Bivariate Piecewise Polynomials on Curved Domains, with Applications to Fully Nonlinear PDE's

ABID SAEED

A Thesis submitted for the degree of Doctor of Philosophy
at the University of Strathclyde in the Faculty of Science.

Department of Mathematics and Statistics
University of Strathclyde
26 Richmond Street, Glasgow G1 1XH, United Kingdom

November 2012

Acknowledgement

First of all I am extremely thankful to the only almighty ALLAH for everything he blessed me with, in particular for the completion of this dissertation.

I would like to express my sincere gratitude to my supervisor Dr. Oleg Davydov for all his continuous help and encouragement, always willing to offer his guidance and useful advices.

I will always remember my parents and my wife for their unconditional love and support. Their prayers were and will always be a hidden support for me.

I am also grateful to Kohat University of Science and Technology, Pakistan for funding me.

Abstract

Using Bernstein-Bézier techniques we construct bivariate polynomial finite element spaces of arbitrary order for curved domains bounded by piecewise conics, which leads to an H^1 conforming isogeometric method to solve Dirichlet problems for second order elliptic partial differential equations. Numerical experiments for several test problems over curved domains show the robustness of the method.

We then construct H^2 conforming polynomial finite element spaces for curved domains by extending the H^1 construction. These spaces are used in Böhmer's method for solving fully nonlinear elliptic equations. Numerical results for several benchmark problems including the Monge-Ampère equation over curved domains confirm the theoretical error bounds given by Böhmer.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 What is the Finite Element Method?

The finite element method (FEM) is one of the most famous numerical techniques to approximate the solution of boundary value problems [12, 18], particularly, due to its ability to handle complicated geometries. There are different versions of the method, with the Ritz-Galerkin version (conforming method) as the simplest one. Let us briefly outline the method, for a full comprehensive treatment see *e.g.* [12, 18]. It will be helpful to take start with some particular partial differential equation (PDE). There is no harm in considering the most familiar Poisson's equation. Assume we want to solve the problem

$$-\Delta u = f \text{ in } \Omega \subset \mathbb{R}^n \tag{1.1.1}$$

$$u = 0 \text{ on } \partial\Omega, \tag{1.1.2}$$

where Δ is the Laplacian operator defined by $\Delta := \sum_{i=1}^{n} \frac{\partial^2}{\partial^2 x_i}$. The first step is conversion of the problem (1.1.1)-(1.1.2) to its equivalent variational form obtained by multiplying (1.1.1) by some smooth test function v, satisfying homogeneous boundary conditions, in conjunction with integrating over Ω. Then using the integration by parts formula ends with

$$a(u,v) \equiv \int_{\Omega} \nabla u^T \nabla v = \int_{\Omega} fv \equiv L(v), \tag{1.1.3}$$

where $\nabla u = \left(\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n} \right)^T$ is the gradient vector of u. For the equality (1.1.3) to make sense we need $u, v \in H_0^1(\Omega)$, where $H_0^1(\Omega)$ is the Sobolev space defined by

$$H_0^1(\Omega) := \left\{ u \in L^2(\Omega) \,:\, \frac{\partial u}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, n, \; u|_{\partial\Omega} = 0 \right\},$$

with

$$L^2(\Omega) := \left\{ u \,:\, \Omega \to \mathbb{R} \,;\, \int_{\Omega} u^2 < \infty \right\}.$$

$a(u,v)$ is usually referred to as a *bilinear form* over a space $H_0^1(\Omega)$, that also defines an inner product on the same space $H_0^1(\Omega)$, and $L(v) := \int_{\Omega} fv \,:\, H_0^1(\Omega) \to \mathbb{R}$, as a linear form.

Thus the variational form corresponding to (1.1.1)-(1.1.2) can be formulated in an abstract form as follows:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = L(v) \; \forall \, v \in H_0^1(\Omega). \tag{1.1.4}$$

The Riesz representation theorem tells us that (1.1.4) has a unique solution [12, Theorem 2.5.6]. The Ritz-Galerkin method, actually, is to approximate u in a finite dimensional subspace S^h of $H_0^1(\Omega)$ *i.e.* (1.1.4) is reformulated as

$$\text{Find } u^h \in S^h \text{ such that } a(u^h, v) = L(v) \; \forall \, v \in S^h. \tag{1.1.5}$$

The finite dimensional space S^h is usually called a solution/approximating/discretizing space for the Galerkin method. Using the basis $\{\phi_1, \ldots, \phi_m\}$ of S^h, the problem (1.1.5) is transformed to a finite dimensional linear algebra problem of the form

$$KU = F,$$

where

$$K_{ij} := \int_\Omega \nabla\phi_i^T \nabla\phi_j \text{ and } F_i := \int_\Omega f\phi_i, i, j = 1, \ldots, m,$$

and $U = [u_1, \ldots, u_m]$ is the vector of unknowns such that

$$u^h := \sum_{i=1}^m u_i\phi_i$$

is an approximate solution to problem (1.1.1)-(1.1.2). The matrix K is, usually, called the *stiffness matrix* and F the *load vector*. Thus choice of S^h plays a crucial role in getting the corresponding system of linear equations that could be solved computationally efficiently. The requirements a space S^h must have can be met if S^h is chosen to be a finite dimensional space of *piecewise polynomials*(*splines*). Definition of piecewise polynomial requires definition of partition of the physical domain Ω which is defined as follows:

Definition 1.1.1. *Let $\triangle := \{T_1, \ldots, T_k\}$ be a set of simplices obtained by subdividing the polygonal domain Ω into subdomains. Then $\triangle$ is called a triangulation of Ω if the intersection of any two neighboring simplices T_i and T_j is a common face, a common vertex or a common edge.*

Also a subdomain T_i is, usually, called an *element* for all i.

