

**A PIECEWISE HERMITE BICUBIC
FINITE ELEMENT GALERKIN METHOD
FOR THE BIHARMONIC DIRICHLET PROBLEM**

by

David Brian Knudson

ARTHUR LAKES LIBRARY
COLORADO SCHOOL OF MINES
GOLDEN, CO 80401

ProQuest Number: 10796744

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10796744

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Mathematical and Computer Sciences).

Golden, Colorado
Date 10/21/97

Signed: David B. Knudson
David Brian Knudson

Approved: B. Bialecki
Dr. Bernard Bialecki
Thesis Advisor

Golden, Colorado
Date 10/21/97

Graeme Fairweather
Dr. Graeme Fairweather
Professor and Head
Department of Mathematical and
Computer Sciences

ABSTRACT

We consider the Ciarlet-Raviart mixed finite element Galerkin method with piecewise Hermite bicubics for the solution of the biharmonic Dirichlet problem $\Delta^2 u = f$ on the unit square Ω with $u = \partial u / \partial n = 0$ on $\partial\Omega$. We prove existence and uniqueness of the Galerkin solution. We use a Schur complement approach to reduce the Galerkin problem to a Schur complement system involving the approximation to Δu on the two vertical sides of $\partial\Omega$ and to an auxiliary Galerkin problem for a related biharmonic problem with Δu instead of $\partial u / \partial n$ specified on the two vertical sides of $\partial\Omega$. The Schur complement system with a symmetric and positive definite matrix is solved by the preconditioned conjugate gradient method. A preconditioner is obtained from the Galerkin problem for a related biharmonic problem with Δu instead of $\partial u / \partial n$ specified on the two horizontal sides of $\partial\Omega$. We conjecture that the preconditioner is spectrally equivalent to the Schur complement matrix. On an $N \times N$ partition the cost of solving the preconditioned system and the cost of multiplying the Schur complement matrix by a vector are $O(N^2)$ each. With the number of iterations proportional to $\log_2 N$, the cost of solving the Schur complement system is $O(N^2 \log_2 N)$. The solution to the auxiliary Galerkin problem is obtained

using separation of variables and fast Fourier transforms at a cost of $O(N^2 \log_2 N)$. Hence the total computational cost of solving the Galerkin problem is $O(N^2 \log_2 N)$. Numerical results indicate that the L^2 and H^1 norm errors in the approximations to u and Δu are of optimal fourth and third orders, respectively. Convergence at the nodes is fourth order for the approximations to u and Δu and third order for the approximations to the first order derivatives of u and Δu .

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES	vii
ACKNOWLEDGEMENTS	viii
1 INTRODUCTION	1
2 PRELIMINARIES	7
2.1 Partition, Basis Functions, Stiffness and Mass Matrices	7
2.2 Tensor Product Notation	13
2.3 Generalized Eigenvalue Problem	16
3 GALERKIN METHOD FOR THE BIHARMONIC DIRICHLET PROBLEM	34
3.1 Mixed Finite Element Galerkin Method	34
3.2 Matrix-Vector Form	37
4 METHOD FOR SOLVING THE GALERKIN PROBLEM	46

4.1	Derivation of Algorithm for Solving the Galerkin Problem	46
4.2	Solving $M_{11}\vec{w} = \vec{b}$	54
4.3	Related Biharmonic Problem	69
4.4	Solving the Schur Complement System	74
4.4.1	Computing $S\vec{z}_v$	74
4.4.2	Preconditioner and its properties	81
4.4.3	Solving the preconditioned system	92
4.4.4	Cost of solving the Schur complement system	96
4.5	Cost of Solving the Galerkin Problem	97
5	NUMERICAL RESULTS	99
5.1	Gauss Quadrature	99
5.2	Numerical Experiments	103
6	CONCLUSION	111
6.1	Summary	111
6.2	Future Work	113

LIST OF FIGURES

1.1	Biharmonic Dirichlet Problem on Domain Ω	2
4.1	Related Biharmonic Problem I.	70
4.2	Related Biharmonic Problem II.	82

ACKNOWLEDGEMENTS

I owe my thesis adviser, Dr. Bernard Bialecki, profound expressions of regard and thanks in recognition of his steadfast guidance, patience, and professionalism. I appreciate the encouragement and advice of Dr. Graeme Fairweather, department head and committee member, which was freely offered and gratefully received. My thanks to my committee members, Dr. Steven Pruess, Dr. Erik Van Vleck, and Dr. Ilya Tsvankin, for their support and encouragement during this long process. Sincere regards are tendered to committee member Dr. Robert Woolsey, for his longstanding concern with my success in academia. I would be remiss if I did not also recognize the friendship and goodwill unfailingly offered by the other faculty, and fellow graduate students, of the Math and Computer Sciences Department.

Of course, my love and gratitude go to the clan Knudson and its unwavering faith. My prime motivators, daughter Krista and son Gunnar, deserve special thanks for being good helpers during a difficult time. I thank my brother Steve, and sisters Kathy, Julie, and Amy, for their continued love and support. Finally, I dedicate this work to Ken and Rosemarie Knudson, family role models of cheerful perseverance.

Chapter 1

INTRODUCTION

In this dissertation, we consider the biharmonic Dirichlet problem:

$$\begin{aligned}\Delta^2 u &= f(x, y) \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega,\end{aligned}\tag{1.1}$$

where $\Omega = (0, 1) \times (0, 1)$, $\partial\Omega$ is the boundary of Ω , and $\partial/\partial n$ is the outer normal derivative on $\partial\Omega$ (cf. Figure 1.1). This problem is important in fluid dynamics where u and Δu represent stream function and vorticity, respectively (cf. Section 10.4 of [19]). It is also fundamental to the aviation industry (cf. Section 2.1 of [21]), where u and Δu represent vertical displacement of a plate and bending moment. Problem (1.1) and the use of mixed finite element Galerkin methods to solve it have

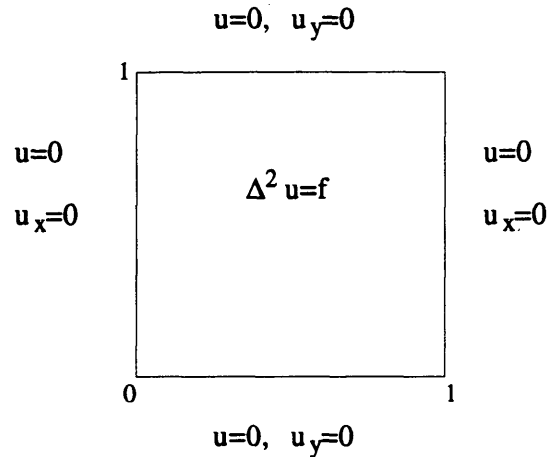


Figure 1.1: Biharmonic Dirichlet Problem on Domain Ω .

been examined extensively since Ciarlet and Raviart [8] first introduced the approach which has come to be called the Ciarlet-Raviart method. In this method the auxiliary function $v = \Delta u$ is introduced, allowing the fourth order problem to be reduced to two second order problems. It is then possible, in particular for irregular Ω , to use finite elements of class C^0 instead of elements of class C^1 which are required for fourth order problems. Ciarlet and Glowinski [7] note that another merit of the method is that it gives a continuous approximation to not only u but also to Δu . Brezzi and Raviart [4] improved convergence results of [8], and Falk and Osborn [11] derive abstract error estimates which are applicable to the Ciarlet-Raviart method. Finally,

Brezzi and Fortin (Section 4.4 of [5]) give a recent summary of finite element methods and their convergence analysis for solving (1.1).

The antecedents for this dissertation consist of three strands of work which we combine into an accurate and efficient piecewise Hermite bicubic solution of (1.1). The first strand is based on reducing the Galerkin problem for (1.1) to a linear system involving either approximation to u in Ω or approximation to Δu on $\partial\Omega$. Braess and Peisker [3], Gustafsson [14], and Hanisch [15] developed approaches seeking an approximation to u in Ω . Glowinski and Pironneau [12] and Peisker [18] reduced the Galerkin problem to finding an approximation to Δu on $\partial\Omega$. We go a step further in seeking an approximation to Δu on the two vertical sides of $\partial\Omega$.

The second strand of our work involves solving the auxiliary Galerkin problem for a related biharmonic problem with Δu instead of $\partial u/\partial n$ specified on the two vertical sides of $\partial\Omega$. The most relevant contribution in this direction is due to Bjørstad [2] who used separation of variables and FFTs to solve the corresponding finite difference problem with cost $O(N^2 \log_2 N)$ on an $N \times N$ partition.

In order to use separation of variables in our approach we need the third strand of work, Lyashko and Soloviev's method [17] for finding the explicit solution of the generalized eigenvalue problem that arises in the piecewise Hermite cubic Galerkin approximation of the continuous eigenvalue problem for the second order derivative. Since the resulting eigenvector matrix is composed of sines and cosines, we can also

use FFTs for matrix-vector multiplications.

Our method of finding the piecewise Hermite bicubic approximation to (1.1) is based on using a Schur complement approach to reduce the Galerkin problem to a Schur complement system and to an auxiliary Galerkin problem. The symmetric, positive definite Schur complement system involves approximation to Δu on the two vertical sides of $\partial\Omega$ and is solved using the PCG method. We do not form the Schur complement matrix explicitly since its special structure is used to compute the required matrix-vector product with cost $O(N^2)$. We use the Galerkin problem for a related biharmonic problem with Δu instead of $\partial u/\partial n$ specified on the two horizontal sides of $\partial\Omega$ to obtain a preconditioner. We show this preconditioner to be symmetric and positive definite, and conjecture spectral equivalence to the Schur complement matrix. The cost of solving a preconditioned system is $O(N^2)$. The number of the PCG iterations is proportional to $\log_2 N$ and hence total cost of solving the Schur complement system is $O(N^2 \log_2 N)$. The auxiliary Galerkin problem is related to a biharmonic problem with Δu instead of $\partial u/\partial n$ specified on the two vertical sides of $\partial\Omega$. Using the solution of the generalized eigenvalue problem we diagonalize the auxiliary Galerkin problem in one direction and use FFTs to solve it with cost $O(N^2 \log_2 N)$. Therefore the cost of solving the Galerkin problem for (1.1) is $O(N^2 \log_2 N)$.

We note that this approach combines both efficiency and accuracy in the same

algorithm. Computational costs are as low as those for Bjørstad's [2] algorithm, or Peisker's [18] efficient linear finite element algorithm, which also uses Schur complement approach and PCG method. But our approach is more accurate than the second order approximations found in those methods. Cooper and Prenter [9], Sun [20], and Lou et al [16], developed different piecewise Hermite bicubic orthogonal spline collocation approaches with fourth and third order accuracy in the L^2 and H^1 norms, respectively. However, the most efficient of these methods, the capacitance matrix method of Lou et al [16], still requires $O(N^3)$ operations.

An outline of this dissertation is as follows. Chapter 2 introduces partitioning, basis functions, and mass and stiffness matrices required for deriving the Galerkin problem for (1.1). It also gives the solution of the generalized eigenvalue problem for later use in separation of variables. In Chapter 3, we define the piecewise Hermite bicubic mixed finite element Galerkin problem for (1.1) and derive its matrix-vector form. In Chapter 4, we formulate the algorithm for solving the Galerkin problem by reducing it to the Schur complement system and an auxiliary Galerkin problem. We verify necessary properties of the Schur complement matrix and its preconditioner, and finally give computational costs associated with solving the Galerkin problem. In Chapter 5, we explain methodology for the numerical integration of the right-hand side functions required in the Galerkin method. We close with numerical results indicating optimal fourth and third order convergence of the L^2 and H^1 norm errors

in the approximations to u and Δu . In the discrete maximum norm, we observe fourth order convergence for the approximations to u and Δu , and third order convergence of errors in approximations to first order derivatives.

Chapter 2

PRELIMINARIES

2.1 Partition, Basis Functions, Stiffness and Mass Matrices

Let N be a positive integer and $\{x_i\}_{i=0}^N, \{y_j\}_{j=0}^N$ be uniform partitions of $[0, 1]$ such that $x_i = ih, y_j = jh, i, j = 0, \dots, N$, where $h = 1/N$ is the step size. Let \mathcal{M}_h be the space of piecewise Hermite cubics on $[0, 1]$ such that

$$\mathcal{M}_h = \{v \in C^1[0, 1] : v|_{[x_{i-1}, x_i]} \in P_3, i = 1, \dots, N\}, \quad (2.1)$$

where P_3 denotes the set of polynomials of degree ≤ 3 , and let

$$\mathcal{M}_h^0 = \{v \in \mathcal{M}_h : v(0) = v(1) = 0\}. \quad (2.2)$$

For each $x_i, i = 0, \dots, N$, consider the “value” and scaled “slope” basis func-

tions $v_i, s_i \in \mathcal{M}_h$, respectively, such that

$$\begin{aligned} v_i(x_j) &= \delta_{ij}, & v_i'(x_j) &= 0, \\ s_i(x_j) &= 0, & s_i'(x_j) &= h^{-1}\delta_{ij}, \end{aligned} \quad i, j = 0, \dots, N,$$

where δ_{ij} is the Kronecker delta.

Introduce $g_1(x) = -2x^3 + 3x^2$, $g_2(x) = x^3 - x^2$. Then (cf. Section 2.3 of [10])

$$v_i(x) = \begin{cases} g_1((x - x_{i-1})/h), & x \in [x_{i-1}, x_i], \\ g_1(1 - (x - x_i)/h), & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise,} \end{cases} \quad (2.3)$$

and

$$s_i(x) = \begin{cases} g_2((x - x_{i-1})/h), & x \in [x_{i-1}, x_i], \\ -g_2(1 - (x - x_i)/h), & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise,} \end{cases} \quad (2.4)$$

for $i = 1, \dots, N - 1$, with obvious modifications for $i = 0, N$.

With different orderings of v_i and s_i , we obtain two bases, $\{\phi_n\}_{n=0}^{2N+1}$ and $\{\psi_n\}_{n=0}^{2N+1}$ for \mathcal{M}_h , such that

$$\{\phi_0, \dots, \phi_{2N+1}\} = \{v_0, v_1, \dots, v_{N-2}, v_{N-1}, s_0, s_1, \dots, s_{N-2}, s_{N-1}, s_N, v_N\}, \quad (2.5)$$

$$\{\psi_0, \dots, \psi_{2N+1}\} = \{v_0, s_0, v_1, s_1, \dots, v_{N-1}, s_{N-1}, s_N, v_N\}. \quad (2.6)$$

By removing the first and last basis functions from (2.5) and (2.6), we obtain two bases for \mathcal{M}_h^0 :

$$\{\phi_1, \dots, \phi_{2N}\} = \{v_1, \dots, v_{N-2}, v_{N-1}, s_0, s_1, \dots, s_{N-2}, s_{N-1}, s_N\}, \quad (2.7)$$

and

$$\{\psi_1, \dots, \psi_{2N}\} = \{s_0, v_1, s_1, \dots, v_{N-1}, s_{N-1}, s_N\}. \quad (2.8)$$

Let (\cdot, \cdot) denote the usual L^2 inner product on the interval $(0, 1)$, that is,

$$(f, g) = \int_0^1 f(x)g(x)dx.$$

For later reference it will be important to note the following formulas for v_i of (2.3) and s_i of (2.4) (cf. Section 2.3 of [10]):

$$\begin{aligned} (v'_i, v'_i) &= h^{-1}\alpha_1, & (v'_i, s'_i) &= 0, & (v'_i, v'_{i+1}) &= h^{-1}\alpha_3, & (v'_i, s'_{i+1}) &= h^{-1}\alpha_5, \\ (s'_i, s'_i) &= h^{-1}\alpha_2, & (s'_i, v'_{i+1}) &= -h^{-1}\alpha_5, & (s'_i, s'_{i+1}) &= h^{-1}\alpha_4, \\ (v'_0, v'_0) &= h^{-1}\alpha_1/2, & (v'_0, s'_0) &= h^{-1}\alpha_5, & (v'_N, v'_N) &= h^{-1}\alpha_1/2, & (v'_N, s'_N) &= -h^{-1}\alpha_5, \\ (s'_0, s'_0) &= h^{-1}\alpha_2/2, & & & & & (s'_N, s'_N) &= h^{-1}\alpha_2/2, \end{aligned} \quad (2.9)$$

$$\begin{aligned}
(v_i, v_i) &= h\beta_1, & (v_i, s_i) &= 0, & (v_i, v_{i+1}) &= h\beta_3, & (v_i, s_{i+1}) &= h\beta_5, \\
(s_i, s_i) &= h\beta_2, & (s_i, v_{i+1}) &= -h\beta_5, & (s_i, s_{i+1}) &= h\beta_4, \\
(v_0, v_0) &= h\beta_1/2, & (v_0, s_0) &= h\beta_6, & (v_N, v_N) &= h\beta_1/2, & (v_N, s_N) &= -h\beta_6, \\
(s_0, s_0) &= h\beta_2/2, & & & (s_N, s_N) &= h\beta_2/2,
\end{aligned} \tag{2.10}$$

where

$$\begin{aligned}
\alpha_1 &= 12/5, & \alpha_2 &= 4/15, & \alpha_3 &= -6/5, & \alpha_4 &= -1/30, & \alpha_5 &= 1/10, \\
\beta_1 &= 26/35, & \beta_2 &= 2/105, & \beta_3 &= 9/70, & \beta_4 &= -1/140, & \beta_5 &= -13/420, \\
\beta_6 &= 11/210.
\end{aligned} \tag{2.11}$$

Now we introduce stiffness and mass matrices which will be used throughout this paper. Let us introduce the $2N \times 2N$ matrices A_x and B_x given by

$$A_x = (a_{i,k}^x)_{i=1,k=1}^{2N,2N}, \quad a_{i,k}^x = \int_0^1 (\phi_i' \phi_k')(x) dx, \tag{2.12}$$

$$B_x = (b_{i,k}^x)_{i=1,k=1}^{2N,2N}, \quad b_{i,k}^x = \int_0^1 (\phi_i \phi_k)(x) dx, \tag{2.13}$$

where $\{\phi_i\}_{i=1}^{2N}$ is defined in (2.7). To present the structure of A_x and B_x , we introduce

for any γ_i , $i = 1, \dots, 5$, the tridiagonal matrices

$$T_\gamma = \begin{bmatrix} \gamma_1 & \gamma_3 & & & \\ \gamma_3 & \gamma_1 & \gamma_3 & & \\ & & & & \\ & & \gamma_3 & \gamma_1 & \gamma_3 \\ & & & \gamma_3 & \gamma_1 \end{bmatrix}, \quad \tilde{R}_\gamma = \begin{bmatrix} \gamma_2/2 & \gamma_4 & & & \\ \gamma_4 & \gamma_2 & \gamma_4 & & \\ & & & & \\ & & \gamma_4 & \gamma_2 & \gamma_4 \\ & & & \gamma_4 & \gamma_2/2 \end{bmatrix}, \quad (2.14)$$

which are of sizes $N-1$ and $N+1$, respectively, and the $(N-1) \times (N+1)$ rectangular matrix S_γ given by

$$S_\gamma = \begin{bmatrix} -\gamma_5 & 0 & \gamma_5 & & & \\ & -\gamma_5 & 0 & \gamma_5 & & \\ & & & & & \\ & & & -\gamma_5 & 0 & \gamma_5 \\ & & & & -\gamma_5 & 0 & \gamma_5 \end{bmatrix}. \quad (2.15)$$

(Notation T_γ , \tilde{R}_γ , and S_γ was introduced in [17] and will be used in Section 2.3.)

Then using (2.7), (2.9), and (2.10) it is easy to see that

$$A_x = h^{-1} \begin{bmatrix} T_\alpha & S_\alpha \\ S_\alpha^T & \tilde{R}_\alpha \end{bmatrix}, \quad B_x = h \begin{bmatrix} T_\beta & S_\beta \\ S_\beta^T & \tilde{R}_\beta \end{bmatrix}, \quad (2.16)$$

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,M_2}B \\ \cdots & a_{i,j}B & \cdots \\ a_{M_1,1}B & \cdots & a_{M_1,M_2}B \end{bmatrix}. \quad (2.22)$$

Properties of matrix tensor products include:

$$(A \otimes B)^T = A^T \otimes B^T, \quad (2.23)$$

and provided matrix products are defined,

$$(A_1 \otimes A_2)(B_1 \otimes B_2) = (A_1 B_1) \otimes (A_2 B_2). \quad (2.24)$$

Consider evaluating the numbers

$$z_{i,j} = \sum_{k \in K} a_{i,k} \sum_{l \in L} b_{j,l} v_{k,l}, \quad i \in I, j \in J, \quad (2.25)$$

where $\{v_{k,l}\}_{k \in K, l \in L}$ are given. If we introduce vectors

$$\vec{z} = [z_{1,1}, \dots, z_{1,N_1}, \dots, z_{M_1,1}, \dots, z_{M_1,N_1}]^T, \quad \vec{v} = [v_{1,1}, \dots, v_{1,N_2}, \dots, v_{M_2,1}, \dots, v_{M_2,N_2}]^T,$$

and

$$\vec{v}_k = [v_{k,1}, \dots, v_{k,N_2}]^T, \quad k \in K,$$

then the matrix-vector form of (2.25) is

$$\begin{aligned} \vec{z} &= \begin{bmatrix} a_{1,1}B\vec{v}_1 + & \cdots & +a_{1,M_2}B\vec{v}_{M_2} \\ a_{i,1}B\vec{v}_1 + & \cdots & +a_{i,M_2}B\vec{v}_{M_2} \\ a_{M_1,1}B\vec{v}_1 + & \cdots & +a_{M_1,M_2}B\vec{v}_{M_2} \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}I_{N_1} & \cdots & a_{1,M_2}I_{N_1} \\ a_{i,1}I_{N_1} & \cdots & a_{i,M_2}I_{N_1} \\ a_{M_1,1}I_{N_1} & \cdots & a_{M_1,M_2}I_{N_1} \end{bmatrix} \begin{bmatrix} B\vec{v}_1 \\ B\vec{v}_2 \\ \vdots \\ B\vec{v}_{M_2} \end{bmatrix} \\ &= (A \otimes I_{N_1})(I_{M_2} \otimes B)\vec{v} = (A \otimes B)\vec{v}, \end{aligned} \tag{2.26}$$

where A and B are as in (2.21), and where here and in what follows I_p denotes the $p \times p$ identity matrix.

