Quadratic Spline Collocation Methods for Systems of Elliptic PDEs

by

Kit Sun Ng

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Quadratic Spline Collocation Methods for

Systems of Elliptic PDEs

Kit Sun Ng

Master of Science

Graduate Department of Computer Science

University of Toronto

2000

We consider Quadratic Spline Collocation (QSC) methods for solving systems of two linear second-order PDEs in two dimensions. Optimal order approximation to the solution is obtained, in the sense that the convergence order of the QSC approximation is the same as the order of the quadratic spline interpolant.

We study the matrix properties of the linear system arising from the discretization of systems of two PDEs by QSC. We give sufficient conditions under which the QSC linear system is uniquely solvable and the optimal order of convergence for the QSC approximation is guaranteed. We develop fast direct solvers based on Fast Fourier Transforms (FFTs) and iterative methods using multigrid or FFT preconditioners for solving the above linear system.

Numerical results demonstrate that the QSC methods are fourth order locally on certain points and third order globally, and that the computational complexity of the linear solvers developed is almost asymptotically optimal. The QSC methods are compared to conventional second order discretization methods and are shown to produce smaller approximation errors in the same computation time, while they achieve the same accuracy in less time.

Dedication

To my family

Acknowledgements

First and foremost, I would like to thank my supervisor Professor C. Christara for her hard work and dedication in guiding me throughout this thesis. Without her help, this thesis would be far inferior.

Secondly, I would like to thank my second reader Professor W. Enright for his helpful comments.

Finally, I would like to thank the University of Toronto, OGSST, and Professor C. Christara for the financial support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many physical phenomena are mathematically modelled by systems of Partial Differential Equations (PDEs). For example, modelling the elastic deformation of a solid subject to external forces in two dimensions results in a system of two second-order PDEs involving two unknown functions. Moreover, high order PDEs may be transformed into a system of lower order PDEs. Therefore, there is a great interest in the scientific and engineering community to solve such problems accurately and efficiently.

In this thesis we consider the numerical solution of Boundary Value Problems (BVPs) described by a 2×2 system of linear second-order elliptic PDEs in two dimensions

$$\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \text{ in } \Omega \equiv (ax, bx) \times (ay, by) \tag{1.1}$$

subject to mixed boundary conditions

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \text{ on } \partial\Omega \equiv \text{ boundary of } \Omega \tag{1.2}$$

where, for $i = 1, 2$ and $j = 1, 2$, $L_{ij}u \equiv a_{ij}u_{xx} + b_{ij}u_{xy} + c_{ij}u_{yy} + d_{ij}u_x + e_{ij}u_y + f_{ij}u$, $B_{ij}u \equiv \alpha_{ij}u + \beta_{ij}u_n$, $a_{ij}, b_{ij}, c_{ij}, d_{ij}, e_{ij}, f_{ij}, g_i, \alpha_{ij}, \beta_{ij}$ and γ_i are given functions of x and y, u and v are the unknown functions of x and y, and u_n and v_n denote the normal derivatives of u and v, respectively.

A finite element method, called the quadratic spline collocation (QSC) method, was developed in [2] to solve a single PDE problem. The collocation points for this method were chosen to be the midpoints of a uniform grid partition of the domain. It was shown that this method has optimal order of convergence. Various solvers, such as domain decomposition and iterative methods, for the resulting system of equations were studied in [3]. The parallel implementation of these solvers was also studied in the same paper. Fast Fourier Transform (FFT) [5] and multigrid [4] methods were developed for the QSC equations.

Russell [13] develops and analyzes collocation methods for systems of first order one-dimensional BVPs. The results of [13] are extension of the work by de Boor and Swartz [6], in which collocation at Gaussian points is analyzed. Ascher et al [14] extend this work by solving mixed order one-dimensional BVPs. The application of collocation at Gaussian points for systems of PDEs has not been studied extensively. However, we note that collocation at Gaussian points requires more than one collocation point per interval of the partition. Therefore, it results in larger linear (or non-linear) systems of equations to solve, than the method of smooth spline collocation.

In this chapter, we will review the properties and solvers for the system of equations arising from the application of the QSC method to a single PDE. In Chapter 2, we extend the optimal QSC method in [2] to a method for solving (1.1)-(1.2). We study the properties of the linear system arising, give sufficient conditions for its solvability, and obtain a priori error bounds for the approximation and its derivatives. Numerical results verify the convergence of the method even for some problems that do not satisfy the conditions obtained by the mathematical analysis. In Chapter 3, we derive an efficient method using FFTs to solve the system of equations arising from the QSC method applied to a system of two PDEs. In the same chapter, we derive another efficient method based on the multigrid techniques developed in [4] to solve the same system of equations. Numerical results verify that the asymptotic computational complexity of the FFT and multigrid solvers is almost optimal. Finally in Chapter 4, we summarize the results of this thesis; the limitations of our work; and future considerations for the QSC method.

1.1 The optimal QSC method for a single PDE

In this section, we summarize some results from [2], that will be used in subsequent chapters, and give the relevant background notation.

Consider a BVP in a rectangular domain $\Omega \equiv (ax, bx) \times (ay, by)$ described by a linear second-order elliptic PDE

$$Lu \equiv au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \text{ in } \Omega, \tag{1.3}$$

subject to boundary conditions defined on the boundary, $\partial\Omega$, of Ω,

$$Bu \equiv \alpha u + \beta u_n = \gamma \text{ on } \partial\Omega, \tag{1.4}$$

where $a, b, c, d, e, f, g, \alpha, \beta$ and γ are given functions of x and y, and u is the unknown function of x and y.

Let $\triangle_x \equiv \{ax = x_0 < x_1 < \ldots < x_M = bx\}$, and $\triangle_y \equiv \{ay = y_0 < y_1 < \ldots < y_N = by\}$ be uniform partitions of the intervals $[ax, bx]$ and $[ay, by]$, with mesh sizes $h_x = \frac{bx-ax}{M}$, $h_y = \frac{by-ay}{N}$, respectively. Then, $\triangle \equiv \triangle_x \times \triangle_y$ is the induced grid partition of $\overline{\Omega} \equiv \Omega \cup \partial\Omega \equiv [ax, bx] \times [ay, by]$.

Based on partition $\triangle$, we define the bi-quadratic spline space – the space of bi-quadratic piecewise polynomials of continuity $\mathbf{C^1}$ on the nodes of partition $\triangle$ – and a set of basis functions for it. Let $\phi_i^x(x) \equiv \frac{1}{2}\psi(\frac{x-ax}{h_x} - i + 2)$, $i = 0, \ldots, M+1$, and $\phi_j^y(y) \equiv \frac{1}{2}\psi(\frac{y-ay}{h_y} - j + 2)$, $j = 0, \ldots, N+1$, where the quadratic function ψ is defined by

$$\psi(x) \equiv \begin{cases} x^2 & \text{if } 0 \le x \le 1, \\ -3 + 6x - 2x^2 & \text{if } 1 \le x \le 2, \\ 9 - 6x + x^2 & \text{if } 2 \le x \le 3, \\ 0 & \text{elsewhere.} \end{cases} \tag{1.5}$$

Note that $\phi_i^x(x)$ and $\phi_j^y(y)$ are the quadratic B-splines defined with respect to partitions $\triangle_x$ and $\triangle_y$, respectively. Then, a set of basis functions for the bi-quadratic spline space defined with respect to partition $\triangle$ is the tensor product of quadratic B-splines $\Phi \equiv \{\phi_i^x(x)\phi_j^y(y)\}_{i=0,j=0}^{M+1,N+1}$. Any bi-quadratic spline $u_\triangle(x,y)$ can be represented as a linear combination of the basis functions, that is, $u_\triangle(x,y) = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\theta_{ij}\phi_i^x(x)\phi_j^y(y)$, where θ_{ij}, $i = 0,\ldots,M+1$, $j = 0,\ldots,N+1$, are the *coefficients* or *degrees of freedom* of $u_\triangle$. The representation of a bi-quadratic spline $u_\triangle(x,y)$ as a linear combination of the basis functions is often referred to as the *finite element representation* of $u_\triangle(x,y)$. In the rest of the thesis, for brevity and convenience, we will omit the superscripts x and y from the basis functions, and use $u_\triangle$ for $u_\triangle(x,y)$.

A set of data points, called *collocation* points, is prescribed in Ω and on $\partial\Omega$. Throughout, let τ_i^x, $i = 1,\ldots,M$, be the midpoints of $\triangle_x$, τ_j^y, $j = 1,\ldots,N$, be the midpoints of $\triangle_y$, $\tau_0^x = x_0$, $\tau_{M+1}^x = x_M$, $\tau_0^y = y_0$ and $\tau_{N+1}^y = x_N$. Then, $T \equiv \{(\tau_i^x,\tau_j^y), i = 0,\ldots,M+1, j = 0,\ldots,N+1\}$ is the set of bi-quadratic spline collocation points in $\overline{\Omega}$. Let also $T_\partial \equiv T \cap \partial\Omega$ denote the collocation points on $\partial\Omega$.

A discretization method for determining an approximate solution to (1.3)-(1.4) that belongs to a piecewise polynomial space is optimal, if the approximate solution has the same order of convergence as point interpolation for the same set of data points in the same approximation space.

The conventional formulation of bi-quadratic spline collocation (QSC) methods, based on finding a bi-quadratic spline $u_\triangle^{(1)}$ such that

$$Lu_\triangle^{(1)} = g \text{ in } T - T_\partial \quad\text{and}\quad Bu_\triangle^{(1)} = \gamma \text{ on } T_\partial \tag{1.6}$$

is non-optimal. Optimal methods have been constructed [2] based on appropriate perturbations P_L and P_B of the operators L and B, respectively. Two methods were formulated. The first method, called *one-step* or *extrapolated* QSC, involves determining a bi-quadratic spline w such that

$$Lw + P_L w = g \text{ in } T - T_\partial \text{ and } Bw + P_B w = \gamma \text{ on } T_\partial. \tag{1.7}$$

The second method, called *two-step* or *deferred-correction* QSC, involves determining first a bi-quadratic spline $u_\triangle^{(1)}$ that satisfies (1.6) and then a biquadratic spline $u_\triangle^{(2)}$ such that

$$Lu_\triangle^{(2)} = g - P_L u_\triangle^{(1)} \text{ in } T - T_\partial \text{ and } Bu_\triangle^{(2)} = \gamma - P_B u_\triangle^{(1)} \text{ on } T_\partial. \tag{1.8}$$

The formulae that define the perturbation operators P_L and P_B are given in [2] and in more details in [1].

Both methods are optimal, but the deferred-correction method is more efficient in terms of time and memory requirements. The higher efficiency of the deferred-correction method is due to the fact that the linear system arising from (1.6) has a smaller bandwidth, and fewer nonzero entries per row than the system arising from (1.7). We also note that the matrix of the linear system arising from (1.8) is the same as the matrix of the linear system arising from (1.6). In this thesis we develop optimal deferred-correction QSC methods to solve (1.1)-(1.2).

1.2 Properties of the QSC matrix for a single PDE

Let $u_\triangle^{(1)} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{\triangle ij}^{(1)} \phi_i(x)\phi_j(y)$ and $u_\triangle^{(2)} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{\triangle ij}^{(2)} \phi_i(x)\phi_j(y)$ be the finite element representation of the bi-quadratic splines defined by equations (1.6) and (1.8), respectively.

Equations (1.6) result in a linear system of the form

$$K\overline{\theta_\triangle^{(1)}} = \overline{g}, \tag{1.9}$$

where K is an $(M+2)(N+2) \times (M+2)(N+2)$ matrix, with at most nine nonzero entries per row, $\overline{\theta_\triangle^{(1)}}$ the $(M+2)(N+2) \times 1$ vector of unknown coefficients of $u_\triangle^{(1)}$ and $\overline{g}$ the $(M+2)(N+2) \times 1$ vector of values of g at the collocation points.

Equations (1.8) result in a linear system

$$K\overline{\theta_{\triangle}^{(2)}} = \overline{g} - \overline{Pu_{\triangle}^{(1)}}, \tag{1.10}$$

with the same matrix K as in (1.9), while $\overline{\theta_{\triangle}^{(2)}}$ is the vector of unknown coefficients of $u_{\triangle}^{(2)}$ and $\overline{Pu_{\triangle}^{(1)}}$ is the perturbation of $\overline{g}$ defined based on the perturbation operators P_L and P_B and the solution $u_{\triangle}^{(1)}$ of the first step of the QSC method.

We adopt a bottom-up then left-to-right natural ordering of the collocation points. Since there is an one-to-one mapping of points to equations and unknown coefficients of the bi-quadratic splines that are to be determined, the ordering of the points specifies the ordering of the equations and the unknowns, and thus of the components of vectors $\overline{g}$, $\overline{\theta_{\triangle}^{(1)}}$, $\overline{Pu_{\triangle}^{(1)}}$ and $\overline{\theta_{\triangle}^{(2)}}$.

With this ordering, the matrix K of the QSC equations (1.9) (as well as that of (1.10)) is block-tridiagonal, with tridiagonal $(N+2) \times (N+2)$ blocks, and has semi-bandwidth $N+3$. Figure 1.1 shows the sparsity pattern of the QSC matrix for $N = M = 4$.

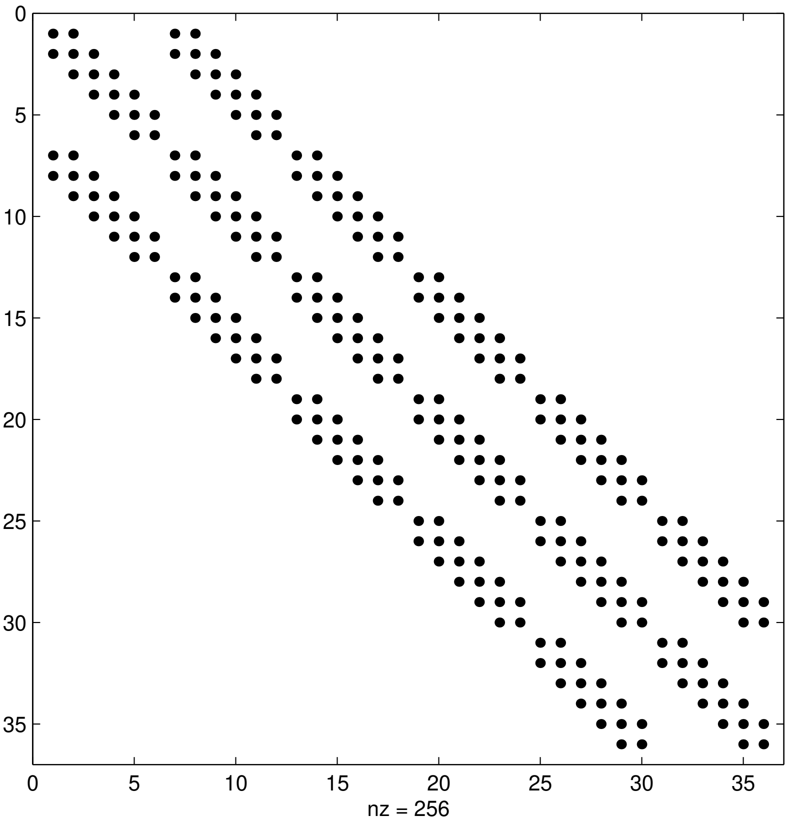

Figure 1.1: The structure of the QSC matrix arising from the discretization of a single PDE (1.3) with general boundary conditions (1.4), for $N = M = 4$. The natural ordering is adopted. The number of non-zero entries is denoted by nz.

1.3 Error bounds of optimal QSC for a single PDE

In this section, we summarize the a priori bounds that the QSC approximations $u_\triangle^{(1)}$ and $u_\triangle^{(2)}$ defined by equations (1.6) and (1.8), respectively, satisfy. In [2], it is shown that, under certain conditions, the QSC equations (1.6) and (1.8) are uniquely solvable. Moreover, if $\epsilon_\triangle^{(1)} = u - u_\triangle^{(1)}$ and $\epsilon_\triangle^{(2)} = u - u_\triangle^{(2)}$ are the errors of the QSC approximations $u_\triangle^{(1)}$ and $u_\triangle^{(2)}$, respectively, the following error bounds hold:

$$\begin{aligned}
|\epsilon_\triangle^{(1)}(x_i, y_j)| &= O(h^2), \quad ||\epsilon_\triangle^{(1)}||_\infty = O(h^2)\\
|\epsilon_\triangle^{(1)}(\tau_i^x, y_j)| &= O(h^2),\\
|\epsilon_\triangle^{(1)}(x_i, \tau_j^y)| &= O(h^2),\\
|\epsilon_\triangle^{(1)}(\tau_i^x, \tau_j^y)| &= O(h^2),
\end{aligned} \tag{1.11}$$

$$\begin{aligned}
|\epsilon_\triangle^{(2)}(x_i, y_j)| &= O(h^4), & ||\epsilon_\triangle^{(2)}||_\infty &= O(h^3)\\
|\epsilon_\triangle^{(2)}(\tau_i^x, y_j)| &= O(h^4), & &\\
|\epsilon_\triangle^{(2)}(x_i, \tau_j^y)| &= O(h^4), & &\\
|\epsilon_\triangle^{(2)}(\tau_i^x, \tau_j^y)| &= O(h^4), & &\\
|D_x\epsilon_\triangle^{(2)}(x_i - \lambda h_x, .)| &= O(h^3), & ||D_x\epsilon_\triangle^{(2)}||_\infty &= O(h^2)\\
|D_y\epsilon_\triangle^{(2)}(., y_i - \lambda h_y)| &= O(h^3), & ||D_y\epsilon_\triangle^{(2)}||_\infty &= O(h^2)\\
|D_{xy}\epsilon_\triangle^{(2)}(x_i - \lambda h_x, y_i - \lambda h_y)| &= O(h^2), & ||D_{xy}\epsilon_\triangle^{(2)}||_\infty &= O(h^2)\\
|D_x^2\epsilon_\triangle^{(2)}(\tau_i^x, \tau_j^y)| &= O(h^2), & ||D_x^2\epsilon_\triangle^{(2)}||_\infty &= O(h)\\
|D_y^2\epsilon_\triangle^{(2)}(\tau_i^x, \tau_j^y)| &= O(h^2), & ||D_y^2\epsilon_\triangle^{(2)}||_\infty &= O(h)
\end{aligned} \tag{1.12}$$

where $h = \max(h_x, h_y)$, $\lambda = (3 \pm \sqrt{3})/6$ and D_z^n denotes the nth derivative operator with respect to the variable z.

