UCLA
COMPUTATIONAL AND APPLIED MATHEMATICS

Domain Decomposition Algorithms
(Ph.D. Thesis)

Jian Ping Shao

October 1993

CAM Report 93-38

Department of Mathematics
University of California, Los Angeles
Los Angeles, CA. 90024-1555

UNIVERSITY OF CALIFORNIA

Los Angeles

Domain Decomposition Algorithms

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Mathematics

by

Jian-Ping Shao

1993

The dissertation of Jian-Ping Shao is approved.

Christopher Anderson

Adrienne G. Lavine

Stanley Osher

Owen I. Smith

Tony F. Chan, Committee Chair

University of California, Los Angeles

1993

TABLE OF CONTENTS

1 Introduction 1

1.1 Sobolev Spaces and the Finite Element Method for Elliptic Problems 6

1.2 Space Decomposition 9

1.3 Additive and Multiplicative DD 13

2 Efficient Variants of Vertex Space DD Algorithm 18

2.1 An Elliptic Problem and Its Many Sub-domain Decomposition . . 19

2.1.1 Block Partition of Elliptic Problem 20

2.1.2 Iterative Solution of the Block Partitioned System 22

2.1.3 Some Properties of the Schur Complement S 23

2.2 The BPS and VS Preconditioners for S 26

2.2.1 Notations for a Partition of the Interface B 26

2.2.2 The BPS Preconditioner 28

2.2.3 The Vertex Space Algorithm of Smith and Nepomnyaschikh 31

2.3 Two Variants of the Vertex Space Method 32

2.3.1 Fourier Approximations 34

2.3.2 Probe Approximations 41

2.4 Numerical Results 54

3 Complexity of DD Algorithms 68

3.1 Introduction 68

3.2 Optimal Computational Complexity: Sequential Case 70

3.3 Optimal Computational Complexity: Parallel Case 73

3.4 Higher Dimensional Problems 75

3.5 Concluding Remarks 75

4 DD Methods for Coupled Elliptic Systems 77

4.1 Coupled Elliptic Systems and Two-Level Schwarz Methods 78

4.1.1 Coupled Elliptic Systems 79

4.1.2 Two-Level Schwarz Methods 81

4.2 Cai and Widlund Framework for Coupled Elliptic Problems 85

4.3 Interface Schur Complement Approximations 88

4.3.1 Abstract Fourier Bound Deduced from Sobolev Inequalities 89

4.3.2 Model Coupled Problem and Its Capacitance Interface Matrix 92

4.3.3 Fourier Analysis on Model Problem 96

4.3.4 Probe Technique Applied to the Coupled Elliptic System . 107

4.3.5 Numerical Results 110

5 Domain Decomposition for Singular Neumann Boundary Value Problems 120

5.1 General Framework for Neumann Boundary Problems 123

5.1.1 Additive Schwarz Framework 123

5.1.2 Elliptic and Elasticity problem 132

5.2 The Modified Domain Decomposition Method 135

5.2.1 The Modified Additive Schwarz Method 135

5.2.2 Schur Complement on Γ 139

5.2.3 Description of the Modified VS Algorithm 143

5.3 The Variants of the Vertex Space Method 147

5.3.1 The Symmetric Probe Technique 148

5.3.2 The Other Simple Vertex Space Approximations 155

5.4 The Modified BPS Method for Neumann Boundary Value Problems 156

5.4.1 The BPS Method and Its Modification 157

5.4.2 Theoretical Results 163

5.5 Numerical Results 166

6 Algebraic Domain Decomposition Methods 175

6.1 General Chaotic Relaxation Schwarz Methods 175

6.2 Algebraic Additive and Multiplicative Schwarz Methods 181

6.2.1 Direct Subsolver for All Sub-problems 182

6.2.2 Iterative Subsolver for All Sub-problems 190

Bibliography 200

LIST OF FIGURES

2.1 The vertex space partitioning of the interface. 27

2.2 Simultaneous probe vectors 44

2.3 Numbering of Edges. 45

2.4 Ordering of unknowns within each vertex sub-region V_k 48

2.5 Discontinuous coefficients $a(x,y)$ 55

4.1 Domain Ω and Sub-domains Ω_i 93

5.1 The Model Rectangular Geometry With Boundary Value 149

5.2 Probe Vectors in Modified VS method 154

ACKNOWLEDGMENTS

I wish to express my sincerest gratitude to my advisor Professor Tony F. Chan, whose encouragement, patience and expert guidance have made this dissertation possible. I also want to thank the other members of the committee, Professor Christopher Anderson, Professor Adrienne G. Lavine, Professor Stanley Osher, and Professor Owen I. Smith, for their interest and precious time. Special thanks go to Professor Li-shan Kang, Tarek P. Mathew, Barry Smith and Panayot Vassilevski for their valuable suggestions and discussions. Finally, I would like to express my thanks to Winnie Wong and Theodore Szeto for their proof-reading and corrections.

ABSTRACT OF THE DISSERTATION

Domain Decomposition Algorithms

by

Jian-Ping Shao

Doctor of Philosophy in Mathematics

University of California, Los Angeles, 1993

Professor Tony F. Chan, Chair

Domain decomposition (DD) has been widely used to design parallel efficient algorithms for solving elliptic problems. In this thesis, we focus on improving the efficiency of DD methods and applying them to more general problems. Specifically, we propose efficient variants of the vertex space DD method and minimize the complexity of general DD methods. In addition, we apply DD algorithms to coupled elliptic systems, singular Neumann boundary problems and linear algebraic systems.

We successfully improve the vertex space DD method of Smith by replacing the exact edge, vertex dense matrices by approximate sparse matrices. It is extremely expensive to calculate, invert and store the exact vertex and edge Schur complement dense sub-matrices in the vertex space DD algorithm. We propose several approximations for these dense matrices, by using *Fourier approximation* and an algebraic *probing* technique. Our numerical and theoretical results show that these variants retain the fast convergence rate and greatly reduce the computational cost.

We develop a simple way to reduce the overall complexity of domain decomposition methods through choosing the coarse grid size. For sub-domain solvers with different complexities, we derive the optimal coarse grid size H_{opt}, which asymptotically minimizes the total computational cost of DD methods under the sequential and parallel environments. The overall complexity of DD methods is significantly reduced by using this optimal coarse grid size.

We apply the additive and multiplicative Schwarz algorithms to solving coupled elliptic systems. Using the Dryja-Widlund framework, we prove that their convergence rates are independent of both the mesh and the coupling parameters. We also construct several approximate interface sparse matrices by using Sobolev inequalities, Fourier analysis and probe technique.

We further discuss the application of DD to the singular Neumann boundary value problems. We extend the general framework to these problems and show how to deal with the null space in practice. Numerical and theoretical results show that these modified DD methods still have optimal convergence rate.

By using the DD methodology, we propose algebraic additive and multiplicative Schwarz methods to solve general sparse linear algebraic systems. We analyze the eigenvalue distribution of the iterative matrix of each each algebraic DD method to study the convergence behavior.

CHAPTER 1

Introduction

The rapid development of advanced-architecture computers (concurrent multiprocessors) has had a very significant impact on all aspects of scientific computation. "Divide and Conquer" is a basic strategy used in practice to design parallel numerical methods which can be most effectively implemented on parallel machines. Domain decomposition is a natural way for distributing programs across processors and data across memory. Domain decomposition methods often provide suitable techniques in designing efficient parallel algorithms for solving large linear systems of equations arised from discretizing partial differential problems. Moreover, these methods can be advantageous for the efficient and local treatment of irregular geometries, discontinuous coefficients, local grid refinement, boundary layers and coupling between equations of different type, see [38, 14, 15, 16, 44].

Domain decomposition methods are the generalization of the alternative methods of Schwarz [62] proposed more than 120 years ago. In recent years, research on these methods has become very active; see [38, 14, 15, 16, 44] and references therein. The earlier work by P. L. Lions [48, 49, 50] has made an important impetus on the development of space decomposition. Later a general abstract framework on additive Schwarz methods was developed by Dryja and Widlund

[32, 33, 34, 35, 73], Matsokin and Nepomnyaschikh [53], Nepomnyaschikh [55] and others. Recently, an abstract theory for multiplicative Schwarz methods has been obtained by Bramble, Pasciak, Wang and Xu [11] and Xu [75]. The extension of uniform theories to certain nonsymmetric elliptic equations was given by Cai and Widlund [12, 13]. Iterative sub-structuring methods, which decompose the given domain into non-overlapping sub-domains, have been studied by Bramble, Pasciak and Schatz [8, 10], Glowinski, Le Tallec, de Roeck et al. [6, 60], Mandel [52, 51] Dryja and Widlund [35] and Smith [65].

Domain decomposition (DD) methods refer to a class of techniques for solving elliptic boundary value problems in which the solution is obtained by iteratively solving smaller versions of the original problem on smaller (overlapping or non-overlapping) sub-domains. In most domain decomposition methods, a coarse problem is introduced to provide global data exchange in order to produce an optimal method. Domain decomposition methods, as preconditioned iterative methods, were classified into additive Schwarz method and multiplicative Schwarz method according to the ordering of solving subproblems. It has been proved that the convergence rate of most additive Schwarz methods and multiplicative Schwarz methods is independent of the coarse grid size and the fine grid size [32, 33, 34, 35, 73, 11, 75]. Domain decomposition methods can also be classified into overlapping domain decomposition and non-overlapping domain decomposition. Generally, overlapping DD converges faster than nonoverlapping DD. Although the convergence rate of the iterative sub-structuring (non-overlapping DD) algorithms depends slightly on the mesh parameters, it is insensitive to the jump (discontinuity) of the coefficients across subdomains.

In this dissertation, we start with the derivations of several efficient variants of vertex space domain decomposition method. These methods include variants of the vertex space algorithm (VS) proposed by Smith [66] and Nepomnyaschikh [54], and an algorithm of Bramble, Pasciak and Schatz (BPS) [8]. All of these methods are based on non-overlapping sub-regions, in which the reduced Schur complement system on the interface is solved using a generalized block Jacobi type preconditioner with the blocks corresponding to the vertex space, edges and a coarse grid. Constructing these dense sub-block matrices and their inverses is extremely expensive. Therefore, we replace these exact dense matrices by approximate sparse matrices. We consider two kinds of approximations for the edge and vertex space sub-blocks, one is based on *Fourier approximation*, the other is based on an algebraic *probing* technique in which sparse approximations to these sub-blocks are computed. Our motivation is to improve efficiency of the algorithms without sacrificing the optimal convergence rate. Our numerical and theoretical results on the performance of these algorithms, show that these variants greatly reduce the computation in each iteration and converge with optimal rate.

Secondly, we develop a simple way to reduce the overall complexity of domain decomposition methods. We analyze the complexity of domain decomposition on serial and parallel machines. It has been observed empirically [42, 64] that the total cost of a method can depend sensitively on the choice of the coarse grid size, H, in addition to the obvious dependence on the efficiency of the subdomain solver. A small H generally improves the convergence rate at the cost of a more costly coarse grid solve, whereas a large H has the opposite effect. Therefore, an optimal value often exists. For sub-domain solvers with different complexities, we derive the optimal values H_{opt}, which asymptotically minimize the total computational cost of DD methods by considering the number of floating point operations in the sequential case and the execution time in the parallel case. The overall complexity of domain decomposition methods is substantially reduced by just using the optimal coarse grid size H_{opt}.

Thirdly, we will study the applications of domain decomposition methods to coupled elliptic systems arising from many practical problems in such as semiconductor model and elasticity. By choosing a fine enough coarse grid, we show that the convergence rates of additive and multiplicative Schwarz methods are independent of both the mesh parameters and the coupling parameters. We also discuss the approximate sparse interface sub-matrices for iterative substructuring domain decomposition for the coupled system. In these DD methods, the reduced Schur complement system on the interface is solved by using a generalized block Jacobi type preconditioner, with the blocks corresponding to the vertex space, edges and a coarse grid. These edge and vertex matrices are expensive to form explicitly. Therefore, we propose several approximate sparse interface matrices. These approximate matrices are constructed through using a Fourier approximation and a probing technique. We show that the exact interface Schur complement matrix is spectrally equivalent to the approximate matrix by using Sobolev inequalities and Fourier analysis. The numerical results show that these Fourier approximations are successful.

Fourthly, we consider the applications of vertex space DD methods to singular Neumann boundary condition. We carefully present how to deal with the null space in practice. We show that the modified DD methods for these singular problems still have the same optimal convergence rate. We also prove that the convergence rate of the modified BPS method is insensitive to highly discontinuous coefficient across the substructures. Numerical experiments have been conducted for the Neumann boundary value problems with various coefficients. These numerical results verify our theoretical results.

Finally, we apply the domain decomposition methodology to design algorithms for the general linear algebraic systems. We propose two kinds of methods: one is algebraic additive Schwarz (AAS) method and the other is algebraic multiplicative Schwarz (AMS) method. We analyze the eigenvalue distribution of the iterative matrices in order to know the convergence factor of these algebraic domain decomposition methods.

For the rest of Chapter 1, we will review some basic Sobolev spaces and discuss the general framework of the domain decomposition for elliptic problems. In Chapter 2 , we will discuss the development of efficient variants of vertex space domain decomposition. In Chapter 3, we will analyze the serial and parallel complexity of domain decomposition. In Chapter 4, we will apply the DD method to coupled elliptic systems. In Chapter 5, we will apply the vertex space domain decomposition to the singular Neumann boundary value problems. In Chapter 6, we present the algebraic domain decomposition method.

1.1 Sobolev Spaces and the Finite Element Method for Elliptic Problems

Assume that $\Omega \in R^d$ is a bounded Lipschitz domain. Let C and c be generic constants. We introduce a Hilbert space V, which is one of the following Hilbert spaces

$$H^m(\Omega) = \{v | D^\alpha v \in L^2(\Omega), \quad \text{if } |\alpha| \leq m\}$$

where the multi-index $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$ is an integer vector and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. The inner product space $L^2(\Omega)$ is defined by

$$(u, v) = \int_\Omega uv dx,$$

and the corresponding L^2 norm is

$$||u||^2_{L^2(\Omega)} = (u, u).$$

Analogously, we can, respectively, define the semi-norm and norm of $H^m(\Omega)$ by

$$|v|^2_{H^m(\Omega)} = \int_\Omega \sum_{|\alpha|=m} D^\alpha u D^\alpha u dx,$$

and

$$||v||^2_{H^m(\Omega)} = \int_\Omega \sum_{|\alpha|\leq m} D^\alpha u D^\alpha u dx.$$

Let H_Ω be the diameter of domain Ω. The following inequalities establish the equivalences of certain norms in the Hilbert space $H^1(\Omega)$.

Lemma 1.1 (Friedrichs' inequality) *There exists a constant $C(\Omega) > 0$, which depends only on the Lipschitz constant of the boundary of Ω, such that, for all $v \in H_0^1(\Omega)$,*

$$||v||_{L^2(\Omega)} \leq C(\Omega) H_\Omega |v|_{H^1(\Omega)}. \tag{1.1}$$

Lemma 1.2 (Poincaré's inequality) *There exists a constant $C(\Omega) > 0$, which depends only on the Lipschitz constant of the boundary of Ω, such that, for all $v \in H_0^1(\Omega)$,*

$$||v||^2_{H^1(\Omega)} \leq C(\Omega)(H_\Omega^2 |v|^2_{H^1(\Omega)} + \frac{1}{H_\Omega^d}(\int_\Omega v dx)^2).$$

The proofs of these inequalities may be found in [57, 58].

Many elliptic problems can be mathematically represented by the following variational problems: Find $u \in V$ such that

$$a(u,v) = f(v) \qquad \text{for all } v \in V, \tag{1.2}$$

where $a(\cdot,\cdot) : V \times V \to \mathbf{R}$ is a continuous bilinear form, and $f : V \to \mathbf{R}$ is a continuous linear form. Let $||\cdot||_V$ denote the norm of the Hilbert space V.

We usually assume that the bilinear form $a(\cdot,\cdot)$ is symmetric

$$a(u,v) = a(v,u) \qquad \text{for all } u, v \in V,$$

continuous

$$|a(u,v)| \leq C||u||_V||v||_V \qquad \text{for all } u, v \in V,$$

and strongly elliptic (coercive)

$$a(v,v) \geq c||v||_V^2 \qquad \text{for all } v \in V.$$

From these properties of $a(\cdot,\cdot)$, we can define an equivalent norm in space V as $||v||_a = \sqrt{a(v,v)}$.

The finite element formulation is obtained by replacing the infinite dimensional space V with a finite dimensional space V^h. Then the discrete approximate problem can be formed by using the Galerkin method. Find $u_h \in V^h$ such that

$$a(u_h, v_h) = f(v_h) \qquad \text{for all } v_h \in V^h. \tag{1.3}$$

Let the domain Ω be partitioned into non-overlapping regions called elements. Generally, these elements are triangles or rectangles with the approximate diameter h. Let V^h be the space of piecewise polynomials on the triangulation. Let $\{\phi_i\}_{i=1}^n$ be the chosen basis of V^h, and $u_h = \sum_{i=1}^N x_i\phi_i$. Then, (1.3) leads to a system of linear algebraic equations:

$$Ax = b$$

where $b_i = f(\phi_i)$. Generally, the condition number κ of this stiffness matrix A tends to infinity when h tends to zero. We usually have $\kappa(A) \approx O(h^{-s})$ with $s > 0$. Therefore, solving this linear system, especially when its size is very large, can be very expensive. Many preconditioners have been designed for A in order to obtain more efficient methods for solving this problem.

1.2 Space Decomposition

Let $\{V_i^h\}_{i=0}^J$ be subspaces of V^h. Assume that

$$V^h = V_0^h + V_1^h + \cdots + V_J^h. \tag{1.4}$$

Suppose $T_i, \quad i = 0, 1, \cdots, J$, are operators, $\quad T_i : V^h \to V_i^h$.

In domain decomposition, the functions in the subspace V_i^h are defined only on the subdomain Ω_i, and the operators T_i are usually chosen to be projector from V^h into subspace V_i^h. Concrete V_i^h and T_i will be given later. Now we briefly describe the additive Schwarz methods and multiplicative Schwarz methods. The detail description of these two kinds of methods can be found in Dryja and Widlund's papers [32, 33], and Bramble, Pasciak, Wang and Xu's papers [11, 75].

Additive Schwarz (AS) Method [32, 33, 53, 55]

To solve equation (1.3) efficiently, we introduce an auxiliary problem:

$$Tu_h = \sum_{i=0}^{J} T_i u_h = g \tag{1.5}$$

which has the same solution as equation (1.3). The additive Schwarz method of solving this auxiliary problem (1.5) is to apply Conjugate gradient (CG) method [40] to this problem.

Generally, one of the most important goals in designing an algorithm is to make the condition number of operator T as small as possible so that the algorithm converges fast. In domain decomposition methods, $T_i u_h$ corresponds to solving the subproblem restricted on subdomain Ω_i. Therefore, the new function $g = \sum_{i=0}^{J} T_i u_h$ can be generally calculated without knowing the solution u_h in practice.

Multiplicative Schwarz (MS) Method [11, 75]

Let $u_h^0 \in V^h$ be given. Assume that $u_h^k \in V^h$ is obtained. Then u^{k+1} is defined by

$$u_h^{k+(i+1)/(J+1)} = u_h^{k+i/(J+1)} + T_i(A^{-1}f - u_h^{k+i/(J+1)})$$

for $i = 0, \cdots, J$.

The error propagation operator E_J for one complete iteration is given by

$$E_J = (I - T_J) \cdots (I - T_1)(I - T_0). \tag{1.6}$$

It is expected that

$$||E_J||_a \leq \delta < 1,$$

when the MS Method converges to the solution of equation (1.3). Note that in practice, we do not need to calculate $A^{-1}f$ while computing the action $T_i A^{-1} f$.

The following assumption allows us to estimate the condition number of operator T and the norm of the error propagation operator E_J. We follow the abstract theories developed by Dryja and Widlund [32, 33], Bramble, Pasick, Wang and Xu [11, 75] to give a simple proof on the bound of condition number and the norm.

Assumption 1.1 *Let T_i be symmetric positive definite. There exists a positive* *constant* $0 < \omega < 2$ *such that*

$$(1.7) \qquad a(T_i v_h, v_h) \geq \omega^{-1} a(T_i v_h, T_i v_h) \qquad \forall v_h \in V \quad i = 0, \cdots, J.$$

From Assumption 1.1, we can easily show that

$$(1.8) \qquad ||T_i||_a \leq \omega \quad \text{and} \quad ||I - T_i||_a \leq 1.$$

Definition 1.1 *Let* $\mathcal{E} = \{\varepsilon_{i,j}\}_{i,j=1}^{J}$ *be the matrix of strengthened Cauchy-Schwarz coefficients, namely:*

$$(1.9) \qquad |a(v_{h,i}, v_{h,j})| \leq \varepsilon_{i,j} ||v_{h,i}||_a ||v_{h,j}||_a \qquad \forall v_{h,i} \in V_i^h, \quad \forall v_{h,j} \in V_j^h.$$

Define $\rho(\mathcal{E})$ *to be the spectral radius of the matrix* $\mathcal{E}$.

Note that $\varepsilon_{i,i} = 1$ and that $0 \leq \varepsilon_{i,j} \leq 1$ by Cauchy-Schwarz inequalities. By Gershgorin's theorem, the spectral radius $\rho(\mathcal{E})$ of the matrix $\mathcal{E}$ is bounded by J, i.e. $\rho(\mathcal{E}) \leq J$.

It is not difficult to show that

$$(1.10) \quad ||\sum_{i=1}^{J} T_i||_a \leq \omega \rho(\mathcal{E}), \quad \text{and} \quad \sum_{i=1}^{J} a(T_i v_h, T_i v_h) \leq \omega \rho(\mathcal{E})^{1/2} a(v_h, v_h)$$

from Definition 1.1.

The inequalities (1.8) and (1.10) imply that

$$(1.11) \qquad ||\sum_{i=0}^{J} T_i||_a \leq \omega(\rho(\mathcal{E}) + 1).$$

This inequality gives an upper bound for $||\sum_{i=0}^{J} T_i||_a$ which is required in analyzing the additive Schwarz method.

Assumption 1.2 *There exists a constant $C_0 > 0$, such that*

$$\sum_{j=0}^{J} a(T_j v_h, v_h) \geq C_0^{-2} a(v_h, v_h) \qquad \forall v_h \in V^h. \tag{1.12}$$

The inequalities (1.11) and (1.12) imply the following theorem.

Theorem 1.1 (Dryja and Widlund [32, 33]) *Under Assumption 1.1 and 1.2, we have*

$$\kappa(T) \leq C_0^2 \omega(\rho(\mathcal{E}) + 1), \tag{1.13}$$

where $\kappa(T)$ is the condition number of operator T.

Now we briefly estimate the norm of error propagation operator E_J. We introduce operator sequence:

$$E_j = (I - T_j) \cdots (I - T_0), \qquad E_{-1} = I.$$

Then, we have

$$E_{j-1} - E_j = T_j E_{j-1}$$

which leads to

$$\begin{aligned} ||E_{j-1} v_h||_a^2 - ||E_j v_h||_a^2 &= ||T_j E_{j-1} v_h||_a^2 + 2a(T_j E_{j-1} v_h, (I - T_j) E_{j-1} v_h) \\ &\geq (2 - \omega) a(T_j E_{j-1} v_h, E_{j-1} v_h). \end{aligned}$$

Thus,

$$||v_h||_a^2 - ||E_J v_h||_a^2 \geq (2-\omega)\sum_{j=0}^{J} a(T_j E_{j-1} v_h, E_{j-1} v_h),$$

$$\text{and} \qquad I - E_{j-1} = \sum_{i=0}^{j-1} T_i E_{i-1} = T_0 + \sum_{i=1}^{j-1} T_i E_{i-1}.$$

In order to get bound on the norm of E_J, we need to estimate the righthandside term by v_h as follows. Let $d_j = a(T_j v_h, v_h)^{1/2}$ and $q_j = a(T_j E_{j-1} v_h, E_{j-1} v_h)^{1/2}$. From the above equations and inequality (1.7), we can deduce that, for $j > 0$,

$$\begin{aligned}
a(T_j v_h, v_h) &= a(T_j v_h, E_{j-1} v_h) + a(T_j v_h, T_0 v_h) + a(T_j v_h, \textstyle\sum_{i=1}^{j-1} T_i E_{i-1} v_h) \\
d_j^2 &\leq d_j q_j + \omega d_j q_0 + \omega \textstyle\sum_{i=1}^{j-1} \varepsilon_{ij} d_j q_i \\
(\textstyle\sum_{j=1}^{J} d_j^2)^{1/2} &\leq (\textstyle\sum_{j=1}^{J} q_j^2)^{1/2} + \omega\sqrt{J} q_0 + \omega\rho(\mathcal{E})(\textstyle\sum_{j=1}^{J} q_j^2)^{1/2} \\
(\textstyle\sum_{j=0}^{J} d_j^2)^{1/2} &\leq 2\max\{1+\omega\rho(\mathcal{E}), 1+\omega\sqrt{J}\}(\textstyle\sum_{j=0}^{J} q_j^2)^{1/2}.
\end{aligned}$$

Thus, we arrive at the following estimation on the norm of E_J.

Theorem 1.2 (Bramble, Pasciak, Wang and Xu [11, 74]) *Assume that Assumption 1.1 and 1.2 are valid. Then, we have*

$$||E_J||_a \leq \sqrt{1 - \frac{2-\omega}{4\max\{1+\omega\rho(\mathcal{E}), 1+\omega\sqrt{J}\}^2 C_0^2}}. \tag{1.14}$$

1.3 Additive and Multiplicative DD

In this section, the operator T_i is constructed by using domain decomposition methods which will satisfy Assumption 1.1 and 1.2. The discussion mainly follows Dryja and Widlund's general framework [32, 33].

Let $\{\Omega_i\}_{i=1}^N$ be a shape regular, coarse finite element triangulation of Ω with H as the maximum diameter of Ω_i. The domain Ω is further divided into elements with diameters of order h. The finite element spaces of continuous, piecewise linear function on these triangulation are denoted by V^H and V^h, respectively. Thus, $V^H \subset V^h \subset V$. For simplicity, we let $V = H_0^1(\Omega)$. Each sub-region Ω_i is extended to a larger region Ω_i^{ext}. We also assume that the boundary $\partial\Omega_i^{ext}$ does not cut through any h-level elements.

The sub-regions Ω_i^{ext} are colored by using colors $1, \cdots, J$ in such a way that no neighboring sub-regions have the same color. The number of colors J, used here should be chosen as small as possible. For example, $J = 4$ suffices, when $\Omega \subset \mathbf{R}^2$. Then we merge all sub-regions of the same color and denote the resulting sets by $\Omega'_1, \cdots, \Omega'_J$.

Now we introduce subspaces $V_j^h = V^h \cap H_0^1(\Omega'_j)$ for $j = 1, \cdots, J$, and denote $V_0^h = V^H$. Then the discrete space V^h can be written as a sum of the $J+1$ subspace;

$$V^h = V_0^h + V_1^h + \cdots + V_J^h.$$

Let $b_j(\cdot,\cdot)$ be a symmetric, positive definite bilinear form on $V_j^h \times V_j^h$. Here, we use b_j as preconditioner of a on the subspace V_j^h. Assume that the bilinear form $b_j(\cdot,\cdot)$ satisfies:

1. For any $u_h \in V^h$, there exists a representation $u_h = \sum_{j=0}^J u_{h,j}, u_{h,j} \in V_j^h$,

with

$$\sum_{j=0}^{J} b_j(u_{h,j}, u_{h,j}) \le C_0^2 a(u_h, u_h) \tag{1.15}$$

2. Let ω_b be the minimum constant such that

$$a(v_h, v_h) \le \omega_b b_j(v_h, v_h) \qquad \forall v_h \in V_j^h, \quad j = 0, 1, \cdots, J \tag{1.16}$$

After introducing the preconditioner $b_j(\cdot,\cdot)$, approximating the $a(\cdot,\cdot)$ on subspace V_j^h, we define projections $T_j : V^h \to V_j^h$ by

$$b_j(T_j u_h, v_h) = a(u_h, v_h) \qquad \forall v_h \in V_j^h, \quad j = 1, \cdots, J. \tag{1.17}$$

Thus $T_j u_h$ can be directly computed by solving subproblems

$$b_j(T_j u_h, v_h) = (f, v_h) \qquad \forall v_h \in V_j^h, \quad j = 1, \cdots, J.$$

By using equation (1.17) and inequality (1.16), we can easily verify Assumption 1.1:

$$a(u_h, T_j u_h) = b_j(T_j u_h, T_j u_h) \ge \frac{1}{\omega_b} a(T_j u_h, T_j u_h).$$

The verification of Assumption 1.2 follows directly from equation (1.17) and inequality (1.15) :

$$\begin{aligned}
a(u_h, u_h) &= \textstyle\sum_{j=0}^J a(u_h, u_{h,j}) = \sum_{j=0}^J b_j(T_j u_h, u_{h,j}) \\
&\le (\textstyle\sum_{j=0}^J b_j(T_j u_h, T_j u_h))^{1/2} (\sum_{j=0}^J b_j(u_{h,j}, u_{h,j}))^{1/2} \\
&\le (\textstyle\sum_{j=0}^J b_j(T_j u_h, T_j u_h))^{1/2} (C_0^2 a(u_h, u_h))^{1/2}, \\
a(u_h, u_h) &\le C_0^2 \textstyle\sum_{j=0}^J b_j(T_j u_h, T_j u_h) = C_0^2 \sum_{j=0}^J a(u_h, T_j u_h).
\end{aligned}$$

This shows the following important Lions' Lemma [49].

Lemma 1.3 (Lions [49]) *Assume that inequalities (1.15) are true. Then*

$$a(u_h, u_h) \leq C_0^2 \sum_{j=0}^{J} b_j(T_j u_h, T_j u_h) = C_0^2 \sum_{j=0}^{J} a(u_h, T_j u_h).$$

We now use an example to show how inequalities (1.15) and (1.16) can be satisfied by choosing the proper preconditioning bilinear form $b_j(\cdot,\cdot)$. Let the bilinear forms $b_j(u_h, v_h) = a(u_h, v_h) = (\nabla u_h, \nabla v_h)$. Then, inequality (1.16) follows directly. The verification of inequality (1.15), shown by Dryja and Widlund [32, 33], is presented in Lemma 1.4. Therefore, the condition number of the additive Schwarz operator T is bounded and the norm of the propagation matrix E_J is less than 1 according to the above abstract theory on space decomposition.

Lemma 1.4 (Dryja and Widlund [32, 33]) *For all $u_h \in V^h$, there exist $u_{h,j} \in V_j^h$ with $u_h = \sum_{j=0}^{J} u_{h,j}$ such that*

$$\sum_{j=0}^{J} ||u_{h,j}||_a^2 \leq C_0^2 ||u_h||_a^2,$$

where constant C_0 is independent of u_h, h and H.

Proof *From Strang [67], there exists a linear map $\hat{I}_H : V^h \rightarrow V^H$ such that*

$$||u_h - \hat{I} u_h||_{L_2(\Omega)}^2 \leq C H^2 |u_h|_{H^1(\Omega)}^2,$$

$$\textit{and} \qquad |u_h - \hat{I} u_h|_{H^1(\Omega)}^2 \leq C |u_h|_{H^1(\Omega)}^2.$$

We then define $w_h = u_h - \hat{I} u_h$ and $u_h^0 = \hat{I} u_h$. and $u_{h,j} = I_h(\theta_j w_h)$. Here I_h is the interpolation operator into the space V^h and θ_j define a partition of unity with $\theta_j \in C_0^\infty(\Omega_j'), 0 \leq \theta_j \leq 1$ and $\sum_{j=1}^{J} \theta_j = 1$. Because of the generous overlap between *sub-regions, these functions can be chosen so that* $|\nabla\theta_j|^2_{L^\infty} \leq CH^2$*. By using the linearity of* I_h*, we can easily show that* $u_h = \sum_{j=0}^{J} u_{h,j}$*. In order to estimate the semi-norm of* $u_{h,j}$*, we work on one element* K *at a time. Let* $\bar{\theta}_j$ *be the average of* θ_j *over* K *. Then,* $||\theta_j - \bar{\theta}_j||^2_{L^\infty(K)} \leq C(h/H)^2$*. By using this inequality and an inverse inequality , we obtained*

$$\begin{aligned} |u_{h,j}|^2_{H^1(K)} &\leq 2|\bar{\theta}_j w_h|^2_{H^1(K)} + 2|I_h(\theta_j - \bar{\theta}_j)w_h|^2_{H^1(K)} \\ &\leq 2|w_h|^2_{H^1(K)} + Ch^{-2}||I_h(\theta_j - \bar{\theta}_j)w_h||^2_{L^2(K)} \\ &\leq 2|w_h|^2_{H^1(K)} + CH^{-2}||w_h||^2_{L^2(K)}. \end{aligned}$$

After summing over all elements K *in* Ω'_j*, we arrived at the inequality*

$$|u_{h,j}|^2_{H^1(\Omega'_j)} \leq 2|w_h|^2_{H^1(\Omega'_j)} + CH^{-2}||w_h||^2_{L^2(\Omega'_j)}.$$

By using Lemma 1.1 and summing over j*, we obtain that*

$$\sum_{j=1}^{J} |u_{h,j}|^2_{H^1(\Omega)} \leq C(|w_h|^2_{H^1(\Omega)} + H^{-2}||w_h||^2_{L^2(\Omega)}).$$

Thus, from the bound of w_h *and* $\bar{I}_h$*, inequality (1.15) follows directly*

$$\sum_{j=0}^{J} |u_{h,j}|^2_{H^1(\Omega)} \leq C_0^2 |u_h|^2_{H^1(\Omega)}.$$

CHAPTER 2

Efficient Variants of Vertex Space DD Algorithm

In this chapter, we primarily focus on the development of efficient versions of divide and conquer type domain decomposition algorithms based on non-overlapping sub-regions for solving self adjoint elliptic problems in two dimensions. The algorithms we are going to describe are variants of the vertex space algorithm (VS) proposed by Smith [66] and Nepomnyaschikh [54], and an algorithm of Bramble, Pasciak and Schatz (BPS) [8]. In both cases, a block Jacobi type preconditioner is used to solve the reduced Schur complement system on the interface. The blocks in the BPS algorithm correspond to the nodes on the edges separating the subdomains and to the collection of vertices of the sub-regions. While in the vertex space algorithm with additional overlapping blocks, centered about each vertex consisting of nodes on the interface close to the vertex, are included to account for coupling amongst the non-overlapping blocks.

In order to implement the original version of the VS preconditioner [66], the subblocks of the Schur complement, which are dense matrices, need to be computed and inverted using direct methods. It can, however, be easily shown that if these sub-blocks are replaced by spectrally equivalent approximations, then the rate of convergence of these algorithms remains asymptotically the same. In order to reduce overhead cost, we therefore focus on constructing approximations which are inexpensive to construct, and which are inexpensive to invert.

Two kinds of approximations will be considered, one based on Fourier approximations of the interface operators, and another based on sparse algebraic approximation of the interface operators by a *probing* technique. The Fourier based approximations can be shown to be spectrally equivalent with respect to mesh size variations. However, their performance can be sensitive to the coefficients. On the other hand, the probing based algorithms adapt well to the coefficients, but can be sensitive to mesh size variations.

In 2.1, we construct the Schur complement on the interfaces. In 2.2, we describe the original versions of the BPS and VS preconditioners for the Schur complement on the interface. In 2.3, we describe the two variants, one based on Fourier approximations, and the other based on the *probing* technique. In 2.4, we present numerical results comparing the rates of convergence of the various preconditioners.

2.1 An Elliptic Problem and Its Many Sub-domain Decomposition

Here we describe the block structure obtained when a self-adjoint elliptic problem is discretized on a domain Ω partitioned into many non-overlapping sub-domains Ω_i with an interface B separating the sub-domains. A reduced Schur complement system is derived for the unknowns on the interface. Some properties of this Schur complement system and an iterative procedure for solving the elliptic problem are described.

2.1.1 Block Partition of Elliptic Problem

We consider the following 2nd order self adjoint elliptic problem on a polygonal domain $\Omega \in R^2$:

$$\left\{ \begin{array}{rcll} -\nabla \cdot (a(x,y)\nabla u) & = & f & \text{in } \Omega \\ u & = & 0 & \text{on } \partial\Omega, \end{array} \right. \tag{2.1}$$

where $a(x,y) \in R^{2\times 2}$ is a symmetric, uniformly positive definite matrix function having $L^\infty(\Omega)$ entries, and $f \in L^2(\Omega)$.

We assume that the domain Ω is partitioned into N non-overlapping sub-domains $\Omega, \cdots, \Omega_N$ of diameter H, which form the elements of a *quasi-uniform* coarse grid triangulation τ^H, see Fig. 2.1. We also assume that the sub-domains Ω_i are refined to produce a fine grid *quasi-uniform* triangulation τ^h having elements of diameter h. Corresponding to the coarse grid and fine grid triangulations, we discretize (2.1) either by using finite elements, see [27], or by using finite difference methods, see [68], resulting in a symmetric positive definite linear system

$$A_h u_h = f_h, \tag{2.2}$$

on the fine grid and

$$A_H u_H = f_H, \tag{2.3}$$

on the coarse grid.

Let I denote the union of the interiors of the sub-domains, and let B denote the interface separating the sub-domains:

$$I = \cup_i \Omega_i, \quad B \equiv (\cup_i \partial\Omega_i) - \partial\Omega.$$

Then, grouping the unknowns in the interior of the sub-domains in the vector u_I and the unknowns on the interface B in the vector u_B, we obtain a reordering of the fine grid problem:

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} u_I \\ u_B \end{bmatrix} = \begin{bmatrix} f_I \\ f_B \end{bmatrix}. \tag{2.4}$$

Here A_{II} corresponds to the coupling between nodes in the interior of the sub-domains. For most discretizations, including five point discretizations, the interior nodes in Ω_i are coupled only to the nodes on the interface B, and not to adjacent sub-domains. In such cases, $A_{II} \equiv blockdiag(A_{11}, \cdots, A_{NN})$ is a block diagonal matrix.

Eliminating interior unknowns u_I, we obtain u_I in terms of u_B:

$$u_I = A_{II}^{-1}\left(f_I - A_{IB}u_B\right), \tag{2.5}$$

and substituting this in the 2nd block row of (2.4) yields an equation for u_B:

$$Su_B = f_B - A_{IB}^T A_{II}^{-1} f_I, \tag{2.6}$$

where $S = A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}$ is referred to as the Schur complement or interface matrix. Some properties of the Schur complement will be discussed in 2.1.3. First, we will outline a procedure for solving (2.4).

2.1.2 Iterative Solution of the Block Partitioned System

System (2.4) can be solved as follows. First problem (2.6) is solved for u_B and then (2.5) is solved for u_I. If direct methods are used to solve (2.6) then S needs to be computed explicitly, and this can be expensive in general (though this is standard practise in the substructuring methods used to solve linear elasticity problems), since it involves computing the action of A_{II}^{-1} on the all columns of A_{IB}. This can be implemented more efficiently through subassembly, see [66], requiring only as many solves on each Ω_i as there are unknowns on $\partial\Omega_i \cap B$. Even if the matrix S has been assembled, it is often preferable to solve (2.6) by an iterative method, since direct methods to solve (2.6) require significant memory storage and computational complexity.

Due to the expense of computing S and solving (2.6) by direct methods, we consider solving (2.6) by a preconditioned iterative method such as the conjugate gradient method, see [40], without the explicit construction of S. In this case only matrix vector products with S are required, and each such matrix vector product requires the solution of one problem on each sub-domain Ω_i. The Schur complement, however, is ill-conditioned with $\kappa(S) \approx O(h^{-1})$, see [5, 8], and therefore requires a preconditioner M; the construction of efficient preconditioners M for S will be the main focus of this paper.

First, we note that the procedure to solve the linear system (2.4) by solving the reduced Schur complement system (2.6) corresponds to a block LU factorization based solution:

$$A = LU = \begin{bmatrix} A_{II} & 0 \\ A_{IB}^T & I \end{bmatrix} \begin{bmatrix} I & A_{II}^{-1}A_{IB} \\ 0 & S \end{bmatrix}, \tag{2.7}$$

for $S = A_{BB} - A_{IB}^T A_{II}^{-1} A_{IB}$. Thus

$$A^{-1} = \begin{bmatrix} I & -A_{II}^{-1}A_{IB}S^{-1} \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} A_{II}^{-1} & 0 \\ -A_{IB}^T A_{II}^{-1} & I \end{bmatrix},$$

and backsolving requires solving two systems with coefficient matrices A_{II} and one system with coefficient matrix S, which will be done using a preconditioned conjugate gradient method. We note that, it is possible to construct a global preconditioner $\tilde{A}$ for A by replacing A_{II} by preconditioner $\tilde{A}_{II}$, and by replacing S by preconditioner M. In this case the inverse of the global preconditioner $\tilde{A}$ has the form:

$$\tilde{A}^{-1} = \begin{bmatrix} I & -\tilde{A}_{II}^{-1}A_{IB}M^{-1} \\ 0 & M^{-1} \end{bmatrix} \begin{bmatrix} \tilde{A}_{II}^{-1} & 0 \\ -A_{IB}^T \tilde{A}_{II}^{-1} & I \end{bmatrix}.$$

Approximations to the sub-matrices $\tilde{A}_{ii}$ can be obtained for instance by replacing it either with a scaled version of the Laplacian, or by other preconditioners, such as ILU, see [18].

2.1.3 Some Properties of the Schur Complement S

The Schur complement matrix S is a discrete approximation to a Steklov-Poincare operator, see [1], which enforces *transmission boundary* conditions on the interface B. In the continuous problem, these transmission boundary conditions correspond to the requirement that the solution u be continuos across the interface and that the flux $\vec{n} \cdot (a(x,y)\nabla u)$ also be continuous across the interface. In the discrete case, the action of the Schur complement on a grid function u_B on B is the same as the action of the discrete operator A_h on the *discrete harmonic extension* of u_B into the sub-domains; More specifically, let $E^h u_B$ denote the *discrete harmonic extension* on B to the interior of the sub-domains:

$$E^h u_B \equiv \left[-A_{II}^{-1} A_{IB} u_B, u_B \right], \tag{2.8}$$

then

$$\begin{bmatrix} A_{II} & A_{IB} \\ A_{IB}^T & A_{BB} \end{bmatrix} \begin{bmatrix} -A_{II}^{-1} A_{IB} u_B \\ u_B \end{bmatrix} = \begin{bmatrix} 0 \\ S u_B \end{bmatrix}.$$

Thus, if R_B denotes the pointwise restriction of nodal values of a grid function onto the nodes on B, then $S u_B = R_B A_h E^h u_B$. In addition,

$$x_B^T S x_B = (E x_B)^T A_h (E x_B). \tag{2.9}$$

This property shows the positive definiteness of the Schur complement. In addition to S being positive definite, it is an M-matrix when A_h is an M-matrix, i.e., $S_{ij} \leq 0$ for $i \neq j$ and $(S^{-1})_{ij} \geq 0$ for all i, j, see [68, 22].

Remark. For finite element discretizations, let $A^{(i)}$ denote the stiffness matrix obtained by integrating the bilinear form on Ω_i, i.e., the discretization of the Neumann problem on Ω_i. For finite difference methods, let $A^{(i)}$ correspond to the discretization with discontinuous coefficients which is $a(x,y)$ in Ω_i and zero outside Ω_i. Then, the energy x^TAx can be partitioned as

$$x^T A_h x = \sum_{i=1}^{N} x^T A^{(i)} x, \tag{2.10}$$

and correspondingly, the Schur complement S can be partitioned:

$$x_B^T S x_B = \sum_{i=1}^{N} x_B^T S^{(i)} x_B, \tag{2.11}$$

where

$$S^{(i)} = R_B A^{(i)} E^h. \tag{2.12}$$

Each $S^{(i)}$ is a map of the Dirichlet values u_B to the *normal derivatives* on $\partial\Omega_i \cap B$ of the discrete harmonic extension $E^h u_B$, and this is not a local operator, i.e., the matrix $S^{(i)}$ is dense on $\partial\Omega_i \cap B$, see [5]. In the two sub-domain case, $S = S^{(1)} + S^{(2)}$ is thus a map of the Dirichlet value u_B to the jump in the normal derivatives on B of the discrete harmonic extension $E^h u_B$, which corresponds to the discrete approximations of the transmission boundary condition. In the two dimensional case, the entries of S decay as $|S_{ij}| = 0(\frac{1}{|i-j|^2})$, see Golub-Mayer [39], and preconditioners for S have been studied extensively, see [5, 17, 9, 29, 19]. The important properties of the two sub-domain Schur complement is that its entries decay away from its main diagonal, and that it is uniformly spectrally equivalent to the square root of the Laplace operator on B, as the mesh size goes to zero. Due to this connection, it can be shown that its condition number grows as $\kappa(S) = O(\frac{1}{h})$, see [5]. Applications of both these properties will be discussed in 2.3.1 and 2.3.2.

2.2 The BPS and VS Preconditioners for S

We will describe two preconditioners for S in this Section, one introduced by Bramble, Pasciak and Schatz (BPS) [8], and another, the vertex space preconditioner (VS) introduced by Smith [66] and Nepomnyaschikh [54]. Both these can be interpreted as generalized block Jacobi type preconditioners for (2.6) with overlapping blocks and involving residual correction on a coarse grid. Variants of these preconditioners will be discussed in 2.3.

2.2.1 Notations for a Partition of the Interface B

In the case of many sub-domains, the interface B can be partitioned as a union of edges E_{ij} and cross-points V, see Fig. 2.1.

