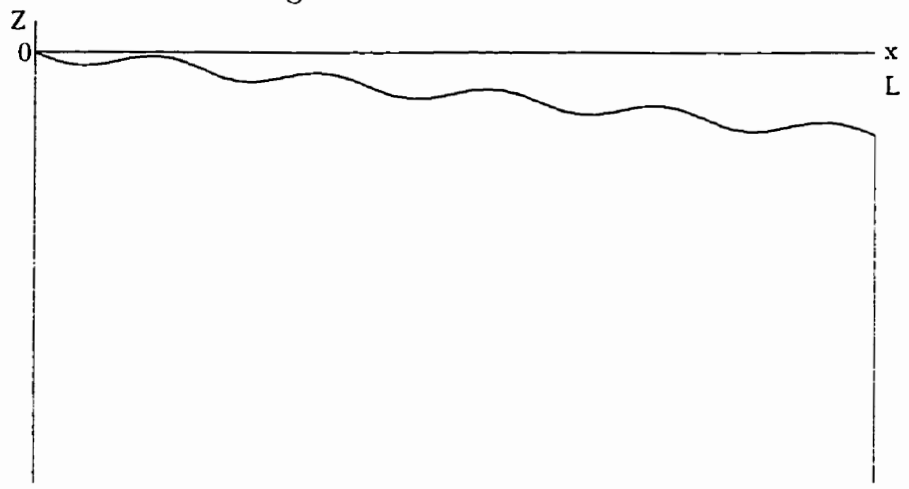
## 6.2 Mathematical model

The hydraulic head, $o(x, z)$, which is the hydraulic potential divided by the constant gravitational acceleration, satisfies the elliptic partial differential equation

$$\nabla \cdot \left( e^{dz} \nabla o(x, z) \right) = 0 \tag{6.1}$$

where $\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j}$. $z$ denotes the height of a point (relative to a vertical scale chosen so that $z = 0$ at one corner of our region). and $d \geq 0$. The hydraulic conductivity is $K = \alpha e^{dz}$. where $\alpha$ is a positive constant. which is in qualitative agreement with the generally observed decrease in conductivity with depth in a well. The region under consideration for equation (6.1) is given by

Figure 6.1: Semi-infinite domain




(6.2) $0 < x < L$. *and* $-\infty < z < g(x) = -\left[\frac{ax}{L} + V \sin\left(\frac{2\pi nx}{L}\right)\right]$.

where $d$. $L$. $a$. $V$ and $n$ are real constants (parameters) with $d \geq 0$. $L > 0$. $a \geq 0$. and $n$ is a positive integer.

The boundary conditions are given by

(6.3) $$\frac{\partial \phi}{\partial x}|_{x=0} = \frac{\partial \phi}{\partial x}|_{x=L} = 0.$$

(6.4) $$\frac{\partial \phi}{\partial z} \to 0 \quad as \quad z \to -\infty. \quad \phi(x,z) \quad is \quad bounded \quad on \quad z \le g(x).$$

and

(6.5) $$\phi(x,z) = z \quad on \quad z = g(x).$$

where $g(x)$ is defined in (6.2).

The solution to (6.1) represents the hydraulic head (or potential) and thus the equi-potential lines, are given by $\phi(x,z) =$ constant. Of interest also are the streamlines, which are the orthogonal trajectories to $\phi(x,z) =$ constant and are the solution of

(6.6) $$\frac{dz}{dx} = \frac{\partial \phi}{\partial z} \Big/ \frac{\partial \phi}{\partial x}.$$

Therefore, we solve (6.6) with $z(x_0) = z_0$, where $(x_0, z_0)$ is an arbitrary given point in the region; the solution is the orthogonal trajectory that passes through $(x_0, z_0)$, and will be of the form, $\psi(x,z) =$ constant.

## 6.3 Formal solution

Expanding (6.1), we obtain

(6.7) $$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial z^2} + d\frac{\partial \phi}{\partial z} = 0$$

Using the method of separation of variables, we set $\phi(x,z) = X(x)Z(z)$. Then (6.7) gives $\frac{X''}{X} = -\frac{Z''+dZ'}{Z} = -\mu$, where $\mu$ is the separation constant, from which

we obtain

$$X'' + \mu X = 0. \tag{6.8}$$

and

$$Z'' + dZ' - \mu Z = 0. \tag{6.9}$$

From (6.3). $X'(0) = X'(L) = 0$. This. together with (6.8) implies that

$$X(x) = \gamma_m \cos\left(\frac{m\pi x}{L}\right).$$

and $\mu = \left(\frac{m\pi}{L}\right)^2$. for $m = 0. 1. 2. ...$: (the $\gamma_m$ are arbitrary constants).

The solution of (6.9) is given by $Z(z) = e^{rz}$ where $r$ satisfies $r^2 + dr - \mu = 0$. However. from the boundary condition (6.4). we have

$$Z(z) \quad is \quad bounded \quad as \quad z \to \infty. \tag{6.10}$$

Thus. the solution $Z$ is given by

Case (1) $\mu = 0. d = 0$: $Z = \delta_0 + \delta_1 z$. and thus. $Z = \delta_0$ by (6.10): (the $\gamma_m$ are arbitrary constants).

