ITERATIVE METHOD
FOR
MIXED FINITE ELEMENT FORMULATION
OF
BIHARMONIC EQUATION

AGAH D. GARNADI

**ITERATIVE METHOD
FOR
MIXED FINITE ELEMENT FORMULATION
OF BIHARMONIC EQUATION**

Agah D. Garnadi

February 1993

Submitted in partial fulfilment of the requirement
for the degree of Graduate Diploma in Science
the Department of Mathematics
the Australian National University

Acknowledgement

I would like to thank AIDAB, the Ministry of Foreign Affairs and International Trade, that made me possible to study at ANU. I'd like to thank my supervisor, Stephen G. Roberts, for his patience and time. Also I'd like to thank my friend, Ahmad Mukhlis whose in his remarks in one of our 'talk' sessions between Canberra-Madison, mentioned that matrix aspect of PDE solver is still an interesting areas for exploration. His remarks made me motivated and spun off the study with the results of the second part of this thesis. Lastly, I'd like to thank my family for their support, especially my two girls, who were willingly accompanied their parents with high hopes but eventually they must have to face the arduous life and high stress daily so far.

Contents

1	Finite Element Method: A Tutorial	6
1.1	Function Spaces and Finite Element Spaces	6
1.2	Basic Error Estimates	10
1.3	Finite Element Approximation: Model Problem	12
1.3.1	Variational Problem	12
1.3.2	Model Problem	13
1.4	Some Results on H^1 and L^2 Estimates for Poisson's Equation	15
2	The Biharmonic Equation: Source of the Problem	18
2.1	Plate Equation	18
2.2	Visual Surface Reconstruction	22
2.3	Low Reynolds Number Flows	24
3	Mixed Finite Element Formulation for the Biharmonic Equation	26
3.1	Variational Problem	26
3.2	Abstract Result	29
3.3	Application to First Biharmonic Equation	38
3.4	Application to Second Biharmonic Equation	43
4	Iterative Methods for Non-Singular Systems of Linear Equations	46
4.1	Conjugate Gradient	47
4.1.1	Convergence Behavior of Conjugate Gradient	50
4.2	Conjugate Gradient for Normal Equation	51
4.3	Diagonal Scaling as Preconditioner	52
4.4	BiConjugate Gradient	53

4.5	Conjugate Gradient Squared	55
4.6	BiCGSTAB	57
4.6.1	Preconditioning	59
4.7	Stopping Criteria for Conjugate Gradient Type Iteration . . .	61
5	Diagonal Scaling as Preconditioner for Preconditioned BiCGSTAB Methods to Solve Linear Equations from Saddle-Point Problems	62
5.1	Conditioning by Scaling of Saddle-point Problem	63
5.2	Eigenvalue Distribution of Saddle-point Problem	65
5.3	Rate Convergence Estimate of BiCGSTAB for Saddle-point Problem	67
5.4	Preconditioning of Saddle-Point Problem	69
5.5	Choice of Diagonal Preconditioner	70
6	Numerical Results of Iterative Methods to Solve the Biharmonic Equation	71
6.1	Review	72
6.2	Result on the First Biharmonic equation	73
6.3	Result on the Second Biharmonic equation	77
6.4	Concluding Remarks	84

Introduction

The main objective of first chapter is to introduce the basic ideas that underlie the finite element method. First we introduce finite element space in the setting of a larger function space of which the problem usually posed. Then we continue by restating some basic error estimates concerning of approximation of an element in a larger space with its interpolant in finite element space. To highlight how the finite element method to be applied, we give an application of finite element approximation to the second order linear partial differential equation as a model problem.

In Chapter 2 we present three sources of problem in which the resulting equation is a fourth order equation, specifically the biharmonic equation. The problems arises from a model of a plate deformation, reconstruction problem of surface of an object in computer vision process, and a slow flow fluid motion respectively.

In Chapter 3 we discuss the mixed finite element method for Biharmonic problem. First we will take a look at two variational formulations of the first Biharmonic problem. The first formulation is a direct variational formulation, and second is mixed variational formulation. It is shown that this formulation is a special case of abstract formulation that we discuss in the following section. This result is applied to the first and second Biharmonic problems in section 3 and 4 respectively.

In Chapter 4 we consider a family of methods to solve systems of linear equations which arise from the application of the finite element method. We seek an efficient solution method for these systems which are characterized by their sparsity and large number of unknowns. Iterative methods have the property that they can utilize the sparsity structure of the systems, and leave it unchanged. Thus the non-zero elements can be stored efficiently. In this chapter, a family of iterative methods are described, namely BCG, CGS, and BiCGSTAB. These methods are known to be parameter free. First we describe the classical Conjugate Gradient method for a symmetric positive definite linear equation and its extension to solve non-singular linear equations. Then we discuss preconditioning by diagonal scaling. The following sections describe three interrelated methods used to solve a nonsingular linear equation, and finally we close the chapter by discussing a stopping criteria for iterative processes.

Chapter 5 is arranged as follows. The first section concerns a particular type of diagonal preconditioner in relation with the condition number of a

particular form symmetric matrix, known as saddle-point problem. In the second section, we discuss the distribution of eigen-values of the matrix and its relationship with extreme singular-values of some blocks. The result of the second section will be used to estimate the rate of convergence of the BiCGSTAB algorithm when used to solve the saddle-point problem, this is discussed in the third section. The result of this section will be used in the fourth section to build a preconditioner to accelerate the iterative method. Finally, the last section gave the range of diagonal preconditioner for saddle-point problem based on the discussion previously.

The last chapter reports the result of solving the linear equation arising from discretization of the Biharmonic equation using mixed finite element formulation as proposed in Chapter 3. The CGS, BiCGSTAB algorithm and preconditioned variant of BiCGSTAB is employed to solve the linear equation.

Chapter 1

Finite Element Method: A Tutorial

The main objective of this chapter to introduce the basic ideas that underlie the finite element method.

The first two sections basically introduce terminology used in finite element approximation. The first section introduce finite element space in the setting of a larger function space of which the problem usually posed. The second section is restatement some basic error estimates concerning of approximation of an element in larger space from its interpolant in finite element space.

The last two section of this chapter we give an application of finite element approximation to the second order linear partial differential equation as a model problem, to highlight how the finite element method to be applied. Moreover, the results in these sections will be used to generate some results in Chapter 3.

1.1 Function Spaces and Finite Element Spaces

In order to introduce the finite element method, first we need to define some function spaces, and associated norms and forms operating on elements of those spaces.

Let Ω denote an open bounded subset of R^2 , and let $\partial\Omega$ denote its boundary. Denote by $L^2(\Omega)$ the space of functions that are square integrable over

Ω equipped with the following inner product and induced norm:

$$\langle u, v \rangle_0 = \int_{\Omega} uv, \forall u, v \in L_2(\Omega)$$

$$\|u\|_0 = \langle u, u \rangle_0^{\frac{1}{2}}, \forall u \in L_2(\Omega)$$

We define the Sobolev spaces:

$$H^k(\Omega) = \{u \in L^2(\Omega); D^s u \in L^2(\Omega), s = 1, \dots, k\}$$

where D^s denote partial derivatives up to order s . So $H^k(\Omega)$ consists of square integrable functions, with all derivatives up to order k belonging to $L^2(\Omega)$. This space is equipped with the norm:

$$\|u\|_k = \{\|u\|_0^2 + \sum_{s=1}^k \|D^s u\|_0^2\}^{1/2}$$

Also in this space we introduce k -th seminorm:

$$|u|_k = \{\sum_{s=k}^{\infty} \|D^s u\|_0^2\}^{1/2},$$

i.e the square root of sums of L^2 -norms for k -th derivatives of u .

Note that

$$H^0(\Omega) = L^2(\Omega)$$

In conjunction with a second order model problem (Laplace operator), of particular interest is the space $H^1(\Omega)$ and the subspace of $H^1(\Omega)$, with elements which vanish on the complement of Ω :

$$H_0^1(\Omega) = \{u \in H^1(\Omega); u|_{\partial\Omega} = 0\}$$

Since $H^k(\Omega)$ is a vector space with respect to pointwise addition and multiplication by scalar of real number, we may equip the function space with forms. First is the linear form defined as follows:

$$f(g) = \int_{\Omega} fg, \forall g \in H^k(\Omega)$$

for given function $f \in L^2(\Omega)$. Another class of forms of particular interest are the bilinear forms, which are defined as a mapping : $a : V_1 \times V_2 \mapsto R$, where V_1, V_2 are functions spaces of interest, where the map a satisfies:

$$a(u, \alpha v_1 + \beta v_2) = \alpha a(u, v_1) + \beta a(u, v_2), \forall \alpha, \beta \in R, \forall u \in V_1, \forall v_1, v_2 \in V_2$$

Figure 1.1: A Forbidden triangulation of a skewed square

$$a(\alpha u_1 + \beta u_2, v) = \alpha a(u_1, v) + \beta a(u_2, v), \forall \alpha, \beta \in R, \forall u_1, u_2 \in V_1, \forall v \in V_2$$

For example, let $V_1 = V_2 = H^1(\Omega)$ and one of particular interest of bilinear form a on $V_1 \times V_2$ is:

$$a(u, v) = \int \nabla u \nabla v, \forall u, v \in H^1(\Omega)$$

Now let us shift to discuss the idea of finite element sub-space of function space V , i.e to construct a finite dimensional sub-space V_h of V , where h is known as discretization parameter of domain Ω . Assume that $\partial\Omega$ is a polygonal curve, for simplicity, in which case we say Ω is a polygonal domain. Now we make a partition of Ω using triangles by subdividing Ω into collection of triangles $T_h = \{K_1, \dots, K_n\}$ here :

$$\Omega = \cup_{K \in T_h} K$$

such that no vertex of triangles lies in the edge of another triangle. For example, Figure 1.1 shows a triangulation of a skewed square that is not permissible.

Let us define the mesh parameter:

$$h = \max_K \rho(K),$$

where $\rho(K)$ is the longest edge of triangle K .

The triangulations T_h are called *quasi-uniform* (of size h) if the area of any triangle $K \in T_h$ is bounded by ch^2 , where $c > 0$ is independent of h and K .

Define the finite dimensional subspace V_h , called the finite element space, as follows:

$$V_h = \{v; v \in C(\Omega), v|_K \in P_1\}$$

and its subspace:

$$V_{h,0} = \{v; v \in C(\Omega), v|_K \in P_1, v|_{\partial\Omega} = 0\}$$

Where P_1 denote affine function in two variables. Instead of using affine function, we may use higher order polynomials over triangle for basis function, for example piece-wise quadratic function. The space $V_{h,0}$ consist of all continuous functions, that are linear on every triangle K and vanish on $\partial\Omega$. Note that $V_{h,0} \subset V$. We choose values $v(N_i)$ of v at nodes $N_i, i = 1, \dots, m$ of T_h in interior of Ω as parameter to describe function $v \in V_h$, note that we exclude nodes at boundary since $v = 0$ on $\partial\Omega$. Then the corresponding basis function $\phi_j \in V_{h,0}, j = 1, \dots, m$ are defined by:

$$\phi_j(N_i) = \delta_{ij}$$

where δ_{ij} is Kronecker delta.

Observe that the support of ϕ_j , i.e $\text{supp}(\phi_j) = \{x \in \Omega : \phi_j(x) \neq 0\}$ consists of all triangles that shares node N_j as their common vertex. Now, the element $v_h \in V_h$ can be written in the form:

$$v_h(x) = \sum_{j=1}^m \alpha_j \phi_j(x), \alpha_j = v_h(N_j), \forall x \in \Omega$$

Next we will discuss some of the approximation properties of this finite element space. In order to do so, first let us introduce the notion of the piecewise interpolant of function.

Definition

For $u \in C(\Omega)$, the interpolant $u_h \in V_h$ of u is defined to be unique continuous piecewise linear function with respect to the triangulation which is equal to values of u at the nodes of triangulation T_h . #

So the previous definition stated that the interpolant can be written as:

$$u_h(x) = \sum_{j=1}^m u(N_j) \phi_j(x)$$

Denote by $\| - \|_{k,K}$ is k -th order Sobolev norm over triangle K of T_h .

Theorem 1:[STW90]

Let K be any triangle in T_h If $u \in H^2(K)$ then:

$$\|u - u_h\|_{0,K} + h\|\nabla(u - u_h)\|_{0,K} \leq ch^2|u|_{2,K}$$

where c is independent of h, K #

As a consequence, we have:

Corollary 2:[STW90]

If $u \in H^2(\Omega)$ (respectively $H_0^1 \cap H^2$) then: $u_h \in V_h$ (respectively $V_{h,0}$) satisfies:

$$\|u - u_h\|_{0,\Omega} + h\|\nabla(u - u_h)\|_{0,\Omega} \leq ch^2|u|_{2,\Omega}$$

where c is independent of h, K #

Now, we turn to one of the fundamental properties to the subspace V_h , usually referred to as inverse property.

Theorem 3:[STW90]

Let T_h be a quasi-uniform family of triangulation and V_h as above. There exist c such that for any $\chi_h \in V_h$:

$$\|\chi_h\|_{1,\Omega} \leq ch^{-1}\|\chi_h\|_{0,\Omega}$$

where c is independent of h, χ_h and Ω #

1.2 Basic Error Estimates

In this section we will discuss error estimates for orthogonal projections in the $L^2(\Omega)$ and $H_0^1(\Omega)$.

Let V be a Hilbert Space with inner products $\langle -, - \rangle_V$ and its induced norm $\| - \|_V$. Let V_m be a finite dimensional subspace of V , with basis $\{\phi_1, \dots, \phi_m\}$.

Lemma 4:

For every $u \in V$, there exist $u_h \in V_m$ satisfying:

$$\langle u - u_h, v \rangle = 0, \forall v \in V$$

u_h is the orthogonal projection of u with respect to $\langle -, - \rangle_V$ into V_m . Equivalently, u_h is the best approximation in V_m to u in the norms of V :

$$\|u - u_h\|_V = \inf_{\chi \in V_m} \|u - \chi\|_V. \#$$

The simplest projection method is L^2 -projection. So let $V = L^2$ and $u \in V$. We look for the best approximant of u in some subspace V_m , in this case finite element subspace of V , i.e $V_m = V_h$.

Let us set:

$$u_h = P_h u$$

where P_h is L^2 -projection of u into V_h . So $P_h u$ is defined by:

$$\langle u - P_h u, v \rangle_0 = 0, \forall v \in V_h(\Omega)$$

or equivalently :

$$\|u - u_h\|_0 = \inf_{\chi_h \in V_h} \|u - \chi_h\|_0$$

Another orthogonal projection that has importance for Dirichlet problem for Poisson's equation is the orthogonal projection in H_0^1 . In this case we take $V = H_0^1(\Omega)$ with inner product:

$$\langle u, v \rangle_1 = \int \nabla u \nabla v$$

For V_m we choose $V_m = V_{h,0}$ and for u_h , the best approximant in $V_{h,0}$ to u , define *Ritz projection* :

$$u_h = R_h u$$

where R_h satisfies:

$$\langle u - R_h u, \phi \rangle_1 = \langle \nabla(u - R_h u), \nabla \phi \rangle_1 = 0, \forall \phi \in V_{h,0}(\Omega)$$

or equivalently :

$$\begin{aligned} \|u - u_h\|_1 &= \|\nabla(u - u_h)\|_0 \\ &= \inf_{\chi_h \in V_{h,0}} \|\nabla(u - \chi_h)\|_0 \\ &= \inf_{\chi_h \in V_{h,0}} \|u - \chi_h\|_1 \end{aligned}$$

The following basic error estimates are an immediate consequence of the best approximation properties of these projections and the properties of the interpolant.

Theorem 5:[STW90]

i) Let $u \in H^2(\Omega)$, then :

$$\|u - P_h u\|_0 \leq ch^2 |u|_2$$

ii) Let $u \in H_0^1(\Omega) \cap H^2(\Omega)$, then :

$$\|u - R_h u\|_1 = \|\nabla(u - R_h u)\|_0 \leq ch|u|_2$$

#

The previous theorem gives estimates for the error in the norms of the spaces in which the projection were taken. We may ask what the error of these projection are in other norms. For example, what is the error of the L^2 -projection in the space H^1 . A crucial assumption in determining an optimal order of accuracy of this H^1 error is that the mesh be quasi-uniform. In this case the inverse property holds on the subspace. From the basic estimates in the previous theorem for $(u - P_h u)$ we obtain the following estimates.

Theorem 6:[STW90]

Let $u \in H^2(\Omega)$, then

$$\|\nabla(u - P_h u)\|_0 = \|u - P_h u\|_1 \leq ch|u|_2$$

#

The L_2 error estimates for Ritz projection will be given in the last section in relation to the error estimates for Poisson's equation.

1.3 Finite Element Approximation: Model Problem

1.3.1 Variational Problem

In this subsection we give a variational formulation of problems in a complete function space. This formulation is useful in explaining how the finite element method works.

Let $V = V_1 = V_2$, and $a(-, -)$ be a bilinear form on $V \times V$. The bilinear form $a(-, -)$ is said to be *symmetric* if:

$$a(u, v) = a(v, u), \forall u, v \in V$$

Theorem 7: [STW90]

Let V be a complete vector space equipped with norm $\| \cdot \|_V$ and let $f : V \mapsto R$ be a continuous linear form. Let $a : V \times V \mapsto R$ be a symmetric, continuous bilinear form, i.e there exist $c_1 > 0$ such that:

$$|a(u, v)| \leq c_1 \|u\|_V \|v\|_V$$

Moreover, suppose that there exist $c_2 > 0$ such that:

$$a(v, v) \geq c_2 \|v\|_V^2$$

(*V-ellipticity condition*).

We will consider the following problem (P_V):

Find $u \in V$ such that:

$$a(u, v) = f(v), \forall v \in V$$

Then the problem (P_V) has a unique solution. #

The following theorem is an equivalent form of the previous theorem.

Theorem 8: [STW90]

Let the assumption of the previous theorems be fulfilled. Then the problem (P_V) is equivalent to the problem:

Find $u \in V$ such that:

$$J(u) = \inf_{v \in V} J(v)$$

where:

$$J(u) = \frac{1}{2} a(u, u) - f(u) \#$$

The function $J(v) : V \mapsto \mathbb{R}$ in the previous theorem is called *the energy functional* and the norm defined by $|a(-, -)|^{\frac{1}{2}}$, which is equivalent to $\| - \|_V$ by continuity and V -ellipticity, is called *the energy norm*.

1.3.2 Model Problem

In this section we apply the finite element method to solve a model problem, the Poisson's equation with Dirichlet boundary conditions:

$$-\Delta u = f, \text{ in } \Omega$$

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

where: Ω is an open bounded set in $\mathbb{R}^2$ with boundary $\partial\Omega$, and f is a given function, while Δ denotes the Laplacian operator. Many problems in Physics and mechanics are modeled by Poisson's equation, for examples: heat potential, fluids potential, or the displacement of an elastic membrane

under transversal load f with fixed boundary. Now we shall give a variational formulation of our model problem. First we shall show that if u is a solution of the model problem, then u is the solution of the following problem (P):

Find $u \in V = H_0^1(\Omega)$ such that :

$$a(u, v) = \langle f, v \rangle, \forall v \in V$$

where:

$$a(u, v) = \int_{\Omega} \nabla u \nabla v$$

$$\langle f, v \rangle = \int_{\Omega} f v$$

Let $u \in C^2(\Omega) \cap H_0^1(\Omega) \subset H^1(\Omega)$ be a solution of Poisson's equation with zero Dirichlet conditions. Let $v \in H_0^1$ arbitrary. Then we have:

$$-\int_{\Omega} \Delta u \cdot v = \int_{\Omega} f \cdot v$$

Applying Green's formula, we find that:

$$\int_{\Omega} \nabla u \cdot \nabla v - \int_{\partial\Omega} (\nabla u \cdot n) v ds = \int_{\Omega} f \cdot v$$

Since $v \in H_0^1$, then:

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f \cdot v$$

So if u is a solution then it is also solution of problem (P). Note that the variational formulation only demands first derivative of u which is advantageous for approximation.