1.4 Boundary conditions

Consider a BVP in a rectangular domain $\Omega \equiv (ax, bx) \times (ay, by)$ described by a linear second-order elliptic PDE (1.3) subject to *homogeneous Dirichlet* boundary conditions,

$$Bu \equiv u = 0 \text{ on } \partial\Omega. \tag{1.13}$$

Using the functions $\phi_i(x), i = 1, \ldots, M$, and $\phi_j(y), j = 1, \ldots, N$, we can construct a set of bi-quadratic spline basis functions that satisfy the homogeneous Dirichlet boundary conditions. More specifically, if

$$\xi_1(x) = \phi_1(x) - \phi_0(x); \xi_i(x) = \phi_i(x), i = 2, \ldots, M-1; \xi_M(x) = \phi_M(x) - \phi_{M+1}(x),$$

and

$$\xi_1(y) = \phi_1(y) - \phi_0(y); \xi_j(y) = \phi_j(y), j = 2, \ldots, N-1; \xi_N(y) = \phi_N(y) - \phi_{N+1}(y),$$

the set of basis functions $\{\xi_i(x)\xi_j(y)\}_{i=1,i=1}^{M,N}$ satisfies the homogeneous Dirichlet boundary conditions (1.13). The optimal two-step QSC method applied to PDE problems with homogeneous Dirichlet boundary conditions (1.3)-(1.13) involves first determining a bi-quadratic spline $u_\triangle^{(1)} = \sum_{i=1}^{M} \sum_{j=1}^{N} \theta_{\triangle ij}^{(1)} \xi_i(x)\xi_j(y)$ that satisfies

$$Lu_\triangle^{(1)} = g \text{ in } T - T_\partial, \tag{1.14}$$

then a bi-quadratic spline $u_\triangle^{(2)} = \sum_{i=1}^{M} \sum_{j=1}^{N} \theta_{\triangle ij}^{(2)} \xi_i(x)\xi_j(y)$ that satisfies

$$Lu_\triangle^{(2)} = g - P_L u_\triangle^{(1)} \text{ in } T - T_\partial. \tag{1.15}$$

By the construction of the basis functions $\xi_i(x)\xi_j(y), i = 1, \ldots, M, j = 1, \ldots, N$, both $u_\triangle^{(1)}$ and $u_\triangle^{(2)}$ satisfy the homogeneous Dirichlet conditions $u_\triangle^{(1)} = 0$ on T_∂ and $u_\triangle^{(2)} = 0$ on T_∂. This implementation of the QSC method produces a smaller size system and can still be formulated as a two-step method. For the remainder of the thesis, we will refer to this formulation as *interior* QSC method. The matrix of the equations arising from the application of the interior QSC method to the PDE problem (1.3)-(1.13), assuming the above set of basis functions, is of size $MN \times MN$, with at most nine nonzero entries per row. We adopt a bottom-up then left-to-right natural ordering of the collocation points. With this ordering, this matrix of QSC equations is block-tridiagonal, with tridiagonal $N \times N$ blocks, and has semi-bandwidth $N+1$. Figure 1.2 shows the sparsity pattern of this QSC matrix for $N = M = 4$.

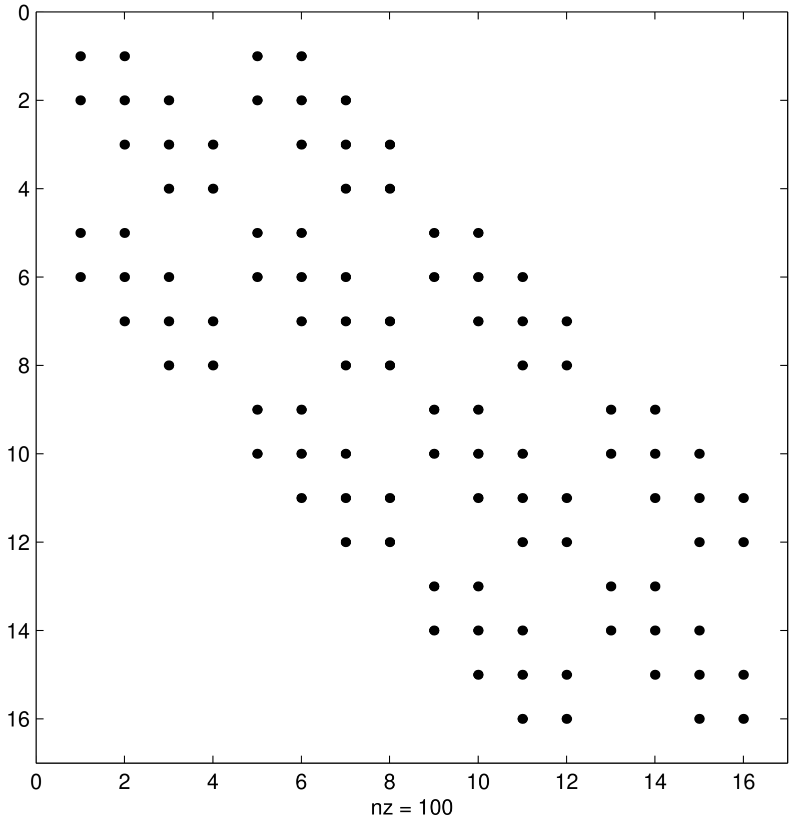

Figure 1.2: The structure of the interior QSC matrix arising from the discretization of a single PDE (1.3) with homogeneous Dirichlet boundary conditions (1.13), for $N = M = 4$. The natural ordering is adopted. The number of non-zero entries is denoted by nz.

In a similar way, we can adjust the basis functions so that they satisfy *Neumann* boundary conditions, i.e.

$$Bu \equiv u_n = 0 \text{ on } \partial\Omega,$$

where u_n denotes the normal derivative of u. The details of this adjustment are found in [2].

Finally, the optimal QSC method has been developed for *periodic* boundary conditions, i.e.

$$u(ax, y) = u(bx, y) \text{ and } u_x(ax, y) = u_x(bx, y), \forall y \text{ in } [ay, by]$$
$$u(x, ay) = u(x, by) \text{ and } u_y(x, ay) = u_y(x, by), \forall x \text{ in } [ax, bx].$$

The details are found in [5]. In this thesis, we will deal with either general mixed boundary conditions, for which we apply the general formulation of the QSC method, or with homogeneous Dirichlet conditions, for which we apply either the general formulation or the interior QSC method. However, it is instructive to note that our results can be easily extended to take advantage of the above two types of boundary conditions.

1.5 Helmholtz operator with constant coefficients

Consider the Helmholtz equation

$$Lu \equiv au_{xx} + cu_{yy} + fu = g \text{ in } \Omega, \tag{1.16}$$

subject to homogeneous Dirichlet boundary conditions (1.13), where a, c and f are constants. All of the properties reviewed in the remainder of this section are shown in [2].

Let K be the $MN \times MN$ matrix of collocation equations arising from the application of the interior QSC method to the PDE problem (1.16)-(1.13). The eigenvalues of K are

$$\rho^{lm} = a\frac{1}{h_x^2}\lambda_l^x(\lambda_m^y + 8) + c\frac{1}{h_y^2}(\lambda_l^x + 8)\lambda_m^y + \frac{1}{8}f(\lambda_l^x + 8)(\lambda_m^y + 8), \tag{1.17}$$

where

$$\lambda_l^x = -4\sin^2(\frac{l\pi}{2M}), \lambda_m^y = -4\sin^2(\frac{m\pi}{2N}); l = 1, \ldots, M, m = 1, \ldots, N.$$

A set of linearly independent eigenvectors of K is given by

$$\{\overline{\delta^{lm}} = \overline{\delta_l^x} \otimes \overline{\delta_m^y}; l = 1, \ldots, M, m = 1, \ldots, N\}, \tag{1.18}$$

where $\otimes$ is the tensor product (Kronecker product) operator and

$$\begin{aligned} (\overline{\delta_l^x})_j &= \kappa_l^x \sin \frac{(2j-1)l\pi}{2M}; j = 1, \ldots, M, \\ (\overline{\delta_m^y})_j &= \kappa_m^y \sin \frac{(2j-1)m\pi}{2N}; j = 1, \ldots, N, \end{aligned} \tag{1.19}$$

where κ_l^x and κ_m^y are non-zero constants. Furthermore, it is shown in [5] that $\{\delta^{lm}\}_{l=1,m=1}^{M,N}$ can be normalized with respect to the Euclidean norm. To normalize the set, for each $\overline{\delta^{lm}} = \overline{\delta_l^x} \otimes \overline{\delta_m^y}, l = 1, \ldots, M, m = 1, \ldots, N$, first set

$$\kappa_l^x = \sqrt{\frac{2}{M}}, \text{ for } l = 1, \ldots, M-1, \text{ and } \kappa_M^x = \sqrt{\frac{1}{M}},$$

for $\overline{\delta_l^x}$ in (1.19). Then set

$$\kappa_m^y = \sqrt{\frac{2}{N}}, \text{ for } m = 1, \ldots, N-1, \text{ and } \kappa_N^y = \sqrt{\frac{1}{N}},$$

for $\overline{\delta_m^y}$ in (1.19).

1.6 Fast Fourier transform solvers for the QSC equations

Methods based on Fast Fourier Transforms (FFTs) are efficient for solving the linear systems arising from the application of the QSC method to the single PDE problem (1.3)-(1.4). It is shown in [5] that the FFT solvers are well-suited for solving Helmholtz PDE problems that have constant coefficients, such as (1.16)-(1.13), because the linear system arising from the application of the QSC method to these problems has eigenvalues and eigenvectors that are given by sine and cosine formulae. For general PDE problems, FFT solvers can be used as preconditioners. In this section, we will review FFT solvers for Helmholtz PDE problems and discuss a variety of FFT preconditioners.

1.6.1 Discrete sine transform II

The *Discrete Sine Transform II* (DST-II) is the computation of the matrix-vector product $y = S_n \cdot x$, where $x = [x_1, x_2, \ldots, x_n]^T$ is a given vector and S_n is the $n \times n$ matrix with entries

$$(S_n)_{lj} = \sin\left(\frac{(2j-1)l\pi}{2n}\right); l, j = 1, \ldots, n.$$

That is

$$y_l = \sum_{j=1}^{n} \sin\left(\frac{(2j-1)l\pi}{2n}\right) x_j; l = 1, \ldots, n.$$

The matrix S_n is orthogonal and can be scaled to become orthonormal.

It is shown in [11] that if n is a power of 2, then there is an algorithm that computes the DST-II of a vector in $2.5n\log_2 n$ real single flops (counting additions and multiplications separately), rather than the standard n^2 flops. We will refer to this algorithm as a *Fast Sine Transform II* (FST-II).

The *inverse Discrete Sine Transform II* (iDST-II) is the computation of the matrix-vector $y = S_n^{-1} \cdot x$, where x and S_n are defined above. It is shown in [11] that if n is a power of 2, then there is an algorithm that computes the iDST-II of a vector in $2.5n\log_2 n$ real single flops rather than the standard n^2 flops. We will refer to this algorithm as an *inverse Fast Sine Transform II* (iFST-II).

The FST-II and iFST-II are used for solving QSC equations arising from the application of the interior QSC method to Helmholtz PDE problems with constant coefficients and homogeneous Dirichlet conditions. In [5] other FFTs are used for other boundary conditions.

1.6.2 Properties of tensor products of matrices

Let $A \in \mathbf{C}^{M\times N}, B \in \mathbf{C}^{J\times L}$. Then the tensor product (Kronecker product) $A \otimes B \in \mathbf{C}^{M\cdot J\times N\cdot L}$ is defined by

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1N}B \\ a_{21}B & a_{22}B & \dots & a_{2N}B \\ \dots & \dots & \dots & \dots \\ a_{M1}B & a_{M2}B & \dots & a_{MN}B \end{bmatrix}.$$

Let $x \in \mathbf{C}^{L\cdot N}$, where L and N are positive integers. We denote by $x_{L\times N}$ the matrix X of size $L \times N$ for which $X_{lj} = x_{(j-1)L+l}, j = 1,\dots,N, l = 1,\dots,L$, that is, each column of X is a block of L (consecutive) components of x.

The following tensor properties are shown in [11] and [5]:

Kron5: If $B \in \mathbf{C}^{L\times L}$, I_N the identity matrix of size N and $x \in \mathbf{C}^{L\cdot N}$, then

$$(I_N \otimes B)x = y \Rightarrow y_{L\times N} = B \cdot x_{L\times N}.$$

Kron6: If $A \in \mathbf{C}^{N\times N}$, I_L the identity matrix of size L and $x \in \mathbf{C}^{L\cdot N}$, then

$$(A \otimes I_L)x = y \Rightarrow y_{L\times N} = x_{L\times N} \cdot A^T.$$

Kron56: If $A \in \mathbf{C}^{N\times N}, B \in \mathbf{C}^{L\times L}$ and $x \in \mathbf{C}^{L\cdot N}$, then

$$(A \otimes B)x = y \Rightarrow y_{L\times N} = B \cdot x_{L\times N} \cdot A^T = B(A(x_{L\times N})^T)^T.$$

1.6.3 FFT solvers for the Helmholtz PDE problem

Let K be the $MN \times MN$ matrix of equations arising from the application of the interior QSC method to the Helmholtz PDE problem (1.16)-(1.13), with the basis functions chosen so that the boundary conditions are satisfied by construction. Let $\overline{\delta_l^x}, l = 1,\dots,M$, and $\overline{\delta_m^y}, m = 1,\dots,N$, be normalized vectors after selecting appropriate κ_l^x and κ_m^y in (1.19). Let V_M be the matrix whose lth column is $\overline{\delta_l^x}, l = 1,\dots,M$. Let V_N be the matrix whose mth column is $\overline{\delta_m^y}, m = 1, \ldots, N$. Let $V = V_M \otimes V_N$ be a matrix of normalized eigenvectors of K, and Λ be a diagonal matrix such that the kth entry is an eigenvalue of K corresponding to the kth column eigenvector of V, $k = 1, \ldots, MN$.

It is shown in [5] that $V_M = S_M^{-1} D_M^{-1}$, and $V_M^{-1} = D_M S_M$, where D_M is some diagonal matrix whose non-zero entries are scaling factors. So, we have $V = (S_M^{-1} \otimes S_N^{-1}) D_M^{-1} D_N^{-1}$ and $V^{-1} = D_N D_M (S_M \otimes S_N)$. Since the columns of V are the MN linearly independent eigenvectors of K, if we assume that K is solvable, we have $KV = V\Lambda \Rightarrow K = V\Lambda V^{-1}$, and so $K^{-1} = V\Lambda^{-1}V^{-1}$. Hence, if $Kx = \overline{g}$ then

$$\begin{aligned} x &= V\Lambda^{-1}V^{-1}\overline{g} \\ &= (S_M^{-1} \otimes S_N^{-1}) D_M^{-1} D_N^{-1} \Lambda^{-1} D_N D_M (S_M \otimes S_N) \\ &= (S_M^{-1} \otimes S_N^{-1}) \Lambda^{-1} (S_M \otimes S_N)\overline{g}. \end{aligned} \tag{1.20}$$

Taking into account property **Kron56** of the previous section, the following algorithm, referred to as *FST2(M,N)*, is used to compute $\overline{g^{(2)}} = (S_M \otimes S_N)\overline{g}$:

Algorithm FST2(M,N)

Step 1: Perform FST-II of size M to each of the N columns of the vector $(\overline{g}_{N\times M})^T$ to obtain $\overline{g^{(1)}}_{M\times N} = S_M(\overline{g}_{N\times M})^T$.

Step 2: Perform FST-II of size N to each of the M columns of the vector $(\overline{g^{(1)}}_{M\times N})^T$ to obtain $\overline{g^{(2)}}_{N\times M} = S_N(\overline{g^{(1)}}_{M\times N})^T$, or equivalently, $\overline{g^{(2)}} = (S_M \otimes S_N)\overline{g}$.

The above two steps require approximately $2.5MN\log_2 M$ and $2.5MN\log_2 N$ real single flops, respectively. Hence, FST2(M,N) requires approximately $2.5MN\log_2(MN)$ real single flops.

Taking into account property **Kron56** of the previous section, the following algorithm, referred to as *iFST2(M,N)*, is used to compute $x = (S_M^{-1} \otimes S_N^{-1})\overline{g^{(2)}}$:

Algorithm iFST2(M,N)

Step 1: Perform iFST-II of size M to each of the N columns of the vector $(\overline{g^{(2)}}_{N\times M})^T$ to obtain $\overline{g^{(3)}}_{M\times N} = S_M^{-1}(\overline{g^{(2)}}_{N\times M})^T$.

Step 2: Perform iFST-II of size N to each of the M columns of the vector $(\overline{g^{(3)}}_{M\times N})^T$ to obtain $x_{N\times M} = S_N^{-1}(\overline{g^{(3)}}_{M\times N})^T$, or, equivalently, $x = (S_M^{-1} \otimes S_N^{-1})\overline{g^{(2)}}$.

The above two steps require approximately $2.5MN\log_2 M$ and $2.5MN\log_2 N$ real single flops, respectively. Hence, iFST2(M,N) requires approximately $2.5MN\log_2(MN)$ real single flops.

The computation involved in (1.20) can be performed in the following steps:

Algorithm: FFT solution of a Helmholtz PDE with constant coefficients

Step 1: Perform FST2(M,N) to $\overline{g}$ to obtain $\overline{g^{(2)}} = (S_M \otimes S_N)\overline{g}$.

Step 2: Divide each of the components of $\overline{g^{(2)}}$ by the respective eigenvalue of K to obtain vector $\overline{g^{(3)}} = \Lambda^{-1}\overline{g^{(2)}}$.

Step 3: Perform iFST2(M,N) to $\overline{g^{(3)}}$ to obtain $x = (S_M^{-1} \otimes S_N^{-1})\Lambda^{-1}(S_M \otimes S_N)\overline{g}$.

The total number of real single flops required is approximately $2 \times 2.5MN\log_2(MN) = 5MN\log_2(MN)$.

1.6.4 Preconditioners for QSC equations

For solving the equations arising from the application of the interior QSC method to general PDEs with homogeneous boundary conditions (1.3)-(1.13), we consider preconditioned iterative methods. It is shown in [5] that the matrix H arising from the discretization of the Helmholtz operator $u_{xx} + u_{yy} - u$ subject to the homogeneous boundary conditions (1.13) can be an effective preconditioner.