Case (2) $\mu = 0. d > 0$: $Z = \delta_0 + \delta_1 e^{-dz}$. and thus. $Z = \delta_0$ by (6.10).

Case (3) $\mu > 0$: $r = \frac{-d+\sqrt{d^2+4\mu}}{2} > 0$. The negative sign on the root would make $r < 0$ and would violate (6.10). Thus. $Z(z) = \delta_m e^{rz}$. where $\delta_m$ is a constant.

Thus. the most general solution of (6.1) that satisfies (6.3) and (6.4) is given by

$$\phi(x, z) = \sum_{m=0}^{\infty} \beta_m \cos\left(\frac{m\pi x}{L}\right) e^{\rho_m z}. \tag{6.11}$$

where

$$\rho_m = \frac{1}{2}\left[-d + \sqrt{d^2 + \frac{4m^2\pi^2}{L^2}}\right] > 0. \quad \beta_m = \gamma_m \delta_m. \tag{6.12}$$

The remaining boundary condition (6.5). is equivalent to

$$-\left[\frac{ax}{L} + V \sin\left(\frac{2\pi n x}{L}\right)\right] = \sum_{m=0}^{\infty} \beta_m \cos\left(\frac{m\pi x}{L}\right) exp\left[-\rho_m\left\{\frac{ax}{L} + V \sin\left(\frac{2\pi n x}{L}\right)\right\}\right]. \tag{6.13}$$

The main problem now is to determine the coefficients $\{\beta_m : m = 0, 1, 2..\}$ in (6.13). Our aim is to pursue analytical methods as far as possible. and then to use numerical techniques at the last stage.

We make (6.13) non-dimensional by putting $y = x/L$. $\alpha_m = \beta_m/L$. $\sigma_m = \rho_m L$. $\tilde{a} = a/L$ and $\tilde{V} = V/L$. and dividing both sides by $L$. This gives. for $0 < y < 1$.

$$-\left[\tilde{a}y + \tilde{V}\sin(2\pi n y)\right] = \sum_{m=0}^{\infty} \alpha_m \cos(m\pi y) exp\left[-\sigma_m(\tilde{a}y + \tilde{V}\sin(2\pi n y)\right]. \tag{6.14}$$

Define

$$u_k(y) = \cos(k\pi y) exp\left[-\sigma_k(\tilde{a}y + \tilde{V}\sin(2\pi n y)\right]. \tag{6.15}$$

Then (6.14) can be written as

$$-\left[\tilde{a}y + \tilde{V}\sin(2\pi n y)\right] = \sum_{m=0}^{\infty} \alpha_m u_m(y). \quad 0 < y < 1. \tag{6.16}$$

We multiply each side of (6.16) by $u_k(y)$ and integrate with respect to $y$ over

$[0.1]$ to obtain

$$-\int_0^1 \left[\tilde{a}y + \tilde{\Gamma}\sin(2\pi ny)\right] u_k(y)dy =$$

$$\sum_{m=0}^{\infty} \alpha_m \int_0^1 u_m(y)u_k(y)dy. \quad k = 0.1.2.... \tag{6.17}$$

where we assume that the series in (6.17) converges uniformly so that term-by-term integration is justified. (6.17) is of the form

$$\sum_{m=0}^{\infty} b_{km}\alpha_m = c_k. \quad k = 0.1.2.... \tag{6.18}$$

where, for $k.m = 0.1.2....$

$$\begin{aligned} b_{km} &= \int_0^1 u_m(y)u_k(y)dy \\ &= \int_0^1 \cos(m\pi y)\cos(k\pi y)exp\left[-(\sigma_m + \sigma_k)\left\{\tilde{a}y + \tilde{\Gamma}\sin(2\pi ny)\right\}\right] dy. \end{aligned} \tag{6.19}$$

and

$$\begin{aligned} c_k &= \int_0^1 \left[\tilde{a}y + \tilde{\Gamma}\sin(2\pi ny)\right] u_k(y)dy \\ &= \int_0^1 \left[\tilde{a}y + \tilde{\Gamma}\sin(2\pi ny)\right]\cos(k\pi y)exp\left[-\sigma_k\left\{\tilde{a}y + \tilde{\Gamma}\sin(2\pi ny)\right\}\right] dy. \end{aligned} \tag{6.20}$$

Using (6.12).

$$\sigma = \rho_m L = \frac{2m^2\pi^2}{dL + \sqrt{d^2L^2 + 4m^2\pi^2}}. \tag{6.21}$$

The infinite matrix $\mathbf{B} = [b_{km}]_{k,m=0,1,2..}$ as defined in (6.19) is called the Gram-

mian of the set $\{u_k : k = 0, 1, 2...\}$. $\mathbf{B}$ is a symmetric, positive definite (and thus, invertible) matrix, where we assume that $\{u_k : k = 0, 1, 2...\}$ is linearly independent.

We note that $\sigma_0 = 0$, $u_0(y) = 1$, $b_{00} = 1$ and $c_0 = -\frac{\tilde{a}}{2}$.