To apply the theorem in the previous subsection, we have to verify whether all assumptions of the theorem are fulfilled.

The linearity of $f(-)$ and bilinearity of $a(-, -)$ on H_0^1 are obvious. The continuity of $f(-)$ follows from Cauchy-Schwarz inequality:

$$|f(v)| = \left| \int_{\Omega} f v \right|$$

$$\leq \|f\|_0 \|v\|_0 \leq \|f\|_0 \|v\|_1$$

The bilinear form $a(-, -)$ is symmetric and we show that it is continuous:

$$|a(u, v)| = \left| \int_{\Omega} \nabla u \nabla v \right|$$

$$\leq \|\nabla u\|_0 \|\nabla v\|_0 \leq \|u\|_1 \|v\|_1$$

Applying Poincare's inequality, then we obtain V -ellipticity condition:

$$a(v, v) = \int_{\Omega} \nabla v \nabla v \geq \|v\|_1^2$$

So by virtue of the theorem in previous subsection, problem (P) for Poisson's equation has unique solution.

Starting from the variational formulation, we can formulate the finite element method for our model problem, denoted (P_h) :

Find $u_h \in V_h$ such that:

$$a(u_h, v_h) = \langle f, v_h \rangle, \forall v_h \in V_h$$

Note that this is equivalent with a linear system of equations:

$$A\xi = b$$

where A is $m \times m$ matrix with entry $a_{ij} = a(\phi_i, \phi_j)$ and $\xi = (\alpha_i)$, $b = (b_i)$ with entries $\alpha_i = u_h(N_i)$ and $b_i = \langle f, \phi_i \rangle$ respectively. The matrix A , called stiffness matrix, is symmetric and positive definite, thus non-singular, so that the discrete problem (P_h) , admits a unique solution.

1.4 Some Results on H^1 and L^2 Estimates for Poisson's Equation

In this section we shall gathered some results on some estimates for approximate solution of Poisson's equation. In order to obtain this estimates we will need some information regarding the smoothness of u . In general, this depends on the smoothness of the data and properties of boundary $\partial\Omega$. From the previous result, problem (P) for Poisson's equation has unique solution in H_0^1 . Since $f \in L^2$, it makes sense if we asks whether $u \in H^2$?. This regularity question is well known for some classes of domain Ω , such as given in the following theorem.

Theorem 9:[STW90]

Let $u \in H_0^1(\Omega)$ be a variational solution of problem (P).

If Ω is a convex domain and $f \in L^2(\Omega)$, then $u \in H^2(\Omega)$,

$$\|u\|_2 \leq c\|f\|_0$$

H^1 estimates

Let $u_h \in V_{h,0}$ be the solution of discrete problem (P_h) , then in view of problem (P), it follows that:

$$a(u - u_h, v_h) = 0, \forall v_h \in V_{h,0}$$

hence it is clear :

$$u_h = R_h u$$

is the best approximant in $H_0^1(\Omega)$ -norms to u . Then from Corollary 1.2 and the definition of P_h and R_h projection respectively we obtain the following basic estimates.

Theorem 10:[STW90]

i) Let $u \in H^2(\Omega)$, then :

$$\|u - P_h u\|_0 \leq ch^2|u|_2$$

ii) Let $u \in H_0^1(\Omega) \cap H^2(\Omega)$, then :

$$\|u - R_h u\|_1 = \|\nabla(u - R_h u)\|_0 \leq ch|u|_2$$

The following Theorem summarizing the discrete solvability and approximant property of Ritz projection in H_0^1 -space.

Theorem 11:[STW90]

Let $T_h = \{K_1, \dots, K_N\}$ be a quasi-uniform triangulation of Ω , a convex polygonal domain.

i). For each $f \in L^2(\Omega)$, the discrete problem (P_h) have unique solution $u_h \in V_{h,0}$.

ii) If u is the solution of the variational problem (P) and $u_h \in V_{h,0}$ is the best solution of discrete problem (P_h) , then :

$$a(u - u_h, v_h) = 0, \forall v_h \in V_{h,0}$$

Thus $u_h = R_h u$ is the best approximant in $V_{h,0}$ to u in H_0^1 -norm. Hence:

$$\|\nabla(u - u_h)\|_0 = \inf_{\chi_h \in V_h} \|\nabla(u - \chi_h)\|_0$$

L^2 estimates

As a consequence of Theorem 9 and Theorem 11 previously, we obtain the following L^2 -estimates for Ritz projection.

Theorem 12:[STW90]

Let the assumption in the previous theorem be fulfilled.

If Ω is convex and $f \in L^2$, then:

$$\|(u - R_h u)\|_0 \leq ch^2 |u|_2 \leq ch^2 \|f\|_0$$

Chapter 2

The Biharmonic Equation: Source of the Problem

This chapter presents three sources of problem in which the resulting equation is a fourth order equation, specifically the biharmonic equation. The first problem, arise from a physical problems of modelling deflection of a square plates due to deformation. The second problem arise from reconstruction problem of two-dimensional surface of an object in computer vision process. The novelty of this model is adapting plate deformation problem to model the representation of the surfaces we want to reconstruct. To close this chapter, we examine a problem arise from fluid mechanics that describes a slow flow of a fluid motion.

2.1 Plate Equation

Consider a square thin plate which undergoes a deformation, its potential energy under deformation is given by an integral of a quadratic form in the *principle curvatures* of the plate[LL86]. Let κ_1 and κ_2 denotes the principle curvatures of the deformed plate, the potential energy density is given by:

$$\begin{aligned}\frac{A}{2}(\kappa_1^2 + \kappa_2^2) + B\kappa_1\kappa_2 &= 2A\left(\frac{\kappa_1 + \kappa_2}{2}\right)^2 - (A - B)\kappa_1\kappa_2 \\ &= 2AH^2 - (A - B)G\end{aligned}$$

where A and B are constants determined by the plate material, and H and G are the mean and Gaussian curvature respectively.

Suppose $u(x, y)$ denotes the deflection of the plate normal to the region Ω on the plane. Assuming that the deflection and its partial derivatives only undergoes small deformation, then [Re69] :

$$\frac{\kappa_1 + \kappa_2}{2} \approx \Delta u$$

and,

$$\kappa_1 \kappa_2 \approx \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2$$

So, the potential energy density can be written in the form :

$$\begin{aligned} e(u) &= 1/2(\Delta u)^2 - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 \right) \\ &= \frac{1}{2} \left(\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right) + \sigma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 \right) \\ &= \frac{1}{2} \{ \sigma (\Delta u)^2 + (1 - \sigma) \left(\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 \right) \} \end{aligned}$$

The constant σ denotes *Poisson ratio*.

By integrating the energy density we obtain the desired potential energy of deformation which is given by:

$$J_1(u) = \int_{\Omega} \frac{1}{2} (\Delta u)^2 - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 \right)$$

Due to any external forces $g(x, y)$ applied to the surfaces of the plate (say gravity), we have to add the potential energy of external forces to the potential energy of deformation, which is given by:

$$J_2(u) = - \int_{\Omega} g u$$

Around the boundary $\partial\Omega$ of Ω , we may apply external forces and bending moments. The energy due to the force density $p(s)$ on the boundary (s denotes arc length of boundary) is:

$$J_3(u) = - \int_{\partial\Omega} p(s) u \, ds$$

while the energy to bending moments is given by:

$$J_4(u) = - \int_{\partial\Omega} m(s) \frac{\partial u}{\partial n} ds$$

where $m(s)$ is the density of applied bending moment to the boundary curve and $\frac{\partial}{\partial n}$ denotes normal derivatives at boundary with outward orientation.

The equilibrium state of the plates can be obtained as the solution of the following minimization problem:

The deflection of the plate at equilibrium is that the function u from a suitable function Space V , for which the total potential energy :

$$J(v) = J_1(v) + J_2(v) + J_3(v) + J_4(v)$$

is minimal

The admissible function space we require in the above assertion is in general a subspace of $H^2(\Omega)$. If $u \in V$ is the minimizer of $J(-)$, then $J(u) \leq J(u + \epsilon v)$ for any $v \in V$ and $\epsilon > 0$. So the necessary condition for the minimizer, after calculating the variation, is :

$$\begin{aligned} \int_{\Omega} \Delta u \Delta v - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} - 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right) \left(\frac{\partial^2 v}{\partial x \partial y} \right) \right) \\ - \int_{\Omega} g v - \int_{\partial\Omega} p(s) v ds - \int_{\partial\Omega} m(s) \frac{\partial v}{\partial n} ds = 0 \end{aligned}$$

Defining the bilinear form :

$$a(u, v) = \int_{\Omega} \Delta u \Delta v - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} - 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right) \left(\frac{\partial^2 v}{\partial x \partial y} \right) \right)$$

and linear form :

$$f(v) = - \int_{\Omega} g v - \int_{\partial\Omega} p(s) v ds - \int_{\partial\Omega} m(s) \frac{\partial v}{\partial n} ds$$

we obtain that the total potential energy for the thin plate deformation can be written in the form :

$$J(u) = a(u, u) - f(u)$$

Now, we will derive the Euler-Lagrange Equation for the thin plate deformation. Assume that $u \in H^4(\Omega)$, applying Green's identity:

$$\int_{\Omega} v \Delta u - \int_{\Omega} u \Delta v = \int_{\partial\Omega} u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} ds$$

to the term $\Delta u \Delta v$ in total potential energy, we obtain:

$$\int_{\Omega} \Delta u \Delta v = \int_{\Omega} v \Delta^2 u + \int_{\partial\Omega} \Delta u \frac{\partial v}{\partial n} - v \frac{\partial(\Delta u)}{\partial n} ds$$

where $\Delta^2 u = \Delta \Delta u$. On the other hand the Gaussian curvature term can be transformed using identity [Ci78]:

$$\begin{aligned} \int_{\Omega} \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} - 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right) \left(\frac{\partial^2 v}{\partial x \partial y} \right) \right) = \\ \int_{\partial\Omega} \left(\frac{\partial^2 u}{\partial t^2} \frac{\partial v}{\partial n} - \left(\frac{\partial^2 u}{\partial n \partial t} \right) \left(\frac{\partial v}{\partial t} \right) \right) ds \end{aligned}$$

where $\frac{\partial}{\partial t}$ denotes tangential derivative at boundary. If the boundary is sufficiently smooth, it can be shown [Re80] that the second term of boundary integral for Gaussian curvature can be written as:

$$\int_{\partial\Omega} \left(\frac{\partial^2 u}{\partial n \partial t} \right) \left(\frac{\partial v}{\partial t} \right) ds = \int_{\partial\Omega} \frac{d}{ds} \left(\frac{\partial^2 u}{\partial n \partial t} \right) (v) ds$$

Substituting this result to the bilinear forms, we get :

$$a(u, v) = \int_{\Omega} \Delta^2 u v + \int_{\partial\Omega} P(u) v ds + \int_{\partial\Omega} M(u) \frac{\partial v}{\partial n} ds$$

where:

$$\begin{aligned} P(u) &= -\frac{\partial(\Delta u)}{\partial n} + (1 - \sigma) \frac{d}{ds} \left(\frac{\partial^2 u}{\partial n \partial t} \right) \\ M(u) &= \Delta^2 u - (1 - \sigma) \left(\frac{\partial^2 u}{\partial t^2} \right) \end{aligned}$$

Thus the necessary condition can be written as:

$$\int_{\Omega} (\Delta^2 u - g) v + \int_{\partial\Omega} (P(u) - p(s)) v ds + \int_{\partial\Omega} (M(u) - m(s)) \frac{\partial v}{\partial n} ds = 0$$

Since v and $\frac{\partial v}{\partial n}$ are arbitrary on Ω and the equation must hold, then we must have:

$$\Delta^2 u = g$$

So this is the equation that governs the small deflection of a thin plate at equilibrium, and it is satisfied by u inside Ω regardless of boundary conditions on $\partial\Omega$. Moreover, u must satisfy the natural boundary conditions :

$$(P(u) = p(s)) \text{ and } (M(u) = m(s)) \text{ on } \partial\Omega.$$

The necessary condition for the minimum implies another consequence, that is the natural boundary conditions satisfied by u are determined by an essential boundary conditions which may be imposed on v . Generally, we can impose up to two essential boundary conditions, one on v and the other on $\frac{\partial v}{\partial n}$. First, suppose that we set the essential boundary condition $v = 0$ and $\frac{\partial v}{\partial n} = 0$ on $\partial\Omega$. Then :

$$u = 0,$$

$$\frac{\partial u}{\partial n} = 0, \text{ on } \partial\Omega$$

which is the case of clamped plate. This condition is well known as First Biharmonic Problem.

Next consider the case of simply supported plate where the essential boundary condition $v = 0$ is imposed on $\partial\Omega$ but $\frac{\partial v}{\partial n}$ is left unconstrained. Thus we have the Neumann boundary conditions:

$$u = 0, M(u) = m(s), \text{ on } \partial\Omega$$

Note that for the case $\sigma = 1$, the boundary conditions became $\Delta u = 0$ on $\partial\Omega$. This type conditions is known as Second Biharmonic Problem.

2.2 Visual Surface Reconstruction

One of the problems in computer vision is the reconstruction of a visible surface of 3-dimensional object from sampled data measurement. The task is to reconstruct a visible surface of 3-dimensional objects from finite point depth measurement. One of the approaches to retrieving the surface is to use a thin plate analogy, as proposed by Terzopoulos [Te82][Te86].

The approach taken is as follows. Assuming that the measurement taken as a parallel projection onto the image plane for simplicity. Let $z(x_1, y_1), \dots, z(x_n, y_n)$ be a depth values of measurement of the surface over $(x_1, y_1), \dots, (x_n, y_n)$ a set of image coordinates. Imagine that the depth measurement be represented by a set of vertical pins. Suppose that we take a thin plate that is planar in the absence of external forces, and constrain it to pass near the tip of the pins by attaching ideal springs between the tip of pins and the surface of the plate.

Intuitively, the equilibrium state of the plate is a 'reasonable' approximating surface in terms of the smoothness of surface representation between constraints. Moreover, this model give way a good strategies for solving the problem.

Now, we may proceed to the mathematical characterization of the visual reconstruction. Obviously, we have to add potential energy due to attached springs to the total potential energy of the plates deformation discussed in the previous section. Let C denote the set of coordinates in Ω at which the imaginary pins are located. Suppose the stiffness of imaginary springs are $\beta(x_i, y_i)$ for $(x_i, y_i) \in C$. According to Hooke's law for the ideal spring, the total potential energy of deformation in the spring is given by:

$$J_5(u) = \sum_{(x_i, y_i)} \beta(x_i, y_i) [u(x_i, y_i) - z(x_i, y_i)]^2$$

So the total potential energy for the problem now is of the form :

$$J(v) = J_1(v) + J_2(v) + J_3(v) + J_4(v) + J_5(v)$$

Assuming that the attached imaginary springs have the same material properties, i.e same stiffness $\beta(x_i, y_i) = \beta$, we have the necessary condition for the minimizer is of the form :

$$\begin{aligned} \int \Delta u \Delta v - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} - 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right) \left(\frac{\partial^2 v}{\partial x \partial y} \right) \right) \\ - \int_{\Omega} g v - \int_{\partial \Omega} p(s) v \, ds \\ - \int_{\partial \Omega} m(s) \frac{\partial v}{\partial n} \, ds + \beta \sum_{(x_i, y_i)} [u(x_i, y_i) - z(x_i, y_i)] v(x_i, y_i) = 0 \end{aligned}$$

Defining the bilinear form:

$$a(u, v) = \int \Delta u \Delta v - (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} - 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right) \left(\frac{\partial^2 v}{\partial x \partial y} \right) \right) \\ + \beta \sum_{(x_i, y_i)} u(x_i, y_i) v(x_i, y_i)$$

and the linear form:

$$f(v) = - \int_{\Omega} g v - \int_{\partial \Omega} p(s) v ds - \int_{\partial \Omega} m(s) \frac{\partial v}{\partial n} ds \\ + \beta \sum_{(x_i, y_i)} [u(x_i, y_i) v(x_i, y_i) - (z(x_i, y_i))^2]$$

Clearly that the variational problem of the vision reconstruction problem is to minimize $J(v) = a(v, v) - f(v)$ over an appropriate subspace V of H^2 .

Following the argument for thin plate deformation in the previous section, if $u \in H^4(\Omega)$, then our problem is looking for solution of:

$$\Delta^2 u = g, \text{ on } \Omega$$

under constraint:

$$u(x_i, y_i) = z(x_i, y_i), (x_i, y_i) \in C$$

Boundary values to be imposed on the edges for this problem can be chosen from analogies of clamped plate, simply supported plate, or free plate. For the later case, our admissible function space is the whole space H^2 . Terzopoulos [Te82] gave necessary conditions such that the bilinear form $a(-, -)$ be H^2 -elliptic; it is necessary that the set of constraint point C contain at least three non-collinear points.

2.3 Low Reynolds Number Flows

Consider a viscous fluid occupying a bounded 2-dimensional domain Ω . It is known that by introducing a stream function Ψ such that:

$$\frac{\partial \Psi}{\partial y} = v_1, \frac{\partial \Psi}{\partial x} = v_2$$

where $\mathbf{v} = (v_1, v_2)$ denotes velocity fields, the Navier-Stokes equation can be written as [LL89]:

$$\frac{\partial}{\partial t} \Delta \Psi - \frac{\partial \Psi}{\partial x} \frac{\partial \Delta \Psi}{\partial y} + \frac{\partial \Psi}{\partial y} \frac{\partial \Delta \Psi}{\partial x} - \nu \Delta^2 \Psi = 0$$

where $\frac{1}{\nu}$ denotes Reynolds number. Suppose that the fluids is in the steady state condition, then the equation becomes:

$$\Delta^2 \Psi = \frac{1}{\nu} \frac{\partial \Psi}{\partial x} \frac{\partial \Delta \Psi}{\partial y} + \frac{\partial \Psi}{\partial y} \frac{\partial \Delta \Psi}{\partial x}$$

In many applications, we encounter a problem which is known as slow flow viscous fluid problem, in which case the Reynolds number is very small. Then the right hand side of the previous equation can be dropped, so the equation becomes:

$$\Delta^2 \Psi = 0$$

i.e a biharmonic equations. Numerous boundary conditions can be imposed. Consider a square open cavity with vertices at OABC, where BC denotes its opening. The boundary condition of this situation is:

$$\Psi = 0, \frac{\partial \Psi}{\partial n} = 0, \text{ on } OA$$

$$\Psi = 0, \frac{\partial \Psi}{\partial n} = 0, \text{ on } AB$$

$$\Psi = 0, \frac{\partial \Delta \Psi}{\partial n} = 0, \text{ on } BC$$

$$\Psi = 0, \frac{\partial \Psi}{\partial n} = -1, \text{ on } OC$$

This problem is known as the open cavity flow problem.