The preconditioners considered in [5] are

1. $P = H$,
2. $P = DH$,
3. $P = HD$,
4. $P = \sqrt{D}H\sqrt{D}$,

where D is the diagonal of the matrix system we want to solve. Among these preconditioners, the preconditioner $P = DH$ was found to be slightly more effective than the rest. Therefore, in this thesis we emphasize on this type of preconditioner.

1.7 Multigrid solvers for the QSC equations

Multigrid Methods (MGMs) are efficient methods for accelerating the convergence of iterative methods for solving the linear systems arising from the discretization of PDEs. They have been primarily developed and analyzed for linear systems arising from finite difference methods. The paper [4] develops MGMs for QSC equations arising from one and two-dimensional second-order PDEs.

MGMs involve converting solution or residual vectors from coarse to fine grids or vice-versa. The difficulty in developing MGMs for QSC equations lies in this conversion. In the traditional use of MGMs for finite difference methods, the arising vectors are estimated solution or residual vectors, with components representing values of some underlying function on the data points. Thus, the conversion can be done by interpolating the components of the vector since these components represent function values. However, because the components of the solution vector of the QSC method represent the coefficients of the basis splines and not function values, this method of conversion is not applicable. Instead, a clever relation between the structures of the basis splines for coarse and fine grids is used to formulate the extension and restriction operators, E and R, respectively, for one-dimensional PDE problems.

By modifying the operators E and R in a natural way, the operators can be extended for two-dimensional PDE problems. Alternative extension and restriction operators are presented in [4] for conversion between fine and coarse grids whose step size ratio may be 2 or 4. In addition, in the same paper, numerical results are presented from the application of the MGM to QSC equations arising from problems with homogeneous Dirichlet or general mixed boundary conditions and it is noted that the extension and restriction operators used are adjusted appropriately to the boundary conditions.

1.7.1 Multigrid Methods for a single PDE

Consider solving the equations arising from the application of the interior QSC method to general PDEs with homogeneous boundary conditions (1.3)-(1.13) by an iterative method. The MGM is incorporated into the iterative method as a preconditioner.

At any iteration, the two-grid scheme involves: first applying R to the residual vector to obtain a restricted residual vector corresponding to the coarse grid; then solving the coarse interior QSC system with the restricted residual vector on the right side to find a solution called the corrected residual; and finally applying E to the corrected residual vector to obtain the corrected residual vector corresponding to the fine grid. After the above steps, the corrected residual is added to the approximate solution to form a new approximate solution on the fine grid. We represent this coarse grid correction scheme by the preconditioner P, with $P^{-1} = EK'^{-1}R$, where K' represents the coarse grid QSC matrix. The two-grid scheme can be extended to a scheme on several grids, giving rise to the multigrid method. For convenience, whenever there is no clarity problem, we will use the generic term "multigrid" for both the two-grid as well as the multigrid method.

There are two types of multigrid preconditioning schemes: *multiplicative* and *additive*. For the multiplicative scheme, we assume that, at some iteration k, given an approximate solution $\theta^{(k)}$ to the interior QSC system, we first apply a simple relaxation scheme to get an approximation $\widehat{\theta^{(k)}}$ and the respective residual $\widehat{r^{(k)}}$. Then we apply P^{-1} to $\widehat{r^{(k)}}$

to obtain the preconditioned residual $\widehat{s^{(k)}}$. The approximation $\theta^{(k+1)}$ is obtained by $\theta^{(k+1)} = \widehat{\theta^{(k)}} + \widehat{s^{(k)}}$.

For the additive scheme, we assume that, at some iteration k, given an approximate solution $\theta^{(k)}$ and the respective residual $r^{(k)}$, we apply a simple relaxation scheme to get an approximation $\widehat{\theta^{(k)}}$. Then we apply P^{-1} to $r^{(k)}$ to obtain the preconditioned residual $s^{(k)}$. The approximation $\theta^{(k+1)}$ is obtained by $\theta^{(k+1)} = \widehat{\theta^{(k)}} + s^{(k)}$.

There are several ways of parsing the various grids from the finest to the coarsest and back to the finest. The V-cycle consists of recursive applications of the two-grid method until a coarsest grid level has been reached, while the full multigrid method applies a V-cycle at each grid level on the way up.

Chapter 2

The optimal QSC method for systems of two PDEs

2.1 Formulation of the optimal QSC method

In order to describe the formulation of the optimal QSC method for the system of two PDEs (1.1)-(1.2), we adopt the notation of Chapter 1 and consider the grid partition $\triangle$ of $\overline{\Omega}$, the set Φ of basis functions for the bi-quadratic spline space, and the set T of collocation points, that were defined in Chapter 1.

The optimal two-step QSC method applied to (1.1)-(1.2) consists of first determining the bi-quadratic splines $u^{(1)} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{ij}^{(1)} \phi_i(x)\phi_j(y)$ and $v^{(1)} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \zeta_{ij}^{(1)} \phi_i(x)\phi_j(y)$ such that

$$\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} u^{(1)} \\ v^{(1)} \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \text{ in } T - T_\partial \tag{2.1}$$

and

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} u^{(1)} \\ v^{(1)} \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \text{ on } T_\partial \tag{2.2}$$

then determining the bi-quadratic splines $u^{(2)} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{ij}^{(2)} \phi_i(x)\phi_j(y)$ and $v^{(2)} =$

$\sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\zeta_{ij}^{(2)}\phi_i(x)\phi_j(y)$ such that

$$\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}\begin{bmatrix} u^{(2)} \\ v^{(2)} \end{bmatrix} = \begin{bmatrix} g_1 - P_{L_{11}}u^{(1)} - P_{L_{12}}v^{(1)} \\ g_2 - P_{L_{21}}u^{(1)} - P_{L_{22}}v^{(1)} \end{bmatrix} \text{ in } T - T_\partial \tag{2.3}$$

and

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}\begin{bmatrix} u^{(2)} \\ v^{(2)} \end{bmatrix} = \begin{bmatrix} \gamma_1 - P_{B_{11}}u^{(1)} - P_{B_{12}}v^{(1)} \\ \gamma_2 - P_{B_{21}}u^{(1)} - P_{B_{22}}v^{(1)} \end{bmatrix} \text{ on } T_\partial, \tag{2.4}$$

where, for $i = 1, 2$ and $j = 1, 2$, $P_{L_{ij}}$ and $P_{B_{ij}}$ are perturbations of the operators L_{ij} and B_{ij}, respectively, given by the formulae in [2], that define the perturbations of the operators L and B, respectively, of the single PDE problem (1.3)-(1.4).

2.2 Properties of the QSC matrix

Equations (2.1)-(2.2) give rise to a $2(M+2)(N+2) \times 2(M+2)(N+2)$ linear system, with a matrix of at most 18 nonzero entries per row. Equations (2.3)-(2.4) give rise to a similar linear system with the same matrix, and a perturbed right-side vector. In order to study the sparsity pattern of the matrix of the above linear systems we need to adopt an ordering of the equations and unknowns.

We adopt a bottom-up then left-to-right natural ordering of the collocation points, as in the case of a single PDE. However, in the case of a system of two PDEs, since two equations and two unknowns correspond to a single point, we consider the following orderings of the equations and unknowns.

According to the *alternating* or *point-by-point* ordering, we order the two equations corresponding to a collocation point one after the other, and the respective two unknowns one after the other. With this ordering, the QSC matrix arising from (2.1)-(2.2) (as well as from (2.3)-(2.4)) is block-tridiagonal and has semi-bandwidth $2(N+2)+3$, with each block being of size $2(N+2) \times 2(N+2)$ and having semi-bandwidth 3. Figure 2.1 shows the sparsity pattern of the QSC matrix with the alternating ordering for $N = M = 4$.

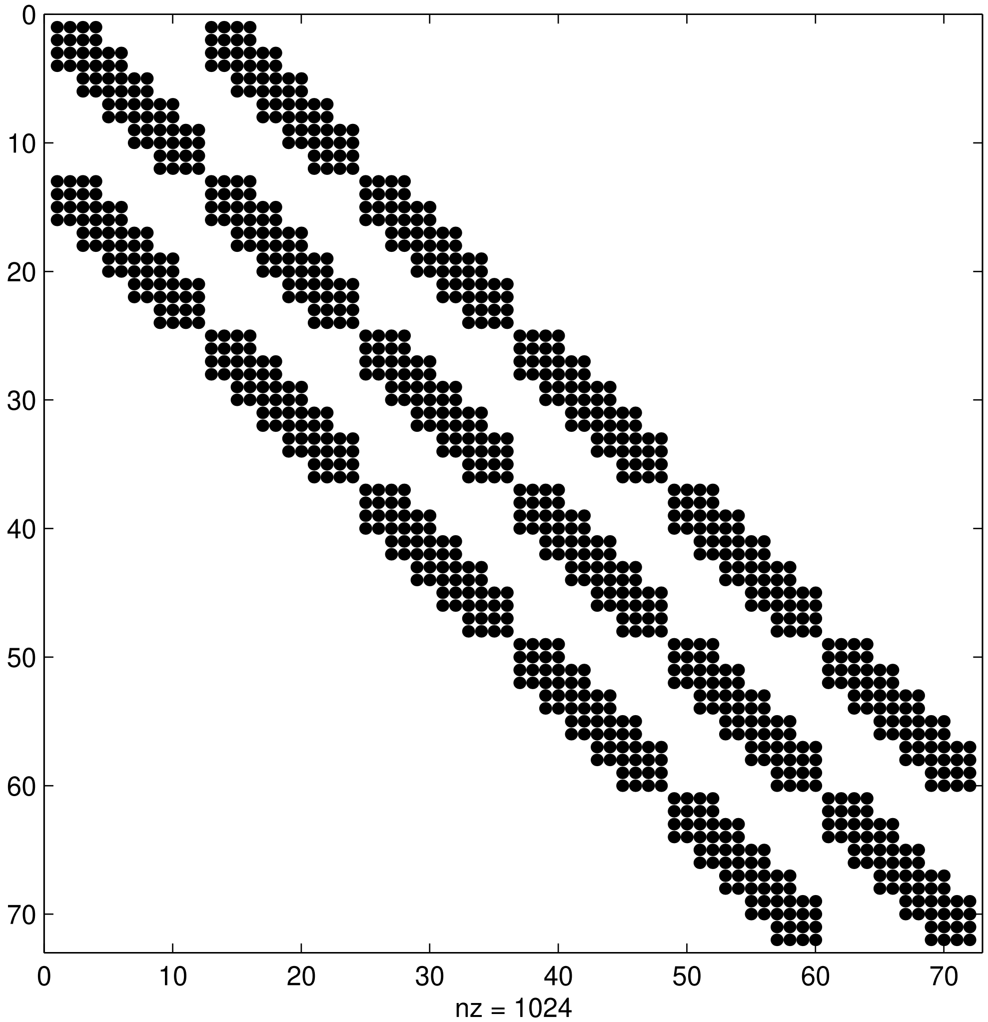

Figure 2.1: The structure of the QSC matrix arising from the discretization of a system of two PDEs (1.1) with general boundary conditions (1.2), for $N = M = 4$. The alternating ordering is adopted. The number of non-zero entries is denoted by nz.

According to the *block* ordering, we order first all linear equations corresponding to the first PDE in (1.1) and first set of boundary conditions in (1.2), always according to the ordering of the points, then all linear equations corresponding to the second PDE in (1.1) and second set of boundary conditions in (1.2), again according to the ordering of the points. The unknown coefficients are ordered similarly, that is, first all those corresponding to $u^{(1)}$, then all those corresponding to $v^{(1)}$. With the block ordering, equations (2.1)-(2.2) result in a 2×2 block linear system. Figure 2.2 shows the sparsity pattern of the QSC matrix with the block ordering for $N = M = 4$.

In order to investigate the details of the block linear system we introduce four auxiliary BVPs. These auxiliary BVPs will also be used in the next section and will facilitate the derivation of error bounds for the QSC approximation to the solution of a system of two PDEs.

Let u and v be the solutions to the problem (1.1)-(1.2). Define the functions g_{ij} and γ_{ij}, for $i = 1, 2$ and $j = 1, 2$, by $g_{11} = L_{11}u$, $\gamma_{11} = B_{11}u$, $g_{21} = L_{21}u$, $\gamma_{21} = B_{21}u$, $g_{12} = L_{12}v$, $\gamma_{12} = B_{12}v$, $g_{22} = L_{22}v$ and $\gamma_{22} = B_{22}u$.

Consider now the following four auxiliary BVPs. The first BVP is defined by the operator equation

$$L_{11}u = g_{11} \text{ in } \Omega \tag{2.5}$$

subject to boundary conditions

$$B_{11}u = \gamma_{11} \text{ on } \partial\Omega. \tag{2.6}$$

By construction, the solution u of (2.5)-(2.6) is the same as the function u that satisfies (1.1)-(1.2).

The second BVP is defined by the operator equation

$$L_{21}u = g_{21} \text{ in } \Omega \tag{2.7}$$

subject to boundary conditions

$$B_{21}u = \gamma_{21} \text{ on } \partial\Omega. \tag{2.8}$$

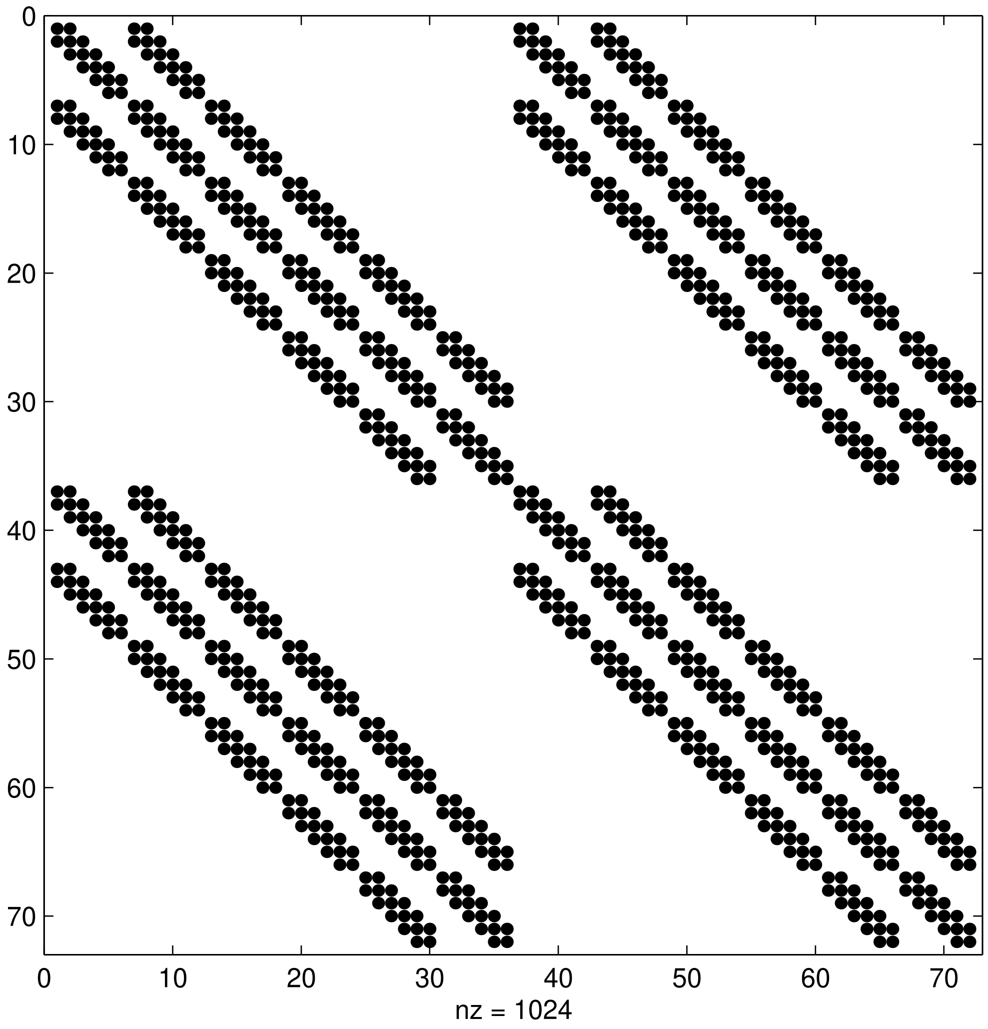

Figure 2.2: The structure of the QSC matrix arising from the discretization of a system of two PDEs (1.1) with general boundary conditions (1.2), for $N = M = 4$. The block ordering is adopted. The number of non-zero entries is denoted by nz.

By construction, the solution u of (2.7)-(2.8) is the same as the function u that satisfies (1.1)-(1.2).

The third BVP is defined by the operator equation

$$L_{12}v = g_{12} \text{ in } \Omega \tag{2.9}$$

subject to boundary conditions

$$B_{12}v = \gamma_{12} \text{ on } \partial\Omega. \tag{2.10}$$

By construction, the solution v of (2.9)-(2.10) is the same as the function v that satisfies (1.1)-(1.2).

Finally, the fourth BVP is defined by the operator equation

$$L_{22}v = g_{22} \text{ in } \Omega \tag{2.11}$$

subject to boundary conditions

$$B_{22}v = \gamma_{22} \text{ on } \partial\Omega. \tag{2.12}$$

By construction, the solution v of (2.11)-(2.12) is the same as the function v that satisfies (1.1)-(1.2).

Consider discretizing the above four BVPs by the optimal two-step QSC method developed in [2] and summarized in Chapter 1. Let $u^{\alpha 1} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\theta_{ij}^{\alpha 1}\phi_i(x)\phi_j(y)$ and $u^{\alpha 2} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\theta_{ij}^{\alpha 2}\phi_i(x)\phi_j(y)$ be the bi-quadratic splines computed by the first and second step, respectively, of the QSC method applied to (2.5)-(2.6). The equations of the first step of the QSC method give rise to a linear system

$$K_{11}\overline{\theta^{\alpha 1}} = \overline{g_{11}}, \tag{2.13}$$

while the QSC equations of the second step give rise to a linear system

$$K_{11}\overline{\theta^{\alpha 2}} = \overline{g_{11}} - \overline{P_{11}u^{\alpha 1}}, \tag{2.14}$$

where $\overline{P_{11}u^{\alpha 1}}$ is the perturbation of $\overline{g_{11}}$ defined based on the perturbation operators $P_{L_{11}}$ and $P_{B_{11}}$ and the solution $u^{\alpha 1}$ obtained in the first step of the QSC method.