To evaluate the integrals in (6.17), or equivalently, in (6.19) and (6.20) analytically, we use the following identity [23]:

$$e^{-p\sin\theta} = I_0(p) + 2\sum_{q=1}^{\infty}(-1)^q I_{2q}(p)\cos(2q\theta) + 2\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(p)\sin(2q-1)\theta. \tag{6.22}$$

where the $I_m$ are the modified Bessel functions given by

$$I_m(x) = \left(\frac{x}{2}\right)^m \sum_{k=0}^{\infty}\frac{(x/2)^{2k}}{k!(k+m)!}, \quad m = 0, 1, 2, \ldots \tag{6.23}$$

Defining $p_{k,m}$, $q_{k,m}$ and $\omega$ as

$$p_{k,m} = \tilde{\Gamma}(\sigma_k + \sigma_m), \quad q_{k,m} = \tilde{a}(\sigma_k + \sigma_m), \quad \omega = 2n\pi,$$

and using (6.22) in (6.19) and performing some simplifications, we obtain the following:

$$\begin{aligned} b_{k,m} &= \frac{1}{2}I_0(p_{k,m})\frac{q_{k,m}\left(1 - e^{-q_{k,m}}(-1)^{k+m}\right)}{(m \pm k)^2\pi^2 + q_{k,m}^2} \\ &+ \frac{1}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q}(p_{k,m})\frac{q_{k,m}\left(1 - e^{-q_{k,m}}(-1)^{k+m}\right)}{[(m \pm k)\pi \pm 2q\omega]^2 + q_{k,m}^2} \end{aligned} \tag{6.24}$$

$$+\frac{1}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(p_{k,m})\frac{[(2q-1)\omega \pm (m \pm k)\pi]\left(1-e^{-q_{k,m}}(-1)^{k+m}\right)}{[(2q-1)\omega \pm (m \pm k)\pi]^2+q_{k,m}^2}.$$

for $k+m>0$. and for $k>0$.

(6.25) $$c_k = -(T_1+T_2+T_3).$$

where

(6.26)

$$\begin{aligned}
T_1 &= \tilde{a}\, I_0(\sigma_k\tilde{r})\left\{-\frac{\sigma_k\tilde{a}\cdot e^{-\sigma_k\tilde{a}}(-1)^k}{(\sigma_k\tilde{a})^2+(k\pi)^2}\frac{[(\sigma_k\tilde{a})^2-(k\pi)^2]\left(1-(-1)^k e^{-\sigma_k\tilde{a}}\right)}{[(\sigma_k\tilde{a})^2+(k\pi)^2]^2}\right\}\\
&\quad +\frac{\tilde{r}}{2}I_0(\sigma_k\tilde{r})\frac{(\omega \pm k\pi)(1-e^{-\sigma_k\tilde{a}}(-1)^k)}{(\omega \pm k\pi)^2+(\sigma_k\tilde{a})^2}\\
T_2 &= \tilde{a}\sum_{q=1}^{\infty}(-1)^q I_{2q}(\sigma_k\tilde{r})\left\{-\frac{\sigma_k\tilde{a}\cdot e^{-\sigma_k\tilde{a}}(-1)^k}{(\sigma_k\tilde{a})^2+(k\pi \pm 2q\omega)^2}\right.\\
&\quad \left.+\frac{[(\sigma_k\tilde{a})^2-(k\pi \pm 2q\omega)^2]\left(1-(-1)^k e^{-\sigma_k\tilde{a}}\right)}{[(\sigma_k\tilde{a})^2+(k\pi \pm 2q\omega)^2]^2}\right\}\\
&\quad +\frac{\tilde{r}}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q}(\sigma_k\tilde{r})\frac{[\pm k\pi \pm 2q\omega+\omega]\left(1-(-1)^k e^{-\sigma_k\tilde{a}}\right)}{[\pm k\pi \pm 2q\omega+\omega]^2+(\sigma_k\tilde{a})^2}\\
T_3 &= \tilde{a}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(\sigma_k\tilde{r})\left\{-\frac{[(2q-1)\omega \pm k\pi]\cdot e^{-\sigma_k\tilde{a}}(-1)^k}{(\sigma_k\tilde{a})^2+[(2q-1)\omega \pm k\pi]^2}\right.\\
&\quad \left.+\frac{2\sigma_k\tilde{a}[(2q-1)\omega \pm k\pi]\left(1-(-1)^k e^{-\sigma_k\tilde{a}}\right)}{[(\sigma_k\tilde{a})^2+((2q-1)\omega \pm k\pi)^2]^2}\right\}\\
&\quad +\frac{\tilde{r}}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(\sigma_k\tilde{r})\left\{\frac{\sigma_k\tilde{a}\left(1-(-1)^k e^{-\sigma_k\tilde{a}}\right)}{[(2q-2)\omega \pm k\pi]^2+(\sigma_k\tilde{a})^2}\right.
\end{aligned}$$

$$\left. - \frac{\sigma_k \tilde{a}\left(1-(-1)^k e^{-\sigma_k \tilde{a}}\right)}{[2q\omega \pm k\pi]^2 + (\sigma_k \tilde{a})^2} \right\}$$