Second, when referring to tensor product notation of function spaces W_1 and W_2 , we use $W_1 \otimes W_2$ to denote the space of functions consisting of all finite linear combinations of products of the form fg , where $f \in W_1$ and $g \in W_2$.

2.3 Generalized Eigenvalue Problem

For A_x and B_x of (2.12) and (2.13), we derive matrices Λ and Z , with Λ diagonal, such that

$$A_x Z = B_x Z \Lambda, \quad (2.27)$$

and

$$Z^T B_x Z = I_{2N}. \quad (2.28)$$

To generate Λ and Z , consider the generalized eigenvalue problem

$$A_x \vec{z} = \lambda B_x \vec{z}, \quad (2.29)$$

which by (2.16) is equivalent to

$$\begin{bmatrix} T_\alpha & S_\alpha \\ S_\alpha^T & \tilde{R}_\alpha \end{bmatrix} \vec{z} = \lambda h^2 \begin{bmatrix} T_\beta & S_\beta \\ S_\beta^T & \tilde{R}_\beta \end{bmatrix} \vec{z}. \quad (2.30)$$

A method of solution of the generalized eigenvalue problem (2.30) with T_α , T_β , \tilde{R}_α , and \tilde{R}_β of the form (2.14) and S_α and S_β of the form (2.15) is demonstrated in [17]. In particular it follows from Lemma 2.2 of [17] that the eigenvalues $\{\lambda_k^\pm\}_{k=1}^N$ of (2.29) are given by

$$\lambda_N^\pm = h^{-2}(\alpha_2 \mp 2\alpha_4)/(\beta_2 \mp 2\beta_4), \quad (2.31)$$

$$\lambda_k^\pm = h^{-2} \Phi^\pm(\alpha_k, \nu_k), \quad k = 1, \dots, N-1, \quad (2.32)$$

where

$$\Phi^\pm(\alpha, \nu) = (-b \pm \sqrt{b^2 - 4ac})/2a, \quad (2.33)$$

$$\begin{aligned} a &= b_1 b_2 - b_3^2, & b &= 2a_3 b_3 - a_1 b_2 - b_1 a_2, & c &= a_1 a_2 - a_3^2, \\ a_1 &= 2\alpha_3 \alpha + \alpha_1, & a_2 &= 2\alpha_4 \alpha + \alpha_2, & a_3 &= 2\alpha_5 \nu, \\ b_1 &= 2\beta_3 \alpha + \beta_1, & b_2 &= 2\beta_4 \alpha + \beta_2, & b_3 &= 2\beta_5 \nu, \end{aligned} \quad (2.34)$$

$$\alpha_k = \cos \frac{k\pi}{N}, \quad \nu_k = \sin \frac{k\pi}{N}, \quad (2.35)$$

and α_i, β_i are defined in (2.11). In comparison with [17], for convenience we switched signs + and - on the right-hand side in (2.31).

Lemma 2.1 *The eigenvalues $\{\lambda_k^\pm\}_{k=1}^N$ given in (2.31)–(2.32) are distinct and positive.*

Proof. For convenience, we express $\Phi^\pm(\alpha, \nu)$ of (2.33) as a function of one variable α , assuming that $\nu = \sqrt{1 - \alpha^2}$ (cf. (2.35)). Using ©Mathematica, we have

$$\Phi^\pm(\alpha, \nu) = \Phi^\pm(\alpha, \sqrt{1 - \alpha^2}) = \Psi^\pm(\alpha) \equiv 6 \frac{p(\alpha) \pm q(\alpha)}{r(\alpha)}, \quad (2.36)$$

where

$$p(\alpha) = 141 - 32\alpha - 4\alpha^2,$$

$$q(\varkappa) = \sqrt{13056 + 3856\varkappa - 7524\varkappa^2 + 1656\varkappa^3 - 19\varkappa^4},$$

and

$$r(\varkappa) = \varkappa^2 - 36\varkappa + 65.$$

We will consider $[\Psi^\pm]'(\varkappa)$ for $-1 < \varkappa < 1$. Using the quotient rule, we can write

$$[\Psi^\pm]'(\varkappa) = 6 \frac{r(\varkappa)[p'(\varkappa) \pm q'(\varkappa)] - [p(\varkappa) \pm q(\varkappa)]r'(\varkappa)}{r^2(\varkappa)}.$$

Using the definitions of $p(\varkappa)$, $q(\varkappa)$, $r(\varkappa)$, and $\text{\textcircled{C}}\text{Mathematica}$, we obtain

$$[\Psi^\pm]'(\varkappa) = \frac{\pm f(\varkappa) + g(\varkappa)q(\varkappa)}{r^2(\varkappa)q(\varkappa)/12}, \quad (2.37)$$

where

$$f(\varkappa) = 297668 - 222882\varkappa + 77838\varkappa^2 - 12377\varkappa^3 - 72\varkappa^4,$$

and

$$g(\varkappa) = 1498 - 401\varkappa + 88\varkappa^2.$$

We want to examine the sign of (2.37) for $-1 < \varkappa < 1$.

For Ψ^+ , the numerator of (2.37) is given by

$$-f(\varkappa) + g(\varkappa)q(\varkappa).$$

For $-1 < \varkappa < 0$,

$$-f'(\varkappa) = 222882 - 155676\varkappa + 37131\varkappa^2 + 288\varkappa^3 > 0.$$

Thus $-f(\varkappa)$ is increasing on $-1 < \varkappa < 0$ and hence

$$-610693 = -f(-1) < -f(\varkappa) < -f(0) = -297668, \quad -1 < \varkappa < 0.$$

For $-1 < \varkappa < 0$,

$$g'(\varkappa) = 176\varkappa - 401 < 0.$$

Thus $g(\varkappa)$ is decreasing on $-1 < \varkappa < 0$ and hence

$$1498 = g(0) < g(\varkappa) < g(-1) = 1987, \quad -1 < \varkappa < 0.$$

For $-1 < \varkappa < 0$,

$$q'(\varkappa) = \frac{3856 - 15048\varkappa + 4968\varkappa^2 - 76\varkappa^3}{2q(\varkappa)} > 0.$$

Thus $q(\varkappa)$ is increasing on $-1 < \varkappa < 0$ and hence

$$1 = q(-1) < q(\varkappa) < q(0) = 16\sqrt{51}, \quad -1 < \varkappa < 0.$$

Bounding $g(\varkappa)q(\varkappa)$ on $-1 < \varkappa < 0$, we obtain

$$1498 < g(\varkappa)q(\varkappa) < 31792\sqrt{51}, \quad -1 < \varkappa < 0.$$

Hence for $-1 < \varkappa < 0$,

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < -297668 + 31792\sqrt{51}.$$

Since

$$31792\sqrt{51} < 31792 \times 8 = 254336,$$

we have

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < 0, \quad -1 < \varkappa < 0. \quad (2.38)$$

For the interval $0 \leq \varkappa < 1$

$$-f'(\varkappa) = 222882 - 155676\varkappa + 37131\varkappa^2 + 288\varkappa^3 > 0.$$

Hence $-f(\varkappa)$ is increasing on $0 \leq \varkappa < 1$. Using the definitions of $g(\varkappa)$ and $q(\varkappa)$, and

©Mathematica, we have

$$\begin{aligned} & [g(\varkappa)q(\varkappa)]' \\ &= \frac{-2347312 - 11292480\varkappa + 10603600\varkappa^2 - 3703400\varkappa^3 + 532905\varkappa^4 - 6688\varkappa^5}{q(\varkappa)} < 0 \end{aligned}$$

for $0 \leq \varkappa < 1$. Therefore $g(\varkappa)q(\varkappa)$ is decreasing on $0 < \varkappa < 1$.

Since $-f(\varkappa)$ is increasing on $0 \leq \varkappa \leq 1/2$, we have

$$-297668 = -f(0) < -f(\varkappa) < -f(1/2) = -\frac{1633079}{8}, \quad 0 \leq \varkappa \leq 1/2.$$

Since $g(\varkappa)q(\varkappa)$ is decreasing on $0 \leq \varkappa \leq 1/2$, we have

$$\frac{2639\sqrt{212941}}{8} = g(1/2)q(1/2) < g(\varkappa)q(\varkappa) < g(0)q(0) = 23968\sqrt{51}, \quad 0 \leq \varkappa \leq 1/2.$$

Hence

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < -\frac{1633079}{8} + 23968\sqrt{51}, \quad 0 \leq \varkappa \leq 1/2.$$

Since

$$8 \times 23968\sqrt{51} < 8 \times 23968 \times 8 = 1533952,$$

we obtain

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < 0, \quad 0 \leq \varkappa \leq 1/2. \quad (2.39)$$

Further,

$$-\frac{1633079}{8} = -f(1/2) < -f(\varkappa) < -f(3/4) = -\frac{10818947}{64}, \quad 1/2 < \varkappa \leq 3/4,$$

and

$$\frac{14961\sqrt{352949}}{64} = g(3/4)q(3/4) < g(\varepsilon)q(\varepsilon) < g(1/2)q(1/2) = \frac{2639\sqrt{212941}}{8}$$

for $1/2 < \varepsilon \leq 3/4$. Hence

$$-f(\varepsilon) + g(\varepsilon)q(\varepsilon) < -\frac{10818947}{64} + \frac{2639\sqrt{212941}}{8}, \quad 1/2 < \varepsilon \leq 3/4.$$

Since

$$8 \times 2639\sqrt{212941} < 8 \times 2639 \times 500 = 10556000,$$

we have

$$-f(\varepsilon) + g(\varepsilon)q(\varepsilon) < 0, \quad 1/2 < \varepsilon \leq 3/4. \quad (2.40)$$

Also,

$$-\frac{10818947}{64} = -f(3/4) < -f(\varepsilon) < -f(1) = -140175, \quad 3/4 < \varepsilon < 1,$$

and

$$124425 = g(1)q(1) < g(\varepsilon)q(\varepsilon) < g(3/4)q(3/4) = \frac{14961\sqrt{352949}}{64}, \quad 3/4 < \varepsilon < 1.$$

Hence

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < -140175 + \frac{14961\sqrt{352949}}{64}, \quad 3/4 < \varkappa < 1.$$

Since

$$-140175 \times 64 = -8971200, \quad 14961\sqrt{352949} < 14961 \times 595 = 8901795,$$

we have

$$-f(\varkappa) + g(\varkappa)q(\varkappa) < 0, \quad 3/4 < \varkappa < 1. \quad (2.41)$$

The denominator in (2.37) is positive for $-1 < \varkappa < 1$ since $r(\varkappa) > 0$ and $q(\varkappa) > 0$ for $-1 < \varkappa < 1$. Thus it follows from (2.37) and (2.38)–(2.41) that

$$[\Psi^-]'(\varkappa) < 0, \quad -1 < \varkappa < 1.$$

For $f(\varkappa)$, the numerator of (2.37) becomes

$$f(\varkappa) + g(\varkappa)q(\varkappa).$$

Since $-f(\varkappa) < 0$ and $g(\varkappa)q(\varkappa) > 0$ for $-1 < \varkappa < 1$, by (2.37) we have

$$[\Psi^+]'\varkappa) > 0, \quad -1 < \varkappa < 1.$$

At this point we see that $[\Psi^-](\varkappa)$ is decreasing over $-1 < \varkappa < 1$, while the function $[\Psi^+](\varkappa)$ is increasing over the same interval. Hence using (2.36), we obtain

$$0 = [\Psi^-](1) < [\Psi^-](\varkappa) < [\Psi^-](-1) = 168/17, \quad -1 < \varkappa < 1,$$

$$10 = [\Psi^+](-1) < [\Psi^+](\varkappa) < [\Psi^+](1) = 42 \quad -1 < \varkappa < 1.$$

Using (2.32), (2.35), and (2.36), we can write

$$\lambda_k^\pm = h^{-2}\Phi^\pm(\varkappa_k, \nu_k) = h^{-2}\Psi^\pm(\varkappa_k), \quad k = 1, \dots, N-1,$$

and hence

$$0 < h^2\lambda_k^- < 168/17, \quad 10 < h^2\lambda_k^+ < 42, \quad k = 1, \dots, N-1. \quad (2.42)$$

Clearly, $\{\lambda_k^-\}_{k=1}^{N-1}$ are distinct since they are given in terms of the decreasing function $h^2[\Psi^-](\varkappa)$ on $-1 < \varkappa < 1$. In a similar way, $\{\lambda_k^+\}_{k=1}^{N-1}$ are distinct since they are given in terms of the increasing function $h^2[\Psi^+](\varkappa)$ on $-1 < \varkappa < 1$. Moreover $\{\lambda_k^-\}_{k=1}^{N-1}$ and $\{\lambda_k^+\}_{k=1}^{N-1}$ are two disjoint sets by (2.42).

Using (2.31) we can easily show that

$$h^2\lambda_N^\pm = (\alpha_2 \mp 2\alpha_4)/(\beta_2 \mp 2\beta_4) = \frac{\frac{4}{15} \pm \frac{1}{15}}{\frac{2}{105} \pm \frac{2}{140}} = \frac{56 \pm 14}{4 \pm 3},$$

and therefore

$$h^2 \lambda_N^+ = 10, \quad h^2 \lambda_N^- = 42. \quad (2.43)$$

By (2.42) these two eigenvalues are clearly different from $\{\lambda_k^\pm\}_{k=1}^{N-1}$. Therefore all eigenvalues are distinct.

All eigenvalues are positive by (2.42) and (2.43). \square

Although explicit formulas for the corresponding eigenvectors $\{z_k^\pm\}_{k=1}^N$ of (2.29) are not given in [17], they can be derived in the following manner. Introduce the vectors

$$\vec{y}^{(k)} = [y_0^{(k)}, \dots, y_N^{(k)}]^T, \quad k = 0, \dots, N, \quad (2.44)$$

$$\vec{x}^{(k)} = [x_1^{(k)}, \dots, x_{N-1}^{(k)}]^T, \quad k = 1, \dots, N-1, \quad (2.45)$$

with components

$$y_i^{(k)} = \alpha_{ik}, \quad i = 0, \dots, N, \quad x_i^{(k)} = \nu_{ik}, \quad i = 1, \dots, N-1, \quad (2.46)$$

where α_i and ν_i are given in (2.35). Let

$$\vec{z}_N^- = [\vec{0}, \vec{y}^{(0)}]^T, \quad \vec{z}_N^+ = [\vec{0}, \vec{y}^{(N)}]^T, \quad (2.47)$$

where $\vec{0}$ is the zero vector with $N - 1$ components and let

$$\vec{z}_k^\pm = [\vec{x}^{(k)}, c_k^\pm \vec{y}^{(k)}]^T, \quad k = 1, \dots, N - 1, \quad (2.48)$$

where the constants c_k^\pm , $k = 1, \dots, N - 1$, are to be specified. Let

$$\gamma_i = \alpha_i - \lambda h^2 \beta_i, \quad i = 1, \dots, 5, \quad (2.49)$$

where α_i and β_i are given in (2.11). It then follows from (2.14) and (2.15) that the vectors \vec{z}_k^\pm of (2.48) satisfy (2.30) if and only if

$$T_\gamma \vec{x}^{(k)} + c_k^\pm S_\gamma \vec{y}^{(k)} = \vec{0}, \quad (2.50)$$

and

$$S_\gamma^T \vec{x}^{(k)} + c_k^\pm \tilde{R}_\gamma \vec{y}^{(k)} = \vec{0}. \quad (2.51)$$

We require the following relationships from the proof of Lemma 2.1 in [17]:

$$S_\gamma \vec{y}^{(k)} = -2\gamma_5 \nu_k \vec{x}^{(k)}, \quad k = 0, \dots, N, \quad (2.52)$$

$$T_\gamma \vec{x}^{(k)} = (2\gamma_3 \varepsilon_k + \gamma_1) \vec{x}^{(k)}, \quad k = 1, \dots, N - 1, \quad (2.53)$$

$$\tilde{R}_\gamma \vec{y}^{(k)} = (2\gamma_4 \varepsilon_k + \gamma_2) \tilde{E} \vec{y}^{(k)}, \quad k = 0, \dots, N, \quad (2.54)$$

where the $N + 1$ dimension matrix \tilde{E} is given by

$$\tilde{E} = \begin{bmatrix} 1/2 & & & & \\ & 1 & & & \\ & \cdot & \cdot & \cdot & \\ & & & 1 & \\ & & & & 1/2 \end{bmatrix}.$$

Further, it is not difficult to verify that

$$S_\gamma^T \tilde{x}^{(k)} = -2\gamma_5 \nu_k \tilde{E} \tilde{y}^{(k)}, \quad k = 1, \dots, N - 1. \quad (2.55)$$

Then from (2.50)–(2.55) we obtain

$$(2\gamma_3 \alpha_k + \gamma_1) \tilde{x}^{(k)} - c_k^\pm 2\gamma_5 \nu_k \tilde{x}^{(k)} = \vec{0}, \quad (2.56)$$

and

$$-2\gamma_5 \nu_k \tilde{E} \tilde{y}^{(k)} + c_k^\pm (2\gamma_4 \alpha_k + \gamma_2) \tilde{E} \tilde{y}^{(k)} = \vec{0}, \quad (2.57)$$

for $k = 1, \dots, N - 1$. Since, by (2.44)–(2.46) and (2.35), we have

$$\tilde{E} \tilde{y}^{(k)} \neq \vec{0}, \quad \tilde{x}^{(k)} \neq \vec{0}, \quad k = 1, \dots, N - 1,$$

(2.56)–(2.57) are equivalent to

$$(2\gamma_3\alpha_k + \gamma_1) - c_k^\pm 2\gamma_5\nu_k = 0, \quad (2.58)$$

and

$$-2\gamma_5\nu_k + c_k^\pm(2\gamma_4\alpha_k + \gamma_2) = 0, \quad (2.59)$$

for $k = 1, \dots, N - 1$. Solving (2.58) for c_k^\pm we obtain

$$c_k^\pm = \frac{2\gamma_3\alpha_k + \gamma_1}{2\gamma_5\nu_k}, \quad k = 1, \dots, N - 1, \quad (2.60)$$

where we assume the denominator of (2.60) to be non-zero. We substitute (2.60) into (2.59) to obtain

$$(2\gamma_3\alpha_k + \gamma_1)(2\gamma_4\alpha_k + \gamma_2) - 4\gamma_5^2\nu_k^2 = 0, \quad k = 1, \dots, N - 1. \quad (2.61)$$

Simple substitution of (2.49) into (2.61) yields

$$a(\lambda h^2)^2 + b\lambda h^2 + c = 0, \quad k = 1, \dots, N - 1,$$

where a , b , and c , are those of (2.34) with α and ν replaced, respectively, with α_k and ν_k of (2.35). Clearly $\{\lambda_k^\pm\}_{k=1}^{N-1}$ of (2.32) are the solutions of this quadratic equation.

Note that the denominator of (2.60) is non-zero for $\lambda = \lambda_k^\pm$, since by (2.49) and (2.11)

$$\gamma_5 \nu_k = (1/10 + \lambda_k^\pm h^2 13/420) \nu_k, \quad k = 1, \dots, N-1,$$

which is positive by (2.42) and (2.35). This shows that $\{z_k^\pm\}_{k=1}^{N-1}$ of (2.48) with c_k^\pm of (2.60) are eigenvectors of (2.29) corresponding to the eigenvalues $\{\lambda_k^\pm\}_{k=1}^{N-1}$ of (2.32).

For \bar{z}_N of (2.47), (2.30) becomes

$$\begin{bmatrix} T_\alpha & S_\alpha \\ S_\alpha^T & \tilde{R}_\alpha \end{bmatrix} \begin{bmatrix} \vec{0} \\ \vec{y}^{(0)} \end{bmatrix} = \lambda h^2 \begin{bmatrix} T_\beta & S_\beta \\ S_\beta^T & \tilde{R}_\beta \end{bmatrix} \begin{bmatrix} \vec{0} \\ \vec{y}^{(0)} \end{bmatrix},$$

which, by (2.15), (2.14), and (2.49), is equivalent to

$$S_\gamma \vec{y}^{(0)} = \vec{0}, \quad \tilde{R}_\gamma \vec{y}^{(0)} = \vec{0}. \quad (2.62)$$

The first equation in (2.62) is satisfied by (2.52) and (2.35), with $k = 0$. By (2.54) and (2.35) with $k = 0$,

$$\tilde{R}_\gamma \vec{y}^{(0)} = (2\gamma_4 + \gamma_2) \tilde{E} \vec{y}^{(0)}.$$

Hence the second equation of (2.62) is equivalent to $2\gamma_4 + \gamma_2 = 0$, the solution of which gives λ_N^- of (2.31). This shows that \bar{z}_N of (2.47) is an eigenvector of (2.29) corresponding to the eigenvalue λ_N^- of (2.31). In the same way it is easy to verify that \bar{z}_N^+ of (2.47) is an eigenvector of (2.29) corresponding to λ_N^+ of (2.31).

An additional indexing of eigenvalues $\{\lambda_k^\pm\}_{k=1}^N$ and eigenvectors $\{\vec{z}_k^\pm\}_{k=1}^N$ of (2.29) will be useful in discussing orthogonality of eigenvectors. Let $\{\lambda_k\}_{k=1}^{2N}$ and $\{\vec{z}_k\}_{k=1}^{2N}$ be defined as

$$\lambda_k = \lambda_k^-, \quad \lambda_{N+k} = \lambda_k^+, \quad (2.63)$$

$$\vec{z}_k = \vec{z}_k^-, \quad \vec{z}_{N+k} = \vec{z}_k^+,$$

where $k = 1, \dots, N$.

Since B_x is symmetric and positive definite, there exists $B_x^{\frac{1}{2}}$ such that $B_x = B_x^{\frac{1}{2}} B_x^{\frac{1}{2}}$. Therefore

$$A_x \vec{z}_k = \lambda_k B_x \vec{z}_k, \quad k = 1, \dots, 2N,$$

can be written as

$$S \vec{w}_k = \lambda_k \vec{w}_k, \quad k = 1, \dots, 2N,$$

where $\vec{w}_k = B_x^{\frac{1}{2}} \vec{z}_k$ and $S = B_x^{-\frac{1}{2}} A_x B_x^{-\frac{1}{2}}$.