Chapter 3

Mixed Finite Element Formulation for the Biharmonic Equation

In this chapter we will discuss the mixed finite element method for Biharmonic problem. The results of which depend heavily on the results of Brezzi and Raviart [BR77]. It should be mentioned that the approach taken by them could accommodate the First and Second Biharmonic problem nicely. They developed their abstract methods initially was to unified some mixed method for plate equation into one abstract formulation. The presentation of this chapter is arranged as follows. In the first section we will take a look at two variational formulations of the first Biharmonic problem. The first formulation is a direct variational formulation, and second is mixed variational formulation due to Ciarlet and Raviart [CR74], it can be shown that their formulation is a special case of Brezzi and Raviart abstract formulation. In the second section, we present an abstract result due to Brezzi and Raviart [BR77]. This result is applied to the first and second Biharmonic problems in section 3 and 4 respectively.

3.1 Variational Problem

Direct Variational Method Formulation

Let $f \in L^2(\Omega)$ be given, consider the problem (P_b) :

Find $u \in V = H_0^2(\Omega)$ which satisfies:

$$a(u, v) = f(v), \forall v \in V$$

where the bilinear form on V defined as :

$$a(u, v) = \int_{\Omega} \left\{ \sigma \Delta u \Delta v + (1 - \sigma) \left(\frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \frac{\partial^2 v}{\partial y^2} + 2 \left(\frac{\partial^2 u}{\partial y \partial x} \frac{\partial^2 v}{\partial x \partial y} \right) \right) \right\}, \forall v \in V$$

and linear form f on V defined as :

$$f(v) = \int_{\Omega} f v, \forall v \in V$$

From discussion on thin plate deformation, this problem is related to minimization of quadratic functional of energy:

$$J(v) = a(v, v) - f(v)$$

Since:

$$a(v, v) = \sigma |\Delta v|_0 + (1 - \sigma) |v|_0, \forall v \in H_0^2$$

then obviously $a(-, -)$ is V -elliptic and continuous over V , hence problem (P_b) is uniquely solvable for every $f \in L^2$, by virtues of section 1.3.1 .

If the solution $u \in H^4 \cap H_0^2$, i.e smooth enough, then this is a solution of the plate equation:

$$\begin{aligned} \Delta^2 u &= f \\ u|_{\partial\Omega} &= 0, \frac{\partial u}{\partial n}|_{\partial\Omega} = 0 \end{aligned}$$

Now, we have to choose the finite element space $V_{h,0} \subset V = H_0^2$ within which we have to solve the discrete problem:

Find $u_h \in V_{h,0}$ which satisfies:

$$a(u_h, v_h) = f(v_h), \forall v_h \in V_{h,0}$$

The inclusion $V_{h,0} \subset H_0^2$ in turn implies that $V_{h,0} \subset C^1(\overline{\Omega})$ which in itself implies requirement of the degree of freedom to generate basis function. For example, we may choose as a basis over triangles is Argyris basis which is a complete polynomial of order 5.

Another Approach

Another approach to solve the first biharmonic equation was proposed by Ciarlet & Raviart [CR74] in the following way. Introduce a new variable: $\psi = \Delta v$. Then the minimization of energy functional of plate equation becomes:

$$\min J(v) = \min \left\{ \int_{\Omega} |\psi|^2 - f.v \right\}$$

subject to : $\psi = \Delta v$.

In order to solve this formulation, let us introduce the following space: $X = H_0^1 \times L^2$. Define on X the bilinear form: $\tilde{a} : X \times X \mapsto R$ by

$$\tilde{a}(x, y) = \int_{\Omega} \phi \psi, \forall (x, y) = ((u, \phi), (v, \psi)) \in X \times X$$

This bilinear form is obviously symmetric and continuous by introducing on X the product norms:

$$\|x\|_X = |u|_1 + |\phi|_0, x = (u, \phi)$$

likewise, the linear form $f(-)$ defined on X by:

$$f(y) = \int_{\Omega} f \psi, y = (v, \psi)$$

is continuous.

Next, let $M = H^1$, defined on $X \times M$ a bilinear form by:

$$b(y, \mu) = \int_{\Omega} \nabla v \nabla \mu - \int_{\Omega} \psi \mu, \forall (y, \mu) = ((v, \psi), \mu) \in X \times M$$

which is clearly continuous. Finally, let $\tilde{X} = \{x = (v, \psi) \in X; \forall \mu \in M, b(x, \mu) = 0\}$. So that $\tilde{X}$ is a Hilbert space since it is a closed subspace of X . Moreover, if $y = (v, \psi) \in X$ satisfy: $b(y, \mu) = 0, \forall \mu \in M$, then we have in particular :

$$\forall \mu \in H_0^1, \int_{\Omega} \nabla v \nabla \mu dx = \int_{\Omega} \psi \mu dx,$$

hence v appears as a solution for Laplace's equation with homogeneous Dirichlet conditions on Ω . Assuming Ω is a convex set, implies $v \in H^2$ by regularity of solutions of second order problem. Using Green's formula:

$$\int_{\Omega} \nabla v \nabla \mu = - \int_{\Omega} \Delta v \mu + \int_{\partial \Omega} \partial_{\nu} v \mu,$$

We deduce that $-\Delta v = \psi$. Observe that since the product norms is equivalent to the norm defined by : $|y| : y \mapsto |\psi|_0, y = (v, \psi) \in \tilde{X}$, by Theorem 7.1.1 of Ciarlet [Ci78], so we can show that the bilinear form $\tilde{a}$ is V-elliptic in the space $\tilde{X}$.

Theorem : [CR74]

Let $u \in H_0^2(\Omega)$ denote the minimizer of energy functional of plates. Then we also have :

$$J(u, -\Delta u) = \inf_{(v, \psi) \in \tilde{X}} J(u, \psi)$$

In addition, the pair $(u, -\Delta u) \in \tilde{X}$ is the unique solution of the above minimization problem. #

3.2 Abstract Result

Let V and W be two (real) Hilbert Spaces with norms $\| \cdot \|_V$ and $\| \cdot \|_W$ respectively. Let V' and W' be the dual spaces of V and W respectively. Denote by $[\cdot, \cdot]$ the pairing between the spaces V' and V or W' and W .

Let $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ be two continuous bilinear forms on $V \times V$ and $V \times W$. Set :

$$\|a\| = \sup_{v \in V} \frac{|a(u, v)|}{\|u\|_V \|v\|_V} \dots (A.1)$$

Assume that the bilinear form $a(\cdot, \cdot)$ is V-elliptic, in a sense that there exist a constant $\alpha > 0$ such that:

$$\forall v \in V, a(v, v) \geq \alpha \|v\|_V^2 \dots (A.2)$$

In addition assume that there exist a constant $\beta > 0$ such that:

$$\forall w \in W, \sup_{v \in V} \frac{b(v, w)}{\|v\|_V} \geq \beta \|w\|_W \dots (A.3)$$

Consider the following problem (P) :

Given $f \in V'$ and $g \in W'$, find a pair $(u, p) \in V \times W$ which satisfies:

$$\forall v \in V, a(u, v) + b(v, p) = [f, v] \quad (3.1)$$

$$\forall w \in W, b(u, w) = [g, w] \quad (3.2)$$

We recall the following results due to Brezzi [Br74]:

Theorem 1 :

Assume that the hypothesis (A.2) and (A.3) hold. Then Problem (P) has unique solution $(u, p) \in V \times W$.

Proof[RT77]

Let $Z^\perp$ denote the subspace of V orthogonal to $Z = \{v \in V; \forall w \in W, b(v, w) = 0\}$, let v^0 and $v^\perp$ denote the orthogonal projections of v onto Z and $Z^\perp$ respectively. With this notation, the problem (P) may be written as the modified problem (P_m) :

Find $u^0, u^\perp$, and p such that :

$$\begin{aligned} u^\perp &\in Z^\perp, u^0 \in Z^0, p \in W \\ \forall v^\perp \in Z^\perp, a(u^\perp, v^\perp) + a(u^0, v^\perp) + b(v^\perp, p) &= [f, v^\perp] \\ \forall v^0 \in Z^0, a(u^\perp, v^0) + a(u^0, v^0) &= [f, v^0] \\ \forall w \in W, b(u^\perp, w) &= [g, w] \end{aligned}$$

Note that we have:

$$\sup_{\{v \in Z^\perp; \|v\|=1\}} b(v, w) = \sup_{\{v \in V; \|v\|=1\}} b(v, w), \forall w \in W$$

also that Z can be characterized as a subset of V :

$$v \in Z \iff \sup_{\{w \in W; \|w\|=1\}} b(v, w) = 0$$

Using hypothesis (A.3) we may establish with the help of V -ellipticity of the bilinear forms a that the linear operator B^* from W to $Z^\perp$ defined by:

$$\langle v, B^*w \rangle_V = b(v, w), \forall v \in V, w \in W$$

is an isomorphism of W into $Z^\perp$. Thus the linear operator B from $Z^\perp$ to W defined by:

$$\langle Bv, w \rangle_W = b(v, w), \forall v \in V, w \in W$$

is an isomorphism of $Z^\perp$ into W .

Finally using Hypothesis (A.1) and (A.2), with the help of Lax-Milgram theorem, we may show that the linear operator A from V to V defined by:

$$\langle Az, v \rangle_V = a(z, v), \forall v \in V, z \in Z$$

is an automorphism. Thus the linear system of modified problem (P_m) , to which we can associate it to a matrix of operator of the form:

$$\begin{pmatrix} \times & \times & B^* \\ \times & A & 0 \\ B & 0 & 0 \end{pmatrix}$$

has unique solution. #

Next we shall give another characterization of the problem (P) well suited for mixed finite element methods. We are given two other (real) Hilbert spaces $\tilde{V}$ and $\tilde{W}$, with norms $\| - \|_{\tilde{V}}$ and $\| - \|_{\tilde{W}}$ respectively, such that $\tilde{V} \subset V, W \subset \tilde{W}$ with continuous imbedding, and W is dense in $\tilde{W}$. Again let $[.,.]$ denote the pairing between $\tilde{W}$ and its dual $\tilde{W}'$. Assume that there exists a continuous bilinear form $\tilde{b}(.,.)$ on $\tilde{V} \times \tilde{W}$. With the following properties :

$$\forall v \in \tilde{V}, \forall w \in W, \tilde{b}(v, w) = b(v, w) \dots (A.4)$$

$$\forall w \in \tilde{W}, \sup_{v \in \tilde{V}} \frac{\tilde{b}(v, w)}{\|v\|_{\tilde{V}}} \geq \tilde{\beta} \|w\|_W \dots (A.5)$$

for some constant $\tilde{\beta} > 0$. Set:

$$\|\tilde{b}\| = \sup_{v \in \tilde{V}, w \in \tilde{W}} \frac{|\tilde{b}(v, w)|}{\|w\|_{\tilde{W}} \|v\|_{\tilde{V}}} \quad (3.3)$$

Assume that $g \in \tilde{W}'$ and consider the following problem $(\tilde{P})$:

Find $(u, p) \in \tilde{V} \times \tilde{W}$ such that :

$$\forall v \in \tilde{V}, a(u, v) + \tilde{b}(v, p) = [f, v] \quad (3.4)$$

$$\forall w \in \tilde{W}, \tilde{b}(u, w) = [g, w] \quad (3.5)$$

Observe that $a(.,.)$ is not a-priori $\tilde{V}$ -elliptic. However we have:

Theorem 2:[BR77]

Assume that conditions (A.4) and (A.5) hold and that the first argument u of the solution (u, p) of problem (P) belongs to $\tilde{V}$. Then (u, p) is the unique solution of problem $(\tilde{P})$. #

Proof

Let (u, p) be the solution of problem (P). If u belongs to $\tilde{V}$, from properties (A.4) and since W is dense in $\tilde{W}$ it follows that (u, p) is a solution to

problem $(\tilde{P})$. Hence it is sufficient to show only the uniqueness of solution of problem $(\tilde{P})$.

Assume that $(u, p) \in \tilde{V} \times \tilde{W}$ satisfies :

$$\forall v \in \tilde{V}, a(u, v) + \tilde{b}(v, p) = 0 \quad (3.6)$$

$$\forall w \in \tilde{W}, \tilde{b}(u, w) = 0 \quad (3.7)$$

Choosing $v = u$ in (1.6) and applying (1.7), we obtain $a(u, u) = 0$, hence $u = 0$ by the V-ellipticity of $a(-, -)$. Therefore we obtain:

$$\forall v \in \tilde{V}, \tilde{b}(v, p) = 0$$

which implies $p = 0$ by (A.5). #

In the sequel, assume that the conclusion of Theorem 2 holds. Let us now consider some general results of approximations of the solutions (u, p) of problem $(\tilde{P})$. Given V_h and W_h two finite dimensional spaces such that:

$$V_h \subset \tilde{V}, W_h \subset \tilde{W} \quad (3.8)$$

Consider the following approximate problem $(\tilde{P}_h)$:

Find $(u_h, p_h) \in V_h \times W_h$ such that :

$$\forall v_h \in V_h, a(u_h, v_h) + \tilde{b}(v_h, p_h) = [f, v_h] \quad (3.9)$$

$$\forall w_h \in W_h, \tilde{b}(u_h, w_h) = [g, w_h] \quad (3.10)$$

As corollary of previous theorem, we have:

Theorem 3:[BR77]

Assume that there exist a constant $\gamma > 0$ such that :

$$\forall w_h \in W_h, \sup_{v_h} \frac{b(v_h, w_h)}{\|v_h\|_{V_h}} \geq \gamma \|w_h\|_{\tilde{W}} \dots (A.6)$$

Then problem $(\tilde{P}_h)$ has unique solution $(u_h, p_h) \in V_h \times W_h$. #

Now we will derive abstract bounds for the errors $e_1 = (u - u_h)$ and $e_2 = (p - p_h)$. The approach is as follows. Since V_h is finite dimensional, there exist $S(h) > 0$, which depends on V_h such that:

$$\forall v_h \in V_h, \|v_h\|_{\tilde{V}} \leq S(h) \|v_h\|_V \quad (3.11)$$

Let us set :

$$Z = \{v \in \tilde{V}; \forall w \in W, \tilde{b}(v, w) = 0\} \quad (3.12)$$

and for any $\phi \in W'$,

$$Z_h(\phi) = \{v_h \in V_h; \forall w_h \in W_h, \tilde{b}(v_h, w_h) = [\phi, w_h]\} \quad (3.13)$$

It is clear that the subspace Z_h of V_h may be regarded as a finite dimensional approximation for Z , even though in general it is not contained in Z .

Theorem 4:[BR77]

Assume hypothesis (A.2) and (A.6) hold. Then there exist a constant $K = K(\alpha, \gamma, \|a\|, \|\tilde{b}\|)$ such that:

$$\|e_1\|_V + \|e_2\|_{\tilde{W}} \leq K \left\{ \inf_{v_h \in V_h} \|e_1\|_{\tilde{V}} + (1 + S(h)) \inf_{w_h \in W_h} \|e_2\|_{W_h} \right\} \quad (3.14)$$

Proof

We may see that (1.10) means u_h belongs to $Z_h(g)$. Moreover, from (1.9) and by definitions of $Z_h(0)$ it follows that :

$$\forall z_h \in Z_h(0), a(u_h, z_h) = [f, z_h]$$

Therefore, applying (1.4) and $Z_h(0)$, we get for all $v_h \in Z_h(\phi)$, and all $w_h \in W_h$,

$$\begin{aligned} a(v_h - u_h, v_h - u_h) &= a(v_h - u, v_h - u_h) + a(u - u_h, v_h - u_h) \\ &= a(v_h - u, v_h - u_h) + \tilde{b}(v_h - u_h, p - w_h) \end{aligned}$$

Hence we have by (A.1), (A.2), and (A.5),

$$\begin{aligned} \alpha \|v_h - u_h\|_V^2 &\leq \|a\| \|v_h - u\|_V \|v_h - u_h\|_V \\ &\quad + \|\tilde{b}\| \|v_h - u_h\|_{\tilde{V}} \|p - w_h\|_{\tilde{W}} \end{aligned}$$

So that by (1.11) :

$$\|v_h - u_h\|_V \leq \frac{1}{\alpha} \{ \|a\| \|v_h - u\|_V + \|\tilde{b}\| S(h) \|p - w_h\|_{\tilde{W}} \}$$

from which we infer the following inequality

$$\|u - u_h\|_V \leq \left(1 + \frac{\|a\|}{\alpha}\right) \inf_{v_h \in Z_h(\phi)} \|u - v_h\|_V + \frac{\|\tilde{b}\|}{\alpha} S(h) \inf_{w_h \in W_h} \|p - w_h\|_{\tilde{W}} \quad (3.15)$$

Next let us show that the hypothesis (A.6) implies:

$$\inf_{v_h \in Z_h(\phi)} \|u - v_h\|_V \leq C_1 \inf_{\mu_h \in W_h} \|u - \mu_h\|_{\tilde{V}} \quad (3.16)$$

for some constant $C_1 = C_1(\gamma, \|\tilde{b}\|) > 0$. Given $\mu_h \in W_h$ by (A.6), there exist at least one $y_h \in V_h$, such that:

$$\forall w_h \in W_h, \tilde{b}(y_h, w_h) = \tilde{b}(u - \mu_h, w_h)$$

and

$$\begin{aligned} \|y_h\|_{\tilde{V}} &\leq \frac{1}{\gamma} \sup_{w_h \in W_h} \frac{\tilde{b}(u - \mu_h, w_h)}{\|w_h\|_{\tilde{W}}} \\ &\leq \frac{\|\tilde{b}\|}{\gamma} \|u - \mu_h\|_{\tilde{V}} \end{aligned}$$

Setting $v_h = y_h + \mu_h$, we get for all $w_h \in W_h$

$$\tilde{b}(v_h, w_h) = \tilde{b}(u, w_h) = [\phi, w_h]$$

So that v_h belongs to $Z_h(\phi)$. Furthermore, we have:

$$\begin{aligned} \|u - v_h\|_{\tilde{V}} &\leq \|u - \mu_h\|_{\tilde{V}} + \|y_h\|_{\tilde{V}} \\ &\leq (1 + \frac{\|\tilde{b}\|}{\gamma}) \|u - \mu_h\|_{\tilde{V}} \end{aligned}$$

from which the inequality (1.16) follows with $c = 1 + \frac{\|\tilde{b}\|}{\gamma}$.

Finally, from (1.4) and (1.9), we obtain from all $v_h \in V_h$ and all $w_h \in W_h$,

$$\tilde{b}(v_h, p_h - w_h) = a(u - u_h, v_h) + \tilde{b}(v_h, p - w_h)$$

Hence using (-) and the inclusion $\tilde{V} \subset V$, we get for some constant $c_2 > 0$:

$$\|p_h - w_h\|_{\tilde{W}} \leq \frac{1}{\gamma} \{C_2 \|a\| \|u - u_h\|_V + \|b\| \|p - w_h\|_{\tilde{W}}\}$$

therefore we have:

$$\|p - p_h\|_{\tilde{W}} \leq \frac{C_2}{\gamma} \|e_1\|_V + (1 + \frac{\|\tilde{b}\|}{\gamma}) \inf_{w_h \in W_h} \|p - w_h\|_{\tilde{W}} \quad (3.17)$$

Combining the three inequalities (1.15), (1.16), (1.17) respectively, we obtain the desired estimates. #

Brezzi and Raviart noted that the error bound in the above theorem is not optimal for the following reasons. The estimates involves the error of approximation $\inf_{V_h} \|u - v_h\|_{\tilde{V}}$ in $\tilde{V}$ instead of the corresponding error $\inf_{V_h} \|u - v_h\|_V$ in V . At the same time the constant $S(h)$ is never bounded independently of the finite dimensional vector spaces V_h but tends to infinity as $\dim(V_h)$ increased. By adding a supplementary hypothesis, however, the error bound can be improved.