Remark: Where the $\pm$ sign appears only in a denominator, we sum over two terms, one for each sign, or over four terms if there are a pair of $\pm$ signs; where one or two $\pm$ signs appear in both the numerator and denominator of some term, sum over all combinations of signs in the numerator and put the same combination of signs in the denominator. For example,

$$\frac{a \pm b}{(a \pm b)^2 + c} = \frac{a+b}{(a+b)^2 + c} + \frac{a-b}{(a-b)^2 + c}.$$

## 6.4 Numerical approximation and error estimation

The analytical solution, $\phi(x, z)$, to our problem is described by (6.11) and (6.12), but involves solving an infinite linear system, which is not readily solvable. To arrive at a practical solution, we adopt a numerical approximation. We consider a truncated solution of the form $\phi_N(y, z) = L \sum_{m=0}^{N} \alpha_m \cos(m\pi y) e^{\rho_m z}$ which satisfies (6.1) and all the boundary value conditions except the one on

$$z = g(y) = -L[\tilde{a}y + \tilde{\Gamma} \sin(2\pi n y)].$$

The difference between these gives an indication of the error. Let

$$e_N(y) = \phi_N(y, z) - z, \text{ with } z = g(y) = -L[\tilde{a}y + \tilde{\Gamma} \sin(2\pi n y)].$$

i.e., $e_N(y) = \phi_N(y, g(y)) - g(y)$.

Substituting $g(y)$ and $\phi_N(y, g(y))$ into $e_N(y)$, we get

$$e_N(y)/L = \tilde{a}y + \tilde{V}\sin(2\pi ny) + \sum_{m=0}^{N} \alpha_m \cos(m\pi y) exp\left[-\sigma_m \left\{\tilde{a}y + \tilde{V}\sin(2\pi ny)\right\}\right].$$

We choose $\{\alpha_m : m = 0, 1, 2, \ldots, N\}$ to minimize the $L_2$-norm of the error function $e_N(y)$, i.e. $\{\alpha_m\}$ is the solution to the minimization problem:

$$\min_{\alpha_0, \alpha_1, \ldots, \alpha_N} \| u(y) + \sum_{m=0}^{N} \alpha_m u_m(y) \|.$$

where $u_k(y) = \cos(k\pi y) exp\left[-\sigma_k \left\{\tilde{a}y + \tilde{V}\sin(2\pi ny)\right\}\right]$, and $u(y) = \tilde{a}y + \tilde{V}\sin(2\pi ny)$. The solution to this problem is determined by satisfying the equation

$$\langle u_k, u + \sum_{m=0}^{N} \alpha_m u_m \rangle = 0,$$

i.e.

$$\sum_{m=0}^{N} b_{km}\alpha_m = c_k.$$

So $\alpha^N = (\alpha_0, \alpha_1, \ldots, \alpha_N)^t$ is the solution to $B_N\alpha^N = c^N$, where $B_N$ is the $N \times N$ truncation of $B$ and $c^N = (c_0, c_1, \ldots, c_N)^t$. We note that $B_N$ is a symmetric, positive definite matrix, where we assume that $\{u_0, u_1, \ldots, u_N\}$ is linearly independent. The above procedure provides the best approximation (or the best matching on the top boundary) for a given $N$. After $\{\alpha_m : m = 0, 1, 2, \ldots, N\}$ has been computed, we set $M = \max_{0 \le y \le 1} \|g(y)\|$; then, $\varepsilon = \max_{0 \le y \le 1} \frac{\|e_N(y)\|}{M}$ can be easily estimated. We increase $N$ until $\varepsilon$ is within a required accuracy. Then the approx-

imation $o_N(y, z)$ is the solution to a perturbed problem, i.e., $o_N(y, z)$ satisfies the differential equation (6.1) on the region (6.2), and the boundary conditions (6.3) and (6.4). In place of the boundary condition (6.5), $o_N(y, z)$ satisfies the perturbed boundary condition:

$$o_N(y, g(y)) = g(y) + e_N(y),$$

with

$$\|e_N(y)\| \leq \varepsilon \max_{0 \leq y \leq 1} \|g(y)\|.$$

## 6.5 Numerical results

We approximate the solution of the boundary value problem (6.1) under the boundary condition (6.3) using the numerical procedure of section 7.4 with the parameters given by:

| | |
|---|---|
| Length of basin: | $L = 80.000.$ |
| Slope of top boundary: | $a/L = 0.1.$ |
| Depth of humps: | $V/L = 0.01.$ |
| Hydraulic conductivity: | $d = 0.$ 0.00235, and .0235 (Figures 6.2, 6.3, 6.4). |