Let $u^{\beta 1} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \theta_{ij}^{\beta 1}\phi_i(x)\phi_j(y)$ and $u^{\beta 2} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \theta_{ij}^{\beta 2}\phi_i(x)\phi_j(y)$ be the bi-quadratic splines computed by the first and second step, respectively, of the QSC method applied to (2.7)-(2.8). The equations of the first step of the QSC method give rise to a linear system

$$K_{21}\overline{\theta^{\beta 1}} = \overline{g_{21}}, \tag{2.15}$$

while the QSC equations of the second step give rise to a linear system

$$K_{21}\overline{\theta^{\beta 2}} = \overline{g_{21}} - \overline{P_{21}u^{\beta 1}}, \tag{2.16}$$

where $\overline{P_{21}u^{\beta 1}}$ is the perturbation of $\overline{g_{21}}$ defined based on the perturbation operators $P_{L_{21}}$ and $P_{B_{21}}$ and the solution $u^{\beta 1}$ obtained in the first step of the QSC method.

Let $v^{\alpha 1} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \zeta_{ij}^{\alpha 1}\phi_i(x)\phi_j(y)$ and $v^{\alpha 2} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \zeta_{ij}^{\alpha 2}\phi_i(x)\phi_j(y)$ be the bi-quadratic splines computed by the first and second step, respectively, of the QSC method applied to (2.9)-(2.10). The equations of the first step of the QSC method give rise to a linear system

$$K_{12}\overline{\zeta^{\alpha 1}} = \overline{g_{12}}, \tag{2.17}$$

while the QSC equations of the second step give rise to a linear system

$$K_{12}\overline{\zeta^{\alpha 2}} = \overline{g_{12}} - \overline{P_{12}v^{\alpha 1}}, \tag{2.18}$$

where $\overline{P_{12}v^{\alpha 1}}$ is the perturbation of $\overline{g_{12}}$ defined based on the perturbation operators $P_{L_{12}}$ and $P_{B_{12}}$ and the solution $v^{\alpha 1}$ obtained in the first step of the QSC method.

Let $v^{\beta 1} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \zeta_{ij}^{\beta 1}\phi_i(x)\phi_j(y)$ and $v^{\beta 2} = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1} \zeta_{ij}^{\beta 2}\phi_i(x)\phi_j(y)$ be the bi-quadratic splines computed by the first and second step, respectively, of the QSC method applied to (2.11)-(2.12). The equations of the first step of the QSC method give rise to a linear system

$$K_{22}\overline{\zeta^{\beta 1}} = \overline{g_{22}}, \tag{2.19}$$

while the QSC equations of the second step give rise to a linear system

$$K_{22}\overline{\zeta^{\beta 2}} = \overline{g_{22}} - \overline{P_{22}v^{\beta 1}}, \tag{2.20}$$

where $\overline{P_{22}v^{\beta 1}}$ is the perturbation of $\overline{g_{22}}$ defined based on the perturbation operators $P_{L_{22}}$ and $P_{B_{22}}$ and the solution $v^{\beta 1}$ obtained in the first step of the QSC method.

We are now ready to show the details of the 2×2 block linear system that arises from the equations (2.1)-(2.2) ordered according to the block ordering. The 2×2 block linear system takes the form

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} \overline{\theta^{(1)}} \\ \overline{\zeta^{(1)}} \end{bmatrix} = \begin{bmatrix} \overline{g_{11}} + \overline{g_{12}} \\ \overline{g_{21}} + \overline{g_{22}} \end{bmatrix}, \tag{2.21}$$

where K_{ij} and $\overline{g_{ij}}$, for $i = 1,2$ and $j = 1,2$, were defined with the help of the four auxiliary BVPs introduced above.

Equations (2.3)-(2.4) ordered according to the block ordering give rise to a 2×2 block linear system of the form

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} \overline{\theta^{(2)}} \\ \overline{\zeta^{(2)}} \end{bmatrix} = \begin{bmatrix} \overline{g_{11}} - \overline{P_{11}u^{(1)}} + \overline{g_{12}} - \overline{P_{12}v^{(1)}} \\ \overline{g_{21}} - \overline{P_{21}u^{(1)}} + \overline{g_{22}} - \overline{P_{22}v^{(1)}} \end{bmatrix}, \tag{2.22}$$

where $\overline{P_{11}u^{(1)}}$ is a perturbation of $\overline{g_{11}}$, defined based on the perturbation operators $P_{L_{11}}$ and $P_{B_{11}}$ and the solution $u^{(1)}$. Similarly, $\overline{P_{21}u^{(1)}}$, $\overline{P_{12}v^{(1)}}$ and $\overline{P_{22}v^{(1)}}$ are appropriate perturbations of $\overline{g_{12}}$, $\overline{g_{21}}$, $\overline{g_{22}}$, respectively.

Note that, if $\overline{g_1}$ and $\overline{g_2}$ are the vectors of values of g_1 and g_2, respectively, at the collocation points in the natural ordering, then by construction

$$\overline{g_{11}} + \overline{g_{12}} = \overline{g_1} \tag{2.23}$$

and

$$\overline{g_{21}} + \overline{g_{22}} = \overline{g_2}. \tag{2.24}$$

2.3 Error bounds of optimal QSC

In this section, we prove that, under certain conditions, the QSC approximations $u^{(1)}$ and $v^{(1)}$ defined by equations (2.1)-(2.2) and corresponding to the system of two PDEs (1.1)-(1.2) satisfy the same error bounds as the QSC approximation $u_{\triangle}^{(1)}$ defined by equations (1.6) and corresponding to the single PDE problem (1.3)-(1.4), and the QSC approximations $u^{(2)}$ and $v^{(2)}$ defined by equations (2.3)-(2.4) and corresponding to the system of two PDEs (1.1)-(1.2) satisfy the same error bounds as the QSC approximation $u_{\triangle}^{(2)}$ defined by equations (1.8) and corresponding to the single PDE problem (1.3)-(1.4).

THEOREM 1 *Assume that* K_{22} *and* $(K_{11}-K_{12}K_{22}^{-1}K_{21})$ *(or* K_{11} *and* $(K_{22}-K_{21}K_{11}^{-1}K_{12})$*) are invertible. Then (2.21) and (2.22) are uniquely solvable. Moreover, assume that* $||K_{ij}^{-1}||_\infty$, $i=1,2$, $j=1,2$, *are bounded independently of* $h=\max(h_x,h_y)$, *and that the spectral norms of* $(K_{11}-K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}$, $(K_{11}-K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}K_{22}^{-1}K_{21}$, $(K_{22}-K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}$, *and* $(K_{22}-K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}K_{11}^{-1}K_{12}$ *are bounded independently of* h. *Let* $\epsilon^{(1)}=u^{(1)}-u$ *and* $\delta^{(1)}=v^{(1)}-v$ *be errors of the QSC approximations* $u^{(1)}$ *and* $v^{(1)}$, *respectively, defined by equations (2.1)-(2.2). Let* $\epsilon^{(2)}=u^{(2)}-u$ *and* $\delta^{(2)}=v^{(2)}-v$ *be errors of the QSC approximations* $u^{(2)}$ *and* $v^{(2)}$, *respectively, defined by equations (2.3)-(2.4). Then, the following a priori bounds hold:*

$$\begin{aligned}
|\epsilon^{(1)}(x_i,y_j)| &= O(h^2), \quad ||\epsilon^{(1)}||_\infty = O(h^2)\\
|\epsilon^{(1)}(\tau_i^x,y_j)| &= O(h^2),\\
|\epsilon^{(1)}(x_i,\tau_j^y)| &= O(h^2),\\
|\epsilon^{(1)}(\tau_i^x,\tau_j^y)| &= O(h^2),
\end{aligned}\tag{2.25}$$

$$\begin{aligned}
|\epsilon^{(2)}(x_i,y_j)| &= O(h^4), & ||\epsilon^{(2)}||_\infty &= O(h^3)\\
|\epsilon^{(2)}(\tau_i^x,y_j)| &= O(h^4), & &\\
|\epsilon^{(2)}(x_i,\tau_j^y)| &= O(h^4), & &\\
|\epsilon^{(2)}(\tau_i^x,\tau_j^y)| &= O(h^4), & &\\
|D_x\epsilon^{(2)}(x_i-\lambda h_x,.)| &= O(h^3), & ||D_x\epsilon^{(2)}||_\infty &= O(h^2)\\
|D_y\epsilon^{(2)}(.,y_i-\lambda h_y)| &= O(h^3), & ||D_y\epsilon^{(2)}||_\infty &= O(h^2)\\
|D_{xy}\epsilon^{(2)}(x_i-\lambda h_x,y_i-\lambda h_y)| &= O(h^2), & ||D_{xy}\epsilon^{(2)}||_\infty &= O(h^2)\\
|D_x^2\epsilon^{(2)}(\tau_i^x,\tau_j^y)| &= O(h^2), & ||D_x^2\epsilon^{(2)}||_\infty &= O(h)\\
|D_y^2\epsilon^{(2)}(\tau_i^x,\tau_j^y)| &= O(h^2), & ||D_y^2\epsilon^{(2)}||_\infty &= O(h)
\end{aligned}\tag{2.26}$$

where $h=\max(h_x,h_y)$, $\lambda=(3\pm\sqrt{3})/6$ *and* D_z^n *denotes the nth derivative operator with respect to the variable* z. *Bounds identical to those of* $\epsilon^{(1)}$ *hold for* $\delta^{(1)}$, *while bounds identical to those of* $\epsilon^{(2)}$ *hold for* $\delta^{(2)}$.

PROOF: We first prove the unique solvability of (2.21) under the conditions of the theorem. Block Gauss elimination applied to the 2×2 block system (2.21) yields

$$(K_{11} - K_{12}K_{22}^{-1}K_{21})\overline{\theta^{(1)}} = \overline{g_{11}} + \overline{g_{12}} - K_{12}K_{22}^{-1}(\overline{g_{21}} + \overline{g_{22}}). \tag{2.27}$$

Since K_{22} and $(K_{11} - K_{12}K_{22}^{-1}K_{21})$ are invertible, $\overline{\theta^{(1)}}$ is uniquely determined by (2.27). Substitution into the second block equation of (2.21) uniquely determines $\overline{\zeta^{(1)}}$.

Without loss of generality, if K_{11} and $(K_{22} - K_{21}K_{11}^{-1}K_{12})$ are invertible, $\overline{\zeta^{(1)}}$ and then $\overline{\theta^{(1)}}$ are uniquely determined. In a similar manner, $\overline{\theta^{(2)}}$ and $\overline{\zeta^{(2)}}$ are uniquely determined. This completes the proof of the first part of the theorem.

We now prove the error bounds (2.25). The boundedness of $||K_{ij}^{-1}||_\infty$, for $i = 1, 2$ and $j = 1, 2$, allows us to apply the results of [2] to the four auxiliary BVPs defined in the previous section and get the error bounds $||u - u^{\alpha 1}||_\infty = O(h^2)$, $||u - u^{\beta 1}||_\infty = O(h^2)$, $||v - v^{\alpha 1}||_\infty = O(h^2)$ and $||v - v^{\beta 1}||_\infty = O(h^2)$. By the triangle inequality, we have $||u^{\alpha 1} - u^{\beta 1}||_\infty = O(h^2)$ and $||v^{\alpha 1} - v^{\beta 1}||_\infty = O(h^2)$. It is known [2] that the inverse of the bi-quadratic interpolation matrix is bounded independently of h. Thus, $||\theta^{\alpha 1} - \theta^{\beta 1}||_\infty = O(h^2)$ and $||\zeta^{\alpha 1} - \zeta^{\beta 1}||_\infty = O(h^2)$. So, we have

$$\overline{\theta^{\alpha 1}} = \overline{\theta^{\beta 1}} + \overline{p}h^2 \text{ and} \tag{2.28}$$

$$\overline{\zeta^{\alpha 1}} = \overline{\zeta^{\beta 1}} + \overline{q}h^2, \tag{2.29}$$

where $\overline{p}$ and $\overline{q}$ are vectors of scalars independent of h. Using (2.15) and (2.28) we get

$$K_{21}\overline{\theta^{\alpha 1}} = K_{21}\overline{\theta^{\beta 1}} + K_{21}\overline{p}h^2 = \overline{g_{21}} + K_{21}\overline{p}h^2 \tag{2.30}$$

and

$$K_{12}K_{22}^{-1}K_{21}\overline{\theta^{\alpha 1}} = K_{12}K_{22}^{-1}\overline{g_{21}} + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2. \tag{2.31}$$

From the above equation and (2.13) we get

$$(K_{11} - K_{12}K_{22}^{-1}K_{21})\overline{\theta^{\alpha 1}} = \overline{g_{11}} - K_{12}K_{22}^{-1}\overline{g_{21}} - K_{12}K_{22}^{-1}K_{21}\overline{p}h^2. \tag{2.32}$$

Subtracting (2.32) from (2.27) results in

$$(K_{11} - K_{12}K_{22}^{-1}K_{21})(\overline{\theta^{(1)}} - \overline{\theta^{\alpha 1}}) = \overline{g_{12}} - K_{12}K_{22}^{-1}\overline{g_{22}} + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2. \tag{2.33}$$

Using (2.17) and (2.19) and then (2.29) the right side of (2.33) becomes $\overline{g_{12}} - K_{12}K_{22}^{-1}\overline{g_{22}} + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2 = K_{12}\overline{\zeta^{\alpha 1}} - K_{12}\overline{\zeta^{\beta 1}} + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2 = K_{12}\overline{q}h^2 + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2$. Thus, from (2.33) we have

$(K_{11} - K_{12}K_{22}^{-1}K_{21})(\overline{\theta^{(1)}} - \overline{\theta^{\alpha 1}}) = K_{12}\overline{q}h^2 + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2$

$\Rightarrow (\overline{\theta^{(1)}} - \overline{\theta^{\alpha 1}}) = (K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}(K_{12}\overline{q}h^2 + K_{12}K_{22}^{-1}K_{21}\overline{p}h^2)$.

Hence, $||\overline{\theta^{(1)}} - \overline{\theta^{\alpha 1}}||_\infty \leq$

$(||(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}||_\infty + ||(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}K_{22}^{-1}K_{21}||_\infty)O(h^2)$.

Under the assumptions of the theorem, the inverses of K_{22} and $(K_{11} - K_{12}K_{22}^{-1}K_{21})$ exist and the norms $||(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}||$, and $||(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}K_{22}^{-1}K_{21}||$ are bounded independently of h. Therefore, $||\overline{\theta^{(1)}} - \overline{\theta^{\alpha 1}}||_\infty = O(h^2)$. This result and the definition of the basis functions prove that $u^{(1)} - u^{\alpha 1}$ satisfies global and local error bounds identical to those in (2.25). The error bounds (2.25) for $\epsilon^{(1)}$ now follow from (1.11), the error bounds for $u^{(1)} - u^{\alpha 1}$ and the use of the triangle inequality.

Similarly, we have

$||\overline{\zeta^{(1)}} - \overline{\zeta^{\alpha 1}}||_\infty \leq$

$(||(K_{22} - K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}||_\infty + ||(K_{22} - K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}K_{11}^{-1}K_{12}||_\infty)O(h^2)$.

Hence, error bounds for $\delta^{(1)}$ identical to those in (2.25) also follow.

Finally, we prove the error bounds (2.26). Consider yet another four auxiliary problems. Let $u^\gamma = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\theta_{ij}^\gamma\phi_i(x)\phi_j(y)$ be the bi-quadratic spline that satisfies

$$L_{11}u^\gamma = g_{11} - P_{L_{11}}u^{(1)} \text{ in } T - T\partial, \;\; B_{11}u^\gamma = g_{11} - P_{B_{11}}u^{(1)} \text{ on } T\partial, \tag{2.34}$$

$u^\delta = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\theta_{ij}^\delta\phi_i(x)\phi_j(y)$ be the bi-quadratic spline that satisfies

$$L_{21}u^\delta = g_{21} - P_{L_{21}}u^{(1)} \text{ in } T - T\partial, \;\; B_{21}u^\delta = g_{21} - P_{B_{21}}u^{(1)} \text{ on } T\partial,$$

$v^\gamma = \sum_{i=0}^{M+1}\sum_{j=0}^{N+1}\zeta_{ij}^\gamma\phi_i(x)\phi_j(y)$ be the bi-quadratic spline that satisfies

$$L_{12}v^\gamma = g_{12} - P_{L_{12}}v^{(1)} \text{ in } T - T\partial, \;\; B_{12}v^\gamma = g_{12} - P_{B_{12}}v^{(1)} \text{ on } T\partial,$$

$v^\delta = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{ij}^\delta \phi_i(x)\phi_j(y)$ be the bi-quadratic spline that satisfies

$$L_{22}v^\delta = g_{22} - P_{L_{22}} v^{(1)} \text{ in } T - T\partial, \;\; B_{22}v^\delta = g_{22} - P_{B_{22}} v^{(1)} \text{ on } T\partial,$$

Note that the bi-quadratic spline $u^{\alpha 2} = \sum_{i=0}^{M+1} \sum_{j=0}^{N+1} \theta_{ij}^{\alpha 2} \phi_i(x)\phi_j(y)$ computed by the second step of the QSC method applied to (2.5)-(2.6), as defined in Section 2.2, satisfies

$$L_{11}u^{\alpha 2} = g_{11} - P_{L_{11}} u^{\alpha 1} \text{ in } T - T\partial, \;\; B_{11}u^{\alpha 2} = g_{11} - P_{B_{11}} u^{\alpha 1} \text{ on } T\partial.$$

From the above relations and (2.34) we get

$$L_{11}(u^{\alpha 2} - u^\gamma) = P_{L_{11}}(u^{(1)} - u^{\alpha 1}) \text{ in } T - T\partial, \;\; B_{11}(u^{\alpha 2} - u^\gamma) = P_{B_{11}}(u^{(1)} - u^{\alpha 1}) \text{ on } T\partial.$$

We have already shown that $||u^{(1)} - u^{\alpha 1}||_\infty = O(h^2)$. Since P_L and P_B are $O(h^2)$ perturbation operators, assuming the terms of the expansion of $u^{(1)} - u^{\alpha 1}$ are sufficiently smooth, we get

$$L_{11}(u^{\alpha 2} - u^\gamma) = O(h^4) \text{ in } T - T\partial, \;\; B_{11}(u^{\alpha 2} - u^\gamma) = O(h^4) \text{ on } T\partial$$

which can be equivalently written in matrix form as follows

$$K_{11}(\overline{\theta^{\alpha 2}} - \overline{\theta^\gamma}) = O(\overline{h^4}).$$

From the above we obtain

$$||\overline{\theta^{\alpha 2}} - \overline{\theta^\gamma}||_\infty = O(h^4).$$

This result and the definition of the basis functions prove that $||u^{\alpha 2} - u^\gamma||_\infty = O(h^4)$. Since $||u - u^{\alpha 2}||_\infty = O(h^4)$ by (1.12), using the triangle inequality, we have $||u - u^\gamma||_\infty = O(h^4)$. Similarly, $||u - u^\delta||_\infty = O(h^4)$, $||v - v^\gamma||_\infty = O(h^4)$ and $||v - v^\delta||_\infty = O(h^4)$. Thus, as before, we can show

$$\begin{aligned} \overline{\theta^\gamma} &= \overline{\theta^\delta} + \overline{p}h^4, \\ \overline{\zeta^\gamma} &= \overline{\zeta^\delta} + \overline{q}h^4, \end{aligned}$$

where $\overline{p}$ and $\overline{q}$ are vectors of scalars independent of h.