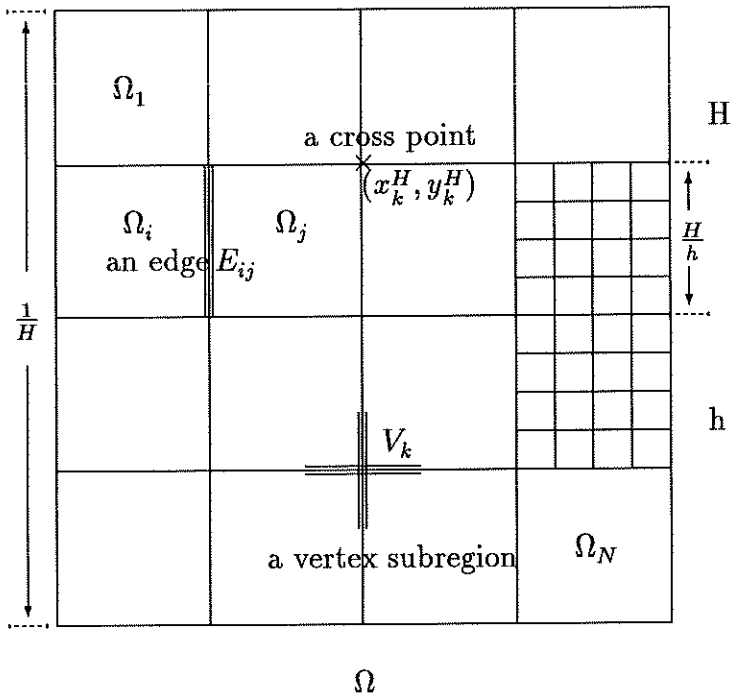
$$B = \cup_{ij} E_{ij} \quad \cup V,$$

where E_{ij} denotes the edge separating sub-domains Ω_i and Ω_j, and V denotes the collection of cross-points (vertices (x_k^H, y_k^H) of the sub-domains).Note that the edges E_{ij} are assumed not to include its endpoints.

For each edge E_{ij} we define $R_{E_{ij}}$ as the pointwise restriction of nodal values to E_{ij}, i.e., if g_B is a grid function defined on B, and if E_{ij} contains n_{ij} interior nodes, then its restriction $R_{E_{ij}} g_B$ is a vector with n_{ij} components defined on E_{ij} by

$$R_{E_{ij}} g_B = g_B \text{ on } E_{ij}.$$

Figure 2.1: The vertex space partitioning of the interface.



Its transpose $R^T_{E_{ij}}$ extends grid functions in E_{ij} by zero to the rest of B:

$$R^T_{E_{ij}} g_{E_{ij}} = \begin{cases} g_{E_{ij}} & \text{on } E_{ij} \\ 0 & \text{on } B - E_{ij} \end{cases} .$$

Similarly, we define R_V as the pointwise restriction map onto the cross-points; if g_B is a grid function on B, and if there are n_V cross-points on B, then $R_V g_B$ is a vector with n_V components defined by

$$R_V g_B = g_B \text{ on } V.$$

Its transpose R^T_V, is thus extension by zero of nodal values in V to B:

$$R^T_V g_V = \begin{cases} g_V & \text{on } V \\ 0 & \text{on } B - V \end{cases} .$$

2.2.2 The BPS Preconditioner

In order to motivate the construction of the BPS preconditioner, we first define a block Jacobi preconditioner M_J consisting of diagonal blocks of the Schur complement S in the following block partitioning of the interface B. Let us suppose there are n edges E_{ij} with some ordering $E_1, \cdots, E_n$. If the unknowns on each edge E_i is grouped together in u_{E_i}, and if the unknowns on the cross-points are grouped in u_V, then S has the following block partitioning corresponding to $(u_{E_1}, \cdots, u_{E_n}, u_V)$:

$$S = \begin{bmatrix} S_{E_1} & \cdots & S_{E_1 E_n} & S_{E_1 V} \\ \vdots & \cdots & \vdots & \vdots \\ S_{E_1 E_n}^T & \cdots & S_{E_n} & S_{E_n V} \\ S_{E_1 V}^T & \cdots & S_{E_n V}^T & S_V \end{bmatrix}.$$

Here, $S_{E_i E_j} \equiv R_{E_i} S R_{E_j}^T$ denotes the coupling in S between nodes on E_i and E_j, and $S_{E_i V} \equiv R_{E_i} S R_V^T$ denotes the coupling in S between nodes on E_i and V. Note that edges E_i and E_j will be coupled in S only if they are part of the boundary of a common sub-domain Ω_k. This can be seen by using the relation between Schur complement and discrete harmonic extensions; since, for instance, discrete harmonic extensions of grid functions on edge E_i is non-zero only in the subdomains that for which E_i is part of its boundary. S is thus a block sparse matrix and corresponding to each edge E_{ij}, the sub-matrix $S_{E_{ij}}$ is identical to the two subdomain Schur complement on interface E_{ij} separating Ω_i and Ω_j. The sub-matrix

S_V which corresponds to coupling in S between cross-points is almost a diagonal matrix since the cross-points are weakly coupled in S. In the case of five point discretizations on rectangular sub-domains, S_V is diagonal since the corner nodes (cross-points) of rectangular domains do not influence the solution in the interior.

For this block partition of S, we define the action of the inverse of the block Jacobi preconditioner M_J:

$$M_J^{-1} g_B = \sum_{\text{edges } ij} R_{E_{ij}}^T S_{E_{ij}}^{-1} R_{E_{ij}} f_B + R_V^T S_V^{-1} R_V f_B. \tag{2.13}$$

This block Jacobi preconditioned system can be shown to have a a condition number satisfying:

$$c_1 H^{-2} \leq \frac{\lambda_{max}(M_J^{-1} S)}{\lambda_{min}(M_J^{-1} S)} \leq c_2 H^{-2} (1 + \log^2(H/h)),$$

where c_1 and c_2 are independent of H and h, see [8, 70]. This indicates that as $H \to 0$, i.e., as the number of sub-domains increases, the rate of convergence deteriorates. This can be attributed to the absence of global communication of information amongst all the edges in the preconditioning step.

The original version of the BPS algorithm [8] involves two changes to this block Jacobi preconditioner. One is that the sub-matrices $S_{E_{ij}}$ are replaced by Fourier based approximations $\tilde{S}_{E_{ij}}$ which will be described in 2.3. The second change is to incorporate global coupling in order to obtain a rate of convergence which does not deteriorate as the number of sub-domains is increased. In order to do this, the cross-points correction term $R_V^T S_V^{-1} R_V$ in (2.13) is replaced by a coarse grid correction term $R_H^T A_H^{-1} R_H$ as in two level multigrid methods (involving weighted restriction and interpolation maps R_H and R_H^T respectively). These are defined below. Let $\phi_{k,H}$ denote the kth coarse grid piecewise linear finite element basis function

$$\phi_{k,H}(x_l^H, y_l^H) = \begin{cases} 1 & \text{if } l = k \\ 0 & \text{if } l \neq k \end{cases},$$

where (x_l^H, y_l^H) is the lth cross-point. Then,

$$(R_H f_B)(x_k^H, y_k^H) \equiv \sum_{(x_j, y_j)} \phi_{k,H}(x_j^H, y_j^H) f_B(x_j^H, y_j^H).$$

Its transpose R_H^T thus denotes linear interpolation of the nodal values on the endpoints of edges E_{ij}:

$$\left(R_H^T g_V\right)(x, y) \equiv \sum_k g_V(x_k^H, y_k^H)\phi_{k,H}(x, y), \quad (x, y) \in B.$$

With these changes, the BPS preconditioner can be defined:

$$M_{BPS}^{-1} f_B = \sum_{\text{edges } i,j} R_{E_{ij}}^T \tilde{S}_{E_{ij}}^{-1} R_{E_{ij}} f_B + R_H^T A_H^{-1} R_H f_B.$$

These changes improve the condition number over that of the block Jacobi version.

Theorem 2.1 *The BPS preconditioner satisfies*

$$\frac{\lambda_{max}(M_{BPS}^{-1} S)}{\lambda_{min}(M_{BPS}^{-1} S)} \leq c_2(1 + \log^2(H/h)),$$

where c_2 is independent of H and h.

Proof *See [8] and [70].*

Remark. It can be easily verified that for five point discretizations of the Laplacian, the coarse grid Schur complement matrix $S_H \equiv R_H S_h R_H^T$ is equal to the coarse grid discretization $A_H = R_H^T A_h R_H$, since piecewise linear interpolation results in grid functions which are discrete harmonic on the sub-domains. In case of more general coefficients, it can be shown that A_H and S_H are spectrally equivalent with respect to coarse grid size H.

2.2.3 The Vertex Space Algorithm of Smith and Nepomnyaschikh

The logarithmic growth in the condition number of the BPS preconditioner can be attributed to the neglect of coupling between adjacent edges of B. The VS preconditioner of Smith [66] and Nepomnyaschikh [54] incorporates some coupling between adjacent edges through the use of certain overlapping blocks of S corresponding to nodes on certain *vertex regions* V_k, which will be defined, and it leads to a condition number independent of mesh parameters.

Let V_k denote the portion of B within a distance of βH from (x_k^H, y_k^H) for some positive fraction $0 < \beta < 1$, see Fig. 2.1. We refer to each V_k as a vertex region or vertex space. We define the corresponding pointwise restriction map R_{V_k} to be

$$R_{V_k} g_B = g_B \text{ on } V_k.$$

Its transpose $R_{V_k}^T$ is thus extension by zero outside V_k:

$$R_{V_k}^T g_{V_k} = \begin{cases} g_{V_k} & \text{on } V_k \\ 0 & \text{on } B - V_k. \end{cases}$$

Corresponding to each vertex region V_k, the sub-matrix S_{V_k} is defined by $S_{V_k} =$

$R_{V_k} S R_{V_k}^T$. The action of the inverse of the vertex space preconditioner M_{VS} involves the inversion of these new overlapping blocks in addition to the blocks used in the BPS preconditioner:

(2.14)

$$M_{VS}^{-1} f_B = R_H^T A_H^{-1} R_H f_B + \sum_{E_{ij}} R_{E_{ij}}^T (S_{E_{ij}})^{-1} R_{E_{ij}} f_B + \sum_{V_k} R_{V_k}^T (S_{V_k})^{-1} R_{V_k} f_B.$$

The following result is proved in [66, 54].

Theorem 2.2 *Suppose the overlap of the vertex regions V_k is βH, then:*

$$\frac{\lambda_{max}(M_{VS}^{-1} S)}{\lambda_{min}(M_{VS}^{-1} S)} \leq C(\beta),$$

where $C(\beta)$ is independent of H and h.

Remark. Other bounds are available for the condition number of the vertex space preconditioned system:

$$\frac{\lambda_{max}(M_{VS}^{-1} S)}{\lambda_{min}(M_{VS}^{-1} S)} \leq \begin{cases} c_1(1 + c_2(1/\beta)), \\ c_3(1 + \log^2(H/h)), \end{cases}$$

where c_1, c_2 are independent of H and h, but may possibly depend on the coefficients $a(x, y)$, while c_3 is independent of H, h and the coefficients $a(x, y)$ provided the coefficients are constant in each sub-domain Ω_i, see [66, 33].

2.3 Two Variants of the Vertex Space Method

An important consideration in the implementation of the algorithms is the expense of computing the edge and vertex matrices $S_{E_{ij}}$ and S_{V_k}, respectively, and the cost of solving the subproblems using direct methods. If there are n_i nodes on each $\partial\Omega_i \cap B$, then computing all the sub-matrices $S_{E_{ij}}$ and S_{V_k} would require solving n_i problems on each Ω_i, and this increases as the mesh size h is reduced. If n_{ij} is the number of nodes on E_{ij}, the cost of using direct methods to solve edge problems is $O(n_{ij}^2)$ once the Cholesky factorizations have been determined, see [66, 65], since the edge sub-matrices $S_{E_{ij}}$ are dense. n_{ij} increases as the mesh size h is reduced.

This expense can be significantly reduced if the exact edge and vertex matrices are replaced by approximations which can be computed at significantly less cost, and which can be inverted at less cost. If these approximations are spectrally equivalent to the exact sub-matrices, then the overall preconditioner would remain spectrally equivalent to the exact VS preconditioner, and the number of iterations required to solve (2.6) would remain independent of h, see Theorem 2.4.

In this Section, we describe two variants of the vertex space and BPS algorithms in which the exact edge and vertex matrices are replaced by approximations. One variant is based on Fourier approximations of both the edge and vertex matrices, while the other variant is based on sparse algebraic approximation of both these matrices using a *probing* technique. Combinations of Fourier and probe approximations are also possible, but will not be considered here for simplicity, e.g., see [21].

2.3.1 Fourier Approximations

Fourier based approximations of the edge and vertex matrices are constructed based on the property that, restricted to simple curves (curves which do not intersect themselves), the Schur complement is spectrally equivalent to the square root of the Laplace operator on it, and this has been studied extensively, see [5, 39, 17, 9, 29, 19].

2.3.1.1 Fourier Edge Approximations

First, we consider Fourier approximations of the edge matrices $S_{E_{ij}}$. Let edge E_{ij} separate Ω_i and Ω_j. Since, the sub-matrix $S_{E_{ij}}$ is identical to the two sub-domain Schur complement on E_{ij}, standard preconditioners for the two sub-domain case can be applied, see [5, 39, 17, 9, 29, 19].

Let J denote the discrete Laplacian on a uniform grid containing n_{ij} interior nodes with mesh size $h = 1/(n_{ij}+1)$:

$$-h^2 \frac{d^2}{dx^2} \approx J \equiv \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

Then, $J^{1/2}$ is uniformly spectrally equivalent to $S_{E_{ij}}$ as the mesh size h is varied, see [5]. Since the discrete Laplacian is diagonalized by the sine transform, $J = W \Lambda W$,

where

$$W_{ij} = \sqrt{2h}\sin(ij\pi h),$$

and $\Lambda = diag(\lambda_i)$ with $\lambda_i = 4\sin^2(\frac{i\pi h}{2})$, it follows that $J^{1/2} = W\Lambda^{1/2}W$. By using Fast Sine Transforms, it is possible to compute the action of the inverse of $J^{1/2}$ in $O(n_{ij}\log(n_{ij}))$ flops.

The Fourier based preconditioners M considered here are all based on the sine transform W, but vary with the choice of eigenvalues:

$$M = W diag(\mu_k) W.$$

The eigenvalues μ_k are chosen to better approximate the eigenvalues of the exact Schur complement $S_{E_{ij}}$. In the model case of Laplace equation on $\Omega_i \cup \Omega_j$ with rectangular sub-domains $\Omega_i = [0,1] \times [0,l_i]$ and $\Omega_j = [0,1] \times [-l_j, 0]$, where m_i and m_j are positive integers with $l_i = (m_i+1)h$ and $l_j = (m_j+1)h$, the eigen-decomposition of the Schur complement is known exactly. These exact eigenvalues are given below in M_{Chan}, along with the eigenvalues of three other preconditioners:

(2.15)

Dryja [31] preconditioner M_D:	$\mu_k = \lambda_k^{1/2}$
Golub-Mayers [39] preconditioner M_{GM}:	$\mu_k = \sqrt{\lambda_k + \frac{1}{4}\lambda_k^2}$
BPS [8] preconditioner M_{BPS}:	$\mu_k = \sqrt{\lambda_k(1 - \frac{\lambda_k}{6})}$
Chan [17] preconditioner M_{Chan}:	$\mu_k = \left(\frac{1-\gamma_k^{m_i+1}}{1+\gamma_k^{m_i+1}} + \frac{1-\gamma_k^{m_j+1}}{1+\gamma_k^{m_j+1}}\right)\sqrt{\lambda_k + \frac{1}{4}\lambda_k^2}$ where $\gamma_k = \frac{1+\frac{\lambda_k}{2}-\sqrt{\lambda_k+\frac{\lambda_k^2}{4}}}{1+\frac{\lambda_k}{2}+\sqrt{\lambda_k+\frac{\lambda_k^2}{4}}}$.

We have the following result.

Lemma 2.1 *Let M denote either the Dryja, Golub-Mayer, BPS or Chan preconditioners for $S_{E_{ij}}$. Then*

$$\frac{\lambda_{max}(M^{-1}S_{E_{ij}})}{\lambda_{min}(M^{-1}S_{E_{ij}})} \le C_1,$$

where C_1 is independent of h. For $S_{E_{ij}}$ corresponding to Laplace equation on the model domain $\Omega_i \cup \Omega_j$ with rectangular sub-domains $\Omega_i = [0,1] \times [0,l_i]$ and $\Omega_j = [0,1] \times [-l_j, 0]$, the condition number of the Dryja, Golub-Mayer and BPS preconditioners satisfy:

$$\frac{\lambda_{max}(M^{-1}S_{E_{ij}})}{\lambda_{min}(M^{-1}S_{E_{ij}})} \le C_2(1 + \frac{1}{l_i} + \frac{1}{l_j}),$$

where C_2 is independent of h, l_i and l_j, while the condition number of the Chan preconditioner satisfies:

$$\frac{\lambda_{max}(M_{Chan}^{-1}S_{E_{ij}})}{\lambda_{min}(M_{Chan}^{-1}S_{E_{ij}})} \le C_2.$$

Proof *See Bjorstad and Widlund [5], Chan [17].*

The Fourier preconditioners described so far do not depend on the coefficients $a(x,y)$ of the elliptic problem, and thus the rate of convergence can be sensitive to the coefficients, see [22]. In order to incorporate some information about the coefficients, we scale the Fourier preconditioners by a scaling matrix. In the original BPS algorithm [8], a scalar coefficient α_{ij} representing the average of the eigenvalues of $a(x,y)$ at a point in Ω_i and a point in Ω_j was used as scaling on each edge E_{ij}. Here, we use a diagonal matrix D_{ij} as scaling, where D_{ij} denotes the diagonal of A_h restricted to E_{ij}, and define the diagonally scaled Fourier preconditioners by

$$\tilde{S}^F_{E_{ij}} \equiv D_{ij}^{1/2} W diag(\mu_k) W D_{ij}^{1/2}. \tag{2.16}$$

For most applications to isotropic coefficients, these diagonally scaled Fourier preconditioners perform well.

2.3.1.2 Fourier Vertex Space Approximation

Next, we describe approximations of the vertex space matrices S_{V_k} based on Fourier techniques. For the case of the discrete Laplacian, it is possible to express the eigen-decomposition of S_{V_k} for cross shaped vertex regions in terms of sine transforms, thereby enabling the use of fast transforms to invert S_{V_k}, see [54]. However, it is not easily generalized to the case of varying coefficients, and instead we construct approximations to the vertex matrices by using a direct sum of smaller matrices that will be described in the following.

We will describe the procedure for the model geometry of Fig. 2.3. Let u_{V_k} be a grid function on B which is zero outside the vertex region V_k, i.e., zero on $B - V_k$. Then, by the property of the Schur complement (2.11), we obtain that

$$u_{V_k}^T S_{V_k} u_{V_k} = \sum_{i=1}^{4} u_{V_k}^T S^{(i)} u_{V_k}, \tag{2.17}$$

where $S^{(i)}$ is the component of the Schur complement originating from Ω_i, as described in (2.12). For $i = 1, 2, 3, 4$, let L_i^k denote the L-shaped segment $V_k \cap \partial\Omega_i$, and further let $R_{L_i^k}$ denote the pointwise restriction onto L_i^k. Then, as in the case for the edges, $(R_{L_i^k} u_B)^T S^{(i)} (R_{L_i^k} u_B)$, is spectrally equivalent to $(R_{L_i^k} u_B)^T M_i^k (R_{L_i^k} u_B)$ where M_i^k is any of the unscaled Fourier approximations to the square root of the

Laplacian on L_i^k, see (2.15). Let D_i^k denote the diagonal of $A^{(i)}$ restricted to L_i^k. Then, by including the effects of coefficients, we define the following scaled Fourier based preconditioner for S_{V_k}:

$$\tilde{S}_{V_k}^F \equiv \sum_{i=1}^{4} R_{L_i^k}^T (D_i^k)^{1/2} M_i^k (D_i^k)^{1/2} R_{L_i^k}. \tag{2.18}$$

For most applications we considered, it was sufficient to choose the number of nodes on the vertex regions V_k to be small, say 5 or 9, and so the matrices $\tilde{S}_{V_k}^F$ can be computed at little expense, and can be inverted inexpensively by direct methods.

Theorem 2.3 *The matrices $\tilde{S}_{V_k}^F$ are spectrally equivalent to to S_{V_k}, i.e., there exists constant c_0, c_1 independent of h such that*

$$c_0 \leq \frac{\lambda_{max}\left((\tilde{S}_{V_k}^F)^{-1} S_{V_k}\right)}{\lambda_{min}\left((\tilde{S}_{V_k}^F)^{-1} S_{V_k}\right)} \leq c_1.$$

Proof *The proof follows trivially by application of the standard result, see [5, 9], that on a simple edge such as L_i^k, the square root of the Laplacian on it is spectrally equivalent to the energy of the local Schur complement, i.e., there exists constants $c_0^{(i)}, c_1^{(i)}$ independent of h such that:*

$$c_0^{(i)} \leq \frac{x_{V_k}^T S_{V_k}^{(i)} x_{V_k}}{x_{V_k}^T M_i^k x_{V_k}} \leq c_1^{(i)}.$$

Similar bounds hold when M_i^k is replaced by $(D_i^k)^{1/2} M_i^k (D_i^k)^{1/2}$, with suitably modified constants $c_0^{(i)}, c_1^{(i)}$, since the entries of D_i^k can be bounded in terms of the upper *and lower bounds for $a(x,y)$ in the neighborhood of L_i^k, independent of h. From this the result follows immediately, since:*

$$\min_i\{c_0^{(i)}\} \leq \frac{\sum_{i=1}^4 x_{V_k}^T S_{V_k}^{(i)} x_{V_k}}{\sum_{i=1}^4 x_{V_k}^T M_i^k x_{V_k}} \leq \max_i\{c_1^{(i)}\},$$

for the suitably modified coefficients $c_0^{(i)}$ and $c_1^{(i)}$.

2.3.1.3 Fourier Based Preconditioner

Based on the approximations $\tilde{S}_{E_{ij}}^F$ and $\tilde{S}_{V_k}^F$, we define the Fourier vertex space preconditioner (FVS) by

$$(2.19) \qquad M_{FVS}^{-1} \equiv R_H^T A_H^{-1} R_H + \sum_{ij} R_{E_{ij}}^T (\tilde{S}_{E_{ij}}^F)^{-1} R_{E_{ij}} + \sum_k R_{V_k}^T (\tilde{S}_{V_k}^F)^{-1} R_{V_k},$$

and the Fourier BPS preconditioner (FBPS) by, see [8]:

$$(2.20) \qquad M_{FBPS}^{-1} \equiv R_H^T A_H^{-1} R_H + \sum_{ij} R_{E_{ij}}^T (\tilde{S}_{E_{ij}}^F)^{-1} R_{E_{ij}}.$$

Note that the Fourier edge approximations $\tilde{S}_{E_{ij}}^F$ can be inverted in $O(n_{ij}\log(n_{ij}))$ flops, using the Fast Sine Transform. Direct methods can be used to solve the Fourier vertex problems $\tilde{S}_{V_k}^F$. The coarse grid matrix problem A_H can be solved using either direct or iterative methods.

Remark. In the original BPS preconditioner [8], the edge approximations were chosen to be

$$\tilde{S}_{E_{ij}} = \alpha_{ij} W diag(\mu_k) W,$$

where α_{ij} is the average of the eigenvalues of $a(x, y)$ at a point from Ω_i and a point from Ω_j, and $\mu_k = \sqrt{\lambda_k(1 - \lambda_k/6)}$, and this differs from the version described in (2.20) because of the scaling matrix D_i^k.

Theorem 2.4 *The Fourier preconditioner M_{FVS} satisfies:*

$$c_0 \leq \frac{\lambda_{max}(M_{FVS}^{-1} S)}{\lambda_{min}(M_{FVS}^{-1} S)} \leq c_1,$$

where c_0, c_1 are independent of H, h, but may depend on the overlap ratio β.

Proof *Bounds for the extreme eigenvalues of $M_{FVS}^{-1} S$ is obtained from bounds for the Rayleigh quotient:*

$$\lambda_{min}(M_{FVS}^{-1} S) \leq \left(\frac{x_B^T S x_B}{x_B^T M_{VS} x_B} \right) \left(\frac{x_B^T M_{VS} x_B}{x_B^T M_{FVS} x_B} \right) \leq \lambda_{max}(M_{FVS}^{-1} S).$$

The fraction $x_B^T S x_B / x_B^T M_{VS} x_B$ has uniform upper and lower bounds, see [66]. It therefore suffices to obtain uniform upper and lower bounds for the fraction $x_B^T M_{VS} x_B / x_B^T M_{FVS} x_B$ or equivalently for

$$\lambda_{min}(M_{FVS} M_{VS}^{-1}) \leq \frac{x_B^T M_{VS}^{-1} x_B}{x_B^T M_{FVS}^{-1} x_B} \leq \lambda_{max}(M_{FVS} M_{VS}^{-1}).$$

By spectral equivalence of the edge Fourier approximations, Lemma 2.1, there exists constants c_{ij} and C_{ij} independent of H and h such that:

$$c_{ij} \leq \frac{x_B^T S_{E_{ij}}^{-1} x_B}{x_B^T (\tilde{S}_{E_{ij}}^F)^{-1} x_B} \leq C_{ij},$$

and similarly for the vertex spaces, by Theorem 2.3, there exists constants c_k and C_k independent of H and h such that:

$$c_k \leq \frac{x_B^T S_{V_k}^{-1} x_B}{x_B^T (\tilde{S}_{V_k}^F)^{-1} x_B} \leq C_k.$$

Letting $C = \max\{C_{ij}, C_k\}$ and $c = \min\{c_{ij}, c_k\}$, we obtain that

$$c \leq \frac{\sum_{ij} x_B^T S_{E_{ij}}^{-1} x_B + \sum_k x_B^T S_{V_k}^{-1} x_B + x_B^T R_H^T A_H^{-1} R_H x_B}{\sum_{ij} x_B^T (\tilde{S}_{E_{ij}}^F)^{-1} x_B + \sum_k x_B^T (\tilde{S}_{E_k}^F)^{-1} x_B + x_B^T R_H^T A_H^{-1} R_H x_B} \leq C,$$

and hence our result follows.

2.3.2 Probe Approximations

Next, we describe another variant of the VS and BPS preconditioners in which the edge and vertex matrices are approximated by sparse matrices obtained using an extension of the *probing* technique of Chan and Resasco [25], Keyes and Gropp [45, 46], and Eisenstat [36]. Unlike Fourier based approximations, the construction of the probe approximations require solving six problems on each sub-domain, and thus has a greater overhead cost than the Fourier approximations, but still considerably less than the exact sub-matrices. An advantage of these approximations is that they often adapt well to coefficient variations and aspect ratios. However a disadvantage is that they do not adapt optimally to mesh size variations.

We will describe construction of these probe approximations for the model rectangular geometry of Fig. 2.1. The techniques are easily extended to more general geometries.

2.3.2.1 Edge Probe Approximations

We first describe how sparse approximations to the edge matrices can be constructed [25]. In its basic form, the *probing* technique consists of approximating each $S_{E_{ij}}$ by a tridiagonal matrix $\tilde{S}_{E_{ij}}$ which is chosen on the assumption that each node on an edge is strongly coupled in S only to nodes adjacent to it and weakly coupled to the other nodes. A heuristic motivation for this is that the entries of each $S_{E_{ij}}$ are known to decay rapidly away from the main diagonals:

$$|(S_{E_{ij}})_{lm}| = O\left(\frac{1}{|l-m|^2}\right),$$

see Golub and Mayer [39].

To obtain a tridiagonal approximation $\tilde{S}_{E_{ij}}$ to $S_{E_{ij}}$, we equate the matrix vector products $S_{E_{ij}}p_i$ to $\tilde{S}_{E_{ij}}p_i$ for the following three *probe* vectors p_i:

$$p_1 = [1,0,0,1,0,0,\cdots]^T, \quad p_2 = [0,1,0,0,1,0,\cdots]^T, \quad p_3 = [0,0,1,0,0,1,\cdots]^T.$$

These matrix vector products $[\tilde{S}_{E_{ij}}p_1, \tilde{S}_{E_{ij}}p_2, \tilde{S}_{E_{ij}}p_3]$ results in:

$$\begin{bmatrix} (\tilde{S}_{E_{ij}})_{11} & (\tilde{S}_{E_{ij}})_{12} & & \\ (\tilde{S}_{E_{ij}})_{21} & (\tilde{S}_{E_{ij}})_{22} & (\tilde{S}_{E_{ij}})_{23} & \\ & (\tilde{S}_{E_{ij}})_{32} & (\tilde{S}_{E_{ij}})_{33} & \ddots \\ & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} (\tilde{S}_{E_{ij}})_{11} & (\tilde{S}_{E_{ij}})_{12} & 0 \\ (\tilde{S}_{E_{ij}})_{21} & (\tilde{S}_{E_{ij}})_{22} & (\tilde{S}_{E_{ij}})_{23} \\ (\tilde{S}_{E_{ij}})_{34} & (\tilde{S}_{E_{ij}})_{32} & (\tilde{S}_{E_{ij}})_{33} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

and equating this with $[S_{E_{ij}}p_1, S_{E_{ij}}p_2, S_{E_{ij}}p_3]$ gives:

$$\begin{bmatrix} (\tilde{S}_{E_{ij}})_{11} & (\tilde{S}_{E_{ij}})_{12} & 0 \\ (\tilde{S}_{E_{ij}})_{21} & (\tilde{S}_{E_{ij}})_{22} & (\tilde{S}_{E_{ij}})_{23} \\ (\tilde{S}_{E_{ij}})_{34} & (\tilde{S}_{E_{ij}})_{32} & (\tilde{S}_{E_{ij}})_{33} \\ \vdots & \vdots & \vdots \end{bmatrix} := [S_{E_{ij}}p_1, S_{E_{ij}}p_2, S_{E_{ij}}p_3], \tag{2.21}$$

from which the non-zero entries of $\tilde{S}_{E_{ij}}$ can be easily read off. In general, $\tilde{S}_{E_{ij}}$ will not preserve the symmetry of $S_{E_{ij}}$, and so we symmetrize it to obtain $\tilde{S}^P_{E_{ij}}$ using a *minimum-modulus* procedure described below:

$$(\tilde{S}^P_{E_{ij}})_{ij} \equiv \begin{cases} (\tilde{S}_{E_{ij}})_{ji} & \text{if } |(\tilde{S}_{E_{ij}})_{ji}| \leq |(\tilde{S}_{E_{ij}})_{ij}| \\ (\tilde{S}_{E_{ij}})_{ij} & \text{if } |(\tilde{S}_{E_{ij}})_{ij}| \leq |(\tilde{S}_{E_{ij}})_{ji}|. \end{cases}$$

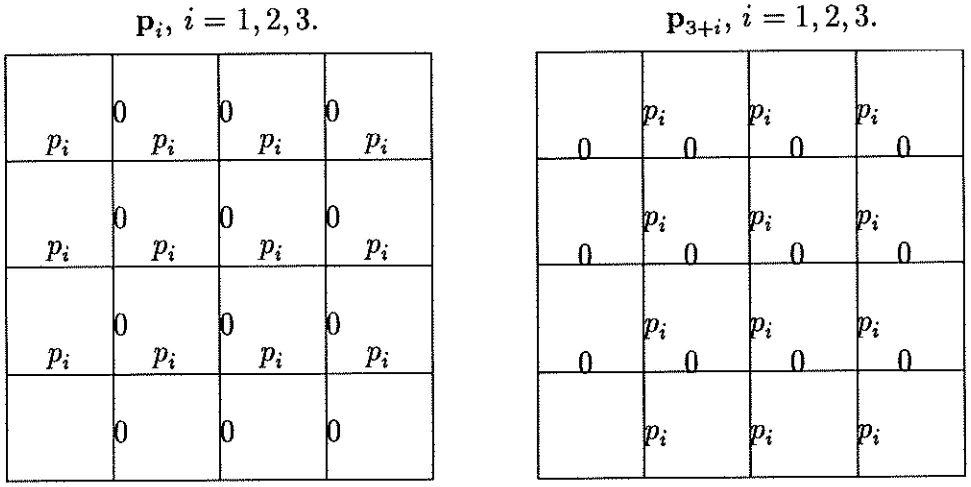
We will denote the construction of $\tilde{S}^P_{E_{ij}}$ from $S_{E_{ij}}p_1, S_{E_{ij}}p_2, S_{E_{ij}}p_3$ by the notation:

$$\tilde{S}^P_{E_{ij}} = \text{PROBE}(S_{E_{ij}}p_1, S_{E_{ij}}p_2, S_{E_{ij}}p_3). \tag{2.22}$$

The resulting approximations can be shown to preserve row-wise diagonal dominance, see [22]. This idea is motivated by Curtis, Powell, and Reid [28]. In an analogous way, using a symmetrised variant of [28], see Powell and Toint [59], it is possible to obtain a symmetric tridiagonal approximation directly using just two probe vectors, see [45, 46].

Computing the three matrix vector products $S_{E_{ij}}p_i$ requires three solves on each sub-domain Ω_i and Ω_j. Thus, in order to compute edge approximations $\tilde{S}^P_{E_{ij}}$ on the edges of all the sub-domains, twelve solves on each sub-domain would be required, since the boundary of rectangular sub-domains consists of four edges.

Figure 2.2: Simultaneous probe vectors


$\mathbf{p}_i$, $i = 1, 2, 3$.
$\mathbf{p}_{3+i}$, $i = 1, 2, 3$.
p_i
0

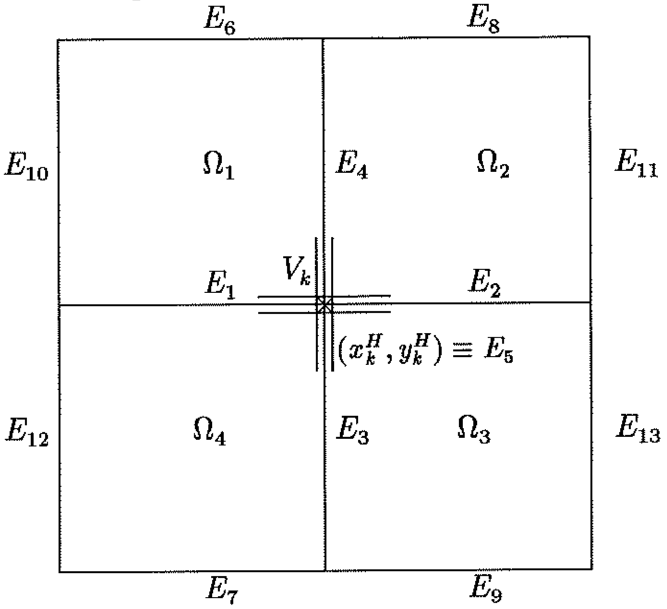
We now describe a procedure for computing all the edge approximations using only six solves on each sub-domain, by simultaneously prescribing boundary conditions on other edges, an idea first used in Keyes and Gropp [45, 46]. To minimize the approximation errors arising from the coupling between vertical and horizontal edges, we will specify probe vectors p_i either on all horizontal edges simultaneously, or on all vertical edges simultaneously. For $i = 1, 2, 3$, see Fig. 2.2, define:

$$\mathbf{p}_i \equiv \begin{cases} p_i & \text{on all horizontal edges} \\ 0 & \text{on all vertical edges ,} \end{cases}$$

$$\mathbf{p}_{3+i} \equiv \begin{cases} 0 & \text{on all horizontal edges} \\ p_i & \text{on all vertical edges.} \end{cases}$$

On the horizontal edges, the probe vectors p_i can be ordered from left to right, and on vertical edges from bottom to top. For these six probe vectors, we compute the discrete harmonic extensions $E^h\mathbf{p}_i = (-A_{II}^{-1}A_{IB}\mathbf{p}_i, \mathbf{p}_i)$, and this involves six solves

Figure 2.3: Numbering of Edges.



on each sub-domain. If E_{ij} is an horizontal edge, we define:

$$\tilde{S}^P_{E_{ij}} = \text{PROBE}(R_{E_{ij}} A_h E^h \mathbf{p}_1, R_{E_{ij}} A_h E^h \mathbf{p}_2, R_{E_{ij}} A_h E^h \mathbf{p}_3).$$

If E_{ij} is a vertical edge, then we define:

$$\tilde{S}^P_{E_{ij}} = \text{PROBE}(R_{E_{ij}} A_h E^h \mathbf{p}_4, R_{E_{ij}} A_h E^h \mathbf{p}_5, R_{E_{ij}} A_h E^h \mathbf{p}_6).$$

We have the following result on the non-singularity and diagonal dominance of the resulting probe approximations.

Theorem 2.5 *If the coefficient matrix A_h for the model rectangular geometry of Fig. 2.1 satisfies the discrete strong maximum principle (as is the case for standard five point discretizations), then the probe approximations $\tilde{S}^P_{E_{ij}}$ obtained above are strictly diagonally dominant.*

Proof *We will prove the diagonal dominance of approximation* $\tilde{S}^P_{E_1}$ *on edge* E_1 *in the model geometry of Fig. 2.3; the proof for the other edge approximations are analogous. By construction,*

$$\tilde{S}^P_{E_1} = PROBE(R_{E_1}A_hE^h\boldsymbol{p}_1, R_{E_1}A_hE^h\boldsymbol{p}_2, R_{E_1}A_hE^h\boldsymbol{p}_3).$$

Due to the effects of the boundary conditions on the adjacent edges, it is easily verified that (see § 3.2 for notation):

$$R_{E_1}A_hE^h\boldsymbol{p}_i = S_{E_1}p_i + S_{E_1E_6}p_i + S_{E_1E_7}p_i, \text{ for } i = 1, 2, 3,$$

and from this we obtain:

$$(2.23)\qquad \begin{aligned} (\tilde{S}^P_{E_1})_{i,i} &\equiv \textstyle\sum_{mod(i-j,3)=0}(S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{i,j}, \\ (\tilde{S}^P_{E_1})_{i,i-1} &\equiv \textstyle\sum_{mod(i-j,3)=1}(S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{i,j}, \\ (\tilde{S}^P_{E_1})_{i,i+1} &\equiv \textstyle\sum_{mod(i-j,3)=-1}(S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{i,j}. \end{aligned}$$

For discretizations A_h *satisfying the discrete strong maximum principle,* S *is a diagonally dominant M-matrix, see [22], and so its off diagonal entries are non-positive, and its row sums are non-negative. Using this in (2.23) we obtain that* $(\tilde{S}^P_{E_1})_{ij} \leq 0$ *for* $j \neq i$ *and the row sum:*

$$(\tilde{S}^P_{E_1})_{i,i-1} + (\tilde{S}^P_{E_1})_{i,i} + (\tilde{S}^P_{E_1})_{i,i+1} = \sum_j (S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{i,j} > 0,$$

which shows that diagonal dominance is preserved. Finally, the min-mod *procedure preserves diagonal dominance by definition.*

2.3.2.2 Probe Vertex Approximations

Next, we describe how sparse algebraic approximations to the vertex submatrices S_{V_k} can be constructed. Unlike the tridiagonal edge approximations $\tilde{S}^P_{E_{ij}}$ which enabled the use of fast direct solvers, the sparse approximations of the vertex matrices are usually small in general and will be solved by direct methods that do not make use of the sparsity of the matrices. The procedure we will describe results from a slight modification of a technique described in [21]. This new variant can be proved to result in non-singular approximations which preserve diagonal dominance.

For simplicity, we will describe this procedure for the vertex region V_k in the center of the sub-domains $\Omega_1, \cdots, \Omega_4$ of Fig. 2.3. We partition V_k into five disjoint regions:

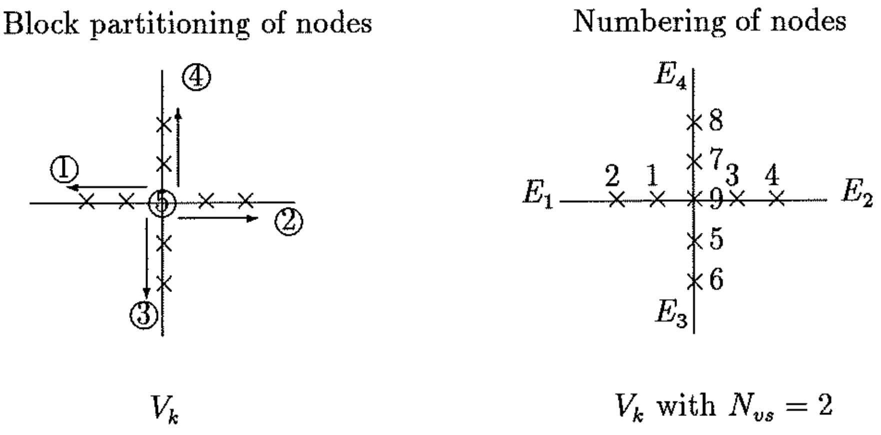
$$V_k = (V_k \cap E_1) \cap (V_k \cap E_2) \cap (V_k \cap E_3) \cap (V_k \cap E_4) \cap (x_k^H, y_k^H), \tag{2.24}$$

and we obtain a corresponding 5 × 5 block partition of the vertex matrix S_{V_k}:

$$S_{V_k} = \begin{bmatrix} S_{11} & 0 & S_{13} & S_{14} & S_{15} \\ 0 & S_{22} & S_{23} & S_{24} & S_{25} \\ S_{13}^T & S_{23}^T & S_{33} & 0 & S_{35} \\ S_{14}^T & S_{24}^T & 0 & S_{44} & S_{45} \\ S_{15}^T & S_{25}^T & S_{35}^T & S_{45}^T & S_{55} \end{bmatrix},$$

where each S_{ij} corresponds to the coupling between nodes in block i and block j. The sub-matrices S_{12} and S_{34} and their transposes are zero, since there is no

Figure 2.4: Ordering of unknowns within each vertex sub-region V_k


Block partitioning of nodes

Numbering of nodes

V_k

V_k with $N_{vs} = 2$

coupling in S between nodes in E_1 and E_2, and between nodes in E_3 and E_4. We will construct a vertex matrix approximation $\tilde{S}^P_{V_k}$ having the same block structure as S_{V_k}, with sub-blocks $\tilde{S}_{ij}$ which will be chosen to be sparse.

To facilitate description of the sparsity pattern, we will use the following ordering of nodes within V_k; for each of the four edge segments $E_i \cap V_k$, the nodes will be numbered to increase away from the cross-point (x^H_k, y^H_k), which is ordered last. This ordering is shown in Fig. 2.4 when each segment $E_i \cap V_k$ contain just two nodes.

Our choice of the sparsity pattern for the sub-blocks $\tilde{S}_{ij}$ is based on the assumption that the elements of S_{V_k} decay with increasing distance between nodes.

Definition and computation of the edge blocks $\tilde{S}_{ii}$ for $i = 1, 2, 3, 4$. Within each edge segment $E_i \cap V_k$ we assume the coupling in S_{V_k} is strong only between adjacent nodes. Based on this assumption, S_{ii} will be approximated by tridiagonal matrices $\tilde{S}_{ii}$ which are chosen to be the sub-matrices of the tridiagonal edge matrices $\tilde{S}^P_{E_i}$ for $i = 1, 2, 3, 4$, which were computed in 2.3.2.

Definition and computation of the blocks $\tilde{S}_{i5}$ for $i = 1, \cdots, 5$. We assume the cross-point (x^H_k, y^H_k) is coupled strongly in S_{V_k} only to the nodes adjacent it. Based on this assumption, we choose the vectors $\tilde{S}_{i5}$ to have zero entries except in the first entry:

$$\tilde{S}_{i5} = \begin{bmatrix} (\tilde{S}_{i5})_1 \\ 0 \\ \vdots \end{bmatrix}, \quad \text{for } i = 1, \cdots, 5.$$

For five point discretizations on the rectangular geometry of Fig. 2.1, it can easily be shown that the last row and column of S_{V_k} is exactly equal to the last row and column of $R_{V_k} A_h R^T_{V_k}$, the matrix A_h restricted to V_k. Therefore, we define

$$\tilde{S}_{i5} \equiv A_{i5} = S_{i5}, \quad i = 1, \cdots, 5$$

$$\tilde{S}_{5i} \equiv A_{5i} = S_{5i}, \quad i = 1, \cdots, 5.$$

To see that $A_{i5} = S_{i5}$, first note that S_{i5} is equal to the restriction of Su_B to the ith edge of V_k, where u_B corresponds to boundary data which is 1 on the kth vertex, and zero elsewhere. Now, recall that $Su_B = R_B A_h E^h u_B$. For five point discretizations on rectangular sub-domains, the boundary conditions on the corner nodes do not influence the solution in the interior. Consequently, the discrete harmonic extension $E^h u_B$ is zero in the interior of sub-domains, and $A_h E^h u_B$ simply gives the column of A_h corresponding to the kth vertex. Thus $S_{i5} = A_{i5}$.

Definition and computation of $\tilde{S}_{ij}$ for $i = 1, 2$ and $j = 3, 4$. We assume the couplings in S_{V_k} between edge segments $E_i \cap V_k$ and $E_j \cap V_k$ is strong only between the nodes which are closest (adjacent) to the cross-point (x_k^H, y_k^H). Based on this assumption, we choose the sub-matrices $\tilde{S}_{13}$, $\tilde{S}_{14}$, $\tilde{S}_{23}$ and $\tilde{S}_{24}$ and their transposes to have all zero entries except for the $(1,1)$-th entry.

$$\tilde{S}_{ij} = \begin{bmatrix} (\tilde{S}_{ij})_{11} & 0 & \cdots \\ 0 & 0 & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}, \quad \text{for } i = 1,2; \quad j = 3,4.$$

So there are only eight non-zero entries to define.