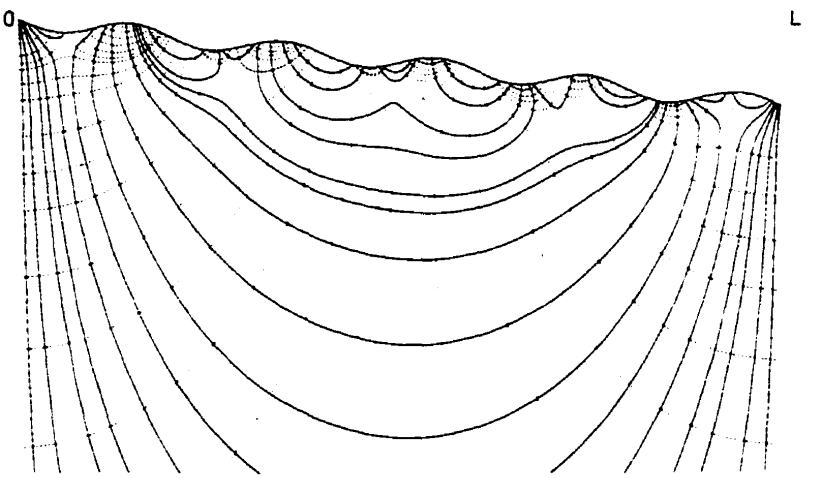
We compute for the given setting the velocity potential $o(x, z)$ at chosen points on the top boundary for comparison with the boundary condition (6.3) at the same points with actual relative errors. The equi-potential lines ($o$ = constant) and the streamlines ($v$ = constant) are plotted in Figures 6.2, 6.3, and 6.4.

Table 6.1: Approximation and actual error on top boundary

$a/L = 0.1,\ V/L = 0.01,\ d = 0.00235,\ L = 80,000$

| $x/L$ | $g(x)$ | $\phi(x, g(x))$ | $\frac{\lvert\phi(x,g(x))-g(x)\rvert}{\max\lvert g(x)\rvert}$ |
|---|---|---|---|
| 0.00000000E+00 | 0.00000000E+00 | -0.23222376E+02 | 0.29027970E-02 |
| 0.50251256E-01 | -0.12019851E+04 | -0.12013125E+04 | 0.84075569E-04 |
| 0.10050251E+00 | -0.79139113E+03 | -0.79113371E+03 | 0.32176654E-04 |
| 0.15075377E+00 | -0.40625443E+03 | -0.40633264E+03 | 0.97767585E-05 |
| 0.20100502E+00 | -0.16332950E+04 | -0.16333726E+04 | 0.96980908E-05 |
| 0.25125628E+00 | -0.28094272E+04 | -0.28096043E+04 | 0.22126669E-04 |
| 0.30150753E+00 | -0.23741860E+04 | -0.23741626E+04 | 0.29233124E-05 |
| 0.35175879E+00 | -0.20152912E+04 | -0.20154705E+04 | 0.22414179E-04 |
| 0.40201005E+00 | -0.32665648E+04 | -0.32664682E+04 | 0.12075630E-04 |
| 0.45226130E+00 | -0.44160725E+04 | -0.44162223E+04 | 0.18718764E-04 |
| 0.50251256E+00 | -0.39570186E+04 | -0.39570276E+04 | 0.11361335E-05 |
| 0.55276381E+00 | -0.36251242E+04 | -0.36253286E+04 | 0.25542214E-04 |
| 0.60301507E+00 | -0.48997842E+04 | -0.48997991E+04 | 0.18538393E-05 |
| 0.65326632E+00 | -0.60219224E+04 | -0.60220771E+04 | 0.19336656E-04 |
| 0.70351758E+00 | -0.55399140E+04 | -0.55396664E+04 | 0.30950067E-04 |
| 0.75376884E+00 | -0.52357517E+04 | -0.52357205E+04 | 0.39036886E-05 |
| 0.80402009E+00 | -0.65329283E+04 | -0.65331829E+04 | 0.31828135E-04 |
| 0.85929647E+00 | -0.76404946E+04 | -0.76408081E+04 | 0.39178008E-04 |
| 0.90452260E+00 | -0.71228975E+04 | -0.71236262E+04 | 0.91096264E-04 |
| 0.95477386E+00 | -0.68471710E+04 | -0.68488603E+04 | 0.21116008E-03 |
| 0.99999999E+00 | -0.79999996E+04 | -0.79545852E+04 | 0.56768093E-02 |

Figure 6.2: Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.0$. $L = 80.000$.



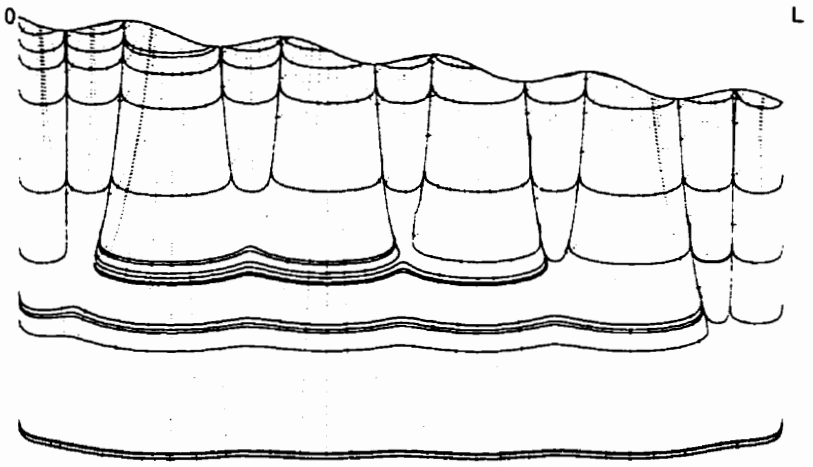

## 6.6 Conclusions and future work

We have developed a method based on a combination of analytic and numerical techniques to solve the groundwater flow problem which is modeled by equation (6.1) in a semi-infinite region with sloping sinusoidal top boundary. Using separation of variables, we reduced the problem to one of matching the formal solution to the given values on the top boundary, which is solved numerically to a required accuracy, and this yields an optimal approximation. The method is simple and mathematically sound. In particular, for the special case of Laplace's equation ($d = 0$) (which is the subject of [26]), our result has the same qualitative behaviour as [26]. However, our solution is more accurate than [26], and is valid on the entire region under consideration.