By defining $\widehat{g_{i1}} = \overline{g_{i1}} - \overline{P_{i1}u^{(1)}}$ and $\widehat{g_{i2}} = \overline{g_{i2}} - \overline{P_{i2}v^{(1)}}$, for $i = 1, 2$, we can proceed similarly as before to show that $||\overline{\theta^{(2)}} - \overline{\theta^{\gamma}}||_{\infty} = O(h^4)$. This result and the definition of the basis functions prove that $u^{(2)} - u^{\gamma}$ and its derivatives satisfy global and local error bounds identical to those in (2.26). The error bounds (2.26) for $\epsilon^{(2)}$ now follow from (1.12), the error bounds for $u^{(2)} - u^{\gamma}$ and the use of the triangle inequality. Error bounds identical to those in (2.26) for $\delta^{(2)}$ can be shown similarly. $\diamondsuit$

2.3.1 Helmholtz operators with constant coefficients

We can extend the interior QSC method for solving one PDE with homogeneous Dirichlet boundary conditions to the interior QSC method for solving a system of PDEs with homogeneous Dirichlet boundary conditions for both unknowns u and v in the same way we extend the QSC method for solving one PDE to the QSC method for solving a system of PDEs. The interior QSC matrix for solving a system of two PDEs with homogeneous Dirichlet boundary conditions for both unknowns u and v gives rise to a $2MN \times 2MN$ linear system of equations, which can be ordered according to either the alternating or the block ordering.

Consider the Helmholtz problems

$$L_{11}u \equiv a_{11}u_{xx} + c_{11}u_{yy} + f_{11}u = g_{11} \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{2.35}$$

$$L_{12}v \equiv a_{12}v_{xx} + c_{12}v_{yy} + f_{12}v = g_{12} \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega, \tag{2.36}$$

$$L_{21}u \equiv a_{21}u_{xx} + c_{21}u_{yy} + f_{21}u = g_{21} \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{2.37}$$

$$L_{22}v \equiv a_{22}v_{xx} + c_{22}v_{yy} + f_{22}v = g_{22} \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega, \tag{2.38}$$

where a_{ij}, c_{ij}, and f_{ij}, $i, j = 1, 2$, are constants. Let $K_{11}, K_{12}, K_{21}, K_{22}$ be the $MN \times MN$ matrices from the application of the interior QSC method to the PDE problems (2.35), (2.36), (2.37), (2.38), respectively.

LEMMA 1 *Let* $p_i(x,y) = a_i x^2 + b_i xy + c_i y^2 + d_i x + e_i y + f_i, i = 1,2$ *where the coefficients* a_i, b_i, c_i, d_i, e_i *and* f_i *are constants and let* $g(x,y) = \frac{p_2(x,y)}{p_1(x,y)}$. *If* $b_1^2 - 4a_1c_1 < 0$ *and* $p_1(x,y) \neq 0 \; \forall x, y$*, then* $g(x,y)$ *is bounded* $\forall x, y$.

PROOF: Since g is a continuous function, it is bounded in any closed interval. In particular, it is bounded in $[-1,1] \times [-1,1]$. Consider the set $\{|x| > 1, |y| > 1\}$. In this set, $|b_2xy| < |b_2|(x^2+y^2), |d_2x| < |d_2|x^2, |e_2y| < |e_2|y^2, |f_2| < |f_2|(x^2+y^2)$. So, $|p_2(x,y)| < Ax^2 + By^2$, where $A = |a_2|+|b_2|+|d_2|+|f_2|$ and $B = |c_2|+|b_2|+|e_2|+|f_2|$. Now, consider the set $\{|x| > 1, |y| \leq 1\}$. In this set, $|a_2x^2| \leq |a_2|x^2, |b_2xy| < |b_2|x^2, |c_2y^2| < |c_2|x^2, |d_2x| < |d_2|x^2, |e_2y| < |e_2|x^2, |f_2| < |f_2|x^2$. So, $|p_2(x,y)| < Cx^2$, where $C = |a_2| + |b_2| + |c_2| + |d_2| + |e_2| + |f_2|$. Similarly, in the set $\{|x| \leq 1, |y| > 1\}$, we have $|p_2(x,y)| < Cy^2$. Thus, we just have to show (by symmetry) that $\frac{x^2}{p_1(x,y)}$ is bounded.

Since $p_1(x,y)$ is non-zero, we may assume without loss of generality that it is strictly positive, that is $a_1, c_1 > 0$. First fix x and then minimize $p_1(x,y)$ over all y by finding its derivative with respect to y and setting it to 0. This minimum is attained when $y = \frac{b_1x+d_1}{-2c_1}, c_1 \neq 0$. At this y, $p_1(x,y)$ can be written as a quadratic $Px^2 + Qx + R$, where $P = \frac{b_1^2 - 4a_1c_1}{-4c_1}$, and Q, R are constants. Since $b_1^2 - 4a_1c_1 < 0$ and $c_1 > 0$, we have $P > 0$.

Now $\frac{x^2}{p_1(x,y)} \leq \frac{x^2}{Px^2+Qx+R}$. The limit of $\frac{x^2}{Px^2+Qx+R}$ as $x \to \pm\infty$ is $1/P, P \neq 0$. Hence $\forall \epsilon > 0$, $\exists t > 0$ such that $|\frac{x^2}{p_1(x,y)}| < 1/P + \epsilon$, whenever $|x| > t$. Since $\frac{x^2}{Px^2+Qx+R}$ is a continuous function, it is also bounded in the closed interval $[-t, t]$. Thus, $\frac{x^2}{p_1(x,y)}$ is bounded $\forall x, y$. $\diamondsuit$

THEOREM 2 *Let*

$$
\begin{aligned}
A_1 &= a_{22}a_{11} - a_{12}a_{21}, & C_1 &= c_{22}c_{11} - c_{12}c_{21}, \\
B_1 &= a_{22}c_{11} + a_{11}c_{22} - a_{12}c_{21} - a_{21}c_{12}, & F_1 &= \frac{1}{64}(f_{22}f_{11} - f_{12}f_{21}) \\
D_1 &= \frac{1}{8}(a_{22}f_{11} + a_{11}f_{22} - a_{12}f_{21} - a_{21}f_{12}), & E_1 &= \frac{1}{8}(c_{22}f_{11} + c_{11}f_{22} - c_{12}f_{21} - c_{21}f_{12}).
\end{aligned} \tag{2.39}
$$

Let

$$
P_1(x,y) = A_1x^2 + B_1xy + C_1y^2 + D_1x + E_1y + F_1. \tag{2.40}
$$

Suppose $B_1^2 - 4A_1C_1 < 0$, and K_{11} and K_{22} are invertible. If $A_1 \cdot P_1(S,T) > 0$, where

$$S = \frac{2C_1D_1 - B_1E_1}{B_1^2 - 4A_1C_1}, \text{ and } T = \frac{2E_1A_1 - B_1D_1}{B_1^2 - 4A_1C_1}, \tag{2.41}$$

then the linear systems (2.21) and (2.22) are uniquely solvable. Moreover, the spectral norms of $(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}$, $(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}K_{22}^{-1}K_{21}$, $(K_{22} - K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}$, and $(K_{22} - K_{21}K_{11}^{-1}K_{12})^{-1}K_{21}K_{11}^{-1}K_{12}$ are bounded independently of $h = \max(h_x, h_y)$. Furthermore, if K_{12} and K_{21} are invertible, then the solutions to the linear systems (2.21) and (2.22) have the same error bounds as in Theorem 1.

PROOF: The matrix $(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}$ has eigenvalues

$$\lambda_{lm} = \frac{\rho_{12}^{lm}}{\rho_{11}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm}/\rho_{22}^{lm}} = \frac{\rho_{12}^{lm}\rho_{22}^{lm}}{\rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm}},$$

where ρ^{lm} is defined in (1.17). Notice that $\rho_{22}^{lm} \neq 0$ because we have assumed that K_{22} is invertible. Dividing both the numerator and denominator of λ_{lm} by $(\lambda_l + 8)(\lambda_m + 8)$, we have

$$\lambda_{lm} = \frac{\gamma_{12}^{lm}\gamma_{22}^{lm}}{\gamma_{11}^{lm}\gamma_{22}^{lm} - \gamma_{12}^{lm}\gamma_{21}^{lm}},$$

where $\gamma_{ij}^{lm} = a_{ij}w_l + c_{ij}w_m + \frac{1}{8}f_{ij}$, $w_l = \frac{\lambda_l}{h_x^2(\lambda_l+8)}$, $w_m = \frac{\lambda_m}{h_y^2(\lambda_m+8)}$. Simplifying this, we have

$$\lambda_{lm} = \frac{P_2(w_l, w_m)}{P_1(w_l, w_m)},$$

where $P_2(w_l, w_m)$ is a second-degree polynomial with variables w_l and w_m, and $P_1(w_l, w_m)$ is given by (2.40) with variables w_l and w_m.

The restriction $B_1^2 - 4A_1C_1 < 0$ implies that P_1 is a paraboloid. The assumptions that $A_1 > 0$ and $P_1(S,T) > 0$, where S and T are given by (2.41), imply that P_1 has a positive minimum, while the assumptions $A_1 < 0$ and $P_1(S,T) < 0$ imply that P_1 has a negative maximum. Hence, $P_1(w_l, w_m) \neq 0$, $\forall w_l, w_m$. Since the eigenvalues of $K_{11} - K_{12}K_{22}^{-1}K_{21}$ and $K_{22} - K_{21}K_{11}^{-1}K_{12}$ are $P_1(w_l, w_m)/\gamma_{22}^{lm}$ and $P_1(w_l, w_m)/\gamma_{11}^{lm}$, respectively, matrices $K_{11} - K_{12}K_{22}^{-1}K_{21}$ and $K_{22} - K_{21}K_{11}^{-1}K_{12}$ have non-zero eigenvalues. Hence by Theorem 1, (2.21) and (2.22) are uniquely solvable.

Moreover since the assumptions of Lemma 1 are satisfied, λ_{lm} is bounded $\forall w_l, w_m$ independently of h. Hence, the spectral norm of $(K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}K_{12}$ is bounded independently of h. Similarly, we can show that the other spectral norms are bounded independently of h.

The final result of this theorem now follows since all the assumptions of Theorem 1 are satisfied. $\diamondsuit$

2.4 Numerical results

In this section we present numerical results to demonstrate the convergence and computational complexity of the QSC method for systems of PDEs. All computations in this section were carried out on a Sun Ultra-4 in double precision. The QSC methods were programmed by us in both FORTRAN and MATLAB. Most results in this thesis are obtained from the FORTRAN implementation. The linear systems arising from Problems 1-4 were solved by banded Gauss elimination without pivoting using the FORTRAN ELLPACK routines *q5bnfa,q5bnsl*. The linear system arising in Problem 5 was solved by Gauss elimination using the backslash operator or the *lu* function in MATLAB, since the finite difference discretization was programmed only in MATLAB.

In all tables, the notation $x.y \pm z$ means $x.y \times 10^{\pm z}$. The errors $\epsilon^{(1)}$, $\delta^{(1)}$, $\epsilon^{(2)}$ and $\delta^{(2)}$ are defined in Theorem 1. The problems are given in the Appendix.

Problem 1 is a system of PDEs with general coefficients. The first component, u, is $x^{13/2}y^{13/2}$, that is, u has finite differentiability in the domain of the problem definition. Table 2.1 shows that the errors and respective orders of convergence of the solution obtained by the first and second step of the QSC method conform with that predicted by Theorem 1. Hence in this experiment, the QSC method produces optimal solutions.

Problem 2 is a system of PDEs with mixed boundary conditions. The coefficients are constants chosen to satisfy $B_1^2 - 4A_1C_1 < 0$, where A_1, B_1, and C_1 are defined in (2.39).

Note that this problem is more general than the one assumed in Theorem 2 in that it has first-order derivative terms and mixed boundary conditions. Table 2.2 shows that, in the second step of the QSC method, the order of convergence of the solution at the grid points and midpoints is optimal, $O(h^4)$. The experimental global order of convergence of $O(h^4)$ in the second step for both components of the solutions is better than the $O(h^3)$ predicted in Theorem 1. It is instructive to note that the assumptions of Theorems 1 and 2 are only sufficient.

Problem 3 is a system of Helmholtz PDEs with homogeneous Dirichlet conditions. The coefficients are constants chosen to satisfy $B_1^2 - 4A_1C_1 < 0$. This problem satisfies all the assumptions of Theorem 2 and thus all the assumptions of Theorem 1. Table 2.3 shows that, in the second step of the interior QSC method, the orders of convergence of the solution at the grid points and midpoints are optimal, $O(h^4)$. The experimental global order of convergence in the second step for both component of the solution is $O(h^3)$. These experimental orders of convergence agree with those predicted by Theorem 1.

Problem 4 is a system of Helmholtz PDEs with homogeneous Dirichlet conditions. The coefficients are constants chosen to satisfy $B_1^2 - 4A_1C_1 > 0$. Note that Problem 4 differs from Problem 3 only in the coefficient of the u_{xx} term in the second PDE. Table 2.4 shows that, in both steps of the interior QSC method, the errors and orders of convergence for the spline solutions are volatile. This experiment and the last one demonstrate that the discriminant $B_1^2 - 4A_1C_1$ can be a good indicator of the quality of the order of convergence for the solution.

Problem 5 was taken from [10] and describes a two-dimensional stress-strain relation. The problem was solved by the standard second order finite difference method and the interior QSC method. Let $\epsilon^{(FD)}$ and $\delta^{(FD)}$ denote the error at the grid points of the u, v components of the the solution obtained by the finite difference method, respectively. Table 2.5 and Figure 2.3 clearly show that the QSC method is superior to the finite difference one in terms of solution errors and order of convergence at the grid points.

Table 2.6 shows the number of floating-point operations (flops) required for the solution of Problem 5 by the finite difference method and the interior QSC method in MATLAB. We used two different solution techniques, the built-in backslash operator "\" and the LU factorization routine *lu* followed by forward and backward (F/B) substitutions performed by the backslash operator. In order to apply the LU factorization effectively, we assume that the QSC matrix is in the alternating ordering since the matrix in alternating ordering has significantly smaller bandwidth than the matrix in block ordering. When we apply the backslash operator, we assume that the QSC matrix is in the block ordering. There is no significant difference between the performance of the backslash operator applied to the matrix in the block ordering and that corresponding to the alternating ordering, except that for large grids, the block ordering seems to be slightly better. It is interesting to note that the LU factorization followed by F/B substitutions is slightly slower than the backslash operator, when both solution techniques are applied to the QSC matrix of the first step of the QSC method. However, the former solver is clearly superior to the latter when used in the two-step interior QSC method, since duplicate work in solving the system is avoided, by saving the LU factorization of the first step and applying only F/B substitutions in the second step. There was no noticeable discrepancy between the performance of the backslash operator and that of the LU factorization in the solution of the finite difference linear system.

In Figure 2.4 we plot the error at the grid points versus the number of flops required by the finite difference and the QSC methods. In this plot, for the QSC method, the error corresponds to the second step and the flops to the sum of the flops of both steps. The computation time is assumed to be proportional to the number of flops. Figure 2.4 clearly shows that the QSC method is asymptotically superior to the finite difference in the sense that it produces smaller approximation errors in the same computation time, while it achieves the same accuracy in less time.

grid size	error	order	error	order	error	order	error	order	error	order
	$\|\epsilon^{(1)}(x_i, y_j)\|$		$\|\|\epsilon^{(1)}\|\|_\infty$		$\|\|D_x\epsilon^{(1)}\|\|_\infty$		$\|D_{xy}\epsilon^{(1)}(s_i, t_j)\|$		$\|D_y^2\epsilon^{(1)}(\tau_i^x, \tau_j^y)\|$	
33×33	1.25-04		1.28-04		1.27-02		3.69-02		1.39-02	
65×65	3.11-05	2.0	3.01-05	2.1	3.23-03	2.0	1.36-02	1.4	4.22-03	1.7
129×129	7.77-06	2.0	7.62-06	2.0	8.12-04	2.0	4.58-03	1.6	1.20-03	1.8
	$\|\delta^{(1)}(x_i, y_j)\|$		$\|\|\delta^{(1)}\|\|_\infty$		$\|\|D_x\delta^{(1)}\|\|_\infty$		$\|D_{xy}\delta^{(1)}(s_i, t_j)\|$		$\|D_y^2\delta^{(1)}(\tau_i^x, \tau_j^y)\|$	
33×33	3.98-06		3.97-06		9.61-05		5.54-04		1.05-03	
65×65	9.93-07	2.0	9.84-07	2.0	2.41-05	2.0	1.59-04	1.8	3.16-04	1.7
129×129	2.48-07	2.0	2.48-07	2.0	6.04-06	2.0	4.37-05	1.9	8.86-05	1.8
	$\|\epsilon^{(2)}(x_i, y_j)\|$		$\|\|\epsilon^{(2)}\|\|_\infty$		$\|\|D_x\epsilon^{(2)}\|\|_\infty$		$\|D_{xy}\epsilon^{(2)}(s_i, t_j)\|$		$\|D_y^2\epsilon^{(2)}(\tau_i^x, \tau_j^y)\|$	
33×33	4.59-06		3.62-05		1.27-02		1.76-02		2.06-02	
65×65	2.97-07	3.9	3.79-06	3.3	3.23-03	2.0	4.78-03	1.9	5.53-03	1.9
129×129	1.89-08	4.0	6.80-07	2.5	8.12-04	2.0	1.24-03	1.9	1.43-03	2.0
	$\|\delta^{(2)}(x_i, y_j)\|$		$\|\|\delta^{(2)}\|\|_\infty$		$\|\|D_x\delta^{(2)}\|\|_\infty$		$\|D_{xy}\delta^{(2)}(s_i, t_j)\|$		$\|D_y^2\delta^{(2)}(\tau_i^x, \tau_j^y)\|$	
33×33	2.34-08		2.62-07		6.85-05		2.95-05		7.17-05	
65×65	1.46-09	4.0	3.06-08	3.1	1.71-05	2.0	7.01-06	2.1	1.71-05	2.1
129×129	9.15-11	4.0	3.75-09	3.0	4.28-06	2.0	1.70-06	2.0	4.20-06	2.0