Since S is symmetric and $\{\lambda_k\}_{k=1}^{2N}$ are distinct by Lemma 2.1, we have

$$(B_x \vec{z}_k, \vec{z}_l)_{\mathbb{R}^{2N}} = (B_x^{\frac{1}{2}} \vec{z}_k, B_x^{\frac{1}{2}} \vec{z}_l)_{\mathbb{R}^{2N}} = (\vec{w}_k, \vec{w}_l)_{\mathbb{R}^{2N}} = 0, \quad k \neq l, \quad (2.64)$$

where here and throughout $(\cdot, \cdot)_{\mathbb{R}^M}$ denotes the standard inner product in \mathbb{R}^M .

To satisfy relationship (2.28), we must select $\{d_k^\pm\}_{k=1}^N$ such that

$$(B_x d_k^\pm \bar{z}_k^\pm, d_k^\pm \bar{z}_k^\pm)_{\mathbb{R}^{2N}} = 1, \quad k = 1, \dots, N. \quad (2.65)$$

For $k = 1, \dots, N - 1$, using (2.16), (2.48), and (2.52)–(2.55) with β in place of γ , we obtain

$$h^{-1} B_x \bar{z}_k^\pm = \begin{bmatrix} T_\beta \bar{x}^{(k)} + c_k^\pm S_\beta \bar{y}^{(k)} \\ S_\beta^T \bar{x}^{(k)} + c_k^\pm \tilde{R}_\beta \bar{y}^{(k)} \end{bmatrix} = \begin{bmatrix} (2\beta_3 \alpha_k + \beta_1 - c_k^\pm 2\beta_5 \nu_k) \bar{x}^{(k)} \\ (c_k^\pm [2\beta_4 \alpha_k + \beta_2] - 2\beta_5 \nu_k) \tilde{E} \bar{y}^{(k)} \end{bmatrix}.$$

Hence using again (2.48), (2.44)–(2.46), and the identity

$$\sum_{i=1}^{N-1} \sin^2 \left(\frac{ik\pi}{N} \right) = \frac{N}{2}, \quad k = 1, \dots, N - 1,$$

we have

$$\begin{aligned} (h^{-1} B_x \bar{z}_k^\pm, \bar{z}_k^\pm)_{\mathbb{R}^{2N}} &= (2\beta_3 \alpha_k + \beta_1 - c_k^\pm 2\beta_5 \nu_k) \sum_{i=1}^{N-1} \sin^2 \left(\frac{ik\pi}{N} \right) \\ &\quad + [c_k^\pm (2\beta_4 \alpha_k + \beta_2) - 2\beta_5 \nu_k] c_k^\pm \left(1 + \sum_{i=1}^{N-1} \cos^2 \left(\frac{ik\pi}{N} \right) \right) \\ &= \frac{N}{2} \left([(2\beta_4 \alpha_k + \beta_2) c_k^\pm - 4\beta_5 \nu_k] c_k^\pm + 2\beta_3 \alpha_k + \beta_1 \right), \quad k = 1, \dots, N - 1. \end{aligned} \quad (2.66)$$

Using (2.16), (2.47), (2.52), (2.54) with $k = 0, N$, and β in place of γ , we obtain

$$h^{-1}B_x\bar{z}_N^- = \begin{bmatrix} S_\beta\bar{y}^{(0)} \\ \tilde{R}_\beta\bar{y}^{(0)} \end{bmatrix} = \begin{bmatrix} \bar{0} \\ (2\beta_4 + \beta_2)\tilde{E}\bar{y}^{(0)} \end{bmatrix},$$

and

$$h^{-1}B_x\bar{z}_N^+ = \begin{bmatrix} S_\beta\bar{y}^{(N)} \\ \tilde{R}_\beta\bar{y}^{(N)} \end{bmatrix} = \begin{bmatrix} \bar{0} \\ (\beta_2 - 2\beta_4)\tilde{E}\bar{y}^{(N)} \end{bmatrix}.$$

Since by (2.44), (2.46), and (2.35), $\bar{y}^0 = [1, 1, \dots, 1]^T$ and $\bar{y}^N = [1, -1, \dots, \pm 1]^T$ with $+1$ for even N and -1 for odd N , respectively, using (2.47) we have

$$(h^{-1}B_x\bar{z}_N^-, \bar{z}_N^-)_{\mathbb{R}^{2N}} = (2\beta_4 + \beta_2)N, \quad (h^{-1}B_x\bar{z}_N^+, \bar{z}_N^+)_{\mathbb{R}^{2N}} = (\beta_2 - 2\beta_4)N. \quad (2.67)$$

Clearly, it follows from (2.66) and (2.67) that to ensure (2.65) we take

$$d_k^\pm = 1/\sqrt{h\frac{N}{2} \left([(2\beta_4\alpha_k + \beta_2)c_k^\pm - 4\beta_5\nu_k]c_k^\pm + 2\beta_3\alpha_k + \beta_1 \right)}, \quad k = 1, \dots, N-1, \quad (2.68)$$

and

$$d_N^- = 1/\sqrt{hN(2\beta_4 + \beta_2)}, \quad d_N^+ = 1/\sqrt{hN(\beta_2 - 2\beta_4)}. \quad (2.69)$$

Now let

$$\Lambda = \text{diag}(\lambda_1^-, \dots, \lambda_N^-, \lambda_1^+, \dots, \lambda_N^+) \quad (2.70)$$

and

$$Z = [d_1^- z_1^-, \dots, d_N^- z_N^-, d_1^+ z_1^+, \dots, d_N^+ z_N^+]^T, \quad (2.71)$$

where $\{\lambda_k^\pm\}_{k=1}^N$ and $\{z_k^\pm\}_{k=1}^N$ are given by (2.31)–(2.32) and (2.47)–(2.48), $\{c_k^\pm\}_{k=1}^{N-1}$ are given in (2.60), and $\{d_k^\pm\}_{k=1}^N$ in (2.68)–(2.69). Then it follows from (2.29), (2.64) and (2.65) that (2.27) and (2.28) are satisfied.

Introducing matrices

$$S = \left(\sin \frac{ik\pi}{N} \right)_{i,k=1}^{N-1}, \quad C = \left(\cos \frac{ik\pi}{N} \right)_{i=0,k=1}^{N,N-1}, \quad \tilde{C} = \left(\cos \frac{ik\pi}{N} \right)_{i,k=0}^N,$$

and

$$\Lambda_d^\pm = \text{diag}(d_1^\pm, \dots, d_{N-1}^\pm), \quad \tilde{\Lambda}_d^+ = \text{diag}(d_N^-, d_1^+, \dots, d_{N-1}^+, d_N^+),$$

$$\Lambda_c^- = \text{diag}(c_1^-, \dots, c_{N-1}^-), \quad \Lambda_c^+ = \text{diag}(1, c_1^+, \dots, c_{N-1}^+, 1),$$

and using (2.47), (2.48), and (2.44)–(2.46), we can rewrite Z of (2.71) as

$$Z = \left[\begin{array}{c|c|c|c} S\Lambda_d^- & \vec{0} & S\Lambda_d^+ & \vec{0} \\ \hline C\Lambda_d^- \Lambda_c^- & & \tilde{C}\tilde{\Lambda}_d^+ \Lambda_c^+ & \end{array} \right]. \quad (2.72)$$

Chapter 3

GALERKIN METHOD FOR THE BIHARMONIC DIRICHLET PROBLEM

In this chapter, we formulate the mixed finite element Galerkin method with piecewise Hermite bicubics for the biharmonic Dirichlet problem (1.1) and develop its matrix-vector form.

3.1 Mixed Finite Element Galerkin Method

If we introduce $v = \Delta u$, then the mixed formulation of (1.1) consists of

$$\left. \begin{aligned} v - \Delta u &= 0 \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega, \end{aligned} \right\} \quad (3.1)$$

and

$$-\Delta v = -f(x, y) \text{ in } \Omega. \quad (3.2)$$

Assume u and v are sufficiently smooth and satisfy (3.1) and (3.2). Then the weak form of (3.1) and (3.2) becomes

$$\int_{\Omega} v\eta dx dy + \int_{\Omega} \nabla u \nabla \eta dx dy = 0, \quad \eta \in H^1(\Omega), \quad (3.3)$$

$$\int_{\Omega} \nabla v \nabla \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in H_0^1(\Omega), \quad (3.4)$$

where $H^1(\Omega)$, $H_0^1(\Omega)$ are classical Sobolev spaces [6]. This form is obtained by multiplying the first equation of (3.1) and (3.2) by $\eta \in H^1(\Omega)$ and $\delta \in H_0^1(\Omega)$, respectively, and then integrating both sides of the resulting equations to obtain

$$\int_{\Omega} v\eta dx dy - \int_{\Omega} (\Delta u) \eta dx dy = 0, \quad \eta \in H^1(\Omega),$$

$$- \int_{\Omega} (\Delta v) \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in H_0^1(\Omega).$$

Applying Green's Formula and using $\frac{\partial u}{\partial n} = 0$ and $\delta = 0$ on $\partial\Omega$ gives

$$- \int_{\Omega} (\Delta u) \eta dx dy = \int_{\Omega} \nabla u \cdot \nabla \eta dx dy - \int_{\partial\Omega} \eta \frac{\partial u}{\partial n} d\partial\Omega = \int_{\Omega} \nabla u \nabla \eta dx dy,$$

$$- \int_{\Omega} (\Delta v) \delta dx dy = \int_{\Omega} \nabla v \cdot \nabla \delta dx dy - \int_{\partial\Omega} \delta \frac{\partial v}{\partial n} d\partial\Omega = \int_{\Omega} \nabla v \nabla \delta dx dy,$$

and hence we obtain (3.3) and (3.4).

With \mathcal{M}_h and \mathcal{M}_h^0 defined by (2.1) and (2.2), respectively, we introduce

$$X_h = \{w \in \mathcal{M}_h \otimes \mathcal{M}_h : w(\alpha, \beta) = 0, \alpha, \beta = 0, 1\},$$

and

$$X_h^0 = \mathcal{M}_h^0 \otimes \mathcal{M}_h^0.$$

Clearly, $X_h^0 \subset X_h$.

The Galerkin solution of (3.1)–(3.2) consists of finding $U \in X_h^0$ and $V \in X_h$ such that

$$\int_{\Omega} V \eta dx dy + \int_{\Omega} \nabla U \cdot \nabla \eta dx dy = 0, \quad \eta \in X_h, \quad (3.5)$$

$$\int_{\Omega} \nabla V \cdot \nabla \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in X_h^0. \quad (3.6)$$

Note that $V \in X_h$ and hence $V(\alpha, \beta) = 0, \alpha, \beta = 0, 1$. We impose these four corner conditions on V since for $v = \Delta u$ we also have

$$v(\alpha, \beta) = u_{xx}(\alpha, \beta) + u_{yy}(\alpha, \beta) = 0, \quad \alpha, \beta = 0, 1.$$

Normally $V, \eta \in \mathcal{M}_h \otimes \mathcal{M}_h$ rather than X_h . Therefore (3.5)–(3.6) can be regarded as a modification of the standard mixed finite element Galerkin method.

Lemma 3.1 *Problem (3.5)–(3.6) has a unique solution.*

Proof. Consider $f(x, y) = 0$. Taking $\eta = V$ in (3.5) and $\delta = U$ in (3.6), we obtain

$$\int_{\Omega} V^2 dx dy + \int_{\Omega} \nabla U \nabla V dx dy = 0,$$

$$\int_{\Omega} \nabla V \cdot \nabla U dx dy = 0,$$

respectively. Clearly

$$\int_{\Omega} V^2 dx dy = 0$$

and hence $V = 0$.

Taking $\eta = U$ in (3.5) and using $V = 0$, we also have

$$\int_{\Omega} \nabla U \cdot \nabla U dx dy = 0.$$

Hence $U = 0$ by the Poincaré inequality [6]. \square

3.2 Matrix-Vector Form

In this section we derive the matrix-vector form of (3.5)–(3.6). Since $\{\phi_i(x)\}_{i=1}^{2N}$ of (2.7) and $\{\psi_i(y)\}_{i=1}^{2N}$ of (2.8) are bases for \mathcal{M}_h^0 , and $\{\phi_i(x)\}_{i=0}^{2N+1}$ of (2.5) and $\{\psi_i(y)\}_{i=0}^{2N+1}$ of (2.6) are bases for \mathcal{M}_h , then $\{\phi_i(x)\psi_j(y)\}_{i=1,j=1}^{2N,2N}$ and

$$\{\phi_i(x)\psi_j(y)\}_{i=1,j=1}^{2N,2N} \cup \{\phi_i(x)\psi_j(y)\}_{i=1,j=0,2N+1}^{2N} \cup \{\phi_i(x)\psi_j(y)\}_{i=0,2N+1,j=1}^{2N}, \quad (3.7)$$

are bases for X_h^0 and X_h , respectively. Hence the Galerkin solutions $U \in X_h^0$ and

$V \in X_h$ of (3.5) and (3.6) are of the form

$$U(x, y) = \sum_{k=1}^{2N} \sum_{l=1}^{2N} u_{k,l} \phi_k(x) \psi_l(y),$$

$$V(x, y) = \sum_{k=1}^{2N} \sum_{l=1}^{2N} v_{k,l} \phi_k(x) \psi_l(y) + \sum_{k=1}^{2N} \sum_{l=0, 2N+1} v_{k,l} \phi_k(x) \psi_l(y) + \sum_{k=0, 2N+1} \sum_{l=1}^{2N} v_{k,l} \phi_k(x) \psi_l(y).$$

Substituting these expressions into (3.5) and (3.6), and taking $\eta = \phi_i(x) \psi_j(y)$, $\delta =$

$\phi_i(x) \psi_j(y)$, we get

$$\begin{aligned} & \sum_{k=1}^{2N} \sum_{l=1}^{2N} \int_{\Omega} \phi_k(x) \psi_l(y) \phi_i(x) \psi_j(y) dx dy v_{k,l} \\ & + \sum_{k=1}^{2N} \sum_{l=0, 2N+1} \int_{\Omega} \phi_k(x) \psi_l(y) \phi_i(x) \psi_j(y) dx dy v_{k,l} + \sum_{k=0, 2N+1} \sum_{l=1}^{2N} \int_{\Omega} \phi_k(x) \psi_l(y) \phi_i(x) \psi_j(y) dx dy v_{k,l} \\ & + \sum_{k=1}^{2N} \sum_{l=1}^{2N} \int_{\Omega} [\phi'_k(x) \psi_l(y) \phi'_i(x) \psi_j(y) + \phi_k(x) \psi'_l(y) \phi_i(x) \psi'_j(y)] dx dy u_{k,l} = 0, \quad (3.8) \end{aligned}$$

where $i, j = 1, \dots, 2N$, $i = 1, \dots, 2N$ for $j = 0, 2N+1$, $i = 0, 2N+1$ for $j = 1, \dots, 2N$,

and

$$\begin{aligned} & \sum_{k=1}^{2N} \sum_{l=1}^{2N} \int_{\Omega} [\phi'_k(x) \psi_l(y) \phi'_i(x) \psi_j(y) + \phi_k(x) \psi'_l(y) \phi_i(x) \psi'_j(y)] dx dy v_{k,l} \\ & + \sum_{k=1}^{2N} \sum_{l=0, 2N+1} \int_{\Omega} [\phi'_k(x) \psi_l(y) \phi'_i(x) \psi_j(y) + \phi_k(x) \psi'_l(y) \phi_i(x) \psi'_j(y)] dx dy v_{k,l} \\ & + \sum_{k=0, 2N+1} \sum_{l=1}^{2N} \int_{\Omega} [\phi'_k(x) \psi_l(y) \phi'_i(x) \psi_j(y) + \phi_k(x) \psi'_l(y) \phi_i(x) \psi'_j(y)] dx dy v_{k,l} \end{aligned}$$

$$= - \int_{\Omega} f(x, y) \phi_i(x) \psi_j(y) dx dy,$$

where $i, j = 1, \dots, 2N$.

Using Fubini's theorem and rearranging the order of the double sum involving $v_{0,l}$ and $v_{2N+1,l}$ with the double sum involving $u_{k,l}$ in (3.8), we obtain

$$\begin{aligned} & \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy u_{k,l} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy u_{k,l} \\ & + \sum_{k=0, 2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} = 0, \end{aligned} \quad (3.9)$$

where $i, j = 1, \dots, 2N$, $i = 1, \dots, 2N$ for $j = 0, 2N+1$, $i = 0, 2N+1$ for $j = 1, \dots, 2N$,

and

$$\begin{aligned} & \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{k,l} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{k,l} \end{aligned}$$

$$\begin{aligned}
& + \sum_{k=0,2N+1} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} \\
& + \sum_{k=0,2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{k,l} \\
& = - \int_0^1 \left[\int_0^1 f(x, y) \phi_i(x) dx \right] \psi_j(y) dy, \tag{3.10}
\end{aligned}$$

where $i, j = 1, \dots, 2N$.

At this point, corresponding to representations of $V(x, y)$ and $U(x, y)$ in terms of basis functions, let us introduce vectors

$$\vec{v}_i = [v_{1,1}, \dots, v_{1,2N}, v_{2,1}, \dots, v_{2,2N}, \dots, v_{2N,1}, \dots, v_{2N,2N}]^T, \tag{3.11}$$

$$\vec{v}_h = [v_{1,0}, v_{1,2N+1}, v_{2,0}, v_{2,2N+1}, \dots, v_{2N-1,0}, v_{2N-1,2N+1}, v_{2N,0}, v_{2N,2N+1}]^T, \tag{3.12}$$

$$\vec{v}_v = [v_{0,1}, \dots, v_{0,2N}, v_{2N+1,1}, \dots, v_{2N+1,2N}]^T, \tag{3.13}$$

and

$$\vec{u} = [u_{1,1}, \dots, u_{1,2N}, u_{2,1}, \dots, u_{2,2N}, \dots, u_{2N,1}, \dots, u_{2N,2N}]^T. \tag{3.14}$$

Also, let

$$\vec{f} = [f_{1,1}, \dots, f_{1,2N}, f_{2,1}, \dots, f_{2,2N}, \dots, f_{2N,1}, \dots, f_{2N,2N}]^T, \tag{3.15}$$

where

$$f_{i,j} = - \int_0^1 \left[\int_0^1 f(x, y) \phi_i(x) dx \right] \psi_j(y) dy. \tag{3.16}$$

Note that \vec{v}_i , \vec{u} , and \vec{f} have $4N^2$ components, while \vec{v}_h and \vec{v}_v have $4N$ components.

The matrix-vector form of (3.5) and (3.6) is then

$$A_{11}\vec{v}_i + A_{12}\vec{v}_h + A_{13}\vec{u} + A_{14}\vec{v}_v = \vec{0}, \quad (3.17)$$

$$A_{21}\vec{v}_i + A_{22}\vec{v}_h + A_{23}\vec{u} + A_{24}\vec{v}_v = \vec{0}, \quad (3.18)$$

$$A_{31}\vec{v}_i + A_{32}\vec{v}_h + A_{34}\vec{v}_v = \vec{f}, \quad (3.19)$$

$$A_{41}\vec{v}_i + A_{42}\vec{v}_h + A_{43}\vec{u} + A_{44}\vec{v}_v = \vec{0}. \quad (3.20)$$

Equation (3.17) corresponds to the $4N^2$ equations of (3.9) with $i, j = 1, \dots, 2N$. The first $2N$ of these equations are obtained by selecting $i = 1$ and $j = 1, \dots, 2N$, with succeeding equations obtained by indexing $i = 2, \dots, 2N$ and $j = 1, \dots, 2N$. By (2.25)–(2.26) and (2.21), the $4N^2 \times 4N^2$ matrix

$$A_{11} = B_x \otimes B_y, \quad (3.21)$$

where B_x and B_y are given by (2.13) and (2.18), respectively. The $4N^2 \times 4N$ matrix

$$A_{12} = B_x \otimes B_y^b, \quad (3.22)$$

where B_x is defined in (2.13), and B_y^b is given by

$$B_y^b = (b_{j,l}^y)_{j=1,l=0,2N+1}^{2N}, \quad b_{j,l}^y = \int_0^1 (\psi_j \psi_l)(y) dy. \quad (3.23)$$

The $4N^2 \times 4N^2$ matrix

$$A_{13} = A_x \otimes B_y + B_x \otimes A_y, \quad (3.24)$$

where A_x , B_y , B_x , and A_y , are defined in (2.12), (2.18), (2.13), and (2.17), respectively.

The $4N^2 \times 4N$ matrix

$$A_{14} = B_x^b \otimes B_y, \quad (3.25)$$

where the $2N \times 2$ matrix B_x^b is given by

$$B_x^b = (b_{i,k}^x)_{i=1,k=0,2N+1}^{2N}, \quad b_{i,k}^x = \int_0^1 (\phi_i \phi_k)(x) dx, \quad (3.26)$$

and B_y is given by (2.18).

Equation (3.18) corresponds to the $4N$ equations in (3.9) with $i = 1$, $j = 0, 2N + 1$, then indexing $i = 2, \dots, 2N$ and taking $j = 0, 2N + 1$. The $4N \times 4N^2$ matrix

$$A_{21} = B_x \otimes (B_y^b)^T, \quad (3.27)$$

where B_x is defined in (2.13) and B_y^b is given by (3.23). The $4N \times 4N$ matrix

$$A_{22} = B_x \otimes B_y^{bb}, \quad (3.28)$$

where B_x is defined in (2.13) and B_y^{bb} is given by

$$B_y^{bb} = (b_{j,l}^y)_{j=0,2N+1,l=0,2N+1}, \quad b_{j,l}^y = \int_0^1 (\psi_j \psi_l)(y) dy. \quad (3.29)$$

The $4N \times 4N^2$ matrix

$$A_{23} = A_x \otimes (B_y^b)^T + B_x \otimes (A_y^b)^T, \quad (3.30)$$

where A_x , B_x , and B_y^b are defined in (2.12), (2.13), and (3.23), respectively, and A_y^b is given by

$$A_y^b = (a_{j,l}^y)_{j=1,l=0,2N+1}^{2N}, \quad a_{j,l}^y = \int_0^1 (\psi_j' \psi_l')(y) dy. \quad (3.31)$$

The $4N \times 4N$ matrix

$$A_{24} = B_x^b \otimes (B_y^b)^T, \quad (3.32)$$

where B_x^b is given by (3.26) and B_y^b is given by (3.23).