Consider for example the entry $(\tilde{S}_{14})_{11}$, which we would like to be an approximation to $(S_{14})_{11}$, the coupling in S between node $(x_k^H - h, y_k^H)$ and node $(x_k^H, y_k^H + h)$. Note that $(S_{14})_{11} = (S\delta_k)(x_k^H - h, y_k^H)$ (i.e. the component of $S\delta_k$ corresponding to the point $(x_k^H - h, y_k^H)$) where δ_k is the boundary data which is 1 on $(x_k^H, y_k^H + h)$ and zero elsewhere, and therefore computing $(S_{14})_{11}$ requires one sub-domain solve. In order to reduce this overhead, we would like to extract an approximation from the sub-domain solves we already used for the probe edge approximations. For example, one could define $(\tilde{S}_{14})_{11} = (S\mathbf{p}_4)(x_k^H - h, y_k^H)$. However, it turns out that this definition can lead to a non-diagonally dominant (and possibly singular) $\tilde{S}_{V_k}$. This can be seen by noting that

$$(S\mathbf{p}_4)(x_k^H - h, y_k^H) = (S_{E_1E_4}p_1 + S_{E_1E_{10}}p_1 + S_{E_1E_3}p_1 + S_{E_1E_{12}}p_1)(x_k^H - h, y_k^H).$$

The last two terms on the right corresponds to extra influence from Ω_4 on the coupling between nodes $(x_k^H - h, y_k^H)$ and $(x_k^H, y_k^H + h)$ (which should only involve couplings within Ω_1). These extra couplings could cause loss of diagonal dominance, since, in case the coefficients are large in Ω_4, the last two terms will dominate the sum on the right. In order to eliminate the influence from Ω_4, we now define

$$(\tilde{S}_{14})_{11} = (S_{E_1E_4}p_1 + S_{E_1E_{10}}p_1)(x_k^H - h, y_k^H) \quad \left(\equiv (R_{E_1}A^{(1)}E^h\mathbf{p}_4)_1\right),$$

where we recall that $A^{(1)}$ is the local stiffness matrix on Ω_1. The last equality comes from the definition of the local Schur complement, and can be extracted from the sub-domain solves used to construct the edge approximations.

Analogously, we define the seven remaining non-zero entries by:

$$\begin{aligned}
(\tilde{S}_{13})_{11} &\equiv (R_{E_1}A^{(4)}E^h\mathbf{p}_4)_1 \\
(\tilde{S}_{24})_{11} &\equiv (R_{E_2}A^{(2)}E^h\mathbf{p}_4)_1 \\
(\tilde{S}_{23})_{11} &\equiv (R_{E_2}A^{(3)}E^h\mathbf{p}_4)_1 \\
(\tilde{S}_{31})_{11} &\equiv (R_{E_3}A^{(4)}E^h\mathbf{p}_1)_1 \\
(\tilde{S}_{32})_{11} &\equiv (R_{E_3}A^{(3)}E^h\mathbf{p}_1)_1 \\
(\tilde{S}_{41})_{11} &\equiv (R_{E_4}A^{(1)}E^h\mathbf{p}_1)_1 \\
(\tilde{S}_{42})_{11} &\equiv (R_{E_4}A^{(2)}E^h\mathbf{p}_1)_1.
\end{aligned} \tag{2.25}$$

Symmetrization of $\tilde{S}_{V_k}$. Finally, in order to obtain a symmetric vertex approximation $\tilde{S}^P_{V_k}$ we use the *minimum-modulus* procedure:

$$(\tilde{S}^P_{V_k})_{ij} \equiv \begin{cases} (\tilde{S}_{V_k})_{ij} & \text{if } |(\tilde{S}_{V_k})_{ij}| \leq |(\tilde{S}_{V_k})_{ji}| \\ (\tilde{S}_{V_k})_{ji} & \text{if } |(\tilde{S}_{V_k})_{ji}| \leq |(\tilde{S}_{V_k})_{ij}|. \end{cases} \tag{2.26}$$

Theorem 2.6 *The vertex matrix approximations $\tilde{S}^P_{V_k}$ are non-singular, diagonally dominant M-matrices.*

Proof *First, we note that since the fifth block row of $\tilde{S}^P_{V_k}$, is identical to the fifth block row of S_{V_k}, it has zero row sum. For any other row of $\tilde{S}^P_{V_k}$ centered about nodes not adjacent to the cross-point, the non-zero entries are the non-zero entries of the diagonal blocks $\tilde{S}_{ii}$, for $i = 1, 2, 3, 4$. These diagonal blocks were chosen as sub-matrices of $\tilde{S}^P_{E_1}$, $\tilde{S}^P_{E_2}$, $\tilde{S}^P_{E_3}$, and $\tilde{S}^P_{E_4}$, respectively, which were shown to be diagonally dominant M-matrices in Theorem 2.5, and therefore these rows are more diagonally dominant than the corresponding rows of S.*

We now prove the diagonal dominance of the rows centered about nodes adjacent to the cross-point (x^H_k, y^H_k). Consider, for instance, the row sum corresponding to node $(x^H_k - h, y^H_k)$ to the left of the cross-point (x^H_k, y^H_k). The non-zero entries of this row are $(\tilde{S}_{11})_{11}$, $(\tilde{S}_{11})_{12}$, $(\tilde{S}_{13})_{11}$, $(\tilde{S}_{14})_{11}$, and $(\tilde{S}_{15})_{11}$. By construction:

$$\begin{aligned}
(\tilde{S}_{11})_{11} &= \textstyle\sum_{mod(j-1,3)=0}(S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{1,j} && > 0,\\
(\tilde{S}_{11})_{12} &= \textstyle\sum_{mod(j-2,3)=0}(S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{1,j} && \leq 0,\\
(\tilde{S}_{13})_{11} &= \textstyle\sum_{mod(j-1,3)=0}(S_{E_1E_3} + S_{E_1E_{10}})_{1,j} && \leq 0,\\
(\tilde{S}_{14})_{11} &= \textstyle\sum_{mod(j-1,3)=0}(S_{E_1E_4} + S_{E_1E_{13}})_{1,j} && \leq 0,\\
(\tilde{S}_{15})_{11} &= (S_{E_1E_5})_{11} && \leq 0.
\end{aligned}$$

By summing all these non-zero entries, we obtain

$$\begin{aligned}\textstyle\sum_j (\tilde{S}_{V_k})_{1j} &= \textstyle\sum_{mod(j-1,3)=0} (S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{1,j} + \\ &\textstyle\sum_{mod(j-2,3)=0} (S_{E_1} + S_{E_1E_6} + S_{E_1E_7})_{1,j} + \\ &\textstyle\sum_{mod(j-1,3)=0} (S_{E_1E_3} + S_{E_1E_{10}})_{1,j} + \\ &\textstyle\sum_{mod(j-1,3)=0} (S_{E_1E_4} + S_{E_1E_{13}})_{1,j} + \\ &(S_{E_1E_5})_{11}.\end{aligned}$$

Since the right hand side is a subset of the corresponding row of S, which is strictly diagonally dominant, this shows that this row of $\tilde{S}^P_{V_k}$ is diagonally dominant. The proof of the diagonal dominance of the other rows centered about the nodes adjacent to (x^H_k, y^H_k) is analogous. Thus $\tilde{S}^P_{V_k}$ is strictly diagonally dominant in all rows except the one corresponding to the cross-point. This last property, together with the fact that $\tilde{S}^P_{V_k}$ has positive diagonal elements and non-positive off-diagonal elements, implies that $\tilde{S}^P_{V_k}$ is a non-singular M-matrix.

2.3.2.3 Probe Based Preconditioner

We now define the Probe vertex space preconditioner (PVS) by

$$(2.27) \qquad M^{-1}_{PVS} \equiv R^T_H A^{-1}_H R_H + \sum_{ij} R^T_{E_{ij}} (\tilde{S}^P_{E_{ij}})^{-1} R_{E_{ij}} + \sum_k R^T_{V_k} (\tilde{S}^P_{V_k})^{-1} R_{V_k},$$

and the Probe BPS preconditioner (PBPS) by:

$$(2.28) \qquad M^{-1}_{PBPS} \equiv R^T_H A^{-1}_H R_H + \sum_{ij} R^T_{E_{ij}} (\tilde{S}^P_{E_{ij}})^{-1} R_{E_{ij}}.$$

2.4 Numerical Results

We now present results of numerical tests on the rate of convergence of the Fourier and Probe variants of the BPS and VS algorithms. The tests were conducted for the following elliptic problem:

$$\begin{cases} -\nabla \cdot (a(x,y)\nabla u) &= f \quad \text{in } \Omega = [0,1]^2 \\ u &= 0 \quad \text{on } \partial\Omega \end{cases}$$

for five choices of coefficients $a(x,y)$, various sub-domain sizes H, and fine grid sizes h. The five coefficients used were:

1. $a(x,y) = I$, the Laplacian, see table 2.1.

2. $a(x,y) = I + 10(x^2 + y^2)I$, slowly varying smooth coefficients, see table 2.2.

3. $a(x,y) = e^{10xy}I$, highly varying smooth coefficients, see table 2.3.

4. $a(x,y) = diag(1,\epsilon)$, anisotropic coefficients, see table 2.4.

5. Highly discontinuous coefficients of Fig. 2.5, see table 2.5.

The elliptic problem was discretized using the standard five-point difference stencil, see [68], on an $(n+1) \times (n+1)$ uniform fine grid with mesh size $h = 1/n$. The sub-domains were chosen to be the sub-rectangles of an $(n_s+1) \times (n_s+1)$ uniform coarse grid with mesh size $H = 1/n_s$. Each sub-domain, therefore consisted of $(n/n_s - 1) \times (n/n_s - 1)$ interior nodes. The coarse grid matrix A_H was chosen to be the five-point difference approximation of the elliptic problem on the coarse grid.

Figure 2.5: Discontinuous coefficients $a(x,y)$

$a = 300$	$a = 10^{-4}$	$a = 31400$	$a = 5$
$a = 0.05$	$a = 6$	$a = 0.07$	$a = 2700$
$a = 10^{6}$	$a = 0.1$	$a = 200$	$a = 9$
$a = 1$	$a = 6000$	$a = 4$	$a = 140000$

The entries of the exact solution were chosen randomly from the uniform distribution on $[-1,1]$ and the initial guess in the conjugate gradient method was chosen to be zero. The estimated condition number, κ, of the preconditioned system, and the number of iterations, $\aleph$, required to reduce the initial residual by a factor of 10^{-5} (i.e., $||r_k||_2/||r_0||_2 \leq 10^{-5}$) are listed in the tables. During each iteration, the coarse grid problem and the sub-domain problems were solved to high precision using a diagonally scaled preconditioned conjugate gradient method. The eigenvalues μ_k in the edge approximations $\tilde{S}^F_{E_{ij}}$ of (2.16) were chosen to be the Bramble, Pasciak and Schatz eigenvalues listed in (2.15), while the eigenvalues of the sub-matrices M_i^k of (2.18) were chosen to be the Dryja eigenvalues in (2.15). The Fourier and Probe BPS versions are denoted by FBPS and PBPS respectively, while the Fourier and Probe versions of the VS algorithms are denoted FVS and PVS, respectively. Unless otherwise stated, the number of nodes of overlap, N_{vs}, in the vertex regions is 1, i.e., there is one node on each vertex segment $V_k \cap E_{ij}$. The overlap ratio $\beta = h/H$ is listed as *Ovlp*.

Discussion. Tables 2.1 through 2.5 compares the performance of the various methods for the five sets of coefficients listed above. Table 2.1 corresponds to the Laplacian. In this case, the exact version of the VS algorithm, denoted by EVS, was also tested, because the eigenvalues of edge matrices $S_{E_{ij}}$ can be computed inexpensively using analytical formulas, see M_{Chan} in (2.15). In agreement with the theory, these results indicate that the Fourier variant FVS, has an observed rate of convergence independent of the mesh parameters H, h for fixed overlap ratio *Ovlp*. Moreover, the actual iteration numbers are quite insensitive to the choice of parameters H, h and *Ovlp*. For the range of sub-domain and fine grid sizes tested, the performance of PVS is very similar to EVS. However, as the number of nodes per edge increases significantly, it is expected that the PVS version would deteriorate, based on properties of the probe preconditioner for two sub-domains in [22]. The condition numbers for the variants of the BPS algorithms grow mildly with H/h, in agreement with theory. In most cases, due to clustering of eigenvalues of the preconditioned system, the number of iterations, $\aleph$, was often better than that predicted by the condition numbers.

Tables 2.2 and 2.3 correspond to smoothly varying coefficients. Here again, the results are similar to those for the Laplacian, and are in agreement with the

Table 2.1: Laplace's equation: $a(x, y) = I$

h^{-1}	Ovlp	FBPS		PBPS		EVS		FVS		PVS	
__H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_2	1/16	14.3	11	9.9	9	3.4	7	5.7	11	3.2	8
32_4	1/ 8	10.0	14	7.4	11	2.6	8	4.5	11	2.5	8
32_8	1/ 4	6.4	12	5.4	11	2.5	8	3.5	10	2.4	8
64_2	1/32	19.3	12	17.1	11	4.3	7	7.2	11	4.0	9
64_4	1/16	14.5	14	11.3	12	3.4	9	5.9	13	3.2	9
64_8	1/ 8	10.3	14	8.0	12	2.8	9	4.6	12	2.7	9
64_16	1/4	6.5	13	5.6	11	2.6	8	3.6	10	2.5	8
128_2	1/64	25.0	13	31.2	13	5.5	8	9.0	11	6.5	11
128_4	1/32	19.8	16	18.4	15	4.4	10	7.4	13	4.1	10
128_8	1/16	14.7	16	12.1	13	3.5	9	5.9	13	3.4	9
128_16	1/ 8	10.4	14	8.3	13	2.8	9	4.6	11	2.7	9
128_32	1/ 4	6.5	13	5.6	11	2.6	8	3.6	10	2.5	8
256_2	1/128	31.5	13	55.9	17	6.8	9	11.0	13	11.6	13
256_4	1/64	25.4	16	33.0	19	5.5	10	9.1	13	7.2	13
256_8	1/32	19.7	16	18.5	15	4.5	10	7.3	13	4.3	10
256_16	1/16	14.7	16	12.4	13	3.5	9	5.9	13	3.3	9
256_32	1/ 8	10.4	14	8.4	13	2.8	9	4.6	11	2.7	9
256_64	1/ 4	6.5	13	5.7	11	2.6	8	3.6	10	2.4	8

theory. Moreover, the rate of convergence of most variants are quite insensitive to the variations in the coefficients $a(x,y)$. In order to see the importance of scalings, in table 2.3 we also tested a variant nsFVS of the FVS preconditioner, in which the edge approximations were not diagonally scaled, but were instead scaled by a scalar α_{ij} on each edge E_{ij}, i.e.

$$\tilde{S}^F_{E_{ij}} \equiv \alpha_{ij} W diag(\mu_k) W,$$

where

$$\alpha_{ij} \equiv \frac{a(x_i,y_i)+a(x_j,y_j)}{2},$$

for some point $(x_i,y_i) \in \Omega_i$ and $(x_j,y_j) \in \Omega_j$. As the results indicate, this variant was sensitive to the variations in the coefficients.

Table 2.4 concerns the case of anisotropic coefficients. Here, the results are qualitatively different from the preceding cases. Note that the rate of convergence of all variants of the VS and BPS algorithms deteriorate to a fixed rate as $\epsilon \rightarrow 0$. The limiting condition numbers seem to depend on the coarse mesh size, as $1/H$. A possible explanation for this deterioration is the following. For $\epsilon = 0$, the unknowns are essentially coupled only along the x axis and adjacent vertical edges are coupled strongly in the Schur complement. This coupling is not represented in the VS preconditioner, and may cause the deterioration in the convergence rate. The results in table 2.4 also indicate that the probe versions perform slightly better than the Fourier versions. This can be explained as follows. For $\epsilon = 0$, the edge matrices $S_{E_{ij}}$ on the horizontal edges become a discrete approximation of

Table 2.2: Mildly varying coefficients: $a(x,y) = (1 + 10(x^2 + y^2))\, I$

h^{-1}	Ovlp	FBPS		PBPS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_2	1/16	15.2	11	10.6	9	6.0	11	3.4	8
32_4	1/ 8	10.2	14	7.6	11	4.6	11	2.6	8
32_8	1/ 4	6.4	12	5.4	11	3.6	10	2.4	8
64_2	1/32	20.4	12	17.8	11	7.5	11	4.2	9
64_4	1/16	14.9	14	11.6	12	5.8	12	3.2	9
64_8	1/ 8	10.3	14	8.1	12	4.6	11	2.7	9
64_16	1/ 4	6.5	13	5.6	11	3.6	10	2.4	8
128_2	1/64	26.3	13	32.1	13	9.4	11	6.7	11
128_4	1/32	20.0	16	18.4	15	7.3	13	4.2	10
128_8	1/16	14.7	16	12.2	13	5.9	13	3.4	9
128_16	1/ 8	10.4	14	8.4	13	4.6	11	2.7	8
128_32	1/ 4	6.5	13	5.6	11	3.6	10	2.4	8
256_2	1/128	32.9	13	57.0	16	11.5	13	11.7	13
256_4	1/64	25.8	17	33.2	19	9.3	13	7.2	13
256_8	1/32	19.9	16	18.6	15	7.3	13	4.3	10
256_16	1/16	14.7	16	12.3	13	5.9	13	3.4	9
256_32	1/ 8	10.4	14	8.4	13	4.6	11	2.7	9
256_64	1/ 4	6.5	13	5.7	11	3.6	10	2.4	8

Table 2.3: Highly varying coefficients: $a(x,y) = e^{10xy}I$

h^{-1}	Ovlp	FBPS		PBPS		nsFVS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_2	1/16	22.5	11	18.4	9	16.1	18	7.5	11	4.4	9
32_4	1/ 8	13.4	15	11.0	13	7.2	13	5.1	11	3.2	9
32_8	1/ 4	7.0	12	6.2	11	4.0	10	3.9	10	2.5	8
64_2	1/32	28.9	12	25.9	11	24.5	23	9.5	11	5.8	9
64_4	1/16	17.6	16	15.5	15	11.3	16	6.5	12	4.0	9
64_8	1/ 8	11.0	12	9.1	12	5.6	12	4.9	11	2.8	8
64_16	1/ 4	6.6	12	5.8	11	3.7	10	3.7	10	2.5	8
128_2	1/64	36.3	13	45.0	14	35.8	28	11.8	12	8.6	11
128_4	1/32	24.4	16	23.3	15	16.1	19	8.4	13	5.1	10
128_8	1/16	15.7	14	13.2	13	7.7	14	6.0	12	3.6	10
128_16	1/ 8	10.4	14	8.4	11	4.7	12	4.6	11	2.8	9
128_32	1/ 4	6.5	12	5.7	11	3.6	10	3.6	10	2.4	8
256_2	1/128	44.2	14	77.2	17	32.0	24	14.4	13	15.1	14
256_4	1/64	29.3	17	41.4	22	16.2	19	10.1	13	8.5	13
256_8	1/32	20.8	16	20.2	15	8.0	14	7.7	13	4.4	10
256_16	1/16	15.0	15	12.4	13	5.0	11	6.1	13	3.3	9
256_32	1/ 8	10.3	14	8.2	12	3.8	10	4.7	12	2.7	8
256_64	1/ 4	6.5	12	5.6	11	2.9	9	3.6	10	2.4	8

$-d^2/dx^2$, while on vertical edges $S_{E_{ij}}$ becomes a nearly diagonal matrix, similar to the identity. The FVS edge matrices $\tilde{S}^P_{E_{ij}}$ approximate the square root of the Laplacian, and are therefore invalid in this case. By construction, the tridiagonal probing technique approximates diagonal and tridiagonal matrices well, and consequently, they perform better than the Fourier versions we tested. The algorithms for anisotropic problems need further study.

Table 2.5 refers to the case of the highly discontinuous coefficients of Fig. 2.5. The performance is similar to the case of smooth coefficients, and the results indicate that the rate of convergence of all variants is quite insensitive to the jumps in the coefficients.

In tables 2.6 and 2.7, we compare various preconditioners for different choices of eigenvalues μ_k in the Fourier approximations (2.16). Here, CFBPS denotes that the eigenvalues of the Fourier edge approximations in the FBPS preconditioner were those of M_{Chan} in (2.15), while CFVS denotes that the same eigenvalues were used in the FVS preconditioner. In agreement with theory, the Fourier versions were spectrally equivalent with respect to variations in H and h, for fixed overlap *Ovlp*. Amongst the various eigenvalues tested, the exact eigenvalues of the Schur complement of the Laplacian used in CFBPS and CFVS gave the best results. Corresponding rates for the probe version are also listed for comparison.

Finally, in tables 2.8, 2.9, 2.10, and 2.11, we present a comparison of the FVS and PVS preconditioners, as the amount of overlap N_{VS} in the vertex regions is increased. Here, $N_{VS} = 0$ indicates that only the vertex node was used, i.e., the

Table 2.4: Anisotropic problem: $\frac{\partial^2 u}{\partial x^2} + \epsilon \frac{\partial^2 u}{\partial y^2} = f$

	$h = 1/64, H = 1/2$				$h = 1/64, H = 1/4$				$h = 1/64, H = 1/16$			
ϵ	PVS		FVS		PVS		FVS		PVS		FVS	
	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
0.1	7.4	10	14.5	17	5.9	12	12.0	18	9.0	16	12.9	19
0.08	8.0	10	16.1	17	6.4	12	13.5	20	10.8	17	15.5	20
0.06	8.9	10	18.4	19	7.5	13	15.2	21	13.6	18	19.7	23
0.04	10.3	10	22.4	21	9.7	14	20.9	24	19.2	22	27.9	26
0.02	13.0	10	31.4	24	13.0	16	29.4	28	34.4	27	50.1	33
0.01	16.3	10	43.6	27	20.7	18	41.7	31	58.4	34	84.8	41
10^{-3}	29.3	8	115.6	38	60.3	25	151.5	47	215.8	59	351.8	73
10^{-4}	39.4	7	179.8	46	81.7	25	250.7	57	352.5	69	591.4	92
10^{-5}	41.8	6	193.8	48	105.1	27	253.6	59	396.6	73	583.6	87
10^{-6}	42.0	6	195.3	49	105.0	26	267.7	59	355.8	71	647.5	93
10^{-7}	42.1	6	195.4	48	102.1	25	273.1	59	405.6	73	654.0	92
10^{-8}	42.1	6	195.2	48	106.2	23	254.9	57	395.6	72	661.7	93

Table 2.5: Discontinuous coefficients: See $a(x, y)$ of Fig. 5.

h^{-1}	Ovlp	FBPS		PBPS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_4	1/ 8	10.2	13	7.5	11	6.1	12	8.1	11
32_8	1/ 4	6.6	12	5.2	10	8.5	13	3.7	9
64_4	1/16	14.7	15	11.1	11	9.3	14	10.1	11
64_8	1/ 8	10.1	14	8.1	12	8.4	14	5.2	10
64_16	1/ 4	6.5	13	5.6	11	6.9	12	4.1	9
128_4	1/32	19.6	17	18.1	16	12.3	14	6.8	11
128_8	1/16	14.4	16	12.1	14	11.5	15	5.9	11
128_16	1/ 8	10.2	14	8.3	13	6.4	13	3.4	9
128_32	1/ 4	6.6	13	5.7	11	6.8	12	4.1	9
256_4	1/64	25.4	19	33.0	17	14.9	15	7.8	13
256_8	1/32	19.3	17	18.7	16	8.8	15	4.9	11
256_16	1/16	14.8	16	12.3	13	12.4	16	6.9	11
256_32	1/ 8	10.3	14	8.4	13	8.6	14	6.0	10
256_64	1/ 4	6.5	13	5.7	11	6.0	12	4.1	9

Table 2.6: Different Edge Fourier Preconditioners for Laplace Equation

h^{-1}	Ovlp	FBPS		CFBPS		FVS		CFVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_2	1/16	14.3	11	9.5	7	5.7	11	4.6	8	3.2	8
32_4	1/ 8	10.0	14	7.3	11	4.5	11	3.6	9	2.5	8
32_8	1/ 4	6.4	12	5.3	11	3.5	10	2.9	9	2.4	8
64_2	1/32	19.3	12	13.4	7	7.2	11	5.8	8	4.0	9
64_4	1/16	14.5	14	10.7	11	5.9	13	4.7	10	3.2	9
64_8	1/ 8	10.3	14	8.1	12	4.6	12	3.7	10	2.7	9
64_16	1/ 4	6.5	13	5.5	11	3.6	10	2.9	9	2.5	8
128_2	1/64	25.0	13	17.8	8	9.0	11	7.3	8	6.5	11
128_4	1/32	19.8	16	14.6	12	7.4	13	5.8	10	4.1	10
128_8	1/16	14.7	16	11.5	14	5.9	13	4.7	10	3.4	9
128_16	1/ 8	10.4	14	8.3	13	4.6	11	3.7	10	2.7	9
128_32	1/ 4	6.5	13	5.5	11	3.6	10	2.9	9	2.5	8
256_2	1/128	31.5	13	23.0	7	11.0	12	8.9	9	11.6	13
256_4	1/64	25.4	16	19.2	13	9.0	14	7.3	10	7.2	13
256_8	1/32	19.7	16	15.6	13	7.2	13	5.9	11	4.3	10
256_16	1/16	14.7	16	11.7	14	5.9	13	4.7	10	3.3	9
256_32	1/ 8	10.4	14	8.4	13	4.6	11	3.8	10	2.7	9
256_64	1/ 4	6.5	13	5.5	11	3.6	10	2.9	9	2.5	8

Table 2.7: Different Edge Fourier Preconditioners for $a(x,y) = e^{10xy}I$

h^{-1}	Ovlp	FBPS		CFBPS		FVS		CFVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_2	1/16	22.5	11	18.1	8	7.5	11	6.2	9	4.4	9
32_4	1/ 8	13.4	15	10.7	13	5.1	11	4.5	10	3.2	9
32_8	1/ 4	7.0	12	5.8	11	3.9	10	3.3	9	2.5	8
64_2	1/32	28.9	12	23.0	9	9.5	11	7.7	9	5.8	9
64_4	1/16	17.6	16	14.7	12	6.5	12	5.4	9	4.0	9
64_8	1/ 8	11.0	12	8.8	11	4.9	11	4.0	10	2.8	8
64_16	1/ 4	6.6	12	5.6	11	3.7	10	3.0	9	2.5	8
128_2	1/64	36.3	13	28.5	9	11.8	12	9.6	9	8.6	11
128_4	1/32	24.4	16	19.4	13	8.4	13	7.0	9	5.1	10
128_8	1/16	15.7	14	12.5	11	6.0	12	5.1	10	3.6	10
128_16	1/ 8	10.4	14	8.5	12	4.6	11	3.8	9	2.8	9
128_32	1/ 4	6.5	12	5.5	11	3.6	10	3.0	9	2.4	8
256_2	1/128	44.2	14	34.7	9	14.4	13	11.6	9	15.1	14
256_4	1/64	29.3	17	23.3	14	10.1	13	8.3	10	8.5	13
256_8	1/32	20.8	16	16.5	13	7.7	13	6.2	10	4.4	10
256_16	1/16	15.0	15	11.9	12	6.1	13	4.8	10	3.3	9
256_32	1/ 8	10.3	14	8.3	12	4.7	12	3.8	10	2.7	8
256_64	1/ 4	6.5	12	5.4	11	3.6	10	2.9	9	2.4	8

Table 2.8: Variation of vertex sizes for $H = 1/2$, $h = 1/128$, and $a(x, y) = I$.

N_{vs}	0	1	2	3	4	5	6	7
κ_{FVS}	7.45	8.97	8.07	7.66	6.85	6.98	6.71	6.53
$\aleph$	10	11	12	12	12	13	12	12

Table 2.9: Variation of vertex sizes for $H = 1/2$, $h = 1/128$, and $a(x, y) = e^{10xy} I$.

N_{VS}	0	1	2	3	4	5	6	7
κ_{FVS}	9.85	11.80	10.25	10.00	9.41	9.01	8.63	8.40
$\aleph$	11	12	12	13	12	12	12	13

Table 2.10: Variation of vertex sizes for $H = 1/2$, $h = 1/128$, $a(x, y) = I$.

N_{VS}	0	1	2	3	4	5	6	7
κ_{PVS}	8.3	6.6	5.6	5.0	4.8	3.2	4.6	4.5
$\aleph$	11	11	11	11	11	9	11	11

Table 2.11: Variation of vertex sizes for $H = 1/2$, $h = 1/128$ and $a(x, y) = e^{10xy} I$.

N_{VS}	0	1	2	3	4	5	6	7
κ_{PVS}	10.8	9.1	7.3	6.6	6.6	6.6	6.8	6.9
$\aleph$	12	11	10	10	10	11	11	11

vertex matrices were 1×1. We note that the improvement in condition number of the VS algorithms as the overlap *Ovlp* is increased is mild, as also noted in [66]. In particular, the performance is quite satisfactory even when the vertex region consists of just one point, see Widlund [72].

Conclusions: Both the Fourier and Probe variants of the vertex space algorithm are designed to be efficient alternatives to the original VS algorithm. Our experiments for a wide range of coefficients and grid sizes show that the efficiency does not come at a price of deteriorated performance. We hope that these variants will provide flexible and efficient methods for solving second order elliptic problems using the domain decomposition approach.

CHAPTER 3

Complexity of DD Algorithms

In most domain decomposition (DD) methods, a coarse grid solve is employed to provide the global coupling required to produce an *optimal* method. The total cost of a method can depend sensitively on the choice of the coarse grid size H. In this chapter, we give a simple analysis of this phenomenon for a model elliptic problem and a variant of Smith's vertex space domain decomposition method [66, 23]. We derive the optimal value H_{opt} which asymptotically minimizes the total cost of method (number of floating point operations in the sequential case and execution time in the parallel case), for subdomain solvers with different complexities. Using the value of H_{opt}, we derive the overall complexity of the DD method, which can be significantly lower than that of the subdomain solver.

3.1 Introduction

The focus of our chapter is on the choice of the *size* H of the coarse grid. It is intuitively obvious that the total cost of a DD method can depend sensitively on this choice, in addition to the obvious dependence on the efficiency of the subdomain solver. A small H generally improves the convergence rate (because the coarse grid problem is a better approximation to the original fine grid problem) at the cost of a more costly coarse grid solve, whereas a large H has the opposite effect. Therefore, an optimal value H_{opt} often exists and indeed has been observed empirically [42, 64]. Surprisingly, there has been almost no systematic study in the literature on this issue. Our approach is to take a simple model elliptic problem and a particular DD method; this allows a simple but complete and easily understood analysis which we think give insights for more general situations.

For concreteness, we focus our analysis on a variant of Smith's vertex space method [66] developed by us earlier [23]. We consider subdomain solvers with different complexity, including banded Gaussian elimination, nested dissection, modified incomplete Cholesky factorization (MIC) and multigrid solvers. For simplicity, we assume the same solver is used for the subdomains and the coarse grid problem, and that these are solved exactly. By expressing the computational complexity as a function of the coarse grid size H and the fine grid size h, we derive the optimal value H_{opt} which *asymptotically* (as h tends to 0) minimizes the total cost of method. Using the value of H_{opt}, we can derive the overall complexity of the DD method as a function of h alone, which can be significantly lower than that of the subdomain solver. That is, through the use of DD, a given solver can be made more efficient for solving the original problem, *by using it to solve smaller (but more) sub-problems.* This is a simple consequence of the divide-and-conquer principle. The assumption of the asymptotic limit is not necessary but does allow a close form expression for H_{opt} from which one can see more clearly the general trend.

3.2 Optimal Computational Complexity: Sequential Case

We now make the assumption that the cost of the FVS method is dominated by that of solving the subdomain problems (in inverting A_{II} in computing the matrix-vector product Su in PCG) and the coarse grid problem (in inverting A_H in the preconditioner). This is a reasonable assumption if h is small enough and H is neither too small or too large, so that the subdomains have a reasonably large "area-to-perimeter" ratio and the coarse grid is not too small.

Let the complexity of the solver used for both the sub-domain problems and the coarse problem be $O(m^p)$ for the preprocessing phase (e.g. factorization) and $O(m^s)$ for the solution phase on an $m \times m$ grid. For example, for banded Gaussian elimination, MIC and multigrid, $p = 4, 2, 2$ and $s = 3, 2.5, 2$ respectively. For nested dissection, $p = 3$ and the solution phase has complexity $O(m^2 \log m)$. We assume that the iteration number $\aleph$ is bounded and *independent* of the fine and coarse grid sizes, which is supported by the theoretical and numerical results in [66, 23]. For example, numerical experiments in [23] indicate that $\aleph$ is between 9 and 15 for a tolerance of 10^{-5} for a wide range of values for h and H and for widely different coefficients of the elliptic problem. It is then easy to see that the leading order terms of the operation count for the FVS/PCG method are given by:

$$\text{flops}(H) \approx \frac{C}{H^2}(\frac{H}{h})^p + \frac{C}{H^p} + \aleph\{\frac{C}{H^2}(\frac{H}{h})^s + \frac{C}{H^s}\}, \tag{3.1}$$

where C is a generic constant that depends on the particular solver. The first two terms are the preprocessing cost (e.g. factorization of A_{II} and A_H) and the last two terms are the cost during the PCG iteration. The leading order terms have the form:

$$\text{flops}(H) \approx \frac{C}{H^2}(\frac{H}{h})^\alpha + \frac{C}{H^\alpha}, \tag{3.2}$$

where $\alpha = \max\{p, s\}$ and the generic constant C may depend on $\aleph$ but is independent of H and h. In other words, the dominant cost consists of solving $1/H^2$ sub-domain problems and one coarse grid problem.

The optimal coarse grid size H_{opt} is obtained by setting the first derivative of function in (3.2) (with respect to H) to zero, giving:

$$H_{opt}(\alpha) = (\frac{\alpha}{\alpha - 2})^{\frac{1}{2\alpha-2}} h^{\frac{\alpha}{2\alpha-2}} \qquad \text{for } \alpha > 2. \tag{3.3}$$

Using this value of H_{opt}, we obtain for the asymptotic complexity:

$$\min_H \text{flops}(H) \approx \text{flops}(H_{opt}) \approx C\{(\frac{\alpha}{\alpha - 2})^{\frac{\alpha-2}{2\alpha-2}} + (\frac{\alpha}{\alpha - 2})^{\frac{-\alpha}{2\alpha-2}}\}h^{\frac{-\alpha^2}{2\alpha-2}}. \tag{3.4}$$

When $\alpha = 2$, i.e. an optimal solver such as a multigrid method,

$$\text{flops}(H) \approx \frac{C}{h^2} + \frac{C}{H^2}, \tag{3.5}$$

which indicates that H should be chosen as large as possible ($O(1)$ in our model problem.)

Table 3.1: The sequential complexity of solvers on an $n \times n$ grid; coarse grid size n_H.

Basic Solver	Complexity	Optimal n_H	Complexity of DD Solver using optimal n_H
Multigrid	$O(n^2)$	1	$O(n^2)$
MIC	$O(n^{2.5})$	$0.58n^{5/6}$	$O(n^{2.08})$
Nested Dissection	$O(n^3)$	$0.76n^{3/4}$	$O(n^{2.25})$
Band-Cholesky	$O(n^4)$	$0.89n^{2/3}$	$O(n^{2.67})$
	$O(n^\alpha), \alpha \rightarrow \infty$	$n^{1/2}$	$O(n^{\alpha/2})$

Note that H_{opt} is independent of the constant C (i.e. the solver). Clearly, H_{opt} depends non-monotonically on the complexity exponent α. For $\alpha = 2.5, 3, 4$, $H_{opt} = 5^{1/3}h^{5/6}, 3^{1/4}h^{3/4}, 2^{1/6}h^{2/3}$ respectively. As $\alpha \rightarrow \infty$, $H_{opt} \rightarrow h^{1/2}$.

The complexity of the FVS algorithm, using H_{opt}, is given by:

$$\text{flops}(H_{opt}) \approx O((\frac{1}{h})^{\gamma(\alpha)}),$$

where $\gamma(\alpha) = \frac{\alpha^2}{2\alpha-2}$. For $\alpha = 2, 2.5, 3, 4$, $\gamma = 2, 2.08, 2.25, 2.67$ respectively. As $\alpha \rightarrow \infty$, $\gamma(\alpha) \rightarrow \alpha/2$. Thus, using a domain decomposition approach results in a substantial reduction in the asymptotic complexity of the solver. The reduction is greater the higher the complexity of the solver is.

We summarize these complexity results in Table 1, where we present the results in terms of an $n \times n$ fine grid ($n = 1/h$) and an $n_H \times n_H$ coarse grid ($n_H = 1/H$).

3.3 Optimal Computational Complexity: Parallel Case

In the parallel case, the operation count model has to be replaced by a true timing model, taking into account both the arithmetic cost and the communication cost. However, in the spirit of the asymptotic analysis used in the last section, we can make some simplifying assumptions which allow us to extract useful information from our model. The most important assumption we shall make is that the communication cost is not dominant over the arithmetic cost, which is valid if the number of unknowns in the interior of a subdomain is not too small compared to those on the boundary (i.e. a small perimeter-to-area ratio) and is consistent with our assumption in Sec. 3. The full treatment with communication cost can be found in [26].

We shall also assume that there are enough processors so that the subdomain problems are solved completely in parallel. A crucial issue is how to solve the coarse grid problem in a parallel environment. According to Gropp [41], one of the best methods is to collect the necessary data on one processor, solve it there and then broadcast the result. Finally, we can do the coarse grid solve either (a) sequentially, after the subdomain solves, or (b) in parallel to the subdomain solves.

Making these assumptions, it is easy to see that the leading order terms of the parallel time of the FVS method is:

$$\text{time}(H) \approx \begin{cases} C(H/h)^\alpha + C(1/H^\alpha) & \text{case (a)} \\ \max\{C(H/h)^\alpha, C(1/H^\alpha)\} & \text{case (b),} \end{cases}$$

where C is a generic constant modeling the time per arithmetic operation. In both cases, the optimal value of H can be easily seen to be:

$$H_{opt} = \sqrt{h},$$

independent of α (i.e. the solver). We note that this optimal choice of H_{opt} implies that *the size of each subdomain problem is equal to the size of the coarse problem.* It also implies that the optimal number of processors is n $(= (1/\sqrt{h})^2)$.

The parallel time of the FVS method using H_{opt} is:

$$\text{time}(H_{opt}) = O(n^{\alpha/2}),$$

and the speed-up is:

$$\text{Speed-up} = O(n^\alpha)/O(n^{\alpha/2}) = O(n^{\alpha/2}).$$

Note that the speed up is greater than $O(n)$ (the number of processors) if $\alpha > 2$. This "superlinear" speed-up is possible because we are not parallelizing a "fixed" algorithm — the FVS algorithm with the optimal coarse grid has different sequential complexity for different n.

Table 3.2: The sequential complexity of solvers on an $n \times n \times n$ grid; coarse grid size n_H.

Basic Solver	Complexity	Optimal n_H	Complexity of DD Solver using optimal n_H
Multigrid	$O(n^3)$	1	$O(n^3)$
MIC	$O(n^{3.5})$	$0.61 n^{7/8}$	$O(n^{3.06})$
Nested Dissection	$O(n^6)$	$0.93 n^{2/3}$	$O(n^4)$
Band-Cholesky	$O(n^7)$	$0.95 n^{7/11}$	$O(n^{4.45})$
	$O(n^\alpha), \alpha \rightarrow \infty$	$n^{1/2}$	$O(n^{\alpha/2})$

3.4 Higher Dimensional Problems

A similar analysis can also be extended to a d-dimensional problem. For a solver of complexity $O(m^\alpha)$ on an m^d grid, the results in the sequential case are:

$$H_{opt} = (\frac{\alpha}{\alpha - d})^{\frac{1}{\alpha - d}} h^{\frac{\alpha}{2\alpha - d}}, \qquad \text{flops}(H_{opt}) = O(h^{-\frac{\alpha^2}{2\alpha - d}}).$$

The $d = 3$ case is summarized in Table 2. The results in Sec. 4 for the parallel case are independent of d, except that the optimal number of processors is $n^{d/2}$.

3.5 Concluding Remarks

The results obtained above should also apply to other *optimal* domain decomposition methods, such as other substructuring methods and overlapping Schwarz methods. The optimal coarse grid size is obtained as a simple balance between the cost of the subdomain solves and the cost of the coarse grid solve. Therefore, the conclusions are valid for any DD method, as long as these costs dominate the overall cost and the convergence rate is independent of H and h.

CHAPTER 4

DD Methods for Coupled Elliptic Systems

In this chapter, we discuss the application of domain decomposition method to certain coupled elliptic systems, arising frequently in the modeling of physics processes. For example, the steady-state drift-diffusion equations in semiconductor modeling have coupling in the lower order derivative terms. We are interested in analysis the relation between the convergence rate of DD method and the coupling parameters. Our purpose, here, is to test whether DD methods are robust as the coupling parameter varying. We prove that the convergence rates of additive and multiplicative Schwarz methods are independent of not only the mesh parameters but also the coupling coefficients when the size H of coarse grid is small enough. Furthermore, we propose several sparse approximations of interface Schur complement dense matrices. Many domain decomposition methods, such as vertex space domain decomposition method and substructuring, require solving the Schur complement systems on the interfaces. It is extremely expensive when calculating the exact Schur complement matrix and its inverse because the matrix is dense. In order to reduce overhead cost, we therefore focus on constructing approximations which are inexpensive to construct and invert. Several approximations of these dense matrices are constructed by using *Fourier* approximations and *probing* technique.

A general framework of domain decomposition method for symmetric and positive definite problems has recently been developed by Bramble, Pasciak, Wang and Xu [11, 74], and Dryja and Widlund [32, 33, 34, 35, 73]. Later, this general framework has been extended to certain non-symmetric and indefinite problems, see Cai and Widlund's paper [12, 13], which prove that the convergence rates of additive and multiplicative Schwarz methods are independent of mesh parameters. For the symmetric positive case, the Fourier and probe approximations of Schur complement has been studied extensively in [5, 39, 17, 9, 29, 19] and [25, 45, 46, 36, 21] respectively.

In 4.1, we describe the nonsymmetric and indefinite problems and define two kinds of operators T_j. Then we show the convergence by using the Cai and Widlund framework [12, 13], described in 4.2. In 4.3, we study several approximations of Schur complement on the interface by using a Fourier approximation and a probing technique.

4.1 Coupled Elliptic Systems and Two-Level Schwarz Methods

In this section, we,first, describe general coupled elliptic system. Then, we discuss the applications of two-level Schwarz domain decomposition methods.

4.1.1 Coupled Elliptic Systems

Consider the Dirichlet problem

$$\mathcal{L}\mathbf{u} = \mathbf{f} \quad \text{in } \Omega \qquad \mathbf{u} = 0 \quad \text{on } \partial\Omega, \tag{4.1}$$

where

$$\mathcal{L}\mathbf{u} = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \mathcal{A}_{i,j}(x) \frac{\partial \mathbf{u}(x)}{\partial x_j} + \sum_{i=1}^{d} \mathcal{B}_i(x) \frac{\partial \mathbf{u}(x)}{\partial x_j} + \mathcal{C}(x)\mathbf{u},$$

with $k \times k$ square block matrices $\mathcal{A}_{i,j}(x), \mathcal{B}_i(x)$ and $\mathcal{C}(x)$, and vector function $\mathbf{u}$. We call $\mathcal{B}_i(x)$ and $\mathcal{C}(x)$ as the lower order coupling coefficients. We notice that the steady-state drift-diffusion equations don't have higher order coupling terms. Let's define the bilinear forms:

$$a(\mathbf{u}, \mathbf{v}) = \sum_{i,j=1}^{d} \int_\Omega \frac{\partial \mathbf{v}^T}{\partial x_i} \mathcal{A}_{i,j} \frac{\partial \mathbf{u}}{\partial x_j} dx,$$

and

$$s(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{d} \int_\Omega \mathbf{u}^T \mathcal{B}_i(x) \frac{\partial \mathbf{v}(x)}{\partial x_i} + \frac{\partial \mathbf{u}^T \mathcal{B}_i(x)}{\partial x_i} \mathbf{v},$$

which correspond to the seconder-order terms and the skew-symmetric part of equation (4.1), respectively. We denote $\mathbf{H}_0^1(\Omega) = (H_0^1(\Omega))^k$. Then, the weak form of equation (4.1) is: Find $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ such that

$$b(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{4.2}$$

where

$$b(\mathbf{u}, \mathbf{v}) = a(\mathbf{u}, \mathbf{v}) + s(\mathbf{u}, \mathbf{v}) + \int_\Omega \tilde{\mathcal{C}}(x)\mathbf{u}\mathbf{v}dx.$$

Here, $\tilde{\mathcal{C}}(x) = \mathcal{C} - \sum_{i=1}^{d} \partial \mathcal{B}_i / \partial x_i$.