Table 2.1: Observed errors and respective orders of convergence for both steps of the two-step QSC method applied to Problem 1. Here, $s_i \equiv x_i - \lambda h_x$ and $t_i \equiv y_i - \lambda h_y$. The constant λ is defined in Theorem 1.

grid size	error	order	error	order	error	order
	$\lvert\epsilon^{(1)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(1)}\rvert\rvert_\infty$	
33×33	7.871-02		7.741-02		7.820-02	
65×65	1.976-02	1.99	1.961-02	1.98	1.970-02	1.99
129×129	4.947-03	2.00	4.928-03	1.99	4.939-03	2.00
	$\lvert\delta^{(1)}(x_i, y_j)\rvert$		$\lvert\delta^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(1)}\rvert\rvert_\infty$	
33×33	6.581-02		6.472-02		6.538-02	
65×65	1.653-02	1.99	1.640-02	1.98	1.647-02	1.99
129×129	4.136-03	2.00	4.121-03	1.99	4.130-03	2.00
	$\lvert\epsilon^{(2)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(2)}\rvert\rvert_\infty$	
33×33	6.356-04		6.268-04		6.335-04	
65×65	4.034-05	3.98	3.998-05	3.97	4.036-05	3.97
129×129	2.555-06	3.98	2.538-06	3.98	2.569-06	3.97
	$\lvert\delta^{(2)}(x_i, y_j)\rvert$		$\lvert\delta^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(2)}\rvert\rvert_\infty$	
33×33	5.339-04		5.262-04		5.311-04	
65×65	3.384-05	3.98	3.354-05	3.97	3.375-05	3.98
129×129	2.140-06	3.98	2.127-06	3.98	2.138-06	3.98

Table 2.2: Observed errors and respective orders of convergence for both steps of the two-step QSC method applied to Problem 2.

grid size	error	order	error	order	error	order
	$\lvert\epsilon^{(1)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(1)}\rvert\rvert_\infty$	
17×17	1.893-04		1.879-04		1.991-04	
33×33	4.724-05	2.00	4.710-05	2.00	4.896-05	2.02
65×65	1.728-04	2.00	1.179-05	2.00	1.194-05	2.04
	$\lvert\delta^{(1)}(x_i, y_j)\rvert$		$\lvert\delta^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(1)}\rvert\rvert_\infty$	
17×17	2.440-05		2.470-05		2.459-05	
33×33	6.149-06	1.99	6.145-06	2.01	6.120-06	2.01
65×65	1.542-06	2.00	1.542-06	1.99	1.533-06	2.00
	$\lvert\epsilon^{(2)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(2)}\rvert\rvert_\infty$	
17×17	1.396-06		3.085-06		2.438-05	
33×33	8.627-08	4.02	1.947-07	3.99	2.787-06	3.13
65×65	5.372-09	4.01	1.221-08	4.00	3.589-07	2.96
	$\lvert\delta^{(2)}(x_i, y_j)\rvert$		$\lvert\delta^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(2)}\rvert\rvert_\infty$	
17×17	1.179-06		1.267-06		1.286-06	
33×33	7.467-08	3.98	8.216-08	3.95	1.113-07	3.53
65×65	4.672-09	4.00	5.155-09	3.99	1.009-08	3.46

Table 2.3: Observed errors and respective orders of convergence for both steps of the interior QSC method applied to Problem 3.

grid size	error	order	error	order	error	order
	$\lvert\epsilon^{(1)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(1)}\rvert\rvert_\infty$	
17×17	3.470-04		2.017+04		2.675+04	
33×33	8.737-05	1.99	1.107+03	4.19	1.468+03	4.19
65×65	2.182-05	2.00	6.269-01	10.79	8.312-01	10.79
	$\lvert\delta^{(1)}(x_i, y_j)\rvert$		$\lvert\delta^{(1)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(1)}\rvert\rvert_\infty$	
17×17	1.891-04		1.592+04		2.112+04	
33×33	4.867-05	1.96	8.742+02	4.19	1.159+03	4.19
65×65	1.223-05	1.99	4.949-01	10.79	6.562-01	10.79
	$\lvert\epsilon^{(2)}(x_i, y_j)\rvert$		$\lvert\epsilon^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\epsilon^{(2)}\rvert\rvert_\infty$	
17×17	4.325+06		1.110+16		1.472+16	
33×33	2.895+05	3.90	1.224+16	-0.14	1.623+16	-0.14
65×65	9.405+02	8.27	1.912+12	12.64	2.536+12	12.64
	$\lvert\delta^{(2)}(x_i, y_j)\rvert$		$\lvert\delta^{(2)}(\tau_i^x, \tau_j^y)\rvert$		$\lvert\lvert\delta^{(2)}\rvert\rvert_\infty$	
17×17	3.412+06		8.766+15		1.162+16	
33×33	2.285+05	3.90	9.662+15	-0.14	1.281+16	-0.14
65×65	7.424+02	8.27	1.510+12	12.64	2.002+12	12.64

Table 2.4: Observed errors and respective orders of convergence for both steps of the interior QSC method applied to Problem 4.

	error on grid points					
grid size	error	order	error	order	error	order
	$\lvert\epsilon^{(1)}\rvert$		$\lvert\epsilon^{(2)}\rvert$		$\lvert\epsilon^{(FD)}\rvert$	
17×17	2.038-04		2.469-06		3.514-04	
33×33	5.110-05	1.99	1.567-07	3.97	8.786-05	2.00
65×65	1.277-05	2.00	9.891-09	3.99	2.194-05	2.00
	$\lvert\delta^{(1)}\rvert$		$\lvert\delta^{(2)}\rvert$		$\lvert\delta^{(FD)}\rvert$	
17×17	1.039-04		8.961-07		3.413-04	
33×33	2.584-05	2.01	7.351-08	3.61	8.586-04	1.99
65×65	6.475-06	2.00	7.351-09	3.32	2.154-05	2.00

Table 2.5: Observed errors on grid points and respective orders of convergence for both steps of the interior QSC method and for standard second order finite differences applied to Problem 5.

	number of floating-point operations					
grid size	\	LU, F/B	\	F/B	\	LU, F/B
	QSC step 1		QSC step 2		FD	
17×17	1.98+6	1.24+6	1.98+6	7.04+4	1.33+6	9.42+5
33×33	2.15+7	1.83+7	2.15+7	5.44+5	2.65+7	1.59+7
65×65	2.21+8	2.80+8	2.21+8	4.27+6	3.04+8	2.62+8

Table 2.6: Number of floating-point operations in MATLAB, required by the interior QSC and standard second order finite difference methods applied to Problem 5. Two solution techniques we used: the backslash operator \ with the block ordering, and the LU factorization followed by forward and backward (F/B) substitutions with the alternating ordering.

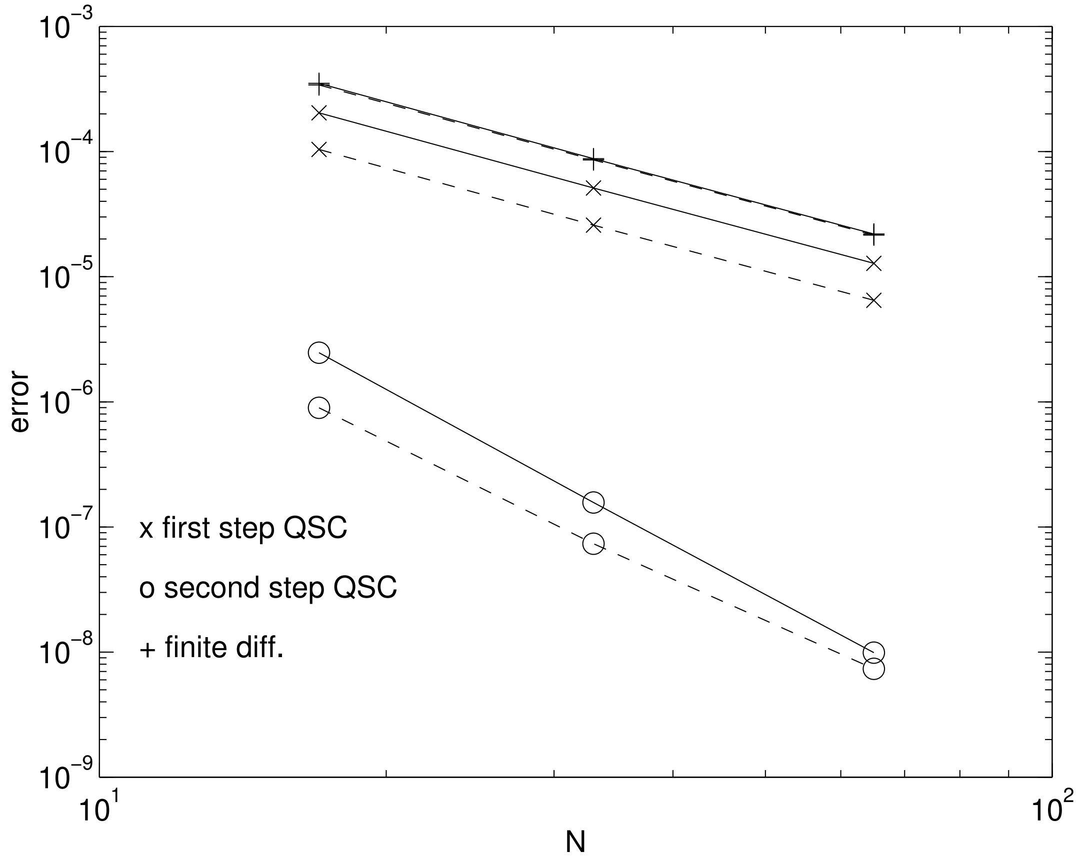

Figure 2.3: Log-log scale plot of observed errors for both steps of the interior QSC method and for standard second order finite differences applied to Problem 5, versus the number N of subintervals in one dimension. The solid lines correspond to the errors for u, while the dashed lines correspond to the errors for v.

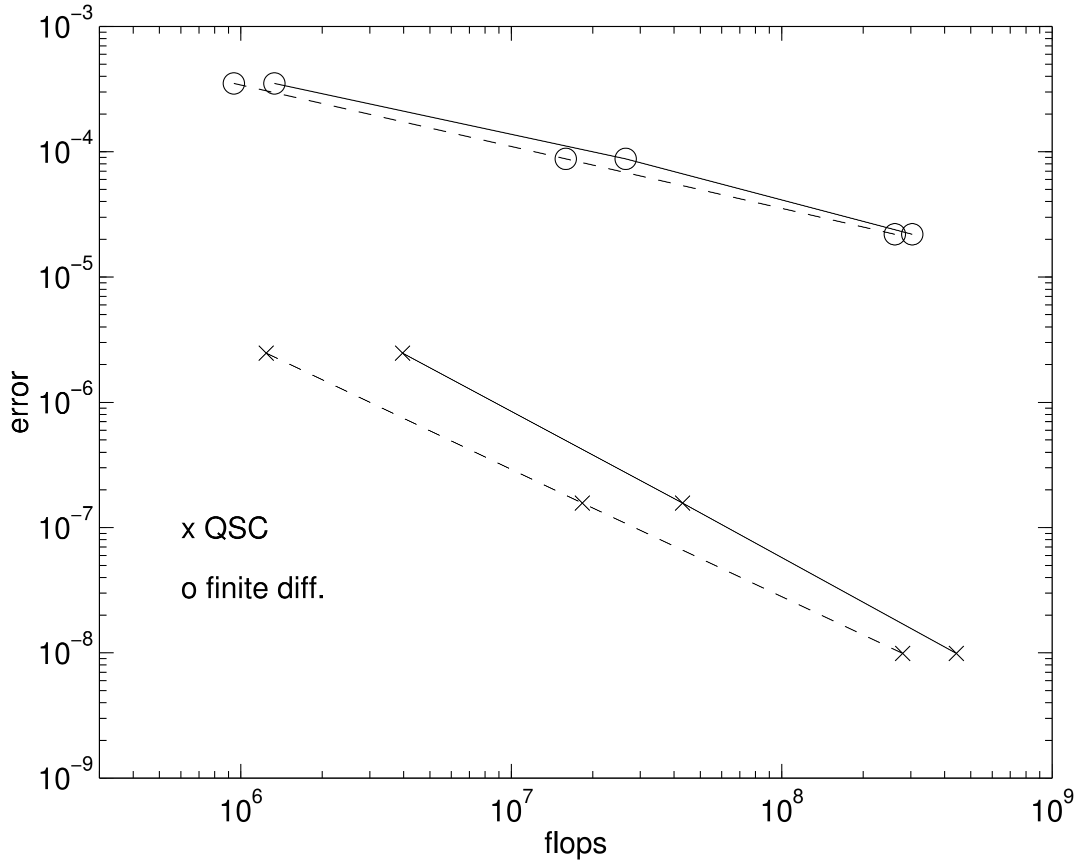

Figure 2.4: Log-log scale plot of observed errors for u of the interior QSC and standard second order finite difference methods applied to Problem 5, versus the number of floating-point operations in MATLAB. The solid lines correspond to the flops using the backslash operator $\backslash$ and the block-ordering, while the dashed lines correspond to the flops using the LU factorization routine and the alternating ordering.

Chapter 3

FFT and Multigrid solvers for systems of two PDEs

In this chapter, we extend the FFT and multigrid solvers of QSC equations for a single PDE to FFT and multigrid solvers for systems of two PDEs (1.1)-(1.2). Let $K_{11}, K_{12}, K_{21}, K_{22}$ be the $MN \times MN$ matrices of equations arising from the application of the interior QSC method to the PDE problems (2.35), (2.36), (2.37), (2.38), respectively. Let

$$H = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

3.1 Eigenvalues and eigenvectors of the QSC matrix for systems of two Helmholtz PDEs

To find the eigenvalues of the matrix H, consider a vector of the form $\left[\overline{\delta^{lm}}, \alpha^{lm}\overline{\delta^{lm}}\right]^T$, for $l = 1, \ldots, M, m = 1, \ldots, N$, where $\overline{\delta^{lm}}$ is an eigenvector of K_{ij}, $i, j = 1, 2$, as defined in (1.18), and α^{lm} is a constant to be determined later. Then by construction

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha^{lm}\overline{\delta^{lm}} \end{bmatrix} = \begin{bmatrix} \rho_{11}^{lm}\overline{\delta^{lm}} + \rho_{12}^{lm}\alpha^{lm}\overline{\delta^{lm}} \\ \rho_{21}^{lm}\overline{\delta^{lm}} + \rho_{22}^{lm}\alpha^{lm}\overline{\delta^{lm}} \end{bmatrix}, \tag{3.1}$$

where $\rho_{i1}^{lm}, \rho_{i2}^{lm}$ are the eigenvalues of K_{i1}, K_{i2}, respectively, given by (1.17).

LEMMA 2 *If* $\rho_{12}^{lm} \neq 0$ *then* $\lambda_{\pm}^{lm} = \frac{1}{2}\left(\rho_{11}^{lm} + \rho_{22}^{lm} \pm \sqrt{(\rho_{11}^{lm} + \rho_{22}^{lm})^2 - 4(\rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm})}\right)$, *and* $\overline{\psi_{\pm}^{lm}} = \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha_{\pm}^{lm}\overline{\delta^{lm}} \end{bmatrix}$ *where* $\alpha_{\pm}^{lm} = \left(\frac{\lambda_{\pm}^{lm} - \rho_{11}^{lm}}{\rho_{12}^{lm}}\right)$, $l = 1, \ldots, M, m = 1, \ldots, N$, *are eigenvalues and the corresponding eigenvectors of* H, *respectively.*

PROOF: If λ^{lm} and $\left[\overline{\delta^{lm}}, \alpha^{lm}\overline{\delta^{lm}}\right]^T$ are an eigenvalue and the corresponding eigenvector of H, respectively, the following holds:

$$\begin{bmatrix} \rho_{11}^{lm}\overline{\delta^{lm}} + \rho_{12}^{lm}\alpha^{lm}\overline{\delta^{lm}} \\ \rho_{21}^{lm}\overline{\delta^{lm}} + \rho_{22}^{lm}\alpha^{lm}\overline{\delta^{lm}} \end{bmatrix} = \lambda^{lm} \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha^{lm}\overline{\delta^{lm}} \end{bmatrix}. \tag{3.2}$$

From this and the fact that $\overline{\delta^{lm}} \neq \overline{0}$, we have

$$\rho_{11}^{lm} + \rho_{12}^{lm}\alpha^{lm} = \lambda^{lm} \tag{3.3}$$

$$\rho_{21}^{lm} + \rho_{22}^{lm}\alpha^{lm} = \alpha^{lm}\lambda^{lm}. \tag{3.4}$$

Since $\rho_{12}^{lm} \neq 0$, we can rewrite (3.3) as $\alpha^{lm} = \frac{\lambda^{lm} - \rho_{11}^{lm}}{\rho_{12}^{lm}}$ and substitute it into (3.4) to obtain $(\lambda^{lm})^2 - (\rho_{11}^{lm} + \rho_{22}^{lm})\lambda^{lm} + \rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm} = 0$. By solving this equation we get

$$\lambda_{\pm}^{lm} = \frac{1}{2}\left(\rho_{11}^{lm} + \rho_{22}^{lm} \pm \sqrt{(\rho_{11}^{lm} + \rho_{22}^{lm})^2 - 4(\rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm})}\right). \tag{3.5}$$

By substituting $\lambda_{\pm}^{lm}$ back into (3.3) we have

$$\alpha_{\pm}^{lm} = \frac{\lambda_{\pm}^{lm} - \rho_{11}^{lm}}{\rho_{12}^{lm}}. \tag{3.6}$$

$\diamondsuit$

For the remainder of this chapter we use the following definitions

$$\Upsilon_{+} = \left\{\lambda_{+}^{lm}\right\}_{l=1,m=1}^{M,N}, \Upsilon_{-} = \left\{\lambda_{-}^{lm}\right\}_{l=1,m=1}^{M,N},$$

and

$$\Psi_{+} = \left\{\overline{\psi_{+}^{lm}}\right\}_{l=1,m=1}^{M,N}, \Psi_{-} = \left\{\overline{\psi_{-}^{lm}}\right\}_{l=1,m=1}^{M,N}.$$

To normalize the eigenvectors in the set Ψ_+ with respect to the Euclidean norm, for each $\overline{\psi_+^{lm}} = \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha_+^{lm}\overline{\delta^{lm}} \end{bmatrix}$ in Ψ_+, where $\overline{\delta^{lm}} = \overline{\delta_l^x} \otimes \overline{\delta_m^y}$, $l = 1, \ldots, M, m = 1, \ldots, N$, first set

$$\kappa_l^x = \sqrt{\frac{2}{M}}, \text{ for } l = 1, \ldots, M-1, \text{ and } \kappa_M^x = \sqrt{\frac{1}{M}}, \tag{3.7}$$

for $\overline{\delta_l^x}$ in (1.19). Then set

$$\kappa_m^y = \sqrt{\frac{2}{N(1+(\alpha_+^{lm})^2)}}, \text{ for } m = 1, \ldots, N-1, \text{ and } \kappa_N^y = \sqrt{\frac{1}{N(1+(\alpha_+^{lN})^2)}}, \tag{3.8}$$

for $\overline{\delta_m^y}$ in (1.19). To normalize the eigenvectors in the set Ψ_- with respect to the Euclidean norm, for each $\overline{\psi_-^{lm}} = \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha_-^{lm}\overline{\delta^{lm}} \end{bmatrix}$ in Ψ_-, where $\overline{\delta^{lm}} = \overline{\delta_l^x} \otimes \overline{\delta_m^y}$, $l = 1, \ldots, M, m = 1, \ldots, N$, first set

$$\kappa_l^x = \sqrt{\frac{2}{M}}, \text{ for } l = 1, \ldots, M-1, \text{ and } \kappa_M^x = \sqrt{\frac{1}{M}}, \tag{3.9}$$

for $\overline{\delta_l^x}$ in (1.19). Then set

$$\kappa_m^y = \sqrt{\frac{2}{N(1+(\alpha_-^{lm})^2)}}, \text{ for } m = 1, \ldots, N-1, \text{ and } \kappa_N^y = \sqrt{\frac{1}{N(1+(\alpha_-^{lN})^2)}}, \tag{3.10}$$

for $\overline{\delta_m^y}$ in (1.19). From now on, we will assume that Ψ_+ and Ψ_- contain these normalized eigenvectors.