Now S^h is said to be a space of piecewise polynomials if it is a space of functions that are polynomials on each triangle in $\triangle$. In modern techniques, usually, *local* basis functions for the space S^h are chosen *i.e.* the basis functions that are non-zero over a small subdomain of Ω. In fact the local basis leads to sparse stiffness matrix helps us solve the resultant system $KU = F$ efficiently. In view of (1.1.3) the space S^h will be a subspace of $V = H_0^1(\Omega)$ if we define it as follows:

$$S^h(\triangle) := \left\{s \in C^0(\Omega) \,:\, s|_{T_i} \in P_d \;\forall i,\; s|_{\partial\Omega} = 0\right\},$$

where P_d is a space of n-variate polynomials of degree at most d. Thus, in other words, we can say that S^h will be a subspace of $V = H_0^1(\Omega)$ if it satisfies two conditions

a) each $s \in S^h$ is continuous over Ω and

b) each $s \in S^h$ vanishes over $\partial\Omega$.

In case Ω is polygonal both of the conditions can be met easily but for Ω with curved boundaries the 2nd condition might be violated [12, 18], depending on what approach is used to construct S^h (see Section 5.2 for the triangulation of curved domains). In the finite element community this violation of 2nd condition is, usually, termed as a *variational crime*. In this case S^h does not remain a subspace of $V = H_0^1(\Omega)$ any more.

1.2 FEM for Domains with Curved Boundaries

In this section we briefly draw a sketch of some techniques that are used to deal with domains having curved boundaries.

The standard and comparatively old technique to deal with curved boundaries is that of isoparametric finite element method [18, 12] . In this approach nonlinear isoparametric transformations $\mathcal{F}_K$ are used that transform a reference triangle $\hat{K}$, with straight sides, onto the triangle K with curved sides (a *pie shaped triangle*), see Figure 1.1. The mappings are defined with the help of polynomials of the same degree as that of the elements over interior elements with straight sides. Obviously theory imposes some conditions on these mappings to satisfy including bijection and invertibility with non-zero Jacobians on curved element. Thus determining such mapping is not very straightforward. Non-zero Jacobian is in fact needed for the integration theory that comes into play while computing entries for element matrices. In case of order d elements the data of $\binom{d+2}{2}$ distinct nodes on the curved element uniquely determine these isoparametric mappings. [18, Theorem

4.3.3] provides sufficient conditions for the nodes to get a mapping with required properties. A subtlety arises while implementing isoparametric method and placing the nodes interior to curved elements. R. Scott [48] presented an algorithm, applicable in 2-D, to compute these nodes on curved elements that satisfy the conditions of [18, Theorem 4.3.3]. The algorithm was later extended to 3-D in [40]. This procedure needs to be followed for each curved element in a mesh. For using higher order elements this obviously gets troublesome and makes the method less attractive. Also it is well known that isoparametric method is not H^1 conforming because it exhibits a variational crime since the non-polynomial shape functions, obtained using nonlinear mappings, satisfy the boundary conditions on approximated curved boundary. Moreover, it is also important to mention that isoparametric approach is problematic when finite element spaces with enhanced smoothness are sought [12, Section 4.7]. Such spaces are used to solve higher order boundary value problems or second order fully nonlinear problems. For example the spaces $S^h \subset C^1(\Omega)$ are required for conforming methods to solve biharmonic equations or second order fully nonlinear equations using Böhmer's method [11]. Bernadou developed C^1 curved element spaces of degree 5 using isoparametric approach [8]. If isoparametric mapping $\mathcal{F}_K$ is defined with the help of polynomials of degree 5 then this construction needs the shape functions of degree 9 on curved elements to ensure global C^1 smoothness, which makes this approach too intricate and hardly practical.

Recently a new approach called the *isogeometric method* is introduced to solve boundary value problems over curved domains [37]. The method has emerged by the combination of finite element and Computer Aided Geometric Design (CAGD) techniques. One of the primary goals of this approach is to be geometrically exact while performing finite element simulations. Thus method is more geometric based

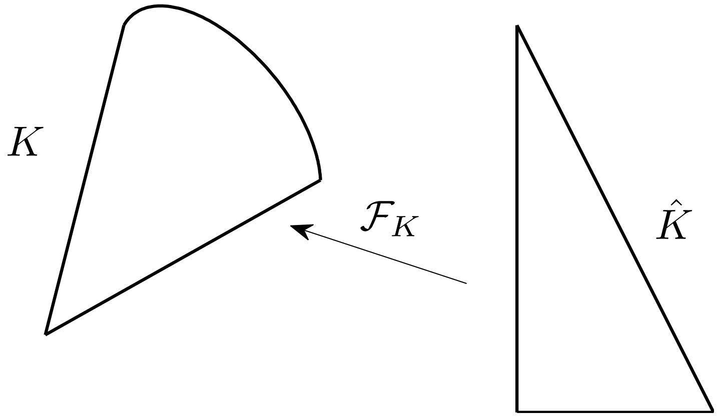

Figure 1.1: Isoparametric mapping $\mathcal{F}_K$.

where the exact geometry of the physical domain is retainned by the appropriate selection of the solution space S^h in conjunction with a geometry mapping that maps the parametric domain onto a physical domain Ω defined with the help of basis functions of the space S^h. Non-Uniform Rational B-Splines (NURBS) are one of the choices to be used as basis for the space S^h that help keep the exact geometry of the domain.

Another newly developed method is the so-called weighted extended B-spline method (or *web-spline method*) [35] . The main idea for web-spline method is the use of weighted B-splines as basis for finite element space where the globally defined weight function vanishes on the boundary which helps impose the homogeneous Dirichlet boundary conditions. But determining such a global weight function is, obviously, not so easy, and there seems to be no natural extension of this approach to deal with the non-homogeneous Dirichlet boundary conditions.

1.3 Piecewise Polynomial Finite Elements on Domains with Piecewise Conic Boundaries

Keeping in mind the difficulties to deal with domains having curved boundaries discussed above, we develop a comparatively simple finite element method that possess the following features.

1. It uses polynomial shape functions on pie shaped triangle.
2. It does not require any kind of geometry/isoparametric mapping.
3. It is isogeometric in the sense that the physical domain is exactly reproduced if it is defined by a piecewise conic curve. Note that NURBS allow an exact representation of conic curves [37, Section 2.9].
4. It is easily extendible to finite element spaces of higher smoothness.