We make the following basic boundedness assumptions.

i) There exist constants $c_1 > 0$ and $c_2 > 0$ such that

$$c_1 ||\mathbf{v}||^2_{\mathbf{H}^1_0(\Omega)} \leq a(\mathbf{v}, \mathbf{v}) \leq c_2 ||\mathbf{v}||^2_{\mathbf{H}^1_0(\Omega)} \qquad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega)$$

ii) Bilinear $s(\mathbf{u}, \mathbf{v})$ is continuous.

$$\begin{aligned} |s(\mathbf{u}, \mathbf{v})| \leq & \; c_3 ||\mathbf{u}||_{\mathbf{H}^1_0(\Omega)} ||\mathbf{v}||_{\mathbf{L}^2(\Omega)} \qquad \forall \mathbf{u}, \mathbf{v} \in \mathbf{H}^1_0(\Omega) \\ \leq & \; c_3 ||\mathbf{u}||_{\mathbf{L}^2(\Omega)} ||\mathbf{v}||_{\mathbf{H}^1_0(\Omega)} \end{aligned}$$

iii) There exists a constant c_4 such that

$$| \int_\Omega \tilde{C}(x) \mathbf{u} \mathbf{v} dx | \leq c_4 ||\mathbf{u}||_{\mathbf{L}^2(\Omega)} ||\mathbf{v}||_{\mathbf{L}^2(\Omega)} \qquad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega).$$

Note the constants c_3 and c_4 depend on the lower order coupling parameters. We will focus on the analysis of relation between the convergence rate and these coupling parameters. As an easy consequence of these assumptions, the following bounds and regularity for the bilinear form $b(\cdot, \cdot)$ can be established.

1) $b(\mathbf{u}, \mathbf{v})$ is continuous. There is a constant $c = c_2 + |\Omega|(c_3 + c_4)$:

$$|b(\mathbf{u}, \mathbf{v})| \leq c ||\mathbf{u}||_{\mathbf{H}^1_0(\Omega)} ||\mathbf{v}||_{\mathbf{H}^1_0(\Omega)} \qquad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega)$$

2) Gårding's inequality holds

$$|b(\mathbf{u}, \mathbf{u})| \geq c_1 ||\mathbf{u}||^2_{\mathbf{H}^1_0(\Omega)} - c_3 ||\mathbf{u}||_{\mathbf{H}^1_0(\Omega)} ||\mathbf{u}||_{\mathbf{L}^2(\Omega)} - c_4 ||\mathbf{u}||^2_{\mathbf{L}^2(\Omega)} \qquad \forall \mathbf{u} \in \mathbf{H}^1_0(\Omega).$$

3) The solution $\mathbf{w}$ of the dual equation

$$b(\mathbf{v}, \mathbf{w}) = (\mathbf{g}, \mathbf{v}), \qquad \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega)$$

satisfies the regularity,

$$||\mathbf{w}||_{\mathbf{H}^{1+\sigma}(\Omega)} \leq C_5||\mathbf{g}||_{\mathbf{L}^2(\Omega)} \qquad \sigma \in [\frac{1}{2}, 1].$$

Let's introduce two triangulations on Ω with elements Ω_i and $\tau_{i,j}$. Then we have two level meshes Ω_H and Ω_h. Two level discrete spaces can be defined by

$$\mathbf{V}^H = \{\mathbf{v}_H | \mathbf{v}_H \in C^0(\Omega), \quad \mathbf{v}_H|_{\Omega_i} \text{ linear, } \mathbf{v}_H = 0 \text{ on } \partial\Omega\}$$

$$\mathbf{V}^h = \{\mathbf{v}_h | \mathbf{v}_h \in C^0(\Omega), \quad \mathbf{v}_h|_{\tau_{i,j}} \text{ linear, } \mathbf{v}_h = 0 \text{ on } \partial\Omega\}.$$

The corresponding discrete problem is: Find $\mathbf{u}_h \in \mathbf{V}^h$, such that

$$b(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}^h. \tag{4.3}$$

4.1.2 Two-Level Schwarz Methods

We extend each substructure Ω_i into a larger region Ω_i^{ext}, whose boundary $\partial\Omega_i^{ext}$ does not cut through any h-level elements. These sub-regions are colored by using minimum colors $1, 2, \cdots, J$ in such a way that no neighbour sub-regions have the same color. Then, all sub-regions of the same color are merged together and denoted as $\Omega'_1, \cdots, \Omega'_J$. Let $\mathbf{V}_i^h = \mathbf{H}_0^1(\Omega'_j) \cap \mathbf{V}^h$, and $\mathbf{V}_0^h = \mathbf{V}^H$. Then the space $\mathbf{V}^h$ can be written as the sum of these subspaces:

$$\mathbf{V}^h = \mathbf{V}_0^h + \mathbf{V}_1^h + \cdots + \mathbf{V}_J^h.$$

Now we construct operator $T_j : \mathbf{V}_j^h \to \mathbf{V}^h$. A natural selection is the projection operators P_j defined as

$$\text{(4.4)} \qquad \text{Find } P_j\mathbf{u}_h \in \mathbf{V}_j^h, \quad b(P_j\mathbf{u}_h, \mathbf{v}_h) = b(\mathbf{u}_h, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_j^h$$

It is often more economical to use approximate projectors rather than exact solvers of the problem on subspaces. We introduce approximate and continuous bilinear forms $b_j(\cdot,\cdot)$ defined on $\mathbf{V}_j^h \times \mathbf{V}_j^h$. We assume that there exists a constant $\omega_b > 0$ such that

$$\text{(4.5)} \qquad a(\mathbf{u}_h, \mathbf{u}_h) \leq \omega_b b_j(\mathbf{u}_h, \mathbf{u}_h) \qquad \forall \mathbf{u}_h, \mathbf{v}_h \in \mathbf{V}_j^h.$$

For example, we can choose $a(\cdot,\cdot)$ as the bilinear $b_j(\cdot,\cdot)$. Then this $b_j(\cdot,\cdot)$ satisfies above assumptions with $\omega_b = 1$. Then, an operator $T_j : \mathbf{V}^h \to \mathbf{V}_j^h$, which approximates P_j, is defined by

$$\text{(4.6)} \qquad b_j(T_j\mathbf{u}_h, \mathbf{v}_h) = b(\mathbf{u}_h, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_j^h.$$

For the theoretical result, we usually take $T_0 = P_0$. We note that $P_j\mathbf{u}_h$ and $T_j\mathbf{u}_h$ can be calculated, without explicit knowledge of solution $\mathbf{u}_h$, by solving a problem in the subspace $\mathbf{V}_j^h$ as follows

$$b(P_j\mathbf{u}_h, \mathbf{v}_h) = b(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_j^h$$

and

$$b_j(T_j\mathbf{u}_h, \mathbf{v}_h) = b(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_j^h,$$

respectively. Then, we can use additive or multiplicative Schwarz method to solve equation (4.3). We follow Cai-Widlund framework to give the relation between convergence rate and the coupling parameters.

Using ii) and Friedrichs' inequality (1.1), we have

$$\begin{aligned} |s(\mathbf{u}_h, P_j\mathbf{u}_h)| &\le C(c_3)||\mathbf{u}_h||_a||P_j\mathbf{u}_h||_{\mathbf{L}^2(\Omega_j)} \qquad \forall \mathbf{u}_h \in \mathbf{V}_j^h \\ &\le CH||\mathbf{u}_h||_a||P_j\mathbf{u}_h||_a \\ &\le C(c_3)H(a(\mathbf{u}_h,\mathbf{u}_h) + a(P_j\mathbf{u}_h, P_j\mathbf{u}_h)) \qquad \text{for } j = 1, \cdots, J. \end{aligned}$$

A direct consequence of these inequalities is

$$|s(\mathbf{u}_h - P_j\mathbf{u}_h, P_j\mathbf{u}_h)| \le C(c_3)H(a(\mathbf{u}_h,\mathbf{u}_h) + a(P_j\mathbf{u}_h, P_j\mathbf{u}_h)) \qquad \text{for } j = 1, \cdots, J.$$

From Schatz's result [61], it follows directly that there exist constants $H_0 > 0$ and $C(H_0)$ such that

$$||\mathbf{u}_h - P_0\mathbf{u}_h||_{\mathbf{L}^2} \le C(H_0, c_3, c_4)H^\sigma||\mathbf{u}_h||_a, \tag{4.7}$$

and

$$||P_0\mathbf{u}_h||_a \le C(H_0)||\mathbf{u}_h||_a, \tag{4.8}$$

when $H \le H_0$. We note that the constant H_0 depends on the coupling coefficient. However, $C(H_0)$ is independent of coupling coefficients. Then, applying Cauchy-Schwarz inequalities and above inequalities gives

$$|s(\mathbf{u}_h - P_0\mathbf{u}_h, P_0\mathbf{u}_h)| \le C(c_3, c_4)H^\sigma(a(\mathbf{u}_h,\mathbf{u}_h) + a(P_0\mathbf{u}_h, P_0\mathbf{u}_h)).$$

From the definition of P_j and above three inequalities, it easily follows that

$$a(P_j\mathbf{u}_h, \mathbf{u}_h) = (1 - C(c_3, c_4)H)a(P_j\mathbf{u}_h, P_j\mathbf{u}_h) - C(c_3, c_4)Ha(\mathbf{u}_h, \mathbf{u}_h),$$

for $j = 1, \cdots, J$, and

$$a(P_0\mathbf{u}_h, \mathbf{u}_h) = (1 - CH^\sigma)a(P_0\mathbf{u}_h, P_0\mathbf{u}_h) - CH^\sigma a(\mathbf{u}_h, \mathbf{u}_h) \qquad \forall \mathbf{u}_h \in \mathbf{V}^h.$$

These two inequalities show that

$$a(\mathbf{u}_h, P_j\mathbf{u}_h) + a(P_j\mathbf{u}_h, (I - P_j)\mathbf{u}_h) \geq \gamma a(P_j\mathbf{u}_h, P_j\mathbf{u}_h) - \delta_j a(\mathbf{u}_h, \mathbf{u}_h),$$

with $\gamma = 1 - 2C(c_3, c_5)H$ and $\delta_j = 2C(c_3, c_5)H$. For general operator T_j, the corresponding results follow from a minor modification of above arguments. Since we assume that $\gamma = 1 - 2C(c_3, c_5)H > 0, \quad \gamma$ is independent of coupling parameters. So it is obvious that only constant $\delta_j = 2C(c_3, c_5)H$ ' depends on coupling coefficients and mesh parameters.

By using the triangle inequality, inequalities (4.7) and (4.8), we obtain

$$||\mathbf{u}_h||^2_{\mathbf{L}^2} \leq C(c_4)(H^{2\sigma}a(\mathbf{u}_h, \mathbf{u}_h) + ||P_0\mathbf{u}_h||^2_{\mathbf{L}^2})$$

and

$$(1 - CH^{2\sigma})a(\mathbf{u}_h, \mathbf{u}_h) \leq b(\mathbf{u}_h, \mathbf{u}_h) + C||P_0\mathbf{u}_h||_a||\mathbf{u}_h||_a.$$

From the definition of the operator P_j and Lemma 1.4, it follows easily that

$$b(\mathbf{u}_h, \mathbf{u}_h) = \sum_{j=0}^{J} b(\mathbf{u}_h, \mathbf{u}_{h,j}) = \sum_{j=0}^{J} b(P_j\mathbf{u}_h, \mathbf{u}_{h,j}).$$

Applying the continuity of $b(\cdot, \cdot)$ and the Cauchy-Schwarz inequality results

$$b(\mathbf{u}_h, \mathbf{u}_h) \leq CC_0(c_3H, c_4H)(\sum_{j=0}^{J} ||P_j\mathbf{u}_h||^2_a)^{1/2}||\mathbf{u}_h||_a.$$

Thus, from this inequality we obtain:

$$a(\mathbf{u}_h, \mathbf{u}_h) \leq CC_0^2 \sum_{j=0}^{J} a(P_j \mathbf{u}_h, P_j \mathbf{u}_h),$$

where CC_0^2 is independent of coupling coefficients when H is small enough.

4.2 Cai and Widlund Framework for Coupled Elliptic Problems

As in Chapter 1, we first represent the discrete space V^h as the sum of subspaces $\{V_j^h\}_{j=0}^J$, and define operator $T_j : V^h \rightarrow V_j^h$. Then the operator of additive Schwarz method can be defined by

$$T = T_0 + T_1 + T_2 + \cdots + T_N,$$

and the error propagation operator of multiplicative Schwarz can be derived as

$$E_J = (I - T_J) \cdots (I - T_1)(I - T_0).$$

Therefore, we need to analyze the condition number of T and the norm of E_J. We assume that the operators T_j satisfy the following two assumptions, which are the extension of those for symmetric positive definite problems.

Assumption 4.1 *There exist a constant $\gamma > 0$ parameters $\delta_j \geq 0$, such that*

$$(4.9) \qquad a(\boldsymbol{u}_h, T_j \boldsymbol{u}_h) + a(T_j \boldsymbol{u}_h, (I - T_j)\boldsymbol{u}_h) \geq \gamma a(T_j \boldsymbol{u}_h, T_j \boldsymbol{u}_h) - \delta_j a(\boldsymbol{u}_h, \boldsymbol{u}_h),$$

and $\sum_{j=0}^J \delta_j$ can be chosen small enough, where $\boldsymbol{u}_h \in V^h$ is a vector and $a(\cdot,\cdot)$ is a symmetric positive definite bilinear operator defined in Chapter 1.

Remarks: We notice that, for the projector P_j, the parameters $\delta_j = C(c_3, c_4)H$ depend on not only the size of coarse grid but also the *coupling coefficients.* As in 1.2, we can obtain the inequalities similar to those in the symmetric case. Denote $\omega = \frac{2}{1+\gamma}$. Then, inequality (4.9) can be rewritten as

$$a(\mathbf{u}_h, T_j\mathbf{u}_h) \geq \omega^{-1} a(T_j\mathbf{u}_h, T_j\mathbf{u}_h) - \frac{\delta_j}{2} a(\mathbf{u}_h, \mathbf{u}_h).$$

It follows easily from this inequality that

$$(4.10) \qquad ||T_j||_a \leq \omega + \frac{\delta_j}{2}, \qquad \text{and} \qquad ||I - T_j||_a \leq 1 + \frac{\delta_j}{2}.$$

The following upper bounds can be obtained straightforwardly by using the strengthened Cauchy -Schwarz inequalities:

$$||\sum_{j=1}^{J} T_j||_a \leq \omega\rho(\mathcal{E}) + \sum_{j=1}^{J} \frac{\delta_j}{2},$$

and

$$\sum_{j=1}^{J} a(T_j\mathbf{u}_h, T_j\mathbf{u}_h) \leq (\omega\rho(\mathcal{E})^{1/2} + \sum_{j=1}^{J} \frac{\delta_j}{2\rho(\mathcal{E})^{1/2}})^2 a(\mathbf{u}_h, \mathbf{u}_h).$$

From these inequalities and (4.10), it is easy to show the following upper bounds:

$$(4.11) \qquad ||\sum_{j=0}^{J} T_j||_a \leq \omega(\rho(\mathcal{E}) + 1) + \sum_{j=0}^{J} \frac{\delta_j}{2},$$

and

$$(4.12) \qquad \sum_{j=0}^{J} ||T_j\mathbf{u}_h||_a^2 \leq (\omega(\rho(\mathcal{E})^{1/2} + 1) + \sum_{j=1}^{J} \frac{\delta_j}{2\rho(\mathcal{E})^{1/2}} + \delta_0/2)^2 ||\mathbf{u}_h||_a^2.$$

Assumption 4.2 *There exists a constant $C_0 > 0$, such that*

$$(4.13) \qquad \sum_{j=0}^{J} a(T_j \boldsymbol{u}_h, T_j \boldsymbol{u}_h) \geq C_0^{-2} a(\boldsymbol{u}_h, \boldsymbol{u}_h).$$

The lower bound for $a(\sum_{j=0}^{J} T_j \mathbf{u}_h, u_h)$ follows directly from Assumption 4.1 and 4.2, and inequalities (4.10):

$$(4.14) \qquad \sum_{j=0}^{J} a(T_j \mathbf{u}_h, \mathbf{u}_h) \geq (C_0^{-2} - \sum_{j=0}^{J} \frac{\delta_j}{2}) a(\mathbf{u}_h, \mathbf{u}_h).$$

The upper bound (4.11) and the lower bound (4.14) give the convergence rate of GMRES with additive Schwarz preconditioner according to the paper [37].

By using definition and Cauchy-Schwarz inequalities, we can easily show that, there is a constant $c > 0$ independent of mesh parameters and coupling coefficient, such that

$$(4.15) \qquad \sum_{j=0}^{J} ||T_j E_{j-1} \mathbf{u}_h||_a^2 \geq c(\omega^2 \rho(\mathcal{E})^2 + (\sum_{j=0}^{J} \delta_j)^2 + 1)^{-1} \sum_{j=0}^{J} ||T_j \mathbf{u}_h||_a^2.$$

It follows directly from Assumption 4.1 that

$$2a(T_j E_{j-1} \mathbf{u}_h, E_{j-1} \mathbf{u}_h) - ||T_j E_{j-1} \mathbf{u}_h||_a^2 \geq \gamma ||T_j E_{j-1} \mathbf{u}_h||_a^2 - \delta_j ||E_{j-1} \mathbf{u}_h||_a^2.$$

The first term on the right can be bounded from below by inequality (4.15). The second term is bounded by $\delta_j \exp(\sum_{j=0}^{J-1} \delta_j)$. Therefore, we have the following theorem.

Theorem 4.1 (Cai and Widlund [12, 13]) *If*

$$\frac{\max_{0 \leq j \leq J} \gamma_j C_0^{-2}}{\omega^2 \rho(\mathcal{E})^2 + (\sum_j \delta_j)^2 + 1}$$

dominates

$$\delta_j \exp(\sum_{j=0}^{J-1} \delta_j)$$

by a sufficiently large constant factor, then the multiplicative Schwarz method converges and

$$||E_J||_a \leq \sqrt{1 - \frac{cC_0^{-2}}{\omega^2 \rho(\mathcal{E})^2 + (\sum_j \delta_j)^2 + 1}}. \tag{4.16}$$

Remarks: The operators T_j, given in 4.1.2, satisfy Assumption 4.1 and 4.2 with $\delta_j = C(c_3, c_4)H$, $\gamma_j = 1 - C(c_3, c_4)H$ and $C_0^{-2} = c/(1 - C(c_3, c_4)H)$. After making these observation, we can easily obtain the following conclusion.

Theorem 4.2 *Assume that the operators T_j are defined in section 4.1.2. Then, the convergence rates of additive and multiplicative Schwarz methods are independent of the lower order coupling parameters when the coarse grid size is fine enough..*

Remarks: By choosing coarse grid size fine enough, we can make DD methods have the fast convergence rates as well as make the DD methods be robust as the coupling parameter varying.

4.3 Interface Schur Complement Approximations

In this section, we consider applying the non-overlapping domain decomposition to the coupled elliptic systems which result from linearization of semiconductor device simulation problems [4, 69, 3, 30]. The main idea is to decompose a domain into many smaller regular sub-domains, reduce the problem on the whole domain to the Schur complement system on the interface and then solve the interface system. Since the Schur complement matrix is expensive to evaluate and to calculate its inverse directly, the reduced Schur complement system is usually solved by iterative methods, such as GMRES or BiCG conjugate gradient type methods. Generally, the Schur complement matrix is not well-conditioned so that a direct application of the iterative method to the system will not be a very efficient algorithm. Therefore, a good preconditioner is required when constructing an efficient algorithm. Our main purpose, here, is to derive interface preconditioners for the coupled elliptic system. We notice that the variation of coupling term may lead the coupled elliptic problem to become indefinite and unsymmetric. Hence, our main interest is in finding the preconditioner that slightly depends on or does not depend on the coupling coefficients. Our efforts on this are based on two approaches. One is Fourier approximation. The other is probe technique.

4.3.1 Abstract Fourier Bound Deduced from Sobolev Inequalities

Let $\tilde{\Omega} = \Omega_i \cup \Omega_j$, where Ω_i and Ω_j have common interface $\tilde{\Gamma}$. Assume that the diameter H of these domains is small enough. The Sobolev space of order one half on $\tilde{\Gamma}$ will be denoted $\mathbf{H}_{00}^{1/2}(\tilde{\Gamma})$. Its corresponding discrete space is written as $\mathbf{V}^h(\tilde{\Gamma})$. For any $\phi_h \in \mathbf{V}^h(\tilde{\Gamma})$, the discrete $b-$harmonic extension on Ω_i is defined by operator $t : V^h(\tilde{\Gamma}) \to V^h(\Omega_i)$

$$\mathbf{u}_h = t\phi_h \in V^h(\Omega_i), \qquad \mathbf{u}_h = \phi_h \text{ on } \tilde{\Gamma} \qquad \mathbf{u}_h = 0 \text{ on } \partial\Omega_i - \tilde{\Gamma}$$

$$\text{and} \qquad b(\mathbf{u}_h, \mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}^{h,0}(\Omega_i),$$

where $\mathbf{V}^{h,0} = \mathbf{H}_0^1(\Omega_i) \cap \mathbf{V}^h$. Then, we introduce a bilinear operator $\beta(\cdot,\cdot)$ on $\mathbf{V}^h(\tilde{\Gamma}) \times \mathbf{V}^h(\tilde{\Gamma})$, defined by

$$\beta(\phi_h, \psi_h) = b(t\phi_h, t\psi_h).$$

Lemma 4.1 *There exists a $H_0 > 0$. such that, if $H \le H_0$, then*

$$c|\phi_h|^2_{\boldsymbol{H}_{00}^{1/2}(\tilde{\Gamma})} \le \beta(\phi_h, \phi_h) \le C|\phi_h|^2_{\boldsymbol{H}_{00}^{1/2}(\tilde{\Gamma})}, \tag{4.17}$$

for any $\phi_h \in \boldsymbol{V}^h(\tilde{\Gamma})$. Here, constants c and C are independent of coupling coefficients.

Proof *Let $\boldsymbol{w} \in \boldsymbol{H}_0^1(\Omega)$ satisfy $w = \phi_h$ on $\tilde{\Gamma}$, $\boldsymbol{w} = 0$ on $\partial\Omega_i \setminus \tilde{\Gamma}$ and*

$$b(\boldsymbol{w}, \boldsymbol{v}) = 0 \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega_i),$$

i.e. $\boldsymbol{w}$ is the $b-$harmonic extension of ϕ_h. We assume that $\boldsymbol{w}_h$ is the discrete $b-$harmonic extension of ϕ_h. Using a well known a priori inequality, we have

$$a(\boldsymbol{w}, \boldsymbol{w}) \le C|\phi_h|^2_{\boldsymbol{H}_{00}^{1/2}(\tilde{\Gamma})}.$$

Then, from the triangle inequality it follows easily that

$$a(\boldsymbol{w}_h, \boldsymbol{w}_h) \le a(\boldsymbol{w} - \boldsymbol{w}_h, \boldsymbol{w} - \boldsymbol{w}_h) + a(\boldsymbol{w}, \boldsymbol{w}).$$

By well-known approximation properties of finite elements, we have that,

$$a(\boldsymbol{w}-\boldsymbol{w}_h,\boldsymbol{w}-\boldsymbol{w}_h)\le Ch^{2\sigma}||\boldsymbol{w}||^2_{\boldsymbol{H}^{1+\sigma}(\Omega_i)}. \qquad \textit{for } 0<\sigma<1/2,$$

Now using a well known a priori inequality (cf. [47]) and an "inverse property", we see that

$$h^{2\sigma}||\boldsymbol{w}||^2_{\boldsymbol{H}^{1+\sigma}(\Omega_i)}\le Ch^{2\sigma}|\phi_h|^2_{\boldsymbol{H}^{1/2+\sigma}(\tilde{\Gamma})}\le C|\phi_h|^2_{\boldsymbol{H}^{1/2}(\tilde{\Gamma})}.$$

Hence, by the continuity of b, the definition of $\boldsymbol{w}_h$ and above inequalities, we can obtain that

$$\beta(\phi_h,\phi_h)=b(\boldsymbol{w}_h,\boldsymbol{w}_h)\le Ca(\boldsymbol{w}_h,\boldsymbol{w}_h)\le C|\phi_h|^2_{H^{1/2}(\tilde{\Gamma})}$$

which shows the upper bound of inequalities (4.17).

To show the lower bound, we apply the trace inequality and Poincaré inequality and obtain that

$$c|\phi_h|^2_{\boldsymbol{H}^{1/2}_{00}(\tilde{\Gamma})}\le a(\boldsymbol{w}_h,\boldsymbol{w}_h).$$

Thus, the lower bound can be easily obtained

$$c|\phi_h|^2_{\boldsymbol{H}^{1/2}_{00}(\tilde{\Gamma})}\le a(\boldsymbol{w}_h,\boldsymbol{w}_h)\le Cb(\boldsymbol{w}_h,\boldsymbol{w}_h)=C\beta(\phi_h,\phi_h),$$

by using Gårding's inequality with small enough H.

Lemma 4.2 *Let l_0 be an operator defined by*

$$<l_0\phi,\psi>=<a\phi',\psi'> \qquad \forall\psi\in\boldsymbol{H}^{1/2}_{00}(\tilde{\Gamma}),$$

where $< \phi, \psi >= \int_{\tilde{\Gamma}} \phi\psi ds$. *There exists a constant* $H_0 > 0$. *Then, the Schur complement is spectrally equivalent to the square root of the Laplace operator on it, when* $H \leq H_0$. *That is,*

$$c < l_0^{1/2}\phi_h, \phi_h > \leq \beta(\phi_h, \phi_h) \leq C < l_0^{1/2}\phi_h, \phi_h > \quad \forall \phi_h \in \boldsymbol{H}_{00}^{1/2}(\tilde{\Gamma}). \tag{4.18}$$

Proof *In [8], it was already shown that*

$$c < l_0^{1/2}\phi_h, \phi_h > \leq |\phi_h|^2_{\boldsymbol{H}_{00}^{1/2}(\tilde{\Gamma})} \leq C < l_0^{1/2}\phi_h, \phi_h > .$$

A direct consequence of Lemma 4.1 shows inequalities (4.18).

Remark: From Theorem 4.1, 4.2, Lemma 4.1 and 4.2, we can conclude that vertex space domain decomposition method with Fourier edge and vertex space approximate matrices has a convergence rate independent of mesh parameters as well as coupling parameters, when coarse grid size H is small enough.

4.3.2 Model Coupled Problem and Its Capacitance Interface Matrix

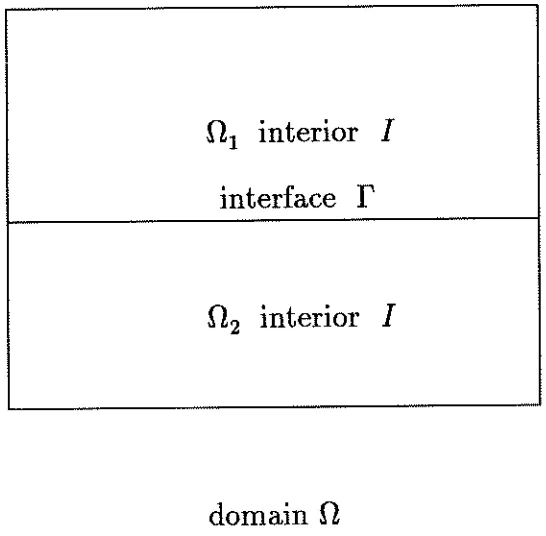
For simplicity, we consider the following linear model coupled elliptic system:

$$\begin{cases} -\Delta u + \bar{a}v = f & \text{in } \Omega \\ \bar{b}u - \Delta v = g & \text{in } \Omega \end{cases} \tag{4.19}$$

with boundary condition

$$u = 0 \quad \text{on } \partial\Omega \qquad v = 0 \quad \text{on } \partial\Omega,$$

Figure 4.1: Domain Ω and Sub-domains Ω_i



where $\bar{a}$ and $\bar{b}$ are coupling parameters. We assume that the coupling parameters $\bar{a}$ and $\bar{b}$ are constant. Since we are only interested in deriving interface preconditioner, we divide the domain Ω, illustrated in Fig. 4.3.2, into two sub-domains Ω_1 and Ω_2 with common interface Γ. Assume a uniform mesh with size h is used on Ω and with n internal grid points in the x-direction, i.e.,

$$h = \frac{1}{n+1}.$$

Suppose that there are m_k internal grid points in Ω_k in y-direction, for $k = 1, 2$, i.e.,

$$l_1 = (m_1 + 1)h \qquad l_2 = (m_2 + 1)h.$$

Denote the mesh point as (x_i, y_j) and $u_{i,j}$ as the approximation to $u(x_i, y_j)$. Let's define $(u_I^{(k)}, v_I^{(k)})$ as the set of interior unknowns in Ω_k for $k = 1, 2$ and (u_B, v_B) as the set of unknowns on the common interfaces $\Gamma = \partial\Omega_1 \cap \Omega_2$. Let

$a = h^2\bar{a}$ and $b = h^2\bar{b}$. By using five point stencil, we discretize the model coupled elliptic system and obtain following linear system in a block form:

(4.20)

$$\begin{pmatrix} A_{II}^{(k)} & a \\ b & A_{II}^{(k)} \end{pmatrix} \begin{pmatrix} u_I^{(k)} \\ v_I^{(k)} \end{pmatrix} + \begin{pmatrix} A_{IB}^{(k)} & 0 \\ 0 & A_{IB}^{(k)} \end{pmatrix} \begin{pmatrix} u_B \\ v_B \end{pmatrix} = \begin{pmatrix} f_I^{(k)} \\ g_I^{(k)} \end{pmatrix} \qquad k = 1, 2$$

$$\begin{pmatrix} A_{BI}^{(1)} & 0 \\ 0 & A_{BI}^{(1)} \end{pmatrix} \begin{pmatrix} u_I^{(1)} \\ v_I^{(1)} \end{pmatrix} + \begin{pmatrix} A_{BI}^{(2)} & 0 \\ 0 & A_{BI}^{(2)} \end{pmatrix} \begin{pmatrix} u_I^{(2)} \\ v_I^{(2)} \end{pmatrix}$$

$$+ \begin{pmatrix} A_{BB} & a \\ b & A_{BB} \end{pmatrix} \begin{pmatrix} u_B \\ v_B \end{pmatrix} = \begin{pmatrix} f_B \\ g_B \end{pmatrix}.$$

Applying block Gauss Elimination to system (4.20), we obtain capacitance interface system only for the common interface unknowns $(u_B, v_B)^T$ as follows:

(4.21)

$$C \begin{pmatrix} u_B \\ v_B \end{pmatrix} = \begin{pmatrix} f_B \\ g_B \end{pmatrix} - \begin{pmatrix} A_{BI}^{(1)} & 0 \\ 0 & A_{BI}^{(1)} \end{pmatrix} \begin{pmatrix} A_{II}^{(1)} & a \\ b & A_{II}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} f_I^{(1)} \\ g_I^{(1)} \end{pmatrix}$$

$$- \begin{pmatrix} A_{BI}^{(2)} & 0 \\ 0 & A_{BI}^{(2)} \end{pmatrix} \begin{pmatrix} A_{II}^{(2)} & a \\ b & A_{II}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} f_I^{(2)} \\ g_I^{(2)} \end{pmatrix},$$

where the capacitance matrix C is defined by

(4.22)

$$C = \begin{pmatrix} A_{BB} & a \\ b & A_{BB} \end{pmatrix} - \begin{pmatrix} A_{BI}^{(1)} & 0 \\ 0 & A_{BI}^{(1)} \end{pmatrix} \begin{pmatrix} A_{II}^{(1)} & a \\ b & A_{II}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} A_{IB}^{(1)} & 0 \\ 0 & A_{IB}^{(1)} \end{pmatrix}$$

$$- \begin{pmatrix} A_{BI}^{(2)} & 0 \\ 0 & A_{BI}^{(2)} \end{pmatrix} \begin{pmatrix} A_{II}^{(2)} & a \\ b & A_{II}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} A_{IB}^{(2)} & 0 \\ 0 & A_{IB}^{(2)} \end{pmatrix}.$$

Note that computing the exact capacitance matrix C needs solving $4n$ the coupled elliptic subsystems on sub-domains Ω_1 and Ω_2. Hence, it is very expensive to evaluate and store the exact capacitance matrix C and its inverse since it is dense. To solve this Schur complement system efficiently, we prefer to use preconditioned iterative method, such as Bi-CG method and GMRES,etc., whose performing procedure only requires the matrix vector product

$$C \begin{pmatrix} u_B \\ v_B \end{pmatrix}.$$

From (4.22), it is easy to see that each evaluation of above matrix vector product needs to solve two coupled elliptic subproblems on the sub-domains Ω_1 and Ω_2 respectively.

4.3.3 Fourier Analysis on Model Problem

In order to propose good preconditioners, we first analyze the eigen-structure of the capacitance matrix C. In a similar way as in [17, 24, 19], we are able to derive the eigen-decomposition of C by using Fourier analysis. Then, according to the distribution of these eigenvalues, we discuss preconditioners which are similar to those proposed by Dryja [31], Golub and Mayer [39], and Chan [17].

4.3.3.1 The Exact Eigen-Decomposition of C

We introduce some notations here:

$$\tau_k = 4\sin^2\frac{k\pi h}{2}, \qquad \text{for} \quad k = 1, \cdots, n$$

$$\mu_k = \frac{2+\tau_k+\sqrt{ab}-\sqrt{(2+\tau_k+\sqrt{ab})^2-4}}{2+\tau_k+\sqrt{ab}+\sqrt{(2+\tau_k+\sqrt{ab})^2-4}} \qquad \text{for} \quad k = 1, \cdots, n,$$

$$\mu_{-k} = \frac{2+\tau_k-\sqrt{ab}-\sqrt{(2+\tau_k-\sqrt{ab})^2-4}}{2+\tau_k-\sqrt{ab}+\sqrt{(2+\tau_k-\sqrt{ab})^2-4}} \qquad \text{for} \quad k = 1, \cdots, n,$$

and

$$\cos\theta_k = \frac{2+\tau_k-\sqrt{ab}}{2} \quad \sin\theta_k = \sqrt{\sqrt{ab}-\tau_k-(\frac{\sqrt{ab}-\tau_k}{2})^2} \quad \text{if } |2+\tau_k-\sqrt{ab}| < 2.$$

Note if $ab \geq 0$, then $\mu_k < 1$ for $k = 1, \cdots, n$. When $ab < 0$, $\sqrt{ab}$ and μ_k, μ_{-k} are complex number.

Let us define Fourier vectors: for $k = 1, \cdots, n$,

$$w^{(k)} = \sqrt{h} \begin{pmatrix} \sin k\pi h \\ \vdots \\ \sin k\pi n h \end{pmatrix}, \qquad F^{(k)} = \begin{pmatrix} w^{(k)} \\ w^{(k)} \end{pmatrix} \qquad F^{(-k)} = \begin{pmatrix} w^{(k)} \\ -w^{(k)} \end{pmatrix}.$$

It is easy to testify that

$$||F^{(k)}||_2 = 1 \quad \text{and} \quad ||F^{(-k)}||_2 = 1.$$

We denote $W = [w^{(1)}, \cdots, w^{(n)}]$ and

$$F = [F^{(1)}, \cdots, F^{(n)}, F^{(-1)}, \cdots, F^{(-n)}] = \begin{pmatrix} W & W \\ W & -W \end{pmatrix}.$$

Note that vectors F^k and F^{-k} are normal and orthogonal, i.e.

$$F^T F = I.$$

Then following theorem can be obtained.

Theorem 4.3 *Define*

$$D = \begin{pmatrix} I & 0 \\ 0 & \sqrt{\frac{b}{a}} I \end{pmatrix} \qquad \textit{and} \qquad \tilde{F} = DF.$$

Then the capacitance matrix C is similar to a diagonal matrix:

$$\tilde{F}^{-1} C \tilde{F} = diag\{\lambda_1, \cdots, \lambda_n, \lambda_{-1}, \cdots, \lambda_{-n}\},$$

i.e. we obtain the Fourier factorization of C

$$C = \tilde{F} diag\{\lambda_1, \cdots, \lambda_n, \lambda_{-1}, \cdots, \lambda_{-n}\} \tilde{F}^{-1} \tag{4.23}$$

Here, for $k = 1, \cdots, n$, λ_k are defined by

$$\lambda_k = \frac{\sqrt{(2+\tau_k+\sqrt{ab})^2-4}}{2}[\frac{1+\mu_k^{m_1+1}}{1-\mu_k^{m_1+1}} + \frac{1+\mu_k^{m_2+1}}{1-\mu_k^{m_2+1}}].$$

If $ab < 0$, then λ_{-k} are simply defined by: for $k = 1, \cdots, n$,

$$\lambda_{-k} = \frac{\sqrt{(2+\tau_k-\sqrt{ab})^2-4}}{2}[\frac{1+\mu_{-k}^{m_1+1}}{1-\mu_{-k}^{m_1+1}} + \frac{1+\mu_{-k}^{m_2+1}}{1-\mu_{-k}^{m_2+1}}].$$

If $ab > 0$, then λ_{-k} are defined by:

$$\lambda_{-k} = \frac{\sqrt{(2+\tau_k-\sqrt{ab})^2-4}}{2}(\frac{1+\mu_{-k}^{m_1+1}}{1-\mu_{-k}^{m_1+1}} + \frac{1+\mu_{-k}^{m_2+1}}{1-\mu_{-k}^{m_2+1}}) \qquad \text{if} \quad |2+\tau_k-\sqrt{ab}| > 2$$

$$\lambda_{-k} = \frac{1}{m_1+1} + \frac{1}{m_2+1} = O(h) \qquad \text{if} \quad \tau_k = \sqrt{ab}$$

$$\lambda_{-k} = \sin\theta_k \left[\frac{1}{\tan((m_1+1)\theta_k)} + \frac{1}{\tan((m_2+1)\theta_k)}\right] \qquad \text{if} \quad |2+\tau_k-\sqrt{ab}| < 2.$$

Proof *To prove this theorem, we need show that*

$$C\tilde{F}^{(k)} = \lambda_k \tilde{F}^{(k)} \qquad \text{and} \qquad C\tilde{F}^{(-k)} = \lambda_{-k}\tilde{F}^{(-k)}.$$

Let first consider the term, contributed from the sub-domain Ω_1,

$$-\begin{pmatrix} A_{BI}^{(1)} & 0 \\ 0 & A_{BI}^{(1)} \end{pmatrix} \begin{pmatrix} A_{II}^{(1)} & a \\ bI & A_{II}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} A_{IB}^{(1)} & 0 \\ 0 & A_{IB}^{(1)} \end{pmatrix} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}}w^{(k)} \end{pmatrix}.$$

The evaluation of this term requires solving the following discrete coupled difference system on sub-domain Ω_1:

$$(4.24) \qquad \begin{cases} 4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} + av_{i,j} = 0 \\ 4v_{i,j} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1} + bu_{i,j} = 0 \end{cases} \qquad \text{on } \Omega_1,$$

with boundary conditions

$$\begin{pmatrix} u_{\cdot,0} \\ \\ v_{\cdot,0} \end{pmatrix} = \tilde{F}^{(k)} \ on\ \Gamma \qquad and \qquad \begin{pmatrix} u_{i,j} \\ \\ v_{i,j} \end{pmatrix} \ on\ \partial\Omega_1/\Gamma.$$

Let us consider the solution of coupled difference system (4.24) in the form

$$\begin{cases} u_{i,j} = \alpha_j \sqrt{h}\sin(k\pi i h) \\ \\ v_{i,j} = \beta_j \sqrt{b/a}\sqrt{h}\sin(k\pi i h) \end{cases} \qquad for\ 0 \le i \le n+1, \quad 0 \le j \le m_1+1,$$

with boundary condition $\alpha_0 = 1, \beta_0 = 1$ *and* $\alpha_{m_1+1} = 0, \beta_{m_1+1} = 0$. *After substituting these form solution into coupled difference system (4.24), we easily obtain that*

$$\begin{cases} (2+\tau_k)\alpha_j - \alpha_{j-1} - \alpha_{j+1} + \sqrt{ab}\beta_j = 0 \\ \\ (2+\tau_k)\beta_j - \beta_{j-1} - \beta_{j+1} + \sqrt{ab}\alpha_j = 0 \end{cases} \qquad for\ 1 \le j \le n.$$

The corresponding characteristic polynomial is

$$r^2 - (2 + \tau_k + \sqrt{ab})r + 1 = 0.$$

The roots of this quadratic polynomial are

$$r_+ = \frac{2 + \tau_k + \sqrt{ab} + \sqrt{(2+\tau_k+\sqrt{ab})^2 - 4}}{2}$$

$$r_- = \frac{2 + \tau_k + \sqrt{ab} - \sqrt{(2+\tau_k+\sqrt{ab})^2 - 4}}{2},$$

with following properties

$$r_+ r_- = 1 \qquad r_+ + r_- = 2 + \tau_k + \sqrt{ab} \qquad r_+ - r_- = \sqrt{(2+\tau_k+\sqrt{ab})^2 - 4}.$$

Then, we have solutions to the coupled difference system as

$$\alpha_j = \beta_j = r_-^j \frac{1 - \mu_k^{m_1+1-j}}{1 - \mu_k^{m_1+1}}.$$

So

$$\begin{pmatrix} A_{BI}^{(1)} & 0 \\ 0 & A_{BI}^{(1)} \end{pmatrix} \begin{pmatrix} A_{II}^{(1)} & aI \\ bI & A_{II}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} A_{IB}^{(1)} & 0 \\ 0 & A_{IB}^{(1)} \end{pmatrix} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix} = \alpha_1^{(1)} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix}$$

where

$$\alpha_1^{(1)} = r_- \frac{1 - \mu_k^{m_1}}{1 - \mu_k^{m_1+1}}.$$

Analogously, we can derive that

$$\begin{pmatrix} A_{BI}^{(2)} & 0 \\ 0 & A_{BI}^{(2)} \end{pmatrix} \begin{pmatrix} A_{II}^{(2)} & aI \\ bI & A_{II}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} A_{IB}^{(2)} & 0 \\ 0 & A_{IB}^{(2)} \end{pmatrix} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix} = \alpha_1^{(2)} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix}$$

where

$$\alpha_1^{(2)} = r_- \frac{1 - \mu_k^{m_2}}{1 - \mu_k^{m_2+1}}.$$

Finally, Direct calculation can lead following equation:

$$\begin{pmatrix} A_{BB} & aI \\ bI & A_{BB} \end{pmatrix} \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix} = (2 + \tau_k + \sqrt{ab}) \begin{pmatrix} w^{(k)} \\ \sqrt{\frac{b}{a}} w^{(k)} \end{pmatrix}.$$

Summarizing all above results, corresponding eigenvalues λ_k can be obtained

$$C\tilde{F}^{(k)} = \lambda_k \tilde{F}^{(k)},$$

where

$$\lambda_k = \sqrt{(2 + \tau_k + \sqrt{ab})^2 - 4}[\frac{1 + \mu_k^{m_1+1}}{1 - \mu_k^{m_1+1}} + \frac{1 + \mu_k^{m_2+1}}{1 - \mu_k^{m_2+1}}].$$

The eigenvalues corresponding to eigenvectors λ_{-k} can be calculated in the same way. So we omit the derivation of eigenvectors λ_{-k}.

It is easy to see that when the coupling parameter $\bar{a}$ and $\bar{b}$ is very small, the coupled discrete system is positive definite. This conclusion can also be easily derived through Sobolev space theory.

Note when $ab < 0$, the eigenvectors and eigenvalues become complex. However, when $ab \geq 0$, all eigenvalues and eigenvectors are real. From above eigenvalues, we notice that some eigenvalues may tend to infinite or 0 if $(m_1 + 1)\theta_k$ tends to $\pi/2$ or $s\pi$. This means that the condition number of the capacitance matrix is very sensitive to the aspect ratio of the domain shape and coupling parameter.

For more general coupled elliptic system:

$$\begin{cases} L_{11}u + L_{12}v = f_1 \\ L_{21}u + L_{22}v = f_2 \end{cases},$$

where L_{11}, L_{12}, L_{21}, and L_{22} are linear elliptical operators with constant coefficients, the eigen-decomposition of its discrete system may be obtained in the analogous way. Fast Domain Decomposition Fourier Transform can be developed for this general coupled elliptic system. Interested reader can pursue this analysis to these more complicated cases.

4.3.3.2 Fourier Based Interface Preconditioners

In this subsection, we will propose several interface preconditioners, similar to those proposed by Dryja [31], Golub and Mayer [39], and Chan [17]. Then we compare these preconditioners. Since the eigenvectors and eigenvalues are complex when $ab < 0$, we usually use the exact Fourier factorization (4.23) as the preconditioner M. To use other kind preconditioners when $ab < 0$, we have to take the real part of the matrix-vector production Mf as the result. Otherwise, the matrix-vector production Mf may become complex number even when the vector f is a real vector. Hence, for simplicity, in the rest of this section, we will only consider the case when $ab > 0$.