Equation (3.19) corresponds to the $4N^2$ equations of (3.10). The equations are obtained by taking $i = 1$ and $j = 1, \dots, 2N$, and then indexing $i = 2, \dots, 2N$ and

$j = 1, \dots, 2N$. It follows by comparing (3.9) and (3.10) that

$$A_{31} = A_{13}, \quad (3.33)$$

where A_{13} is defined by (3.24). The $4N^2 \times 4N$ matrix

$$A_{32} = A_x \otimes B_y^b + B_x \otimes A_y^b, \quad (3.34)$$

where A_x , B_y^b , B_x , and A_y^b are defined in (2.12), (3.23), (2.13), and (3.31), respectively.

The $4N^2 \times 4N$ matrix

$$A_{34} = A_x^b \otimes B_y + B_x^b \otimes A_y, \quad (3.35)$$

where B_y , B_x^b , and A_y are given by (2.18), (3.26), and (2.17), respectively, and the $(2N) \times 2$ matrix A_x^b is given by

$$A_x^b = (a_{i,k}^x)_{i=1,k=0,2N+1}^{2N}, \quad a_{i,k}^x = \int_0^1 (\phi_i' \phi_k')(x) dx. \quad (3.36)$$

Equation (3.20) corresponds to $4N$ equations in (3.9) with $i = 0, j = 1, \dots, 2N$, and $i = 2N + 1, j = 1, \dots, 2N$. The $4N \times 4N^2$ matrix

$$A_{41} = (B_x^b)^T \otimes B_y, \quad (3.37)$$

where B_x^b and B_y are given by (3.26) and (2.18), respectively. The $4N \times 4N$ matrix

$$A_{42} = (B_x^b)^T \otimes B_y^b, \quad (3.38)$$

where B_x^b and B_y^b are given by (3.26) and (3.23), respectively. The $4N \times 4N^2$ matrix

$$A_{43} = (A_x^b)^T \otimes B_y + (B_x^b)^T \otimes A_y, \quad (3.39)$$

where A_x^b , B_y , B_x^b , and A_y are given by (3.36), (2.18), (3.26), and (2.17), respectively.

The $4N \times 4N$ matrix

$$A_{44} = B_x^{bb} \otimes B_y, \quad (3.40)$$

where B_x^{bb} is defined by

$$B_x^{bb} = (b_{i,k}^x)_{i=0,2N+1, k=0,2N+1}, \quad b_{i,k}^x = \int_0^1 (\phi_i \phi_k)(x) dx, \quad (3.41)$$

and B_y is given by (2.18).

Chapter 4

METHOD FOR SOLVING THE GALERKIN PROBLEM

In this chapter, we solve the Galerkin problem (3.5)–(3.6) by reducing its matrix-vector form (3.17)–(3.20) to a symmetric, positive definite Schur complement system. This reduces the size of the problem from $O(N^2)$ to $O(N)$. We also discuss the PCG method for solving the Schur complement system.

4.1 Derivation of Algorithm for Solving the Galerkin Problem

Clearly the matrix-vector form (3.17)–(3.20) can be written as

$$M_{11}\vec{w} + M_{12}\vec{v}_v = \vec{g}, \quad (4.1)$$

$$M_{21}\vec{w} + A_{44}\vec{v}_v = \vec{0}, \quad (4.2)$$

where

$$M_{11} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & 0 \end{bmatrix}, \quad M_{12} = \begin{bmatrix} A_{14} \\ A_{24} \\ A_{34} \end{bmatrix}, \quad M_{21} = \begin{bmatrix} A_{41} & A_{42} & A_{43} \end{bmatrix}, \quad (4.3)$$

and

$$\vec{w} = \begin{bmatrix} \vec{v}_i \\ \vec{v}_h \\ \vec{u} \end{bmatrix}, \quad \vec{g} = \begin{bmatrix} \vec{0} \\ \vec{0} \\ \vec{f} \end{bmatrix}. \quad (4.4)$$

Suppose that M_{11}^{-1} exists (this will be demonstrated in Lemma 4.3). Then multiplying (4.1) by M_{11}^{-1} and solving for \vec{w} , we have

$$\vec{w} = M_{11}^{-1} \vec{g} - M_{11}^{-1} M_{12} \vec{v}_v. \quad (4.5)$$

By substituting (4.5) into (4.2), we obtain the Schur complement system

$$S \vec{v}_v = \vec{c}, \quad (4.6)$$

where

$$S = A_{44} - M_{21} M_{11}^{-1} M_{12}, \quad \vec{c} = -M_{21} M_{11}^{-1} \vec{g}, \quad (4.7)$$

and S is the Schur complement (cf. [13]) of M_{11} in

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & A_{44} \end{bmatrix}. \quad (4.8)$$

Lemma 4.1 *The matrix S of (4.7) is symmetric and positive definite.*

Proof. We see from (2.12), (2.13), (2.17), (2.18), that A_x , B_x , A_y , and B_y are symmetric. Therefore it follows from (3.21), (3.24), and (2.23) that

$$A_{11} = A_{11}^T, \quad A_{13} = A_{13}^T. \quad (4.9)$$

Similarly, (3.29) and (3.41) imply that B_y^{bb} and B_x^{bb} are symmetric. Hence using (3.28), (3.40), and (2.23), we obtain

$$A_{22} = A_{22}^T, \quad A_{44} = A_{44}^T. \quad (4.10)$$

From (3.33) and (4.9), we see that

$$A_{31} = A_{13}^T. \quad (4.11)$$

It also follows from (3.27), (3.22), (3.37), (3.25), (3.34), (3.30), (3.38), (3.32), (3.39), (3.35), and (2.23) that

$$A_{21} = A_{12}^T, \quad A_{41} = A_{14}^T, \quad A_{32} = A_{23}^T, \quad A_{42} = A_{24}^T, \quad A_{43} = A_{34}^T. \quad (4.12)$$

It follows from (4.9)–(4.12), and (4.3) that

$$M_{11} = M_{11}^T, \quad M_{21} = M_{12}^T. \quad (4.13)$$

Therefore, by (4.7), (4.10), and (4.13), we have $S = S^T$.

To show that S is positive definite, let us consider any fixed, non-zero $\vec{v}_v \in \mathbb{R}^{4N}$ of the form (3.13). Let \vec{v}_i , \vec{v}_h , and \vec{u} of the form (3.11), (3.12), and (3.14), be such that

$$A_{11}\vec{v}_i + A_{12}\vec{v}_h + A_{13}\vec{u} + A_{14}\vec{v}_v = \vec{0}, \quad (4.14)$$

$$A_{21}\vec{v}_i + A_{22}\vec{v}_h + A_{23}\vec{u} + A_{24}\vec{v}_v = \vec{0}, \quad (4.15)$$

$$A_{31}\vec{v}_i + A_{32}\vec{v}_h + A_{34}\vec{v}_v = \vec{0}, \quad (4.16)$$

where $A_{11}, A_{12}, A_{13}, A_{14}, A_{21}, A_{22}, A_{23}, A_{24}, A_{31}, A_{32}$, and A_{34} are as in (3.17)–(3.19).

Then by (4.3), (4.14)–(4.16) is equivalent to

$$M_{11} \begin{bmatrix} \vec{v}_i \\ \vec{v}_h \\ \vec{u} \end{bmatrix} = -M_{12}\vec{v}_v. \quad (4.17)$$

We know that \vec{v}_i , \vec{v}_h , and \vec{u} exist since M_{11} is nonsingular by Lemma 4.3, which we present later. In fact, (4.17) gives

$$\begin{bmatrix} \vec{v}_i \\ \vec{v}_h \\ \vec{u} \end{bmatrix} = -M_{11}^{-1}M_{12}\vec{v}_v. \quad (4.18)$$

Now by (4.7), (4.13), (4.18), (4.3), (4.16), (4.11)–(4.12), (4.14), and (4.15), we have

$$\begin{aligned}
(S\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} &= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} - (M_{21}M_{11}^{-1}M_{12}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} - (M_{11}^{-1}M_{12}\vec{v}_v, M_{12}\vec{v}_v)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} + ([\vec{v}_i, \vec{v}_h, \vec{u}]^T, [A_{14}, A_{24}, A_{34}]^T \vec{v}_v)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} + (\vec{v}_i, A_{14}\vec{v}_v)_{\mathbb{R}^{4N}} + (\vec{v}_h, A_{24}\vec{v}_v)_{\mathbb{R}^{4N}} + (\vec{u}, A_{34}\vec{v}_v)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} + (\vec{v}_i, A_{14}\vec{v}_v)_{\mathbb{R}^{4N}} + (\vec{v}_h, A_{24}\vec{v}_v)_{\mathbb{R}^{4N}} - (\vec{u}, A_{31}\vec{v}_i)_{\mathbb{R}^{4N}} - (\vec{u}, A_{32}\vec{v}_h)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{41}\vec{v}_i, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{42}\vec{v}_h, \vec{v}_v)_{\mathbb{R}^{4N}} - (A_{13}\vec{u}, \vec{v}_i)_{\mathbb{R}^{4N}} - (A_{23}\vec{u}, \vec{v}_h)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{41}\vec{v}_i, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{42}\vec{v}_h, \vec{v}_v)_{\mathbb{R}^{4N}} \\
&\quad + (A_{11}\vec{v}_i, \vec{v}_i)_{\mathbb{R}^{4N}} + (A_{12}\vec{v}_h, \vec{v}_i)_{\mathbb{R}^{4N}} + (A_{14}\vec{v}_v, \vec{v}_i)_{\mathbb{R}^{4N}} \\
&\quad + (A_{21}\vec{v}_i, \vec{v}_h)_{\mathbb{R}^{4N}} + (A_{22}\vec{v}_h, \vec{v}_h)_{\mathbb{R}^{4N}} + (A_{24}\vec{v}_v, \vec{v}_h)_{\mathbb{R}^{4N}}.
\end{aligned} \tag{4.19}$$

Introduce

$$z = \tilde{v}_i + \tilde{v}_h + \tilde{v}_v,$$

where the functions

$$\tilde{v}_i = \sum_{k=1}^{2N} \sum_{l=1}^{2N} v_{k,l} \phi_k(x) \psi_l(y), \tag{4.20}$$

$$\tilde{v}_h = \sum_{k=1}^{2N} \sum_{l=0, 2N+1} v_{k,l} \phi_k(x) \psi_l(y), \tag{4.21}$$

$$\tilde{v}_v = \sum_{k=0, 2N+1} \sum_{l=1}^{2N} v_{k,l} \phi_k(x) \psi_l(y). \tag{4.22}$$

With $(\cdot, \cdot)_{L^2(\Omega)}$ denoting the usual L^2 inner product, consider

$$\begin{aligned}
(z, z)_{L^2(\Omega)} &= (\tilde{v}_i + \tilde{v}_h + \tilde{v}_v, \tilde{v}_i + \tilde{v}_h + \tilde{v}_v)_{L^2(\Omega)} \\
&= (\tilde{v}_i, \tilde{v}_i)_{L^2(\Omega)} + (\tilde{v}_h, \tilde{v}_i)_{L^2(\Omega)} + (\tilde{v}_v, \tilde{v}_i)_{L^2(\Omega)} + (\tilde{v}_i, \tilde{v}_h)_{L^2(\Omega)} + (\tilde{v}_h, \tilde{v}_h)_{L^2(\Omega)} + (\tilde{v}_v, \tilde{v}_h)_{L^2(\Omega)} \\
&\quad + (\tilde{v}_i, \tilde{v}_v)_{L^2(\Omega)} + (\tilde{v}_h, \tilde{v}_v)_{L^2(\Omega)} + (\tilde{v}_v, \tilde{v}_v)_{L^2(\Omega)}. \tag{4.23}
\end{aligned}$$

By (4.20), (2.25)–(2.26), (2.21), (2.13), (2.18), and (3.21), we can write

$$\begin{aligned}
(\tilde{v}_i, \tilde{v}_i)_{L^2(\Omega)} &= \int_{\Omega} \tilde{v}_i^2(x, y) dx dy \\
&= \int_{\Omega} \left[\sum_{k=1}^{2N} \sum_{l=1}^{2N} v_{k,l} \phi_k(x) \psi_l(y) \sum_{i=1}^{2N} \sum_{j=1}^{2N} v_{i,j} \phi_i(x) \psi_j(y) \right] dx dy \\
&= \sum_{i=1}^{2N} \sum_{j=1}^{2N} \left[\sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{k,l} \right] v_{i,j} \\
&= ((B_x \otimes B_y) \vec{v}_i, \vec{v}_i)_{\mathbb{R}^{4N}} = (A_{11} \vec{v}_i, \vec{v}_i)_{\mathbb{R}^{4N}}.
\end{aligned}$$

Using the same reasoning, from (4.23) we obtain

$$\begin{aligned}
(z, z)_{L^2(\Omega)} &= (A_{11} \vec{v}_i, \vec{v}_i)_{\mathbb{R}^{4N}} + (A_{12} \vec{v}_h, \vec{v}_i)_{\mathbb{R}^{4N}} + (A_{14} \vec{v}_v, \vec{v}_i)_{\mathbb{R}^{4N}} \\
&\quad + (A_{21} \vec{v}_i, \vec{v}_h)_{\mathbb{R}^{4N}} + (A_{22} \vec{v}_h, \vec{v}_h)_{\mathbb{R}^{4N}} + (A_{24} \vec{v}_v, \vec{v}_h)_{\mathbb{R}^{4N}} \\
&\quad + (A_{41} \vec{v}_i, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{42} \vec{v}_h, \vec{v}_v)_{\mathbb{R}^{4N}} + (A_{44} \vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}}. \tag{4.24}
\end{aligned}$$

From (3.7), the basis functions in (4.20)–(4.22) are linearly independent. Therefore $z \neq 0$ for $\vec{v}_v \neq \vec{0}$ and hence by (4.19) and (4.24), $(S\vec{v}_v, \vec{v}_v)_{\mathbb{R}^{4N}} = (z, z)_{L^2(\Omega)} > 0$. \square

Lemma 4.1 implies that S is nonsingular. Hence, once we have solved (4.6) for \vec{v}_v , then (4.1) implies that \vec{w} can be obtained by solving

$$M_{11}\vec{w} = \vec{g} - M_{12}\vec{v}_v. \quad (4.25)$$

Therefore we arrive at the following algorithm.

ALGORITHM FOR SOLVING (4.1)–(4.2)

1. Compute \vec{c} of (4.7).
2. Solve (4.6) for \vec{v}_v . (4.26)
3. Compute $\vec{g} - M_{12}\vec{v}_v$.
4. Solve (4.25) for \vec{w} .

In Section 4.4 we discuss solving the Schur complement problem in step 2 of the algorithm. First, we observe that steps 1 and 4 of the algorithm require solving linear systems with coefficient matrix M_{11} . In the next section we demonstrate how to solve $M_{11}\vec{w} = \vec{b}$ for arbitrary \vec{b} .

4.2 Solving $M_{11}\vec{w} = \vec{b}$

First, we write \vec{b} in the form

$$\vec{b} = [\vec{b}_1, \vec{b}_2, \vec{b}_3]^T,$$

where

$$\vec{b}_1 = [(b_1)_{1,1}, \dots, (b_1)_{1,2N}, (b_1)_{2,1}, \dots, (b_1)_{2,2N}, \dots, (b_1)_{2N,1}, \dots, (b_1)_{2N,2N},]^T, \quad (4.27)$$

$$\vec{b}_2 = [(b_2)_{1,0}, (b_2)_{1,2N+1}, (b_2)_{2,0}, (b_2)_{2,2N+1}, \dots, (b_2)_{2N,0}, (b_2)_{2N,2N+1},]^T, \quad (4.28)$$

$$\vec{b}_3 = [(b_3)_{1,1}, \dots, (b_3)_{1,2N}, (b_3)_{2,1}, \dots, (b_3)_{2,2N}, \dots, (b_3)_{2N,1}, \dots, (b_3)_{2N,2N},]^T. \quad (4.29)$$

The vectors \vec{b}_1 , \vec{b}_2 , and \vec{b}_3 have lengths $4N^2$, $4N$, and $4N^2$, respectively, to match the number of rows in the corresponding sub-blocks of M_{11} . Using M_{11} of (4.3), (3.21), (3.22), (3.24), (3.27), (3.28), (3.30), (3.33), (3.34), and \vec{w} as in (4.4), $M_{11}\vec{w} = \vec{b}$ becomes

$$(B_x \otimes B_y)\vec{v}_i + (B_x \otimes B_y^b)\vec{v}_h + (A_x \otimes B_y + B_x \otimes A_y)\vec{u} = \vec{b}_1, \quad (4.30)$$

$$[B_x \otimes (B_y^b)^T]\vec{v}_i + (B_x \otimes B_y^{bb})\vec{v}_h + [A_x \otimes (B_y^b)^T + B_x \otimes (A_y^b)^T]\vec{u} = \vec{b}_2, \quad (4.31)$$

$$(A_x \otimes B_y + B_x \otimes A_y)\vec{v}_i + (A_x \otimes B_y^b + B_x \otimes A_y^b)\vec{v}_h = \vec{b}_3, \quad (4.32)$$

where \vec{v}_i , \vec{v}_h , and \vec{u} have lengths $4N^2$, $4N$, and $4N^2$, respectively.

We pre-multiply both sides of (4.30) by $(Z^T \otimes I_{2N})$, where Z is given by (2.72), and use the relationships

$$I_{4N^2} = (Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}, \quad I_{4N} = (Z \otimes I_2)(Z \otimes I_2)^{-1}$$

to obtain

$$\begin{aligned} & (Z^T \otimes I_{2N})(B_x \otimes B_y)(Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}\vec{v}_i \\ & + (Z^T \otimes I_{2N})(B_x \otimes B_y^b)(Z \otimes I_2)(Z \otimes I_2)^{-1}\vec{v}_h \\ & + (Z^T \otimes I_{2N})(A_x \otimes B_y + B_x \otimes A_y)(Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}\vec{u} = (Z^T \otimes I_{2N})\vec{b}_1. \end{aligned} \quad (4.33)$$

In a similar way, we pre-multiply both sides of (4.31) by $(Z^T \otimes I_2)$ to obtain

$$\begin{aligned} & (Z^T \otimes I_2)[B_x \otimes (B_y^b)^T](Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}\vec{v}_i \\ & + (Z^T \otimes I_2)(B_x \otimes B_y^{bb})(Z \otimes I_2)(Z \otimes I_2)^{-1}\vec{v}_h \\ & + (Z^T \otimes I_2)[A_x \otimes (B_y^b)^T + B_x \otimes (A_y^b)^T](Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}\vec{u} = (Z^T \otimes I_2)\vec{b}_2. \end{aligned} \quad (4.34)$$

Equation (4.32) can be treated in the same way as (4.30) to yield

$$(Z^T \otimes I_{2N})(A_x \otimes B_y + B_x \otimes A_y)(Z \otimes I_{2N})(Z \otimes I_{2N})^{-1}\vec{v}_i$$

$$+ (Z^T \otimes I_{2N})(A_x \otimes B_y^b + B_x \otimes A_y^b)(Z \otimes I_2)(Z \otimes I_2)^{-1}\vec{v}_h = (Z^T \otimes I_{2N})\vec{b}_3. \quad (4.35)$$

Corresponding to \vec{v}_i of (3.11), \vec{v}_h of (3.12), and \vec{u} of (3.14), we introduce

$$\begin{aligned} \vec{v}'_i &= (Z \otimes I_{2N})^{-1}\vec{v}_i \\ &= [v'_{1,1}, \dots, v'_{1,2N}, v'_{2,1}, \dots, v'_{2,2N}, \dots, v'_{2N,1}, \dots, v'_{2N,2N}]^T, \end{aligned} \quad (4.36)$$

$$\begin{aligned} \vec{v}'_h &= (Z \otimes I_2)^{-1}\vec{v}_h \\ &= [v'_{1,0}, v'_{1,2N+1}, v'_{2,0}, v'_{2,2N+1}, \dots, v'_{2N-1,0}, v'_{2N-1,2N+1}, v'_{2N,0}, v'_{2N,2N+1}]^T, \end{aligned} \quad (4.37)$$

$$\begin{aligned} \vec{u}' &= (Z \otimes I_{2N})^{-1}\vec{u} \\ &= [u'_{1,1}, \dots, u'_{1,2N}, u'_{2,1}, \dots, u'_{2,2N}, \dots, u'_{2N,1}, \dots, u'_{2N,2N}]^T. \end{aligned} \quad (4.38)$$

Similarly, corresponding to \vec{b}_1 , \vec{b}_2 , and \vec{b}_3 of (4.27)–(4.29), we introduce

$$\begin{aligned} \vec{b}'_1 &= (Z^T \otimes I_{2N})\vec{b}_1 \\ &= [(b'_1)_{1,1}, \dots, (b'_1)_{1,2N}, (b'_1)_{2,1}, \dots, (b'_1)_{2,2N}, \dots, (b'_1)_{2N,1}, \dots, (b'_1)_{2N,2N}]^T \end{aligned} \quad (4.39)$$

$$\begin{aligned}
\vec{b}'_2 &= (Z^T \otimes I_2) \vec{b}_2 \\
&= [(b'_2)_{1,0}, (b'_2)_{1,2N+1}, (b'_2)_{2,0}, (b'_2)_{2,2N+1}, \dots, (b'_2)_{2N,0}, (b'_2)_{2N,2N+1}]^T, \quad (4.40)
\end{aligned}$$

$$\begin{aligned}
\vec{b}'_3 &= (Z^T \otimes I_{2N}) \vec{b}_3 \\
&= [(b'_3)_{1,1}, \dots, (b'_3)_{1,2N}, (b'_3)_{2,1}, \dots, (b'_3)_{2,2N}, \dots, (b'_3)_{2N,1}, \dots, (b'_3)_{2N,2N}]^T \quad (4.41)
\end{aligned}$$