More specifically we develop a finite element method for which we construct spaces that satisfy boundary conditions on exact domains bounded by piecewise conics. Note that a conic is an implicit quadratic polynomial curve which also has a parametric representation as rational quadratic curve, and hence it is a particular case of NURBS curve. To achieve our goal we make use of Bezout's theorem. The theorem says that if an algebraic curve $f = 0$ has infinitely many points of intersection with a given algebraic curve $q = 0$ then f and q have a common factor [34]. Thus if a conic $q = 0$ represents a curved boundary edge of a pie shaped triangle T then we use shape functions in the space $P_{d-1}q$, where P_{d-1} is a space of all polynomials of degree at most $d-1$ and

$$P_{d-1}q = \{pq \, : \, p \in P_{d-1}\} .$$

Then a boundary edge of T is a zero curve for these shape functions. In this context our method is in fact isogeometric method. Note that in contrast to the standard isogeometric approach we do not require any kind of geometry mapping. Thus there is no hurdle that prevents us from using higher order elements if needed.

In the context of the web-spline method we can say that we also use a kind of weighted shape function on curved elements where the weight function is a conic piece of boundary, *i.e.* q, is already known. An advantage of our approach over the web-spline method is that our weight functions are local, already known and are used only on curved elements (not on interior elements with all straight edges).

Using a kind of direct method we need to integrate on the original physical domain. To this end we develop a quadrature rule to approximate integrals on a pie shaped triangle.

The most attractive feature and in fact a motivation for developing our method is its relatively straightforward extension to smooth spaces. In particular we also construct H^2 polynomial finite element spaces by a natural extension of the H^1 construction. These space are required for Böhmer's method to solve fully nonlinear equations on a curved domain bounded by piecewise conics [10].

Before considering curved domains we study C^1 spaces on polygonal domains that possess the properties required in Böhmer's method for fully nonlinear equations.

Bernstein-Bézier techniques [43] are our main tools to construct the required finite element spaces. We have implemented, in MATLAB, a modified Argyris finite element space on polygonal and curved domains and C^0 finite elements of any degree on curved domains. This implementation has been used in all numerical experiments presented in the thesis.

1.4 Outline of the Thesis

We are mainly concerned with the construction of H^1 and H^2 conforming finite element spaces for curved domains bounded by piecewise conics to be used to solve second order elliptic boundary value problems (BVPs).

In Chapter 2 we introduce fully nonlinear equations and formulate several related definitions and results. Later, in the chapter, we also formulate Böhmer's finite element method.

In Chapter 3 the first section is devoted to a brief description of the relevant Bernstein-Bézier techniques that will be used extensively in the sequel. In the second part of the chapter we study C^1 spline spaces on polygonal domains for the possibility to be used in Böhmer's method for fully nonlinear problems.

In Chapter 4 we report the numerical results for Böhmer's method to solve several fully nonlinear benchmark problems on polygonal domains using a modified Argyris space as a discretizing space.

In Chapter 5 we develop a polynomial H^1 conforming finite element method for domains bounded by piecewise conics and present numerical results of our method by considering a few linear elliptic Dirichlet test problems including the eigenvalue problem.

In Chapter 6 we describe a construction of polynomial H^2 conforming finite element spaces for domains bounded by piecewise conics and present numerical results of the Böhmer finite element method applied to several test problems for the Monge-Ampère equation.

In Chapter 7 we compile all our achievements in this work and report some possible future directions.

Chapter 2

Fully Nonlinear PDEs and Böhmer's Method

The numerical solution of fully nonlinear elliptic partial differential equations is a topic of intensive research and great practical interest, see [11]. The motivation behind this interest is the presence of these equations in different fields of science and engineering including differential geometry [3], fluid mechanics [44] and optimal transportation [15].

Several numerical methods have been proposed in the literature for equations of *Monge-Ampère type* such as the *Monge-Ampère equation* $\det(\nabla^2 u) = f(> 0)$, the Gaussian curvature equation and Pucci's equation etc [4, 13, 14, 24, 25, 30, 31, 41, 45, 46, 47].

In [45] the authors proposed a discretization of the Monge-Ampère equation based on the geometric interpretation of the solution and solved the discretized version of the problem using an iterative method that yields a convergent sequence. Loeper [41] presented/proved a convergent Newton's algorithm to solve the Monge-Ampère equation with periodic boundary conditions, where the matrices involved in the algorithm are assembled using finite differences.

Adam Oberman [46, 47] developed finite difference methods of different orders for the numerical solution of fully nonlinear elliptic equations. The methods are proved to converge to the viscosity solution [16] of the problem. Obviously, being finite difference methods, they have limitations. In [30, 31] Feng and Neilan investigated a finite element Galerkin method to approximate the viscosity solutions of the Monge-Ampère equation $\det(\nabla^2 u) = f(> 0)$. The nonlinear problem is approximated by a fourth order quasi-linear equation $-\epsilon\Delta^2 u^\epsilon + \det(\nabla^2 u^\epsilon) = f$ completed by some additional boundary conditions. The additional fourth order term, obviously, introduces an additional approximation error. Recently a finite element penalisation method has been developed and analyzed in [13] to solve the Monge-Ampère equation where Newton method is used to solve the resultant nonlinear algebraic system of equations. Omar Lakkis [42] with his co-author used a kind of non-variational finite element method to solve linear equations in a sequence after applying Newton's method to a nonlinear problem. Dean and Glowinski dedicated many of their papers to the numerical solution of nonlinear problems [24, 25, 26, 27]. They have explored least squares and Lagrangian methods to solve nonlinear problems.

The strong nonlinearity in higher order derivatives in fully nonlinear equations makes the standard Galerkin finite element methods based on a direct weak form formulation inapplicable. Recently, Böhmer [10, 11] introduced a general approach that solves the Dirichlet problem for fully nonlinear elliptic equations numerically with the help of a sequence of linear elliptic equations used within an appropriate Newton scheme. These linear elliptic equations can be solved by the finite element method, but the discretization has to be done by appropriate spaces of C^1 finite elements (splines) that admit a stable splitting into a subspace satisfying zero boundary conditions, and its complement. Such a stable splitting has been developed in [20, 21] for a modified space of the Argyris finite element.