Theorem 5:

Assume the hypothesis of the previous theorem holds. In addition, assume that the inclusion $Z_h \subset Z$ holds. Then there exist a constant $K = K(\alpha, \gamma, \|a\|, \|\tilde{b}\|)$ such that:

$$\|e_1\|_V + \|e_2\|_{\tilde{W}} \leq K \left\{ \inf_{v_h \in Z_h(\phi)} \|e_1\|_V + \inf_{w_h \in W_h} \|e_2\|_{\tilde{W}} \right\}$$

where $Z_h(\phi)$ is defined as above. #

Proof

Since $Z_h(0) \subset Z$, it follows from (1.4) and (1.5) that we have for every $z_h \in Z_h$:

$$a(u - u_h, z_h) = 0$$

Therefore, for every $v_h \in Z_h(\phi)$ we get:

$$\begin{aligned} \alpha \|u - u_h\| &\leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \\ &\leq \|a\| \|u - u_h\|_V \|u - v_h\|_V \end{aligned}$$

so that:

$$\|e_1\| \leq \frac{\|a\|}{\alpha} \|u - v_h\|_V$$

Combining the results with (1.17), we get the desired estimates. #

Instead of the above theorem, we shall use the following consequences:

Corollary.

Assume (A.1), (A.5) and inclusion $Z_h \subset Z$ holds. Assume that there exist an operator $\pi_h : \tilde{V} \rightarrow V_h$ and a constant $c > 0$ such that:

$$\forall w_h \in W_h, \tilde{b}(v - \pi_h v, w_h) = 0 \quad (3.18)$$

$$\forall v \in \tilde{V}, \|\pi_h v\|_{\tilde{V}} \leq c\|v\|_{\tilde{V}} \quad (3.19)$$

Then problem $(\tilde{P}_h)$ has unique solution $(u_h, p_h) \in V_h \times W_h$ and we have :

$$\|e_1\|_V + \|e_2\|_{\tilde{W}} \leq K \{ \|u - \pi_h u\|_V + \inf_{w_h \in W_h} \|e_2\|_{\tilde{W}} \}$$

for some constant $K = K(\alpha, \tilde{\beta}, c, \|a\|, \|\tilde{b}\|)$. #

Proof

Using (1.18), (1.19), and (A.5), for every $w_h \in W_h$ we have:

$$\begin{aligned} \sup_{v_h \in V_h} \frac{\tilde{b}(v_h, w_h)}{\|v_h\|_{\tilde{V}}} &\geq \sup_{v \in \tilde{V}} \frac{\tilde{b}(\pi_h v, w_h)}{\|\pi_h v\|_{\tilde{V}}} \\ &\geq \frac{\tilde{\beta}}{c} \|w_h\|_{\tilde{W}} \end{aligned}$$

So the hypothesis (A.6) holds with $\gamma = \frac{\tilde{\beta}}{c}$. On the other hand, since u belongs to $\tilde{V}$, we have by (1.5) and (1.18):

$$\tilde{b}(\pi_h u, w_h) = \tilde{b}(u, w_h) = \langle \phi, w_h \rangle$$

hence $\pi_h u \in Z_h(\phi)$. So the desired estimates follows from estimates of the previous theorem. #

Let Λ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\Lambda$ and its induced norms $\|\cdot\|_\Lambda$ such that : $\tilde{W} \subset \Lambda$ with continuous imbedding.

Assume that problem (P) is regular in the following sense :

Given $\psi \in \Lambda$, let $(y, \chi) \in V \times \Lambda$ be the solutions of :

$$\forall v \in V, a(v, y) + b(v, \chi) = 0 \quad (3.20)$$

$$\forall w \in W, b(y, w) = \langle \psi, w \rangle_\Lambda \quad (3.21)$$

Then y belongs to the space $\tilde{V}$.

Theorem 6:

Assume problem (P) is regular in the sense (1.20) and (1.21). then we have some constant $K = K(\|a\|, \|\tilde{b}\|) > 0$,

$$\|p - w_h\|_\Lambda \leq \sup_{\psi \in \Lambda} \left\{ \frac{1}{\|\psi\|} \inf_{y_h \in Z_h(\phi), w_h \in W_h, \chi_h \in W_h} [\|e_1\|_V \|y - y_h\|_V] \right\}$$

$$+ \|e_1\|_{\tilde{V}} \|\chi - \chi_h\|_{\tilde{W}} \\ + \|y - y_h\|_{\tilde{V}} \|e_2\|_{\tilde{W}} \}}]$$

where, for any given $\psi \in \Lambda$, (y, χ) is the regular solution in the sense (1.20) and (1.21).

Proof

Let us write:

$$\|p - p_h\| = \sup_{\psi \in \Lambda} \frac{|\langle \psi, p - p_h \rangle|}{\|\psi\|_{\Lambda}}$$

Now let ψ be in Λ . Since problem (P) is regular, note that as in the proof of Theorem (-) that the solution (y, χ) of (-) satisfies:

$$\forall v \in \tilde{V}, a(v, y) + \tilde{b}(v, \chi) = 0$$

$$\forall w \in \tilde{W}, \tilde{b}(y, w) = \langle \psi, w \rangle_{\Lambda}$$

So we get:

$$a(u - u_h, y) + \tilde{b}(u - u_h, \chi) = 0$$

On the other hand, using (1.4), (1.5), (1.9), and (1.10), we have

$$\forall y_h \in V_h, a(u - u_h, y_h) + \tilde{b}(y_h, p - p_h) = 0$$

and

$$\forall \chi_h \in W_h, \tilde{b}(u - u_h, \chi_h) = 0$$

Therefore, we get for every $y_h \in V_h$ and every $\chi_h \in W_h$

$$\begin{aligned} \langle \psi, p - p_h \rangle_{\Lambda} &= \tilde{b}(y, p - p_h) \\ &= a(u - u_h, y - y_h) + \tilde{b}(u - u_h, \chi - \chi_h) + \tilde{b}(y - y_h, p - p_h) \end{aligned}$$

Next, assuming that $y_h \in Z_h(\psi)$. we have for all $w_h \in W_h$

$$\tilde{b}(y - y_h, p - p_h) = \tilde{b}(y - y_h, p - w_h)$$

So we obtain for every $y_h \in Z_h(\psi)$ and all $\chi_h, w_h \in W_h$

$$\begin{aligned} |\langle \psi, p - p_h \rangle_{\Lambda}| &\leq \|a\| \|u - u_h\|_V \|y - y_h\|_V \\ &\quad + \|\tilde{b}\| \|u - u_h\|_{\tilde{V}} \|\chi - \chi_h\|_{\tilde{W}} \\ &\quad + \|\tilde{b}\| \|y - y_h\|_{\tilde{V}} \|p - p_h\|_{\tilde{W}} \end{aligned}$$

from which the desired inequality follows. #

3.3 Application to First Biharmonic Equation

We consider the first biharmonic problem :

$$\Delta^2 u = g, \text{ in } \Omega$$

$$u|_{\partial\Omega} = \frac{\partial u}{\partial n}|_{\partial\Omega} = 0$$

Rewrite the problem as a coupled system as follows

$$w - \Delta u = 0, \text{ in } \Omega$$

$$\Delta w = g, \text{ in } \Omega$$

$$u|_{\partial\Omega} = \frac{\partial u}{\partial n}|_{\partial\Omega} = 0$$

Set :

$$V = L^2(\Omega)$$

$$W = H_0^2(\Omega)$$

$$a(u, v) = \int uv dx$$

$$b(v, w) = \int v \Delta w dx$$

$$f = 0$$

$$\langle \phi, \mu \rangle = - \int \phi \mu dx$$

We will check whether the hypothesis are satisfied. Since $a(u, v)$ in V is an inner product in $L^2(\Omega)$, by taking $\alpha = 1$, then $a(u, v)$ is V -elliptic. So obviously that the first hypothesis is satisfied. And we can verify that there exist $\beta \geq 0$ such that

$$\forall w \in W, \sup_{v \in V} \frac{b(v, w)}{\|v\|_V} \geq \beta \|w\|_W$$

We associate the problem (B.1) with a problem ($\tilde{B}1$) by setting

$$\tilde{V} = H^1(\Omega)$$

$$\tilde{W} = H_0^1(\Omega)$$

$$\tilde{b}(v, w) = - \int \nabla v \nabla w dx$$

The conditions (A.1) and (A.2) hold trivially and by Theorem 2, there exist a pair of functions $(u, p) \in H^1 \times H_0^1$ which is the unique solution of :

$$\int v u dx + \int \nabla v \nabla p dx = 0, \forall v \in H^1(\Omega)$$

$$\int \nabla p \nabla w dx + \int g w dx = 0, \forall w \in H_0^1(\Omega)$$

Let us assume that the solution resides in $H_0^2 \cup H^3$. Suppose that Ω is triangulated by triangulations T_h with triangles K whose sides are less than h . Define finite dimensional space:

$$V_h = \{v_h \in C^0(\Omega); K \in T_h, v_h|_K \in P_1\}$$

$$V_{h,0} = V_h \cap H_0^1$$

$$W_h = V_{h,0}$$

So we take piecewise linear basis. In order to apply Theorem 3, we need to check the hypothesis (A.6).

Lemma

Let there be given spaces V_h and W_h as above. Then there exist a constant $\gamma > 0$ independent of h such that consequences in Theorem 3 holds.

Proof

Let w_h be in W_h . By the definition of V_h and W_h , we may choose $v_h \in V_h$ by taking:

$$v_h = w_h$$

Then we get :

$$\tilde{b}(v_h, w_h) = |w_h|_1^2 \geq c_1 \|w_h\|_1^2$$

Where the constant c_1 is independent of h . Since $\|v_h\|_{\tilde{V}} = \|w_h\|_1$, we obtain the desired inequality with $\gamma = c_1$. #

Let us define Ritz operators $R_h : H^1 \rightarrow V_h$ and $R_{h,0} : H_0^1 \rightarrow V_{h,0}$ as follows

$$\begin{aligned}\tilde{b}(v - R_h v, \eta) &= 0, \forall \eta \in V_h \\ \int (v - R_h v) dx &= 0,\end{aligned}$$

and

$$\tilde{b}(v - R_{h,0} v, \xi) = 0, \forall \xi \in V_{h,0}$$

respectively.

The following Lemma is important to the proof of estimates for piecewise linear basis. The proof uses L^∞ -error estimates for the Ritz approximation of second order problem.

Lemma :[Sc78]

For all $u \in H_0^1 \cap W^{2,\infty}$ and for all $\eta \in V_h$ we have

$$|\tilde{b}(u - R_{h,0} u, \eta)| = ch^{1/2} |lnh| \|\Delta u\|_\infty \|\eta\|_0$$

with c is independent of u, η , and h . #

Proof

Let Ω_h be the union of all triangles K with $K \cap \partial\Omega \neq \emptyset$. We denote by $\xi \in V_{h,0}$ as function which interpolates η at the interior of grid-points of the triangulations T_h . For $\phi = \eta - \xi$, it follows that :

$$\phi(s) = 0, s \in \Omega - \Omega_h$$

Setting : $e = u - R_{h,0} u$ and applying the inverse estimate :

$$\|\phi\|_{W^{1,\infty}(K)} \leq ch^{-1} \|\phi\|_{L^\infty(K)}$$

we get:

$$\begin{aligned}|\tilde{b}(e, \eta)| &= |\tilde{b}(e, \phi)| \\ &= \left| \sum_{T_h} \int_K \nabla e \nabla \phi dx \right| \\ &\leq ch^2 \|e\|_{H^1} \sum_{T_h} \|\phi\|_{W^{1,\infty}(K)} \\ &\leq ch \|e\|_{W^{1,\infty}} \sum_{T_h} \|\phi\|_{L^\infty(K)}\end{aligned}$$

Applying the inverse estimate :

$$\|\eta\|_{L^\infty(K)} \leq ch^{-1}\|\eta\|_{L^2(K)}$$

we get :

$$\begin{aligned} |\tilde{b}(e, \eta)| &\leq \|e\|_{W^{1,\infty}} \sum_{T_h} \|\eta\|_{L^2(K)} \\ &\leq h^{-1/2} \|e\|_{W^{1,\infty}} \|\eta\|_{L^2(\Omega_h)} \end{aligned}$$

Since the number of triangles in Ω is of order h^{-1} . Combining with Nitsche results on L^∞ -estimate for Ritz-approximation:

$$\begin{aligned} \|e\|_{W^{1,\infty}} &= \|u - R_{h,0}u\|_{W^{1,\infty}} \\ &\leq ch|\ln h| \|\Delta u\|_{L^\infty} \end{aligned}$$

we proved the results of the lemma. #

Let $(u_h, p_h) \in V_{h,0} \times V_h$ be the solution of discrete problem. With $e_1 = (u - u_h)$ and $e_2 = (p - p_h)$, we can rewrite equations in (B.1h) in the following form :

$$\begin{aligned} \tilde{b}(e_1, \eta) &= (e_2, \eta), \forall \eta \in V_h \\ \tilde{b}(e_2, \xi) &= 0, \forall \xi \in V_{h,0} \end{aligned}$$

We get the following estimates in the L^2 -norm.

Theorem :[Sc78]

The differences e_1 and e_2 between the exact solution of the first biharmonic problem and the mixed finite element formulation can be estimated by:

$$\|e_1\|_0 + h^{1/2} |\ln h| \|e_2\|_0 \leq ch |\ln h|^2 \|u\|_4$$

where c is independent of (u, p) and h . #

Proof

Let $\phi_1 \in V_{h,0}$ and $\phi_2 \in V_h$ be the Ritz approximations of u and p respectively as defined above. We find that:

$$\begin{aligned} \|p - p_h\|_0^2 &= \langle p - \phi_2, p_h - \phi_2 \rangle - \tilde{b}(u_h - \phi_1, p_h - \phi_2) + \tilde{b}(u_h - \phi_1, p_h - \phi_2) \\ &= \langle p - \phi_2, p_h - \phi_2 \rangle - \tilde{b}(u - \phi_1, p_h - \phi_2) + \tilde{b}(u_h - \phi_1, p - \phi_2) \end{aligned}$$

Using standard error estimates for the Ritz approximation in the L^2 -norm, we can estimate the first term on the right hand side by:

$$|\langle p - \phi_2, p_h - \phi_2 \rangle| \leq ch^2 \|p\|_2 \|p_h - \phi_2\|_0$$

For estimates on the second term, with the help of previous lemma, we get

$$|\tilde{b}(u - \phi_1, p_h - \phi_2)| \leq ch^{1/2} |\ln h| \|\Delta u\|_\infty \|p_h - \phi_2\|_0$$

While the third term is zero by definition of ϕ_2 . Then :

$$\begin{aligned} \|e_2\|_0 &= \|p - p_h\|_0 \\ &\leq ch^2 \|p\|_2^2 + ch^{\frac{1}{2}} |\ln h| \|\Delta u\|_\infty^2 \end{aligned}$$

The other part of the estimates proven by following argument. Let $w \in H_0^2 \cap H^4$ be a solution of:

$$\begin{aligned} \Delta^2 w &= e_1, \text{ in } \Omega \\ w &= w_n, \text{ on } \partial\Omega \end{aligned}$$

Observing (eqno) we get:

$$\|e_1\|_0^2 = -\tilde{b}(e_1, \Delta w - R_h \Delta w) + \langle e_2, \Delta w - R_h \Delta w \rangle + \tilde{b}(e_2, w - R_{h,0} w)$$

Since u_h and ϕ_1 are elements of $V_{h,0}$, we obtain for the first term:

$$\begin{aligned} |\tilde{b}(e_2, w - \Delta w - R_h \Delta w)| &= |\tilde{b}(u - \phi_1, w - \Delta w - R_h \Delta w)| \\ &\leq ch^2 \|u\|_2 \|w\|_4 \end{aligned}$$

and the second term can be estimated by

$$|\langle e_2, \Delta w - R_h \Delta w \rangle| \leq ch^2 \|e_2\|_0 \|w\|_4$$

Finally, again with the help of previous lemma, we get:

$$|\tilde{b}(e_2, w - R_{h,0} w)| \leq ch^2 \|p\|_2 (\|w\|_2 + \|\Delta w\|_\infty) + ch^{1/2} |\ln h| \|e_2\|_0 \|\Delta w\|_\infty$$

Using Sobolev embedding theorem, we get:

$$\|\Delta w\|_\infty \leq \|w\|_4$$

Together with the estimate of e_2 we find :

$$|\tilde{b}(e_2, w - R_{h,0}w)| \leq ch|lnh|^2 \|p\|_2 \|w\|_4$$

combining this inequalities with (eqno) and observing that

$$\|w\|_4 \leq \|e_1\|_0$$

we obtain estimates:

$$\|e_1\|_0 + h^{1/2}|lnh|\|e_2\|_0 \leq ch|lnh|^2 \|u\|_4$$

where c is independent of (u, p) and h as desired. #

Corollary

As a consequence of estimates in the above theorem and combination of the first part of equation (-) we get for e_1 in the H^2 -norm :

$$\|e_1\|_2 \leq ch^{3/4}|lnh|^{3/2} \|u\|_4$$

#

Observe that Scholz results need a higher regularity, i.e $u \in H^4(\Omega) \cap H_0^1(\Omega)$ to show error estimates in L_2 for piecewise linear basis. While for approximation using $P_k, k \geq 2$, error estimates in $O(h^{k-1})$ is achieved, with weaker regularity, $u \in H^3(\Omega) \cap H_0^1(\Omega)$ [Mo87]. Furthermore, Glowinski [Gl73] showed that for a particular triangulation of a square error estimates in $O(h)$ is achieved.

3.4 Application to Second Biharmonic Equation

We consider the second biharmonic problem :

$$\Delta^2 u = g, \text{ in } \Omega$$

$$u|_{\partial\Omega} = \Delta u|_{\partial\Omega} = 0$$

Rewrite the problem as a coupled system as follows

$$w - \Delta u = 0, \text{ in } \Omega$$

$$\begin{aligned}\Delta w &= g, \text{ in } \Omega \\ u|_{\partial\Omega} &= w|_{\partial\Omega} = 0\end{aligned}$$

Set :

$$\begin{aligned}V &= L^2(\Omega) \\ W &= H_0^2(\Omega) \\ a(u, v) &= \int uv dx \\ b(v, w) &= \int v \Delta w dx \\ f &= 0 \\ [\phi, \mu] &= - \int \phi \mu dx\end{aligned}$$

We will check whether the hypothesis are satisfied. Since $a(u, v)$ in V is an inner product in $L^2(\Omega)$, by taking $\alpha = 1$, then $a(u, v)$ is V -elliptic. So obviously that the first hypothesis is satisfied. And we can verify that there exist $\beta \geq 0$ such that

$$\forall w \in W, \sup_{v \in V} \frac{b(v, w)}{\|v\|_V} \geq \beta \|w\|_W$$

We associate the problem(B.2) with a problem $(\tilde{B}2)$ by set

$$\begin{aligned}\tilde{V} &= H_0^1(\Omega) \\ \tilde{W} &= H_0^1(\Omega) \\ \tilde{b}(v, w) &= - \int \nabla v \nabla w dx\end{aligned}$$

The conditions (A.1) and (A.2) are holds trivially and by Theorem 2, there exist pair of functions $(u, h) \in H_0^1 \times H_0^1$ is the unique solution of :

$$\begin{aligned}\int v u dx + \int \nabla v \nabla p dx &= 0, \forall v \in H_0^1(\Omega) \\ \int \nabla p \nabla u dx + \int g w dx &= 0, \forall w \in H_0^1(\Omega)\end{aligned}$$

Let us assume that the solution be reside in $H_0^2 \cup H^3$ Suppose that Ω is triangulated by triangulations T_h with triangles K whose sides are less than h . Define finite dimensional space:

$$V_h = \{v_h \in C^0(\Omega); K \in T_h, v_h|_K \in P_1\}$$

$$V_{h,0} = V_h \cap H_0^1$$

$$W_h = V_{h,0}$$

So we take piecewise linear basis. In order to apply Theorem 3, we need to check the hypothesis (A.6). This is obviously clear from the result from the previous chapter, since $H_0^1 \subset H^1$.