For future work, we propose the following suggestions:

1. Solve the problem with the hydraulic conductivity, $K(z)$, modeled by some

Figure 6.3: Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.00235$. $L = 80.000$.

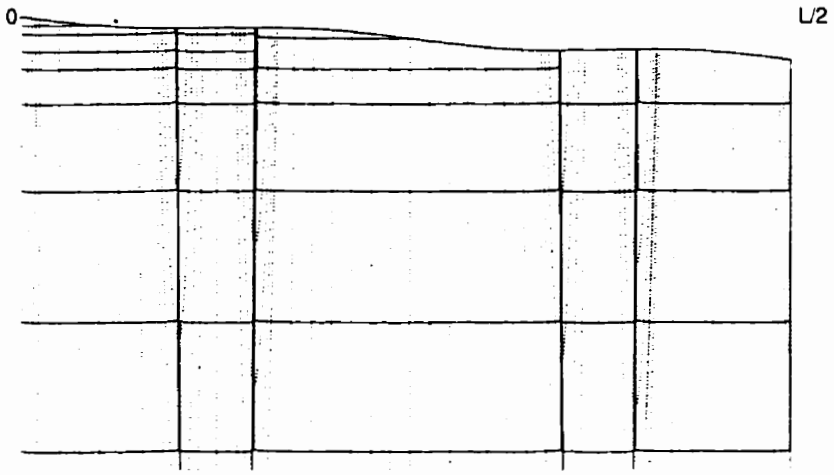



explicitly given formula rather than the exponential function: for example, $K(z)$ could be a rational function or a piecewise defined step function that better models the layer structure in the ground.

2. Solve the problem for other boundary curves. The boundary curve $g(x)$ could be any explicitly given function, or could be found by using curve-fitting on arbitrarily given data. We can use numerical integration instead of Bessel series.

3. The solution could be developed, and the problem solved, for a three- dimensional setting, i.e., $\phi = \phi(x, y, z)$.

Figure 6.4: Level curves for $\phi$ (dotted) and $\psi$ (solid). $a/L = 0.1$. $V/L = 0.01$. $d = 0.0235$. $L = 80.000$.



# Bibliography


[1] R. W. Hockney. *A Solution of Laplace's Equation for A Round Hole in A Square Peg.* J. Soc. Ind. Appl. Math., Vol. 12, No. 2, March 1964. 1.

[2] P. A. Laura. *Conformal Mapping of A Class of Doubly Connected Regions.* NASA Tech. Rep. NO. 8. Catholic University of America. Washington.

[3] M. Z. Narodetskii. D. I. Sherman. *A Problem in Conformal Transformation.* PRIKL. MAT. i MEKH. XIV (1950). 209.

[4] G. T. Symm. *Conformal Mapping of Doubly Connected Domains.* Numerical Mathematics. XIII. 1969. 448.

[5] H. B. Wilson. *A Method of Conformal Mapping and the Determination of Stresses in Solid Propellant Rocket Grains.* Rep. No. S-38. Rohm and Haas Co., Alabama. 1963.

[6] Bengt Fornberg. *A Numerical Method for Conformal Mapping of Doubly Connected Regions.* SIAM Journal of Science. Statistics and Computation. Vol. 5. No. 4. Dec. 1984. pp 771-783.

[7] Klaus Menke. *Conformal Mapping of Doubly Connected Regions.* Complex Variables. Theory and Application. 14. 1990. No.1-4. pp 251-260.

[8] Klaus Menke. *Point System with Extremal properties and Conformal Mapping.* Numerical Mathematics. 54. 1988. No.2. pp 125-143.

[9] Rudolf Wegmann. *An Iterative Method for the Conformal Mapping of Doubly Connected Regions.* Journal of Computational and Applied Mathematics. 14. 1986. No.1-2. pp 79-98.

[10] N. Papamichael. *The Use of Singular Functions for the Approximation Conformal Mapping of Doubly Connected Domain.* SIAM J. Sci. Statist Comput.. 5. 1984. No.3 pp 684-700.

[11] C. Caratheodory. *Theory of Functions of a Complex Variable.* Vol. 2. Chelsea Pub. Co.. N.Y. (1956).

[12] John Drew and Charles R. Johnson. *Strong Forms of Nonsingularity.* Linear Algebra and its Applications. No. 162-164. pp. 187-204. 1992.

[13] K. Fan and A.S. Householder. *A Note Concerning Positive Matrices and M-Matrices.* Monatshefte Math. 63. pp. 265-270. 1959.