2.1 Fully Nonlinear PDEs

Partial differential equations (PDEs) that contain nonlinear terms of the higher order derivatives of unknown function are called fully nonlinear PDEs. To write down a general expression for second order equations let us consider a bounded domain Ω in $\mathbb{R}^n$. Let G be a second order differential operator of the form

$$G(u) = \widetilde{G}(\cdot, u, \nabla u, \nabla^2 u),$$

where $\widetilde{G}$ is a real valued function defined on a domain $\widetilde{\Omega} \times \Gamma$ such that

$$\overline{\Omega} \subset \widetilde{\Omega} \subset \mathbb{R}^n \quad \text{and} \quad \Gamma \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{n \times n},$$

and $\nabla u, \nabla^2 u$ denote the gradient and the Hessian of u, respectively. The points in $\widetilde{\Omega} \times \Gamma$ are denoted by $w = (x, z, p, r)$, with $x \in \widetilde{\Omega}$, $z \in \mathbb{R}$, $p = [p_i]_{i=1}^n \in \mathbb{R}^n$, $r = [r_{ij}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$, to indicate the product structure of this set. We denote by $D(G)$ the domain of the differential operator G. Then the PDE

$$G(u) = \widetilde{G}(\cdot, u, \nabla u, \nabla^2 u) = 0,$$

is fully nonlinear if $\widetilde{G}$ depends nonlinearly on $\nabla^2 u$, otherwise it is either linear, semi-linear or quasi-linear. One of the most prototypical example of such equations is the *Monge-Ampère equation* given by

$$G(u) = \det(\nabla^2 u) - f(x) = 0, \quad x \in \Omega. \tag{2.1.1}$$

For $n = 2$ it can be explicitly written as

$$G(u) = u_{x_1 x_1} u_{x_2 x_2} - u_{x_1 x_2}^2 - f(x) = 0, \quad x \in \Omega. \tag{2.1.2}$$

The operator G is said to be *elliptic* in a subset $\widetilde{\Gamma} \subset \widetilde{\Omega} \times \Gamma$ if the matrix

$$[G_{ij}(w)] = [\frac{\partial \widetilde{G}}{\partial r_{ij}}(w)]_{i,j=1}^{n},$$

is well defined and positive definite for all $w \in \widetilde{\Gamma}$ [11, 33].

Example 2.1.1. In case of Monge-Ampère equation the matrix $G_{ij}(w)$ is the cofactor of $[r_{ij}] = \left[\frac{\partial^2 u}{\partial x_i \partial x_j}\right]$. Accordingly, (2.1.1) is elliptic only for convex functions $u \in C^2(\Omega)$. For such solution of (2.1.1) to exist we must have f positive.

Note that the convexity condition for the solution of the Monge-Ampère equation is also essential for uniqueness. For example consider the Dirichlet boundary condition

$$u|_{\partial\Omega} = \phi. \tag{2.1.3}$$

Then if u is a convex solution to the problem (2.1.2)-(2.1.3), with $\phi = 0$, then $-u$ is another, concave, solution [19, Chapter 4].

2.1.1 Linearization

The linearization of a nonlinear operator is an essential tool used in Böhmer's method to solve fully nonlinear equations. In fact, the *Fréchet derivative* of operator G, if it exists, is usually refered to as linearization of G. For this we formulate the following definition.

Definition 2.1.2. The Fréchet derivative of an operator G at $u = u_0 \in H^\gamma(\Omega)$, $\gamma \geq 2$, is a bounded linear operator $G^{'}(u) : H^\gamma(\Omega) \to L^2(\Omega)$ such that

$$\lim_{h \to 0} \frac{\|G(u+h) - G(u) - G^{'}(u)h\|_{L^2}}{\|h\|_{H^\gamma}} = 0,$$

where $h \in H^\gamma(\Omega)$.

In fact the value of γ in the above definition depends on the operator G. We show in Example 2.1.3 that $\gamma = 5/2$ for the two dimensional Monge-Ampère operator.

In the neighborhood of a fixed function $\hat{u} \in D(G) \subset H^2(\Omega)$ the linear elliptic operator $G'(\hat{u})$ is defined by

$$G'(\hat{u})u = \frac{\partial \widetilde{G}}{\partial z}(\hat{w})u + \sum_{i=1}^{n} \frac{\partial \widetilde{G}}{\partial p_i}(\hat{w})\partial^i u + \sum_{i,j=1}^{n} \frac{\partial \widetilde{G}}{\partial r_{ij}}(\hat{w})\partial^i \partial^j u, \qquad (2.1.4)$$

where $\hat{w} = (x, \hat{u}(x), \nabla \hat{u}(x), \nabla^2 \hat{u}(x))$ is a function of $x \in \Omega$, and ∂^i denotes the partial derivative with respect to the i-th variable [11, Section 5.2.3]. Now if $G : D(G) \subset H^2(\Omega) \to L^2(\Omega)$ is Fréchet differentiable at $\hat{u}$, then $G'(\hat{u}) : D(G) \subset H^2(\Omega) \to L^2(\Omega)$ is its Fréchet derivative. If $G'(\hat{u})$ depends continuously on $\hat{u}$ with respect to the linear operator norm, then G is said to be *continuously differentiable* at $\hat{u}$.

Example 2.1.3. In the case of the two dimensional Monge-Ampère equation, (2.1.4) can explicitly be written as

$$G^{'}(\hat{u})u \;\;=\;\; \left(\frac{\partial^2 \hat{u}}{\partial x_2^2}\right)\frac{\partial^2 u}{\partial x_1^2} - 2\left(\frac{\partial^2 \hat{u}}{\partial x_1 \partial x_2}\right)\frac{\partial^2 u}{\partial x_1 \partial x_2} + \left(\frac{\partial^2 \hat{u}}{\partial x_1^2}\right)\frac{\partial^2 u}{\partial x_2^2} \qquad (2.1.5)$$