The error bounds for first biharmonic problems can be extended to the second biharmonic problems with only minor changes in the proof. Furthermore, error bounds for Poisson equation can be extended to the second biharmonic problems by treating the splitting of Biharmonic equation as a coupled set of Poisson equations. Then we obtain the following theorems as consequences.

Theorem :

i. $u \in H^4(\Omega)$ then:

$$\|u - P_h u\|_0 \leq c_1 h^2 |u|_2$$

$$\|p - P_h p\|_0 \leq c_1 h^2 |p|_2$$

then :

$$\|u - P_h u\|_0 + \|p - P_h p\|_0 \leq h^2 (c_1 |u|_2 + c_2 |p|_2)$$

ii. $u \in H^4(\Omega)$ then:

$$\|u - R_h u\|_0 \leq c_1 h^2 |u|_2$$

$$\|p - R_h p\|_0 \leq c_1 h^2 |p|_2$$

then :

$$\|u - R_h u\|_0 + \|p - R_h p\|_0 \leq h^2 (c_1 |u|_2 + c_2 |p|_2)$$

Similarly from L_2 -estimates of Poisson equation, we have directly the following $L - 2$ -estimates of second biharmonic problems.

Theorem

i.

$$\|p - R_h p\|_0 \leq c_1 h^2 |p|_2 \leq c_1 h^2 \|f\|_0$$

ii.

$$\|u - R_h u\|_0 \leq c_2 h^2 |u|_2 \leq c_2 h^2 \|p\|_0$$

Chapter 4

Iterative Methods for Non-Singular Systems of Linear Equations

In this chapter we consider a family of methods to solve systems of linear equation which arise from the application of finite element method. We seek an efficient solution method for these system which are characterized by its sparsity and large number of unknown. The basic choice is between a direct method , such as Gaussian elimination or Cholesky's method, and an iterative method in which the desired solution is the limiting value of a sequence of approximations.

Direct methods require the storage of both the whole system matrix and the constant vector, or at least the relevant parts, since they are required throughout the solution process. In this way, the capacity of computer storage sets a limit on the order of linear systems, hence economization techniques for storage usage became important. Moreover, these methods often destroy the sparsity structure of linear systems due to fill in.

Iterative methods have the property that they can utilize the sparsity structure of the systems, and leave it unchanged. Thus the non-zero elements can be stored efficiently, and the required storage is lower compared to direct method.

In this chapter, a family of iterative method are described. These methods are known to be parameter free and have a finite termination property, i.e in exact arithmetic will terminates at most of the order of systems being solved.

The exposition of this chapter is arranged as follows: The first and second section described the classical Conjugate Gradient method for a symmetric positive definite linear equation and its extension to solve non-singular linear equation. The third section concerns preconditioning by diagonal scaling. The following three sections describe three interrelated methods to solve a nonsingular linear equation. The final section concerns a stopping criteria for iterative processes.

4.1 Conjugate Gradient

The standard conjugate gradient method for symmetric positive definite linear systems generates a sequence of approximate solutions for $x : x^{(0)}, x^{(1)}, \dots$, by the following recursive process:

$$x^{(k+1)} = x^{(k)} + \alpha_{(k)} p^{(k)}$$

$$\alpha_{(k)} = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle p^{(k)}, Ap^{(k)} \rangle}$$

$$r^{(k+1)} = r^{(k)} - \alpha_{(k)} Ap^{(k)}$$

$$p^{(k+1)} = r^{(k)} + \beta_{(k)} p^{(k)}$$

where parameter $\beta_{(k)}$ will be determined.

Let :

$$S_k = \text{span}(Ar^{(0)}, \dots, A^k r^{(0)})$$

$$K(r^{(0)}, A, k) = \text{span}(r^{(0)}, Ar^{(0)}, \dots, A^{k-1} r^{(0)})$$

a Krylov subspace generated by $r^{(0)}$

$$T_k = \{r; r = r^{(0)} + h, h \in S_k\} \subset K(r^{(0)}, A, k+1)$$

Theorem :[AB85]

If we impose the condition :

$$\|r^k\|_{A^{-1}} = \min_{r \in T_k} \|r\|_{A^{-1}}$$

Then the parameter $\beta_{(k)}$ in the above recursive process is given by:

$$\beta_k = -\frac{\langle r^{(k+1)}, Ap^{(k)} \rangle}{\langle p^{(k)}, Ap^{(k)} \rangle}$$

Furthermore, this condition also implies the relations :

$$\langle r^k, r^l \rangle = 0, k \neq l$$

and

$$\langle p^k, Ap^l \rangle = 0, k \neq l$$

Proof

The condition:

$$\|r^k\|_{A^{-1}} = \min_{r \in T_k} \|r\|_{A^{-1}}$$

is equivalent to:

$$\|r^0 + h^k\|_{A^{-1}} = \min_{h \in S_k} \|r^0 + h\|_{A^{-1}}$$

where $h^k = r^k - r^0$. Now if we regard any $h \in S_k$ as an approximation to the vector $-r^0$, then the approximation error is $h - (-r^0) = h + r^0$. So we have a problem to find the vector r^k in the subspace that closely approximates $-r^0$, with respect to A^{-1} -norm. According to the standard result in linear algebra such a vector exists and is unique, and it also makes the error $h^k + r^0$ is A^{-1} -orthogonal to every h in S_k , i.e :

$$\langle (h^k + r^0), A^{-1}h \rangle = \langle (h^k + r^0), h \rangle_{A^{-1}} = 0$$

For any $r \in T_{k-1}$, the vector $h = Ar$ belongs to S_k . Then r^k has property that :

$$\langle r^k, r \rangle = 0, \forall r \in T_{k-1}$$

and orthogonality for residual follows since we have $r^l \in T_l \subset T_{k-1}$, for $l \leq k-1$

To obtain A-orthogonality of search direction vector, assume $l \leq k$,

$$\langle p^{(k)}, Ap^{(l)} \rangle = \langle Ap^{(k)}, p^{(k)} \rangle$$

Multiplying the recursive formula of $x^{(k+1)}$ by A and subtract b, we obtain:

$$r^{k+1} = r^k + \alpha_k Ap^k$$

or

$$Ap^k = \frac{(r^{k+1} - r^k)}{\alpha_k}$$

substitute this into:

$$\langle p^{(k)}, Ap^{(l)} \rangle = \langle Ap^{(k)}, p^{(k)} \rangle$$

we obtain:

$$\langle p^{(k)}, Ap^{(l)} \rangle = \frac{1}{\alpha_k} \langle (r^{k+1} - r^k), p^{(k)} \rangle.$$

Substituting the search direction formula at l -th step so that:

$$\langle p^{(k)}, Ap^{(l)} \rangle = \frac{1}{\alpha_k} \langle (r^{k+1} - r^k), r^l + \beta_{l-1} p^{(l-1)} \rangle$$

and then applying orthogonality relation of residual we obtain:

$$\langle p^{(k)}, Ap^{(l)} \rangle = \frac{\beta_{l-1}}{\alpha_k} \langle (r^{k+1} - r^k), p^{(l-1)} \rangle$$

Hence by induction :

$$\langle p^{(k)}, Ap^{(l)} \rangle = \frac{(\prod_{i=1}^{l-1} \beta_i)}{\alpha_k} \langle (r^{k+1} - r^k), p^{(0)} \rangle$$

$$\langle p^{(k)}, Ap^{(l)} \rangle = 0$$

since $p^0 = r^0$ by construction.

Now it remains to show the formula for β_k . From A-orthogonality of p^{k+1} to p^k and the formula for search direction, we obtain:

$$0 = \langle p^{(k+1)}, Ap^{(k)} \rangle = \langle r^{k+1} + \beta Ap^{(k)}, p^{(k)} \rangle$$

hence we get the formula for β_k

$$\beta_k = - \frac{\langle r^{(k+1)}, Ap^{(k)} \rangle}{\langle p^{(k)}, Ap^{(k)} \rangle}$$

as required. #

Since:

$$\langle r^k, r^k \rangle_{A^{-1}} = \langle A^{-1} r^k, A^{-1} r^k \rangle_A$$

Then the minimization property is equivalent with:

$$\min_{r \in T_k} \|r\|_{A^{-1}} = \min_{r \in T_k} \|A^{-1} r\|_A$$

$$\begin{aligned}
\min_{r \in T_k} \|r\|_{A^{-1}} &= \min_{h \in S_k} \|A^{-1}(r^0 + h)\|_A \\
\min_{r \in T_k} \|r\|_{A^{-1}} &= \min_{h \in S_k} \|(x - x^0) + A^{-1}h\|_A \\
\min_{r \in T_k} \|r\|_{A^{-1}} &= \min_{h \in S_k} \|x - (x^0 - A^{-1}h)\|_A \\
\min_{r \in T_k} \|r\|_{A^{-1}} &= \min_{u \in x^0 + K(r^0, A, k)} \|x - u\|_A
\end{aligned}$$

So vector $x^{(i)}$ is the vector argument for which $\|x - u\|_A$ is minimal over all $u \in x^{(0)} + K(r^{(0)}, A, i)$, where $\|v\|_A = \langle v, Av \rangle$.

It follows that any $u \in K(r^{(0)}, A, i)$ satisfies : $x - u = q(A)(x - x^{(0)})$ where $q \in \Pi_i^1$, Π_i^1 denotes set all polynomials with constant term 1 and of order at most i .

From minimization property of $x^{(i)}$, it follows that

$$\|x - x^{(i)}\|_A = \min \|q(A)(x - x^{(0)})\|_A$$

where the minimum is taken over all $q \in \Pi_i^1$. Let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ be the eigenvalue and its corresponding normalized eigenvectors $v^1, v^2, \dots, v^m, m \leq n$, of A . Then : $r^{(0)} = \sum_{i=1}^m \mu_i z_i, \mu_i > 0$, for all i in exact arithmetic we may infer that $\dim K(r^{(0)}, A, i) = i, i \leq m$

Let R_i is a matrix which has as its column: $r^{(0)}/\|r^{(0)}\|, r^{(1)}/\|r^{(1)}\|, \dots, r^{(i-1)}/\|r^{(i-1)}\|$, then $R_i^T A R_i$ is representation of A projected onto $K(r^{(0)}, A, i)$. The eigenvalues $\theta_1^{(i)} < \theta_2^{(i)} < \dots < \theta_i^{(i)}$ of $R_i^T A R_i$ are called Ritz values of A with respect to $K(r^{(0)}, A, i)$.

4.1.1 Convergence Behavior of Conjugate Gradient

From the minimization property of $x^{(i)}$ it follows that:

$$\begin{aligned}
\|x - x^{(i)}\|_A &\leq \|q(A)(x - x^{(0)})\|_A \\
&\leq \|q(A)\|_F \|(x - x^{(0)})\|_A \\
&\leq \max |q(\lambda)| \|(x - x^{(0)})\|_A
\end{aligned}$$

for any $q \in \Pi_i^1$, where $\|(-)\|_F$ denotes Frobenius norm and $\|(-)\|_A$ denotes A-norm.

If we take for $q(\lambda)$ the Chebychev polynomials shifted to $[\lambda_1, \lambda_m]$ and normalized such that $q(0) = 1$, then we obtain the upperbound :

$$\|x - x^{(i)}\|_A \leq 2 \left\{ \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right\}^i \|x - x^{(0)}\|_A$$

with $\kappa(A) = \frac{\lambda_m}{\lambda_1}$, condition number of A .

The actual behavior of Conjugate Gradients is not yet understood theoretically, van der Sluis and van der Vorst [SV86],[SV87] studied it empirically and made some observation as follows. The increase of speed of convergence can be explained, as soon as a Ritz value get closer to one of the eigenvalue, say λ_1 , the Conjugate Gradient process behaves from then on as a process corresponds to a starting vector in which z_1 and λ_1 is not present. This process then converge faster with a reduction factor $\frac{(\sqrt{\kappa(A)}-1)}{(\sqrt{\kappa(A)}+1)}$.

Observation of modest approximation of θ_1^i to λ_1 is sufficient to shift to a lower reduction factor. For instances it suffices when $\theta_1^i < \frac{(\lambda_1 + \lambda_2)}{2}$.

The study of Ritz value make a way of understanding fully the behavior of Conjugate Gradient convergence. It was observed that as the gap $\frac{(\lambda_2 - \lambda_1)}{(\lambda_m - \lambda_1)}$ becomes larger the speed of convergence becomes faster, and the convergence of θ_1^i towards λ_1 is monotonically as i increases.

4.2 Conjugate Gradient for Normal Equation

Let A be an $n \times n$ nonsingular matrix. Then the Conjugate Gradient method can be used as an iterative method to solve original linear equation via normal equation :

$$A^T A x = A^T b$$

In this form, obviously $A^T A$ is a positive definite matrix. Hence all convergence properties of Conjugate iteration are still applicable to the normal equation, where now we use the condition number $\kappa(A^T A) = (\kappa(A))^2$

4.3 Diagonal Scaling as Preconditioner

Let $K = LL^T$ be non singular. The linear system $Ax = b$ equivalently can be solved in the form :

$$\hat{A}\hat{x} = \hat{b}$$

where $\hat{A} = L^{-1}AL$, $\hat{b} = L^{-1}b$, and $\hat{x} = Lx$. This is called preconditioning the original system.

The objective of preconditioning is to find a matrix L such that the preconditioned matrix $L^{-1}AL^{-T}$ has a smaller condition number or larger gap ratio for extremal eigenvalues with respect to A . In view of the convergence behavior of conjugate gradient method for A a symmetric positive definite matrix, this is an attempt to reduce iteration steps.

One of the methods of preconditioning is to choose L as a diagonal matrix; this method is called diagonal scaling.

Let $D = \{D; \text{diagonal}, n \times n\}$. Define *minimum condition number* of A to be :

$$\kappa^*(A) = \min_D (DAD)$$

Then obviously $\kappa^*(A) \leq \kappa(A)$. A standard choice for diagonal preconditioner is $D = \text{diag}(a_{11}^{-\frac{1}{2}}, \dots, a_{nn}^{-\frac{1}{2}})$. According to the results of van der Sluis [Sl69], this choice of preconditioner has property:

$$\kappa^*(A) \leq n\kappa(DAD)$$

So, this diagonal preconditioner almost minimizes the condition number of A , symmetric positive definite matrix, over all possible choice of diagonal matrix. Schwartz[Sc88], also recommends to use diagonal preconditioner for symmetric positive definite matrices of the form:

$$d_{11} = 1,$$

$$d_{22} = 1,$$

$$d_{ii} = 1, i = 2j$$

$$d_{ii} = \frac{1}{l}, i = 2j - 1$$

where l , is a scaling parameter.

4.4 BiConjugate Gradient

The residual vector generated by Conjugate Gradient method satisfy a recurrence relation which make the method efficient. When A is not symmetric, the recurrence property for residual vectors is lost. Moreover, if A is not positive definite, then $\| - \|_A$ does not define a norm, hence minimizing $\|(x - x^{(i)})\|_A$ does not makes sense.

The BiConjugate Gradient Method (BCG)[Fl76], first developed by Lanczos to compute eigenvalues and eigenvector of nonsymmetric matrix and Fletcher applied the method to solve symmetric indefinite matrix problems. This method can be regarded as generalized Conjugate Gradient method.

The BCG algorithm approximates the solution by constructing two sets of residual vectors $r^{(i)}$ and $\hat{r}^{(i)}$ (called a biresidual vector) and two sets of directional vector $p^{(i)}$ and $\hat{p}^{(i)}$ (called a bidirectional vector) such that: $r^{(i)}$ is orthogonal to $\hat{r}^{(1)}, \hat{r}^{(2)}, \dots, \hat{r}^{(i-1)}$ (biorthogonality relations) and vice versa. This strategy can be achieved by two recurrence relations for $r^{(i)}$ and $\hat{r}^{(i)}$.

The BCG algorithm proceeds as follows:

Algorihm. BiConjugate Gradient

$x^{(0)}$ is given ; $r^{(0)} = b - Ax^{(0)}$;

$\hat{r}^{(0)}$ is arbitrary such that: $\langle r^{(0)}, \hat{r}^{(0)} \rangle \neq 0$;

$\rho_{(0)} = 1$; $p^{(0)} = \hat{p}^{(0)} = 0$;

for $i = 1, 2, \dots$;

$\rho_i = \langle r^{(i-1)}, \hat{r}^{(i-1)} \rangle$;

$\beta_i = \frac{\rho_i}{\rho_{(i-1)}}$;

$p^i = r^{(i-1)} + \beta_i p^{(i-1)}$;

$\hat{p}^i = \hat{r}^{(i-1)} + \beta_i \hat{p}^{(i-1)}$;

$v^i = Ap^i$;

$\alpha_i = \frac{\rho_i}{\langle \hat{p}^i, v^i \rangle}$;

$x^i = x^{(i-1)} + \alpha_i p^i$;

if x^i is accurate enough then quit ;

$r^i = r^{(i-1)} - \alpha_i v^i$;

$\hat{r}^i = \hat{r}^{(i-1)} - \alpha_i A^T \hat{p}^i$;

end

It can be shown that the expression for $\alpha_{(i)}$ and $\beta_{(i)}$ given above ensures the biconjugacy relations :

$$\langle \hat{p}^{(i)}, Ap^{(j)} \rangle = 0, i \neq j$$

and

$$\langle p^{(i)}, A^T \hat{p}^{(j)} \rangle = 0, i \neq j$$

and the biorthogonality relations :

$$\langle \hat{r}^{(i)}, r^{(j)} \rangle = \langle r^{(i)}, \hat{r}^{(j)} \rangle, i > j$$

and

$$\langle \hat{r}^{(i)}, p^{(j)} \rangle = \langle r^{(i)}, \hat{p}^{(j)} \rangle, i > j$$

The residual vector $r^{(i)}$ at i -th iteration is in the right Krylov subspace generated by $r^{(0)} : r^{(i)} \in K(A, r^{(0)}, i - 1)$ While the biresidual vector $\hat{r}^{(i)}$ at i -th iteration is in the left Krylov subspace generated by $\hat{r}^{(0)} : \hat{r}^{(i)} \in K(A^T, \hat{r}^{(0)}, i - 1)$ Note that if A is a symmetric positive definite matrix, BCG and Conjugate Gradient are identical.

The biorthogonality and biconjugacy relations imply that BCG is a finite method, i.e will terminate at most n steps under exact arithmetic. This scheme also suffer the possibility of early termination due to breakdown caused by division by zero. The breakdown at k -th iteration will occurred if:

$$\langle \hat{p}^{(k-1)}, A p^{(k-1)} \rangle = 0, \text{ for } \hat{p}^{(k-1)} \neq 0 \neq p^{(k-1)}$$

or

$$\langle \hat{r}^{(k-1)}, r^{(k-1)} \rangle = 0, \text{ for } \hat{r}^{(k-1)} \neq 0 \neq r^{(k-1)}$$

The first cause of breakdown, named as *the first kind breakdown*, is that there are no BCG k -th iterates. This breakdown is happens to be that the right Krylov subspace generated from initial residual is a proper invariant subspace of column space of A . [Jo92]

The second cause of breakdown, named as *the second kind breakdown*, is equivalent with the serious breakdown of the classical Lanczos process for unsymmetric eigenvalue problem. [Jo92]

This can be stated in the following theorem [Fl76], [Jo92].