[14] M. Fiedler and V.P. Praha. *On Matrices with Non-Positive Off-Diagonal Elements and Positive Principal Minors.* Czechoslovak Math. J.. Vol 12. pp. 382-400. 1962.

[15] C. Thomassen. *When the Sign Distribution of A Square Matrix Determines Uniquely the Sign Pattern of its Inverse.* Linear Algebra and its Applications. Vol 119. pp. 27-34. 1989.

[16] L.V. Atkinson, P.J. Harley and J.D. Hudson. *Numerical Methods with Fortran 77. A Practical Introduction* . International Computer Science Series. 1989.

[17] D. Kershaw. *Inequalities on the Elements of the Inverse of a Certain Tridiagonal Matrix.* Math. Comp., Vol 24. 1970. pp 155-158.

[18] R. Mattheij. *Estimates for the Inverse of Tridiagonal Matrices Arising in Boundary Value Problems.* Linear Algebra and its Applications. Vol 73. 1986. pp 33-57.

[19] A.M. Ostrowski. *Note on Bounds for Determinants with Dominant Principal Diagonal.* Proc. Amer. Math. Soc., Vol 3. 1952. pp 26-30.

[20] J.M. Varah. *A Lower Bound for the Smallest Singular Value of A Matrix.* Linear Algebra and its Applications. Vol 11. 1975. pp 3-5.

[21] A.E. Taylor. *Introduction to Functional Analysis.* Wiley. 1958.

[22] Jet Wimp. *Computation with Recurrence Relations.* 1984. pp 82-85.

[23] M. Abromovich and I. A. Stegun. *Handbook of Mathematical Functions.* Dover Pub., New York, 1965 (corrected printing, 1972).

[24] R.A. Freeze and P.A. Witherspoon. *Theoretical Analysis of Regional Groundwater Flow: 1. Analytical and Numerical Solutions to the Mathematical Model.* Water Resour. Res., 2(4), 641 - 656. 1966.

[25] J. Tóth. *A Theory of Groundwater Motion in Small Drainage Basins in Central Alberta, Canada.* J. Geophys. Res., 67. 4375 - 4381. 1962.

[26] J. Tóth. *A Theoretical Analysis of Groundwater Flow in Small Drainage Basins.* J. Geophys. Res.. 67. 4795 - 4812. 1963.

[27] G.Polya and A.Weinstein. *On the Torsional Rigidity of Multiply Connected Cross-Sections.* Annals of Mathematics. No.2. 52. 1950. pp. 154-163.

[28] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland Publications. Amsterdam. 1978.

[29] R. A. Adams. *Sobolev Spaces.* Academic Press. New York. 1975.

[30] R.G. Cooke. *Infinite Matrices and Sequence Spaces.* Dover. New York. 1955.

[31] M. Bernkopf. *A History of Infinite Matrices. Archive for History of Exact Sciences.* Vol. 4. Number 4. 1968. pp 308-358.

[32] P.N.Shivakumar. *Dislocations in isotropic cylinders of eccentric circular cross-sections.* Quarterly Journal of Mechanics and Applied Mathematics. Vol.16. Part 2. 1963. pp 129-136.

[33] P.N.Shivakumar. *Diagonally dominant infinite matrices in linear equations.* Utilitas Mathematica. Vol. 1. 1972. pp 235-248.

[34] P.N. Shivakumar and R. Wong. *Linear Equations in Infinite Matrices.* Linear Algebra and its Applications. Vol. 7. 1973. pp 53-62.

[35] P.N.Shivakumar. *Viscous flow in pipes whose cross-sections are doubly connected regions.* Applied Scientific Research. Vol. 27. 1973. pp 355-365.

[36] P.N.Shivakumar and K.H.Chew. *A sufficient condition for the vanishing of determinants*, Proceedings of the American Mathematical Society. Vol. 43. No. 1. 1974. pp 63-66

[37] P.N.Shivakumar and K.H.Chew. *Iterations for diagonally dominant matrices.* Bulletin of the Canadian Mathematical Society. Vol. 19 (3). 1976. pp 375-377

[38] P.N.Shivakumar and K.H.Chew, *On the conformal mapping of a circular disc with a curvilinear polygonal hole*, Proceedings of the Indian Academy of Sciences. Vol. 85. A. No. 6. 1977. pp 406-414

[39] P.N. Shivakumar. J.J. Williams and N. Rudraiah. *Eigenvalues for Infinite Matrices*. Linear Algebra and its Applications. Vol 96. 1987. pp 35-63.

[40] P.N.Shivakumar and J.J.Williams. *An iterative method with truncation for infinite linear systems*. Journal of Computational and Applied Mathematics. 24. 1988. pp 109-207.

[41] P.N.Shivakumar and C.Ji. *On Poisson's equation for doubly connected regions*. Canadian Applied Mathematics Quarterly. Vol.1. No. 4. Fall 1993.