Fréchet differentiability of the Monge-Ampère operator

Let $u, h \in D(G) \subset H^2(\Omega)$ and

$$R(h) = G(u+h) - G(u) - G^{'}(u)h = h_{11}h_{22} - h_{12}^2,$$

where $h_{ij} = \partial^i \partial^j h$. Then it is easy to see, using the Cauchy-Schwarz inequality and the embedding theorem, that

$$\begin{aligned}
\|R(h)\|_{L^2} &= \|h_{11}h_{22} - h_{12}^2\|_{L^2} \\
&\leq \|h_{11}h_{22}\|_{L^2} + \|h_{12}^2\|_{L^2} \\
&\leq \|h_{11}\|_{L^4}\|h_{22}\|_{L^4} + \|h_{12}\|_{L^4}^2 \\
&\leq C\left(\|h_{11}\|_{H^{\frac{1}{2}}}\|h_{22}\|_{H^{\frac{1}{2}}} + \|h_{12}\|_{H^{\frac{1}{2}}}^2\right) \\
&\leq 2C\left(\|h\|_{H^{\frac{5}{2}}}^2\right),
\end{aligned}$$

where C is a constant. Hence G', given in (2.1.5), is Fréchet derivative of Monge-Ampère operator G. It also shows that $D(G) = H^{5/2}(\Omega)$ for a bounded domain $\Omega \subset \mathbb{R}^2$. In other words $G(u) \in L^2(\Omega)$ if u belongs to the Sobolev space $H^{5/2}(\Omega)$ and $f \in L^2(\Omega)$. Moreover, $G : H^\gamma(\Omega) \to L^2(\Omega)$ is continuously differentiable if $\gamma = 5/2$.

2.1.2 Existence, Uniqueness and Regularity

Fully nonlinear equations do not necessarily have a smooth solution even if the data is smooth [24]. Again the Monge-Ampère equation provides a good example in this regard. Consider a Dirichlet problem (2.1.2)-(2.1.3), over $\Omega = [0,1]^2$, with $f = 1$ and any constant function ϕ. Then it can be seen that along the boundary we arrive at the following contradiction

$$-u_{x_1x_2}^2 = 1.$$

We formulate some theorems here that provide sufficient conditions for existence and uniqueness of solutions of fully nonlinear equations.

Definition 2.1.4. A domain Ω satisfies an *exterior sphere condition*, for every $\xi \in \partial\Omega$, if there exist a ball $B = B_R(x) \subset \mathbb{R}^n$ s.t. $\overline{B} \cap \overline{\Omega} = \{\xi\}$.

Remark 2.1.5. The *exterior sphere condition* is clearly satisfied by a convex domain.

We say that $\widetilde{G}$ satisfies the *structure conditions* for $w = (x, z, p, r) \in D(\widetilde{G})$ if

$$0 < \lambda|\xi|^2 \leq G_{ij}(x, z, p, r)\xi_i\xi_j \leq \Lambda|\xi|^2, \tag{2.1.6}$$

and

$$\begin{aligned}
&|\widetilde{G}_z(w)|, |\widetilde{G}_p(w)| \leq \lambda\mu, \\
&|\widetilde{G}_x(w)| \leq \lambda\mu(1 + |p| + |r|), & (2.1.7)\\
&|\widetilde{G}_z(w)|, |\widetilde{G}_p(w)|, |\widetilde{G}_{rx}(w)|, |\widetilde{G}_{px}(w)|, |\widetilde{G}_{zx}(w)| \leq \lambda\mu, \\
&|\widetilde{G}_x(w)|, |\widetilde{G}_{xx}(w)| \leq \lambda\mu(1 + |p| + |r|), & (2.1.8)
\end{aligned}$$

for all non zero $\xi \in \mathbb{R}^2$, where λ is non-increasing function of $|z|$ and Λ and μ are non-decreasing functions of $|z|$.

Example 2.1.6. Again for two dimensional Monge-Ampère equation we have

$$\begin{aligned}
\widetilde{G}(x, z, p, r) &= \det(r) - f(x) \\
&= r_{11}r_{22} - r_{12}^2 - f(x).
\end{aligned}$$

Hence

$$\begin{aligned}
&|\widetilde{G}_z(w)| = |\widetilde{G}_p(w)| = |\widetilde{G}_{rx}(w)| = |\widetilde{G}_{px}(w)| = |\widetilde{G}_{zx}(w)| = 0, \\
&|\widetilde{G}_x(w)| = |\frac{\partial f}{\partial x}(x)|, |\widetilde{G}_{xx}(w)| = |\frac{\partial^2 f}{\partial^2 x}(x)|,
\end{aligned}$$

where $\frac{\partial f}{\partial x}$ and $\frac{\partial^2 f}{\partial^2 x}$ are the gradient and Hessian of f respectively. Thus the structure conditions will be satisfied by $\widetilde{G}$ if

$$|\frac{\partial f}{\partial x}(x)| \ , \ |\frac{\partial^2 f}{\partial^2 x}(x)| \leq \lambda\mu(1 + |p| + |r|).$$

Theorem 2.1.7. ([11, Theorem 2.80], [33, Theorem 17.17]). *Assume Ω is a bounded domain in $\mathbb{R}^n$ and $\phi \in C(\partial\Omega)$. Let Ω satisfy an exterior sphere condition for every $\xi \in \partial\Omega$ and $\widetilde{G}$ satisfy the conditions* (2.1.6), (2.1.8), *and let $\widetilde{G} \in C^2(\Omega \times \Gamma)$ be concave (or convex) w.r.t. (z,p,r), and nonincreasing w.r.t. z. Then the classical Dirichlet problem*

$$\begin{aligned} G(u_0) &= 0 \quad , \quad x \in \Omega & (2.1.9) \\ u_0 &= \phi \quad , \quad x \in \partial\Omega & (2.1.10) \end{aligned}$$

has a unique solution $u_0 \in C^2(\Omega) \cap C(\overline{\Omega})$.

The following theorem provides sufficient conditions required for the existence of a global smooth solution.

Theorem 2.1.8. ([11, Theorem 2.79],[33, Theorem 17.12]). *Assume Ω is a bounded domain in $\mathbb{R}^2$ and $\phi \in C^3(\partial\Omega)$. Let $\widetilde{G}$ satisfy the structure conditions* (2.1.6), (2.1.7), *and let $\widetilde{G}_z \leq 0$ in $D(\widetilde{G})$. Then the classical Dirichlet problem*

$$\begin{aligned} G(u_0) &= 0 \quad , \quad x \in \Omega & (2.1.11) \\ u_0 &= \phi \quad , \quad x \in \partial\Omega & (2.1.12) \end{aligned}$$

has a unique solution $u_0 \in C^{2,\alpha}(\overline{\Omega})$ for $0 < \alpha < 1$.