Let the Fourier based preconditioners $M_?$ considered here have the form

$$M_? = \tilde{F}^{-1} \begin{pmatrix} \lambda_1^? & & & & & \\ & \ddots & & & & \\ & & \lambda_n^? & & & \\ & & & \lambda_{-1}^? & & \\ & & & & \ddots & \\ & & & & & \lambda_{-n}^? \end{pmatrix} \tilde{F}.$$

for different choices of $\lambda_k^?$ and $\lambda_{-k}^?$:

- Dryja type preconditioner M_D excluding coupling parameter:

$$\lambda_k^D = \sqrt{\tau_k} \qquad \lambda_{-k}^D = \sqrt{\tau_k}$$

- Dryja type preconditioner M_{DC} including coupling parameter:

$$\lambda_k^{DC} = \sqrt{\tau_k + \sqrt{ab}} \qquad \lambda_{-k}^{DC} = \sqrt{|\tau_k - \sqrt{ab}|}$$

- Golub and Mayer type preconditioner M_{GM} :

$$\lambda_k^{GM} = \sqrt{(\tau_k + \sqrt{ab}) + \frac{(\tau_k + \sqrt{ab})^2}{4}}$$

$$\lambda_{-k}^{GM} = \sqrt{|\tau_k - \sqrt{ab} + \frac{(\tau_k - \sqrt{ab})^2}{4}|}$$

- Approximate Chan type preconditioner M_{C1} :

$$\lambda_k^{C1} = \sqrt{\tau_k + \sqrt{ab} + \frac{(\tau_k + \sqrt{ab})^2}{4}} \qquad \text{for} \quad k = 1, \cdots, n$$

$$\lambda_{-k}^{C1} = \sqrt{\tau_k - \sqrt{ab} + \frac{(\tau_k - \sqrt{ab})^2}{4}} \qquad \text{if} \quad |2 + \tau_k - \sqrt{ab}| > 2$$

$$\lambda_{-k}^{C1} = h \qquad \text{if} \quad \tau_k = \sqrt{ab}$$

$$\lambda_{-k}^{C1} = \left[\frac{1}{\tan((m_1+1)\theta_k)} + \frac{1}{\tan((m_2+1)\theta_k)}\right] \sin\theta_k \qquad \text{if} \quad |2 + \tau_k - \sqrt{ab}| < 2$$

- Chan type preconditioner M_{C2} : exact eigen-decomposition (4.23) is used as preconditioner.

Remark 1: If the coupling parameter $\bar{a}$ and $\bar{b}$ are so small that the coupled system becomes positive definite, then Sobolev space theory can be easily used to show that the Dryja type preconditioner M_D is spectrally equivalent to the capacitance matrix. This equivalence, independent of mesh size h, can be satisfied even when the domain is not square and the interface is not a straight line. The proof of this is analogous to that for Poission equation [9, 31]. However, this equivalence is not true for large coupling parameter.

Remark 2: By comparing exact eigen-decomposition (4.23) with these preconditioners of the coupled elliptic system, we notice that preconditioner M_{C1} best approximates exact eigen-decomposition (4.23); preconditioner M_{GM} approximates to M_{C1} and preconditioner M_{DC} approximates to M_{GM}. We have to use the absolute value under the square root in the expressions of preconditioners M_{GM} and M_{DC} to avoid the appearance of complex number when $ab > 0$. For simplicity and clarity, we consider the eigenvalue distributions of $M_{MG}^{-1}C$ and M_{C1}^{-1}, which relates to the convergence speed of corresponding preconditioned iterative method. We hope that the absolute values of all eigenvalues are bounded from above and below. This bound should not depend on mesh size h and coupling parameter if possible.

Lemma 4.3 *The exact eigenvalues of $M_{MG}^{-1}C$ are:*

$$\lambda_k = \frac{1+\mu_k^{m_1+1}}{1-\mu_k^{m_1+1}} + \frac{1+\mu_k^{m_2+1}}{1-\mu_k^{m_2+1}} \qquad \mathit{for} \quad k = 1, \cdots, n$$

$$\lambda_{-k} = \frac{1+\mu_{-k}^{m_1+1}}{1-\mu_{-k}^{m_1+1}} + \frac{1+\mu_{-k}^{m_2+1}}{1-\mu_{-k}^{m_2+1}} \qquad \mathit{if} \quad |2+\tau_k-\sqrt{ab}| > 2$$

$$\lambda_{-k} = \frac{1}{m_1+1} + \frac{1}{m_2+1} = O(h) \qquad \mathit{if}\ \tau_k = \sqrt{ab}$$

$$\lambda_{-k} = \frac{1}{\tan((m_1+1)\theta_k)} + \frac{1}{\tan((m_2+1)\theta_k)} \qquad \mathit{if} \quad |2+\tau_k-\sqrt{ab}| < 2$$

From the distribution of these eigenvalue, we could explain the general trend of numerical results presented in this paper. Note $\mu_k < 1$ and $\mu_{-k} < 1$ or > 1. Then, $1 \leq \lambda_k \leq 3$ for all $k = 1, \cdots, n$. So we have following results on the eigenvalues of M_{GM}^{-1}.

- When $\sqrt{ab} \leq \tau_k$, then $\mu_{-k} < 1$ and the corresponding absolute eigenvalues of $M_{GM}^{-1}C$ are bounded from above and below by 4 and 1.

- When $\sqrt{ab} > 4 + \tau_k$, then $\mu_{-k} > 1, \mu_k < 1$ and the absolute values of all eigenvalues are bounded from above and below by 4 and 1 respectively.

- If $\tau_k < \sqrt{ab} < 4 + \tau_k$, then some eigenvalue will be expressed by tan functions which are difficult to be bounded.

Lemma 4.4 *The exact eigenvalues of $M_{C1}^{-1}C$ are:*

$$\lambda_k = \frac{1+\mu_k^{m_1+1}}{1-\mu_k^{m_1+1}} + \frac{1+\mu_k^{m_2+1}}{1-\mu_k^{m_2+1}} \qquad \textit{for} \quad k = 1, \cdots, n$$

$$\lambda_{-k} = \frac{1+\mu_{-k}^{m_1+1}}{1-\mu_{-k}^{m_1+1}} + \frac{1+\mu_{-k}^{m_2+1}}{1-\mu_{-k}^{m_2+1}} \qquad \textit{if} \quad |2 + \tau_k - \sqrt{ab}| > 2$$

$$\lambda_{-k} = 1 \qquad \textit{if} \quad |2 + \tau_k - \sqrt{ab}| \leq 2$$

Now we analyze the properties of these eigenvalues of $M_{C1}^{-1}C$. If h is fixed, then

$$\lim_{m_1, m_2 \to \infty} \lambda_k = 2 \qquad \lim_{m_1, m_2 \to \infty} \lambda_{-k} = \begin{cases} 2 & \text{if } \tau_k - \sqrt{ab} > 0 \\ -2 & \text{if } \tau_k - \sqrt{ab} < -4 \end{cases} .$$

Therefore, the absolute values of the eigenvalues of $M_{C1}^{-1}C$ are larger than 1 and smaller than 3.

Let us show this result by denoting that

$$\delta_k = \frac{\sqrt{(2+\tau_k+\sqrt{ab})^2 - 4}}{2 + \tau_k + \sqrt{ab}} \qquad \text{for } k = 1, \cdots, n,$$

$$\delta_{-k} = \frac{\sqrt{(2+\tau_k-\sqrt{ab})^2 - 4}}{2 + \tau_k - \sqrt{ab}} \qquad \text{for } |2 + \tau_k - \sqrt{ab}| > 2.$$

Then,

$$\mu_k^{m_1+1} = (1 - \frac{2\delta_k}{1+\delta_k})^{l_1/h} \qquad \mu_{-k}^{m_1+1} = (1 - \frac{2\delta_{-k}}{1+\delta_{-k}})^{l_1/h},$$

and

$$\mu_k^{m_2+1} = (1 - \frac{2\delta_k}{1+\delta_k})^{l_2/h} \qquad \mu_{-k}^{m_2+1} = (1 - \frac{2\delta_{-k}}{1+\delta_{-k}})^{l_2/h}.$$

It can be verified that μ_k is a decreasing function of k. By using the fact that

$$\lim_{x\to 0}(1 + xf(x))^{\frac{1}{x}} = e^{\lim_{x\to 0} f(x)},$$

we have that

$$\lim_{h\to 0} \mu_1^{m_1+1} = e^{-2l_1\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}} \qquad \lim_{h\to 0} \mu_1^{m_2+1} = e^{-2l_2\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}}$$

Since $\tau_k - \sqrt{ab} < -4$ does not satisfy as h tends to zero, we can assume that $\tau_k - \sqrt{ab} > 0$ for $k \geq s$. It is easy to show that μ_{-k} is a decreasing function of $k \geq s$ and

$$\lim_{h\to 0} \mu_{-1}^{m_1+1} = e^{-2l_1\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}} \qquad \lim_{h\to 0} \mu_{-1}^{m_2+1} = e^{-2l_2\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}}.$$

Therefore, we have obtained that

$$\lim_{h\to 0} \lambda_k \leq \frac{1 + e^{-2l_1\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}}}{1 - e^{-2l_1\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}}} + \frac{1 + e^{-2l_2\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}}}{1 - e^{-2l_2\sqrt{\pi^2+\sqrt{\bar{a}\bar{b}}}}} \qquad \text{for } k = 1, \cdots, n,$$

$$\lim_{h\to 0} \lambda_{-k} \leq \frac{1 + e^{-2l_1\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}}}{1 - e^{-2l_1\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}}} + \frac{1 + e^{-2l_2\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}}}{1 - e^{-2l_2\sqrt{(s\pi)^2-\sqrt{\bar{a}\bar{b}}}}} \qquad \text{for } k \geq s$$

and $\lim_{h\to 0} \lambda_{-k} = 1$ for $k = 1, \cdots, s$.

To apply these Fourier based preconditioners to the coupled elliptic system with variable coefficients, we will use scaled Fourier approximation with $\bar{a}$ and $\bar{b}$ as the mean values of coupling variable on the interface. Let D denote the diagonal of A_{BB}. Then, preconditioner for the problem with variable coefficients can be defined by:

$$\tilde{M} = \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix}^{\frac{1}{2}} M_? \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix}^{\frac{1}{2}},$$

with $\bar{a} = \sum_{i=1}^{n} \bar{a}(x_i, \frac{n+1}{2}h)/n$ and $\bar{b} = \sum_{i=1}^{n} \bar{b}(x_i, \frac{n+1}{2}h)/n$ in the definition of $M_?$ given before. Since $M_?$ may be complex, we choose the real part of matrix vector product $M_?^{-1}r$ as its result of preconditioning step. The related numerical results will be shown in the last section.

4.3.4 Probe Technique Applied to the Coupled Elliptic System

The basic idea of probe technique is to construct approximate interface matrix through using a few matrix vector products to capture the strongest coupling of the exact interface matrix. This algebraic technique can easily be applied to any operator having decay properties. In this section, we propose several preconditioners for coupled interface capacitance matrix C by using probe technique.

We introduce the notation

$$M_d = \mathrm{PROBE}(C, d),$$

to denote the probe procedure of creating the banded approximation M_d of matrix C, in which M_d is a banded matrix with bandwidth d. In most case, we interest the tridiagonal probe preconditioner. Here, we illustrate the PROBE procedure for the case $d = 1$. We use following probe vectors used commonly: $v_1 = (1,0,0,1,0,0,\cdots)^T, v_2 = (0,1,0,0,1,0,\cdots)^T$ and $v_3 = (0,0,1,0,0,1,\cdots)^T$. Since M_1 is tridiagonal, direct calculation of $[M_1v_1, M_1v_2, M_1v_3]$ leads to:

$$\begin{pmatrix} m_{11} & m_{12} & & & \\ m_{21} & m_{22} & m_{23} & & \\ & m_{32} & m_{33} & m_{34} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & 0 \\ m_{21} & m_{22} & m_{23} \\ m_{34} & m_{32} & m_{33} \\ \vdots & \vdots & \vdots \end{pmatrix}.$$

The probe algorithm reconstructs the nonzero entries m_{ij} by equating above right hand side to matrix vector product $[Cv_1, Cv_2, Cv_3]$. From this probe procedure, we need three matrix vector products to obtain tridiagonal approximate matrix M.

Here, we discuss how to apply probe technique to the interface matrix C of coupled elliptic system. Since Fourier transform diagonalize the matrix C, Fourier transform is combined with probe technique to construct preconditioner of the capacitance matrix C.

We first assume that the capacitance matrix C has the properties that its entries decay from the diagonal when the coupling parameter is small. We approximate the capacitance matrix C by

$$P_1 = \text{PROBE}(C, 1),$$

which is used as preconditioner in the preconditioned iterative method.

We notice that when the interface capacitance matrix C is written in the form of 2×2 block square sub-matrices, i.e.

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

entries of each square sub-matrix decay from their diagonals. We can construct the preconditioner P_2 as:

$$P_2 = \begin{pmatrix} \text{PROBE}(C_{11}, 1) & \text{PROBE}(C_{12}, 1) \\ \text{PROBE}(C_{21}, 1) & \text{PROBE}(C_{22}, 1) \end{pmatrix}.$$

It is natural to combine Fourier method with probe method. Since matrix $\tilde{F}^{-1} C \tilde{F}$ becomes diagonal for the constant coupling parameter from theorem 3.1, probe procedure can be used to find tridiagonal approximation to $\tilde{F}^{-1} C \tilde{F}$. Hence, we propose following preconditioner:

$$\tilde{P}_3 = \tilde{F} * \text{PROBE}(\tilde{F}^{-1} C \tilde{F}, 1) * \tilde{F}^{-1},$$

is defined as a preconditioner. Note that we substitute $\sqrt{b/a}$ by $\sqrt{|b/a|}$ in the definition of matrix $\tilde{F}$ in order to avoid the appearance of complex numbers.

We can also define a simpler Fourier probe preconditioner by:

$$P_3 = F * \text{PROBE}(F^T C F, 1) * F^T.$$

If $F^T C F$ is written in the form:

$$F^T C F = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

we assume that M_{ij} can be approximated by tridiagonal matrix. We use probe technique to find this approximation:

$$P_4 = F \begin{pmatrix} \mathrm{PROBE}(M_{11}, 1) & \mathrm{PROBE}(M_{12}, 1) \\ \mathrm{PROBE}(M_{21}, 1) & \mathrm{PROBE}(M_{22}, 1) \end{pmatrix} F^T.$$

Let us write $\tilde{F}^{-1}C\tilde{F}$ in the form:

$$\tilde{F}^{-1}C\tilde{F} = \begin{pmatrix} \tilde{M}_{11} & \tilde{M}_{12} \\ \tilde{M}_{21} & \tilde{M}_{22} \end{pmatrix}.$$

We assume that $\tilde{M}_{ij}$ can be well approximated by tridiagonal matrices. So probe procedure is used to construct these tridiagonal approximations:

$$\tilde{P}_4 = \tilde{F} \begin{pmatrix} \mathrm{PROBE}(\tilde{M}_{11}, 1) & \mathrm{PROBE}(\tilde{M}_{12}, 1) \\ \mathrm{PROBE}(\tilde{M}_{21}, 1) & \mathrm{PROBE}(\tilde{M}_{22}, 1) \end{pmatrix} \tilde{F}^{-1}.$$

4.3.5 Numerical Results

In this section, we will test the convergence behavior of preconditioned BiCG method with above preconditioners. The test were conducted for the model coupled elliptic system (4.19) with various choices of coupling parameter $\bar{a}$ and $\bar{b}$ and the mesh size $h = 1/(n+1)$. The domain Ω is a unit square $\Omega = [0,1]^2$ partitioned into an $n \times n$ with mesh size $h = 1/(n+1)$. In our test, the stopping criterion was chosen to be

$$\frac{||r_k||}{||r_0||} \leq 10^{-5}.$$

The entries of the exact solution were chosen randomly from a uniform distribution.

Since coupled elliptic system (4.19) with constant coupling parameters can be easily reduced to symmetric (i.e., $\bar{a} = \bar{b}$) or unsymmetric (i.e., $\bar{a} = -\bar{b}$), see [30], the test were performed for the system with these two case. For the variable coupling parameter $\bar{a}(x, y)$ and $\bar{b}(x, y)$, we will request their mean values should be equal or have different sign. Then the preconditioner $\tilde{P}_3$ is equivalent to P_3, and $\tilde{P}_4$ is equivalent to P_4. Therefore, we only test the preconditioners P_k for $k = 1, 2, 3, 4$ in our numerical experiments.

We first test Fourier based preconditioners and list the results in the following tables. Since we are interested in the relationship between mesh size, coupling parameters $\bar{a}$ and $\bar{b}$ and convergence rate, we change the mesh size and these constant coupling parameters with $\bar{a} = \bar{b}$ or $\bar{a} = -\bar{b}$. We also try to apply the scaled Fourier based preconditioners to the coupled elliptic system with variable coupling parameter.

Remark on the Fourier based preconditioners:

1. The iteration number N does not depend on the mesh size h but strongly depends on the coupling parameter.

2. The iteration number of preconditioner M_{DC} increase slower than that of preconditioner M_D as the coupling parameter increasing.

3. After the iteration number reaching the maximum, the iteration numbers of preconditioner M_{DC} will quickly decrease as the coupling parameter α

Table 4.1: Fourier based preconditioner for constant coupling parameters

Preconditioners		M_D	M_{DC}	M_{GM}	M_{C1}
Grid Points	$\bar{a} = \bar{b}$	Iteration Number			
$n \times n$	$\bar{a}$	N	N	N	N
7×7	5	5	4	3	3
	50	10	7	4	3
	100	11	9	5	3
	1000	13	8	1	2
	10000	13	4	1	2
15×15	5	5	5	3	3
	50	12	8	4	3
	100	12	9	5	3
	1000	30	20	14	3
	10000	25	6	1	2
25×25	5	6	5	3	5
	50	12	8	5	3
	100	13	9	5	3
	1000	28	17	12	3
	10000	39	8	2	2

Table 4.2: Fourier based preconditioner for constant coupling parameters

Preconditioners		M_D	M_{DC}	M_{GM}	M_{C1}
Grid Points	$\bar{a} = -\bar{b}$	Iteration Number			
$n \times n$	$\bar{a}$	N	N	N	N
7×7	5	10	10	7	6
	50	10	12	7	4
	100	10	12	8	4
	1000	14	6	6	2
	10000	14	4	4	2
15×15	5	12	12	6	6
	50	12	14	7	4
	100	12	9	5	3
	1000	18	11	8	2
	10000	24	4	4	2
25×25	5	12	12	6	6
	50	14	14	6	4
	100	14	14	6	4
	1000	22	14	8	2
	10000	22	14	8	2

Table 4.3: Fourier based preconditioner for variable coupling parameters

Preconditioners		M_D	M_{DC}	M_{GM}	M_{C1}	M_{C2}
$n \times n$	$\bar{a}, \quad \bar{b}$	N	N	N	N	N
7×7		9	9	7	8	7
15×15	$\bar{a} = 100x^2$	9	9	7	8	7
25×25	$\bar{b} = 100(1-x)^2$	8	8	7	8	7
7×7		12	12	8	8	8
15×15	$\bar{a} = 100x^2$	14	14	8	8	8
25×25	$\bar{b} = -100(1-x)^2$	14	14	8	8	8
7×7		5	5	5	5	5
15×15	$\bar{a} = e^{20(x-0.5)y}$	7	7	6	6	6
25×25	$\bar{b} = \bar{a}$	6	6	6	6	5
7×7		12	12	8	8	6
15×15	$\bar{a} = e^{20(x-0.5)y}$	12	12	8	8	6
25×25	$\bar{b} = -\bar{a}$	14	14	8	8	6
7×7		12	12	6	8	6
15×15	$\bar{a} = e^{20(x-0.5)y}$	12	12	8	8	6
25×25	$\bar{b} = -100(0.5-x)^2$	12	12	6	8	6

increasing.

4. However, for preconditioner M_D, the iteration number slightly decreases and remains the same large iteration number.

5. preconditioner M_{GM} has a convergence rate, independent of the mesh size h, but dependent on the coupling parameters.

6. Preconditioner M_{C1} has iteration number, which is independent of mesh size h and the coupling parameters.

Finally, we test probing based preconditioners and list their results. The constant coupling parameters have been tested and their results are shown in Table 4.4 and Table 4.5. The smoothly varying and highly varying coupling coefficients are considered and their results are presented in Table 4.6.

Remark on the probing based preconditioners:

1. From the last column of Table 4.4, Table 4.5 and Table 4.6, we can see that the iteration number increases as the mesh size h decreases when BiCG method is used without preconditioner. We also notice that the iteration number first increases and then decreases as the coupling parameters $\bar{a}$ and $\bar{b}$ increases. The reason for this phenomenon is that small coupling parameters and Laplace operator make the problem become more ill condition, but enough large coupling parameters make the discrete Laplace become dormant so that the condition number is improved and the iteration number

Table 4.4: Probe based preconditioner for constant coupling parameters

Preconditioners		P_1	P_2	P_3	P_4	I
Grid Points	$\bar{a} = \bar{b}$	Iteration Number				
$n \times n$	$\bar{a}$	N	N	N	N	N
7×7	5	5	4	1	1	12
	50	10	9	1	1	12
	100	13	6	1	1	13
	1000	15	1	1	1	11
	10000	12	1	1	1	6
15×15	5	6	6	1	1	18
	50	10	24	1	1	21
	100	11	8	1	1	21
	1000	18	18	1	1	27
	10000	4	1	1	1	8
25×25	5	7	7	1	1	22
	50	10	34	1	1	26
	100	11	12	1	1	28
	1000	28	19	1	1	29
	10000	54	2	1	1	12

Table 4.5: Probe based preconditioner for constant coupling parameters

Preconditioners		P_1	P_2	P_3	P_4	I
Grid Points	$\bar{a} = -\bar{b}$	Iteration Number				
$n \times n$	$\bar{a}$	N	N	N	N	N
7×7	5	4	11	5	1	13
	50	8	> 80	8	1	14
	100	11	> 80	11	1	14
	1000	13	1	11	1	10
	10000	12	1	8	1	6
15×15	5	6	> 80	3	1	21
	50	8	> 80	6	1	21
	100	11	> 80	8	1	22
	1000	19	23	20	1	18
	10000	3	1	3	1	8
25×25	5	8	> 80	4	1	25
	50	10	> 80	8	1	26
	100	11	> 80	9	1	27
	1000	19	> 80	21	1	20
	10000	50	1	37	1	12

Table 4.6: Probe based preconditioner for variable coupling parameters

Preconditioners		P_1	P_2	P_3	P_4	I
$n \times n$	$\bar{a}, \quad \bar{b}$	N	N	N	N	N
7×7		8	> 80	8	> 80	14
15×15	$\bar{a} = 100x^2$	8	> 80	7	> 80	23
25×25	$\bar{b} = 100(1-x)^2$	9	> 80	8	> 80	26
7×7		7	> 80	8	> 80	14
15×15	$\bar{a} = 100x^2$	9	> 80	7	> 80	21
25×25	$\bar{b} = -100(1-x)^2$	12	> 80	7	> 80	27
7×7		5	4	4	4	12
15×15	$\bar{a} = e^{20(x-0.5)y}$	7	9	7	7	19
25×25	$\bar{b} = \bar{a}$	6	7	5	5	22
7×7		6	> 80	5	> 80	14
15×15	$\bar{a} = e^{20(x-0.5)y}$	6	> 80	5	> 80	21
25×25	$\bar{b} = -\bar{a}$	7	> 80	5	> 80	25
7×7		6	> 80	6	> 80	14
15×15	$\bar{a} = e^{20(x-0.5)y}$	6	> 80	6	> 80	21
25×25	$\bar{b} = -100(0.5-x)^2$	8	> 80	5	> 80	27

decreases.

2. When the coupling parameters $\bar{a}$ and $\bar{b}$ are constant and $\bar{a} = \bar{b}$, the probe based preconditioners P_3 and P_4 greatly improve the convergence rate of BiCG method.

3. When the constant coupling parameters $\bar{a} = -\bar{b}$, P_4 accelerates the convergence speed of BiCG method. When $\bar{a}$ is small, the convergence has also been improved when P_1 or P_3 is used as the preconditioner in BiCG method.

4. As for the variable coupling parameters, P_3 and P_1 speed up the convergence of BiCG method. Their corresponding iteration numbers are independent of mesh size h. The preconditioners P_2, P_4 and $\tilde{P}_4$ make the BiCG method diverge.

CHAPTER 5

Domain Decomposition for Singular Neumann Boundary Value Problems

In this chapter, we consider the problem of solving the very large systems of symmetric and semi-positive definite algebraic equation, arising from the discretization of elliptic problems with Neumann boundary conditions by finite differences or finite elements. We will apply BPS method and vertex space domain decomposition (VSDD) method to this kind problem, We will discuss their convergence for these singular Neumann boundary value problems. We also further improve probe technique in our applications. Only four probe vectors are needed to form approximate edge and vertex matrix in our calculation.

The aim of this chapter is that by giving and using the generalized framework of domain decomposition [35], we modify and analyze the additive Schwarz [32], the vertex space (VS) domain decomposition method [66] and BPS [8], developed by Dryja and Widlund, Smith and Bramble et al., respectively so that these methods can be applied to the symmetric semi-positive definite systems of linear algebraic equations, which result from finite element approximation of elliptic system with Neumann boundary condition. We also focus on the improvement of probe technique in the construction of edge and vertex approximations. The difficulty of these problems is how we could efficiently modify domain decomposition preconditioner so that the result of preconditioner is orthogonal to the kernel space. These modified domain decomposition methods still have condition numbers which are independent of the mesh size. VS method can be viewed as a kind of the additive Schwarz method (ASM) [35], which forms one of the most important classes of parallel domain decomposition methods. Comparing additive Schwarz with VS and BPS methods, we found that additive Schwarz is easier constructed than VS and BPS methods but VS and BPS methods are more suitable for solving problem with highly jumping coefficients. Theoretically, we can only show that the BPS method for this kind singular problems has a condition number $C(1+\ln(H/h))^2$ independent of certain discontinuous coefficients. A brief general framework is presented which may be quite useful in designing, extending and analyzing variants of domain decomposition methods for these symmetric semi-positive definite systems. As we know, it is very expensive to construct the exact edge and vertex matrices in the vertex space method [66] and BPS method [8]. In order to keep the quick convergence of these methods for the problem with highly varying coefficients, we use several variants of edge and vertex approximation matrices [23, 25]. These variants are based on Fourier approximation [31, 39, 8] and probe technique [22, 21, 23]. Here, we further improve the probe technique in our applications. Only four probe vectors instead of six probe vectors have been used to multiply Schur complement [23] and to form edge and vertex approximations. Each product of Schur complement with probe vector includes solving all subproblems on sub-domains. Hence, we save $\frac{1}{3}$ computational cost while we are constructing the edge and vertex approximation matrices by using 4 probe vectors in stead of 6 probe vectors in the probe technique.

One of our motivations in generalizing the framework of additive Schwarz and modifying vertex space method is to design an optimal parallel algorithm for Navier-Stokes equation. As we know, one of the most difficult parts of parallel computation for this nonlinear differential equation is that we have to solve elliptic problems with some boundary conditions in each time step. Generally, there are two types of finite difference schemes for this nonlinear problem. One is based on the velocity and stream function. The other is based on the velocity and pressure formulation. In the second scheme, we will have to solve symmetric semi-positive system, resulting from Laplace equation with "Neumann like" boundary conditions, in each time step. The application and theory of domain decomposition technique for positive definite and symmetric system has been widely discussed, [38, 14, 15, 16]. However, not so much concrete work has been done for symmetric semi-positive system [7, 2, 56, 63]. The difficulty of these problems is how to modify domain decomposition methods so that an approximate solution orthogonal to the kernel space can be efficiently obtained. It is not obvious whether the preconditioner problem is well defined even when the null space of domain decomposition preconditioners is the same as that of original problem. In this paper, the additive Schwarz, the vertex space domain decomposition and BPS methods have been successfully extended for this kind singular problem so that these methods still have quick convergent property. For simplicity and without loss generality, we restrict ourselves to the scalar elliptic problems or linear elasticity problems with Neumann boundary condition.

5.1 General Framework for Neumann Boundary Problems

In this section we first give additive Schwarz framework for variational problem with natural boundary condition. Then we present elasticity problem and elliptic problem with Neumann boundary condition.

5.1.1 Additive Schwarz Framework

Assume that the connected polygonal domain Ω in R^d is the union of disjoint regions Ω_k, which are either quadrilaterals or triangles;

$$\bar{\Omega} = \cup_k \bar{\Omega}_k \qquad \text{and} \quad \Omega_i \bigcap \Omega_j = \emptyset \qquad \text{if } i \neq j.$$

We denote the boundary of each sub-domain as $\partial\Omega_k$, and the union of these boundaries as $\Gamma = \cup_k \partial\Omega_k$. In the Sobolev space $\mathbf{V} = (H^1(\Omega))^q$, we introduce a bilinear form $a(\cdot,\cdot) : \mathbf{V} \times \mathbf{V} \rightarrow R$, which is symmetric, bounded and semi-positive definite. Let $(\cdot,\cdot)$ be the inner product in $(L^2(\Omega))^q$:

$$(\mathbf{f}, \mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v} dx \tag{5.1}$$

and $(\mathbf{f},\mathbf{v}) : \mathbf{V} \rightarrow R$ be a continuous linear functional. The kernel space is defined by

$$KerA = \{\mathbf{u}|\mathbf{u} \in \mathbf{V}, \qquad a(\mathbf{u},\mathbf{v}) = 0, \qquad \forall \mathbf{v} \in \mathbf{V}\}.$$

To define a Neumann boundary problem, we shall assume that the compatible condition is satisfied:

$$(5.2) \qquad (\mathbf{f},\mathbf{v}) = 0 \qquad \forall \mathbf{v} \in KerA.$$

We consider a general variational problem with a natural boundary condition in the Hilbert space $\mathbf{V} = (H^1(\Omega))^q$: Find $\mathbf{u} \in \mathbf{V}$, and $\mathbf{u} \perp KerA$ such that

$$(5.3) \qquad a(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v}), \qquad \forall \mathbf{v} \in \mathbf{V}.$$

For problem (5.3), we introduced two levels of triangulations of Ω. One is the coarse triangulation defined by the substructures Ω_i of diameter $0(H)$. The other is the fine triangulation defined by further dividing the substructures into elements of diameter $0(h)$. Assume that these triangulations are shape regular in the sense common to finite element theory; cf Ciarlet [27].

Let $\mathbf{V}^h(\Omega)$ and $\mathbf{V}^H(\Omega)$ be the finite element space of continuous, piecewise linear functions defined on the fine grid and coarse grid respectively. $\mathbf{V}^h(\Omega)$ and $\mathbf{V}^H(\Omega)$ are the subspaces of Sobolev space $(H^1(\Omega))^q$. In these two subspaces, we respectively define following two kernel spaces:

$$KerA_h = \{\mathbf{v}_h|\mathbf{v}_h \in \mathbf{V}^h, \quad a(\mathbf{v}_h,\mathbf{u}_h) = 0, \forall \mathbf{u}_h \in \mathbf{V}^h(\Omega)\}$$

and

$$KerA_H = \{\mathbf{v}_H | \mathbf{v}_H \in \mathbf{V}^h, \quad a(\mathbf{v}_H, \mathbf{u}_H) = 0, \forall \mathbf{u}_H \in \mathbf{V}^H(\Omega)\}.$$

Then, the discrete formula of the problem (5.3) is of the form: Find $\mathbf{u}_h \in \mathbf{V}^h$ such that

$$\begin{cases} a(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) & \forall \mathbf{v}_h \in \mathbf{V}^h(\Omega) \\ (\mathbf{u}_h, \mathbf{v}_h) = 0, & \forall \mathbf{v}_h \in KerA_h \end{cases} \tag{5.4}$$

Let $\{\pi_k\}$ be the standard finite element basis functions of $\mathbf{V}^h$. Representing the solution as $\mathbf{u}_h = \sum x_k \pi_k$ results in a symmetric semi-positive definite system:

$$\mathbf{A}x = \mathbf{f}_h, \tag{5.5}$$

where A is the stiffness matrix with entries $a(\pi_i, \pi_k)$, and x and f are vectors with elements x_k and (f, π_k) respectively.

If we denote the subspace $\tilde{\mathbf{V}}^h(\Omega)$ as

$$\tilde{\mathbf{V}}^h(\Omega) = \{\mathbf{v}_h | \mathbf{v}_h \in \mathbf{V}^h(\Omega), \quad (\mathbf{v}_h, \mathbf{u}_h) = 0, \quad \forall \mathbf{u}_h \in KerA_h\},$$

there is an equivalent discrete formula of the equation (5.4) defined by: Find $\mathbf{u}_h \in \tilde{\mathbf{V}}^h(\Omega)$ such that

$$a(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \tilde{\mathbf{V}}^h(\Omega). \tag{5.6}$$

Since the bilinear form $a(\cdot, \cdot)$ is symmetric positive definite in the space $\tilde{V}^h(\Omega)$, by the Lax-Milgram theorem this problem has an unique solution in $\tilde{V}^h(\Omega)$.

To define an additive Schwarz (AS) method for problem (5.4), we represent the discrete space $\mathbf{V}^h(\Omega)$ as the sum of $N+1$ subspaces.

$$\mathbf{V}^h(\Omega) = \mathbf{V}_0^h + \mathbf{V}_1^h + \mathbf{V}_2^h + \cdots + \mathbf{V}_N^h.$$

Here $\mathbf{V}_k^h$ are the subspaces of $\mathbf{V}^h(\Omega)$. We usually choose the first subspace $\mathbf{V}_0^h$ to be the finite element subspace $\mathbf{V}^H(\Omega)$, defined on the coarse mesh (i.e. sub-domains), in order to improve the global communication of algorithms in some way. The other N subspaces $\mathbf{V}_k^h$ (usually consisting of restriction functions on sub-domains) can be chosen quite arbitrarily. With each of these subspaces $\mathbf{V}_k^h$, there is a corresponding projection operator $P_k : \mathbf{V}^h(\Omega) \rightarrow \mathbf{V}_k^h$ defined by:

$$a(P_k\mathbf{u}_h, \mathbf{v}_h) = a(\mathbf{u}_h, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_k^h. \tag{5.7}$$

We note that $P_k\mathbf{u}_h = 0, \quad \forall \mathbf{u}_h \in KerA_h$ and $P_k\mathbf{u}_h$ can be determined if we can uniquely solve problems: Find $P_k\mathbf{u}_h \in \mathbf{V}_k^h$ such that

$$a(P_k\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_k^h. \tag{5.8}$$

However, when $V_0^h = V^H(\Omega)$, problem (5.8) has many solutions in $\mathbf{V}_0^h$. Let $\tilde{\mathbf{V}}^H$ be a space defined by:

$$\tilde{\mathbf{V}}^H(\Omega) = \{\mathbf{v}_H | \mathbf{v}_H \in \mathbf{V}^H(\Omega), \quad (\mathbf{v}_H, \mathbf{u}_H) = 0, \quad \forall \mathbf{u}_H \in KerA_H\}.$$

$\mathbf{u}_H \in \tilde{\mathbf{V}}^H$ can be uniquely determined from equation

$$a(\mathbf{u}_H, \mathbf{v}_H) = a(\mathbf{u}_h, \mathbf{v}_H) = (\mathbf{f}, \mathbf{v}_H) \qquad \forall \mathbf{v}_H \in \mathbf{V}^H,$$

which leads to linear system on the coarse grid:

$$A_H x_H = f_H. \tag{5.9}$$

Then every element in the set of $\mathbf{u}_H + KerA_H$ is the solution of problem (5.8), when $k = 0$. We have to choose an appropriate solution $P_0 u_h \in V^H(\Omega)$ such that $\sum_{k=0}^N P_k \mathbf{u}_h$ is orthogonal to the kernel space $KerA_h$. Since $KerA_h$ usually consists of constant and linear function for most Neumann boundary problems, we can assume that the Neumann boundary problems, considered here, satisfy $KerA_h = KerA_H$. Because of this assumption , we can find $\mathbf{w}_H \in KerA_H$ and let $P_0\mathbf{u}_h = \mathbf{u}_H + \mathbf{w}_H$ such that $\sum_{k=0}^N P_k \mathbf{u}_h$ is orthogonal to the kernel space $KerA_h$. Thus, all $P_k \mathbf{u}_h$ for $k = 0, 1, \cdots, N$ have been uniquely defined.

We summarize the process of calculating $P\mathbf{u}_h$ if $\mathbf{f}$ is known here but $\mathbf{u}_h$ is not known:

Calculating $P\mathbf{u}_h$

1. Calculate $P_k \mathbf{u}_h$ by solving equation (5.7): Find $P_k \mathbf{u}_h \in \mathbf{V}_k^h$ such that

$$a(P_k \mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_k^h, \qquad \text{for } k = 1, \cdots, N;$$

2. Find the unique solution $\mathbf{u}_H \in \mathbf{V}^H$ such that $\mathbf{u}_H \perp KerA_H$ and

$$a(\mathbf{u}_H, \mathbf{v}_H) = (\mathbf{f}, \mathbf{v}_H) \qquad \forall \mathbf{v}_H \in \mathbf{V}^H.$$

3. Let $\mathbf{w}_h = \mathbf{u}_H + P_1 \mathbf{u}_h + \cdots + P_N \mathbf{u}_h$ and find $\bar{\mathbf{w}}_H \in KerA_H$ such that $\mathbf{w}_h + \bar{\mathbf{w}}_H$ is orthogonal to $KerA_h$. Thus, we obtain $P\mathbf{u}_h = (\bar{\mathbf{w}}_H + \mathbf{u}_H) + P_1 \mathbf{u}_h + \cdots + P_N \mathbf{u}_h$ with $P_0 \mathbf{u}_h = \bar{\mathbf{w}}_H + \mathbf{u}_H$.

Problem (5.4) can be replaced by a well defined operator equation of the form

$$P\mathbf{u} = (P_0 + P_1 + \cdots + P_N)\mathbf{u}_h = \mathbf{g} \tag{5.10}$$

where the right-hand side is equal to $\mathbf{g} = \sum_{i=0}^{N} \mathbf{g}_i$ with $\mathbf{g}_i = P_i\mathbf{u}_h$. We note that the kernel space of operator P is the same as $KerA_h$ and g is orthogonal to the kernel space $KerA_h$. Therefore, problem (5.10) is well defined. It is straightforward that operator P is always automatically symmetric semi-positive definite in the discrete space $\mathbf{V}^h$ with respect to the bilinear form $a(\cdot,\cdot)$ and is positive definite in the space $\tilde{\mathbf{V}}^h$. Hence, we often use iterative method, such as conjugate gradient method, to calculate the approximate solution of operator equation (5.10) in the space $\tilde{\mathbf{V}}^h$. It is well known that the number of iterations required, to decrease an appropriate norm of the error of this iteration method by a fixed factor, depends on the condition number $\kappa(P)$ of operator P in the space $\tilde{\mathbf{V}}^h$. For the conjugate gradient (CG) iterative method, the iteration number is proportional to $\sqrt{\kappa(P)}$: cf Golub and Van Loan [40]. We, therefore, need derive inequalities

$$\lambda_0 a(\mathbf{v}_h, \mathbf{v}_h) \leq a(P\mathbf{v}_h, \mathbf{v}_h) \leq \lambda_1 a(\mathbf{v}_h, \mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \tilde{\mathbf{V}}^h \tag{5.11}$$

in order to get estimation of condition number κ of operator P by $\kappa \leq \lambda_1/\lambda_0$. An upper bound λ_1 for the eigenvalues of P can be easily given by $N+1$ since P is the sum of projections. This upper bound can be improved by using following lemma which is obvious.

Lemma 5.1 *Denote N_k as the total number of the other different subspaces V_j^h*

which satisfies $V_k^h \cap V_j^h \neq 0$ *for* $j \neq k$. *Let* $p = \max_{1 \leq k \leq N} N_k + 1$. *Then, we have*

$$\lambda_1 = \lambda_{\max} \leq p.$$

A lower bound λ_0 can conveniently be estimated by using the useful lemma, given by Lions [49]; a proof is also given in Widlund [71].

Lemma 5.2 *Let* $\boldsymbol{u}_h \in \tilde{\boldsymbol{V}}^h$ *be written as* $\boldsymbol{u}_h = \sum_{k=0}^N \boldsymbol{u}_h^k$, *where* $\boldsymbol{u}_h^k \in \boldsymbol{V}_k^h$, *be a partition of an element of* $\boldsymbol{V}^h(\Omega) = \boldsymbol{V}_0^h + \boldsymbol{V}_1^h + \cdots + \boldsymbol{V}_N^h$. *If this partition can be chosen so that*

$$\sum_{k=0}^N a(\boldsymbol{u}_h^k, \boldsymbol{u}_h^k) \leq C_0^2, \qquad \forall \boldsymbol{u}_h \in \tilde{\boldsymbol{V}}^h(\Omega),$$

then $\lambda_0 = \lambda_{min} \geq C_0^{-2}$.

In practice, we usually use the other spectrally equivalent bilinear form $b_k(\cdot, \cdot)$ in the space $\mathbf{V}_k^h$ in stead of $a(\cdot, \cdot)$. We can describe the AS method in a more abstract way by introducing new generalized projection operator $\hat{P}_k : \mathbf{V}^h(\Omega) \to \mathbf{V}_k^h$, defined by

$$b_k(\hat{P}_k \mathbf{u}_h, \mathbf{v}_h) = a(\mathbf{u}_h, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_k^h \tag{5.12}$$

where the bilinear form $b_k : \mathbf{V}_k^h \times \mathbf{V}_k^h \to R$. As above discussion, for Neumann boundary problem, $\hat{P}_k \mathbf{u}_h$ for $k = 1, \cdots, N$ can be uniquely determined but $\hat{P}_0 \mathbf{u}_h$ should carefully be chosen from a set of solutions when $\mathbf{V}_0^h = \mathbf{V}^H$. Let $\mathbf{u}_H \in \tilde{\mathbf{V}}^H$ be the unique solution of equation

$$b_k(\mathbf{u}_H, \mathbf{v}_H) = a(\mathbf{u}_h, \mathbf{v}_H) \qquad \forall \mathbf{v}_H \in \mathbf{V}_0^h.$$

Assume $\hat{P}_0\mathbf{u}_h = \mathbf{u}_H + \mathbf{w}_H$ where $\mathbf{w}_H \in KerA_H$ makes $\sum_{k=0}^N \hat{P}_k\mathbf{u}_h$ orthogonal to $KerA_h$. Note that $\hat{P}_k\mathbf{u}_h = 0, \quad \forall \mathbf{u}_h \in KerA_h$ for all $k = 0, 1, \cdots, N$. Therefore, we can replace the projection operator P_k by $\hat{P}_k$ and well define a new projection operator equation:

$$\hat{P}\mathbf{u} = (\hat{P}_0 + \hat{P}_1 + \cdots + \hat{P}_N)\mathbf{u}_h = \hat{\mathbf{g}}$$

where $\hat{\mathbf{g}} = \sum_{k=0}^N \hat{\mathbf{g}}_k$ and $\hat{\mathbf{g}}_k = \hat{P}_k\mathbf{u}_h$. For the same reason as before, we should estimate

$$\lambda_0 a(\mathbf{v}_h, \mathbf{v}_h) \leq a(\hat{P}\mathbf{v}_h, \mathbf{v}_h) \leq \lambda_1 a(\mathbf{v}_h, \mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \tilde{\mathbf{V}}^h,$$

in order to get estimation of condition number κ of operator $\hat{P}$. A lower bound for the eigenvalues of $\hat{P}$ can be easily estimated from following useful lemma which is the extension of Lions' lemma [49] and is developed by Dryja and Widlund [35]. Since the proof is quite short, it is included.

Lemma 5.3 *If there exists a positive constant $C_0 > 0$ such that for all $\boldsymbol{v}_h \in \tilde{\boldsymbol{V}}^h(\Omega)$, there exists a decomposition $\boldsymbol{v}_h = \sum_{k=0}^N \boldsymbol{v}_h^k$ where $\boldsymbol{v}_h^k \in \boldsymbol{V}_k^h$ for $k = 0, 1, \cdots, N$ such that*

$$\sum_{k=0}^N b_k(\boldsymbol{v}_h^k, \boldsymbol{v}_h^k) \leq C_0^2 a(\boldsymbol{v}_h, \boldsymbol{v}_h).$$

Then $\lambda_0 \geq C_0^{-2}$. and

$$a(\hat{P}\boldsymbol{u}_h, \boldsymbol{u}_h) \geq C_0^{-2} a(\boldsymbol{u}_h, \boldsymbol{u}_h).$$

Proof *For any $\boldsymbol{u}_h \in \tilde{\boldsymbol{V}}^h$, we have*

$$a(\boldsymbol{u}_h, \boldsymbol{u}_h) = \sum_{k=0}^N a(\boldsymbol{u}_h, \boldsymbol{u}_h^k) = \sum_{k=0}^N b_k(\hat{P}_k\boldsymbol{u}_h, \boldsymbol{u}_h^k).$$

Therefore, by Cauchy inequality, it follows that

$$a(\boldsymbol{u}_h, \boldsymbol{u}_h) \leq (\sum_{k=0}^{N} b_k(\hat{P}_k \boldsymbol{u}_h, \hat{P}_k \boldsymbol{u}_h))^{1/2} (\sum_{k=0}^{N} b_k(\boldsymbol{u}_h^k, \boldsymbol{u}_h^k))^{1/2}.$$

By the assumption of the lemma,

$$a(\boldsymbol{u}_h, \boldsymbol{u}_h) \leq C_0^2 \sum_{k=0}^{N} b_k(\hat{P}_k \boldsymbol{u}_h, \hat{P}_k \boldsymbol{u}_h) = C_0^2 \sum_{k=0}^{N} a(\hat{P}_k \boldsymbol{u}_h, \boldsymbol{u}_h) = C_0^2 a(\hat{P} \boldsymbol{u}_h, \boldsymbol{u}_h),$$

and the lemma is established.

The following lemma, presented in [35], is satisfied for Neumann boundary problem. This lemma gives an upper bound of eigenvalue of operator $\hat{P}$.