THEOREM 3 *If $(\rho_{11}^{lm}+\rho_{22}^{lm})^2-4(\rho_{11}^{lm}\rho_{22}^{lm}-\rho_{12}^{lm}\rho_{21}^{lm}) \neq 0$ and $\rho_{12}^{lm} \neq 0$, for $l = 1, \ldots, M, m = 1, \ldots, N$, then the sets Ψ_+ and Ψ_- contain the $2MN$ linearly independent normalized eigenvectors of H. Furthermore, the sets Υ_- and Υ_+ span all the eigenvalues of H.*

PROOF: The condition $\rho_{12}^{lm} \neq 0$, for $l = 1, \ldots, M, m = 1, \ldots, N$, and Lemma 2 ensure that Ψ_+ and Ψ_- contain eigenvectors of H. Without loss of generality, suppose that for some p, q

$$\begin{bmatrix} \overline{\delta^{pq}} \\ \alpha_+^{pq}\overline{\delta^{pq}} \end{bmatrix} = \sum_{l=1,l\neq p}^{M} \sum_{m=1,m\neq q}^{N} a_{lm} \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha_+^{lm}\overline{\delta^{lm}} \end{bmatrix} + \sum_{l=1}^{M}\sum_{m=1}^{N} b_{lm} \begin{bmatrix} \overline{\delta^{lm}} \\ \alpha_-^{lm}\overline{\delta^{lm}} \end{bmatrix}, \tag{3.11}$$

where a_{lm} and b_{lm}, $l = 1, \ldots, M, m = 1, \ldots, N$, are constants that are not all zero. Note that all vectors in (3.11) consist of two blocks, thus (3.11) can be split into two block relations. From the first block relation, we can derive

$$(1 - b_{pq})\overline{\delta^{pq}} - \sum_{l=1, l\neq p}^{M} \sum_{m=1, m\neq q}^{N} (a_{lm} + b_{lm})\overline{\delta^{lm}} = 0.$$

But as the set $\{\overline{\delta^{lm}}\}_{l=1,m=1}^{M,N}$ contains NM linearly independent vectors, we must have $1 - b_{pq} = 0$ and $a_{lm} + b_{lm} = 0$, $\forall (l, m) \neq (p, q)$. Substituting into the second block relation arising from (3.11) we get

$$\alpha_{+}^{pq}\overline{\delta^{pq}} = \alpha_{-}^{pq}\overline{\delta^{pq}} + \sum_{l=1, l\neq p}^{M} \sum_{m=1, m\neq q}^{N} a_{lm}(\alpha_{+}^{lm} - \alpha_{-}^{lm})\overline{\delta^{lm}}$$

The linear independence of the vectors in the set $\{\overline{\delta^{lm}}\}_{l=1,m=1}^{M,N}$ leads to the relation $\alpha_{+}^{pq} = \alpha_{-}^{pq}$. This implies that $(\rho_{11}^{pq} + \rho_{22}^{pq})^2 - 4(\rho_{11}^{pq}\rho_{22}^{pq} - \rho_{12}^{pq}\rho_{21}^{pq}) = 0$, which contradicts an assumption of the theorem. Thus Ψ_{+} and Ψ_{-} contain $2MN$ linearly independent eigenvectors of H.

Now, every eigenvector of H must be in an eigenspace spanned by Ψ_{+} and Ψ_{-}. Since every eigenvector in Ψ_{+} and Ψ_{-} corresponds to an eigenvalue in Υ_{+} or Υ_{-}, the union of the sets Υ_{+} and Υ_{-} spans all the eigenvalues of H. $\diamondsuit$

COROLLARY 1 *Let A_1, B_1, C_1, D_1, E_1, and F_1 be defined as in (2.39). Let $P_1(x, y)$ be defined as in (2.40). If $\rho_{12}^{lm} \neq 0$, for $l = 1, \ldots, M, m = 1, \ldots, N$, $B_1^2 - 4A_1C_1 < 0$, $A_1 < 0$ and $P_1(S, T) < 0$, where S, T are defined in (2.41), then Ψ_{+} and Ψ_{-} contain $2MN$ the linearly independent normalized eigenvectors of H. Furthermore, the sets Υ_{-} and Υ_{+} contain all the eigenvalues of H.*

PROOF: Under the assumptions, we have $(\rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm}) < 0$, for $l = 1, \ldots, M, m = 1, \ldots, N$. So $(\rho_{11}^{lm} + \rho_{22}^{lm})^2 - 4(\rho_{11}^{lm}\rho_{22}^{lm} - \rho_{12}^{lm}\rho_{21}^{lm}) > 0$. The result now follows from Theorem 3. $\diamondsuit$

3.2 FFT solvers for systems of two Helmholtz PDEs

Let V_M, V_N and V be defined as in Section 1.6. It is clear that the columns in V are the MN linearly independent eigenvectors of any of the matrices K_{ij}, $i, j = 1, 2$. If the conditions of Theorem 3 are satisfied, then the columns of the matrix

$$W = \begin{bmatrix} V & V \\ VD_+ & VD_- \end{bmatrix}, \tag{3.12}$$

where $D_+ = \mathrm{diag}\{\alpha_+^{lm}\}_{l=1,m=1}^{M,N}$, and $D_- = \mathrm{diag}\{\alpha_-^{lm}\}_{l=1,m=1}^{M,N}$, are the $2MN$ linearly independent eigenvectors of H from the sets Ψ_+ and Ψ_-.

To find W^{-1} set

$$\begin{bmatrix} V & V \\ VD_+ & VD_- \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

where I is the $MN \times MN$ identity matrix. Simplifying, we get

$$W^{-1} = \begin{bmatrix} (I + D_\pm^{-1} D_+)V^{-1} & -D_\pm^{-1} V^{-1} \\ -D_\pm^{-1} D_+ V^{-1} & D_\pm^{-1} V^{-1} \end{bmatrix},$$

where $D_\pm = D_- - D_+$.

Let Σ be a diagonal matrix such that its kth entry is an eigenvalue of H that corresponds to the eigenvector in the kth column of W, $k = 1, \ldots, 2MN$. The diagonalization of H is given by $H = W\Sigma W^{-1}$, and if H is solvable, $H^{-1} = W\Sigma^{-1}W^{-1}$. Hence, if $Hx = \overline{g}$, where $\overline{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$, g_1 and g_2 contain the first MN and last MN entries of $\overline{g}$, respectively, then

$$\begin{aligned} x &= W\Sigma^{-1}W^{-1}\overline{g} \\ &= \begin{bmatrix} V & V \\ VD_+ & VD_- \end{bmatrix} \Sigma^{-1} \begin{bmatrix} (I + D_\pm^{-1} D_+)V^{-1} & -D_\pm^{-1} V^{-1} \\ -D_\pm^{-1} D_+ V^{-1} & D_\pm^{-1} V^{-1} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \\ &= \begin{bmatrix} (S_M^{-1} \otimes S_N^{-1}) & (S_M^{-1} \otimes S_N^{-1}) \\ (S_M^{-1} \otimes S_N^{-1})D_+ & (S_M^{-1} \otimes S_N^{-1})D_- \end{bmatrix} \Sigma^{-1} \end{aligned}$$

$$\begin{bmatrix} (I + D_{\pm}^{-1}D_{+})(S_M \otimes S_N) & -D_{\pm}^{-1}(S_M \otimes S_N) \\ -D_{\pm}^{-1}D_{+}(S_M \otimes S_N) & D_{\pm}^{-1}(S_M \otimes S_N) \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \tag{3.13}$$

where S_M and S_N are the discrete sine transform II matrices of size M and N, respectively, as defined in Section 1.6.

We now provide an algorithm for performing the computation involved in (3.13):

Algorithm: FFT solution of a system of two Helmholtz PDEs with constant coefficients

Step 1: Perform FST2(M, N) to g_1 to obtain $g_1^{(1)} = (S_M \otimes S_N)g_1$.

Step 2: Perform FST2(M, N) to g_2 to obtain $g_2^{(1)} = (S_M \otimes S_N)g_2$.

Step 3: Multiply each component of $g_1^{(1)}$ by the corresponding diagonal entries of the matrix $(I + D_{\pm}^{-1}D_{+})$ to obtain $g_1^{(2)} = (I + D_{\pm}^{-1}D_{+})g_1^{(1)}$.

Step 4: Divide each component of $g_2^{(1)}$ by the corresponding diagonal entries of the matrix $-D_{\pm}$ to obtain $g_2^{(2)} = -D_{\pm}^{-1}g_2^{(1)}$.

Step 5: Add the components of $g_1^{(2)}$ and $g_2^{(2)}$ to obtain $g_1^{(3)} = (I + D_{\pm}^{-1}D_{+})(S_M \otimes S_N)g_1 - D_{\pm}^{-1}(S_M \otimes S_N)g_2$.

Step 6: Multiply each component of $g_1^{(1)}$ by the corresponding diagonal entries of the matrix $-D_{\pm}^{-1}D_{+}$ to obtain $g_1^{(4)} = -D_{\pm}^{-1}D_{+}g_1^{(1)}$.

Step 7: Divide each component of $g_2^{(1)}$ by the corresponding diagonal entries of the matrix $D_{\pm}$ to obtain $g_2^{(4)} = D_{\pm}^{-1}g_2^{(1)}$.

Step 8: Add the components of $g_1^{(4)}$ and $g_2^{(4)}$ to obtain $g_2^{(3)} = -D_{\pm}^{-1}D_{+}(S_M \otimes S_N)g_1 + D_{\pm}^{-1}(S_M \otimes S_N)g_2$.

Step 9: Divide each component of $g_1^{(3)}$ by the corresponding first MN diagonal entries of the matrix Σ to obtain $g_1^{(5)}$.

Step 10: Divide each component of $g_2^{(3)}$ by the corresponding last MN diagonal entries of the matrix Σ to obtain $g_2^{(5)}$. Hence we have $\begin{bmatrix} g_1^{(5)} \\ g_2^{(5)} \end{bmatrix} = \Sigma^{-1} \begin{bmatrix} g_1^{(3)} \\ g_2^{(3)} \end{bmatrix}$.

Step 11: Perform iFST2(M, N) to $g_1^{(5)}$ to obtain $g_1^{(6)} = (S_M^{-1} \otimes S_N^{-1}) g_1^{(5)}$.

Step 12: Perform iFST2(M, N) to $g_2^{(5)}$ to obtain $g_2^{(6)} = (S_M^{-1} \otimes S_N^{-1}) g_2^{(5)}$.

Step 13: Add the components of $g_1^{(6)}$ and $g_2^{(6)}$ to obtain $x_1 = (S_M^{-1} \otimes S_N^{-1}) g_1^{(5)} + (S_M^{-1} \otimes S_N^{-1}) g_2^{(5)}$.

Step 14: Multiply each component of $g_1^{(5)}$ by the corresponding diagonal entries of the matrix D_+ to obtain $g_1^{(7)} = D_+ g_1^{(5)}$.

Step 15: Multiply each component of $g_2^{(5)}$ by the corresponding diagonal entries of the matrix D_- to obtain $g_2^{(7)} = D_- g_2^{(5)}$.

Step 16: Perform iFST2(M, N) to $g_1^{(7)}$ to obtain $g_1^{(8)} = (S_M^{-1} \otimes S_N^{-1}) g_1^{(7)}$.

Step 17: Perform iFST2(M, N) to $g_2^{(7)}$ to obtain $g_2^{(8)} = (S_M^{-1} \otimes S_N^{-1}) g_2^{(7)}$.

Step 18: Add the components of $g_1^{(8)}$ and $g_2^{(8)}$ to obtain $x_2 = (S_M^{-1} \otimes S_N^{-1}) D_+ g_1^{(5)} + (S_M^{-1} \otimes S_N^{-1}) D_- g_2^{(5)}$. Hence $x \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = W \Sigma^{-1} W^{-1} \overline{g}$.

There are 2 performances of FST2(M, N) and 4 performances of iFST2(M, N). Since these form the dominant part of the computation of (3.13), the total asymptotic computation cost is $15MN \log_2(MN)$ flops, or about 3 times more costly than to use FFTs to solve the single Helmholtz problem (1.16)-(1.13).

3.2.1 Preconditioners for QSC equations

For solving the equations arising from the application of the QSC method to systems of general PDEs with homogeneous boundary conditions, we consider preconditioned iterative methods.

Consider the system of Helmholtz problems

$$\begin{aligned} 10u_{xx} + 10u_{yy} - 10u \quad & +v_{xx} + v_{yy} - v &= g_1 \\ u_{xx} + u_{yy} - u \quad & +10v_{xx} + 10v_{yy} - 10v &= g_2 \end{aligned} \quad \text{in } \Omega \equiv (ax, bx) \times (ay, by) \qquad (3.14)$$

subject to homogeneous Dirichlet boundary conditions for both unknowns u and v

$$\begin{aligned} u &= 0 \\ v &= 0 \end{aligned} \quad \text{on } \partial\Omega \equiv \text{ boundary of } \Omega \qquad (3.15)$$

Let H be the $2MN \times 2MN$ matrix from the application of the interior QSC method to the system of PDEs (3.14)-(3.15).

Let K be the $2MN \times 2MN$ matrix from the application of the interior QSC method to a general system of two PDEs (1.1) subject to homogeneous Dirichlet boundary conditions for both unknowns u and v (3.15).

To effectively solve K, we consider the following preconditioners:

1. $P = H$,
2. $P = DH$,
3. $P = HD$,
4. $P = \sqrt{D} H \sqrt{D}$,

where D is the diagonal of K.

It is worth noting that the choice of coefficients appearing in the Helmholtz PDEs (3.14) seems a bit arbitrary at this point. However, one needs to note that the system (3.14) safely satisfies the assumptions of Theorem 2, and that choosing all coefficients of second order derivatives to be equal to 1 and all coefficients of unknown functions equal to -1 (i.e. a direct extension of the technique in [5]) does not lead to a uniquely solvable system.

Another note arises with respect to the ordering used for K and H. We chose to adopt the alternating ordering in all experiments presented here, mainly for programming convenience. It does not seem that the choice of ordering, alternating or block, significantly affects the convergence of iterative methods, however, we plan to study the effects of ordering further in future work.

3.3 Multigrid Methods for QSC equations

For solving the equations arising from the application of the QSC method to systems of PDEs, we consider iterative methods, where the multigrid technique is incorporated as a preconditioner.

Let again K be the $2MN \times 2MN$ matrix arising from the application of the interior QSC method to a general system of two PDEs (1.1), subject to homogeneous Dirichlet boundary conditions (3.15) for both unknowns u and v. In order to give the form of the extension and restriction operators for multigrid preconditioning of the QSC equations arising from the above system of two PDEs, let $E_2 = \begin{bmatrix} E & Z_E \\ Z_E & E \end{bmatrix}$ and $R_2 = \begin{bmatrix} R & Z_R \\ Z_R & R \end{bmatrix}$, where E and R are the extension and restriction operators, respectively, designed for a single PDE with homogeneous Dirichlet conditions, as mentioned in Section 1.7, and described in detail in [4], and Z_E and Z_R are matrices of zeroes of the same sizes as E and R, respectively. For convenience, assume also that K is in the block ordering. Then E_2 and R_2 act as extension and restriction operators, respectively, and P, with $P^{-1} = E_2 K'^{-1} R_2$, where K' represents the coarse grid interior QSC matrix for the system of two PDEs, acts as preconditioner for K. We can use this preconditioner in both additive and multiplicative schemes.

We note that, in the above discussion, the block ordering was chosen for convenience, since it allows the presentation of the extension and restriction operators, E_2 and R_2, in a condensed form. However, in the numerical experiments, we used the alternating ordering, for programming convenience and a fair comparison with the FFT preconditioners described in the previous section.

We also need to note that, in all our numerical experiments, the multigrid preconditioners are applied to the system of interior QSC equations. This allows a fair comparison between the FFT and the multigrid preconditioners. However, we note that the application of the multigrid preconditioner is not limited to homogeneous Dirichlet boundary conditions. With appropriate adjustment of the extension and restriction operators, the multigrid preconditioner is applicable to PDEs subject to the most general boundary conditions.