2.2 Böhmer's Method for Fully Nonlinear Elliptic PDEs

Bohmer's method is, in fact, a finite element method to solve general fully nonlinear Dirichlet problems. It is based on the ellipticity of the corresponding linearized problem. We formulate the method in this section. The full theoretical justification for the method can be seen in [11, 10].

We consider the Dirichlet problem for the nonlinear operator G: Find u such that

$$\begin{aligned} G(u) &= 0, \quad x \in \Omega && (2.2.13) \\ u &= \phi, \quad x \in \partial\Omega && (2.2.14) \end{aligned}$$

where ϕ is a continuous function defined on $\partial\Omega$. Under certain assumptions, discussed in Section 2.1.2, there exists a unique solution for this problem.

2.2.1 Spline Spaces and Stable Splitting

As usual in the finite element method, the discretization of the Dirichlet problem is done with the help of spaces of piecewise polynomial functions (splines). To define the spline spaces, let us assume that $\triangle$ is a triangulation of a polyhedral domain $\Omega \subset \mathbb{R}^n$, *i.e.* $\triangle$ is a partition of Ω into simplices such that the intersection of every pair of simplices is either empty or a common face. The space of multivariate splines of degree d and smoothness r is defined by

$$S_d^r(\triangle) = \{s \in C^r(\Omega) : s|_T \in P_d \text{ for all simplices } T \text{ in } \triangle\}, \qquad (2.2.15)$$

where $d > r \geq 0$ and P_d is the space of polynomials of total degree d in n variables.

Definition 2.2.1. The *star* of a vertex v of $\triangle$, denoted by $\text{star}(v) = \text{star}^1(v)$, is the union of all triangles $T \in \triangle$ attached to v. We define $\text{star}^j(v)$, $j \geq 2$, inductively as the union of the stars of all vertices of $\triangle$ contained in $\text{star}^{j-1}(v)$.

Let $\{\triangle^h\}_{h \in H}$ be a family of triangulations of Ω, where h is the maximum edge length in $\triangle^h$. The triangulations in the family are said to be *quasi-uniform* if there is an absolute constant $c > 0$ such that $\rho_T \geq ch$ for all $T \in \triangle^h$, where ρ_T denotes the radius of the inscribed sphere of the simplex T.

Let $S^h \subset S_d^r(\triangle^h)$ be a linear subspace with basis $s_1, \ldots, s_N$ and dual functionals $\lambda_1, \ldots, \lambda_N$ such that $\lambda_i s_j = \delta_{ij}$. This basis is *stable* and *local* if there are three constants $m \in \mathbb{N}$ and $C_1, C_2 > 0$ independent of h such that

(a). $\operatorname{supp} s_k$ is contained in $\operatorname{star}^m(v)$ for some vertex v of $\triangle^h$, for each $k = 1, \ldots, N$,

(b). $\|s_k\|_{L^\infty(\Omega)} \leq C_1$, $k = 1, \ldots, N$, and

(c). $|\lambda_k s| \leq C_2 \|s\|_{L^\infty(\operatorname{supp} s_k)}$, $k = 1, \ldots, N$, for all $s \in S^h$, [11, Section 4.2.6].

To handle the Dirichlet boundary conditions, the following subspace of S^h is important:

$$S_0^h := \left\{ s \in S^h : s|_{\partial\Omega} = 0 \right\}.$$

Moreover, Böhmer's method of solving (2.2.13)–(2.2.14) proposed in [10, 11] requires an additional property called *stable splitting* of S^h into a direct sum

$$S^h = S_0^h + S_b^h,$$

such that a stable local basis $\{s_1, \ldots, s_N\}$ for S^h can be split into two parts

$$\{s_1, \ldots, s_N\} = \{s_1, \ldots, s_{N_0}\} \cup \{s_{N_0+1}, \ldots, s_N\},$$

where $\{s_1, \ldots, s_{N_0}\}$ and $\{s_{N_0+1}, \ldots, s_N\}$ are bases for S_0^h and S_b^h, respectively. Note that the space S_b^h is not uniquely defined by the pair S^h, S_0^h. It was shown in [20] (see also [11, Section 4.2.6]) how the stable splitting can be achieved for a modified space of Argyris finite element using nodal techniques. Also see Chapter 3 for the description of stable splitting using Bernstein-Bézier techniques and discussion of other spaces that possess this property.

2.2.2 Böhmer's Method

Let $u = \hat{u}$ be the solution of (2.2.13)–(2.2.14). According to [10], its approximation $\hat{u}^h \approx \hat{u}$ is sought as a solution of the following problem: Find $\hat{u}^h \in S^h$ such that

$$(G(\hat{u}^h), v^h)_{L^2(\Omega)} = 0 \quad \forall v^h \in S_0^h, \quad \text{and} \tag{2.2.16}$$

$$(\hat{u}^h, v_b^h)_{L^2(\partial\Omega)} = (\phi, v_b^h)_{L^2(\partial\Omega)} \quad \forall v_b^h \in S_b^h, \tag{2.2.17}$$

where $(\cdot,\cdot)$ denotes the inner products in the respective Hilbert spaces. Since S_0^h and S_b^h are finite dimensional linear spaces, the problem (2.2.16)–(2.2.17) is equivalent to a system of algebraic equations with respect to the coefficients of $\hat{u}^h$ in a basis of S^h.