From (2.27) and (2.28), we can write

$$Z^T A_x Z = \Lambda. \quad (4.42)$$

Then utilizing relationships (2.24), (2.28), (4.42), and (4.36)–(4.41) in (4.33)–(4.35), we obtain

$$(I_{2N} \otimes B_y) \vec{v}'_i + (I_{2N} \otimes B_y^b) \vec{v}'_h + (\Lambda \otimes B_y + I_{2N} \otimes A_y) \vec{u}' = \vec{b}'_1, \quad (4.43)$$

$$[I_{2N} \otimes (B_y^b)^T] \vec{v}'_i + (I_{2N} \otimes B_y^{bb}) \vec{v}'_h + [\Lambda \otimes (B_y^b)^T + I_{2N} \otimes (A_y^b)^T] \vec{u}' = \vec{b}'_2, \quad (4.44)$$

$$(\Lambda \otimes B_y + I_{2N} \otimes A_y) \vec{v}'_i + (\Lambda \otimes B_y^b + I_{2N} \otimes A_y^b) \vec{v}'_h = \vec{b}'_3. \quad (4.45)$$

Since Λ is diagonal, each of (4.43), (4.44), and (4.45) splits into a collection of $2N$

independent linear systems. In fact, using (2.70), we have

$$B_y(\vec{v}'_i)_k + B_y^b(\vec{v}'_h)_k + (\lambda_k B_y + A_y)\vec{u}'_k = (\vec{b}'_1)_k, \quad (4.46)$$

$$(B_y^b)^T(\vec{v}'_i)_k + B_y^{bb}(\vec{v}'_h)_k + [\lambda_k(B_y^b)^T + (A_y^b)^T]\vec{u}'_k = (\vec{b}'_2)_k, \quad (4.47)$$

$$(\lambda_k B_y + A_y)(\vec{v}'_i)_k + (\lambda_k B_y^b + A_y^b)(\vec{v}'_h)_k = (\vec{b}'_3)_k, \quad (4.48)$$

for $k = 1, \dots, 2N$, where λ_k is defined by (2.63) and (2.31)–(2.32), and where, by (4.36)–(4.38),

$$(\vec{v}'_i)_k = [v'_{k,1}, \dots, v'_{k,2N}]^T, \quad (\vec{v}'_h)_k = [v'_{k,0}, v'_{k,2N+1}]^T, \quad \vec{u}'_k = [u'_{k,1}, \dots, u'_{k,2N}]^T,$$

and by (4.39)–(4.41)

$$(\vec{b}'_1)_k = [(b'_{1,k,1}, \dots, b'_{1,k,2N})^T, (\vec{b}'_2)_k = [(b'_{2,k,0}, b'_{2,k,2N+1})^T, (\vec{b}'_3)_k = [(b'_{3,k,1}, \dots, b'_{3,k,2N})^T]^T.$$

Thus, for fixed $k = 1, \dots, 2N$, (4.46), (4.47), and (4.48) form the linear system

$$B_k \begin{bmatrix} (\vec{v}'_i)_k \\ (\vec{v}'_h)_k \\ \vec{u}'_k \end{bmatrix} = \begin{bmatrix} (\vec{b}'_1)_k \\ (\vec{b}'_2)_k \\ (\vec{b}'_3)_k \end{bmatrix}, \quad (4.49)$$

where B_k is a 3×3 block-matrix of the form

$$B_k = \begin{bmatrix} B_y & B_y^b & \lambda_k B_y + A_y \\ (B_y^b)^T & B_y^{bb} & \lambda_k (B_y^b)^T + (A_y^b)^T \\ \lambda_k B_y + A_y & \lambda_k B_y^b + A_y^b & 0 \end{bmatrix}. \quad (4.50)$$

We now prove the following lemma.

Lemma 4.2 *For $k = 1, \dots, 2N$, B_k given by (4.50), where λ_k are defined in (2.31)–(2.32), is nonsingular.*

Proof. Consider the one-dimensional boundary value problem

$$-u''(y) + \lambda u(y) + v(y) = 0, \quad y \in (0, 1), \quad (4.51)$$

$$-v''(y) + \lambda v(y) = 0, \quad y \in (0, 1), \quad (4.52)$$

where $\lambda \geq 0$ and

$$u(0) = u(1) = 0, \quad u'(0) = u'(1) = 0. \quad (4.53)$$

Multiplying (4.51) by $\eta \in H^1(0, 1)$ and (4.52) by $\delta \in H_0^1(0, 1)$ and then integrating we get

$$-\int_0^1 u''(y)\eta(y)dy + \lambda \int_0^1 u(y)\eta(y)dy + \int_0^1 v(y)\eta(y)dy = 0, \quad \eta \in H^1(0, 1),$$

$$-\int_0^1 v''(y)\delta(y)dy + \lambda \int_0^1 v(y)\delta(y)dy = 0, \quad \delta \in H_0^1(0,1).$$

Integration by parts and the boundary values $u'(0) = u'(1) = 0$ and $\delta(0) = \delta(1) = 0$ give

$$\int_0^1 u'(y)\eta'(y)dy + \lambda \int_0^1 u(y)\eta(y)dy + \int_0^1 v(y)\eta(y)dy = 0, \quad \eta \in H^1(0,1),$$

$$\int_0^1 v'(y)\delta'(y)dy + \lambda \int_0^1 v(y)\delta(y)dy = 0, \quad \delta \in H_0^1(0,1).$$

With \mathcal{M}_h and \mathcal{M}_h^0 of (2.1) and (2.2), respectively, the Galerkin solution of (4.51)–(4.53) consists of $U \in \mathcal{M}_h^0$ and $V \in \mathcal{M}_h$ such that

$$\int_0^1 U'\eta'dy + \lambda \int_0^1 U\eta dy + \int_0^1 V\eta dy = 0, \quad \eta \in \mathcal{M}_h, \quad (4.54)$$

$$\int_0^1 V'\delta'dy + \lambda \int_0^1 V\delta dy = 0, \quad \delta \in \mathcal{M}_h^0. \quad (4.55)$$

Next we demonstrate that the only solution of (4.54)–(4.55) is $U = V = 0$.

Taking $\eta = V$ and $\delta = U$, we obtain

$$\int_0^1 U'V'dy + \lambda \int_0^1 UV dy + \int_0^1 V^2 dy = 0,$$

$$\int_0^1 V'U'dy + \lambda \int_0^1 VU dy = 0.$$

Subtracting the second equation from the first gives us $\int_0^1 V^2 dy = 0$, and therefore $V = 0$. With $\eta = U$ and $V = 0$, (4.54) becomes

$$\int_0^1 (U')^2 dy + \lambda \int_0^1 U^2 dy = 0.$$

Since $\lambda \geq 0$, the last equation implies $\int_0^1 (U')^2 dy = 0$ and hence by the Poincaré inequality $U = 0$.

Since $\{\psi_j(y)\}_{j=1}^{2N}$ of (2.8) is a basis for \mathcal{M}_h^0 , and $\{\psi_j(y)\}_{j=0}^{2N+1}$ of (2.6) is a basis for \mathcal{M}_h , we have

$$U(y) = \sum_{l=1}^{2N} u_l \psi_l(y), \quad V(y) = \sum_{l=0}^{2N+1} v_l \psi_l(y).$$

Substituting these expressions into (4.54) and (4.55), taking $\eta = \psi_j(y)$, $\delta = \psi_j(y)$, and re-arranging the order of the sums in (4.54), we get

$$\sum_{l=0}^{2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_l + \sum_{l=1}^{2N} \int_0^1 [(\psi_j' \psi_l')(y) + \lambda(\psi_j \psi_l)(y)] dy u_l = 0, \quad (4.56)$$

where $j = 0, \dots, 2N + 1$, and

$$\sum_{l=0}^{2N+1} \int_0^1 [(\psi_j' \psi_l')(y) + \lambda(\psi_j \psi_l)(y)] dy v_l = 0, \quad (4.57)$$

where $j = 1, \dots, 2N$. With

$$\vec{v} = [\vec{v}_i, \vec{v}_b]^T, \quad \vec{v}_i = [v_1, \dots, v_{2N}]^T, \quad \vec{v}_b = [v_0, v_{2N+1}]^T, \quad \vec{u} = [u_1, \dots, u_{2N}]^T,$$

the matrix-vector form of (4.54)–(4.55) is

$$B_y \vec{v}_i + B_y^b \vec{v}_b + (\lambda B_y + A_y) \vec{u} = \vec{0}, \quad (4.58)$$

$$(B_y^b)^T \vec{v}_i + B_y^{bb} \vec{v}_b + [\lambda (B_y^b)^T + (A_y^b)^T] \vec{u} = \vec{0}, \quad (4.59)$$

$$(\lambda B_y + A_y) \vec{v}_i + (\lambda B_y^b + A_y^b) \vec{v}_b = \vec{0}. \quad (4.60)$$

Equation (4.58) corresponds to the $2N$ equations of (4.56) with $j = 1, \dots, 2N$. The matrices B_y , B_y^b , and A_y are given by (2.18), (3.23), and (2.17), respectively. Equation (4.59) corresponds to the 2 equations in (4.56) with $j = 0, 2N + 1$. The matrices B_y^{bb} and A_y^b are given by (3.29) and (3.31), respectively. Equation (4.60) corresponds to the $2N$ equations of (4.57) with $j = 1, \dots, 2N$.

Since the only solution to (4.54)–(4.55) is $U = V = 0$, the only solution to (4.58)–(4.60) is $\vec{v}_i = \vec{u} = \vec{0}$, $\vec{v}_b = \vec{0}$. Hence the matrix in the linear system (4.58)–(4.60) is nonsingular. Comparing this matrix with B_k and using the fact that $\lambda_k > 0$ (Lemma 2.1), we conclude that each B_k is nonsingular. \square

In B_k of (4.50), B_y and A_y are $2N \times 2N$ square matrices defined by (2.18) and

(2.17), respectively, with the block tridiagonal structure given in (2.20) and (2.19), respectively. The $2N \times 2$ rectangular matrices B_y^b and A_y^b are defined by (3.23) and (3.31), respectively. Using (2.6), (2.10), and (2.9), we have

$$B_y^b = h \begin{bmatrix} \beta_6 & 0 \\ \beta_3 & 0 \\ \beta_5 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \beta_3 \\ 0 & -\beta_5 \\ 0 & -\beta_6 \end{bmatrix}, \quad A_y^b = h^{-1} \begin{bmatrix} \alpha_5 & 0 \\ \alpha_3 & 0 \\ \alpha_5 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \alpha_3 \\ 0 & -\alpha_5 \\ 0 & -\alpha_5 \end{bmatrix}. \quad (4.61)$$

Similarly by (3.29),

$$B_y^{bb} = h \begin{bmatrix} \beta_1/2 & 0 \\ 0 & \beta_1/2 \end{bmatrix}.$$

Hence B_k has the form

$$\dots, v'_{k,2N-2}, v'_{k,2N-1}, u'_{k,2N-2}, u'_{k,2N-1}, v'_{k,2N+1}, v'_{k,2N}, u'_{k,2N+1}, u'_{k,2N}]^T. \quad (4.62)$$

For the right-hand sides we introduce a $(4N + 4)$ -vector

$$\vec{g}_k = [(b'_2)_{k,0}, (b'_1)_{k,1}, 0, (b'_3)_{k,1}, (b'_1)_{k,2}, (b'_1)_{k,3}, (b'_3)_{k,2}, (b'_3)_{k,3},$$

$$\dots, (b'_1)_{k,2N-2}, (b'_1)_{k,2N-1}, (b'_3)_{k,2N-2}, (b'_3)_{k,2N-1}, (b'_2)_{k,2N+1}, (b'_1)_{k,2N}, 0, (b'_3)_{k,2N}]^T,$$

where the 0's correspond to $u'_{k,0} = u'_{k,2N+1} = 0$. Then (4.49) can be rewritten as

$$C_k \vec{w}_k = \vec{g}_k, \quad (4.63)$$

where the matrix C_k has the following block tridiagonal form

4×4 blocks. The 4×12 blocks of C_k are given by

$$\begin{bmatrix} l_{1,1} & l_{1,2} & l_{1,3}^{(k)} & l_{1,4}^{(k)} & l_{1,5} & 0 & l_{1,7}^{(k)} & 0 & l_{1,9} & l_{1,10} & l_{1,11}^{(k)} & l_{1,12}^{(k)} \\ l_{2,1} & l_{2,2} & l_{2,3}^{(k)} & l_{2,4}^{(k)} & 0 & l_{2,6} & 0 & l_{2,8}^{(k)} & l_{2,9} & l_{2,10} & l_{2,11}^{(k)} & l_{2,12}^{(k)} \\ l_{3,1}^{(k)} & l_{3,2}^{(k)} & 0 & 0 & l_{3,5}^{(k)} & 0 & 0 & 0 & l_{3,9}^{(k)} & l_{3,10}^{(k)} & 0 & 0 \\ l_{4,1}^{(k)} & l_{4,2}^{(k)} & 0 & 0 & 0 & l_{4,6}^{(k)} & 0 & 0 & l_{4,9}^{(k)} & l_{4,10}^{(k)} & 0 & 0 \end{bmatrix},$$

where

$$\begin{aligned} l_{1,1} &= h\beta_3, & l_{1,2} &= -h\beta_5, & l_{1,3}^{(k)} &= \lambda_k h\beta_3 + h^{-1}\alpha_3, \\ l_{1,4}^{(k)} &= -\lambda_k h\beta_5 - h^{-1}\alpha_5, & l_{1,5} &= h\beta_1, & l_{1,7}^{(k)} &= \lambda_k h\beta_1 + h^{-1}\alpha_1, \\ l_{1,9} &= h\beta_3, & l_{1,10} &= h\beta_5, & l_{1,11}^{(k)} &= \lambda_k h\beta_3 + h^{-1}\alpha_3, \\ l_{1,12}^{(k)} &= \lambda_k h\beta_5 + h^{-1}\alpha_5, & l_{2,1} &= h\beta_5, & l_{2,2} &= h\beta_4, \\ l_{2,3}^{(k)} &= \lambda_k h\beta_5 + h^{-1}\alpha_5, & l_{2,4}^{(k)} &= \lambda_k h\beta_4 + h^{-1}\alpha_4, & l_{2,6} &= h\beta_2, \\ l_{2,8}^{(k)} &= \lambda_k h\beta_2 + h^{-1}\alpha_2, & l_{2,9} &= -h\beta_5, & l_{2,10} &= h\beta_4, \\ l_{2,11}^{(k)} &= -\lambda_k h\beta_5 - h^{-1}\alpha_5, & l_{2,12}^{(k)} &= \lambda_k h\beta_4 + h^{-1}\alpha_4, & l_{3,1}^{(k)} &= \lambda_k h\beta_3 + h^{-1}\alpha_3, \\ l_{3,2}^{(k)} &= -\lambda_k h\beta_5 - h^{-1}\alpha_5, & l_{3,5}^{(k)} &= \lambda_k h\beta_1 + h^{-1}\alpha_1, & l_{3,9}^{(k)} &= \lambda_k h\beta_3 + h^{-1}\alpha_3, \\ l_{3,10}^{(k)} &= \lambda_k h\beta_5 + h^{-1}\alpha_5, & l_{4,1}^{(k)} &= \lambda_k h\beta_5 + h^{-1}\alpha_5, & l_{4,2}^{(k)} &= \lambda_k h\beta_4 + h^{-1}\alpha_4, \\ l_{4,6}^{(k)} &= \lambda_k h\beta_2 + h^{-1}\alpha_2, & l_{4,9}^{(k)} &= -\lambda_k h\beta_5 - h^{-1}\alpha_5, & l_{4,10}^{(k)} &= \lambda_k h\beta_4 + h^{-1}\alpha_4. \end{aligned}$$

The top 4×8 block of (4.64) is

$$\begin{bmatrix} l_{1,5}/2 & l_{1,6} & l_{1,7}^{(k)}/2 & l_{1,8}^{(k)} & l_{1,9} & l_{1,10} & l_{1,11}^{(k)} & l_{1,12}^{(k)} \\ l_{2,5} & l_{2,6}/2 & l_{2,7}^{(k)} & l_{2,8}^{(k)} & l_{2,9} & l_{2,10} & l_{2,11}^{(k)} & l_{2,12}^{(k)} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ l_{4,5}^{(k)} & l_{4,6}/2 & 0 & 0 & l_{4,9}^{(k)} & l_{4,10}^{(k)} & 0 & 0 \end{bmatrix},$$

where

$$\begin{aligned} l_{1,6} &= \beta_6, & l_{1,8}^{(k)} &= \lambda_k \beta_6 + \alpha_5, & l_{2,5} &= \beta_6, \\ l_{2,7}^{(k)} &= \lambda_k \beta_6 + \alpha_5, & l_{4,5}^{(k)} &= \lambda_k \beta_6 + \alpha_5. \end{aligned}$$

The bottom 4×8 block of (4.64) is

$$\begin{bmatrix} l_{1,1} & l_{1,2} & l_{1,3}^{(k)} & l_{1,4}^{(k)} & l_{1,5}/2 & -l_{1,6} & l_{1,7}^{(k)}/2 & -l_{1,8}^{(k)} \\ l_{2,1} & l_{2,2} & l_{2,3}^{(k)} & l_{2,4}^{(k)} & -l_{2,5} & l_{2,6}/2 & -l_{2,7}^{(k)} & l_{2,8}^{(k)}/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ l_{4,1}^{(k)} & l_{4,2}^{(k)} & 0 & 0 & -l_{4,5}^{(k)} & l_{4,6}^{(k)}/2 & 0 & 0 \end{bmatrix}.$$

Thus we obtain the following algorithm:

ALGORITHM FOR SOLVING $M_{11}\vec{w} = \vec{b}$

1. Compute \vec{b}_1 , \vec{b}_2 , and \vec{b}_3 , using (4.39)–(4.41). (4.65)
2. For $k = 1, \dots, 2N$, solve (4.63) for \vec{w}_k .
3. Compute \vec{v}_i , \vec{v}_h , and \vec{u} using (4.36)–(4.38).

We can now discuss implementation and cost of this algorithm. Because of (2.72), we can use FFT routines to multiply a vector by Z^T in step 1 at a cost of $O(N \log_2 N)$; $4N + 2$ such multiplications give a total cost for step 1 of $O(N^2 \log_2 N)$.

In the same way, FFT routines can be used to perform the multiplications by Z in step 3. Therefore the cost of step 3 is also $O(N^2 \log_2 N)$. In step 2, each $(4N + 4) \times (4N + 4)$ block tridiagonal system (4.63) can be solved effectively at a cost $O(N)$ using LAPACK [1] routine DGBTRS which implements band Gauss elimination. Therefore the cost of step 2 is $O(N^2)$. Hence total cost of the algorithm for solving $M_{11} \vec{w} = \vec{b}$ is $O(N^2 \log_2 N)$.

4.3 Related Biharmonic Problem

Bjørstad observed in [2] that for the finite difference discretization the biharmonic problem with Δu rather than $\partial u / \partial n$ specified on the two vertical sides of $\partial\Omega$ one can use separation of variables in the x -direction. We show that in our case solving $M_{11} \vec{v} = \vec{b}$ is equivalent to finding the Galerkin solution of the following related biharmonic problem I:

$$\left. \begin{aligned} v - \Delta u &= g(x, y) \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega_h, \end{aligned} \right\} \quad (4.66)$$

and

$$\left. \begin{aligned} -\Delta v &= -f(x, y) \text{ in } \Omega, \\ v &= 0 \text{ on } \partial\Omega_v, \end{aligned} \right\} \quad (4.67)$$

where $\partial\Omega_h$ is the union of the two horizontal sides of $\partial\Omega$, and $\partial\Omega_v$ is the union of the two vertical sides of $\partial\Omega$ (cf. Figure 4.1).

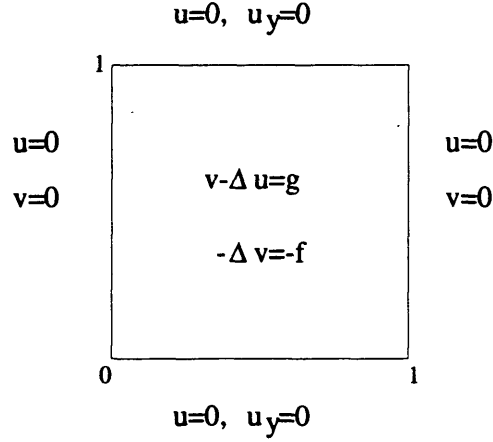


Figure 4.1: Related Biharmonic Problem I.

Assume u and v are sufficiently smooth and satisfy (4.66) and (4.67). Then the weak form of (4.66) and (4.67) becomes (cf. (3.3) and (3.4))

$$\int_{\Omega} \nabla u \cdot \nabla \eta \, dx \, dy + \int_{\Omega} v \eta \, dx \, dy = \int_{\Omega} g(x, y) \eta \, dx \, dy, \quad \eta \in H^1(\Omega), \quad \eta = 0 \text{ on } \partial\Omega_v, \quad (4.68)$$

$$\int_{\Omega} \nabla v \cdot \nabla \delta \, dx \, dy = - \int_{\Omega} f(x, y) \delta \, dx \, dy, \quad \delta \in H_0^1(\Omega). \quad (4.69)$$

The Galerkin solution of (4.66)–(4.67) consists of $U \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0$, $V \in \mathcal{M}_h^0 \otimes \mathcal{M}_h$ such that

$$\int_{\Omega} \nabla U \cdot \nabla \eta \, dx \, dy + \int_{\Omega} V \eta \, dx \, dy = \int_{\Omega} g(x, y) \eta \, dx \, dy, \quad \eta \in \mathcal{M}_h^0 \otimes \mathcal{M}_h, \quad (4.70)$$

$$\int_{\Omega} \nabla V \cdot \nabla \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0. \quad (4.71)$$

Using $\{\phi_i(x)\psi_j(y)\}_{i=1, j=1}^{2N, 2N}$ as a basis for $\mathcal{M}_h^0 \otimes \mathcal{M}_h^0$ and $\{\phi_i(x)\psi_j(y)\}_{i=1, j=0}^{2N, 2N+1}$ as a basis for $\mathcal{M}_h^0 \otimes \mathcal{M}_h$, we may write

$$U(x, y) = \sum_{k=1}^{2N} \sum_{l=1}^{2N} u_{kl} \phi_k(x) \psi_l(y),$$

$$V(x, y) = \sum_{k=1}^{2N} \sum_{l=0}^{2N+1} v_{kl} \phi_k(x) \psi_l(y).$$

Substituting these expressions into (4.70) and (4.71), taking $\eta = \phi_i(x)\psi_j(y)$, $\delta = \phi_i(x)\psi_j(y)$, using Fubini's theorem and rearranging the order of double sums involving $u_{k,l}$ and $v_{k,l}$ we obtain (cf. (3.9) and (3.10))

$$\begin{aligned} & \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\ & + \sum_{k=1}^{2N} \int_0^1 (\phi_i' \phi_k')(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy u_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j' \psi_l')(y) dy u_{kl} \\ & = \int_0^1 \left[\int_0^1 g(x, y) \phi_i(x) dx \right] \psi_j(y) dy, \end{aligned} \quad (4.72)$$

where $i, j = 1, \dots, 2N$,

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy u_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy u_{kl} \\
& = \int_0^1 \left[\int_0^1 g(x, y) \phi_i(x) dx \right] \psi_j(y) dy, \tag{4.73}
\end{aligned}$$

where $i = 1, \dots, 2N$, $j = 0, 2N + 1$, and

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{kl} \\
& = - \int_0^1 \left[\int_0^1 f(x, y) \phi_i(x) dx \right] \psi_j(y) dy, \tag{4.74}
\end{aligned}$$

where $i, j = 1, \dots, 2N$.