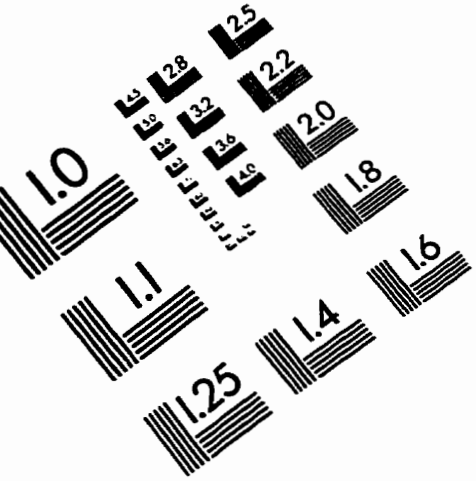
[42] P.N.Shivakumar and C.Ji. *Upper and lower bounds for inverse elements of finite and infinite tri- diagonal matrices*. Linear Algebra and its Applications 247. 1996. pp 297-316.

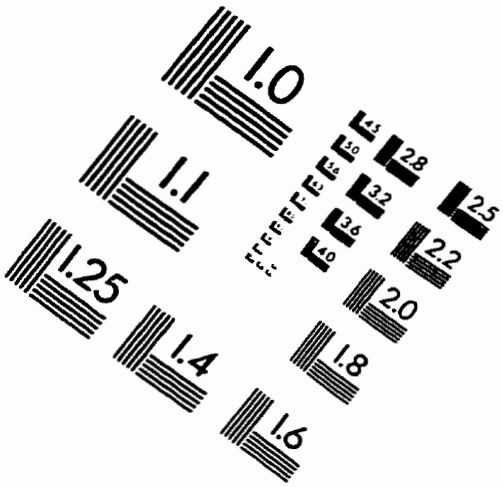
[43] P.N.Shivakumar and C.Ji. *On the nonsingularity of matrices with certain sign patterns*. final revision submitted to Linear Algebra and its Applications.

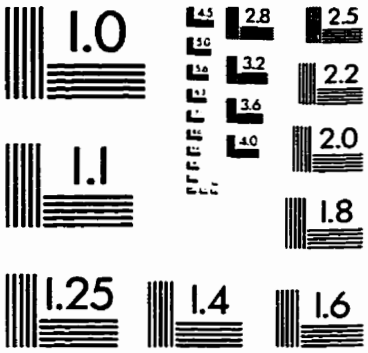
[44] N.Yan, P.N.Shivakumar and C.Ji, *A finite element method for double integrals with application to a Dirchlet problem arising in fluid dynamics*, accepted for publication, Communications in Numerical Methods in Engineering.

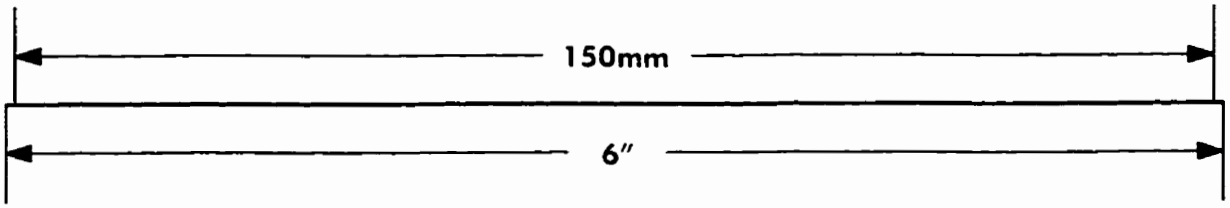
[45] C.Ji, P.N.Shivakumar, Q.Ye and J.J.Williams, *An analysis of groundwater flow in an infinite region with a sinusoidal top*, final revision submitted to International Journal for Numerical and Analytical methods in Geomechanics.

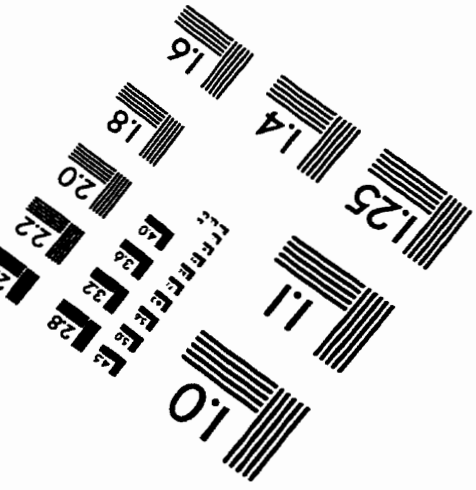
# IMAGE EVALUATION
# TEST TARGET (QA-3)

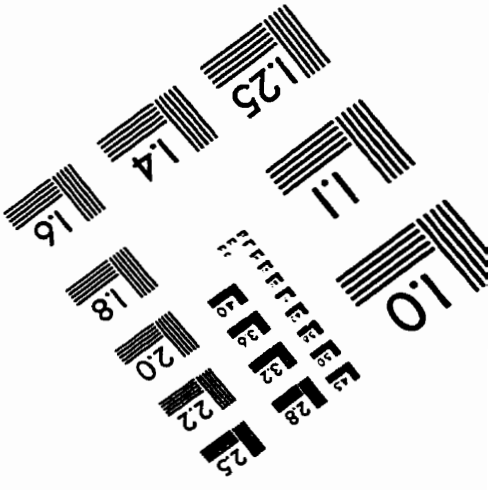



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989