Lemma 5.4 *Assume that*

1. *there exists a constant ω such that for $k = 0, \cdots, N$,*

$$a(\boldsymbol{u}_h, \boldsymbol{u}_h) \leq \omega b_k(\boldsymbol{u}_h, \boldsymbol{u}_h) \qquad \forall \boldsymbol{u}_h \in \boldsymbol{V}_k^h;$$

2. *there exist constants ε_{ij}, for $i, j = 1, \cdots, N$ such that*

$$a(\boldsymbol{u}_h^{(i)}, \boldsymbol{u}_h^{(j)}) \leq \varepsilon_{ij} a(\boldsymbol{u}_h^{(i)}, \boldsymbol{u}_h^{i})^{1/2} a(\boldsymbol{u}_h^{(j)}, \boldsymbol{u}_h^{j})^{1/2}, \quad \forall \boldsymbol{u}_h^{(i)} \in \boldsymbol{V}_i^h \quad \forall \boldsymbol{u}_h^{(j)} \in \boldsymbol{V}_j^h.$$

Then

$$a(\hat{P} \boldsymbol{u}_h, \boldsymbol{u}_h) \leq (\rho(\varepsilon) + 1)\omega a(\boldsymbol{u}_h, \boldsymbol{u}_h), \qquad \forall \boldsymbol{u}_h \in \tilde{\boldsymbol{V}}^h(\Omega).$$

where $\rho(\omega)$ is the spectral radius of the matrix $\varepsilon = \{\varepsilon_{ij}\}_{ij=1}^{N}$. Hence, $\lambda_{\max}(\hat{P}) \leq (\rho(\varepsilon) + 1)\omega$.

5.1.2 Elliptic and Elasticity problem

As a model problem for the second order elliptic equation with normal derivative boundary condition, we consider:

$$(5.13)\qquad Lu = f \quad \text{in } \Omega \quad \text{and} \quad \tfrac{\partial u}{\partial N} = 0 \quad \text{on } \partial\Omega$$

where

$$Lv = -\sum_{i,j} \frac{\partial}{\partial x_i}(a_{ij}\frac{\partial v}{\partial x_j})$$
$$\frac{\partial v}{\partial N} = \sum_{i,j} a_{ij}(x)\frac{\partial v}{\partial x_j}\cos(\vec{n},\vec{e}_i)$$

with a_{ij} uniformly positive definite, bounded and piecewise smooth on external normal direction of boundary of Ω and $\vec{e}_i$ is the unit direction of the ith axis. In the Sobolev space $H^1(\Omega)$, this equation can be written into the weak variational form as problem (5.3) with

$$(5.14)\qquad a(u,v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}dx, \quad \text{and} \quad (f,v) = \int_\Omega fv dx.$$

The kernel space of this problem is a subspace with only $\{1\}$ as its basis:

$$KerA = KerA_h = KerA_H = \{1\}.$$

Hence, step 2 and 3 in the process of calculating $P_0 u_h$ become:

2 Find $u_H \in V^H$ such that

$$\int_\Omega u_H dx = 0 \quad \text{and} \quad a(u_H, v_H) = (f, v_H), \quad \forall v_H \in V^H;$$

3 Compute the mean value of $w_h = u_H + \sum_{k=1}^{N} P_K u_h$ and let $\bar{w} = -\frac{1}{|\Omega|}\int_\Omega w_h dx$. We then obtain $Pu_h = \sum_{k=0}^{N} P_k u_h$ by defining $P_0 u_h = \bar{w} + u_H$.

Now we consider the model elasticity problem with Neumann boundary condition:

$$(5.15) \quad \begin{cases} \mu\nabla\vec{u} + (\lambda+\mu)\text{grad}(\text{div}\vec{u}) = \vec{f} & \text{in} \quad \Omega \\ B_i\vec{u} \equiv \sum_{k=1}^{3} \mu\cos(\vec{n},\vec{e_k})2\xi_{ik}(\vec{u}) + \lambda\cos(\vec{n},\vec{e_k})\text{div}\vec{u} = 0 & \text{on} \quad \partial\Omega \\ \qquad\qquad \text{for } i = 1,2,3 & \end{cases}$$

where $\xi_{ik}(\vec{v}) = \frac{1}{2}(\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i})$, and λ and μ are positive constants. In the Sobolev space $\mathbf{V} = (H^1(\Omega))^3$, we form an equivalent variational equation as problem (5.3) with

$$a(\vec{u},\vec{v}) = \int_\Omega \{\lambda\text{div}(\vec{u})\text{div}(\vec{v}) + 2\mu\sum_{i,j}\xi_{ij}(\vec{u})\xi_{ij}(\vec{v})\}dx$$

and $(\vec{f},\vec{v}) = \int_\Omega(\vec{f}\cdot\vec{v})dx$. The kernel space of this elasticity problem is

$$KerA = \{\vec{u}|\vec{u} = \vec{a} + \vec{b}\times\vec{r}, \qquad \text{with } \vec{r} = (x_1,x_2,x_3)^T\}$$

where $\vec{a}$ and $\vec{b}$ are constant vectors in R^3. The basis functions of this kernel space are

$$(5.16) \quad \begin{array}{lll} \vec{e}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, & \vec{e}^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, & \vec{e}^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ \vec{e}^{(4)} = \begin{pmatrix} 0 \\ -z \\ y \end{pmatrix}, & \vec{e}^{(5)} = \begin{pmatrix} z \\ 0 \\ -x \end{pmatrix}, & \vec{e}^{(6)} = \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}. \end{array}$$

Since these six functions are linear vector functions, it follows that $\vec{e}^{(i)} \in \mathbf{V}^h \cap \mathbf{V}^H$ for $i = 1, \cdots, 6$ and $KerA = KerA_h = KerA_H$ is the subspace with these six functions $\{\vec{e}^{(i)}\}_{i=1}^{6}$ as its basis. Hence, the concrete computation of $P\vec{u}_h$ is to do step 1 described in the previous section with following step 2 and step 3:

2 Find $\vec{u_H} \in \mathbf{V}^H$ such that

$$\vec{u_H} \perp KerA_H \quad \text{and} \quad a(\vec{u}_H, \vec{v}_H) = (\vec{f}, \vec{v}_H) \qquad \forall \vec{v}_H \in \mathbf{V}^H;$$

3 Let $\vec{w}_h = (w_1, w_2, w_3)^T = \vec{u}_H + \sum_{k=1}^{N} P_k \vec{u}_h$ and find α_i such that $\vec{w}_h + \sum_{i=1}^{6} \alpha_i \vec{e}^{(i)} \perp KerA_h$. These α_i can be determined by solving following small symmetric positive definite system:

$$O\vec{\alpha} = \vec{\beta} \tag{5.17}$$

where

$$O = \begin{bmatrix} |\Omega| & 0 & 0 & 0 & \int_\Omega x_3 & -\int_\Omega x_2 \\ 0 & |\Omega| & 0 & -\int_\Omega x_3 & 0 & \int_\Omega x_1 \\ 0 & 0 & |\Omega| & \int_\Omega x_2 & -\int_\Omega x_1 & 0 \\ 0 & -\int x_3 & \int x_2 & \int(x_3^2 + x_2^2) & -\int x_1 x_2 & -\int x_1 x_3 \\ \int x_3 & 0 & -\int x_1 & -\int x_1 x_2 & \int(x_3^2 + x_1^2) & -\int x_2 x_3 \\ -\int x_2 & \int x_1 & 0 & -\int x_1 x_3 & -\int x_2 x_3 & \int(x_1^2 + x_2^2) \end{bmatrix}$$

is symmetric and positive definite, and

$$\vec{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} \qquad \vec{\beta} = \begin{pmatrix} -\int_\Omega w_1 \\ -\int_\Omega w_2 \\ -\int_\Omega w_3 \\ -\int_\Omega (x_2 w_3 - x_3 w_2) \\ -\int_\Omega (x_3 w_1 - x_1 w_3) \\ -\int_\Omega (x_1 w_2 - x_2 w_1) \end{pmatrix}.$$

5.2 The Modified Domain Decomposition Method

In this section, we discuss concrete preconditioners for the elliptic scalar problem with Neumann boundary condition. We first describe how to construct a preconditioner from additive Schwarz method. Then, we derive Schur complement S for the unknowns on the boundaries Γ by using block Gauss elimination method and discuss its properties. For this Schur complement system, we extend the vertex space domain decomposition method. Finally, we give theoretical results which show that the condition number of the modified VS method is still independent of grid sizes H and h .

5.2.1 The Modified Additive Schwarz Method

Each substructure Ω_k is extended into a larger region $\hat{\Omega}_k$ so that $\hat{\Omega}_k$ has overlap stripe only with its neighbor substructures. We assume that the overlap stripe has a generous width of $\delta = 0(H)$ and $\partial\hat{\Omega}_k$ does not cut through any element. Let $\mathbf{V}_k^h = \mathbf{V}^h(\Omega) \cap (\hat{H}_0^1(\hat{\Omega}_k))^q$ for $k = 1, \cdots, N$, and $\mathbf{V}_0^h = \mathbf{V}^H(\Omega)$ where

$$\hat{H}_0^1(\hat{\Omega}_k) = \{u | u \in H^1(\Omega), \text{ and } u(x) = 0 \text{ if } x \notin \hat{\Omega}_k\}.$$

Then, the discrete space $\mathbf{V}^h$ is decomposed into the sum of N+1 subspaces:

$$\mathbf{V}^h = \mathbf{V}_0^h + \mathbf{V}_1^h + \cdots + \mathbf{V}_N^h.$$

On each substructure $\hat{\Omega}_k$, we introduce $R_k : \mathbf{V}^h(\Omega) \to \mathbf{V}_k^h$ as the pointwise restriction operator which returns only those unknowns which are associated with $\hat{\Omega}_k$. Then, the projection P_k from $\mathbf{V}^h$ into $\mathbf{V}_k^h$ can be defined by

$$P_k = R_k^T A_k^{-1} R_k A \qquad \text{for } k = 1, \cdots, N$$

where $A_k = R_k A R_k^T$. It is obvious that A_k is a symmetric positive definite matrix. Furthermore, the computation of $A_k^{-1}\mathbf{f}$ corresponds to solving the elliptic problem restricted on sub-domain $\hat{\Omega}_k$ with Dirichlet boundary condition on $\partial\hat{\Omega}_k$. Let R_0^T to be denoted linear interpolation from $\mathbf{V}^H \to V^h$. This definition of R_0^T ensures the important equation

$$R_0^T(KerA_H) = KerA_h$$

which makes it possible to extend the additive Schwarz method for Neumann boundary problem. We can define the coarse projection operator by $P_0 : \mathbf{V}^h \to \mathbf{V}^H$

$$P_0 = R_0^T A_H^{-1} R_0 A = R_0^T A_0^{-1} R_0 A \quad \text{and} \quad A_0 = A_H$$

where A_H is the stiffness matrix resulted from discrete variational problem (5.4) through treating the substructures $\{\Omega_k\}$ as elements in finite element scheme. Like the stiffness matrix A defined by equation (5.5), A_H is symmetric and semi-positive definite. Therefore, we will meet following problems whether the right hand side of the coarse problem is orthogonal to $KerA_H$ and which solution in $\mathbf{V}^H$ is suitable for our problem. Since in general $KerA_h = KerA_H$ is satisfied for Neumann boundary problem, the coarse grid problem can be well defined through choosing suitable restriction R_0 such that $R_0 f \perp KerA_H$ if $f \perp KerA_H$. Assume $KerA_h = KerA_H$ and the inverse of each A_k is well defined. By writing the operator P as

$$P = M^{-1}A = (\sum_{k=0}^{N} R_k^T A_k^{-1} R_k)A,$$

we can easily extract the preconditioner

$$M^{-1} = \sum_{k=0}^{N} R_k^T A_k^{-1} R_k$$

from this expression for Neumann boundary problem. We summarize the process of Preconditioned Conjugate Gradient method by only outlining the steps for performing the preconditioner $M^{-1}f$ since the standard procedure of Conjugate Gradient method can be easily found [40]:

Find x such that $Mx = f$ where f is orthogonal to $kerA_h$.

Additive Schwarz Preconditioner

1. Solve each subproblem on each sub-domain $\hat{\Omega}_k$

$$A_k x^{(k)} = R_k f \qquad \text{for } k = 1, \cdots, N;$$

2. Find $x^{(0)} \perp KerA_H$ so that :

$$A_H x^{(0)} = R_0 f;$$

3. Let

$$w_h = R_0^T x^{(0)} + \sum_{k=1}^{N} R_k x^{(k)}$$

and find $\bar{w}_h \in KerA_h$ such that $\bar{w}_h + w_h$ is orthogonal to $KerA_h$. Then

$$M^{-1} f = \bar{w}_h + R_0^T x^{(0)} + \sum_{i=1}^{N} R_i x^{(i)}$$

is obtained.

Remark: Step 1 and 2 can be performed simultaneously. Since the linear interpolation R_0^T maps linear function into linear function and $KerA_h = KerA_H$ consists of linear functions, $R_0 f$ is orthogonal to $KerA_H$ if $f \perp KerA_h$. Hence, the coarse grid problem is well defined in the Additive Schwarz Preconditioner. Like step 3 of computing $P\mathbf{u}_h$ process, step 3 in the Additive Schwarz Preconditioner calculates $\bar{w}_h \in KerA_h$ and adds $\bar{w}_h$ to w_h so that $M^{-1} f \perp KerA_h$. For elliptic problem (5.13), $-\bar{w}_h$ is the mean value of w_h on the whole fine grid. For elasticity problem (5.15), $\bar{w}_h = \sum_{i=1}^{6} \alpha^{(i)} \vec{e}^{(i)} \in kerA_h$ where $\alpha^{(i)}$ is the solution of linear system (5.17).

We give following theorem which shows that the condition number of this modified additive Schwarz method is still independent of the sizes of coarse grid and fine grid.

Theorem 5.1 *For all* $\mathbf{u}_h \in \tilde{V}^h(\Omega)$, *there exists a partition* $\mathbf{u}_{h,k} \in V_k^h$ *for all*

$0 \leq k \leq N$ such that

$$\boldsymbol{u}_h = \sum_{k=0}^{N} \boldsymbol{u}_{h,k} \quad \textit{and} \quad \sum_{k=0}^{N} |u_{h,k}|^2_{(H^1(\Omega))^q} \leq C_0^2 |u_h|^2_{(H^1(\Omega))^q},$$

where C_0 is independent of $\boldsymbol{u}_h, h$ and H. Therefore,

$$\lambda_{min} \geq \frac{1}{C_0^2} \quad \textit{and} \quad \lambda_{max} \leq 5.$$

Hence, $\kappa(P) \leq \lambda_{max}/\lambda_{min} = 5/C_0^2$.

Following theorem refines above result.

Theorem 5.2 *Suppose the overlap size of each region $\hat{\Omega}_k$ is $\beta_k H$. Then for all $\boldsymbol{u}_h \in \tilde{\boldsymbol{V}}^h(\Omega)$, there exists a partition $\boldsymbol{u}_{h,k} \in \boldsymbol{V}_k^h$ for all $0 \leq k \leq N$ such that*

$$\boldsymbol{u}_h = \sum_{k=0}^{N} \boldsymbol{u}_{h,k} \quad \textit{and} \quad \sum_{k} |\boldsymbol{u}_{h,k}|^2_{(H^1(\Omega))^q} \leq C(1 + \max_k \frac{C}{\beta_k}) |\boldsymbol{u}_h|^2_{(H^1(\Omega))^q},$$

where C is independent of u^h, h, H and overlap size.

5.2.2 Schur Complement on Γ

In this section, we derive the Schur complement on Γ and analyze its properties for the Neumann boundary problem. Let $\mathbf{u}_H$ and $\mathbf{u}_h$ be expanded in terms of the standard nodal basis functions of discrete spaces $\mathbf{V}^H(\Omega)$ and $\mathbf{V}^h$ respectively. Denote the corresponding kernel spaces as $KerA_H = \{x_H | A_H x_H = 0\}$ and $kerA =$

$\{x|Ax = 0\}$ where A_H and A are stiff matrices. Then the discrete problem (5.9) and (5.5) can be rewritten as the following linear systems respectively:

$$\text{Find } x_H \perp KerA_H \quad \text{such that} \quad A_H x_H = f_H, \tag{5.18}$$

and

$$\text{Find } x \perp KerA \quad \text{such that} \quad Ax = f, \tag{5.19}$$

where stiffness matrix A can be written as block matrix according to interior points and edge points:

$$A = \begin{pmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{pmatrix} \qquad x = \begin{pmatrix} x_I \\ x_B \end{pmatrix} \qquad f = \begin{pmatrix} f_I \\ f_B \end{pmatrix}$$

and the unknowns x are split into two subgroups. The first subgroup x_I is the set of unknowns corresponding to the interior points in $\cup_{k=0}^{N}\Omega_k$ and the second subgroup x_b corresponds to the variables on the boundaries $\Gamma = \cup_k \partial\Omega_k$. f is orthogonal to the space $KerA$. Let $A^{(k)}$ be the stiffness matrix of the bilinear $a_{\Omega_k}(u_h, v_h)$ which represents the contribution of the substructure Ω_k to the integral $a_\Omega(u_h, v_h)$. Then the entire matrix A can be obtained by using the method of sub-assembly:

$$\begin{pmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{BB} \end{pmatrix} \begin{pmatrix} x_I \\ x_B \end{pmatrix} = \sum_k \begin{pmatrix} A_{II}^{(k)} & A_{IB}^{(k)} \\ A_{BI}^{(k)} & A_{BB}^{(k)} \end{pmatrix} \begin{pmatrix} x_I^{(k)} \\ x_B^{(k)} \end{pmatrix} \tag{5.20}$$

$$x^{(k)} = \begin{pmatrix} x_I^{(k)} \\ x_B^{(k)} \end{pmatrix} \qquad x = \sum_k x^{(k)}.$$

Here $x^{(k)}$ is the subgroup of nodal parameters associated with $\bar{\Omega}_k$, $x_I^{(k)}$ is the set of unknowns corresponding to the interior variables of substructure Ω_k and $x_B^{(k)}$ is associated with the nodal points of $\partial\Omega_k$. Because each interior variable $x_I^{(k)}$ is associated with only one of the substructures, it can be eliminated locally and simultaneously. The reduced global equation, called Schur complement on Γ, can be written in assembly:

$$Sx_B = (A_{BB} - A_{BI}A_{II}^{-1}A_{IB})x_B = \sum_k S^{(k)}x_B^{(k)} = g \tag{5.21}$$

where

$$S^{(k)} = A_{BB}^{(i)} - A_{BI}^{(k)}{A_{II}^{(k)}}^{-1}A_{IB}^{(k)},$$

corresponds to the contribution from Ω_k to the boundary $\partial\Omega_k \subset \Gamma$, and

$$g = \sum_k g^{(k)} = \sum_k f_B^{(k)} - A_{BI}^{(k)}{A_{II}^{(k)}}^{-1}f_I^{(k)},$$

comes from the value of f on interior points Ω_k and $\partial\Omega_k$. As we know, the action of inverse of $A_{II}^{(k)}$ is associated with solving a local problem on Ω_k with Dirichlet boundary condition and small size. A fast Dirichlet solver can be used to implement the inverse action of $A_{II}^{(k)}$. Note the reduced Schur complement is still singular i.e.

$$KerS = \{x_B | Sx_B = 0\} \neq \{0\}.$$

Hence, we could find an unique solution $x_B \perp KerS$ of problem (5.21) only if g satisfies compatible condition $g \perp KerS$. To deal with this difficulty for this Schur complement system, we rely on the following properties of capacitance matrix S.

Lemma 5.5 *If element $x = (x_I^T, x_B^T)^T \in KerA$, then $x_B \in KerS$. If $x_B \in KerS$, there exists x_I such that $x = (x_I^T, x_B^T)^T \in KerA$.*

Lemma 5.6 *If f is orthogonal to $KerA$, then $g = f_B - A_{BI}A_{II}^{-1}f_I$ is orthogonal to $KerS$.*

The proof of these two lemmas is quite easy [2] so we omit it here. We have following simple lemmas.

Lemma 5.7 *If A is a symmetric semi-positive matrix, then for any x_B*

$$x_B^T S x_B = \min_{x_I} x^T A x$$

where $x = (x_I^T, x_B^T)^T$. Hence, the capacitance matrix S is also symmetric semi-positive definite.

Lemma 5.8 *For any x_B, let $x = (x_I^T, x_B^T)^T$ be discrete harmonic extension of x_B, i.e. $(A_{II} \quad A_{IB})x = 0$. Then $x_B^T S x_B = x^T A x$.*

From these lemmas, we can conclude that the direct restriction of $KerA$ to the pseudo-boundary Γ equals $KerS$. We can roughly think that $KerS = KerA$. Therefore, for elliptic problem (5.13), $KerS = \{c | c \text{ is any constant on } \Gamma\}$. For elasticity problem, $KerS$ is the linear space on Γ with six basis functions defined in (5.16).

5.2.3 Description of the Modified VS Algorithm

To describe the modified VS algorithm for the Schur complement system on the boundaries Γ, we need work on the discrete trace subspace $V(\Gamma)$ of trace space $H^{1/2}(\Gamma)$ and introduce some notations. Let Γ be partitioned into three pieces: the substructure vertices $\{v_l\}$, the edges between substructure vertices and the faces $\Gamma^{F_{ij}}$ of the substructures. Note in two dimensions there are no faces, only vertices and edges. To obtain overlapping regions, we define $\Gamma^{E_{ij}}$ as the regions consisting of an edge and an overlap of order H onto adjacent faces and Γ^{v_l} as the regions consisting of a vertex and an overlap of order H onto adjacent faces and edges. We restrict the overlap so that no portion of Γ is covered more than p times, usually $p = 4$. We can now introduce the sum representation of discrete trace space $V^h(\Gamma)$ as follows :

$$V^h(\Gamma) = \sum V^h_{F_{ij}} + \sum V^h_{E_{ij}} + \sum V^h_{v_l},$$

where $V^h_{F_{ij}} = V^h(\Gamma) \cap H^{1/2}_{00}(\Gamma^{F_{ij}})$, $V^h_{E_{ij}} = V^h(\Gamma) \cap H^{1/2}_{00}(\Gamma^{E_{ij}})$ and $V^h_{v_l} = V^h(\Gamma) \cap H^{1/2}_{00}(\Gamma^{v_l})$. Here $H^{1/2}_{00}(\hat{\Gamma})$ with $\hat{\Gamma} \subset \Gamma$, is denoted as the space of functions $v \in H^{1/2}(\Gamma)$ which vanish at all nodal points not on $\hat{\Gamma}$.

For each sub-region $\hat{\Gamma}$, we introduce $R_{\hat{\Gamma}}$ as the pointwise restriction operator which returns only those unknowns which are associated with $\hat{\Gamma}$. The projection operator $P_{\hat{\Gamma}}$ onto each face, vertex or edge subspace $V^h(\hat{\Gamma})$ is defined by

$$P_{\hat{\Gamma}} = R_{\hat{\Gamma}}^T S_{\hat{\Gamma}}^{-1} R_{\hat{\Gamma}} S$$

where $S_{\hat{\Gamma}} = R_{\hat{\Gamma}} S R_{\hat{\Gamma}}^T$.

For the coarse space V^H, let R_H be the weighted restriction operator from $V^h(\Gamma)$ to V^H. Then R_H^T is the corresponding linear interpolation operator from V^H to $V^h(\Gamma)$. The criteria of choosing the restriction operator R_H is that for any function $g \in \tilde{V}^h(\Gamma)$ i.e. $g \perp KerS$, then $R_H g$ is orthogonal to the space $KerA_H$. For the Neumann boundary problem in two dimension, we choose the restriction operator R_H as follows: Let $\Psi_1, \Psi_2, ... \Psi_M$ be the piecewise linear basis functions of coarse space V^H, where M is the number of vertices on Γ and $\Psi_k(v_l) = \delta_{kl}$ for any vertex point v_l where δ_{lk} equals one if $l = k$ and zero otherwise. Then,

$$R_H g(v_l) = \sum_{(x_i, y_j) \in \Gamma} \Psi_l(x_i, y_j) g(x_i, y_j).$$

A straightforward computation gives that $R_H g$ is orthogonal to the kernel space $KerA_H$ when g is orthogonal to the kernel space $KerS$, because $KerS$ consists of linear functions for elliptic problem (5.13) or elasticity problem (5.15). Therefore, the coarse problem $A_H x_H = R_H g$ is well defined and has only one solution $x_H \perp KerA_H$. For Schur complement system (5.21), we solve it by using preconditioned conjugate gradient iterative method with preconditioner M. The action of the inverse of the preconditioner M involves following block calculations.

VS Preconditioner (Calculate $M^{-1}g$)

1. Solve subproblems on faces $\Gamma^{F_{ij}}$

$$S_{F_{ij}} x_{F_{ij}} = R_{F_{ij}} g;$$

2. Solve subproblems on edges $\Gamma^{E_{ij}}$:

$$S_{E_{ij}} x_{E_{ij}} = R_{E_{ij}} g;$$

3. Solve subproblems on the vertex space Γ^{v_l}

$$S_{v_l} x_{v_l} = R_{v_l} g;$$

4. Find $x_H \perp Ker A_H$ so that:

$$A_H x_H = R_H g;$$

5. Let

$$w = R_H^T x_H + \sum_{ij} R_{F_{ij}}^T x_{F_{ij}} + \sum_{ij} R_{E_{ij}}^T x_{E_{ij}} + \sum_{l} R_{v_l}^T x_{v_l},$$

and find $\bar{w} \in KerS$ such that $\bar{w} + w$ is orthogonal to $KerS$. Then

$$M^{-1} g = \bar{w} + R_H^T x_H + \sum_{ij} R_{F_{ij}}^T x_{F_{ij}} + \sum_{ij} R_{E_{ij}}^T x_{E_{ij}} + \sum_{l} R_{v_l}^T x_{v_l}.$$

Remark: Step 1, 2, 3 and 4 can be calculated simultaneously. This preconditioning procedure can be rewritten in a short form:

$$\begin{aligned}(5.22) \quad M^{-1} g \; = \; & R_H^T A_H^{-1} R_H g + \bar{w} + \\ & \sum_{ij} R_{F_{ij}}^T S_{F_{ij}}^{-1} R_{F_{ij}} g + \sum_{ij} R_{E_{ij}}^T S_{E_{ij}}^{-1} R_{E_{ij}} g + \sum_{l} R_{v_l}^T S_{v_l}^{-1} R_{v_l} g\end{aligned}$$

where $\bar{w} \in KerS$ is determined by making $M^{-1}g$ be orthogonal to the kernel space $KerS$.

After obtaining the approximation solution on the boundaries Γ through using PCG iterative method, we can calculate the approximate solution of problem (5.4) on the whole domain by solving concurrently all the Dirichlet boundary subproblems defined on the substructures $\{\Omega_k\}$ with boundary value u_h on $\partial\Omega_k \subset \Gamma$:

$$a(u_h^k, v_h^k) = (f, v_h^k) \qquad \forall v_h^k \in V_0^h(\Omega_k),$$

where $V_0^h(\Omega_k) = \{v_h | v_h \in V_k^h, \quad v_h|_\Gamma = 0\}$. However, such extension solution u_h, defined on the whole physical domain Ω, is not orthogonal to the kernel space $KerA_h$. Therefore, we have to find $w_h \in KerA_h$ such that the approximation solution $u_h + w_h$ is orthogonal to $\perp KerA_h$. The following theorem gives an estimation of condition number κ of the modified vertex space method. The proof of this theorem is similar to that in [66].

Theorem 5.3 *Suppose the overlap size of each vertex region Γ^{v_k} is βH, then:*

$$\kappa(P) = \kappa(M^{-1}S) \leq \frac{\lambda_{\max}(M^{-1}S)}{\lambda_{\min}(M^{-1}S)} \leq C(\beta),$$

where $C(\beta)$ is a function of β which is independent of H and h.

A concrete estimation of upper bound $C(\beta)$ of condition number is given in the follows.

Theorem 5.4 *Let the overlap size of each vertex region Γ^{v_l} is $\beta_l H$, then:*

$$\kappa(P) = \kappa(M^{-1}S) \leq \frac{\lambda_{\max}(M^{-1}S)}{\lambda_{\min}(M^{-1}S)} \leq C(1 + C \max_l \frac{1}{\beta_l}),$$

where C is independent of H and h.

5.3 The Variants of the Vertex Space Method

As described before, the application of the modified vertex space method for scale elliptic problem in two dimensions involves forming exact edge matrices $S_{E_{ij}}$ and vertex matrices S_{v_k}. The computation of these matrices associated with $\partial\Omega_k$ can be very expensive since it requires to solve n_k subproblems on Ω_k if there are n_k nodes along the $\partial\Omega_k$. Big memory size is also required for storing all these dense matrices $S_{E_{ij}}$ and S_{v_k}. This expense on computation and storage can be significantly reduced if these exact edge and vertex matrices are replaced by approximations which can be computed or inverted at much less cost. If these approximations are spectrally equivalent to the exact sub-matrices, then the overall new preconditioner would remain spectrally equivalent to the exact VS preconditioner. Hence, the condition number of these variants of vertex space method is still independent of mesh size. In this section, we are looking for the spectrally equivalent edge matrices $\tilde{S}_{E_{ij}}$ and $\tilde{S}_{v_k}$ such that

$$cx^T_{E_{ij}} S_{E_{ij}} x_{E_{ij}} \leq x^T_{E_{ij}} \tilde{S}_{E_{ij}} x_{E_{ij}} \leq C x^T_{E_{ij}} S_{E_{ij}} x_{E_{ij}},$$

and

$$c x_{v_k}^T S_{v_k} x_{v_k} \leq x_{v_k}^T \tilde{S}_{v_k} x_{v_k} \leq C x_{v_k}^T S_{v_k} x_{v_k}$$

with positive constants c and C independent of h and H. Then the corresponding preconditioner $\tilde{M}$ satisfies that

$$c \leq \frac{\lambda_{max}(\tilde{M}^{-1}S)}{\lambda_{min}(\tilde{M}^{-1}S)} \leq C,$$

if the exact preconditioner M satisfies

$$c \leq \frac{\lambda_{max}(M^{-1}S)}{\lambda_{min}(M^{-1}S)} \leq C.$$

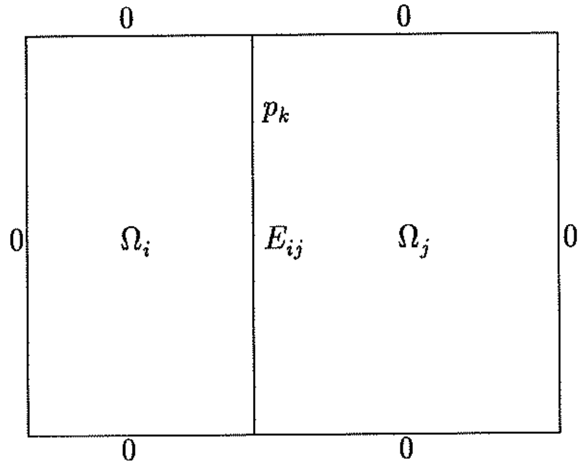
Based on the Fourier approximation or probe technique, several approximate edge and vertex matrices have been discussed in [23, 21]. For simplicity, we don't consider the combination of these two approximation methods.

In this section, we are going to discuss using 4 probe vectors instead of 6 probe vectors [23] to construct the vertex and edge matrices which would save much computation work in forming these approximate matrices. The Fourier approximation described in Chapter 2 can be directly applied here. For vertex space matrices S_{v_k}, we discuss the two other approximations: tangential and scale diagonal approximations.

5.3.1 The Symmetric Probe Technique

In this section, we describe how to use four probe vectors instead of six probe vectors [23] in the modified VS method to obtain the approximate edge matrices

Figure 5.1: The Model Rectangular Geometry With Boundary Value



and vertex matrices in order to save computational work. The probe technique has been extensively discussed by Chan and Resasco [25], Keyes and Gropp [45, 46], Chan and Mathew [22, 21], and Eicenstate [36]. In [23], the probe technique was applied to construct approximate edge and vertex matrices through using six probe vectors. Here, we only use four probe vectors to form all approximate edge and vertex matrices. An advantage of these approximations is that they often adapt well to coefficient variations and aspect ratios. However, a disadvantage is that the construction of probe approximation matrices costs much higher than that of Fourier method, but still considerably less than that of exact sub-matrices.

For simplicity, we first describe the procedure of construction of these probe approximations for the model rectangular geometry as Fig. 5.1 . The technique can be easily extended to more general geometries.

On this sub-region with two sub-domain Ω_i and Ω_j, we construct a symmetric tridiagonal approximation $\tilde{S}_{E_{ij}}$ to the exact Schur complement $S_{E_{ij}}$ by using matrix vector products of $S_{E_{ij}}$ with two probe vectors. A heuristic motivation for using the tridiagonal approximations $\tilde{S}_{E_{ij}}$ to the edge matrices $S_{E_{ij}}$ is that the entries of each $S_{E_{ij}}$ decay rapidly away from the diagonal. For model problems and geometries, it has been shown that

$$(S_{E_{ij}})_{lm} = 0(\frac{1}{(l-m)^2})$$

for l, m away from the diagonal ,[39].

To obtain a symmetric tridiagonal approximation $\tilde{S}_{E_{ij}}$ to $S_{E_{ij}}$, we equate the matrix vector products $S_{E_{ij}}p_k$ to $\tilde{S}_{E_{ij}}p_k$ for the following two probe vectors p_k:

$$p_1 = [1, 0, 1, 0, \cdots]^T \quad \text{and} \quad p_2 = [0, 1, 0, 1, \cdots]^T.$$

To make the description simple, we assume that the tridiagonal matrix $\tilde{S}_{E_{ij}}$ can be written as

$$\tilde{S}_{E_{ij}} = \begin{bmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & b_3 & a_4 & \ddots \\ & & & \ddots & \ddots \end{bmatrix}_{n_{ij} \times n_{ij}} .$$

The matrix vector products $\tilde{S}_{E_{ij}}p_1, \tilde{S}_{E_{ij}}p_2$ can be written as :

$$\begin{bmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & b_2 & a_3 & b_3 & \\ & & b_3 & a_4 & \ddots \\ & & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ b_1 + b_2 & a_2 \\ a_3 & b_2 + b_3 \\ b_3 + b_4 & a_4 \\ \vdots & \vdots \end{bmatrix}.$$

Furthermore, computing the matrix vector product $S_{E_{ij}}p_k$ requires solving one problem on each sub-domain Ω_i and Ω_j, see fig 5.1. Hence, the tridiagonal approximation $\tilde{S}_{E_{ij}}$ can be obtained from the equation

$$[S_{E_{ij}}p_1, S_{E_{ij}}p_2] = [\tilde{S}_{E_{ij}}p_1, \tilde{S}_{E_{ij}}p_2]$$

through using following algorithm, proposed in [45, 46]:

Symmetric Probe Algorithm

```
For    $l = 1, \cdots, n_{ij}$
    $a_l = \begin{cases} (S_{E_{ij}}p_1)_l & \text{if } l \text{ is odd} \\ (S_{E_{ij}}p_2)_l & \text{if } l \text{ is even} \end{cases}$
$b_1 = (S_{E_{ij}}p_2)_1$
For    $l = 2, \cdots, n_{ij} - 1$
    $b_l = \begin{cases} (S_{E_{ij}}p_1)_l - b_{l-1} & \text{if } l \text{ is even} \\ (S_{E_{ij}}p_2)_l - b_{l-1} & \text{if } l \text{ is odd .} \end{cases}$
```

Because $S_{E_{ij}}$ is a symmetric M-matrix and strictly diagonally dominant matrix

[22] and each off diagonal element of exact $S_{E_{ij}}$ decays away from the diagonal, each $\tilde{S}_{E_{ij}}$ of these probe approximations satisfies:

Theorem 5.5 *Assume that $S_{E_{ij}}$ is a diagonal dominant M-matrix and*

$$|(S_{E_{ij}})_{l,l}| \geq |(S_{E_{ij}})_{l,l+1}| \geq \cdots \geq |(S_{E_{ij}})_{l,n_{ij}}| \qquad \textit{for } l = 1, 2, \cdots, n_{ij}$$

Then $\tilde{S}_{E_{ij}}$ is still an M-matrix.

Proof *From the assumption, we can obtain*

$$(5.23) \qquad |(S_{E_{ij}})_{k,l}| \geq |(S_{E_{ij}})_{s,t}| \qquad \textit{if } s \leq k, l \leq t.$$

Without loss generality, we assume $n_{ij} = 2k + 1$. Since $S_{E_{ij}}$ is strictly diagonal dominant, we have for $1 \leq l \leq n_{ij}$:

$$a_l = \begin{cases} (S_{E_{ij}})_{l,1} + (S_{E_{ij}})_{l,3} + \cdots + (S_{E_{ij}})_{l,2k+1} & \geq 0, \quad \textit{if } l \textit{ is odd} \\ (S_{E_{ij}})_{l,2} + (S_{E_{ij}})_{l,4} + \cdots + (S_{E_{ij}})_{l,2k} & \geq 0, \quad \textit{if } l \textit{ is even} . \end{cases}$$

$$b_1 = (S_{E_{ij}})_{1,2} + (S_{E_{ij}})_{1,4} + \cdots + (S_{E_{ij}})_{1,2k} \leq 0$$

follows from the property of M-matrix $S_{E_{ij}}$. Then,

$$|a_1| - |b_1| = (S_{E_{ij}})_{1,1} + (S_{E_{ij}})_{1,2} + (S_{E_{ij}})_{1,3} + \cdots + (S_{E_{ij}})_{1,n_{ij}} \geq 0.$$

The equation

$$b_1 + b_2 = (S_{E_{ij}})_{2,1} + (S_{E_{ij}})_{2,3} + \cdots + (S_{E_{ij}})_{2,2k+1} \leq 0,$$

implies

$$\begin{aligned} b_2 &= (S_{E_{ij}})_{2,1} + (S_{E_{ij}})_{2,3} + \cdots + (S_{E_{ij}})_{2,2k+1} \\ &- (S_{E_{ij}})_{1,2} - (S_{E_{ij}})_{1,4} - \cdots - (S_{E_{ij}})_{1,2k} \qquad \leq 0, \end{aligned}$$

by using inequality (5.23). Therefore,

$$\begin{aligned} |a_2| - |b_1| - |b_2| &= (S_{E_{ij}})_{2,2} + (S_{E_{ij}})_{2,4} + \cdots + (S_{E_{ij}})_{2,2k} \\ &+ (S_{E_{ij}})_{1,2} + (S_{E_{ij}})_{1,4} + \cdots + (S_{E_{ij}})_{1,2k} \\ &+ (S_{E_{ij}})_{2,1} + (S_{E_{ij}})_{2,3} + \cdots + (S_{E_{ij}})_{2,2k+1} \\ &- (S_{E_{ij}})_{1,2} - (S_{E_{ij}})_{1,4} - \cdots - (S_{E_{ij}})_{1,2k} \\ &= (S_{E_{ij}})_{2,1} + (S_{E_{ij}})_{2,2} + (S_{E_{ij}})_{2,3} + \cdots + (S_{E_{ij}})_{2,n_{ij}} \qquad \geq 0. \end{aligned}$$

By induction, for any $3 \leq l \leq n_{ij}$, we have

$$\begin{aligned} b_l &= (S_{E_{ij}})_{l,1} + (S_{E_{ij}})_{l,3} + \cdots + (S_{E_{ij}})_{l,2k+1} \\ &- (S_{E_{ij}})_{l-1,2} - (S_{E_{ij}})_{l-1,4} - \cdots - (S_{E_{ij}})_{l-1,2k} \\ &+ b_{l-2} \qquad \text{for} \quad l \quad \text{even}\ , \end{aligned}$$

and

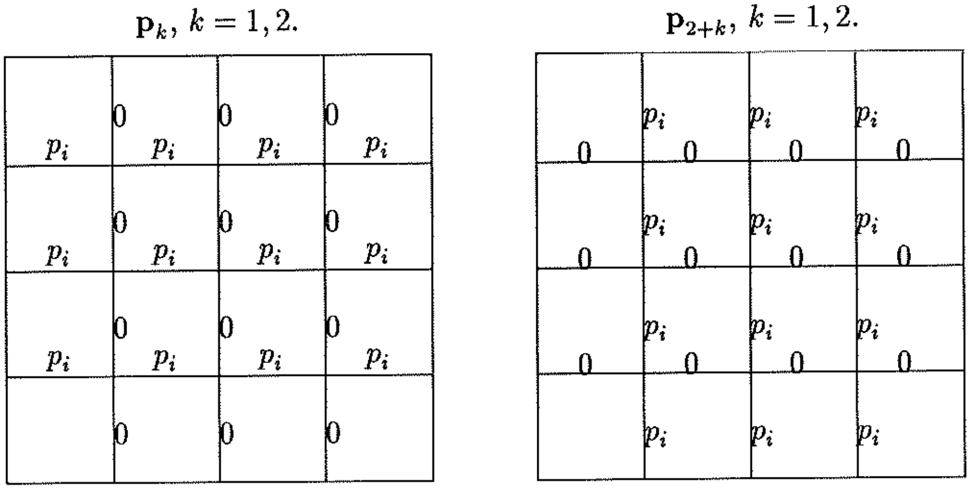
$$\begin{aligned} b_l &= (S_{E_{ij}})_{l,2} + (S_{E_{ij}})_{l,4} + \cdots + (S_{E_{ij}})_{l,2k} \\ &- (S_{E_{ij}})_{l-1,1} - (S_{E_{ij}})_{l-1,3} - \cdots - (S_{E_{ij}})_{l-1,2k+1} \\ &+ b_{l-2} \qquad \text{for} \quad l \quad \text{odd.} \end{aligned}$$

So, we can obtain $b_l \leq 0$ through using $b_{l-1} \leq 0$ and inequality (5.23). Hence,

$$\begin{aligned} |a_l| - |b_{l-1}| - |b_l| &= a_l + b_{l-1} + b_l \\ &= (S_{E_{ij}})_{l,1} + (S_{E_{ij}})_{l,2} + (S_{E_{ij}})_{l,3} + \cdots + (S_{E_{ij}})_{l,n_{ij}} \qquad \geq 0. \end{aligned}$$

Thus, the tridiagonal symmetric matrix $\tilde{S}_{E_{ij}}$ is positive definite. Note that above inequalities can be replaced by strict inequalities, if $S_{E_{ij}}$ is strictly diagonally dominant.

Figure 5.2: Probe Vectors in Modified VS method


$\mathbf{p}_k$, $k = 1, 2$.

$\mathbf{p}_{2+k}$, $k = 1, 2$.

Now, we describe how to use only four probe vectors to form all approximate edge matrices. To minimize the computational work and the approximation errors arising from boundary value on other edges in the constructing procedure of all approximate edge matrices, we will specify the same probe vectors p_k either on all horizontal edges simultaneously or on all vertical edges simultaneously. Let's define $\mathbf{p}_k$ for $k = 1, 2$:

$$\mathbf{p}_k = \begin{cases} p_k & \text{on all horizontal edges} \\ 0 & \text{on all vertical edges} \end{cases}$$

and

$$\mathbf{p}_{k+2} = \begin{cases} 0 & \text{on all horizontal edges} \\ p_k & \text{on all vertical edges} \end{cases}$$

which are drawn as fig (5.2).

Analogously, these approximations $\tilde{S}_{E_{ij}}$ resulted from above simultaneous probe vectors $\mathbf{p}_k$ preserve strictly diagonally dominant and positive definite. Since the edge matrices $\tilde{S}_{E_{ij}}$ are tridiagonal, it is cheap and easy to invert $\tilde{S}_{E_{ij}}$.

Theorem 5.6 *If the Schur complement S on Γ is a strictly diagonally dominant M-matrix, then edge approximations $\tilde{S}_{E_{ij}}$ obtained from above are strictly diagonally dominant and positive definite.*

Proof *From the assumption that M-matrix S is strictly diagonally dominant, $\tilde{S}_{E_{ij}}$ can be proved to be strictly diagonally dominant in the same way as the proof of theorem 5.5.*

The probe approximate vertex matrices $\tilde{S}_{v_k}$ can be easily constructed from the results of matrix simultaneous probe vector products $\{S\mathbf{p}_k\}_{k=1}^{4}$. The detail of this technique can be found in [23]. The same argument as in [23] can be used to show that the vertex approximations $\tilde{S}_{v_k}$ obtained from above probe procedure are diagonally dominant if Schur complement S is a diagonally dominant M-matrix.

5.3.2 The Other Simple Vertex Space Approximations

It is easy to find spectrally equivalent vertex approximations $\tilde{S}_{v_k}$ when vertex size is small. Numerical experiment [66, 23] has already shown that increasing vertex size makes little improvement on the convergence speed. Hence, in practice the vertex size is chosen small in order to save computational work. Here, we suppose that the vertex size is small. Like forming Fourier vertex approximations [23], **tangential vertex approximations** can be constructed as the sum of four tangential approximations [20, 42] on the four L-shape segments around the vertex point. The simplest other vertex approximations are **diagonal vertex approximations** which are diagonal matrices with entries consisting of diagonal entries of original matrix A associated with vertex regions. The advantage of these tangential or diagonal vertex approximations is that they are very sparse so that they can be easily inverted. The disadvantage is that these approximation become deteriorated while increasing the size of vertex space . Numerical tests of the modified VS method with these simple approximations indicate that the modified VS method retains the optimal convergence for elliptic scalar problems with Neumann boundary conditions.

5.4 The Modified BPS Method for Neumann Boundary Value Problems

In this section, we modify BPS method, a preconditioner developed by J.H. Bramble et al. and based on the domain decomposition method, so that the resulted preconditioned conjugate gradient method can be used to solve these singular discrete systems. The condition numbers of the modified methods have still been shown to be $C(1 + \ln(H/h))^2$ in R^2. We have tested the several approximate edge matrices, such as probing technique and Fourier approximation, in the numerical experiment. The numerical results have been reported for Neumann boundary condition problems with various coefficients, such as highly varying and jumping coefficients. The results show that the estimation of condition number is fully realized in practice. The modified BPS algorithm is highly parallelizable.