With \vec{v}_i , \vec{v}_h , and \vec{u} defined as in (3.11), (3.12), and (3.14), the matrix-vector

form of (4.72) and (4.74) is (cf. derivation of (3.17)–(3.20))

$$A_{11}\vec{v}_i + A_{12}\vec{v}_h + A_{13}\vec{u} = \vec{b}_1, \quad (4.75)$$

$$A_{21}\vec{v}_i + A_{22}\vec{v}_h + A_{23}\vec{u} = \vec{b}_2, \quad (4.76)$$

$$A_{31}\vec{v}_i + A_{32}\vec{v}_h = \vec{b}_3, \quad (4.77)$$

where A_{11} , A_{12} , A_{13} , A_{21} , A_{22} , A_{23} , A_{31} , and A_{32} are as in (3.17)–(3.19), and

$$\vec{b}_1 = [(b_1)_{1,1}, \dots, (b_1)_{1,2N}, (b_1)_{2,1}, \dots, (b_1)_{2,2N}, \dots, (b_1)_{2N,1}, \dots, (b_1)_{2N,2N}]^T,$$

$$\vec{b}_2 = [(b_2)_{1,0}, (b_2)_{1,2N+1}, (b_2)_{2,0}, (b_2)_{2,2N+1}, \dots, (b_2)_{2N-1,0}, (b_2)_{2N-1,2N+1}, (b_2)_{2N,0}, (b_2)_{2N,2N+1}]^T.$$

$$\vec{b}_3 = [(b_3)_{1,1}, \dots, (b_3)_{1,2N}, (b_3)_{2,1}, \dots, (b_3)_{2,2N}, \dots, (b_3)_{2N,1}, \dots, (b_3)_{2N,2N}]^T,$$

with

$$(b_1)_{i,j} = \int_0^1 \left[\int_0^1 g(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i, j = 1, \dots, 2N,$$

$$(b_2)_{i,j} = \int_0^1 \left[\int_0^1 g(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i = 1, \dots, 2N, \quad j = 0, 2N+1,$$

$$(b_3)_{i,j} = - \int_0^1 \left[\int_0^1 f(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i, j = 1, \dots, 2N.$$

Clearly the matrix in (4.75)–(4.77) is the same as M_{11} of (4.3).

Lemma 4.3 M_{11} is nonsingular.

Proof. The Galerkin solution of (4.70)–(4.71) is unique (proof similar to that of Lemma 3.1). Hence the matrix of (4.75)–(4.77) and also M_{11} are nonsingular. \square

4.4 Solving the Schur Complement System

To solve (4.6), it is possible to form the matrix S of (4.7) explicitly and then compute \vec{v}_v using, for example, Cholesky decomposition. However this would be too expensive at a cost of $O(N^3)$. Instead we seek an iterative scheme for solving (4.6). Since we know that S is symmetric and positive definite, a good candidate for solving (4.6) would be the PCG method described in [13]. Since every iteration of the PCG method requires multiplying S with an arbitrary vector, subsection 1 is concerned with the computational cost of this operation. Subsection 2 is devoted to the definition and properties of the preconditioner. Subsection 3 deals with solving the preconditioned system. Subsection 4 summarizes the cost of solving the Schur complement system.

4.4.1 Computing $S\vec{z}_v$

We discuss computation of $S\vec{z}_v$ for arbitrary

$$\vec{z}_v = [z_{0,1}, \dots, z_{0,2N}, z_{2N+1,1}, \dots, z_{2N+1,2N}]^T.$$

Using (4.7), we first consider $A_{44}\vec{z}_v$, which, by (3.40) and (2.24), can be written as

$$A_{44}\vec{z}_v = (B_x^{bb} \otimes I_{2N})(I_2 \otimes B_y)\vec{z}_v.$$

The computation of $(I_2 \otimes B_y)\vec{z}_v$ requires two matrix-vector multiplications of B_y with the two subvectors of \vec{z}_v , which by (2.20) can be accomplished at a cost of $O(N)$. Then multiplication by $B_x^{bb} \otimes I_{2N}$ involves $2N$ multiplications by the 2×2 matrix B_x^{bb} of (3.41), which requires $O(N)$ operations. Therefore the total cost of computing $A_{44}\vec{z}_v$ is $O(N)$.

By (4.7), each multiplication by S also requires computing $M_{21}M_{11}^{-1}M_{12}\vec{z}_v$, which is equivalent to computing $\vec{b} = M_{12}\vec{z}_v$, solving $M_{11}\vec{w} = \vec{b}$ for \vec{w} , and then computing $M_{21}\vec{w}$. We show that the unique structure associated with each of these operations insures a cheaper cost than expected.

First, let

$$\vec{b} = M_{12}\vec{z}_v = [\vec{b}_1, \vec{b}_2, \vec{b}_3]^T \tag{4.78}$$

with \vec{b}_1 , \vec{b}_2 , and \vec{b}_3 of the form (4.27), (4.28), and (4.29), respectively. Then by (4.3), (3.25), (3.32), and (3.35), we have

$$\vec{b}_1 = A_{14}\vec{z}_v = (B_x^b \otimes I_{2N})(I_2 \otimes B_y)\vec{z}_v, \tag{4.79}$$

$$\vec{b}_2 = A_{24}\vec{z}_v = (B_x^b \otimes I_2)(I_2 \otimes (B_y^b)^T)\vec{z}_v, \tag{4.80}$$

$$\vec{b}_3 = A_{34}\vec{z}_v = (A_x^b \otimes I_{2N})(I_2 \otimes B_y)\vec{z}_v + (B_x^b \otimes I_{2N})(I_2 \otimes A_y)\vec{z}_v. \quad (4.81)$$

To determine the cost of computing \vec{b}_1 and \vec{b}_3 , we observe that the computation of $(I_2 \otimes B_y)\vec{z}_v$ and $(I_2 \otimes A_y)\vec{z}_v$ requires two matrix-vector multiplications with B_y and A_y , respectively, which by (2.20) and (2.19), can be accomplished at a cost of $O(N)$. By (3.26), (3.36), (2.5), and (2.10), we know that

$$B_x^b = h \begin{bmatrix} \beta_3 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \beta_3 \\ \beta_6 & 0 \\ \beta_5 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & -\beta_5 \\ 0 & -\beta_6 \end{bmatrix}, \quad A_x^b = h^{-1} \begin{bmatrix} \alpha_3 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \alpha_3 \\ \alpha_5 & 0 \\ \alpha_5 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & -\alpha_5 \\ 0 & -\alpha_5 \end{bmatrix}, \quad (4.82)$$

where non-zero elements in each matrix are located in the first column at rows 1, N , and $N + 1$, while non-zero elements in the second column are located in rows $N - 1$, $2N - 1$, and $2N$. Since multiplication by $B_x^b \otimes I_{2N}$ and $A_x^b \otimes I_{2N}$ involves $2N$ multiplications by B_x^b and A_x^b , respectively, from (4.82), this will require $O(N)$ operations. Therefore the total cost of the computation of \vec{b}_1 and \vec{b}_3 is $O(N)$.

To determine the cost of computing \vec{b}_2 , we know the computation of $(I_2 \otimes (B_y^b)^T)\vec{z}_v$ requires two matrix-vector multiplications by $(B_y^b)^T$ which from (4.61) can be accomplished at a cost of $O(1)$. Again, we know that multiplication by $B_x^b \otimes I_2$

involves 2 multiplications by B_z^b , which from (4.82), will require $O(1)$ operations. Total cost for computation of \vec{b}_2 is $O(1)$. Therefore the total cost of computing \vec{b} is $O(N)$.

Solving $M_{11}\vec{w} = \vec{b}$ for \vec{w} involves \vec{b} of (4.78). By (4.79)–(4.81), and (4.82), the only possible non-zero elements of \vec{b}_1 and \vec{b}_3 are

$$(b_1)_{i,j}, (b_3)_{i,j}, \quad i = 1, N-1, N, N+1, 2N-1, 2N, \quad j = 1, \dots, 2N, \quad (4.83)$$

and the only possible non-zero elements of \vec{b}_2 are

$$(b_2)_{i,j}, \quad i = 1, N-1, N, N+1, 2N-1, 2N, \quad j = 0, 2N+1. \quad (4.84)$$

Recall that after computing \vec{w} we must calculate $M_{21}\vec{w}$. Because of the special structure of M_{21} we will only require certain components of \vec{w} when solving $M_{11}\vec{w} = \vec{b}$. To determine which components of \vec{w} are needed, let us examine first the computation of $M_{21}\vec{w}$. By (4.3) with \vec{w} as in (4.4), we have

$$M_{21}\vec{w} = A_{41}\vec{v}_i + A_{42}\vec{v}_h + A_{43}\vec{u},$$

where \vec{v}_i , \vec{v}_h , and \vec{u} are of the form (3.11), (3.12), and (3.14), respectively. By (3.37),

(3.38), and (3.39), we may write

$$A_{41}\vec{v}_i = (I_{2N} \otimes B_y)((B_x^b)^T \otimes I_{2N})\vec{v}_i, \quad (4.85)$$

$$A_{42}\vec{v}_h = (I_{2N} \otimes B_y^b)((B_x^b)^T \otimes I_{2N})\vec{v}_h, \quad (4.86)$$

$$A_{43}\vec{u} = (I_{2N} \otimes B_y)((A_x^b)^T \otimes I_{2N})\vec{u} + (I_{2N} \otimes A_y)((B_x^b)^T \otimes I_{2N})\vec{u}. \quad (4.87)$$

We observe that in order to calculate $A_{41}\vec{v}_i$ and $A_{43}\vec{u}$, multiplication by $(B_x^b)^T \otimes I_{2N}$ and $(A_x^b)^T \otimes I_{2N}$ each will involve $2N$ multiplications by $(B_x^b)^T$ and $(A_x^b)^T$, respectively. But by (4.82), (3.11), and (3.14), to perform these multiplications we require only knowledge of the elements

$$v_{i,j}, \quad u_{i,j}, \quad i = 1, N-1, N, N+1, 2N-1, 2N, \quad j = 1, \dots, 2N. \quad (4.88)$$

In the same way, to calculate $A_{42}\vec{v}_h$ we observe that the multiplication by $(B_x^b)^T \otimes I_{2N}$ will involve $2N$ multiplications by $(B_x^b)^T$. But, by (4.82) and (3.12), we require only knowledge of the elements

$$v_{i,j}, \quad i = 1, N-1, N, N+1, 2N-1, 2N, \quad j = 0, 2N+1. \quad (4.89)$$

Hence we see, that when solving $M_{11}\vec{w} = \vec{b}$ for \vec{w} , we require only the components in (4.88) and (4.89). As a result, we can now discuss the reduced cost of solving

$M_{11}\vec{w} = \vec{b}$ using Algorithm (4.65). In step 1 of this algorithm, the operations (4.39) and (4.41) each consist of $2N$ matrix-vector multiplications with Z^T , where by (4.83), for $k = 1, 3, j = 1, \dots, 2N$,

$$\begin{bmatrix} (b'_k)_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ (b'_k)_{2N,j} \end{bmatrix} = Z^T \begin{bmatrix} (b_k)_{1,j} \\ 0 \\ \vdots \\ 0 \\ (b_k)_{N-1,j} \\ (b_k)_{N,j} \\ (b_k)_{N+1,j} \\ 0 \\ \vdots \\ 0 \\ (b_k)_{2N-1,j} \\ (b_k)_{2N,j} \end{bmatrix}. \quad (4.90)$$

Then (4.90) can be accomplished with direct matrix-vector multiplication without FFTs, being computations of linear combinations of columns 1, $N-1$, N , $N+1$, $2N-1$, $2N$ of Z^T . Therefore the resulting cost is $O(N^2)$. By (4.84), operation (4.40) is (4.90) with $k = 2$ and $j = 0, 2N+1$, hence the total cost for this step is $O(N)$. Therefore all of step 1 of Algorithm (4.65) can be accomplished at a cost of $O(N^2)$. Calculations in step 2 of Algorithm (4.65) remain unchanged from those required for an arbitrary right-hand side \vec{b} , which gives a cost of $O(N^2)$. In step 3 of Algorithm (4.65), the operations (4.36) and (4.38) each consist of $2N$ multiplications of Z by subvectors of \vec{u}'_i and \vec{u}' , respectively. This is illustrated in (4.91).

$$\begin{bmatrix} v_{1,j} \\ \vdots \\ \vdots \\ v_{N-1,j} \\ v_{N,j} \\ v_{N+1,j} \\ \vdots \\ \vdots \\ v_{2N-1,j} \\ v_{2N,j} \end{bmatrix} = Z \begin{bmatrix} v'_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ v'_{2N,j} \end{bmatrix}, \quad \begin{bmatrix} u_{1,j} \\ \vdots \\ \vdots \\ u_{N-1,j} \\ u_{N,j} \\ u_{N+1,j} \\ \vdots \\ \vdots \\ u_{2N-1,j} \\ u_{2N,j} \end{bmatrix} = Z \begin{bmatrix} u'_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ u'_{2N,j} \end{bmatrix}, \quad j = 1, \dots, 2N. \quad (4.91)$$

Since only the components in (4.88) need to be evaluated, we compute the inner products involving rows 1, $N - 1$, N , $N + 1$, $2N - 1$, and $2N$ of Z during each of the $2N$ multiplications. The resulting cost is then $O(N^2)$. Operation (4.37) consists of two matrix-vector multiplications of Z by subvectors of \vec{v}_h . This is illustrated in (4.92).

$$\begin{bmatrix} v_{1,j} \\ \vdots \\ \vdots \\ v_{N-1,j} \\ v_{N,j} \\ v_{N+1,j} \\ \vdots \\ \vdots \\ v_{2N-1,j} \\ v_{2N,j} \end{bmatrix} = Z \begin{bmatrix} v'_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ v'_{2N,j} \end{bmatrix}, \quad j = 0, 2N + 1. \quad (4.92)$$

Since only the components in (4.89) need to be evaluated, we compute the inner

products involving rows $1, N-1, N, N+1, 2N-1$, and $2N$ of Z during each of the two multiplications. Therefore the total cost for this operation is $O(N)$, and step 3 of Algorithm (4.65) can be accomplished at a cost of $O(N^2)$. Hence the total cost of solving $M_{11}\vec{w} = \vec{b}$ for \vec{w} is $O(N^2)$.

Once the required components of \vec{w} are obtained from solving $M_{11}\vec{w} = \vec{b}$ for \vec{w} , the multiplication $M_{21}\vec{w}$ is carried out. By (4.85)–(4.87), this requires multiplications by $(B_x^b)^T \otimes I_{2N}$ and $(A_x^b)^T \otimes I_{2N}$ which involve $2N$ matrix-vector multiplications by $(B_x^b)^T$ and $(A_x^b)^T$, respectively. By (4.82), this can be accomplished at a cost of $O(N)$. For (4.85) and (4.87), multiplication by $I_{2N} \otimes B_y$ and $I_{2N} \otimes A_y$ involves $2N$ multiplications by B_y and A_y , respectively, which by (2.20) and (2.19) requires $O(N^2)$ operations. For (4.86), multiplication by $I_{2N} \otimes B_y^b$ involves $2N$ multiplications by B_y^b , which by (4.61) incurs a cost of $O(N)$. Hence total cost of computing $M_{21}\vec{w}$ is $O(N^2)$.

We have shown that computing $M_{21}M_{11}^{-1}M_{12}\vec{z}_v$ has a cost of $O(N^2)$, and that $A_{44}\vec{z}_v$ costs $O(N)$. Hence the cost of computing $S\vec{z}_v$ is $O(N^2)$.

4.4.2 Preconditioner and its properties

Next we discuss the selection and properties of the preconditioner for the solution of (4.6) by the PCG method. We obtain a preconditioner for S from the

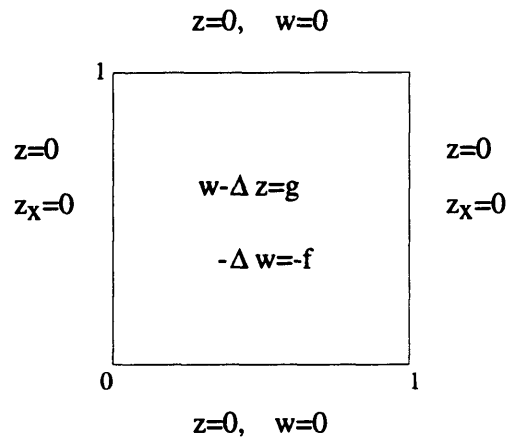


Figure 4.2: Related Biharmonic Problem II.

discretization of the related biharmonic problem II (cf. Figure 4.2):

$$\left. \begin{aligned}
 w - \Delta z &= g(x, y) \text{ in } \Omega, \\
 z &= 0 \text{ on } \partial\Omega, \\
 \frac{\partial z}{\partial n} &= 0 \text{ on } \partial\Omega_v,
 \end{aligned} \right\} \quad (4.93)$$

and

$$\left. \begin{aligned}
 -\Delta w &= -f(x, y) \text{ in } \Omega, \\
 w &= 0 \text{ on } \partial\Omega_h.
 \end{aligned} \right\} \quad (4.94)$$

Assume z and w are sufficiently smooth and satisfy (4.93) and (4.94). Then the weak

form of (4.93) and (4.94) becomes (cf. (3.3)–(3.4) and (4.68)–(4.69))

$$\int_{\Omega} \nabla z \cdot \nabla \eta dx dy + \int_{\Omega} w \eta dx dy = \int_{\Omega} g(x, y) \eta dx dy, \quad \eta \in H^1(\Omega), \quad \eta = 0 \text{ on } \partial\Omega_h,$$

$$\int_{\Omega} \nabla w \cdot \nabla \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in H_0^1(\Omega).$$

The Galerkin solution of (4.93)–(4.94) consists of $Z \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0$, $W \in \mathcal{M}_h \otimes \mathcal{M}_h^0$, such that

$$\int_{\Omega} \nabla Z \cdot \nabla \eta dx dy + \int_{\Omega} W \eta dx dy = \int_{\Omega} g(x, y) \eta dx dy, \quad \eta \in \mathcal{M}_h \otimes \mathcal{M}_h^0, \quad (4.95)$$

$$\int_{\Omega} \nabla W \cdot \nabla \delta dx dy = - \int_{\Omega} f(x, y) \delta dx dy, \quad \delta \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0. \quad (4.96)$$

Using $\{\phi_i(x)\psi_j(y)\}_{i=1, j=1}^{2N, 2N}$ as a basis for $\mathcal{M}_h^0 \otimes \mathcal{M}_h^0$ and $\{\phi_i(x)\psi_j(y)\}_{i=0, j=1}^{2N+1, 2N}$ as a basis for $\mathcal{M}_h \otimes \mathcal{M}_h^0$, we may write

$$Z(x, y) = \sum_{k=1}^{2N} \sum_{l=1}^{2N} z_{kl} \phi_k(x) \psi_l(y),$$

$$W(x, y) = \sum_{k=0}^{2N+1} \sum_{l=1}^{2N} w_{kl} \phi_k(x) \psi_l(y).$$

Substituting these expressions into (4.95) and (4.96), taking $\eta = \phi_i(x)\psi_j(y)$, $\delta = \phi_i(x)\psi_j(y)$, using Fubini's theorem, and rearranging the order of the double sums

involving z_{kl} and w_{kl} , $k, l = 1, \dots, 2N$, we obtain (cf. (3.9)–(3.10))

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy z_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy z_{kl} \\
& + \sum_{k=0, 2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} = \int_0^1 \left[\int_0^1 g(x, y) \phi_i(x) dx \right] \psi_j(y) dy,
\end{aligned} \tag{4.97}$$

where $i, j = 1, \dots, 2N$,

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy z_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy z_{kl} \\
& + \sum_{k=0, 2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} = \int_0^1 \left[\int_0^1 g(x, y) \phi_i(x) dx \right] \psi_j(y) dy,
\end{aligned} \tag{4.98}$$

where $i = 0, 2N + 1$, $j = 1, \dots, 2N$ and

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy w_{kl} \\
& + \sum_{k=0, 2N+1} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{k=0,2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy w_{kl} \\
& = - \int_0^1 \left[\int_0^1 f(x, y) \phi_i(x) dx \right] \psi_j(y) dy,
\end{aligned} \tag{4.99}$$

where $i, j = 1, \dots, 2N$.