Theorem 2.2.2. *Let Ω be a bounded convex polyhedral domain, and let $G : D(G) \to L^2(\Omega)$, with $D(G) \subset H^2(\Omega)$, satisfy Condition H of [11, Section 5.2.3]. Assume that G is continuously differentiable in the neighbourhood of an isolated solution $\hat{u}$ of* (2.2.13)–(2.2.14)*, such that $\hat{u} \in H^\ell(\Omega)$, $\ell > 2$, and $G'(\hat{u}) : D(G) \cap H_0^1(\Omega) \to L^2(\Omega)$ is boundedly invertible. Furthermore, assume that the spline spaces $S^h \subset S_d^1(\triangle^h)$, $d \geq \ell - 1$, on quasi-uniform triangulations $\triangle^h$ possess stable local bases and stable splitting $S^h = S_0^h + S_b^h$, and include polynomials of degree $\ell - 1$. Then the problem* (2.2.16)–(2.2.17) *has a unique solution $\hat{u}^h \in S^h$ as soon as the maximum edge length h is sufficiently small. Moreover,*

$$\|\hat{u} - \hat{u}^h\|_{H^2(\Omega)} \leq C h^{\ell-2} \|\hat{u}\|_{H^\ell(\Omega)}.$$

In particular, Condition H is satisfied by the Monge-Ampère operators on bounded convex polygonal domains in $\mathbb{R}^2$.

The nonlinear problem (2.2.16)–(2.2.17) can be solved iteratively by a Newton method [10], where the initial guess $u_0^h \in S^h$ satisfies the boundary condition

$$(u_0^h, v_b^h)_{L^2(\partial\Omega)} = (\phi, v_b^h)_{L^2(\partial\Omega)} \quad \forall v_b^h \in S_b^h,$$

and the sequence of approximations $\{u_k^h\}_{k\in\mathbb{N}}$ of $\hat{u}^h$ is generated by

$$u_{k+1}^h = u_k^h - w^h, \quad k = 0, 1, \dots,$$

with $w^h \in S_0^h$ being the solution of the linear elliptic problem:

$$\text{Find } w^h \in S_0^h \text{ such that } (G'(u_k^h)w^h, v^h)_{L^2(\Omega)} = (G(u_k^h), v^h)_{L^2(\Omega)} \; \forall v^h \in S_0^h.$$

Clearly, w^h can be found by using the standard finite element method. Under some additional assumptions on G, it is proved in [10, Theorem 9.1] that u_i^h converges to $\hat{u}^h$ quadratically. Note that in the case when $G(u)$ is only conditionally elliptic (e.g. elliptic only for a convex u for Monge-Ampère equation) the ellipticity of the above linear problem is only guaranteed for u_k^h sufficiently close to the exact solution $\hat{u}$.

Chapter 3

Stable Splitting of Bivariate Smooth Splines using Bernstein-Bézier Methods

In this chapter we systematically study the problem of stable splitting for the spaces of bivariate C^1 splines on triangulations of low degree using the Bernstein-Bézier methods. It turns out that stable splitting can be easily formulated as splitting of the minimal determining sets (MDS). We revisit the modified Argyris space studied in [20] by a different technique, and show that its modification is necessary at least if the convenient MDS splitting approach is used. In addition we answer the question whether there are lower order C^1 spaces that possess stable splitting so that they can be considered as other possible candidates with the modified Argyris space to be used in Böhmer's numerical method to solve fully nonlinear partial differential equations. We show that Clough-Tocher, Powell-Sabin and quadrilateral macro-element spaces are among such spaces.

This chapter is organised as follows. Section 3.1 introduces necessary definitions from the theory of Bernstein-Bézier techniques [43], and define the stable splitting of a minimal determining set. In Section 3.2 we discuss the stable splitting for the Argyris space and modified Argyris space and show why stable splitting for the Argyris space is not possible. Section 3.3 is devoted to C^1 macro-element spaces.

3.1 Bernstein-Bézier Techniques

We use Bernstein-Bézier methods, as a tool, for the construction of spline spaces (discussed in the next chapters) and implementation of Böhmer's finite element method. Bernstein polynomials were first introduced by Sergei Natanovich Bernstein in 1912 to provide a simple proof of the Weierstrass theorem regarding approximation of continuous functions by polynomials on a bounded domains [9]. R. T. Farouki has recently compiled a brief survey that provides a historical development, current state of the theory and applications of Bernstein polynomials [28] . He praises this invention by the words

> "...,*methods introduced to facilitate theoretical proof of seemingly limited scope and practical interest may eventually flourish into useful tools that gain widespread acceptance in diverse practical computations. This category undoubtedly includes the Bernstein polynomial basis,...*"

In the form of Bézier curves/surfaces Bernstein polynomials enjoyed their appearance/use in computer graphics (to model smooth curves and surfaces), engineering design (to develop a quantitative description of the geometry of different products) and CAGD [29, 32, 36] due to many of their desired properties and efficient algorithms. Note that Bézier curves/surfaces are linear combination of

Bernstein polynomials with control points as coefficients. The use of Bernstein polynomials in finite elements remained negligible till the end of 20th century. But currently it gained attention among the finite element community particularly in the context of assembling element matrices efficiently for higher order elements[1, 50, 38]. In [1] efficient algorithms are presented to assemble element system matrices and to compute Bernstein-Bézier moments of coefficients that come from partial differential equation under consideration. These properties add enough attraction to Bernstein-Bézier polynomials to be used in FEM.

Let us recall some of the key concepts of Bernstein-Bézier techniques in this section that will be used in the following chapters. A comprehensive treatment of these techniques can be found in [43].

3.1.1 Bernstein-Bézier Methods for Bivariate Polynomials

As we will use bivariate polynomials we restrict ourselves to $\mathbb{R}^2$. We start by defining *barycentric coordinates* of a point. Let $T := \langle v_1, v_2, v_3 \rangle$ be a nondegenerate triangle. The *barycentric coordinates* of any point $v \in \mathbb{R}^2$ with respect to the triangle T are given by a unique triplet (b_1, b_2, b_3) such that

$$v = \sum_{i=1}^{3} b_i v_i, \quad \sum_{i=1}^{3} b_i = 1,$$

and

$$B^d_{ijk}(v) := \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k, \quad i + j + k = d,$$

are the *Bernstein-Bézier basis polynomials* (*BB-polynomials*) of degree d associated with triangle T. Bernstein-Bézier basis polynomials possess many nice properties. For example one can readily see that they are non-negative over T and sum up to unity as

$$\sum_{i+j+k=d} \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k = (b_1 + b_2 + b_3)^d = 1.$$