5.4.1 The BPS Method and Its Modification

As we know, it is very important to make a 'good' choice for preconditioner B in order to construct an efficient PCG algorithm for problem (5.4). The bilinear form $b(\cdot,\cdot)$ should have two properties. Firstly, it should be easy to obtain the solution $w_h \in \tilde{V}^h(\Omega)$ of

$$b(w_h, v_h) = (g, v_h), \qquad \text{for all } v_h \in \tilde{V}^h(\Omega) \tag{5.24}$$

for a given function g. Secondly, the bilinear form $b(\cdot,\cdot)$ should be spectrally equivalent to the bilinear form $a(\cdot,\cdot)$, i.e. there exist positive constants λ_0 and λ_1 such that

$$\lambda_0 b(v_h, v_h) \leq a(v_h, v_h) \leq \lambda_1 b(v_h, v_h), \qquad \text{for all } v_h \in \tilde{V}^h(\Omega). \tag{5.25}$$

The first property guarantees that the computational work in each iteration step is small. The second property implies that the condition number κ of corresponding PCG method is less than λ_1/λ_0 . It is well known that the number of steps required to decrease an appropriate norm of the error of a conjugate gradient iteration by a fixed factor is proportional to the condition number κ; see Golub and Van Loan [40]. Therefore, if the condition number κ is a small positive number and slightly depends on the size of grid and substructure, then the resulting algorithm is an efficient method. We are going to construct a preconditioner B so that above two properties could be satisfied in some way. We first introduce an approximate bilinear form $\tilde{a}(\cdot,\cdot)$ by

$$\tilde{a}_k(u_h, v_h) = \int_{\Omega_k} \sum_{i,j=1}^{2} a_{ij}^k \frac{\partial u_h}{\partial x_i} \frac{\partial v_h}{\partial x_j} dx$$

for each k and then define

$$\tilde{a}(u_h, v_h) = \sum_k \tilde{a}_k(u_h, v_h).$$

Here $a_{i,j}^k$ can be chosen as piecewise smooth uniformly positive definite for each Ω_k so that the inequalities

$$C_0\tilde{a}(u_h, u_h) \leq a(u_h, u_h) \leq C_1\tilde{a}(u_h, u_h) \qquad \text{for all } u_h \in V^h(\Omega)$$

are satisfied for positive constants C_0 and C_1 (independent of h, H, Ω_k) and the problem (5.24) should be easily solvable. From these inequalities, it follows that $Ker\tilde{A} = KerA$. Thus, the problem of finding a preconditioner for a is the same as finding one for $\tilde{a}$.

Denote a subspace

$$V_0^h(\Omega, \Gamma) = \{u_h | u_h \in V^h(\Omega), \qquad u(x) = 0 \text{ on } \Gamma\}$$

which can be represented as the sum of orthogonal subspaces

$$V_0^h(\Omega, \Gamma) = V_0^h(\Omega_1) + V_0^h(\Omega_2) + \cdots + V_0^h(\Omega_n)$$

where

$$V_0^h(\Omega_i) = \{u_h | u_h \in V^h(\Omega),\ u_h = 0,\ x \notin \Omega_i\}.$$

To construct a preconditioner b of a, we first decompose the functions u_h in $\tilde{V}^h(\Omega)$ as $u_h = u_h^I + u_h^B$ where $u_h^I \in V_0^h(\Omega, \Gamma)$ satisfies

$$\tilde{a}_k(u_h^I, v_h) = \tilde{a}_k(u_h, v_h) \qquad \text{for all } v_h \in V_0^h(\Omega_k)$$

for each k , and $u_h^B = u_h$ on Γ and

$$\tilde{a}_k(u_h^B, v_h) = 0 \qquad \text{for all } v_h \in V_0^h(\Omega_k)$$

for all k . We refer to such a function u_h^B as 'discrete $\tilde{a}_k$-harmonic'. It is obvious that such decomposition of function is orthogonal in the $\tilde{a}$-inner product

$$\tilde{a}(u_h, u_h) = \tilde{a}(u_h^I + u_h^B, u_h^I + u_h^B) = \tilde{a}(u_h^I, u_h^I) + \tilde{a}(u_h^B, u_h^B).$$

So we will define $b(\cdot, \cdot)$ by replacing the $\tilde{a}(u_h^B, u_h^B)$ term in above equation. We next further divide the u_h^B as the sum of two functions $u_h^B = u_h^E + u_h^v$ where $u_h^E \in V^h(\Omega)$ is the discrete $\tilde{a}_k$ -harmonic function with zero values at the vertices and $u_h^v \in V^h(\Omega)$ is also the discrete $\tilde{a}_k$-harmonic with linear function values along each edge Γ_{ij} and with the same values as u_h at the vertices. Before defining the preconditioner b , we denote $V_0^h(\Gamma_{ij})$ as the subspace of trace space $V_0^h(\partial\Omega_k)$ whose functions have the supports on the edge Γ_{ij} and introduce an operator $\tilde{l}_0$ defined on each $V_0^h(\Gamma_{ij})$ by

$$< \alpha^{-1}\tilde{l}_0 u_h, v_h >_{\Gamma_{ij}} = < \alpha u_h', v_h' >_{\Gamma_{ij}} \qquad \text{for all } v_h \in V_0^h(\Gamma_{ij})$$

where the prime denotes the differentiation with respect to the arc length s along Γ_{ij}, and the inner product $< \cdot, \cdot >$ on the edge Γ_{ij} is defined as

$$< u_h, v_h >_{\Gamma_{ij}} = \int_{\Gamma_{ij}} u_h v_h ds.$$

By denoting the vertices as v_i or v_j, we could define the preconditioner B by

$$
\begin{aligned}
(5.26) \qquad b(u_h, v_h) \;&=\; \tilde{a}(u_h^I, v_h^I) + \sum_{\Gamma_{ij}} \alpha_{ij} < \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, v_h^E >_{\Gamma_{ij}} \\
&\quad + \sum_{\Gamma_{ij}} \alpha_{ij} (u_h^v(v_i) - u_h^v(v_j))(v_h^v(v_i) - v_h^v(v_j)).
\end{aligned}
$$

Now we present a detail description of the process used to solve problem (5.24) for any given function $g \perp KerA$ (the mean value of function g is zero). In fact, solving problem (5.24) is equivalent to finding the corresponding decomposition functions u_h^I and u_h^B. The restriction of function u_h^I on Ω_k could be uniquely determined by solving the small size Dirichlet subproblem with zero boundary condition on Ω_k:

$$
(5.27) \qquad \tilde{a}_k(u_h^I, v_h) = (g, v_h) \qquad \text{for all } v_h \in V_0^h(\Omega_k).
$$

Therefore, u_h^I could be obtained on the whole domain $\overline{\Omega}$ by solving subproblems on each sub-domain. Since all these subproblems are independent of each other, they can be solved in parallel. With u_h^I now known, the problem reduces to finding u_h^B from following equation:

$$
\begin{aligned}
\sum_{\Gamma_{ij}} \alpha_{ij} < \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, w_h^E > \;&+\; \sum_{\Gamma_{ij}} \alpha_{ij} (u_h^v(v_i) - u_h^v(v_j))(w_h^v(v_i) - w_h^v(v_j)) \\
(5.28) \qquad &=\; (g, w_h) - \tilde{a}(u_h^I, w_h^I) \\
&=\; (g, w_h) - \tilde{a}(u_h^I, w_h) \qquad \text{for all } w_h \in V^h(\Omega).
\end{aligned}
$$

Denote

$$
(\tilde{g}, w_h) = (g, w_h) - \tilde{a}(u_h^I, w_h) \qquad \text{for all } w_h \in V^h(\Omega).
$$

Then $\tilde{g}$ is obviously orthogonal to the subspace $KerA$. Note that the value of $(\tilde{g}, w_h)$ only depends on the value of w_h on each Γ_{ij}. Let w_h be in the subspace of $V^h(\Omega)$ whose elements vanish in the interior mesh points of every Ω_k and all vertices. Then the problem (5.28) decouples into the independent problems of finding $u_h^E \in V_0^h(\Gamma_{ij})$ such that

$$(5.29) \qquad \alpha_{ij} < a^{-1}\tilde{l}_0^{1/2} u_h^E, w_h >_{\Gamma_{ij}} = (g, w_h) - \tilde{a}(u_h^I, w_h) \qquad \forall w_h \in V_0^h(\Gamma_{ij})$$

on each Γ_{ij}. All these subproblems have unique solutions and could be solved concurrently. In practice we use Dryja approximation matrix [31, 25] or probing edge matrix [21, 22, 23] instead of $\tilde{l}_0^{1/2}$ in above problem. Right now only function u_h^v is left unknown. To determine the function u_h^v , we introduce a subspace of $V^h(\Omega)$ consisting of functions which are linear between the endpoints of each edge Γ_{ij} and vanish at the interior mesh points of each Ω_k. It is clear that for each $w_h \in V_0^h$ in this subspace, the corresponding w_h^E should be zero. In this subspace, the problem (5.28) reduces to the problem

$$(5.30) \qquad \sum_{\Gamma_{ij}} \alpha_{ij}(u_h^v(v_i) - u_h^v(v_j))(w_h^v(v_i) - w_h^v(v_j)) = (g, w_h) - \tilde{a}(u_h^I, w_h)$$

which only has u_h^v as the unknown function. Choose a basis $\Phi_1, \Phi_2, \cdots, \Phi_{N_v}$ in this subspace where N_v is the number of vertices on Γ and $\Phi_i(v_j)$ is one if $i = j$ and zero otherwise. Under this basis, problem (5.30) reduces to a difference equation on the coarse mesh for the elliptic problem with Neumann boundary condition. Therefore, problem (5.30) has many solutions when the restricted $\tilde{g}$ on the coarse grid satisfies compatible condition. To find an appropriate solution, we look for the unique solution $\tilde{u}_h^v$ which has zero mean value. Then the values of $\tilde{u}_h^v$ at vertices uniquely determine its values on the edges. Note this extension method should change constant function on the whole coarse grid to the constant function on the whole edges Γ. Extend the sum $\tilde{u}_h^B = \tilde{u}_h^v + u_h^E$ into substructures Ω_k so that

$$\tilde{a}_k(\tilde{u}_h^B, v_h) = 0 \quad \forall v_h \in V_0^h(\Omega_k) \quad \forall k.$$

By subtracting the mean value c of $\tilde{u}_h = \tilde{u}_h^B + u_h^I$ from $\tilde{u}_h$, we obtain

$$u_h = \tilde{u}_h - c = \tilde{u}_h^B + u_h^I - c$$

which has zero mean value. Hence, for any given function g, the solution u_h of (5.24) is unique and belongs to $\tilde{V}^h$. Then a preconditioner for problem (5.4) in $\tilde{V}^h$ has been well defined.

Note that problem (5.29) and problem (5.30) are independent. Hence, these two kind subproblems can be solved at the same time. Because all the subproblems have almost the same small size and the computational work for obtaining the solutions to these problem is almost the same small, we have good balance on the working load for each processor when this method is used on the multi-processor computer.

To make the proceed of the inverse of preconditioner b clear, we outline the steps of calculating $u_h \in \tilde{V}^h(\Omega)$ such that

$$b(u_h, v_h) = (g, v_h) \qquad \forall v_h \in \tilde{V}^h(\Omega).$$

where g satisfies compatible condition.

Algorithm:

1. Find $u_h^I \in V_0^h(\Omega, \Gamma)$ such that

$$\tilde{a}_k(u_h^I, v_h) = (g, v_h) \forall v_h \in V_0^h(\Omega_k) \forall k.$$

2. Find u_h^E on each Γ_{ij} by solving one dimensional problem

$$\alpha_{ij} < \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, v_h >_{\Gamma_{ij}} = (g, v_h) - \tilde{a}(u_h^I, v_h) \forall v_h \in V_0^h(\Gamma_{ij}).$$

3. Solve coarse problem to get $\tilde{u}_h^v$ with zero mean value such that equation (5.30) is true under the same coarse base functions, i.e.

$$\sum_{\Gamma_{ij}} \alpha_{ij}(u_h^v(v_i) - u_h^v(v_j))(w_h^v(v_i) - w_h^v(v_j)) = (g, w_h) - \tilde{a}(u_h^I, w_h).$$

4. Extend the $\tilde{u}_h^v$ to the edges piecewise linearly.

5. Calculate $\tilde{u}_h^B$ such that $\tilde{u}_h^B|_\Gamma = \tilde{u}_h^v + u_h^E$ and

$$\tilde{a}_k(\tilde{u}_h^B, v_h) = 0 \qquad \forall v_h \in V_0^h(\Omega_k) \forall k.$$

6. Compute the mean value c of $u_h^I + \tilde{u}_h^B$ on $\bar{\Omega}$ and let $u_h = u_h^I + \tilde{u}_h^B - c \in \tilde{V}^h(\Omega)$.

5.4.2 Theoretical Results

In this section we estimate the condition number of the modified BPS method for Neumann boundary value problem through proving the inequalities in (5.25). We will use the approach, stated in [8] and [34], to prove inequalities (5.25).

Since there exist positive constant C_0 and C_1 such that

$$C_0\tilde{a}(u_h, u_h) \le a(u_h, u_h) \le C_1\tilde{a}(u_h, u_h) \qquad \forall u_h \in \tilde{V}^h(\Omega),$$

it suffices to compare $\tilde{a}(u_h, u_h)$ with $b(u_h, u_h)$. We decompose $u_h \in \tilde{V}^h(\Omega)$ into $u_h = u_h^I + u_h^B$ as the previous section where $u_h^B = \tilde{u}_h^B - c$. Then, we have

$$\tilde{a}(u_h, u_h) = \tilde{a}(u_h^I, u_h^I) + \tilde{a}(u_h^B, u_h^B).$$

From the equality

$$b(u_h, u_h) = \tilde{a}(u_h^I, u_h^I) + b(u_h^B, u_h^B),$$

the proof of the inequalities in (5.25) will be obtained if the following inequalities are true.

$$\tilde{a}(u_h^B, u_h^B) \le Cb(u_h^B, u_h^B) \tag{5.31}$$

and

$$b(u_h^B, u_h^B) \le C(1 + \ln(H/h))^2\tilde{a}(u_h^B, u_h^B). \tag{5.32}$$

In order to prove these inequalities, we further decompose u_h^B into $u_h^E + u_h^v$ with $u_h^v = \tilde{u}_h^v + c$. Hence, if these inequalities

$$\tilde{a}_k(u_h^B, u_h^B) \le C \sum_{ij\in\beta_k} \alpha_{ij}(< \alpha^{-1}\tilde{l}_0^{1/2}u_h^E, u_h^E >_{\Gamma_{ij}} + (u_h^v(v_i) - u_h^v(v_j))^2) \tag{5.33}$$

and

$$\begin{aligned} \sum_{ij\in\beta_k} \alpha_{ij}(< \alpha^{-1}\tilde{l}_0^{1/2}u_h^E, u_h^E >_{\Gamma_{ij}} \;&+\; (u_h^v(v_i) - u_h^v(v_j))^2) \\ &\le\; (1 + \ln(H/h))^2\tilde{a}_k(u_h^B, u_h^B), \end{aligned} \tag{5.34}$$

are satisfied on each substructure, then summing these inequalities with respect to k gives the inequalities in (5.31) and (5.32). On each substructure Ω_k, using the results in [8] leads

$$\tilde{a}_k(\tilde{u}_h^B, \tilde{u}_h^B) \leq C \sum_{ij \in \beta_k} \alpha_{ij}(< \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, u_h^E >_{\Gamma_{ij}} + (\tilde{u}_h^v(v_i) - \tilde{u}_h^v(v_j))^2)$$

and

$$\begin{aligned} \sum_{ij \in \beta_k} \alpha_{ij}(< \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, u_h^E >_{\Gamma_{ij}} &+ (\tilde{u}_h^v(v_i) - \tilde{u}_h^v(v_j))^2) \\ &\leq (1 + \ln(H/h))^2 \tilde{a}_k(\tilde{u}_h^B, \tilde{u}_h^B). \end{aligned}$$

For any constant c, we can prove that

$$\begin{aligned} \tilde{a}_k(\tilde{u}_h^B + c, \tilde{u}_h^B + c) \leq \; & C \textstyle\sum_{ij \in \beta_k} \alpha_{ij}(< \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, u_h^E >_{\Gamma_{ij}} \\ & + (\tilde{u}_h^v(v_i) + c - (\tilde{u}_h^v(v_j) + c))^2) \end{aligned}$$

and

$$\begin{aligned} \sum_{ij \in \beta_k} \alpha_{ij}(< \alpha^{-1} \tilde{l}_0^{1/2} u_h^E, u_h^E >_{\Gamma_{ij}} &+ (\tilde{u}_h^v(v_i) + c - (\tilde{u}_h^v(v_j) + c))^2) \\ &\leq (1 + \ln(H/h))^2 \tilde{a}_k(\tilde{u}_h^B + c, \tilde{u}_h^B + c). \end{aligned}$$

Therefore, it is obvious that the inequalities in (5.33) and (5.34) follow from these inequalities. Summing the inequalities in (5.33) and (5.34) over all sub-domains gives the inequalities in (5.31) and (5.32). Hence, we have following estimation on condition number.

Theorem 5.7 *The above preconditioner b satisfies : for all $u_h \in \tilde{V}^h(\Omega)$*

$$\frac{C_0}{(1+\ln^2(H/h))} b(u_h, u_h) \leq a(u_h, u_h) \leq C_1 b(u_h, u_h) \tag{5.35}$$

for positive constants C_0 and C_1 which are independent of h and H. Thus, the condition number of corresponding preconditioned conjugate gradient method grows at most like $\kappa \leq C(1+\ln^2(H/h))$ as h tends to zero.

5.5 Numerical Results

Now we present the numerical results of tests on the convergence rates of the modified VS and BPS methods with various edge and vertex approximations. The tests were performed for the scalar elliptic problem with Neumann boundary condition:

$$\begin{cases} Lu \equiv -\nabla \cdot (\alpha(x,y)\nabla u(x,y)) = f(x,y) & \text{in } \Omega = [0,1]^2 \\ \frac{\partial u}{\partial N} = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.36}$$

The following four coefficients have been used in our tests:

1. $\alpha(x,y) = 1$, the Laplace operator, see Table 5.1.
2. $\alpha(x,y) = 1 + 10(x^2 + y^2)$, slowly varying smooth coefficients, see Table 5.2.
3. $\alpha(x,y) = e^{10xy}$, highly varying smooth coefficients, see Table 5.3
4. Highly discontinuous coefficients of Fig. 2.5.

The square domain $[0,1]^2$ was first divided into small square sub-domains with uniform size H. Then each square sub-domain was triangulated into finite element with uniform mesh size h on the square domain. These problems were discretized by standard finite element method with five stencil.

u is randomly generated solution of the scalar elliptic problem normalized so that the mean value of u is zero. The integer $\aleph$ is defined to be the number of iterations required to reduce the A-norm of the error $e_n = u - u_n$ by a factor 10^{-5} . We will list the iteration number $\aleph$ and estimated condition number κ for these discrete problems with various coarse mesh size H and fine mesh size h in following tables. The modified VS space method and BPS method with Fourier approximations on edges and vertex spaces are denoted as FVS and FBPS, respectively. FDVS stands for the modified VS method with Fourier approximations on the edges and diagonal vertex approximations. PVS and PBPS respectively represent the modified VS space method and BPS with probe approximations on edges and vertex spaces.

In our program the problems on coarse grid and on sub-domains are solved with high precision. The size of vertex space matrix is chosen as 5×5 matrices in our experiment except in Table 5.5 the size of vertex matrices is 1. Since the size of vertex matrices is small, they can be inexpensively inverted.

To compare six probe method with four probe method, we will list the results in Table 5.6 for the problem (5.36) with harmonic Dirichlet boundary condition and various coefficient defined above.

Table 5.1: Comparison for Laplace equation

h^{-1}	Ovlp	FBPS		PBPS		FDVS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
64_2	1/32	21	24.49	19	17.91	9.81	14	9.12	14	6.47	12
64_4	1/16	19	17.99	16	12.34	8.02	14	7.33	14	5.12	11
64_8	1/8	17	12.36	14	8.42	6.13	12	5.51	13	3.68	10
64_16	1/4	14	7.23	12	5.44	4.58	11	4.43	12	3.01	9
64_32	1/2	11	4.08	10	3.11	6.00	13	5.72	13	5.45	13
128_2	1/64	24	32.79	21	33.69	11.56	15	11.50	15	11.23	15
128_4	1/32	21	24.28	19	17.83	10.49	15	9.49	15	6.53	12
128_8	1/16	19	17.91	16	11.89	8.16	13	7.38	14	4.90	11
128_16	1/8	16	11.95	14	8.49	6.05	12	5.54	13	3.66	10
128_32	1/4	14	7.09	12	5.46	4.51	11	4.38	11	2.94	9
128_64	1/2	11	4.05	10	3.12	5.84	13	5.67	13	5.35	13
256_2	1/128	24	41.09	28	59.84	15.17	15	13.65	16	19.80	17
256_4	1/64	23	31.64	22	31.95	12.70	15	11.54	16	11.03	15
256_8	1/32	21	23.94	18	17.69	10.41	14	9.58	15	6.24	12
256_16	1/16	18	17.45	16	12.22	8.16	14	7.31	14	4.58	11
256_32	1/8	16	11.89	14	8.40	6.08	12	5.55	13	3.60	10
256_64	1/4	14	7.09	12	5.47	4.50	11	4.37	12	2.86	9
256_128	1/2	11	4.07	10	3.12	5.79	13	5.65	13	5.33	13

Table 5.2: Comparison for $\alpha = 1 + 10(x^2 + y^2)$

h^{-1}	Ovlp	FBPS		PBPS		FDVS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
64_2	1/32	25	34.38	20	19.41	10.22	15	9.59	15	6.76	12
64_4	1/16	20	17.70	16	12.03	8.29	14	7.78	15	5.27	12
64_8	1/8	17	12.15	15	8.49	6.26	13	5.64	13	3.78	10
64_16	1/4	15	7.33	12	5.48	4.67	11	4.48	12	3.00	9
64_32	1/2	11	4.08	10	3.13	6.04	13	5.75	13	5.53	13
128_2	1/64	24	35.95	23	32.95	12.22	15	11.86	16	11.51	15
128_4	1/32	21	23.86	19	17.60	10.68	15	9.91	16	6.48	12
128_8	1/16	19	17.45	17	12.27	8.30	14	7.42	14	5.05	11
128_16	1/8	16	11.90	14	8.35	6.05	12	5.53	13	3.73	10
128_32	1/4	14	7.09	12	5.51	4.57	11	4.44	12	3.00	9
128_64	1/2	11	4.05	10	3.12	5.88	13	5.68	13	5.36	13
256_2	1/128	26	50.60	28	58.48	15.78	16	14.24	17	20.02	18
256_4	1/64	24	32.05	22	32.10	12.77	16	11.76	17	11.16	15
256_8	1/32	21	24.30	17	17.30	10.62	15	9.70	15	6.06	12
256_16	1/16	18	17.51	16	12.16	8.17	14	7.38	14	4.61	11
256_32	1/8	16	11.91	14	8.42	6.10	12	5.55	13	3.66	10
256_64	1/4	14	7.13	12	5.46	4.55	11	4.37	11	2.87	9
256_128	1/2	11	4.06	10	3.12	5.80	13	5.65	13	5.34	13

Table 5.3: Comparison for $\alpha = e^{10xy}$

h^{-1}	Ovlp	FBPS		PBPS		FDVS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
64_2	1/32	22	83.13	19	56.49	21.84	15	20.85	16	10.87	11
64_4	1/16	21	24.72	17	16.61	10.97	14	10.79	15	7.09	12
64_8	1/8	19	13.30	15	9.11	7.32	14	6.92	14	4.69	11
64_16	1/4	15	7.48	12	5.44	5.18	12	5.29	12	3.36	10
64_32	1/2	11	4.13	10	3.10	6.07	13	5.54	13	5.89	13
128_2	1/64	24	101.22	24	83.61	27.78	16	26.40	16	18.28	14
128_4	1/32	22	31.72	17	22.36	15.53	16	14.76	17	7.86	12
128_8	1/16	20	18.78	15	12.92	8.72	14	8.06	14	6.02	11
128_16	1/8	16	12.07	15	8.72	6.70	13	6.34	14	4.16	10
128_32	1/4	15	7.40	12	5.59	4.91	12	4.78	12	3.31	9
128_64	1/2	11	4.08	9	3.09	6.00	13	5.65	13	5.59	13
256_2	1/128	32	119.78	27	141.88	34.91	17	35.19	19	32.47	18
256_4	1/64	25	39.69	22	38.50	17.60	17	16.90	19	13.33	16
256_8	1/32	22	25.00	17	18.60	11.65	16	11.11	16	7.46	12
256_16	1/16	20	17.69	17	12.28	8.41	14	7.52	14	4.91	11
256_32	1/8	17	11.96	15	8.57	6.33	13	5.73	13	3.67	10
256_64	1/4	15	7.29	12	5.47	4.72	11	4.50	12	3.11	9
256_128	1/2	11	4.04	10	3.12	6.03	13	5.69	13	5.36	13

Table 5.4: Comparison for highly jumping coefficient

h^{-1}	Ovlp	FBPS		PBPS		FDVS		FVS		PVS	
_H^{-1}	h/H	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$	κ	$\aleph$
32_4	1/8	15	12.59	14	8.84	23.66	16	18.85	16	9.90	12
32_8	1/4	14	7.17	12	5.60	15.60	16	12.64	14	6.76	11
32_16	1/2	11	3.99	9	3.04	9.90	16	6.88	14	6.60	13
64_4	1/16	18	18.54	16	12.80	33.31	18	27.20	18	13.84	13
64_8	1/8	16	11.86	14	8.84	23.59	18	19.10	16	9.75	12
64_16	1/4	14	7.15	12	5.43	8.32	14	11.39	14	6.29	11
64_32	1/2	11	4.14	10	3.13	7.63	14	6.37	13	5.52	13
128_4	1/32	21	24.29	17	18.36	43.43	19	35.89	20	8.50	13
128_8	1/16	18	18.24	15	12.72	32.35	20	26.62	19	13.31	13
128_16	1/8	16	12.03	13	8.58	23.28	18	18.81	16	9.84	12
128_32	1/4	14	7.20	12	5.46	15.25	16	12.10	14	6.63	11
128_64	1/2	11	4.10	10	3.12	7.64	15	6.26	13	5.54	13
256_4	1/64	23	32.50	22	32.92	54.73	21	45.68	21	27.38	16
256_8	1/32	21	24.57	17	17.98	44.17	22	36.06	20	18.31	15
256_16	1/16	18	17.44	14	12.05	28.89	19	24.62	18	13.12	13
256_32	1/8	16	11.80	14	8.50	22.66	17	18.89	16	9.56	12
256_64	1/4	14	7.10	12	5.46	6.70	13	4.95	12	5.76	10
256_128	1/2	11	4.06	10	3.12	7.22	14	6.10	13	5.38	13

Table 5.5: Comparison when size of vertex matrices is 1 × 1

Coefficient		Laplace		$1+10(x^2+y^2)$		e^{10xy}		disc.	
h^{-1}	Ovlp	FVS	PVS	FVS	PVS	FVS	PVS	FVS	PVS
_H^{-1}	h/H	$\aleph$	$\aleph$	$\aleph$	$\aleph$	$\aleph$	$\aleph$	$\aleph$	$\aleph$
64_4	1/16	14	12	14	12	14	13	16	14
64_8	1/8	12	11	12	11	12	11	13	12
64_16	1/4	10	9	10	10	10	10	11	9
64_32	1/2	9	9	9	9	9	9	9	9
128_4	1/32	15	14	15	14	16	14	18	16
128_8	1/16	13	12	13	12	13	12	15	14
128_16	1/8	12	11	12	10	12	10	14	12
128_32	1/4	10	9	10	9	10	10	12	11
128_64	1/2	9	9	9	9	9	9	9	9
256_4	1/64	16	16	16	16	18	16	19	19
256_8	1/32	14	13	14	13	15	14	17	15
256_16	1/16	13	12	13	12	14	12	15	14
256_32	1/8	11	10	12	11	12	11	13	12
256_64	1/4	10	9	10	9	10	9	11	10
256_128	1/2	9	8	9	9	9	9	9	8

Table 5.6: Comparison 6 with 4 Probe Vectors in Probe Method

Coefficient		Laplace		$1 + 10(x^2 + y^2)$		e^{10xy}		disc.	
Probe Vec.		6	4	6	4	6	4	6	4
h^{-1}	Ovlp	PVS		PVS		PVS		PVS	
__H^{-1}	h/H	ℵ	ℵ	ℵ	ℵ	ℵ	ℵ	ℵ	ℵ
64_4	1/16	9	10	9	10	9	10	11	12
64_8	1/8	9	9	9	9	8	9	10	10
64_16	1/4	8	9	8	9	8	8	9	10
64_32	1/2	13	13	13	13	13	13	13	13
128_4	1/32	10	10	10	10	10	11	11	12
128_8	1/16	9	10	9	10	9	10	11	10
128_16	1/8	9	9	8	9	8	9	9	10
128_32	1/4	8	9	8	9	8	8	9	10
128_64	1/2	13	13	13	13	13	13	13	13
256_4	1/64	13	13	13	13	13	11	13	13
256_8	1/32	10	11	10	11	10	11	11	11
256_16	1/16	9	10	9	10	9	10	11	11
256_32	1/8	9	9	9	9	8	9	10	9
256_64	1/4	8	9	8	9	8	8	9	10
256_128	1/2	13	13	13	13	13	13	13	13

Conclusions: The modified BPS method and VSDD method for the singular Neumann boundary value problems still converge very fast. Both the Fourier and Probe variants of the vertex space algorithm and BPS method are also efficient. The numerical results in all tables demonstrate that the condition number of the vertex space method slightly depends on the size of overlapping, and the BPS method has a condition number $C(1 + \ln^2(H/h))$ with $0 < C \leq 5$. We also notice the convergence rate does not significantly deteriorate even for the highly jumping and varying coefficient.

CHAPTER 6

Algebraic Domain Decomposition Methods

In this chapter, by using the domain decomposition methodology, we construct several algebraic domain decomposition (ADD) methods [43] for certain algebraic systems with sparse matrix. These methods are highly parallelizable. We show that these methods are convergent. We also discuss the eigenvalue distributions of the corresponding iterative matrices in order to analyze the convergence factors of these methods.

6.1 General Chaotic Relaxation Schwarz Methods

Let's consider the linear algebraic system:

$$Au = f, \tag{6.1}$$

where matrix A is a $(2p-1) \times (2p-1)$ block square matrix, denoted as

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,2p-1} \\ \vdots & \cdots & \vdots \\ A_{2p-1,1} & \cdots & A_{2p-1,2p-1} \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_{2p-1} \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ \vdots \\ f_{2p-1} \end{bmatrix},$$

u_i and f_i are vectors. As in the domain decomposition methods, we partition the unknown vector u into p new subvectors with overlapping. We let $\tilde{u}$ denoted as a vector defined by these p new sub-vectors:

$$(6.2)\ \tilde{u} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \text{ with } x_1 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad x_p = \begin{bmatrix} \tilde{u}_{2p-2} \\ u_{2p-1} \end{bmatrix} \quad x_i = \begin{bmatrix} \tilde{u}_{2i-2} \\ u_{2i-1} \\ u_{2i} \end{bmatrix},$$

with $2 \le i \le p-1$. Note that $\tilde{u}_{2i}$ are the unknown vectors associated with the overlapping to the subvector u_{2i} for $i = 1, \cdots, p-1$. In the same way, we can introduce a new vector $\tilde{f}$ from righthandside vector f. Thus, the matrix A is divided into $p \times p$ corresponding block submatrices with overlapping. Then, a corresponding new matrix $\tilde{A}$ can be defined by these $p \times p$ block submatrices:

$$(6.3) \qquad \tilde{A} = \begin{bmatrix} \tilde{A}_{1,1} & \cdots & \tilde{A}_{1,p} \\ \vdots & \cdots & \vdots \\ \tilde{A}_{p,1} & \cdots & \tilde{A}_{p,p} \end{bmatrix}.$$

Here $\tilde{A}$ is a $p \times p$ block square matrix defined by:

$$\tilde{A}_{1,1} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \qquad \tilde{A}_{p,p} = \begin{bmatrix} A_{2p-2,2p-2} & A_{2p-2,2p-1} \\ A_{2p-1,2p-2} & A_{2p-1,2p-1} \end{bmatrix},$$

$$\tilde{A}_{1,p} = \begin{bmatrix} 0 & A_{1,2p-1} \\ 0 & A_{2,2p-1} \end{bmatrix}, \qquad \tilde{A}_{1,i} = \begin{bmatrix} 0 & A_{1,2i-1} & A_{1,2i} \\ 0 & A_{2,2i-1} & A_{2,2i} \end{bmatrix}, \text{ for } 2 \le i \le p-1.$$

$$\tilde{A}_{p,1} = \begin{bmatrix} A_{2p-2,1} & 0 \\ A_{2p-1,1} & 0 \end{bmatrix} \qquad \tilde{A}_{p,i} = \begin{bmatrix} A_{2p-2,2i-2} & A_{2p-2,2i-1} & 0 \\ A_{2p-1,2i-2} & A_{2p-1,2i-1} & 0 \end{bmatrix} \qquad \text{for } 2 \le i \le p-1,$$

$$\tilde{A}_{i,1} = \begin{bmatrix} A_{2i-2,1} & 0 \\ A_{2i-1,1} & 0 \\ A_{2i,1} & 0 \end{bmatrix} \qquad \tilde{A}_{i,p} = \begin{bmatrix} 0 & A_{2i-2,2p-1} \\ 0 & A_{2i-1,2p-1} \\ 0 & A_{2i,2p-1} \end{bmatrix}, \text{ for } 2 \le i \le p-1.,$$

$$\tilde{A}_{i,j} = \begin{bmatrix} 0 & A_{2i-2,2j-1} & A_{2i-2,2j} \\ 0 & A_{2i-1,2j-1} & A_{2i-1,2j} \\ 0 & A_{2i,2j-1} & A_{2i,2j} \end{bmatrix}, \qquad \text{for } 2 \le i < j \le p-1,$$

$$\tilde{A}_{i,j} = \begin{bmatrix} A_{2i-2,2j-2} & A_{2i-2,2j-1} & 0 \\ A_{2i-1,2j-2} & A_{2i-1,2j-1} & 0 \\ A_{2i,2j-2} & A_{2i,2j-1} & 0 \end{bmatrix} \qquad \text{for } 2 \le j < i \le p-1,$$

and

$$\tilde{A}_{i,i} = \begin{bmatrix} A_{2i-2,2i-2} & A_{2i-2,2i-1} & A_{2i-2,2i} \\ A_{2i-1,2i-2} & A_{2i-1,2i-1} & A_{2i-1,2i} \\ A_{2i,2i-2} & A_{2i,2i-1} & A_{2i,2i} \end{bmatrix} \qquad \text{for } 2 \le i \le p-1.$$

From these definition, we obtain a new linear system

$$\tilde{A}\tilde{u} = \tilde{f}, \tag{6.4}$$

which is associated with system (6.1). The following theorem gives the relation between equation (6.1) and system (6.4).

Theorem 6.1 *Suppose that $A_{2i,2i}$ are nonsingular for all $i = 1, \cdots, p-1$. Then, the solution u of system (6.1) can be constructed from the solution $\tilde{u}$ of the problem (6.4) and vice versa.*

Proof *Let u be the solution of system (6.1). We can construct a new vector $\tilde{u}$ as above from u by letting the overlapping subvectors $\tilde{u}_{2i} = u_{2i}$ for $i = 1, \cdots, p-1$. Then, this $\tilde{u}$ is the solution of system (6.4).*

Suppose that $\tilde{u}$ is the solution of problem (6.4). Because $A_{2i,2i}$ are nonsingular for $i = 1, \cdots, p-1$, we have $u_{2i} = \tilde{u}_{2i}, \quad i = 1, \cdots, p-1$, from the equation $A_{2i,2i}(u_{2i} - \tilde{u}_{2i}) = 0$. A direct consequence is that the vector $u = (u_1^T, \cdots, u_{2p-1})^T$ is the solution of system (6.1).

Theorem 6.1 implies that if the solution of (6.1) is unique, then problem (6.4) has only one solution. The following theorem compares the eigenvalues of A with those of matrix $\tilde{A}$. Let us denote $\lambda\{A\}$ to be the set of eigenvalues of matrix A.

Theorem 6.2

$$\lambda\{\tilde{A}\} = \lambda\{A\}(\cup_{i=1}^{p-1}\lambda\{A_{2i,2i}\})$$

Proof *We first prove $\lambda\{\tilde{A}\} \subseteq \lambda\{A\}(\cup_{i=1}^{p-1}\lambda\{A_{2i,2i}\})$. From $\tilde{A}\tilde{u} = \tilde{\lambda}\tilde{u}$, we have*

$$A_{2i,2i}(u_{2i} - \tilde{u}_{2i}) = \tilde{\lambda}(u_{2i} - \tilde{u}_{2i}) \qquad i = 1, \cdots, p-1.$$

If $u_{2i} \neq \tilde{u}_{2i}$ for some $i \in \{1, \cdots, p-1\}$, then $\tilde{\lambda} \in \lambda\{A_{2i,2i}\}$. Otherwise, if $u_{2i} = \tilde{u}_{2i}$ for all $1 \leq i \leq p-1$, then $\tilde{\lambda} \in \lambda\{A\}$.

Now we prove that $\lambda\{A\}(\cup_{i=1}^{p-1}\lambda\{A_{2i,2i}\}) \subseteq \lambda\{\tilde{A}\}$. From $Au = \lambda u$, we have $\tilde{A}\tilde{u} = \lambda\tilde{u}$ by letting $\tilde{u}_{2i} = u_{2i}$ in the definition form (6.2). If $A_{2i,2i}v_{2i} = \lambda v_{2i}$ for some $i \in \{1, \cdots, p\}$, then we can construct a vector $\tilde{u}$ in the following way such that $\tilde{A}\tilde{u} = \lambda\tilde{u}$. Let $w = -(0, \cdots, 0, v_{2i}^T A_{2i+1,2i}^T, \cdots, v_{2i}^T A_{2p-1,2i})^T$. If this $\lambda \notin \lambda\{A\}$, *then the equation*

$$(A - \lambda I)u = w$$

has only one solution u. By setting

$$\tilde{u}_{2j} = \begin{cases} u_{2j} & \textit{if } j \neq i \\ u_{2i} + v_{2j} & \textit{when } j = i \end{cases}$$

For this $\tilde{u}$, we can easily show that $\tilde{A}\tilde{u} = \lambda\tilde{u}$. Hence, $\lambda\{A\}(\cup_{i=1}^{p-1}\lambda\{A_{2i,2i}\}) \subseteq \lambda\{\tilde{A}\}$.

Before giving a definition of the asynchronous Schwarz algorithms for problem (6.1), we, first, decompose linear system (6.1) into p subproblems: To find $\underline{\mathrm{x}}_i$ such that

$$\tilde{A}_{i,i}\underline{\mathrm{x}}_i = \tilde{f}_i - \sum_{j \neq i} \tilde{A}_{i,j}x_j \qquad \text{for } 1 \leq i \leq p, \tag{6.5}$$

where $\tilde{u} = (x_1^T, \cdots, x_p^T)^T$ is regarded as a known vector and $\underline{\tilde{\mathrm{u}}} = (\underline{\mathrm{x}}_1^T, \cdots, \underline{\mathrm{x}}_p^T)^T$ as an unknown vector. Let ϕ_i denote as an solver for subproblem (6.5). This ϕ_i is either direct solver or iterative solver. Now we list several possible choices of ϕ_i as the examples,

(a) ϕ_i is a director solver for the ith problem, i.e.

$$\underline{\tilde{\mathrm{u}}} = \phi_i(\tilde{u}) \qquad \text{and} \qquad \tilde{A}_{i,i}\underline{\tilde{\mathrm{u}}} = \tilde{f}_i - \sum_{j \neq i} \tilde{A}_{i,j}x_j;$$

(b) $\underline{\tilde{\mathrm{u}}} = \phi_i\tilde{u}$ is defined by

$$r_i = \tilde{f}_i - \sum_{1 \leq j \leq p} \tilde{A}_{i,j}x_j$$

$$\alpha_i = (r_i, r_i)/(\tilde{A}_{i,i}r_i, r_i) \qquad \underline{\mathrm{x}}_i = x_i + \alpha_i r_i;$$

(c) Let $\beta_i$be the maximum eigenvalue of $\tilde{A}_{i,i}$, Then $\underline{\tilde{u}} = \phi_i \tilde{u}$ is defined by

$$\underline{x}_i = x_i - (\sum_{1 \le j \le p} \tilde{A}_{i,j} x_j - tildef_i)$$

(d) Let $\tilde{A}_{ii} = M_i - N_i$ where M_i is an invertible matrix. Then $\underline{\tilde{u}} = \phi_i \tilde{u}$

$$\underline{x}_i = x_i + M_i^{-1}(\tilde{f}_i - \sum_{1 \le j \le p} \tilde{A}_{i,j} x_j)$$

Let's distribute the computation on the MIMD machine with p processors. Each processor is assigned to solve one subproblem (6.5). Denote **N** as the whole positive integer set. If we let all processor keep on calculating by using the most recent available data from neighbour processors, then we have following asynchronous Schwarz methods:

Chaotic Relaxation Schwarz Methods:

Let $\tilde{u}^{(0)}$ be a given initial guess vector. The vector sequence $\{\tilde{u}^{(k)}\}$ will be defined by the recursion

$$(6.6) \qquad \begin{array}{ll} \underline{x}_i^{(k+1)} = \phi_i(x_1^{s_1(k)}, \cdots, x_p^{s_p(k)}) & \\ & if\, i \in J(k) \\ x_i^{(k+1)} = x_i^{s_i(k)} + \omega_i(\underline{x}_i^{(k+1)} - x_i^{s_i(k)}) & \\ x_i^{(k+1)} = x_i^{(k)} & if\, i \notin J(k) \\ \tilde{u}^{(k+1)} = (x_1^{(k+1)}, \cdots, x_p^{(k+1)}) & \end{array},$$

where $\{J(k)\}_{k \in N}$ is a sequence of nonempty subsets of $\{1, \cdots, p\}$. In fact, $J(k)$ is the set of subvectors to be updated at step k. Here,

$$S = \{s_1(k), \cdots, s_p(k)\}$$

is a sequence of elements of N^p with the following properties:

$$s_i(k) \leq k \qquad \forall k \in \boldsymbol{N} \qquad \forall i \in \{1, \cdots, p\};$$

$$\lim_{k \to \infty} s_i(k) = \infty, \qquad \forall i \in \{1, \cdots, p\},$$

$$i \text{ occurs infinitely often in the sets } J(k).$$

Such procedure is called chaotic relaxation Schwarz (CRS) algorithm and identified by $(\phi_i, \tilde{u}^{(0)}, J, S)$.

6.2 Algebraic Additive and Multiplicative Schwarz Methods

Now we give several special cases of CRS algorithm by selecting the set $J(k)$ and set $S = \{s_1(k), \cdots, s_p(k)\}$.

Algebraic Multiplicative Schwarz (AMS) Algorithm

If $s_i(k) = k$ and $J(K) = 1 + k \mod (p)$, for all k, then the CRS method is called Algebraic Multiplicative Schwarz (AMS) Algorithm.

Algebraic Additive Schwarz (AAS) Algorithm

When $s_i(k) = k$ and $J(K) = \{1, \cdots, p\}$ for all k, we call the CRS method as the algebraic additive Schwarz (ASA) method.

6.2.1 Direct Subsolver for All Sub-problems

We use direct solver for the subproblems in our CRS method. In order to analyze the convergence factor, we discuss the eigenvalue distribution of iterative matrix of AMS and AAS methods. Let us first express $\tilde{A}$ as the sum of block diagonal matrix $\tilde{D}$, block lower triangular matrix $\tilde{L}$ and block upper triangular matrix $\tilde{U}$,

$$\tilde{A} = \tilde{D} + \tilde{L} + \tilde{U}.$$

Here

$$\tilde{D} = \begin{bmatrix} \tilde{A}_{1,1} & & \\ & \ddots & \\ & & \tilde{A}_{p,p} \end{bmatrix}$$

$$\tilde{L} = \begin{bmatrix} 0 & & & \\ \tilde{A}_{2,1} & 0 & & \\ \vdots & \ddots & \ddots & \\ \tilde{A}_{p,1} & \cdots & \tilde{A}_{p,p-1} & \end{bmatrix} \qquad \tilde{U} = \begin{bmatrix} 0 & \tilde{A}_{1,2} & \cdots & \tilde{A}_{1,p} \\ & \ddots & \ddots & \vdots \\ & & 0 & \tilde{A}_{p-1,p} \\ & & & 0 \end{bmatrix}.$$

Assume for $k = 1, 2, \cdots,,$ we define a new sequence $y^{(k)}$ by

$$y^{(k)} = \begin{bmatrix} x_1^{(k-1)p+1} \\ \vdots \\ x_p^{kp} \end{bmatrix} \qquad \text{with } y^{(0)} = \tilde{u}^{(0)}$$

If p subproblems are all solved by the direct solver ϕ_i with $\omega_i = \omega$, the AMS method can be described in one simple form

$$\tilde{H}y^{(k+1)} = \tilde{B}y^{(k)} + \omega\tilde{f}$$

where

$$\tilde{H} = \tilde{D} + \omega\tilde{L} \qquad \text{and} \qquad \tilde{B} = ((1-\omega)\tilde{D} - \omega\tilde{U}),$$

and the AAS method can be rewritten as

$$\tilde{D}\tilde{u}^{(k+1)} = (\tilde{D} - \omega\tilde{A})\tilde{u}^{(k)} + \omega\tilde{f}.$$

Then the iterative matrices of the AMS method and AAS method satisfy the following theorem.