Introduce

$$\vec{w}_i = [w_{1,1}, \dots, w_{1,2N}, w_{2,1}, \dots, w_{2,2N}, \dots, w_{2N,1}, \dots, w_{2N,2N}]^T, \tag{4.100}$$

$$\vec{w}_v = [w_{0,1}, \dots, w_{0,2N}, w_{2N+1,1}, \dots, w_{2N+1,2N}]^T, \tag{4.101}$$

and

$$\vec{z} = [z_{1,1}, \dots, z_{1,2N}, z_{2,1}, \dots, z_{2,2N}, \dots, z_{2N,1}, \dots, z_{2N,2N}]^T. \tag{4.102}$$

Then the matrix-vector form of (4.97)–(4.99) is (cf. derivation of (3.17)–(3.20))

$$A_{11} \vec{w}_i + A_{13} \vec{z} + A_{14} \vec{w}_v = \vec{c}_1, \tag{4.103}$$

$$A_{31} \vec{w}_i + A_{34} \vec{w}_v = \vec{c}_3, \tag{4.104}$$

$$A_{41} \vec{w}_i + A_{43} \vec{z} + A_{44} \vec{w}_v = \vec{c}_4, \tag{4.105}$$

where A_{11} , A_{13} , A_{14} , A_{31} , A_{34} , A_{41} , A_{43} , and A_{44} are as in (3.17), (3.19), and (3.20),

and

$$\begin{aligned}\vec{c}_1 &= [(c_1)_{1,1}, \dots, (c_1)_{1,2N}, (c_1)_{2,1}, \dots, (c_1)_{2,2N}, \dots, (c_1)_{2N,1}, \dots, (c_1)_{2N,2N}]^T, \\ \vec{c}_3 &= [(c_3)_{1,1}, \dots, (c_3)_{1,2N}, (c_3)_{2,1}, \dots, (c_3)_{2,2N}, \dots, (c_3)_{2N,1}, \dots, (c_3)_{2N,2N}]^T, \\ \vec{c}_4 &= [(c_4)_{0,1}, \dots, (c_4)_{0,2N}, (c_4)_{2N+1,1}, \dots, (c_4)_{2N+1,2N}]^T,\end{aligned}\quad (4.106)$$

with

$$\begin{aligned}(c_1)_{i,j} &= \int_0^1 \left[\int_0^1 g(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i, j = 1, \dots, 2N, \\ (c_3)_{i,j} &= - \int_0^1 \left[\int_0^1 f(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i, j = 1, \dots, 2N, \\ (c_4)_{i,j} &= \int_0^1 \left[\int_0^1 g(x,y) \phi_i(x) dx \right] \psi_j(y) dy, \quad i = 0, 2N+1, \quad j = 1, \dots, 2N.\end{aligned}$$

We can write (4.103)–(4.105) as

$$B_{11}\vec{e} + B_{12}\vec{w}_v = \vec{d}, \quad (4.107)$$

$$B_{21}\vec{e} + A_{44}\vec{w}_v = \vec{c}_4, \quad (4.108)$$

where

$$B_{11} = \begin{bmatrix} A_{11} & A_{13} \\ A_{31} & 0 \end{bmatrix}, \quad B_{12} = \begin{bmatrix} A_{14} \\ A_{34} \end{bmatrix}, \quad B_{21} = \begin{bmatrix} A_{41} & A_{43} \end{bmatrix}, \quad (4.109)$$

and

$$\vec{e} = \begin{bmatrix} \vec{w}_i \\ \vec{z} \end{bmatrix}, \quad \vec{d} = \begin{bmatrix} \vec{c}_1 \\ \vec{c}_3 \end{bmatrix}.$$

Lemma 4.4 *The matrix B_{11} of (4.109) is nonsingular.*

Proof. Consider

$$A_{11}\vec{w}_i + A_{13}\vec{z} = \vec{0}, \quad (4.110)$$

$$A_{31}\vec{w}_i = \vec{0}, \quad (4.111)$$

where A_{11} , A_{13} , and A_{31} , are as in (4.103) and (4.104). This is the matrix-vector form of the Galerkin problem

$$\int_{\Omega} \nabla Z \cdot \nabla \eta dx dy + \int_{\Omega} W \eta dx dy = 0, \quad \eta \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0, \quad (4.112)$$

$$\int_{\Omega} \nabla W \cdot \nabla \delta dx dy = 0, \quad \delta \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0, \quad (4.113)$$

where $Z \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0$, $W \in \mathcal{M}_h^0 \otimes \mathcal{M}_h^0$. The only solution of the Galerkin problem (4.112)–(4.113) is $Z = W = 0$ (proof similar to that of Lemma 3.1), therefore the only solution of (4.110)–(4.111) is $\vec{w}_i = \vec{z} = \vec{0}$. Hence the coefficient matrix of (4.110)–(4.111), B_{11} , is nonsingular. \square

Solving (4.107)–(4.108) for \vec{w}_v (cf. (4.6)), we obtain the Schur complement

system

$$P\vec{w}_v = \vec{c}_4 - B_{21}B_{11}^{-1}\vec{d}, \quad (4.114)$$

where

$$P = A_{44} - B_{21}B_{11}^{-1}B_{12} \quad (4.115)$$

is the Schur complement of B_{11} in

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & A_{44} \end{bmatrix}.$$

As a preconditioner for S in the Schur complement system (4.6), we select P of (4.115).

Lemma 4.5 *The matrix P of (4.115) is symmetric and positive definite.*

Proof. It follows from (4.109), (4.9), (4.11), and (4.12), that

$$B_{11} = B_{11}^T, \quad B_{12} = B_{21}^T. \quad (4.116)$$

Therefore, by (4.115), (4.10), and (4.116), we have that P is symmetric.

To show that P is positive definite, let us consider any fixed, non-zero $\vec{w}_v \in$

\mathbb{R}^{4N} of the form (4.101). Let \vec{w}_i and \vec{z} of the form (4.100) and (4.102) be such that

$$A_{11}\vec{w}_i + A_{13}\vec{z} + A_{14}\vec{w}_v = \vec{0}, \quad (4.117)$$

$$A_{31}\vec{w}_i + A_{34}\vec{w}_v = \vec{0}, \quad (4.118)$$

where A_{11} , A_{13} , A_{14} , A_{31} , and A_{34} , are as in (4.103) and (4.104). Then by (4.109), (4.117)–(4.118) is equivalent to

$$B_{11} \begin{bmatrix} \vec{w}_i \\ \vec{z} \end{bmatrix} = -B_{12}\vec{w}_v. \quad (4.119)$$

Since B_{11} is nonsingular it follows that

$$\begin{bmatrix} \vec{w}_i \\ \vec{z} \end{bmatrix} = -B_{11}^{-1}B_{12}\vec{w}_v. \quad (4.120)$$

Now, by (4.115), (4.116), (4.120), (4.109), (4.118), (4.11)–(4.12), and (4.117),

$$\begin{aligned} (P\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} &= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} - (B_{21}B_{11}^{-1}B_{12}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} \\ &= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} - (B_{11}^{-1}B_{12}\vec{w}_v, B_{12}\vec{w}_v)_{\mathbb{R}^{4N}} \\ &= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} + ([\vec{w}_i, \vec{z}]^T, [A_{14}, A_{34}]^T \vec{w}_v)_{\mathbb{R}^{4N}} \end{aligned}$$

$$\begin{aligned}
&= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} + (\vec{w}_i, A_{14}\vec{w}_v)_{\mathbb{R}^{4N}} + (\vec{z}, A_{34}\vec{w}_v)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} + (\vec{w}_i, A_{14}\vec{w}_v)_{\mathbb{R}^{4N}} - (\vec{z}, A_{31}\vec{w}_i)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} + (A_{41}\vec{w}_i, \vec{w}_v)_{\mathbb{R}^{4N}} - (A_{13}\vec{z}, \vec{w}_i)_{\mathbb{R}^{4N}} \\
&= (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} + (A_{41}\vec{w}_i, \vec{w}_v)_{\mathbb{R}^{4N}} + (A_{11}\vec{w}_i, \vec{w}_i)_{\mathbb{R}^{4N}} + (A_{14}\vec{w}_v, \vec{w}_i)_{\mathbb{R}^{4N}}. \quad (4.121)
\end{aligned}$$

Introduce

$$v = \tilde{w}_i + \tilde{w}_v,$$

where

$$\tilde{w}_i = \sum_{k=1}^{2N} \sum_{l=1}^{2N} w_{k,l} \phi_k(x) \psi_l(y), \quad (4.122)$$

$$\tilde{w}_v = \sum_{k=0, 2N+1}^{2N} \sum_{l=1}^{2N} w_{k,l} \phi_k(x) \psi_l(y). \quad (4.123)$$

Consider

$$\begin{aligned}
(v, v)_{L^2(\Omega)} &= (\tilde{w}_i + \tilde{w}_v, \tilde{w}_i + \tilde{w}_v)_{L^2(\Omega)} \\
&= (\tilde{w}_i, \tilde{w}_i)_{L^2(\Omega)} + (\tilde{w}_v, \tilde{w}_v)_{L^2(\Omega)} + (\tilde{w}_i, \tilde{w}_v)_{L^2(\Omega)} + (\tilde{w}_v, \tilde{w}_i)_{L^2(\Omega)}. \quad (4.124)
\end{aligned}$$

Using the same reasoning as in Lemma 4.1, we can write (4.124) as

$$(v, v)_{L^2(\Omega)} = (A_{11}\vec{w}_i, \vec{w}_i)_{\mathbb{R}^{4N}} + (A_{14}\vec{w}_v, \vec{w}_i)_{\mathbb{R}^{4N}} + (A_{41}\vec{w}_i, \vec{w}_v)_{\mathbb{R}^{4N}} + (A_{44}\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}}. \quad (4.125)$$

From (2.5), (2.8), and (3.7), the basis functions in (4.122) and (4.123) are linearly independent. Therefore $v \neq 0$ for $\vec{w}_v \neq \vec{0}$ and hence by (4.121) and (4.125),

$$(P\vec{w}_v, \vec{w}_v)_{\mathbb{R}^{4N}} = (v, v)_{L^2(\Omega)} > 0.$$

Therefore we have shown that P is positive definite. \square

A conjecture which remains to be proven theoretically is

$$\gamma_1(P\vec{v}, \vec{v})_{\mathbb{R}^{4N}} \leq (S\vec{v}, \vec{v})_{\mathbb{R}^{4N}} \leq \gamma_2(P\vec{v}, \vec{v})_{\mathbb{R}^{4N}}, \quad \forall \vec{v} \in \mathbb{R}^{4N}, \quad (4.126)$$

where $0 < \gamma_1 < \gamma_2$ are independent of h . The significance of this conjecture is that when true, the number of iterations of the PCG method required to reduce initial error to ϵ is dependent only on ϵ , and not on the size of the problem we are trying to solve.

4.4.3 Solving the preconditioned system

To solve the preconditioned system

$$P\vec{w}_v = \vec{c}_4 \quad (4.127)$$

for \vec{w}_v , where arbitrary \vec{c}_4 is of the form (4.106), we set $\vec{d} = \vec{0}$ in (4.107)–(4.108) to obtain

$$\begin{aligned} B_{11}\vec{e} + B_{12}\vec{w}_v &= \vec{0}, \\ B_{21}\vec{e} + A_{44}\vec{w}_v &= \vec{c}_4. \end{aligned} \quad (4.128)$$

If \vec{e} and \vec{w}_v satisfy (4.128), then by (4.114) \vec{w}_v satisfies (4.127). Therefore, to solve (4.127), it suffices to solve (4.128) for \vec{w}_v . By (4.103)–(4.105) and (4.107)–(4.108), (4.128) is equivalent to

$$\begin{aligned} A_{11}\vec{w}_i + A_{13}\vec{z} + A_{14}\vec{w}_v &= \vec{0}, \\ A_{31}\vec{w}_i + A_{34}\vec{w}_v &= \vec{0}, \\ A_{41}\vec{w}_i + A_{43}\vec{z} + A_{44}\vec{w}_v &= \vec{c}_4, \end{aligned}$$

which, by (4.97)–(4.99), is the matrix-vector form of

$$\sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl}$$

$$\begin{aligned}
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy z_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy z_{kl} \\
& \quad + \sum_{k=0,2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} = 0, \tag{4.129}
\end{aligned}$$

for $i, j = 1, \dots, 2N$,

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy z_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy z_{kl} \\
& \quad + \sum_{k=0,2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} = (c_4)_{i,j}, \tag{4.130}
\end{aligned}$$

for $i = 0, 2N + 1, j = 1, \dots, 2N$, and

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy w_{kl} \\
& \quad + \sum_{k=0,2N+1} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy w_{kl} \\
& \quad + \sum_{k=0,2N+1} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy w_{kl} = 0, \tag{4.131}
\end{aligned}$$

for $i, j = 1, \dots, 2N$.

To develop a method for solving (4.129)–(4.131), we consider the following

variant of (4.72)–(4.74):

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy u_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy u_{kl} \\
& = 0, \tag{4.132}
\end{aligned}$$

where $i, j = 1, \dots, 2N$,

$$\begin{aligned}
& \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0, 2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy u_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy u_{kl} \\
& = (b_2)_{i,j}, \tag{4.133}
\end{aligned}$$

where $i = 1, \dots, 2N$, $j = 0, 2N + 1$, and

$$\sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=1}^{2N} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{kl}$$

$$\begin{aligned}
& + \sum_{k=1}^{2N} \int_0^1 (\phi'_i \phi'_k)(x) dx \sum_{l=0,2N+1} \int_0^1 (\psi_j \psi_l)(y) dy v_{kl} \\
& + \sum_{k=1}^{2N} \int_0^1 (\phi_i \phi_k)(x) dx \sum_{l=0,2N+1} \int_0^1 (\psi'_j \psi'_l)(y) dy v_{kl} \\
& = 0, \tag{4.134}
\end{aligned}$$

where $i, j = 1, \dots, 2N$. To solve (4.129)–(4.131) using (4.132)–(4.134), we take $(b_2)_{i,j}$ of (4.133) so that

$$(b_2)_{i,j} = (c_4)_{j,2i}, \quad (b_2)_{N-1+i,j} = (c_4)_{j,2i-1}, \quad j = 0, 2N+1, \quad i = 1, \dots, N-1, \tag{4.135}$$

$$(b_2)_{2N-1,j} = (c_4)_{j,2N-1}, \quad (b_2)_{2N,j} = (c_4)_{j,2N}, \quad j = 0, 2N+1. \tag{4.136}$$

Having solved (4.132)–(4.134) for $\{v_{k,l}\}_{k=1,l=0,2N+1}^{2N}$, we define $\{w_{l,k}\}_{l=0,2N+1,k=1}^{2N}$ as follows:

$$w_{l,2k} = v_{k,l}, \quad w_{l,2k-1} = v_{N-1+k,l}, \quad l = 0, 2N+1, \quad k = 1, \dots, N-1,$$

$$w_{l,2N-1} = v_{2N-1,l}, \quad w_{l,2N} = v_{2N,l}, \quad l = 0, 2N+1.$$

Then using (2.5)–(2.6), it can be shown that $\{w_{k,l}\}_{k=0,2N+1,l=1}^{2N}$ is a part of the solution to (4.129)–(4.131). Therefore it follows that, in order to solve (4.128) for \vec{w}_v , it suffices to solve (4.75)–(4.77) for \vec{v}_h , with $\vec{b}_1 = \vec{b}_3 = \vec{0}$, and also components of

\vec{b}_2 defined by (4.135)–(4.136). This is done using Algorithm (4.65). To perform step 1 of this algorithm, we need only compute $\vec{b}'_2 = (Z^T \otimes I_2)\vec{b}_2$ of (4.40) since $\vec{b}_1 = \vec{b}_3 = \vec{0}$. This involves two multiplications with Z^T which can be accomplished at a cost of $O(N \log_2 N)$. The cost of step 2 of (4.65) remains $O(N^2)$, while step 3 requires computing $\vec{v}_h = (Z \otimes I_2)\vec{v}'_h$ of (4.37). We can use FFTs for the two multiplications with Z for a cost of $O(N \log_2 N)$. Thus the cost of Algorithm (4.65) in this special case is $O(N^2)$.

4.4.4 Cost of solving the Schur complement system

Using conjecture (4.126), we determine the cost of solving the Schur complement system by the PCG method with preconditioner P .

The PCG method applied to (4.6) produces iterates $\vec{v}_v^{(0)}, \vec{v}_v^{(1)}, \dots$, satisfying

$$\|\vec{v}_v^{(n)} - \vec{v}_v\|_S \leq 2\rho^n \|\vec{v}_v^{(0)} - \vec{v}_v\|_S, \quad n = 1, 2, \dots,$$

where

$$\|\vec{v}\|_S = \sqrt{(S\vec{v}, \vec{v})_{\mathbb{R}^{4N}}}, \quad \forall \vec{v} \in \mathbb{R}^{4N},$$

and

$$\rho = \frac{1 - \sqrt{\gamma_1/\gamma_2}}{1 + \sqrt{\gamma_1/\gamma_2}},$$

(cf. Section 10.2.7 of [13]). Then if $0 < \epsilon < 1$ and we require

$$\|\vec{v}_v^{(n)} - \vec{v}_v\|_S \leq \epsilon \|\vec{v}_v^{(0)} - \vec{v}_v\|_S,$$

this can be accomplished with a number of iterations

$$n \geq \frac{\log_2(2/\epsilon)}{\log_2(1/\rho)}. \quad (4.137)$$

With $\epsilon = O(h^\rho)$, the number of iterations is $O(\log_2 N)$. In our numerical experiments, we used $n = 2\log_2 N$. It follows from subsections 4.4.1 and 4.4.3 that the cost of one PCG iteration is $O(N^2)$. Hence the total cost of solving the Schur complement system is $O(N^2 \log_2 N)$ under the assumption that P is spectrally equivalent to S .

4.5 Cost of Solving the Galerkin Problem

Finally we discuss the cost of Algorithm (4.26) to solve (3.5)–(3.6). Step 1 of Algorithm (4.26) consists of first solving $M_{11}\vec{w} = \vec{g}$ for \vec{w} , which by Section 4.2 incurs a cost of $O(N^2 \log_2 N)$. Step 1 is completed by computing $M_{21}\vec{w}$, which by Section 4.4.1 involves a cost of $O(N^2)$. Therefore step 1 of Algorithm (4.26) has a total cost of $O(N^2 \log_2 N)$. By Section 4.4.4, step 2 of Algorithm (4.26) is accomplished by the PCG method for a cost of $O(N^2 \log_2 N)$. Step 3 of Algorithm (4.26) requires multiplication $M_{12}\vec{v}_v$, which by Section 4.4 has a total cost of $O(N)$. Finally, step 4

of Algorithm (4.26) requires solving (4.25) for \vec{w} , at a cost shown in Section 4.2 to be $O(N^2 \log_2 N)$. Therefore the total computational cost of Algorithm (4.26) for solving (3.5)–(3.6) is $O(N^2 \log_2 N)$.

Chapter 5

NUMERICAL RESULTS

In this chapter, we show how we perform numerical integration to compute (3.16). We close with results of numerical tests performed on several test problems.

5.1 Gauss Quadrature

In general, it is necessary to evaluate the integrals (3.16) approximately. In order to preserve the $O(h^3)$ convergence rate in the H^1 -norm, we use M-point Gauss quadrature with $M \geq 3$.

For an arbitrary function $g(t)$, we know that the M-point Gauss quadrature for the interval $(0, 1)$ has the form

$$\int_0^1 g(t) dt \approx \sum_{k=1}^M w_k g(\xi_k), \quad (5.1)$$

where $\{w_k\}_{k=1}^M$ and $\{\xi_k\}_{k=1}^M$ are the given weights and nodes, respectively.

For fixed i and j , we have for $f_{i,j}$ of (3.16),

$$\begin{aligned}
f_{i,j} &= - \int_{y_{j-1}}^{y_{j+1}} \left[\int_{x_{i-1}}^{x_{i+1}} f(x, y) \phi_i(x) dx \right] \psi_j(y) dy \\
&= - \int_{y_{j-1}}^{y_j} \left[\int_{x_{i-1}}^{x_i} f(x, y) \phi_i(x) dx + \int_{x_i}^{x_{i+1}} f(x, y) \phi_i(x) dx \right] \psi_j(y) dy \\
&\quad - \int_{y_j}^{y_{j+1}} \left[\int_{x_{i-1}}^{x_i} f(x, y) \phi_i(x) dx + \int_{x_i}^{x_{i+1}} f(x, y) \phi_i(x) dx \right] \psi_j(y) dy.
\end{aligned}$$

Using the transformations $x = x_{i-1} + ht$ and $x = x_i + ht$ in the x -direction, we obtain

$$\begin{aligned}
f_{i,j} &= -h \int_{y_{j-1}}^{y_j} \left[\int_0^1 f(x_{i-1} + ht, y) \phi_i(x_{i-1} + ht) dt \right. \\
&\quad \left. + \int_0^1 f(x_i + ht, y) \phi_i(x_i + ht) dt \right] \psi_j(y) dy \\
&\quad -h \int_{y_j}^{y_{j+1}} \left[\int_0^1 f(x_{i-1} + ht, y) \phi_i(x_{i-1} + ht) dt \right. \\
&\quad \left. + \int_0^1 f(x_i + ht, y) \phi_i(x_i + ht) dt \right] \psi_j(y) dy.
\end{aligned}$$

Using the transformations $y = y_{j-1} + hs$ and $y = y_j + hs$ in the y -direction, we obtain

further

$$\begin{aligned}
f_{i,j} &= -h^2 \int_0^1 \left[\int_0^1 f(x_{i-1} + ht, y_{j-1} + hs) \phi_i(x_{i-1} + ht) dt \right. \\
&\quad \left. + \int_0^1 f(x_i + ht, y_{j-1} + hs) \phi_i(x_i + ht) dt \right] \psi_j(y_{j-1} + hs) ds \\
&\quad -h^2 \int_0^1 \left[\int_0^1 f(x_{i-1} + ht, y_j + hs) \phi_i(x_{i-1} + ht) dt \right.
\end{aligned}$$

$$+ \int_0^1 f(x_i + ht, y_j + hs) \phi_i(x_i + ht) dt \Big] \psi_j(y_j + hs) ds.$$

Finally using (5.1) to approximate integrals over $(0, 1)$, we obtain

$$\begin{aligned} f_{i,j} \approx & -h^2 \sum_{l=1}^M w_l \left[\sum_{k=1}^M w_k f(x_{i-1} + h\xi_k, y_{j-1} + h\xi_l) \phi_i(x_{i-1} + h\xi_k) \right. \\ & \left. + \sum_{k=1}^M w_k f(x_i + h\xi_k, y_{j-1} + h\xi_l) \phi_i(x_i + h\xi_k) \right] \psi_j(y_{j-1} + h\xi_l) \\ & - h^2 \sum_{l=1}^M w_l \left[\sum_{k=1}^M w_k f(x_{i-1} + h\xi_k, y_j + h\xi_l) \phi_i(x_{i-1} + h\xi_k) \right. \\ & \left. + \sum_{k=1}^M w_k f(x_i + h\xi_k, y_j + h\xi_l) \phi_i(x_i + h\xi_k) \right] \psi_j(y_j + h\xi_l). \end{aligned}$$

If $\phi_i = v_i$, then using (2.3) we have

$$\phi_i(x_{i-1} + h\eta_k) = g_1\left(\frac{x_{i-1} + h\eta_k - x_{i-1}}{h}\right) = g_1(\eta_k),$$

and

$$\phi_i(x_i + h\eta_k) = g_1\left(\frac{x_{i+1} - x_i - h\eta_k}{h}\right) = g_1(1 - \eta_k).$$

If $\phi_i = s_i$, then in the same way using (2.4) we obtain

$$\phi_i(x_{i-1} + h\eta_k) = g_2(\eta_k), \quad \phi_i(x_i + h\eta_k) = -g_2(1 - \eta_k).$$

It is apparent that in the y -direction we obtain similar results. For $\psi_j = v_j$,

$$\psi_j(y_{j-1} + h\eta_l) = g_1(\eta_l), \quad \psi_j(y_j + h\eta_l) = g_1(1 - \eta_l),$$

and for $\psi_j = s_j$,

$$\psi_i(y_{j-1} + h\eta_l) = g_2(\eta_l), \quad \psi_i(y_j + h\eta_l) = -g_2(1 - \eta_l).$$

In our numerical tests, we used the 3-point Gauss quadrature whose weights and nodes are

$$\begin{aligned} & \frac{5}{18}, \quad \frac{8}{18}, \quad \frac{5}{18}, \\ & \frac{5 - \sqrt{15}}{10}, \quad \frac{1}{2}, \quad \frac{5 + \sqrt{15}}{10}, \end{aligned}$$

respectively. We also used the 4-point Gauss quadrature with weights

$$\frac{1}{4} - \frac{1}{24}\sqrt{\frac{10}{3}}, \quad \frac{1}{4} + \frac{1}{24}\sqrt{\frac{10}{3}}, \quad \frac{1}{4} + \frac{1}{24}\sqrt{\frac{10}{3}}, \quad \frac{1}{4} - \frac{1}{24}\sqrt{\frac{10}{3}},$$

and nodes

$$\frac{1 - \sqrt{\frac{1}{7}(3 + 4\sqrt{\frac{3}{10}})}}{2}, \quad \frac{1 - \sqrt{\frac{1}{7}(3 - 4\sqrt{\frac{3}{10}})}}{2}, \quad \frac{1 + \sqrt{\frac{1}{7}(3 - 4\sqrt{\frac{3}{10}})}}{2}, \quad \frac{1 + \sqrt{\frac{1}{7}(3 + 4\sqrt{\frac{3}{10}})}}{2}.$$

3 and 4-point Gauss quadratures were compared to determine if error from the cal-

ulation of the right-hand side was obscuring convergence results. Since there was no significant difference between the two quadratures, all results in the next section were obtained using 3-point Gauss quadrature.