Theorem

Suppose the algorithm has succesfully generated $x^{(k-1)} \neq x^*$.

Then the following are equivalent:

- (i) The algorithm breakdown at step k .
- (ii) Either $[K(A^T, \hat{r}^{(0)}, k - 1)]^T [K(A, r^{(0)}, k - 1)]$ or $[K(A^T, \hat{r}^{(0)}, k)]^T A [K(A, r^{(0)}, k)]$ is singular. #

Joubert [Jo92] proposed a strategy to handle the case of nearbreakdown occurrence. If this occurrence is detected, he recommends that the iteration to be restarted, where the current residual becomes the initial residual. To detect the occurrence of breakdown he propose a criterion :

$$\frac{|\langle \hat{p}^{(n)}, Ap^{(n)} \rangle|}{\|\hat{p}^{(n)}\| \|Ap^{(n)}\|} \leq \epsilon^{1/2}$$

to be satisfied, where ϵ is the unit round off of the machine.

4.5 Conjugate Gradient Squared

Sonneveld [So89] made an observation that for BiConjugate Gradient, in case of convergence, the residual vector $r^{(i)}$ and biresidual vector $\hat{r}^{(i)}$ both converge to zero, while in the process only $r^{(i)}$ is exploited. He proposes a modification for BiConjugate Gradient, that concentrates only on $r^{(i)}$ for convergence effort. This modification turns out to be that the residual vector at i-th iteration :

$$r^{(i)} = (P_i(A))^2 r^{(0)}$$

where $P_i(A)$ is the polynomial of degree i , such that $P_i(A)r^{(0)}$ is residual vector of BiConjugate Gradient at i-th step.

The proposed method only uses the single residual vector $r^{(i)}$ and if the polynomial $P_i(A)$ viewed as a reductor acting on $r^{(0)}$, then the method apply it twice for the same iterands, hence suggested the name Conjugate Gradient Squared. The prefix Bi is dropped since the method only uses a single residual vector.

For BCG vectors $r^{(i)}$ and $\hat{r}^{(i)}$ can be written as :

$$r^{(i)} = P_i(A)r^{(0)}$$

and

$$\hat{r}^{(i)} = P_i(A^T)\hat{r}^{(0)}$$

by the biorthogonality of $r^{(i)}$ and $\hat{r}^{(j)}$, for $i < j$, we obtain :

$$\begin{aligned} \langle r^{(i)}, \hat{r}^{(j)} \rangle &= \langle P_i(A)r^{(0)}, P_j(A^T)\hat{r}^{(0)} \rangle \\ &= \langle P_j(A)P_i(A)r^{(0)}, \hat{r}^{(0)} \rangle, i < j \end{aligned}$$

The parameter of BCG iterations are recovered from the above constructions of inner products between residual and biconjugate residual. In this way, Sonneveld observed that we can construct $\hat{r}^{(j)} = (P_j(A))^2 r^{(0)}$, by using only the previous form of inner products to recover BCG parameters. By doing this, it is not necessary to compute $\hat{r}^{(j)}$, and the multiplication by transpose of A is avoided.

The resulting algorithm can be represented by the following scheme:

Algorithm. Conjugate Gradient Square

$x^{(0)}$ is given ; $r^{(0)} = b - Ax^{(0)}$;

$\hat{r}^{(0)}$ is arbitrary such that: $\langle r^{(0)}, \hat{r}^{(0)} \rangle \neq 0$;

$\rho_{(0)} = 1$; $p^{(0)} = \hat{p}^{(0)} = 0$;

for $i = 1, 2, \dots$;

$\rho_i = \langle r^{(i-1)}, \hat{r}^{(0)} \rangle$;

$\beta_i = \frac{\rho_i}{\rho_{(i-1)}}$;

$u = r^{(i-1)} + \beta_i q^{(i-1)}$;

$p^i = u + \beta_i (q^{(i-1)} + \beta_i p^{(i-1)})$;

$v^i = Ap^i$;

$\alpha_i = \frac{\rho_i}{\langle \hat{r}^{(0)}, v^i \rangle}$;

$q^i = u - \alpha_i v$;

$w = u + q^i$;

$x^i = x^{(i-1)} + \alpha_i w$;

if x^i is accurate enough then quit ;

$r^i = r^{(i-1)} - \alpha_i Aw$;

end

Since the CGS is derived from BCG scheme, then the breakdown phenomenon still apparent. Hence the restarting strategy for BCG scheme can be extended to CGS. The criterion proposed to detect near breakdown occurrence for CGS to be satisfied is [Jo92]:

$$\frac{|\langle r^{(0)}, Ap^{(n)} \rangle|}{\|r^{(0)}\| \|Ap^{(n)}\|} \leq \epsilon^{1/2}$$

where ϵ is the unit round off of the machine.

4.6 BiCGSTAB

As previously mentioned, in the CGS method, at the i -th iteration, the residual at that step satisfies the relation : $r^{(i)} = (P_i(A))^2 r^{(0)}$, where $P_i(A)r^{(0)} = r_{BCG}^{(i)}$ is the residual of BCG at some iteration. We have that for $i < j$, $\langle P_i(A)r^{(0)}, P_j(A^T)\hat{r}^{(0)} \rangle = 0$ by construction, which expresses that $P_i(A)r^{(0)}$ is orthogonal to the left Krylov subspace $K(A^T, \hat{r}^{(0)}, j)$. Hence in principle we can recover the BCG parameters for each iteration by requiring for example, the orthogonality of $r^{(i)}$ to $P_j(A^T)\hat{r}^{(0)}$, or equivalently $\langle \hat{P}_j(A)P_i(A)r^{(0)}, \hat{r}^{(0)} \rangle = 0$ for a suitable polynomial $\hat{P}_j$ of degree j . The polynomial $\hat{P}_j$ taken in BCG is $\hat{P}_j = P_j$, that is $\hat{r}^{(j)} = P_j(A^T)r^{(0)}$. As indicated before, in CGS this is exploited by a recursion relation for the vector $(P_j(A))^2 r^{(0)}$ that can be derived from the form $P_j(A)r^{(0)}$.

Using the same line of thought, we can construct other iteration methods, by which $x^{(i)}$ are generated such that $r^{(i)} = \tilde{P}_i(A)P_i(A)r^{(0)}$, with i -th degree polynomials, for example using Chebychev polynomials. Since the optimal parameter of Chebychev polynomials in general are not easily obtainable and the recurrence relations for the resulting method are more complicated than CGS, the polynomials $\tilde{P}_j$ chosen for the candidates is of the form :

$$Q_i(x) = (1 - \omega_1 x)(1 - \omega_2 x) \dots (1 - \omega_i x)$$

and select the suitable constant ω_i . The choice of this form leads to a more simple recurrence relation for Q_i . The constant ω_i are selected such that $r^{(i)}$ is minimized with respect to ω_i .

Rewriting the BCG scheme, in which the residual $Q_i(A)P_i(A)r^{(0)}$ are generated, which results is an algorithm that more stable and efficient. Since this algorithm has connection with CGS and BCG, van der Vorst named this algorithm as BiCGSTAB [Vo92].

From the BCG scheme, it follows that $r^{(i)} = P_i(A)r^{(0)}$ and $p^{(i+1)} = T_i(A)r^{(0)}$ where P_i and T_i denotes polynomials with degree i at most. The relations between these polynomials in BCG algorithms defined as :

$$T_i(A)r^{(0)} = (P_i(A) + \beta_{(i+1)}T_{(i-1)}(A))r^{(0)}$$

and

$$P_i(A)r^{(0)} = (P_{(i-1)}(A) - \alpha_{(i)}AT_{(i-1)}(A))r^{(0)}$$

In the BiCGSTAB algorithm we want to have a recurrence relations for

$$r^{(i)} = (Q_{(i)}(A)P_{(i)}(A))r^{(0)}$$

where Q_i is a multiplication of a monomial as $(-)$. Employing factors of P_i and T_i in the BCG relations, we have :

$$\begin{aligned} Q_i(A)P_i(A)r^{(0)} &= ((1 - \omega_i A)Q_i(A)(P_{(i-1)}(A) - \alpha_{(i)}AT_{(i-1)}(A))r^{(0)}) \\ &= \{Q_{(i-1)}(A)P_{(i-1)}(A) - \alpha_{(i)}AQ^{(i-1)}T_{(i-1)}(A)\}r^{(0)} \\ &\quad - \omega_i A\{Q_{(i-1)}(A)P_{(i-1)}(A) - \alpha_{(i)}AQ^{(i-1)}T_{(i-1)}(A)\}r^{(0)} \end{aligned}$$

We also need a relation for product $Q_i(A)T_i(A)$. From the BCG relations, we obtain:

$$\begin{aligned} Q_i(A)T_i(A)r^{(0)} &= Q_i(A)(P_{(i)}(A) + \beta_{(i+1)}T_{(i-1)}(A))r^{(0)} \\ &= Q_i(A)P_{(i)}(A)r^{(0)} + \beta_{(i+1)}(1 - \omega_i A)Q_{(i-1)}(A)T_{(i-1)}(A)r^{(0)} \\ &= Q_i(A)P_{(i)}(A)r^{(0)} + \beta_{(i+1)}Q_{(i-1)}(A)T_{(i-1)}(A)r^{(0)} \\ &\quad - \beta_{(i+1)}\omega_i AQ_{(i-1)}(A)T_{(i-1)}(A)r^{(0)} \end{aligned}$$

In terms of the new vector we have generated, the BCG constants $\rho_{(i)}$, $\beta_{(i)}$, and $\alpha_{(i)}$ can be recovered via inner products as follows. First, let us compute parameter $\beta_{(i)}$. Define parameter :

$$\begin{aligned} \hat{\rho}_{(i+1)} &= \langle r^{(0)}, Q_i(A)P_i(A)r^{(0)} \rangle \\ &= \langle Q_i(A^T)r^{(0)}, P_i(A)r^{(0)} \rangle \end{aligned}$$

By construction, $P_i(A)r^{(0)}$ is orthogonal to all vectors $U_{(i-1)}(A^T)r^{(0)}$ where $U_{(i-1)}$ is an arbitrary polynomial of degree $(i-1)$ at most. Hence we left only to take into consideration the highest order term of $Q_i(A^T)r^{(0)}$ when computing $\hat{\rho}_{(i+1)}$. This term is given by $(-1)^i \omega_1 \omega_2 \dots \omega_i (A^T)^i$. We want to compute BCG parameter :

$$\rho_{(i+1)} = \langle P_i(A^T)\hat{r}^{(0)}, P_i(A)r^{(0)} \rangle$$

By similar argument as previously, it is left only the highest order term of $P_i(A^T)$ given by $(-1)^i \alpha_1 \alpha_2 \dots \alpha_i (A^T)^i$, it follows that :

$$\beta_{(i+1)} = \frac{\tilde{\rho}_{(i)} \alpha_{(i-1)}}{\tilde{\rho}_{(i-1)} \omega_{(i-1)}}$$

By writing $r^{(i)}$ for $Q_i(A)P_i(A)r^{(0)}$ and $p^{(i)}$ for $Q_{(i-1)}(A)T_{(i-1)}(A)r^{(0)}$ and compute ω_i such that $r^{(i)}$ is minimized in 2-norm as function of ω_i , we obtain the following scheme :

Algorithm. BiCGSTAB

$x^{(0)}$ is given ; $r^{(0)} = b - Ax^{(0)}$;
 $\hat{r}^{(0)}$ is arbitrary such that: $\langle r^{(0)}, \hat{r}^{(0)} \rangle \neq 0$;
 $\rho_{(0)} = \alpha_{(0)} = \omega_{(0)} = 1$; $p^{(0)} = v^{(0)} = 0$;
for $i = 1, 2, \dots$;
 $\rho_i = \langle r^{(i-1)}, \hat{r}^{(0)} \rangle$;
 $\beta_i = \frac{\rho_i}{\rho_{(i-1)} \omega_{(i-1)}}$;
 $p^i = u + \beta_i(p^{(i-1)} + \omega_{(i-1)}v^{(i-1)})$;
 $v^i = Ap^i$;
 $\alpha_i = \frac{\rho_i}{\langle \hat{r}^{(0)}, v^i \rangle}$;
 $s = r^{(i-1)} - \alpha_i v^i$;
 $t = As$;
 $\omega_i = \frac{\langle t, s \rangle}{\langle t, t \rangle}$;
 $x^i = x^{(i-1)} + \alpha_i p^i + \omega_i t$;
 if x^i is accurate enough then quit ;
 $r^i = s - \omega_i t$;

end

The BiCGSTAB algorithm is a finite method, it will terminates at $m \leq n$ iterations at most. This is due to the orthogonality property : $\langle P_i(A)r^{(0)}, Q_j(A)\hat{r}^{(0)} \rangle = 0$ for $j < i$.

For each iteration in the BiCGSTAB algorithm, it requires 2 matrix-vector multiplication, 4 inner product, and 12 n flops vector updates. It requires 1 matrix storage and 7 vector storage.

Since BiCGSTAB is deived from BCG schemes, then the Krylov subspace generated in the process is identical. Hence the breakdown phenomena is also unavoidable. For this reason, the criterion used in CGS to detect the occurence of breakdown can be used.

4.6.1 Preconditioning

Let K be a suitable preconditioning for A , i.e $K \approx A$, then we write: $K = K_1 K_2$ and we may apply the BiCGSTAB methods to the explicitly

preconditioned system:

$$\tilde{A}y = \tilde{b}$$

where: $\tilde{A} = K_1^{(-1)}AK_2^{(-1)}$, $x = K_2^{(-1)}y$, and $\tilde{b} = K_1^{(-1)}b$. Now we write the BiCGSTAB method for the explicit preconditioned system, with change of variables :

$$\begin{aligned}\tilde{p}^i &\Rightarrow K_1^{(-1)}p^i, \tilde{v}^i \Rightarrow K_1^{(-1)}v^i \\ \tilde{r}^i &\Rightarrow K_1^{(-1)}r^i, \tilde{s}^i \Rightarrow K_1^{(-1)}s^i \\ \tilde{t}^i &\Rightarrow K_1^{(-1)}t^i, \tilde{x}^i \Rightarrow K_1^{(-1)}x^i \\ \tilde{r}^{(0)} &\Rightarrow K_1^T r^{(0)}\end{aligned}$$

We obtain the following scheme for preconditioned BiCGSTAB:

Algorithm. Preconditioned BiCGSTAB

$x^{(0)}$ is given ; $r^{(0)} = b - Ax^{(0)}$;

$\hat{r}^{(0)}$ is arbitrary such that: $\langle r^{(0)}, \hat{r}^{(0)} \rangle \neq 0$;

$\rho_{(0)} = \alpha(0) = \omega(0) = 1$; $p^{(0)} = v^{(0)} = 0$;

for $i = 1, 2, \dots$;

$$\rho_i = \langle r^{(i-1)}, \hat{r}^{(0)} \rangle ;$$

$$\beta_i = \frac{\rho_i}{\rho_{(i-1)} \omega_{(i-1)}} ;$$

$$p^i = u + \beta_i(p^{(i-1)} + \omega_{(i-1)}v^{(i-1)}) ;$$

solve y from $Ky = p^i$;

$$v^i = Ay ;$$

$$\alpha_i = \frac{\rho_i}{\langle \hat{r}^{(0)}, v^i \rangle} ;$$

$$s = r^{(i-1)} - \alpha_i v^i ;$$

solve z from $Kz = s$;

$$t = Az ;$$

$$\omega_i = \frac{\langle K_1^{-1}t, K_1^{-1}s \rangle}{\langle K_1^{-1}t, K_1^{-1}t \rangle} ;$$

$$x^i = x^{(i-1)} + \alpha_i y + \omega_i z ;$$

if x^i is accurate enough then quit ;

$$r^i = s - \omega_i t ;$$

end

Note that this algorithm delivers the variables x^i and r^i corresponding to the original systems : $Ax = b$, i.e $r^i = b - Ax^i$.

Van der Vorst [Vo92] made an observation that in a preconditioned CGS schemes, preconditioning corresponds only to a different choice for initial

residual $r^{(0)}$. This observation does not carry on to the BiCGSTAB, he mentioned that this is attributed to different expression for ω_i . We might also compute another value for ω_i :

$$\omega_i = \frac{\langle t, s \rangle}{\langle t, t \rangle}$$

If we compute ω_i in this form, we are minimizing the current residual of original linear system rather than the preconditioned one. Preconditioned BiCGSTAB that used this form of ω_i is named as BiCGSTAB-P.

4.7 Stopping Criteria for Conjugate Gradient Type Iteration

Iteration process to obtain solution must be terminated at some stages. One criterion to terminates the iteration, under assumption that early breakdown has not occurred, is checking whether the current residual is small enough compared to the initial residual. This criterion is popular and is used in current practice. Arioli e.a [ADR92] investigates three stopping criteria for iterative schemes, one of which is the previously mentioned criterion. For the BiCGSTAB schemes, van der Vorst[Vo92] gave another alternative for regular termination criterion. He gave the norm of current s iterates as a measure of closeness of approximate solution to the true solution. He points out that the current s represent residual after BiCG steps, hence it is a reasonable measure to terminate the current iteration.

Chapter 5

Diagonal Scaling as Preconditioner for Preconditioned BiCGSTAB Methods to Solve Linear Equations from Saddle-Point Problems

In this chapter we will discuss the construction of a diagonal preconditioner for linear equation :

$$\hat{C}x = b$$

where:

$$\hat{C} = \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix}$$

is an $(n + m) \times (n + m)$ non-singular symmetric indefinite matrix. This form of linear equations arise from Linear Programming as the Karush-Kuhn-Tucker necessary condition and augmented least squares problem [ADR89] also from finite element discretization of linear elliptic equation using mixed finite element formulation as discussed in Chapter 3. This problem is well known as the saddle-point problem.

We may solved saddle-point problem in an equivalent form by permuting

the block such that the matrix will be of the form :

$$C = \begin{pmatrix} A^T & 0 \\ M & A \end{pmatrix}$$

So the preconditioned BiCGSTAB and its variant can be used without modification.

This chapter is arranged as follows. The first section concerns a particular type of diagonal preconditioner in relation with the condition number of C . In the second section, we discuss the distribution of eigenvalues of C and its relationship with extreme singular-values of M and A . The result of the second section will be used to estimate the rate of convergence of BiCGSTAB algorithm when used to solve the saddle-point problem, this is discussed in the third section. The result of this section will be used in the fourth section to build a preconditioner to accelerate the iterative method. Finally, the last section gave the range of diagonal preconditioner for saddle-point problem based on the discussion previously.

5.1 Conditioning by Scaling of Saddle-point Problem

In this section, we consider conditioning of a specific form of saddle-point problem, where M is identity matrix.