Theorem 6.3 *If $A_{i,2j} = 0$, for $|i - 2j| \geq 2$, then*

(i) The iterative matrix of the AMS method satisfies

$$\lambda\{\tilde{H}^{-1}\tilde{B}\} \subseteq \lambda\{(\dot{D} + \omega\dot{L})^{-1}((1-\omega)\dot{D} - \omega\dot{U})\}$$

and

$$\lambda\{(\dot{D} + \omega\dot{L})^{-1}((1-\omega)\dot{D} - \omega\dot{U})\} \setminus \{1-\omega\} \subseteq \lambda\{\tilde{H}^{-1}\tilde{B}\}.$$

(ii) The iterative matrix of the AAS method satisfies

$$\lambda\{\tilde{D}^{-1}(\tilde{D} - \omega\tilde{A}\} \subseteq \lambda\{\dot{D}^{-1}(\dot{D} - \omega\dot{A})\},$$

and

$$\lambda\{\dot{D}^{-1}(\dot{D} - \omega\dot{A})\} \setminus \{1-\omega\} \subseteq \lambda\{\tilde{D}^{-1}(\tilde{D} - \omega\tilde{A})\}.$$

Here the block diagonal matrix $\dot{D}$, lower triangular matrix $\dot{L}$ and upper triangular matrix $\dot{U}$ in the sum expression

$$\dot{A} = \dot{D} + \dot{L} + \dot{U}$$

have the forms:

$$\dot{D} = \begin{bmatrix} \dot{A}_{1,1} & & & \\ & \dot{A}_{3,3} & & \\ & & \ddots & \\ & & & \dot{A}_{2p-1.2p-1} \end{bmatrix}$$

$$\dot{L} = \begin{bmatrix} 0 & & & \\ \dot{A}_{3,1} & 0 & & \\ \dot{A}_{5,1} & \dot{A}_{5,3} & 0 & \\ \vdots & \ddots & \ddots & \\ \dot{A}_{2p-1,1} & \cdots & \dot{A}_{2p-1,2p-3} & 0 \end{bmatrix} \qquad \dot{U} = \begin{bmatrix} 0 & \dot{A}_{1,3} & \cdots & \dot{A}_{1,2p-1} \\ & \ddots & \cdots & \vdots \\ & & \dot{A}_{2p-5,2p-3} & \dot{A}_{2p-5,2p-1} \\ & & 0 & \dot{A}_{2p-3,2p-1} \\ & & & 0 \end{bmatrix},$$

where

$$\dot{A}_{1,2j-1} = A_{1,2j-1} - A_{1,2}A_{2,2}^{-1}A_{2,2j-1} \qquad \textit{for} \quad 1 \le j \le p,$$

$$\dot{A}_{2p-1,2j-1} = A_{2p-1,2j-1} - A_{2p-1,2p-2}A_{2p-2,2p-2}^{-1}A_{2p-2,2j-1} \qquad \textit{for} \quad 1 \le j \le p,$$

and for $\quad 1 < i < p, \quad 1 \le j \le p,$

$$\dot{A}_{2i-1,2j-1} = A_{2i-1,2j-1} - A_{2i-1,2i-2}A_{2i-2,2i-2}^{-1}A_{2i-2,2j-1} - A_{2i-1,2i}A_{2i,2i}^{-1}A_{2i,2j-1}.$$

Proof *Suppose that* $\lambda \in \lambda\{\tilde{H}^{-1}\tilde{B}\}$ *and* $\tilde{u}$ *is the corresponding eigenvector, i.e.* $\tilde{H}^{-1}\tilde{B}\tilde{u} = \lambda\tilde{u}$. *So* $\tilde{B}\tilde{u} = \lambda\tilde{H}\tilde{u}$. *Let* $\dot{u} = (u_1^T, \cdots, u_{2p-1}^T)^T$ *be the vector defined by the subvectors of* $\tilde{u}$. *Then, we have*

$$\lambda(\dot{D} + \omega\dot{L})\dot{u} = ((1-\omega)\dot{D} - \omega\dot{U})\dot{u}.$$

Hence, $\lambda \in \lambda\{(\dot{D} + \omega\dot{L})^{-1}((1-\omega)\dot{D} - \omega\dot{U})\}$. *Thus,*

$$\lambda\{\tilde{H}^{-1}\tilde{B}\} \subseteq \lambda\{(\dot{D} + \omega\dot{L})^{-1}((1-\omega)\dot{D} - \omega\dot{U})\}.$$

Let $\lambda \in \lambda\{(\dot{D}+\omega\dot{L})^{-1}((1-\omega)\dot{D}-\omega\dot{U})\}$ and $\dot{u}$ be the corresponding eigenvector, i.e.

$$\lambda(\dot{D}+\omega\dot{L})\dot{u} = ((1-\omega)\dot{D}-\omega\dot{U})\dot{u}.$$

Now we construct an eigenvector $\tilde{u}$ of $\tilde{H}^{-1}\tilde{B}$ from this eigenvector $\dot{u}$. Let the subvectors of $\tilde{u}$ be defined by the corresponding subvectors of $\dot{u}$. The other subvectors of $\tilde{u}$ are uniquely determined by the equation $\lambda\tilde{H}\tilde{u} = \tilde{B}\tilde{u}$, if $\lambda \neq 1-\omega$. It follows that $\lambda\{\tilde{H}^{-1}\tilde{B}\}$. Thus,

$$\lambda\{(\dot{D}+\omega\dot{L})^{-1}((1-\omega)\dot{D}-\omega\dot{U})\} \setminus \{1-\omega\} \subseteq \lambda\{\tilde{H}^{-1}\tilde{B}\}.$$

Remarks: It is obvious that, the larger the dimension of $A_{2i,2i}$ is, the smaller the dimension of $\dot{A}$ is. Then, the small dimension of $\dot{A}$ usually implied that the spectrum of the matrix $(\dot{D}+\omega\dot{L})^{-1}((1-\omega)\dot{D}-\omega\dot{U})$ is small. Hence, we can conclude that, the larger the overlapping is , the quicker the method converges. The assumption of Theorem 6.3 implies that the matrix A is not too dense.

We rewrite some of the block submatrices of A as follows:

$$A_{1,1} = \begin{bmatrix} \bar{A}_{0,0} & \bar{A}_{0,1} \\ \bar{A}_{1,0} & \bar{A}_{1,1} \end{bmatrix}, \qquad A_{2p-1,2p-1} = \begin{bmatrix} \bar{A}_{2p-1,2p-1} & \bar{A}_{2p-1,2p} \\ \bar{A}_{2p,2p-1} & \bar{A}_{2p,2p} \end{bmatrix},$$

$$A_{2p-1,1} = \begin{bmatrix} \bar{A}_{2p-1,0} & \bar{A}_{2p-1,1} \\ \bar{A}_{2p,0} & \bar{A}_{2p,1} \end{bmatrix}, \qquad A_{1,2p-1} = \begin{bmatrix} \bar{A}_{0,2p-1} & \bar{A}_{0,2p} \\ \bar{A}_{1,2p-1} & \bar{A}_{1,2p} \end{bmatrix},$$

$$A_{1,i} = \begin{bmatrix} \bar{A}_{0,i} \\ \bar{A}_{1,i} \end{bmatrix}, \quad A_{i,1} = (\bar{A}_{i,0}, \bar{A}_{i,1}) \quad A_{2p-1,i} = \begin{bmatrix} \bar{A}_{2p-1,i} \\ \bar{A}_{2p,i} \end{bmatrix},$$

for $1 < i < 2p-1$, and for $1 < i < 2p-1$

$$A_{i,2p-1} = (\bar{A}_{i,2p-1}, \bar{A}_{i,2p}) \qquad u_1 = \begin{bmatrix} \bar{u}_0 \\ \\ \bar{u}_1 \end{bmatrix} \quad u_{2p-1} = \begin{bmatrix} \bar{u}_{2p-1} \\ \\ \bar{u}_{2p} \end{bmatrix}.$$

After these rewriting of those sub-matrices along the boundary of matrix A, we can obtain the following theorem, which requires less zero sub-matrices in its assumption. The proof of this theorem is very similar to that of Theorem 6.3. So the proof is omitted.

Theorem 6.4 *If*

$$A_{i,2j} = 0 \quad \textit{for } |i-2j| \geq 2, \qquad \bar{A}_{i,0} = 0 \quad \textit{for } 2 \leq i \leq 2p$$

and

$$\bar{A}_{i,2p} = 0 \qquad \textit{for} \quad 1 \leq i \leq 2p-2,$$

then

(i) The iterative matrix of the AMS method satisfies

$$\lambda\{\tilde{H}^{-1}\tilde{B}\} \subseteq \lambda\{(\ddot{D} + \omega\ddot{L})^{-1}((1-\omega)\ddot{D} - \omega\ddot{U})\}$$

and

$$\lambda\{(\ddot{D} + \omega\ddot{L})^{-1}((1-\omega)\ddot{D} - \omega\ddot{U})\} \setminus \{1-\omega\} \subseteq \lambda\{\tilde{H}^{-1}\tilde{B}\}.$$

(ii) The iterative matrix of the AAS method satisfies

$$\lambda\{\tilde{D}^{-1}(\tilde{D} - \omega\tilde{A}\} \subseteq \lambda\{\ddot{D}^{-1}(\ddot{D} - \omega\ddot{A})\},$$

and

$$\lambda\{\ddot{D}^{-1}(\ddot{D} - \omega\ddot{A})\} \setminus \{1-\omega\} \subseteq \lambda\{\tilde{D}^{-1}(\tilde{D} - \omega\tilde{A})\}.$$

Here the block diagonal matrix $\ddot{D}$, lower triangular matrix $\ddot{L}$ and upper triangular matrix $\ddot{U}$ in the sum expression

$$\ddot{A} = \ddot{D} + \ddot{L} + \ddot{U}$$

have the forms:

$$\ddot{D} = \begin{bmatrix} \ddot{A}_{1,1} & & & \\ & \ddot{A}_{3,3} & & \\ & & \ddots & \\ & & & \ddot{A}_{2p-1.2p-1} \end{bmatrix}$$

$$\ddot{L} = \begin{bmatrix} 0 & & & \\ \ddot{A}_{3,1} & 0 & & \\ \ddot{A}_{5,1} & \ddot{A}_{5,3} & 0 & \\ \vdots & \ddots & \ddots & \\ \ddot{A}_{2p-1,1} & \cdots & \ddot{A}_{2p-1,2p-3} & 0 \end{bmatrix} \qquad \ddot{U} = \begin{bmatrix} 0 & \ddot{A}_{1,3} & \cdots & \ddot{A}_{1,2p-1} \\ & \ddots & \cdots & \vdots \\ & & \ddot{A}_{2p-5,2p-3} & \ddot{A}_{2p-5,2p-1} \\ & & 0 & \ddot{A}_{2p-3,2p-1} \\ & & & 0 \end{bmatrix},$$

where, for $1 \le i \le p$,

$$\ddot{A}_{1,2i-1} = \bar{A}_{1,2i-1} - \bar{A}_{1,0}\bar{A}_{0,0}^{-1}\bar{A}_{0,2i-1} - \bar{A}_{1,2}A_{2,2}^{-1}A_{2,2i-1},$$

$$\ddot{A}_{2p-1,2i-1} = \bar{A}_{2p-1,2i-1} - \bar{A}_{2p-1,2p}\bar{A}_{2p,2p}^{-1}\bar{A}_{2p,2i-1} - \bar{A}_{2p-1,2p-2}A^{-1}_{2p-2,2p-2}A_{2p-2,2i-1},$$

and for $2 \le i \le p-1, \qquad m = 1$ or $2p-1$,

$$\ddot{A}_{2i-1,m} = \bar{A}_{2i-1,m} - A_{2i-1,2i-2}A^{-1}_{2i-2,2i-2}\bar{A}_{2i-2,m} - A_{2i-1,2i}A^{-1}_{2i,2i}\bar{A}_{2i,m}.$$

From Theorem 6.3 and 6.4, we established the eigenvalue relation between the matrix A and $\dot{A}$ or $\ddot{A}$. In order to obtain the convergence factors of the AMS method and AAS method, we further discuss the relation between A and $\dot{A}$, as well as between A and $\ddot{A}$ in the following lemma.

Lemma 6.1 *Suppose that*

$$A_{i,2j} = 0 \qquad \textit{and} \qquad A_{2j,i} = 0 \qquad \textit{for } |i-2j| \geq 2.$$

Then, we have the following results:

(i) If A is symmetric, then $\dot{A}$ is symmetric.

If A is positive definite, then $\dot{A}$ is positive definite .

(ii) Assume, for $2 \leq i \leq 2p$, $\quad \bar{A}_{i,0} = 0, \quad \bar{A}_{0,i} = 0,$

and for $1 \leq i \leq 2p-2, \quad \bar{A}_{i,2p} = 0 \quad \bar{A}_{2p,i} = 0.$

If A is symmetric, then $\ddot{A}$ is symmetric.

If A is positive definite, then $\ddot{A}$ is positive definite .

(iii) If A is an M-matrix, then $\dot{A}$ is an M-matrix.

(iv) Assume , for $2 \leq i \leq 2p$, $\quad \bar{A}_{i,0} = 0, \quad \bar{A}_{0,i} = 0,$

and for $1 \leq i \leq 2p-2, \quad \bar{A}_{i,2p} = 0 \quad \bar{A}_{2p,i} = 0.$

If A is an M-matrix, then $\ddot{A}$ is an M-matrix.

Proof *Since the proof of (ii) and (iv) is similar to that of (i) and (iii), we only prove (i) and (ii) here.*

(i) Because $A_{i,2j} = 0$ for $|i-2j| \geq 2$, it is obvious that $\dot{A}$ is symmetric.

From $\dot{u}^T \dot{A} \dot{u} = \sum_{i,j=1}^p \dot{u}_{2i-1}^T \dot{A}_{2i-1,2j-1} \dot{u}_{2j-1}$, we let

$$u_{2i} = -A_{2i,2i}^{-1}(A_{2i,2i-1}\dot{u}_{2i-1} + A_{2i,2i+1}\dot{u}_{2i+1}),$$

and construct a vector u by putting these subvectors $\dot{u}_{2i-1}$ and u_{2i} according to the order of their index. Then, for this new vector u , it is easy to see that

$$\dot{u}^T \dot{A} \dot{u} = \sum_{i,j=1}^{p} \dot{u}_{2i-1}^T \dot{A}_{2i-1,2j-1} \dot{u}_{2j-1} = \sum_{i,j=1}^{2p-1} u_i^T A_{i,j} u_j.$$

Hence, the positive definite of A implies the positive definite of $\dot{A}$.

(iii) For any

$$\dot{b} = (\dot{b}_1^T, \dot{b}_3^T, \cdots, \dot{b}_{2p-1}^T)^T \geq 0,$$

there exists a vector $\dot{u}$ such that $\dot{A}\dot{u} = \dot{b}$. Let u be a vector whose subvectors are defined from the vector of $\dot{u}$, and the solutions u_{2i} of

$$A_{2i,2i}u_{2i} = -(A_{2i,2i-1}\dot{u}_{2i-1} + A_{2i,2i+1}\dot{u}_{2i+1}).$$

Then this new vector u satisfies

$$Au = b \qquad \text{where} \qquad b = (\dot{b}_1^T, 0, \dot{b}_3^T, 0, \cdots, 0, \dot{b}_{2p-1}^T)^T.$$

Since $A^{-1} \geq 0$, and $b \geq 0$, it follows that $u = A^{-1}b \geq 0$. Thus, $\dot{A}^{-1} \geq 0$. Note that $\dot{A}$ is positive definite. So the diagonal elements of $\dot{A}$ must be positive. Now we prove that the off-diagonal elements of $\dot{A}$ are negative. Let $\dot{u}$ be the vector that only one component of $\dot{u}$ is one and the other components are zero. Let

$$\dot{A}\dot{u} = \dot{b} \qquad \text{where } \dot{b} = (\dot{b}_1^T, \dot{b}_3^T, \cdots, \dot{b}_{2p-1}^T)^T,$$

and

$$A_{2i,2i}u_{2i} = -(A_{2i,2i-1}\dot{u}_{2i-1} + A_{2i,2i+1}\dot{u}_{2i+1}) \qquad \textit{for } 1 \le i \le p-1.$$

Because $A_{2i,2i-1}\dot{u}_{2i-1} + A_{2i,2i+1}\dot{u}_{2i+1} \le 0$ *and* $A^{-1}_{2i,2i} \ge 0$*, we have* $u_{2i} \ge 0$ *for* $1 \le i \le p-1$*. Then a vector* u *is defined and* $Au = b$*, with* $b = (\dot{b}_1^T, 0, \dot{b}_3^T, 0, \cdots, 0, \dot{b}_{2p-1}^T)^T$*. From this equation, we can obtain that the component of* $\dot{b}$ *corresponding to the nonzero components of* $\dot{u}$ *must be strictly positive and other components of* $\dot{b}$ *are negative. Hence, the off-diagonal elements of* $\dot{A}$ *are negative. Then* $\dot{A}$ *is an M-matrix.*

Remarks: By using the Varga's results [68], we can use the spectrum ρ of the block Jacobi iterative matrices of $\dot{A}$ and $\ddot{A}$ to get the optimal

$$\omega_{opt} = \frac{2}{1 + \sqrt{1-\rho^2}}.$$

This choice makes the convergence factor $\lambda = \omega - 1$ minimum. Since $\rho < 1$, we prefer to choose $1 < \omega < 2$ in AMS and AAS methods in practice.

6.2.2 Iterative Subsolver for All Sub-problems

In this section, we assume that all subproblems are solved by iterative methods.

We write the matrix A as the sum of diagonal matrix D, lower triangular matrix L and upper triangular matrix U,

$$A = D + L + U = D + C,$$

where

$$D = \begin{bmatrix} D_{1,1} & & \\ & \ddots & \\ & & D_{2p-1,2p-1} \end{bmatrix} \qquad L = \begin{bmatrix} L_{1,1} & & & \\ A_{2,1} & L_{2,2} & & \\ \vdots & \ddots & \ddots & \\ A_{2p-1,1} & \cdots & A_{2p-1,2p-2} & L_{2p-1,2p-1} \end{bmatrix},$$

and

$$U = \begin{bmatrix} U_{1,1} & A_{1,2} & \cdots & A_{1,2p-1} \\ & U_{2,2} & \ddots & \vdots \\ & & \ddots & A_{2p-2,2p-1} \\ & & & U_{2p-1,2p-1} \end{bmatrix}.$$

Then, the matrix $\tilde{A}$ has a corresponding decomposition denoted as,

$$\tilde{A} = \hat{D} + \hat{L} + \hat{U} = \hat{D} + \hat{C},$$

where

$$\hat{D} = \begin{bmatrix} \hat{D}_{1,1} & & \\ & \ddots & \\ & & \hat{D}_{p,p} \end{bmatrix} \qquad \hat{L} = \begin{bmatrix} \hat{L}_{1,1} & & & \\ \tilde{A}_{2,1} & \hat{L}_{2,2} & & \\ \vdots & \ddots & \ddots & \\ \tilde{A}_{p,1} & \cdots & \tilde{A}_{p,p-1} & \hat{L}_{p,p} \end{bmatrix},$$

and

$$\hat{U} = \begin{bmatrix} \hat{U}_{1,1} & \tilde{A}_{1,2} & \cdots & \tilde{A}_{1,p} \\ & \hat{U}_{2,2} & \ddots & \vdots \\ & & \ddots & \tilde{A}_{p-1,p} \\ & & & \hat{U}_{p,p} \end{bmatrix}.$$

Here we let

$$\hat{D}_{1,1} = \begin{bmatrix} D_{1,1} & \\ & D_{2,2} \end{bmatrix}, \qquad \hat{D}_{p,p} = \begin{bmatrix} D_{2p-2,2p-2} & \\ & D_{2p-1,2p-1} \end{bmatrix}$$

$$\hat{L}_{1,1} = \begin{bmatrix} L_{1,1} & \\ A_{2,1} & L_{2,2} \end{bmatrix}, \qquad \hat{L}_{p,p} = \begin{bmatrix} L_{2p-2,2p-2} & \\ A_{2p-1,2p-2} & L_{2p-1,2p-1} \end{bmatrix}$$

$$\hat{U}_1 = \begin{bmatrix} U_{1,1} & A_{1,2} \\ & U_{2,2} \end{bmatrix}, \qquad \hat{U}_p = \begin{bmatrix} U_{2p-2,2p-2} & A_{2p-2,2p-1} \\ & U_{2p-1,2p-1} \end{bmatrix}$$

$$\hat{D}_i = \begin{bmatrix} D_{2i-2,2i-2} & & \\ & D_{2i-1,2i-1} & \\ & & D_{2i,2i} \end{bmatrix},$$

$$\hat{L}_i = \begin{bmatrix} L_{2i-2,2i-2} & & \\ A_{2i-1,2i-2} & L_{2i-1,2i-1} & \\ A_{2i,2i-2} & A_{2i,2i-1} & L_{2i,2i} \end{bmatrix},$$

and

$$\hat{U}_i = \begin{bmatrix} U_{2i-2,2i-2} & A_{2i-2,2i-1} & A_{2i-2,2i} \\ & U_{2i-1,2i-1} & A_{2i-1,2i} \\ & & U_{2i,2i} \end{bmatrix} \qquad \text{for } 2 \le i \le p-1.$$

Suppose that p subproblems are solved by using *point Jacobi iterative method* with $\omega_i = \omega$ for $i = 1, \cdots, p$. Then, the AAS algorithm can be written in the form:

$$\hat{D}\tilde{u}^{(k+1)} = (\hat{D} - \omega\tilde{A})\tilde{u}^{(k)} + \omega f.$$

The AMS method can be represented by

$$(\hat{D}+\omega\tilde{L})y^{(k+1)} = (\hat{D}-\omega(\tilde{D}+\tilde{U}))y^{(k)}+\omega f.$$

Theorem 6.5 *Suppose that $D_{i,i}$ are invertible for $i=1,\cdots,2p-1$. Then we have*

$$\lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\} = \lambda\{D^{-1}(D-\omega A)\}(\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}).$$

Proof *We first show*

$$\lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\} \subseteq \lambda\{D^{-1}(D-\omega A)\}(\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}).$$

Suppose that $\lambda \in \lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\}$. and $\tilde{u}$ is the corresponding eigenvector, i.e.

$$\lambda\hat{D}\tilde{u} = (\hat{D}-\omega\tilde{A})\tilde{u}.$$

If $u_{2i} \neq \tilde{u}_{2i}$, for some $1 \leq i \leq p-1$, then we have

$$\lambda \in \lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}.$$

If $u_{2i} = \tilde{u}_{2i}$, for all $i=1,\cdots,p-1$, then we have

$$\lambda \in \lambda\{D^{-1}(D-\omega A)\}.$$

Therefore, we already show that

$$\lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\} \subseteq \lambda\{D^{-1}(D-\omega A)\}(\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}).$$

Now we prove that

$$\lambda\{D^{-1}(D-\omega A)\}(\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}) \subseteq \lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\}.$$

Assume that

$$\lambda \in \lambda\{D^{-1}(D-\omega A)\}(\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\}).$$

If $\lambda \in \lambda\{D^{-1}(D-\omega A)\}$ *and* u *is the associated eigenvector, i.e.* $-\lambda D u + (D - \omega A)u = 0$, *then we construct an eigenvector* $\tilde{u}$ *of* $\hat{D}^{-1}(\hat{D}-\omega\tilde{A})$. *from the eigenvector* u *through letting* $\tilde{u}_{2i} = u_{2i}$ *for* $i = 1, \cdots, p-1$. *If* $\lambda \in (\cup_{i=1}^{p}\lambda\{D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})\})$, *and* $\lambda \notin \lambda\{D^{-1}(D-\omega A)\}$, *then an eigenvector* $\tilde{u}$ *of matrix* $\hat{D}^{-1}(\hat{D}-\omega\tilde{A})$ *is constructed by the following procedure. Assume that* v_{2i} *is the solution of equation*

$$-\lambda D_{2i}v_{2i} + (D_{2i} - \omega A_{2i,2i})v_{2i} = 0,$$

i.e. v_{2i} *is the eigenvector of* $D_{2i}^{-1}(D_{2i}-\omega A_{2i,2i})$. *By solving equation*

$$-\lambda D u + (D-\omega A)u = w$$

where $w = (0^T, \cdots, 0^T, \omega v_{2i}^T A_{2i+1,2i}, \cdots, \omega v_{2i}^T A_{2p-1,2i}^T$, *we obtain a vector* u. *Define* $\tilde{u}_{2i} = v_{2i} + u_{2i}$. *The other subvectors of* $\tilde{u}$ *are defined by* $\tilde{u}_{2j} = u_{2j}$ *for* $j = 1, \cdots, i-1, i+1 \cdots, p-1$. *Then, this* $\tilde{u}$ *satisfies*

$$-\lambda\hat{D}\tilde{u} + (\hat{D}-\omega\tilde{A})\tilde{u} = 0.$$

Therefore, $\lambda \in \lambda\{\hat{D}^{-1}(\hat{D}-\omega\tilde{A})\}$.

Theorem 6.6 *Suppose that A is an M-matrix,*

$$\left(D + \omega \begin{bmatrix} 0 & & & \\ A_{2,1} & & & \\ \vdots & & & \\ A_{2p-1,1} & \cdots & A_{2p-1,2p-2} & 0 \end{bmatrix} \right)^{-1}$$

exists and is nonnegative, and matrix

$$D - \omega \begin{bmatrix} A_{1,1} & \cdots & A_{1,2p-1} \\ & \ddots & \vdots \\ & & A_{2p-1,2p-1} \end{bmatrix}$$

is nonnegative. Then

$$\tilde{A}^{-1} \qquad \text{and} \qquad (\hat{D} + \omega \tilde{L})^{-1}$$

exist and are nonnegative, and

$$\hat{D} - \omega(\tilde{D} + \tilde{U})$$

is nonnegative. So the spectrum of the AMS iterative matrix

$$(\hat{D} + \omega \tilde{L})^{-1}(\hat{D} - \omega(\tilde{D} + \tilde{U}))$$

is less than 1.

The proof of this theorem is similar to that of Lemma 6.1.

Let all subproblems be solved by *the SOR method* with $\omega_i = 1, \quad i = 1, \cdots, p$.

Then, the AAS method can be expressed by

$$\left(\hat{D} + \omega \begin{bmatrix} \hat{L}_{1,1} & & \\ & \ddots & \\ & & \hat{L}_{p,p} \end{bmatrix}\right) u^{(k+1)} = ((1-\omega)\hat{D} - \omega(\tilde{L} + \hat{U}))u^{(k)} + \omega f.$$

Hence, the corresponding iterative matrix is

$$\hat{J} = \left(\hat{D} + \omega \begin{bmatrix} \hat{L}_{1,1} & & \\ & \ddots & \\ & & \hat{L}_{p,p} \end{bmatrix}\right)^{-1} ((1-\omega)\hat{D} - \omega(\tilde{L} + \hat{U})).$$

The AMS method can also be written in the simple form:

$$(\hat{D} + \omega\hat{L})y^{(k+1)} = ((1-\omega)\hat{D} - \omega\hat{U})y^{(k)} + \omega f,$$

with the iterative matrix

$$\hat{S} = (\hat{D} + \omega\hat{L})^{-1}((1-\omega)\hat{D} - \omega\hat{U}).$$

Theorem 6.7 *Assume that A is an M-matrix,*

$$\left(D + \omega \begin{bmatrix} L_{1,1} & & \\ & \ddots & \\ & & L_{2p-1,2p-1} \end{bmatrix}\right)^{-1}$$

exists and is nonnegative, and

$$(1-\omega)D-\omega\left[\begin{array}{llll} 0 & & & \\ A_{2,1} & & & \\ \vdots & \ddots & \ddots & \\ A_{2p-1,1} & \cdots & A_{2p-1,2p-2} & 0 \end{array}\right)-\omega U$$

is nonnegative. Then

$$\tilde{A}^{-1} \qquad \textit{and} \qquad \left(\hat{D}+\omega\left[\begin{array}{ccc} \hat{L}_{1,1} & & \\ & \ddots & \\ & & \hat{L}_{p,p} \end{array}\right]\right)^{-1}$$

exist and are nonnegative. So the spectrum of the AAS iterative matrix $\hat{J}$ is strictly less than 1.

The proof is similar to that of Lemma 6.1.

Denote

$$M=(D+\omega L)^{-1}((1-\omega)D-\omega U),$$

and

$$M_{2i,2i}=(D_{2i,2i}+\omega L_{2i,2i})^{-1}((1-\omega)D_{2i,2i}-\omega U_{2i,2i}).$$

Theorem 6.8 *Assume that all $D_{i,i}$ are nonsingular. Then, we have*

$$\lambda\{\hat{S}\}=\lambda\{M\}\cup(\cup_{i=1}^{p-1}\lambda\{M_{2i,2i}\}).$$

Proof *We first prove that*

$$\lambda\{\hat{S}\} \subseteq \lambda\{M\} \cup (\cup_{i=1}^{p-1} \lambda\{M_{2i,2i}\}).$$

Assume that $\lambda \in \lambda\{\hat{S}\}$, *and* $\tilde{u}$ *be the corresponding eigenvector, i.e.*

$$\lambda\tilde{u} = \hat{S}\tilde{u} \qquad \lambda(\hat{D} + \omega\hat{L})\tilde{u} = ((1-\omega)\hat{D} - \omega\hat{U})\tilde{u}.$$

If $u_{2i} \neq \tilde{u}_{2i}$ *for some* $1 \leq i \leq p-1$, *then* $\lambda \in \lambda\{M_{2i,2i}\}$. *If* $u_{2i} = \tilde{u}_{2i}$ *for all* $1 \leq i \leq p-1$, *then* $\lambda \in \lambda\{M\}$. *Hence,*

$$\lambda\{\hat{S}\} \subseteq \lambda\{M\} \cup (\cup_{i=1}^{p-1} \lambda\{M_{2i,2i}\}).$$

Now we show that

$$\lambda\{M\} \cup (\cup_{i=1}^{p-1} \lambda\{M_{2i,2i}\}) \subseteq \lambda\{\hat{S}\}.$$

Let $\lambda \in \lambda\{M\}$ *and* u *be the associated eigenvector, i.e.*

$$\lambda u = Mu \qquad \text{and} \qquad \lambda(D + \omega L)u = ((1-\omega)D - \omega U)u.$$

Define $\tilde{u}$ *by letting* $\tilde{u}_{2i} = u_{2i}$. *Then, this* $\tilde{u}$ *is the eigenvector of* $\hat{S}$ *and* $\lambda\tilde{u} = \hat{S}\tilde{u}$. *Thus,*

$$\lambda\{M\} \subseteq \lambda\{\hat{S}\}.$$

Assume $\lambda \in \lambda\{M_{2i,2i}\}$, $\lambda \notin \lambda\{M\}$. *Denote* v_{2i} *to be the corresponding eigenvector, i.e.* $\lambda v_{2i} = M_{2i,2i}v_{2i}$. *We solve the following equation and get a solution vector* u *from*

$$(\lambda(D + \omega L) + \omega U + (\omega - 1)D)u = w,$$

where $w = -\lambda\omega(0^T, \cdots, 0^T, v_{2i}^T A_{2i+1,2i}, \cdots, v_{2i}^T A_{2p-1,2i})^T$. *Since* $\lambda \notin \lambda\{M\}$, *this problem has only one solution. We define a new vector tildeu by letting*

$$\tilde{u}_{2j} = \begin{cases} v_{2i} + u_{2i} & \text{for } j = i \\ u_{2j} & \text{for } j \neq i \end{cases} .$$

This $\tilde{u}$ *satisfies* $\lambda\tilde{u} = \hat{S}\tilde{u}$. *Hence,* $\tilde{u}$ *is the eigenvector of* $\hat{S}$. *The result of the theorem follows.*

Remarks: From above theorems and lemmas, we conclude that the convergence factors of the AAS method and AMS method are almost the same as the block Jacobi method and the SOR method. The reason is the subsolver provides only local information. So it is crucial to introduce a kind of coarse problem to provide the global communication, while designing DD methods for these problems.

Bibliography

[1] V. I. AGOSHKOV, *Poincaré-Steklov operators and domain decomposition methods in finite dimensional spaces*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., Philadelphia, 1988, SIAM.

[2] C. ANDERSON, *Manipulating fast solvers-changeing their boundary conditions and putting them on multiple processor computers*, Tech. Rep. CAM 88-37, Department of Mathematics, UCLA, 1988.

[3] R. BANK, T. F. CHAN, J. W.M. COUGHRAN, AND R. SMITH, *The alternate-block-factorization procedure for systems of partial differential equations*, Tech. Rep. CAM 89-13, Department of Mathematics, UCLA, 1989.

[4] R. BANK, D. ROSE, AND W.FICHTNER, *Numerical methods for semiconductor device simulation*, IEEE Trans. Electron Devices, ED-30 (1983), pp. 1031–1041.

[5] P. E. BJØRSTAD AND O. B. WIDLUND, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1093 – 1120.

[6] J.-F. Bourgat, R. Glowinski, P. Le Tallec, and M. Vidrascu, *Variational formulation and algorithm for trace operator in domain decomposition calculations*, in Domain Decomposition Methods, T. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., Philadelphia, 1988, SIAM.

[7] J. H. Bramble, R. E. Ewing, J. E. Pasciak, and A. H. Schatz, *A preconditioning technique for the efficient solution of problems with local grid refinement*, Comput. Meth. Appl. Mech. Engin., 67 (1988), pp. 149 –159.

[8] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring, I*, Math. Comp., 47 (1986), pp. 103– 134.

[9] ——, *An iterative method for elliptic problems on regions partitioned into substructures*, Math. Comp., 46 (1986), pp. 361–369.

[10] ——, *The construction of preconditioners for elliptic problems by substructuring, IV*, Math. Comp., 53 (1989), pp. 1 – 24.

[11] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decompositions*, Math. Comp., 57 (1991), pp. 1–21.

[12] X. C. Cai and O. B. Widlund, *Domain decomposition algorithms for indefinite elliptic problems*, SIAM J. Sci. Statist. Comput., 13 (1992).

[13] ——, *Multiplicative schwarz algorithm for some nonsymmetric and indefinite elliptic problems*, SIAM J. Numer. Anal., 30 (1993).

[14] T. Chan, R. Glowinski, , J. Périaux, and O. Widlund, eds., *Domain Decomposition Methods*, Philadelphia, 1989, SIAM. Proceedings of the Second International Symposium on Domain Decomposition Methods , Los Angeles, California , January 14 - 16, 1988.

[15] ——, eds., *Domain Decomposition Methods*, Philadelphia, 1990, SIAM. Proceedings of the Third International Symposium on Domain Decomposition Methods , Hoston, Texas, 1989.

[16] ——, eds., *Domain Decomposition Methods*, Philadelphia, 1991, SIAM. Proceedings of the Fourth International Symposium on Domain Decomposition Methods , Moscow, USSR, 1990.

[17] T. F. Chan, *Analysis of preconditioners for domain decomposition*, SIAM J. Numer. Anal., 24 (1987), pp. 382–390.

[18] T. F. Chan and D. Goovaerts, *A note on the efficiency of domain decomposed incomplete factorizations*, SIAM J. Sci. Stat. Comput., 11 (July 1990), pp. 794 – 803.

[19] T. F. Chan and T. Y. Hou, *Domain decomposition interface preconditioners for general 2nd order elliptic problems*, Tech. Rep. CAM 88-16, Department of Mathematics, UCLA, 1990.

[20] T. F. Chan and D. F. Keyes, *Interface preconditioning for domain-decomposed convection-diffusion operators*, tech. rep., CAM 89-28, Department of Mathematics, University of California Los Angeles, 1989. Pages 245 – 262, Proceedings of the Third International Symposium on Domain Decomposition Methods, Houston, Texas, April, 1989.

[21] T. F. Chan and T. P. Mathew, *An application of the probing technique to the vertex space method in domain decomposition*, Tech. Rep. CAM 90-22, Department of Mathematics, UCLA, 1990. To appear in Proceedings of 5th Domain Decomposition Conference, Moscow, April 1990.

[22] ——, *The interface probing technique in domain decomposition*, SIAM J. Matrix Analysis and Applications, 13 (1992).

[23] T. F. Chan, T. P. Mathew, and J. P. Shao, *Efficient variants of the vertex space domain decomposition algorithm*, Tech. Rep. CAM 92-07, Department of Mathematics, UCLA, 1992. To appear in SIAM J. Sci. Comp.

[24] T. F. Chan and D. C. Resasco, *A domain-decomposed fast Poisson solver on a rectangle*, Tech. Rep. /DCS/RR-409, Yale University, 1985.

[25] ——, *A survey of preconditioners for domain decomposition*, Tech. Rep. /DCS/RR-414, Yale University, 1985.

[26] T. F. Chan and J. Shao, *The optimal parallel complexity of domain decomposition*, tech. rep., Department of Mathematics, UCLA, 1993. In preparation.

[27] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, 1978.

[28] A. R. Curtis, M. J. Powell, and J. K. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Maths. Applics., 13 (1974), pp. 117–120.

[29] A. Q. D. Funaro and P. Zanolli, *An iterative procedure with interface relaxation for domain decomposition methods*, SIAM J. Numer. Anal., 25 (1988), pp. 1213 –1236.

[30] J. Donato, *Iterative methods for scale and coupled systems of elliptic elliptic equations*, Tech. Rep. CAM 91-20, Mathematics Department, UCLA, September 1991. UCLA doctoral dissertation.

[31] M. Dryja, *A capacitance matrix method for Dirichlet problem on polygon region*, Numer. Math., 39 (1982), pp. 51 – 64.

[32] M. Dryja and O. B. Widlund, *An additive variant of the Schwarz alternating method for the case of many subregions*, Tech. Rep. 339, also Ultracomputer Note 131, Department of Computer Science, Courant Institute, 1987.

[33] ——, *Some domain decomposition algorithms for elliptic problems*, in Proceedings of the Conference on Iterative Methods for Large Linear Systems held in Austin, Texas, October 1988, to celebrate the Sixty-fifth Birthday of David M. Young, Jr., Academic Press, Orlando, Florida, 1989., 1989.

[34] ——, *Towards a unified theory of domain decomposition algorithms for elliptic problems*, Tech. Rep. 486, also Ultracomputer Note 167, Department of Computer Science, Courant Institute, 1989.

[35] ——, *Additive schwarz methods for elliptic finite element problems in three dimensions*, in The Fifth Internatinal Symposium on Domain Decomposition Methods for Partial Diferential Equations, T. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, and R. G. Voigt, eds., Philadelphia, 1992, SIAM.

[36] S. C. EISENSTAT, Personal Communication, 1985.

[37] S. C. EISENSTAT, H. C. ELMAN, AND M. H. SCHULTZ, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20(2) (1983), pp. 345–357.

[38] R. GLOWINSKI, G. H. GOLUB, G. A. MEURANT, AND J. PÉRIAUX, eds., *Domain Decomposition Methods for Partial Differential Equations*, Philadelphia, 1988, SIAM. Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, Paris, France, January 1987.

[39] G. GOLUB AND D. MAYERS, *The use of preconditioning over irregular regions*, in Computing Methods in Applied Sciences and Engineering, VI, R. Glowinski and J. L. Lions, eds., Amsterdam, New York, Oxford, 1984, North-Holland, pp. 3–14. Proceedings of a conference held in Versailles,

France, December 12-16,1983.

[40] G. H. GOLUB AND C. F. V. LOAN, *Matrix Computations*, Johns Hopkins Univ. Press, 1989. Second Edition.

[41] W. D. GROPP, *Parallel computing and domain decomposition*, in Domain Decomposition Methods, T. Chan, D. E. Keyes, G. Meurant, J. S. Scroggs, and R. G. Voigt, eds., Philadelphia, 1992, SIAM.

[42] W. D. GROPP AND D. E. KEYES, *Domain decomposition with local mesh refinement*, Tech. Rep. RR-726, Yale University, Dept. of Comp. Sci., August 1989. Accepted for SIAM J. Sci. Stat. Comp.

[43] L. S. KANG, ed., *Parallel Algorithms and Domain Decomposition*, Wuhan, China, 1987, Wuhan University Press.

[44] D. E. KEYES, T. F. CHAN, G. MEURANT, J. S. SCROGGS, AND R. G. VOIGT, eds., *Domain Decomposition Methods*, Philadelphia, 1992, SIAM. Proceedings of the Fifth International Symposium on Domain Decomposition Methods, Norfolk, VA, 1991.

[45] D. E. KEYES AND W. D. GROPP, *A comparison of domain decomposition techniques for elliptic partial differential equations and their parallel implementation*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s166 – s202.

[46] ——, *Domain decomposition techniques for the parallel solution of nonsymmetric systems of elliptic bvps*, in Domain Decomposition Methods, T. Chan,

R. Glowinski, J. Périaux, and O. Widlund, eds., Philadelphia, 1989, SIAM.

[47] J. L. Lions and E. Magenes, *Nonhomogenous Boundary Value Problems and Applications*, vol. I, Springer, New York, Heidelberg, Berlin, 1972.

[48] P. L. Lions, *Interprétation stochastique de la méthode alternée de Schwarz*, C. R. Acad. Sci. Paris, 268 (1978), pp. 325 – 328.

[49] ——, *On the Schwarz alternating method. I.*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., Philadelphia, 1988, SIAM.

[50] ——, *On the Schwarz alternating method. II.*, in Domain Decomposition Methods, T. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., Philadelphia, 1989, SIAM.

[51] J. Mandel, *Iterative solvers by substructuring for the p-version finite element method*, Comput. Meth. Appl. Mech. Engin., (1989). To appear in a special issue as Proceedings of an International Conference on Spectral and High Order Methods, Como, Italy, June 1989.

[52] ——, *Two-level domain decomposition preconditioning for the p-version finite element version in three dimensions*, Int. J. Numer. Meth. Engin., (1989). To appear.

[53] A. M. Matsokin and S. V. Nepomnyaschikh, *A Schwarz alternating method in a subspace*, Soviet Mathematics, 29(10) (1985), pp. 78 – 84.

[54] S. V. Nepomnyaschikh, *On the application of the method of bordering for elliptic mixed boundary value problems and on the difference norms of* $w_2^{1/2}(s)$, Tech. Rep. 106, Computing Center of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk, 1984. In Russian.

[55] ——, *Domain decomposition and schwarz methods in a subspace for the approximate solution of elliptic boundary value problems*, tech. rep., Computing Center of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk, USSR, 1986. Ph.D thesis.

[56] ——, *Schwarz alternating method for solving the singular Neumann problem*, Soviet J. Numer. Anal. Math. Modelling, 5(1) (1990), pp. 69 – 78.

[57] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Academia, Prague, 1967.

[58] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley and Sons, New York, 1982.

[59] M. J. Powell and P. L. Toint, *On the estimation of sparse Hessian matrices*, SIAM J. Numerical Analysis, 16 (1979), pp. 1060–1074.

[60] Y. H. D. Roeck and P. L. Tallec, *Analysis and test of a local domain decomposition preconditioner*, in Fourth International Domain Decomposition

Methods for Partial Differential Equations, R. Glowinski, Y. A. Kuznetsov, G. Meurant, J. Périaux, and O. Widlund, eds., Philadelphia, 1991, SIAM.

[61] A. H. SCHATZ, *An observation concerning ritz-galerkin methods with indefinite forms*, Math. Comp., 28 (1974), pp. 959 – 962.

[62] H. A. SCHWARZ, *Gesammelete Mathematische Abhandlungen*, vol. 2, Springer, Berlin, 1890, pp. 133–143. First published in Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich, volume 15, 1870, pp. 272–286.

[63] J. P. SHAO, *The application of a domain decomposition method to solving singular Neumann boundary problems*, Tech. Rep. CAM 93-09, Department of Mathematics, UCLA, 1993.

[64] B. F. SMITH, *Domain decomposition algorithms for partial differential equations of linear elasticity*, Tech. Rep. 517, Department of Computer Science, Courant Institute, 1990 Sept. Ph.D thesis.

[65] ——, *A domain decomposition algorithm for elliptic problems in three dimensions*, Numer. Math., 60 (1991), pp. 219–234.

[66] ——, *An optimal domain decomposition preconditioner for the finite element solution of linear elasticity problems*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 364–378.

[67] G. STRANG, *Approximation in the finite element method*, Numer. Math., 19 (1972), pp. 81–98.

[68] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, 1962.

[69] W.FICHTNER, D. ROSE, AND R. BANK, *Semiconductor device simulation*, IEEE Trans. Electron Devices, ED-30 (Sept. 1983), pp. 1018–1041.

[70] O. B. WIDLUND, *Iterative substructuring methods: Algorithms and theory for elliptic problems in the plane*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., Philadelphia, 1988, SIAM.

[71] ——, *Optimal iterative refinement methods*, in Domain Decomposition Methods, T. Chan, R. Glowinski, J. Périaux, and O. Widlund, eds., Philadelphia, 1989, SIAM.

[72] ——, Personal Communication, 1990.

[73] ——, *Some schwarz methods for symmetric and nosymmetric elliptic problems*, in The Fifth Internatinal Symposium on Domain Decomposition Methods for Partial Diferential Equations, T. Chan, D. E. Keyes, G. A. Meurant, J. S. Scroggs, and R. G. Voigt, eds., Philadelphia, 1992, SIAM.

[74] J. XU, *Theory of Multilevel Methods*, PhD thesis, Cornell University, May 1989.

[75] ——, *Iterative methods by space decomposition and subspace correction*, SIAM Review, 34 (1992), pp. 581– 613.