5.2 Numerical Experiments

In this section, we present the results of numerical experiments conducted on several test problems. We use a 3500 line Fortran 77 implementation, coupled with a called LAPACK routine DGBTRS [1] which implements banded Gauss elimination. We ran our algorithm on a Silicon Graphics, Inc. Indigo 2 computer in double precision.

In the first experiment, the right-hand side function f of (1.1) was calculated so that the exact solution is

$$u(x, y) = (1 - x)^2 x^2 (1 - y)^2 y^2.$$

In Table 1, we present errors in H^1 and L^2 norms for this exact solution using five uniform $N \times N$ meshes.

Table 1.
 H^1, L^2 norm errors for $u(x, y) = (1 - x)^2 x^2 (1 - y)^2 y^2$.

N	4	8	16	32	64
$\ u - U\ _{H^1}$.106(-3)	.143(-4)	.184(-5)	.234(-6)	.294(-7)
$\ u - U\ _{L^2}$.385(-5)	.272(-6)	.179(-7)	.115(-8)	.725(-10)
$\ v - V\ _{H^1}$.247(-2)	.323(-3)	.414(-4)	.525(-5)	.661(-6)
$\ v - V\ _{L^2}$.867(-4)	.611(-5)	.402(-6)	.257(-7)	.163(-8)

To determine convergence rates, we assume that the error $e_N \approx Ch^p$, where C is a constant independent of the stepsize $h = 1/N$, and where N for the next finer grid is twice as large as N for the previous grid. To determine p we observe that

$$\frac{\log_2 \left[\frac{e_N}{e_{2N}} \right]}{\log_2 2} \approx \frac{\log_2 \left[\frac{C(2h)^p}{Ch^p} \right]}{\log_2 2} = \frac{\log_2 \left[\left(\frac{2h}{h} \right)^p \right]}{\log_2 2} = p.$$

In Table 2, we give values of

$$\rho = \log_2 \left(\frac{e_N}{e_{2N}} \right) / \log_2 2$$

where the subscripts on ρ refer to the H^1 and L^2 norms.

Table 2.
Convergence Rates
for $u(x, y) = (1 - x)^2 x^2 (1 - y)^2 y^2$.

N	4	8	16	32	64
$\rho_{H^1}(u)$		2.90	2.95	2.98	2.99
$\rho_{L^2}(u)$		3.82	3.93	3.97	3.98
$\rho_{H^1}(v)$		2.93	2.96	2.98	2.99
$\rho_{L^2}(v)$		3.83	3.93	3.97	3.98

The results clearly demonstrate the optimal fourth order accuracy for both u and $v = \Delta u$ in the L^2 norm, while u and v are approximated with optimal third order accuracy in the H^1 norm.

In Tables 3 and 4, we give discrete maximum norm errors and convergence rates at the partition nodes where

$$\|w\|_{C(\bar{\Omega}_h)} = \max_{0 \leq i, j \leq N} |w(x_i, y_j)|$$

represents the discrete maximum norm of w at the partition nodes.

Table 3. Discrete Maximum Norm Errors
at Partition Nodes for $u(x, y) = (1 - x)^2 x^2 (1 - y)^2 y^2$.

N	4	8	16	32	64
$\ u - U\ _{C(\bar{\Omega}_h)}$.152(-4)	.101(-5)	.635(-7)	.397(-8)	.248(-9)
$\ (u - U)_x\ _{C(\bar{\Omega}_h)}$.277(-3)	.345(-4)	.432(-5)	.539(-6)	.674(-7)
$\ (u - U)_y\ _{C(\bar{\Omega}_h)}$.277(-3)	.345(-4)	.432(-5)	.539(-6)	.674(-7)
$\ (u - U)_{xy}\ _{C(\bar{\Omega}_h)}$.835(-3)	.104(-3)	.132(-4)	.166(-5)	.207(-6)
$\ v - V\ _{C(\bar{\Omega}_h)}$.246(-3)	.163(-4)	.110(-5)	.863(-7)	.631(-8)
$\ (v - V)_x\ _{C(\bar{\Omega}_h)}$.929(-2)	.114(-2)	.140(-3)	.174(-4)	.217(-5)
$\ (v - V)_y\ _{C(\bar{\Omega}_h)}$.929(-2)	.114(-2)	.140(-3)	.174(-4)	.217(-5)
$\ (v - V)_{xy}\ _{C(\bar{\Omega}_h)}$.139	.182(-1)	.231(-2)	.292(-3)	.367(-4)

Table 4. Convergence Rates of
Maximum Norm Errors
for $u(x, y) = (1 - x)^2 x^2 (1 - y)^2 y^2$.

N	4	8	16	32	64
$\rho_{C(\bar{\Omega}_h)}(u)$		3.91	4.00	4.00	4.00
$\rho_{C(\bar{\Omega}_h)}(u_x)$		3.00	3.00	3.00	3.00
$\rho_{C(\bar{\Omega}_h)}(u_y)$		3.00	3.00	3.00	3.00
$\rho_{C(\bar{\Omega}_h)}(u_{xy})$		3.01	2.98	2.99	3.00
$\rho_{C(\bar{\Omega}_h)}(v)$		3.92	3.89	3.67	3.77
$\rho_{C(\bar{\Omega}_h)}(v_x)$		3.03	3.02	3.01	3.01
$\rho_{C(\bar{\Omega}_h)}(v_y)$		3.03	3.02	3.01	3.01
$\rho_{C(\bar{\Omega}_h)}(v_{xy})$		2.94	2.97	2.99	2.99

Note that the accuracy for u and v at the nodes is demonstrated to be fourth order, while errors for u_x , u_y , u_{xy} , v_x , v_y , and v_{xy} , are all demonstrated to be of third order. As a comparison, experiments were conducted with the “non-symmetric”

polynomial exact solution

$$u(x, y) = (1 - x)^2 x^5 (1 - y)^3 y^4.$$

Results are shown in Tables 5-8. Note that while the results are similar, the same orders of accuracy develop more slowly.

Table 5.

H^1, L^2 -norm errors for $u(x, y) = (1 - x)^2 x^5 (1 - y)^3 y^4$.

N	4	8	16	32	64
$\ u - U\ _{H^1}$.177(-4)	.298(-5)	.441(-6)	.605(-7)	.794(-8)
$\ u - U\ _{L^2}$.511(-6)	.492(-7)	.397(-8)	.285(-9)	.192(-10)
$\ v - V\ _{H^1}$.100(-2)	.154(-3)	.213(-4)	.281(-5)	.362(-6)
$\ v - V\ _{L^2}$.303(-4)	.267(-5)	.198(-6)	.135(-7)	.882(-9)

Table 6.

Convergence Rates

for $u(x, y) = (1 - x)^2 x^5 (1 - y)^3 y^4$.

N	4	8	16	32	64
$\rho_{H^1}(u)$		2.57	2.76	2.87	2.93
$\rho_{L^2}(u)$		3.38	3.63	3.80	3.90
$\rho_{H^1}(v)$		2.70	2.85	2.92	2.96
$\rho_{L^2}(v)$		3.50	3.75	3.88	3.94

Table 7. Discrete Maximum Norm Errors
at Partition Nodes for $u(x, y) = (1 - x)^2 x^5 (1 - y)^3 y^4$.

N	4	8	16	32	64
$\ u - U\ _{C(\bar{\Omega}_h)}$.192(-4)	.273(-6)	.247(-7)	.189(-8)	.138(-9)
$\ (u - U)_x\ _{C(\bar{\Omega}_h)}$.158(-3)	.297(-4)	.472(-5)	.653(-6)	.860(-7)
$\ (u - U)_y\ _{C(\bar{\Omega}_h)}$.760(-4)	.176(-4)	.304(-5)	.449(-6)	.606(-7)
$\ (u - U)_{xy}\ _{C(\bar{\Omega}_h)}$.715(-3)	.123(-3)	.186(-4)	.261(-5)	.343(-6)
$\ v - V\ _{C(\bar{\Omega}_h)}$.854(-4)	.133(-4)	.181(-5)	.157(-6)	.120(-7)
$\ (v - V)_x\ _{C(\bar{\Omega}_h)}$.365(-2)	.980(-3)	.153(-3)	.204(-4)	.257(-5)
$\ (v - V)_y\ _{C(\bar{\Omega}_h)}$.110(-1)	.255(-2)	.424(-3)	.606(-4)	.809(-5)
$\ (v - V)_{xy}\ _{C(\bar{\Omega}_h)}$.321	.840(-1)	.150(-1)	.223(-2)	.304(-3)

Table 8. Convergence Rates of
Maximum Norm Errors
for $u(x, y) = (1 - x)^2 x^5 (1 - y)^3 y^4$.

N	4	8	16	32	64
$\rho_{C(\bar{\Omega}_h)}(u)$		2.82	3.47	3.70	3.78
$\rho_{C(\bar{\Omega}_h)}(u_x)$		2.41	2.65	2.85	2.93
$\rho_{C(\bar{\Omega}_h)}(u_y)$		2.11	2.53	2.76	2.89
$\rho_{C(\bar{\Omega}_h)}(u_{xy})$		2.54	2.72	2.83	2.93
$\rho_{C(\bar{\Omega}_h)}(v)$		2.69	2.87	3.53	3.71
$\rho_{C(\bar{\Omega}_h)}(v_x)$		1.90	2.68	2.90	2.99
$\rho_{C(\bar{\Omega}_h)}(v_y)$		2.10	2.59	2.81	2.91
$\rho_{C(\bar{\Omega}_h)}(v_{xy})$		1.93	2.49	2.75	2.88

Tables 9-12 present the results of the experiment for the non-polynomial solution

$$u(x, y) = \sin^2(\pi x) \sin^2(\pi y).$$

Table 9.
 H^1, L^2 -norm errors for $u(x, y) = \sin^2(\pi x) \sin^2(\pi y)$.

N	4	8	16	32	64
$\ u - U\ _{H^1}$.281(-1)	.438(-2)	.614(-3)	.807(-4)	.103(-4)
$\ u - U\ _{L^2}$.882(-3)	.772(-4)	.579(-5)	.391(-6)	.252(-7)
$\ v - V\ _{H^1}$.160(+1)	.246	.343(-1)	.450(-2)	.575(-3)
$\ v - V\ _{L^2}$.478(-1)	.425(-2)	.322(-3)	.218(-4)	.141(-5)

Table 10.
 Convergence Rates
 for $u(x, y) = \sin^2(\pi x) \sin^2(\pi y)$.

N	4	8	16	32	64
$\rho_{H^1}(u)$		2.68	2.83	2.93	2.97
$\rho_{L^2}(u)$		3.51	3.74	3.89	3.95
$\rho_{H^1}(v)$		2.70	2.84	2.93	2.97
$\rho_{L^2}(v)$		3.49	3.72	3.89	3.95

Table 11. Discrete Maximum Norm Errors
 at Partition Nodes for $u(x, y) = \sin^2(\pi x) \sin^2(\pi y)$.

N	4	8	16	32	64
$\ u - U\ _{C(\bar{\Omega}_h)}$.431(-2)	.430(-3)	.313(-4)	.204(-5)	.129(-6)
$\ (u - U)_x\ _{C(\bar{\Omega}_h)}$.813(-1)	.145(-1)	.212(-2)	.276(-3)	.349(-4)
$\ (u - U)_y\ _{C(\bar{\Omega}_h)}$.813(-1)	.145(-1)	.212(-2)	.276(-3)	.349(-4)
$\ (u - U)_{xy}\ _{C(\bar{\Omega}_h)}$.480	.463(-1)	.667(-2)	.868(-3)	.110(-3)
$\ v - V\ _{C(\bar{\Omega}_h)}$.229	.248(-1)	.184(-2)	.120(-3)	.761(-5)
$\ (v - V)_x\ _{C(\bar{\Omega}_h)}$.567(+1)	.827	.124	.163(-1)	.207(-2)
$\ (v - V)_y\ _{C(\bar{\Omega}_h)}$.567(+1)	.827	.124	.163(-1)	.207(-2)
$\ (v - V)_{xy}\ _{C(\bar{\Omega}_h)}$.435(+2)	.591(+1)	.853	.111	.154(-1)

Table 12. Convergence Rates of
Maximum Norm Errors
for $u(x, y) = \sin^2(\pi x) \sin^2(\pi y)$.

N	4	8	16	32	64
$\rho_{C(\bar{\Omega}_h)}(u)$		3.33	3.78	3.94	3.98
$\rho_{C(\bar{\Omega}_h)}(u_x)$		2.48	2.78	2.94	2.98
$\rho_{C(\bar{\Omega}_h)}(u_y)$		2.48	2.78	2.94	2.98
$\rho_{C(\bar{\Omega}_h)}(u_{xy})$		3.37	2.80	2.94	2.99
$\rho_{C(\bar{\Omega}_h)}(v)$		3.21	3.75	3.93	3.98
$\rho_{C(\bar{\Omega}_h)}(v_x)$		2.78	2.73	2.93	2.98
$\rho_{C(\bar{\Omega}_h)}(v_y)$		2.78	2.73	2.93	2.98
$\rho_{C(\bar{\Omega}_h)}(v_{xy})$		2.88	2.79	2.94	2.85

Results in Tables 1, 2, 5, 6, 9, and 10 confirm that our method is of optimal fourth order and third order accuracy in the L^2 and H^1 norms, respectively. Tables 3, 4, 7, 8, 11, and 12 demonstrate that at partition nodes the accuracy for functions u and v is fourth order, while the convergence rates for first order derivatives of those functions are third order.

Chapter 6

CONCLUSION

In this chapter, we give concluding remarks on what we have accomplished in our approach of solving the biharmonic Dirichlet problem. We also outline areas of future work.

6.1 Summary

A Ciarlet-Raviart mixed finite element Galerkin method with piecewise Hermite bicubics is used to solve the biharmonic Dirichlet problem. This enables us to reduce the fourth order problem to two second order equations. The key component of our algorithm is then the Schur complement approach, which reduces the discrete problem to two more easily handled problems: an auxiliary Galerkin problem and a Schur complement system. In the auxiliary Galerkin problem, the boundary condition $\partial u / \partial n = 0$ on the two vertical sides of Ω is replaced with $\Delta u = 0$. We therefore solve this problem using separation of variables through solution of a generalized eigenvalue problem in one direction, and fast Fourier transforms. We solve the symmetric and

positive definite Schur complement system iteratively, using a preconditioned conjugate gradient method. We gain computational savings by the special structure of the Schur complement matrix. We limit the number of iterations through a symmetric and positive definite preconditioner which we conjecture is spectrally equivalent to the Schur complement matrix. This preconditioner is related to a biharmonic problem with Δu specified on the two horizontal sides of Ω in place of $\partial u/\partial n$. Computational efficiency overall is $O(N^2 \log_2 N)$. We proved the existence and uniqueness of the Galerkin solution, and numerical results show that convergence in the L^2 and H^1 norms is fourth and third order, respectively.

These results compare very favorably with other methods of solving the biharmonic Dirichlet problem. Using finite differences, Bjørstad [2] developed an algorithm of complexity $O(N^2 \log_2 N)$, however convergence is only second order. Peisker [18] used a Schur complement approach to reduce the size of the mixed Galerkin problem, resulting in an efficient algorithm also of cost $O(N^2 \log_2 N)$. But use of linear finite elements restricts accuracy to second order. On the other hand, Cooper and Prenter [9], Sun [20], and Lou et al [16], developed different piecewise Hermite bicubic orthogonal spline collocation methods which are fourth and third order accurate in the L^2 and H^1 norms, respectively. But the computational cost of the most efficient of these methods, by Lou et al [16], is still $O(N^3)$.

Immediate extensions:

- Discretization of nonhomogeneous boundary conditions in the biharmonic Dirichlet problem (cf. [16]).

- Approximate discretization with non-uniform partitioning in the y-direction.

6.2 Future Work

In this section we list open problems and a future extension of this approach to solving the biharmonic Dirichlet problem:

- Theoretical proof of spectral equivalence of the preconditioner with the Schur complement matrix.

- Theoretical convergence analysis in the L^2 and H^1 norms.

- Extension to variable coefficient problems.

Bibliography

- [1] E. Anderson et al., *LAPACK Users' Guide*, SIAM, Philadelphia, 1992.
- [2] P. Bjørstad, *Fast numerical solution of the biharmonic Dirichlet problem on rectangles*, SIAM J. Numer. Anal., **20** (1983), 59–71.
- [3] D. Braess and P. Peisker, *On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning*, IMA J. Numerical Analysis, **6** (1986), 393–404.
- [4] F. Brezzi and P.A. Raviart, *Mixed finite element methods for 4-th order elliptic equations*, Topics in Numer. Anal. III, J. Miller, ed., Academic Press, New York, 1978, 33–56.
- [5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [6] P. G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, North-Holland Publishing Company, Amsterdam, 1978.

- [7] P. G. Ciarlet and R. Glowinski, *Dual iterative techniques for solving a finite element approximation of the biharmonic equation*, *Comp. Meths. Appl. Mech. Eng.*, **5** (1975), 277–295.
- [8] P. G. Ciarlet and P.A. Raviart, *A mixed finite element method for the biharmonic equation*, *Symposium on Mathematical Aspects of Finite Elements in Partial Differential Equations*, C. de Boor, ed., Academic Press, New York, 1974, 125–143.
- [9] K. D. Cooper and P. M. Prenter, *A coupled double splitting ADI scheme for first biharmonic using collocation*, *Numer. Math. PDE.*, **6** (1990), 321–333.
- [10] G. Fairweather, *Finite Element Galerkin Methods for Differential Equations*, *Lecture Notes in Pure and Applied Mathematics*, Vol. 34, Marcel Dekker, New York, 1978.
- [11] R. S. Falk and J. E. Osborn, *Error estimates for mixed methods*, *RAIRO Numer. Anal.*, **14** (1980), 249–277.
- [12] R. Glowinski and O. Pironneau, *Numerical methods for the first biharmonic equation and for the two-dimensional Stokes problem*, *SIAM Review*, **21** (1979), 167–212.
- [13] G. H. Golub and C. Van Loan, *Matrix Computations, Third Edition*, Johns Hopkins University Press, Baltimore, MD, 1996.

- [14] I. Gustafsson, *A preconditioned iterative method for the solution of the biharmonic problem*, IMA J. Numer. Anal., **4** (1986), 55–67.
- [15] M.R. Hanisch, *A preconditioned iterative method for the solution of the biharmonic problem*, SIAM J. Numer. Anal., **30** (1993), 184–214.
- [16] Z.-M. Lou, B. Bialecki, and G. Fairweather, *Orthogonal spline collocation methods for biharmonic problems*, to appear in Numer. Math.
- [17] A. D. Lyashko and S. I. Solov'yev, *Fourier method of solution of FE systems with Hermite elements for Poisson equation*, Sov. J. Numer. Anal. Math. Modelling, **6** (1991), 121–129.
- [18] P. Peisker, *On the numerical solution of the first biharmonic equation*, Mathematical Modelling and Numerical Analysis, **22** (1988), 655–676.
- [19] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Verlag, New York, 1991.
- [20] W. Sun, *Orthogonal collocation solution of biharmonic equations*, Intern. J. Computer Math., **49** (1993), 221–232.
- [21] M. Vajtersić, *Algorithms for Elliptic Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.