3.4 Numerical results

In this section we present numerical results to compare the effectiveness among the following linear solvers: banded Gauss elimination without pivoting, the FFT solvers and preconditioners described in Section 3.2, and the multigrid preconditioners described in the previous section. All computations in this section were carried out on a Sun Ultra-4 in double precision. The QSC solvers were programmed by us in both FORTRAN and MATLAB, with the exception of the full multigrid method which was programmed only in MATLAB. Most of the results in this chapter are obtained from the FORTRAN implementation.

In the FORTRAN implementation, banded Gauss elimination without pivoting was implemented by using the ELLPACK routines *q5bnfa,q5bnsl*; for the FFT solvers we used the FORTRAN FFT package in [12] and our own FORTRAN code; for the multigrid methods we used our own FORTRAN codes; both the FFT and multigrid preconditioners were coupled with the GMRES acceleration method from the KSP package [9].

The maximum number of previous iteration vectors that GMRES takes into account in one iteration, i.e., the "restart", was set to 20, for all tests performed. The stopping criterion for the iterative solvers was the relative Euclidean norm residual and the tolerance was set to 10^{-8} for the first step and 10^{-6} for the second step of the interior QSC method. The initial guess solution vector in the first step was a vector of zeroes, while in the second step the solution vector of the first step. The FFT preconditioned methods used the DH preconditioner. For the two-grid method, the step size ratio for the conversion between the fine and coarse grids was 4 and the multiplicative scheme was used to couple the coarse grid preconditioner with 6 Jacobi iterations. The coarse system was solved by banded Gauss elimination. All times are user CPU seconds and were obtained using the FORTRAN routine *etime*.

In addition, we present numerical results from the MATLAB implementation of the full multigrid and the two-grid methods. The purpose is to compare the computational complexities of the full multigrid and two-grid methods for solving the linear system arising from the first step of the QSC method. For the MATLAB results from the multigrid methods, the step size ratio for the conversion between the fine and coarse grids was 2 and the multiplicative scheme was used to couple the coarse grid preconditioner with 3 Jacobi iterations. No acceleration method was applied. The restriction operator was based on cubic interpolation as described in [4]. The stopping criterion for the iterative solvers was the relative Euclidean norm residual and the tolerance was set to 10^{-8}. The problems are given in the Appendix.

Problem 6 describes a system of PDEs with variable coefficients, that do not exhibit a large variation. Table 3.1 shows results from the application of various solvers to the system arising from the interior QSC method applied to Problem 6. The first two columns under "Gauss elimination" present the orders of convergence of the errors $\epsilon^{(2)}$ and $\delta^{(2)}$ defined in Theorem 1 and evaluated on the grid points based on the Gauss elimination solution. "It1" and "It2" correspond to the number of iterations the solver required to reach the tolerance in the first and second steps of the interior QSC method, respectively. All times are user CPU seconds. "Total" is the total time, "per It" is the time per iteration, "LU" is the time used for LU factorization, "BF" is the time used for backward-forward substitution, "Update" is the time used for updating the right-hand side in the second step, and "Coarse" is the time that the two-grid solver used to setup and LU factor the coarse system. The orders of convergence of the errors are clearly optimal.

Problem 7 describes a system of PDEs such that some of the coefficients have a large range of values. Table 3.2 shows results from the application of FFT and two-grid solvers to the system arising from the interior QSC method applied to Problem 7. The first two columns under "FFT" present the orders of convergence of the errors $\epsilon^{(2)}$ and $\delta^{(2)}$ defined in Theorem 1 and evaluated on the grid points based on the FFT solution. The rest of the data follow the notation of Table 3.1. The orders of convergence of the errors are optimal.

From both tables, we note that the number of iterations for both iterative solvers to reach the tolerance for the first and second steps of the QSC method is independent of the grid size $n \times n$. For the Gauss elimination solver, the LU factorization is the most time consuming work done and the total time to solve the linear system is proportional to n^4, as expected. The total time for the FFT solver is proportional to the expected complexity of $n^2 \log n$, which is asymptotically close to the optimal $O(n^2)$ behaviour. The two-grid solver uses a significant amount of time in setting up and LU factorizing the coarse system, this is why the total time deviates from the optimal $O(n^2)$. The processing of the coarse system by a direct solver seems to be the bottleneck for the two-grid solver. From both tables, though, it is clear that both iterative solvers are substantially faster than the direct solver, while the FFT is the fastest solver overall.

Table 3.3 shows results from the application of the two-grid and full multigrid solvers to the system arising from the interior QSC method applied to Problem 7. The number of iterations for the multigrid solvers to reach the tolerance is once again independent of the grid size, and almost the same between the two solvers. More importantly, the results show that the computational complexity of the full multigrid method is asymptotically close to the optimal $O(n^2)$ flops. The computational complexity of the full multigrid method is noticeably better than that of the two-grid method.

	Gauss Elimination						FFT				Two-grid				
	conv. order		time				no. iter.		time		no. iter.		time		
grid size	$\lvert\epsilon^{(2)}\rvert$	$\lvert\delta^{(2)}\rvert$	LU	BF	Update	Total	It1	It2	per It	Total	It1	It2	per It	Coarse	Total
17×17			0.07	0.01	0.06	0.15	11	8	0.002	0.10	10	8	0.003	0.01	0.13
33×33	4.04	3.94	1.16	0.04	0.17	1.42	11	8	0.007	0.32	11	9	0.013	0.02	0.44
65×65	4.01	4.00	16.91	0.32	0.65	18.31	11	8	0.035	1.34	11	9	0.065	0.14	2.10
129×129	4.00	4.00	261.47	2.43	2.58	269.72	11	8	0.170	5.84	11	9	0.427	1.40	12.49

Table 3.1: Comparison among direct and iterative solvers for the system arising from the interior QSC method applied to Problem 6.

	FFT						Two-grid				
	conv. order		no. iter.		time		no. iter.		time		
grid size	$\|\epsilon^{(2)}\|$	$\|\delta^{(2)}\|$	It1	It2	per It	Total	It1	It2	per It	Coarse	Total
33×33			11	9	0.008	0.33	11	9	0.013	0.02	0.45
65×65	4.05	3.85	11	9	0.038	1.42	11	9	0.073	0.14	2.29
129×129	3.89	3.92	11	9	0.173	6.03	12	9	0.425	1.39	12.75
257×257	3.93	3.97	11	9	0.758	25.57	12	10	1.986	17.88	71.65

Table 3.2: Comparison between iterative solvers for the system arising from the interior QSC method applied to Problem 7.

	Full MGM		Two-grid method	
grid size	It1	flops	It1	flops
17×17	9	1.54+6	10	1.29+6
33×33	10	8.18+6	11	6.76+6
65×65	10	3.27+7	11	4.29+7
129×129	10	1.32+8	11	3.98+8

Table 3.3: Comparison between the full MGM and the two-grid method for the system arising from the interior QSC method applied to Problem 7. The notation $x.y+z$ means $x.y \times 10^{+z}$.

Chapter 4

Conclusions

This thesis develops a QSC method for solving systems of two linear second-order PDEs in two dimensions. This method, based on an optimal two-step QSC method for a single elliptic PDE described in [2], is optimal in the sense that it is fourth order locally at certain points and third order globally. The analysis shows that the optimal order of convergence holds under certain conditions. Numerical results confirm that the QSC method converges optimally even for some problems that do not satisfy the conditions of the analysis.

We develop analytic formulae for the eigenvalues and eigenvectors of the matrix arising from the application of the QSC method to a system of two PDEs with Helmholtz operators and constant coefficients. Based on these formulae, we develop fast direct solvers for the QSC system of equations using FFTs.

In addition, we develop fast iterative methods for solving the linear systems arising from the application of the QSC method to more general problems. These methods are natural extensions of FFT and multigrid solvers for systems arising from the application of the QSC method to a single PDE and can be used as preconditioners to solve systems of PDEs with general coefficients. Numerical results show that the number of iterations that the resulting iterative solvers need to reach the tolerance for both steps of the QSC method is not sensitive to the grid size. Moreover, the performance of the FFT solver is almost optimal since the computational time is $O(n^2 \log n)$, where n is the grid size in one dimension. However, more research needs to be done in selecting the most suitable FFT preconditioner for solving general systems of PDEs. Although we developed FFT solvers for homogeneous Dirichlet boundary conditions only, we can easily extend FFT solvers for a variety of other boundary conditions. Also, we developed the FFT solvers so that the FFT is applied to both the x and y directions. We can further reduce the time by a factor of two, by applying the FFT in one direction and tridiagonal solves in the other, as long as the boundary conditions are not periodic.

The optimal performance of the two-grid solver was hindered by the time spent on the processing of the coarse system by a direct method. The full multigrid method does not suffer from this problem and exhibits asymptotically optimal performance, with the computational time being $O(n^2)$. In any case, the results show that all iterative solvers are superior to the direct solver.

Appendix A

Problems

Problem 1. Determine $u = u(x,y)$ and $v = v(x,y)$ such that

$$\begin{aligned} a_{11}u_{xx} + c_{11}u_{yy} + f_{11}u + a_{12}v_{xx} + c_{12}v_{yy} + f_{12}v = g_1 \\ a_{21}u_{xx} + c_{21}u_{yy} + f_{21}u + a_{22}v_{xx} + c_{22}v_{yy} + f_{22}v = g_2 \end{aligned} \quad \text{in } \Omega \equiv (0,1)\times(0,1),$$

subject to

$$\begin{aligned} 4u + v = \gamma_1 \\ 2u + 3v = \gamma_2 \end{aligned} \quad \text{on } \partial\Omega,$$

where

$$\begin{cases} a_{11}(x,y) = xy & a_{12}(x,y) = e^{x+y} + 5 \\ c_{11}(x,y) = x - y & c_{12}(x,y) = 4xy + 9 \\ f_{11}(x,y) = \sqrt{2} & f_{12}(x,y) = 1/(1+xy) \\ a_{21}(x,y) = e^{2x+y} + 7 & a_{22}(x,y) = x + y \\ c_{21}(x,y) = (x+3)(y+5) & c_{22}(x,y) = x^2 + y^2 \\ f_{21}(x,y) = -(x+y) & f_{22}(x,y) = x - y. \end{cases}$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y) = x^{13/2}y^{13/2}$ and $v(x,y) = \sin(x)\sin(y)$.

Problem 2. Determine $u = u(x,y)$ and $v = v(x,y)$ such that

$$\begin{aligned} 9u_{xx} + u_{xy} + 4u_{yy} + 3u_x + u_y + u + 9v_{xx} + v_{xy} + 5v_{yy} + 3v_x + 2v_y + 4v = g_1 \\ 4u_{xx} + u_{xy} + 5u_{yy} + u_x + 2u_y + 3u + 3v_{xx} + v_{xy} + 6v_{yy} + 2v_x + 2v_y + 4v = g_2 \end{aligned} \quad \text{in } \Omega,$$

subject to

$$\left.\begin{array}{r} u+2v \\ 4u+3v \end{array}\right\} \text{ on } x=0, \qquad \left.\begin{array}{r} u-u_x+2v-2v_x \\ 4u-4u_x+3v-3v_x \end{array}\right\} \text{ on } x=1,$$

$$\left.\begin{array}{r} u-u_y+2v-2v_y \\ 4u-4u_y+3v-3v_y \end{array}\right\} \text{ on } y=0, \qquad \left.\begin{array}{r} u+2v \\ 4u+3v \end{array}\right\} \text{ on } y=1.$$

Here, $\Omega \equiv (0,1)\times(0,1)$. The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y)=e^{x+y}$ and $v(x,y)=\sin(x)\sin(y)$.

Problem 3. Determine $u=u(x,y)$ and $v=v(x,y)$ such that

$$\begin{array}{r} 6u_{xx}+3u_{yy}+u+7v_{xx}+4v_{yy}+v=g_1 \\ 3u_{xx}+4u_{yy}+u+2v_{xx}+5v_{yy}+v=g_2 \end{array} \quad \text{in } \Omega\equiv(0,1)\times(0,1),$$

subject to

$$\begin{array}{r} u=0 \\ v=0 \end{array} \quad \text{on } \partial\Omega.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y)=(x^2-x)(y^2-y)e^{x+y}$ and $v(x,y)=x^{9/2}(x-1)^2y^{9/2}(y-1)^2$.

Problem 4. Determine $u=u(x,y)$ and $v=v(x,y)$ such that

$$\begin{array}{r} 6u_{xx}+3u_{yy}+u+7v_{xx}+4v_{yy}+v=g_1 \\ 3u_{xx}+4u_{yy}+u+4v_{xx}+5v_{yy}+v=g_2 \end{array} \quad \text{in } \Omega\equiv(0,1)\times(0,1),$$

subject to

$$\begin{array}{r} u=0 \\ v=0 \end{array} \quad \text{on } \partial\Omega.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y)=(x^2-x)(y^2-y)e^{x+y}$ and $v(x,y)=x^{9/2}(x-1)^2y^{9/2}(y-1)^2$.

Problem 5. Determine $u=u(x,y)$ and $v=v(x,y)$ such that

$$\begin{array}{r} \nabla^2u+\frac{1}{1-2\nu}(u_{xx}+v_{xy})=g_1 \\ \nabla^2v+\frac{1}{1-2\nu}(v_{yy}+u_{xy})=g_2 \end{array} \quad \text{in } \Omega\equiv(0,1)\times(0,1),$$

subject to

$$\begin{matrix} u = 0 \\ v = 0 \end{matrix} \quad \text{on } \partial\Omega.$$

Here, $\nu = .22$. The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x,y) = x^{9/2}(x-1)^2 y^{9/2}(y-1)^2$.

Problem 6. Determine $u = u(x,y)$ and $v = v(x,y)$ such that

$$\begin{matrix} L_{11}u + L_{12}v = g_1 \\ L_{21}u + L_{22}v = g_2 \end{matrix} \quad \text{in } \Omega \equiv (0,1) \times (0,1),$$

subject to

$$\begin{matrix} u = 0 \\ v = 0 \end{matrix} \quad \text{on } \partial\Omega,$$

where, for $i,j = 1,2$, $L_{ij}u \equiv a_{ij}u_{xx} + b_{ij}u_{xy} + c_{ij}u_{yy} + d_{ij}u_x + e_{ij}u_y + f_{ij}u$, and

$$\left\{ \begin{array}{ll} a_{11}(x,y) = 6 + xy & a_{12}(x,y) = x + y \\ b_{11}(x,y) = x^2 + y^2 & b_{12}(x,y) = \sqrt{2} \\ c_{11}(x,y) = 3 + e^{x+y} & c_{12}(x,y) = xy \\ d_{11}(x,y) = -(x+y) & d_{12}(x,y) = x - y \\ e_{11}(x,y) = 1/(1+xy) & e_{12}(x,y) = -(x+y) \\ f_{11}(x,y) = \sqrt{2} & f_{12}(x,y) = 1/(1+xy) \\ a_{21}(x,y) = 1/(1+xy) & a_{22}(x,y) = 10 + xy \\ b_{21}(x,y) = xy & b_{22}(x,y) = x - y \\ c_{21}(x,y) = x + y & c_{22}(x,y) = 7(x^2+y^2) + 5 \\ d_{21}(x,y) = x - y & d_{22}(x,y) = \sqrt{2} \\ e_{21}(x,y) = \sqrt{2} & e_{22}(x,y) = x - y \\ f_{21}(x,y) = -(x+y) & f_{22}(x,y) = x - y. \end{array} \right.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x,y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x,y) = x^{9/2}(x-1)^2 y^{9/2}(y-1)^2$.

Problem 7. Determine $u = u(x, y)$ and $v = v(x, y)$ such that

$$\left.\begin{array}{l} 5u_{xx} + 5u_{yy} + \frac{1}{x+10^{-2}} u_x + \frac{1}{y+10^{-2}} u_y + v_{xx} + v_{yy} - v = g_1 \\ u_{xx} + u_{yy} - u + 6v_{xx} + 6v_{yy} + \frac{1}{x+10^{-2}} v_x + \frac{1}{y+10^{-2}} v_y = g_2 \end{array}\right. \quad \text{in } \Omega \equiv (0,1) \times (0,1),$$

subject to

$$\left.\begin{array}{l} u = 0 \\ v = 0 \end{array}\right. \quad \text{on } \partial\Omega.$$

The functions g_1 and g_2 are chosen so that the solution to the problem is $u(x, y) = (x^2 - x)(y^2 - y)e^{x+y}$ and $v(x, y) = x^{9/2}(x-1)^2 y^{9/2}(y-1)^2$.

Bibliography

[1] C. C. Christara. Quadratic spline collocation methods for elliptic partial differential equations. Technical Report 235, Department of Computer Science, University of Toronto, 1990.

[2] C. C. Christara. Quadratic spline collocation methods for elliptic partial differential equations. *BIT*, 34:33–61, 1994.

[3] C. C. Christara. Parallel solvers for spline collocation equations. *Advances in Engineering Software*, 27:71–89, 1996.

[4] C. C. Christara and B. Smith. Multigrid and multilevel methods for quadratic spline collocation. *BIT*, 37(4):781–803, 1997.

[5] A. Constas. Fast Fourier transform solvers for quadratic spline collocation. M.Sc. Thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, July 1996.

[6] C. de Boor and B. Swartz. Collocation at Gaussian points. *SIAM J. Numer. Anal.*, 10(4):582–606, 1973.

[7] E. N. Houstis, C. C. Christara and J. R. Rice. Quadratic-spline collocation methods for two-point boundary value problems. *International Journal for Numerical Methods in Engineering*, 26:935–952, 1988.

[8] E. N. Houstis, J. R. Rice, C. C. Christara and E. A. Vavalis. Performance of scientific software. In J. R. Rice, editor, *Mathematical Aspects of Scientific Software*, pages 123–156. Springer Verlag, 1988.

[9] W. Gropp and B. Smith. *Users Manual for KSP: Data-Structure-Neutral Codes Implementing Krylov Space Methods.* Argonne National Laboratory, August 1993.

[10] C. Hall and T. Porsching. *Numerical Analysis of Partial Differential Equations.* Prentice Hall, 1990.

[11] C. Van Loan. *Computational Frameworks for the Fast Fourier Transform.* Society for Industrial and Applied Mathematics, 1992.

[12] B. Mossberg. *Documentation for the FFT subroutines available in four libraries.* NEC Systems Laboratory, Inc., August 1992.

[13] R. D. Russell. Collocation for systems of boundary value problems. *Numer. Math.*, 23:119–133, 1974.

[14] U. Ascher, J. Christiansen and R. D. Russell. A collocation solver for mixed order systems of boundary value problems. *Mathematics of Computation*, 33(146):659–679, 1979.