Let :

$$C = \begin{pmatrix} I_n & A \\ A^T & 0 \end{pmatrix}$$

be a non-singular matrix, where I_n is $n \times n$ identity and A is $n \times m$ matrix with $m \leq n$ and $\text{rank}(A) = m$. This form of matrix arises from the augmented least squares problem [ADR89]. The inverse of C can be expressed in terms of generalized inverse of A :

$$A^+ = (A^T A)^{-1} A^T$$

and the orthogonal projection matrix on the null space of A^T ,

$$P = I_n - AA^+.$$

Then we can write:

$$C^{(-1)} = \begin{pmatrix} P & (A^+)^T \\ A^+ & -(A^T A)^{-1} \end{pmatrix}$$

The spectral condition number of C is :

$$\kappa(C) = \|C\|_2 \|C^{-1}\|_2$$

is strictly related to σ_1 and σ_m , maximum and minimum singular-values of A respectively. Bjorck [Bj67] showed this relationship is given by:

$$\kappa(C) = \frac{\frac{1}{2}(1 + \sqrt{1 + 4\sigma_1^2})}{\min(1, \frac{1}{2}(\sqrt{1 + 4\sigma_m^2} - 1))}$$

So that $\kappa(C)$ is much larger than the condition number of A, $\kappa(A) = \frac{\sigma_1}{\sigma_m}$ when σ_1 is much less than 1. In order to decrease the condition number of C, Bjork[Bj67] propose scaling as follows:

$$\begin{pmatrix} I_n & 0 \\ 0 & \alpha^{-1} I_m \end{pmatrix} \begin{pmatrix} I_n & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \alpha I_n & 0 \\ 0 & I_m \end{pmatrix} = \begin{pmatrix} \alpha I_n & A \\ A^T & 0 \end{pmatrix}$$

where α is a real positive number. The condition number of :

$$C_\alpha = \begin{pmatrix} \alpha I_n & A \\ A^T & 0 \end{pmatrix}$$

$\kappa(C_\alpha)$ is given by:

$$\kappa(C_\alpha) = \frac{\frac{1}{2}(\alpha + \sqrt{(\alpha^2 + 4\sigma_1^2)})}{\min(\alpha, \frac{1}{2}(\sqrt{(\alpha^2 + 4\sigma_m^2)} - \alpha))}$$

and attains its minimum of $\frac{1}{2}(1 + \sqrt{1 + 8\kappa(A)^2})$ for $\alpha = \frac{\sigma_m}{\sqrt{2}}$

Note that every original saddle-point problems can be transformed into the augmented least squares problem. Since M is symmetric positive definite, then M has Cholesky factorization : $M = LL^T$. Hence :

$$\begin{pmatrix} L^{-1} & 0 \\ 0 & \alpha^{-1} I_m \end{pmatrix} \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \alpha L^{-T} & 0 \\ 0 & I_m \end{pmatrix} = \begin{pmatrix} \alpha I_n & L^{-1} A \\ (L^{-1} A)^T & 0 \end{pmatrix}$$

5.2 Eigenvalue Distribution of Sadddle-point Problem

Since the original linear system arising from saddle point problems is symmetric indefinite and non-singular, then the spectrum contains both positive and negative eigenvalues. In this section we give a result concerning the spectrum of C and its relation to properties of the sub-matrices M and A . This result will be used to analyze the construction of diagonal preconditioner of linear system.

Let M have eigenvalues $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n > 0$ Then :

$$\mu_n \|x\|^2 \leq \langle Mx, x \rangle \leq \mu_1 \|x\|^2, \text{ for } x \in R^n$$

Let A have singular-values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$

$$\sigma_m \|y\| \leq \|Ay\| \leq \sigma_1 \|y\|, \text{ for } y \in R^m$$

and

$$\sigma_m \|x\| \leq \|A^T x\| \leq \sigma_1 \|x\|, \text{ for } x \in N(A^T)^\perp$$

The following lemma gives a connection between the spectrum of C and eigenvalues of M and singular-values of A .

Lemma[RW92]

Let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n > 0$ be eigenvalues of M and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$ be the singular-values of A , and let Λ denote the spectrum of C . Then:

$$\Lambda \subset I = I^+ \cup I^-$$

where

$$I^- = [\frac{1}{2}(\mu_n - \sqrt{(\mu_n^2 + 4\sigma_1^2)}), \frac{1}{2}(\mu_1 - \sqrt{(\mu_1^2 + 4\sigma_m^2)})],$$

and

$$I^+ = [\mu_n, \frac{1}{2}(\mu_1 + \sqrt{(\mu_1^2 + 4\sigma_1^2)})]$$

Proof

Let $\lambda \in \Lambda$ and let (x, y) be the corresponding eigenvector, i.e:

$$A^T x = \lambda y$$

$$Mx + Ay = \lambda x$$

Since $\text{rank}(A) = m$, if $x = 0$ then $y = 0$. So it is impossible that this pair is an eigenvector of C . So $x \neq 0$. Now we divide the proof into two parts. First we compute the bound for the positive eigenvalue, then the negative eigenvalue.

Let λ be a positive eigenvalue. By taking an inner product of the second equation with x and combine it with an inner product of the first equation with y we obtain:

$$\langle Mx, x \rangle + \lambda \|y\|^2 = \lambda \|x\|^2$$

applying bounds of eigenvalue of M we get :

$$(\mu_n - \lambda) \|x\|^2 \leq -\lambda \|y\|^2 \leq 0$$

Since $\|x\|$ is positive, we have $\lambda \geq \mu_n$. To get an upper bound, we use the first equation and inner product of the second equation with x to get:

$$\langle Mx, x \rangle + \frac{1}{\lambda} \|B^T x\|^2 = \lambda \|x\|^2$$

By the bound of inner product (i) and (iii) we get an implication of an inequality :

$$(\lambda^2 - \mu_1 \lambda - \sigma_1^2) \|x\|^2 \leq 0$$

Since $\|x\|$ is positive, we can conclude that : $\lambda \leq \frac{1}{2}(\mu_1 + \sqrt{(\mu_1^2 + 4\sigma_1^2)})$. So we have derived the bounds of positive eigenvalues.

Next, we consider λ be negative eigenvalue of C . The lower bound of I^- can be derived following the proof for upper bound of I^+ . Hence it is only suffice to derive the upperbound for the negative eigenvalue. To Obtain this, Let $x = v + w$, where $v \in N(A^T)^\perp$ and $w \in N(A^T)$ By the first equation and th inner product of the second equation by v we get :

$$\langle Mw, v \rangle = \lambda \|x\|^2 - \langle Mv, v \rangle - \frac{1}{\lambda} \|B^T v\|^2$$

Applying the bound of eigenvalues of M and singular-values of A we obtain:

$$\langle Mw, v \rangle \geq (\lambda - \mu_1 - \frac{\sigma_m^2}{\lambda}) \|v\|^2$$

To proceed we must bound $\langle Mw, v \rangle$. This is obtained by calculating inner product of (ii) with w and using eigenvalues bound of M . Since $w \in N(B^T)$ and $\lambda - \mu_n$ is negative we get :

$$\langle Mw, w \rangle \leq (\lambda - \mu_n)\|w\| \leq 0$$

By the symmetry of M and combines it with the previous inequality, we obtain:

$$0 \leq (\lambda^2 - \mu_1\lambda - \sigma_m^2)\|v\|$$

If $v = 0$, then $y = 0$ and (ii) reduce to $Mw = \lambda w$. But this contradict to the fact that λ is negative, hence it follows that $\lambda^2 - \mu_1\lambda - \sigma^2 \geq 0$. Hence :

$$\lambda \leq \frac{1}{2}(\mu_1 - \sqrt{(\mu_1^2 + 4\sigma_1^2)})$$

. q.e.d.

5.3 Rate Convergence Estimate of BiCGSTAB for Saddle-point Problem

Assume that $I = I^- \cup I^+$ as stated in previous Lemma. From the choice of parameter ω in BiCGSTAB methods, we can derive a convergence estimate of BiCGSTAB for saddle-point problem. Recall that we choose ω such that $r^{(k)} = b - Cx^{(k)}$ is minimum in 2-norms, then :

$$\begin{aligned} \|r^{(k)}\|^2 &= \|b - Cx^{(k)}\|^2 \\ &\leq \|b - Cu\|^2, \text{ for } u \in K(r^{(0)}, C, k) \end{aligned}$$

Using spectral theory [RW92] we obtain :

$$\begin{aligned} \|r^{(k)}\| &\leq \delta_k(\Lambda(C))\|b\| \\ &\leq \delta_k(I)\|b\| \end{aligned}$$

where :

$$\delta_k(\Lambda(C)) = \inf_{p \in \Pi_k^1} \sup_{\lambda \in \Lambda} |p(\lambda)|$$

Observe that the convergence can be estimated from the quantity $\delta_k(\Lambda(C))$. Unfortunately, it is hard to obtain the optimal polynomial to bound the above estimates [RW92], unless the eigenvalues of C are positive. Which in this case the bounds is similar to the bound of Conjugate Gradient Method. However, some properties can be derived. Observe that if the polynomials in Π_k^1 is scaled properly we obtain that $\delta_k(\alpha I) \leq \delta_k(I)$ for any $\alpha \neq 0$. also if $\hat{I} \subset I$, then obviously $\delta_k(\hat{I}) \leq \delta_k(I)$. So The BiCGSTAB convergence rates can be reduced if we can shrink the intervals I .

By scaling the interval I as indicated previously such that the lower bound of I^+ is unity, we can assume that I^+ and I^- are given by:

$$I^- = [\frac{1}{2}(1 - \sqrt{(1 + 4\rho^2\kappa(A)^2)}), \frac{1}{2}(\kappa(M)^2 - \sqrt{(\kappa(M)^2 + 4\rho^2)})],$$

and

$$I^+ = [1, \frac{1}{2}(\kappa(M)^2 + \sqrt{(\kappa(M)^2 + 4\rho^2\kappa(A)^2)})],$$

where $\rho = \rho(A, M)$ denotes relative scaling between the smallest singular values of A and singular values of M :

$$\rho(A, M) = \frac{\sigma_m}{\mu_n}$$

The expression I^- and I^+ as given above indicates that convergence rates of BiCGSTAB can be estimated by three parameters $\kappa(A)$, $\kappa(M)$, and $\rho(A, M)$. Note that the upper limit of I^+ will increase as function of ρ when $\kappa(A)$ and $\kappa(M)$ are fixed. On the other hand, the upper limit of I^- will approach zero as $\rho \rightarrow 0$. Since the polynomials in Π_k^1 must satisfy $p(0) = 1$, obviously this is not desirable in order to make $\delta_k(I)$ small. Therefore the scaling factor $\rho(A, M)$ should be neither too small nor too large, i.e. the matrices A and M should be properly balanced.

However, if the scaling factor $\rho(A, M)$ is fixed, it is clear that a reduction one of $\kappa(A)$ or $\kappa(M)$ will lead to shrinking of the interval I^- and I^+ given above. Therefore, if the matrices A and M are well conditioned and properly balanced, the BiCGSTAB applied to saddle point problems will converge fast.

5.4 Preconditioning of Saddle-Point Problem

From the previous section we know that the rate of convergence of BiCGSTAB for saddle point problems in some sense is governed by three parameters: $\kappa(A)$, $\kappa(M)$, and $\rho(A, M)$. There is no clear connection between these parameters with $\kappa(C)$ the condition number of C .

As previously discussed, the rate of convergence of BiCGSTAB for saddle point problem can be accelerated if we can shrink the interval I^+ and I^- . This shrinking of intervals can be achieved by reducing the condition number of A or M while kept the relative scaling $\rho(A, M)$ fixed. Therefore the preconditioned systems will be designed such that the conditioning of at least one of the matrices is reduced, while the relative scaling between them is kept approximately fixed.

We will use a block diagonal preconditioner for saddle point problems. This type of preconditioner has been proposed by Fortin [Fo89] and explored further by Rusten & Winther [RW92]. The advantages of block diagonal preconditioner is that the preconditioned systems have the same structures with the original linear systems with the property that convergence rate of the iterative method is significantly improved.

Let $L \in R^{n \times n}$ and $R \in R^{m \times m}$ be non-singular matrices, and let $K_1 = \text{diag}(R^T, L)$ and $K_2 = \text{diag}(L^T, R)$. Then the matrix $K_1^{-1}CK_2^{-1}$ is given by:

$$K_1^{-1}CK_2^{-1} = \begin{pmatrix} (L^{-1}AR^{-1})^T & 0 \\ (L^{-1}ML^{-1}) & L^{-1}AR^{-1} \end{pmatrix}$$

Hence the block structure of original linear systems still preserved after preconditioning.

If L and R are such that :

$$\kappa(L^{-1}ML^{-1}) \ll \kappa(M)$$

and/or

$$\kappa(L^{-1}AR^{-1}) \ll \kappa(A)$$

while

$$\rho(L^{-1}AR^{-1}, L^{-1}ML^{-1}) \approx \rho(A, M)$$

then according to discussion in previous section, the preconditioned BiCGSTAB will usually converge faster.

The choice of preconditioning matrices L and R is obviously problem dependent. Since M is symmetric positive definite and the preconditioning of such systems is extensively studied [Gu84], we often know how to find a matrix L with required property that $\kappa(L^{-1}ML^{-1})$ is sufficiently small. In such cases the remaining problem is therefore partly to find a matrix R such that $(L^{-1}AR^{-1})$ is well conditioned, i.e R should be chosen such that $(L^{-1}AR^{-1})$ is approximately a unitary matrix. Observe that this is equivalent to the requirement:

$$A^T(LL^T)A \approx R^T R$$

Hence R should be chosen such that $R^T R$ is a good preconditioner for the positive definite matrix $A^T(LL^T)A$.

5.5 Choice of Diagonal Preconditioner

In this section we proposed three block preconditioner for saddle-point problem as a result of discussion of previous sections.

The first preconditioner M_1 is constructed by taking $L = \alpha \text{diag}(m_{11}^{\frac{1}{2}}, \dots, m_{nn}^{\frac{1}{2}})$, where m_{ii} is diagonal entry of mass matrix and α is scaling parameter to conditioned C . We choose $R = I$. Since the mass matrix is generated using lumped mass integration, we obtain M is diagonal, so L is the Cholesky factor of M for $\alpha = 0$. The best choice of α have to be approximately close to the smallest singular values of $(L^{-1}A)$. Using heuristic arguments, Arioli et. al. [ADR89] gave estimates $\alpha = \frac{\max |l_{ik}^{-1} a_{kj}|}{1000}$.

The second preconditioner M_2 is constructed by taking $L = I_n$ and $R = \text{diag}((a_{1j}a_{j1})^{\frac{1}{2}}, \dots, (a_{nj}a_{jn})^{\frac{1}{2}})$, This preconditioner is to approximate $A^T A$ by its diagonal.

The third preconditioner M_3 is constructed by taking $L = I_n$ and choose R a diagonal matrix from $A^T A$ that is given by Schwartz[Sc85] in the previous chapter.

Chapter 6

Numerical Results of Iterative Methods to Solve the Biharmonic Equation

This chapter reports the result of solving the linear equation :

$$\hat{C}x = b$$

where :

$$\hat{C} = \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix}$$

is an $(n + m) \times (n + m)$ non-singular indefinite matrix derived from the discretization of Biharmonic equation using the mixed finite element formulation as posed in Chapter 3. In the numerical implementation it suffices to use a finite element program for solving second order problem. We shall use here linear finite element.

The CGS, BiCGSTAB algorithm and preconditioned variant of BiCGSTAB is employed to solve an equivalent form of saddle-point problem by permuting the block such that the matrix will be of the form :

$$C = \begin{pmatrix} A^T & 0 \\ M & A \end{pmatrix}$$

Numerical experiment were implemented using FORTRAN-77 in double precision on an Apollo Domain DN10000 workstation.

This chapter arranged as follows. In the first section we give a short review of the method of solution for solving saddle-point problem. In the second section we present numerical results on the first Biharmonic problem, while the numerical results of the second Biharmonic problem is in the third section. The discussion of the results is postponed till the last section in which an indication of possible future work for preconditioning strategy for saddle-point problems is discussed.

6.1 Review

In this section we give a short survey of methods for solving the linear systems resulting from mixed finite element formulation, i.e a symmetric non-singular matrix of the form :

$$\hat{C} = \begin{pmatrix} M & A \\ A^T & 0 \end{pmatrix}$$

In the literature, largely the discussion is centered on the results for second order problems. Nevertheless, since the resulting linear system is a similar form for the Biharmonic problems, the discussion is still true in this case.

Since the submatrix M is symmetric non-singular, we have that $AM^{-1}A^T$ is a symmetric positive definite matrix. So the conjugate gradient method can be used to solve the problem. Unfortunately, $AM^{-1}A^T$ may be a dense matrix. For the second order problem, this difficulties can be circumvented by taking a sparse equivalent matrix that approximate $AM^{-1}A^T$ [JT77]. This strategy also taken by Hanisch [Ha92] to solve first Biharmonic problems in multi-grid context.

Another trick such that the conjugate gradient method can be used is by augmenting the matrix $\hat{C}$ into the form:

$$\begin{pmatrix} M & A \\ A^T & -\epsilon I \end{pmatrix}$$

and choose ϵ such that the augmented matrix is symmetric positive definite.

One popular method to solve $\hat{C}$ is a method that evolved from mathematical programming, which is well known as Uzawa's algorithm. Bourgat [Bo76] studies this method to solve the first Biharmonic problem. Basically, the Uzawa's algorithm consist of two iteration steps, an inner and outer iterations. For every outer iteration a sub-problem is solved using inner iteration.

This method require one parameteer that have to determined apriorily, which is depend on the problem. Hong-Ci [Ho83] modified the Uzawa's method such that this parameter determined adaptively for every outer iteration. He investigates this method for first Biharmonic problem for its rate convergence, but without numerical evidence.

Another approach is by solving iteratively $\hat{C}$ using some modification of conjugate gradient method. This strategy is taken by Rusten and Winther [RW92] to solve second order problem using conjugate residual method, a conjugate gradient like method for symmetric non-singular linear system.

6.2 Result on the First Biharmonic equation

Numerical Integration

In the numerical implementation, we use lumped mass integration for calculation the mass matrix term :[KN90]

$$\int \psi_i \phi_j$$

while for calculation the term :

$$\int \nabla \psi_i \nabla \phi_j$$

we use an exact calculation of basis element over a standard triangle [Sc88].

Triangulation

The triangulation chosen for the unit square are generated by four families of parallel lines by construction of uniform square grid by vertical and horizontal families of parallel lines. Then the triangulation are obtained by adding two families parallel lines in north-east and north-west direction to the square uniform grid obtained previously. In Figure 6.1 shown a particular triangulation generated in the way prescribed previously.

The information about triangulation we use for numerical experiment are given in Table 6.1.

Numerical Test

The CGS, BiCGSTAB and its preconditioned algorithm is initialized with $u_0 = 1$ and $p_0 = 1$, i.e a column vector with entries 1. The detection for near-breakdown is inserted with threshold detection set at 10^{-5} . The algorithm

Figure 6.1: Triangulation T_{h1} of a Square

Triangulation	Triangles	internal	boundary	total
T_{h1}	16	5	8	13
T_{h2}	64	25	16	41
T_{h3}	288	121	32	153
T_{h4}	1024	481	64	545
T_{h5}	4096	1985	128	2113

Table 6.1: Triangulation

is stopped when stopping criterion for residual reaches the bound:

$$\delta \leq 2.0 * 10^{-7}$$

Also the algorithm will be stopped if iteration reaches a maximum iteration bound which is set to the size of $\tilde{C}$, i.e twice of the total number of nodes.

We choose for a numerical test a function on the unit square whose analytical solution were known.

$$\begin{aligned} u(x, y) &= (2^5)(x(1-x) \times y(1-y))^2 \\ \Delta u(x, y) &= (2^5) \times 2 \times (\\ &\quad (y(1-y))^2 \times (-2 \times x(1-x) + (1-2 \times x)^2) + \\ &\quad (x(1-x))^2 \times (-2 \times y(1-y) + (1-2 \times y)^2)) \\ \Delta \Delta u(x, y) &= -2^5 \times 8 \times (3 \times (y(1-y))^2 + 3 \times (x(1-x))^2 + \\ &\quad (1-6x+6x^2) + (1-6y+6y^2)) \end{aligned}$$

The bound of error of Scholz's for linear element in § 3.3 were computed as a constant :

$$c = \frac{\|u - u_h\|_0 + h^{\frac{1}{2}} |\ln h| \|p - p_h\|_0}{h |\ln h|^2}$$

In Table 6.2 , we show the comparison of corresponding iteration of BiCGSTAB and CGS, and their corresponding CPU-time for each case. In the case of CGS for three discretization, near breakdown was detected and the iteration was intentionally aborted. Off course, we may restarted the process from the vector where the near breakdown occuring, but in practice [Jo92] this restarting process usually using other iteration method.

In Table 6.3, we have the error bounds of Scholz's. Also this table show the ratio of errors of u and p with respect to function of discretization parameter h to obtain insight into the order of convergence. This ratio is portrayed since this is the most consistent compared to other ratio that also calculated.

In Table 6.4, we show the results for preconditioned BiCGSTAB, using three types of preconditioner proposed in the last section of the previous chapter. The iteration for third preconditioner was terminated after reaching the maximum iterations bound. CPU run time for each preconditioner are given in Table 6.5.

Triangulation	BiCGSTAB	CPU-Time	CGS	CPU-Time
T_{h1}	11	1.66-2	9	1.66-2
T_{h2}	27	1.00-1	22	1.50-1
T_{h3}	80	1.31-0	*	*
T_{h4}	365	1.95+1	*	*
T_{h5}	1284	2.68+2	*	*

Table 6.2: Comparison of BiCGSTAB and CGS

Triangulation	Bounds	$\frac{\ u-u_h\ _0}{h lnh ^2}$	$\frac{\ p-p_h\ _0}{h^{1/2} lnh }$
T_{h1}	4.340+1	6.932	3.656+1
T_{h2}	3.620+1	2.792	3.341+1
T_{h3}	3.555+1	2.451	3.309+1
T_{h4}	3.914+1	2.870	3.627+1
T_{h5}	4.646+1	3.712	4.267+1

Table 6.3: Some errors indicator for the case of the first Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	11	10	11	+(26)
T_{h2}	27	25	22	+(82)
T_{h3}	80	77	75	+(290)
T_{h4}	365	252	243	+(1090)
T_{h5}	1284	982	563	+(4226)

Table 6.4: Comparison of preconditioned BiCGSTAB

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	1.66-2	3.33-2	1.67-2	5.00-2
T_{h2}	1.00-1	1.67-2	1.33-1	4.50-1
T_{h3}	1.31+0	1.33-0	1.07+0	4.88+0
T_{h4}	1.95+1	1.52+1	1.04+1	6.59+1
T_{h5}	2.68+2	2.25+2	1.29+2	9.75+2

Table 6.5: CPU run time for preconditioned BiCGSTAB

In Table 6.6, we show results of preconditioned BiCGSTABP for three proposed preconditioner. The same thing for preconditioner M_3 happened as for previous preconditioned BiCGSTAB. While its corresponding CPU run time is given in Table 6.7.

6.3 Result on the Second Biharmonic equation

The same numerical integration, triangulation, and stopping criterion for the iterative solver as described in the previous section is used in the numerical experiment for second Biharmonic problem. The solver was initialized using $u_0 = 0$ and $p_0 = 0$

We choose for numerical test two functions on the unit square whose analytical solution were known.

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	11	10	11	+(26)
T_{h2}	27	24	23	+(82)
T_{h3}	80	77	74	+(290)
T_{h4}	365	359	263	+(1090)
T_{h5}	1284	1003	556	+(4226)

Table 6.6: Comparison of preconditioned BiCGSTABP

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	1.66-2	1.67-2	3.33-2	1.00-1
T_{h2}	1.00-1	1.00-1	8.33-2	4.00-1
T_{h3}	1.31+0	1.08-0	1.07+0	4.78+0
T_{h4}	1.95+1	1.89+1	9.85+0	6.37+1
T_{h5}	2.68+2	2.24+2	1.22+2	9.46+2

Table 6.7: CPU run time for preconditioned BiCGSTABP

Triangulation	BiCGSTAB	CPU-Time	CGS	CPU-Time
T_{h1}	9	1.66-2	8	1.66-2
T_{h2}	18	1.00-1	15	1.00-1
T_{h3}	34	5.33-1	28	4.00-1
T_{h4}	59	2.87	51	2.48
T_{h5}	119	2.31+1	92	1.78+1

Table 6.8: Comparison of BiCGSTAB and CGS for first case of the second Biharmonic problem

First case :

$$u(x, y) = x(1 - x) \times \exp(y)$$

$$\Delta u(x, y) = (x(1 - x) - 2) \times \exp(y)$$

$$\Delta \Delta u(x, y) = (x(1 - x) - 4) \times \exp(y)$$

Second case :

$$u(x, y) = \frac{1}{2 \times (2\pi)^2} \times \sin(2\pi \times x) \times \sin(2\pi \times y)$$

$$\Delta u(x, y) = \sin(2\pi \times x) \times \sin(2\pi \times y)$$

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

In Table 6.8, we show the comparison of corresponding iteration of BiCGSTAB and CGS, and their corresponding CPU-time for the first case.

The bound of error of Scholz's for linear element were computed as defined for first Biharmonic problem. In Table 6.9, we have the error bounds of Scholz's. Also this table show the ratio of errors of u and p with respect to function of discretization parameter h to obtain insight into the order of convergence.

In Table 6.10, we show the results for preconditioned BiCGSTAB while its corresponding CPU run time is given in Table 6.11. We show the results for preconditioned BiCGSTABP and its corresponding CPU run time are given in Table 6.12 and Table 6.13 respectively.

While the results for test of case 2 are given in the following six tables. Every methods failed to solve the first triangulation due to early termination.

Triangulation	Bounds	$\frac{\ u-u_h\ _0}{h \ln h ^2}$	$\frac{\ p-p_h\ _0}{h^{1/2} \ln h }$
T_{h1}	3.335	7.257-1	2.609
T_{h2}	2.286	3.124-1	1.9731
T_{h3}	1.070	1.089-1	9.616-1
T_{h4}	1.124	1.343-1	9.901-1
T_{h5}	1.474	2.121-1	1.261

Table 6.9: Some errors indicator for first case of the second Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	9	9	8	11
T_{h2}	18	19	17	24
T_{h3}	34	36	31	44
T_{h4}	59	102	71	89
T_{h5}	119	250	117	165

Table 6.10: Comparison of preconditioned BiCGSTAB for first case of the second Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	1.66-2	3.33-2	1.67-2	1.67-2
T_{h2}	1.00-1	8.33-2	6.67-2	1.33-1
T_{h3}	5.33-1	5.17-1	4.50-1	7.00-1
T_{h4}	2.87+0	3.90+0	4.65+0	4.65+0
T_{h5}	5.39+1	3.28+1	2.44+1	3.53+1

Table 6.11: CPU run time for preconditioned BiCGSTAB

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	9	9	8	10
T_{h2}	18	18	17	24
T_{h3}	34	35	30	46
T_{h4}	59	68	71	78
T_{h5}	119	158	116	170

Table 6.12: Comparison of preconditioned BiCGSTABP for first case of the second Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	1.66-2	1.67-2	1.67-2	1.67-2
T_{h2}	1.00-1	8.33-2	6.66-2	8.33-2
T_{h3}	5.33-1	4.83-1	3.67-1	7.17-1
T_{h4}	2.87+0	3.40+0	3.68+0	4.05+0
T_{h5}	2.31+1	3.28+1	2.35+1	3.49+1

Table 6.13: CPU run time for preconditioned BiCGSTABP

Triangulation	BiCGSTAB	CPU-Time	CGS	CPU-Time
T_{h1}	*	*	*	*
T_{h2}	17	1.00-1	10	6.66-2
T_{h3}	36	4.83-1	23	3.17-1
T_{h4}	71	3.55+0	41	2.10+0
T_{h5}	94	1.82+1	77	1.49+1

Table 6.14: Comparison of BiCGSTAB and CGS for second case of the second Biharmonic problem

Triangulation	Bounds	$\frac{\ u-u_h\ _0}{h lnh ^2}$	$\frac{\ p-p_h\ _0}{h^{1/2} lnh }$
T_{h1}	*	*	*
T_{h2}	7.345-1	1.318-2	7.213-1
T_{h3}	6.918-1	1.171-2	6.801-1
T_{h4}	7.345-1	1.318-2	7.213-1
T_{h5}	8.329-1	1.687-2	8.213-1

Table 6.15: Some errors indicator for second case of the second Biharmonic problem

In Table 6.14, we show the comparison of corresponding iteration of BiCGSTAB and CGS, and their corresponding CPU-time for each case.

In Table 6.15, we have the error bounds of Scholz's. Also this table show the ratio of errors of u and p with respect to function of discretization parameter h to obtain insight into the order of convergence. This ratio is portrayed since this is the most consistent compared to other ratio that also calculated.

In Table 6.16, we show the results for preconditioned BiCGSTAB and its corresponding CPU run time is given in Table 6.17. While the results for preconditioned BiCGSTABP and its corresponding CPU run time are given in Table 6.18 and Table 6.19 respectively.

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	*	*	*	*
T_{h2}	17	18	18	23
T_{h3}	36	37	34	45
T_{h4}	71	72	66	69
T_{h5}	94	137	84	144

Table 6.16: Comparison of preconditioned BiCGSTAB for second case of the second Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	*	*	*	*
T_{h2}	1.00-1	6.67-2	6.67-1	1.17-1
T_{h3}	4.83-1	6.66-1	4.50-1	7.17-1
T_{h4}	3.55+0	3.97+0	3.62+0	3.75+0
T_{h5}	1.82+1	2.94+1	1.76+1	3.08+1

Table 6.17: CPU run time for preconditioned BiCGSTAB

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	*	*	*	*
T_{h2}	17	18	18	23
T_{h3}	36	34	36	44
T_{h4}	71	60	69	89
T_{h5}	94	172	92	141

Table 6.18: Comparison of preconditioned BiCGSTABP for second case of the second Biharmonic problem

Triangulation	no preconditioner	M1	M2	M3
T_{h1}	*	*	*	*
T_{h2}	1.00-1	6.67-2	1.00-1	1.50-1
T_{h3}	4.83-1	4.83-1	4.83-1	6.33-1
T_{h4}	3.55+0	3.15+0	3.65+0	4.70+0
T_{h5}	1.82+1	3.54+1	1.94+1	2.92+1

Table 6.19: CPU run time for preconditioned BiCGSTABP

6.4 Concluding Remarks

From the calculation of error estimates, the results show that estimates given by Scholz are optimal. Since we did not use a triangulation of type considered by Glowinski [Gl73], we did not get polynomial type of error estimates.

For the both Biharmonic problem, preconditioner M_2 is the most consistent preconditioner. This preconditioner was designed to maintain the ratio $\frac{\sigma_m}{\mu_n}$, which is the scaling factor $\rho(A, M)$, and at the same time try to reduce the interval containing the distributions of eigenvalues of $\tilde{C}$. This preconditioner is the best of the three choices studied.

The results on preconditioner M_1 show that the condition number is not the sole factor that determines the rate convergence for saddle-point problem. The performance of this preconditioner, basically shows that even though we obtain a better conditioned matrix, by destroying the scaling factor $\rho(A, M)$, we do not obtain a better convergence rate for a saddle-point problem. This is shown by the results of the second Biharmonic problem, which in both versions of BiCGSTAB have worse performance compare to BiCGSTAB. While this preconditioner improves the rate of convergence for the first Biharmonic problem, which is an ill-conditioned matrix, and shows that condition number is one of the crucial factors obtaining a better rate of convergence. Furthermore, from the performance of M_3 , which has worst performance in every cases, shows that the strategy for maintaining the scaling factor and reducing the interval, is the best way for block preconditioner of saddle-point problems. Since the design behind preconditioner M_3 is basically to approximate the diagonal factor of $A^T A$ and destroy the interval that contains eigenvalues of $\tilde{C}$.

There are two things at least that can be done for future work in area of iterative methods for saddle-point problem. First, observe that we have not made an attempt to take advantage of symmetry of $\tilde{C}$. Freund [Fr92] points out that CGS algorithm can not take advantage of symmetry, so for every iteration we always required two matrix-vector multiplication, and we can not reduce it into one matrix-vector multiplication. Since the BiCGSTAB algorithm is derived from CGS, the argument is still valid. So a contender for BiCGSTAB with takes advantage of symmetry is possibly the QMR algorithm of Freund [Fr92] which is derived from BCG, which is claimed has a better residual error reduction than BCG.

The second is concerned with the construction of a block preconditioner. Observe that the preconditioner M_2 contains a block that is built from a diagonal approximation of $A^T A$, so the best possible thing to do is to use QR factorization of A . Since A is a sparse matrix, then Givens rotations are recommended for this purpose.

Reference

- [AB85] O. Axelsson, & V.A. Barker, 1985, Finite Element Solution of Boundary Value Problem: Theory and Computation, Academic Press.
- [ADR89] M. Arioli, I. S. Duff & P. P. M. deRijk, 1984, On the Augmented System Approach to Sparse Least-square Problems, Numer. Math., **55**, pp 667-684
- [ADR92] M. Arioli, I. Duff, D. Ruiz., 1992, Stopping Criteria for Iterative Solvers, SIAM J. Mat. Anal. Applic., **13**, pp 138-145
- [BF91] F. Brezzi & M. Fortin., 1991, Mixed and Hybrid Finite Element Methods, Springer Verlag, Berlin
- [Bj67] A. Bjork, 1967, Iterative Refinement of Linear Least-Squares Solutions, BIT, **7**, pp 257-278
- [Bo76] J. F. Bourgat., 1976, Numerical Study of A Dual Iterative Method for Solving A Finite Element Approximation of the Biharmonic Equation, Comp. Meth. Appl. Mech. Eng., **9**, 203-218
- [BR77] F. Brezzi & P. Raviart., 1977, Mixed Finite Element Methods for Fourth Order Elliptic Equations, in J. J. H. Miller (ed), Topics in Numerical analysis III, Academic Press, 33-56
- [Br74] F. Brezzi., 1974, On the existence, Uniqueness and Approximation of Saddle-Point Problems Arising from Lagrangian Multipliers, RAIRO, **8**, 129-151
- [CR74] P. Ciarlet & P. Raviart., 1974, A Mixed Finite Element Methods for the Biharmonic Equation, in C. deBoor (ed), Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, 125-145
- [FGN79] G.J. Fix, M. D. Gunzburger, & A.R. Nicolaides., 1979, Theory and Applications of Mixed Finite Element Methods, in C. V. Coffman & G. J. Fix (ed), Constructive Approaches to Mathematical Models, Academic Press, 375-393
- [Fl76] R. Fletcher., 1976, Conjugate Gradient Methods for the Indefinite Systems, in Numerical Analysis Dundee 1975, G. A. Watson, ed., Lect. N. Math. 506, Springer Verlag, New York, pp 73-89
- [Fo89] M. Fortin, 1989, Some Iterative Methods for Incompressible Flow Problems, Comp. Phys. Comm., **53**, pp 393-399
- [Fr92] R. Freund., 1992, Conjugate Gradient-Type Methods for Linear Systems with Complex Symmetric Coefficient Matrices, SIAM J. Sci. Stat. Comp., **13**, 425-448

- [Gl73] R. Glowinski.,1973 , Approximations Externes, par Elements Fini de Lagrange d'ordre Un et Deux, du Probleme de Dirichlet pour l'Operateur Biharmonique, in J. J. H. Miller (ed), Topics in Numerical analysis I, Academic Press, 123-171
- [Gu84] I. Gustafsson,1984, A Preconditioned Iterative Methods for The Solution of Biharmonic Equation., IMA J. Numer. An., **4**, pp 55-67
- [Gu89] I. Gustafsson,1989, A Class of Preconditioned Conjugate Gradient Methods Applied to Finite Element Equation - A Survey on MIC methods, in O. Axelsson & Y. I. Kotilina, eds., Lect. Notes Math, **1457**, 55-67
- [Ha92] M. Hanisch., 1992, Multigrid Preconditioning for the Biharmonic Dirichlet Problems, Mathematical Sciences Institute, Cornell, Preprint.
- [Ho83] H. Hong-Ci., 1983, A Type of Iterative Method for Solving the Finite Element Approximation of Saddle-point Problem, in F. Kang & J. L. Lions (ed), Proc. of the China-French Symposium on Finite Element Method, Gordon-Breach, New-York, pp 306-322
- [Jo92] W. Joubert., 1992, Lanczos Methods for the NonSymmetric Systems of Linear Equations, SIAM J. Mat. Anal. Applic., **13**, pp 926-943
- [Jo90] C. Johnson.,1990 , Numerical Solution of Partial Differential Equations by Finite Element Methods , Cambridge University Press.
- [KN90] M. Krizek & P. Neittaanmaki.,1990 , Finite Element Application of Variational Problem and Applications., Longman
- [LL89] L. Landau & E. Lifshitz, 1989, Fluid Mechanics , Pergamon Press, London, 2nd edition.
- [LL86] L. Landau & E. Lifshitz, 1986, Theory of Elasticity, Pergamon Press, London, 3rd edition.
- [Mo87] P. Monk,1987 , A Mixed Finite Element Method for The Biharmonic Equation , SIAM J. Numer. Anal.,**24**, 737-749
- [Re69] K. Rektorys, 1969, Survey of Applicable Mathematics, Illife, London.
- [Re80] K. Rektorys, 1980, Variational Methods in Mathematics, Science, and Engineering, 2nd ed, Reidel Dodrecht, London.
- [RT77] J. E. Roberts & J.M. Thomas.,1977 , Mixed and Hybrid Methods, , in P. G. Ciarlet & J. L. Lions (ed), Handbook of Numerical analysis , Volume II, Part I., North Holland, 523-639
- [RW92] T. Rusten & R. Winther.,1992, The Preconditioned Iterative Methods for Saddle-Point Problems, SIAM J. Matrix Anal. Applic., **13**, pp 887-905

- [Sc88] H. R. Schwarz , 1988, Finite Element Methods, Academic Press.
- [Sc78] F. Scholz ,1978 , Mixed Methods for Fourth Order Problems using Linear Finite Elements, , RAIRO,v12, 85-90
- [STW90] A.H. Schatz, V. Thomee, and W.L. Wendland .,1990 , Mathematical Theory of Finite and Boundary Element Methods , Birkhauser
- [Sl69] A. van der Sluis ., 1969, Condition Numbers and Equilibration of Matrices, Num. Math.,14, pp 14-23
- [So89] P. Sonneveld , 1989, CGS: A Fast Lanczos-Type Solver for Non-symmetric Linear Systems, SIAM J. Sci. Stat. Comput., 10, pp 36-52
- [SV87] A. van der Sluis & H. van der Vorst., 1987, The Convergence Behavior of Ritz Values in the Presence of Close Eigenvalues Lin. Alg. Applic.,88, pp 651-694
- [SV86] A. van der Sluis & H. van der Vorst., 1986, The Rate of Convergence of Conjugate Gradients, Num. Math.,48, pp 543-560
- [Te86] D. Terzopoulos, 1986, Regularization of Inverse Visual Problems Involving Discontinuities, IEEE Trans. Pat. An. Mach. Intel., PAMI-8, pp 413-424
- [Te82] D. Terzopoulos, 1982, Multilevel Reconstruction of Visual Surfaces. Variational Principles and Finite Element Representation, Artificial-Intelligence Lab. Memo 671, MIT, Cambridge, Mass.
- [Vo92] H. van der Vorst., 1992, BiCGSTAB: a Fast and Smoothly Converging Variant of BiCG for the Solution of NonSymmetric Linear Equations, SIAM J. Sci. Stat. Comput., 13, pp 631-644