And their most important property is that they form a basis for the space P_d of polynomials of degree d [43, Theorem 2.4]. In other words every polynomial p of degree d can be written in the *BB-form* as

$$p = \sum_{i+j+k=d} c_{ijk} B^d_{ijk}, \tag{3.1.1}$$

where c_{ijk} are called the *Bézier coefficients* of p. For given $d \geq 1$, the set $D_{d,T}$

$$D_{d,T} := \left\{ \xi_{ijk} = \frac{iv_1 + jv_2 + kv_3}{d} : \ i+j+k = d, \ i,j,k \geq 0 \right\} \tag{3.1.2}$$

is usually named as the set of *domain points* associated with Bézier coefficients c_{ijk} for triangle $T := \langle v_1, v_2, v_3 \rangle$. To store, evaluate, multiply, differentiate and integrate the polynomials in form (3.1.1) we just need to store a vector c of $\binom{d+2}{2}$ Bézier coefficients of the polynomial for each triangle. Hence we must agree on some ordering of these coefficients to store them when dealing with more than one triangle as in finite element methods over triangulations. For this we use lexicographic order [43, p. 23] to arrange these coefficients throughout the dissertation. That is the indices (i, j, k) are ordered according to the function

$$\varsigma(i,j,k) = \binom{j+k+1}{2} + k + 1,$$

where $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ as usual. Note that we take $\binom{n}{r} = 0$ for $n < r$. Then, for example, for $d = 2$ the ordering is

$$c_{200}, c_{110}, c_{101}, c_{020}, c_{011}, c_{002}.$$

3.1.2 The de Casteljau Algorithm

To evaluate polynomials in BB-form at any point $v \in \mathbb{R}^2$ we have an efficient and numerically stable algorithm called *de Casteljau algorithm*. The algorithm is actually a recursive method based on an obvious recurrence relation for B^d_{ijk} given by

$$B^d_{ijk} = b_1 B^{d-1}_{i-1,j,k} + b_2 B^{d-1}_{i,j-1,k} + b_3 B^{d-1}_{i,j,k-1}, \quad i+j+k=d.$$

Note that expressions with negative subscripts are considered to be zero. The following theorem describes the algorithm in detail.

Theorem 3.1.1 ([43, Theorem 2.8]). *Let p be a polynomial in BB-form with coefficients*

$$c^{(0)}_{ijk} := c_{ijk}, \quad i+j+k=d,$$

Suppose v has a triplet (b_1, b_2, b_3) as its barycentric coordinates, and for all $\ell = 1, \dots, d$, let

$$c^{(\ell)}_{ijk} := c^{(\ell-1)}_{i+1,j,k} + c^{(\ell-1)}_{i,j+1,k} + c^{(\ell-1)}_{i,j,k+1}, \textit{ for } i+j+k = d-\ell,$$

then

$$p(v) = \sum_{i+j+k=d-\ell} c^{(\ell)}_{ijk} B^{d-\ell}_{ijk}(v), \ \forall 0 \leq \ell \leq d.$$

In particular,

$$p(v) = c^d_{000}.$$

It is clear that bivariate polynomials restricted to a straight line are univariate polynomials but the restriction of polynomials in BB-form to the edges of the associated triangle T is more interesting in a sense that it keeps the univariate BB-form. This makes evaluation more efficient at points on edges of T using a simplified version of the de Casteljau algorithm [43, Remark 2.5].

3.1.3 Products and Integrals of BB-polynomials

One of the interesting properties of BB-polynomials is that their product again results in a scaled BB-polynomial. For example for B^d_{ijk} and B^q_{rst} we have

$$B^d_{ijk} B^q_{rst} := \frac{\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}}{\binom{d+q}{d}} B^{d+q}_{i+r,j+s,k+t}.$$

As a consequence if

$$p_1 = \sum_{i+j+k=d} c_{ijk} B^d_{ijk} \text{ and } p_2 = \sum_{r+s+t=q} \tilde{c}_{rst} B^q_{rst} \tag{3.1.3}$$

are two polynomials of degree d and q then simple algebra leads us to the formula

$$\begin{aligned} p_1 p_2 &= \left(\sum_{i+j+k=d} c_{ijk} B^d_{ijk} \right) \left(\sum_{r+s+t=q} \tilde{c}_{rst} B^q_{rst} \right) \\ &= \sum_{\substack{i+j+k=d \\ r+s+t=q}} c_{ijk} \tilde{c}_{rst} \frac{\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}}{\binom{d+q}{d}} B^{d+q}_{i+r,j+s,k+t}. \end{aligned}$$

It is easy to see that

$$\frac{\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}}{\binom{d+q}{d}} \leq 1, \text{ for } i+j+k=d,\ r+s+t=q.$$

Hence the Bézier coefficients for the product polynomial $p_1 p_2$ remain bounded provided the Bézier coefficients for p_1 and p_2 are bounded. In other words the multiplication of polynomials in BB-form is a stable process.

Similarly, a simple and efficient formula for integrals of BB-polynomials is formulated in the following theorem.

Theorem 3.1.2 ([43, Theorem 2.33]). *Let p be a polynomial in BB-form* (3.1.1) *over a triangle T, then*

$$\int_T p\,dxdy = \frac{|T|}{\binom{d+2}{2}} \sum_{i+j+k=d} c_{ijk},$$

where $|T|$ is the area of triangle T.

As a consequence of this theorem we get a formula for the inner product of two polynomials p_1 and p_2, mentioned above in (3.1.3), as

$$\begin{aligned}
\int_T p_1 p_2 dxdy &= \frac{|T|}{\binom{d+q}{d}\binom{d+q+2}{2}} \sum_{\substack{i+j+k=d\\ r+s+t=q}} c_{ijk}\tilde{c}_{rst}\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}, \\
&= \frac{|T|}{\binom{d+q}{d}\binom{d+q+2}{2}} c^T Q \tilde{c},
\end{aligned}$$

with c and $\tilde{c}$ are the (lexicographically) ordered vectors of coefficients of p_1 and p_2 respectively. Note that a matrix Q depends only on the degrees of two polynomials but not on the shape of T. Hence Q can be computed once, for given d and q, when using this formula. (particularly needed while computing entries for element matrices in finite element methods).

3.1.4 Gradient of BB-polynomials

As one often needs to compute gradients of basis functions while assembling element matrices we formulate an expression for the gradient of BB-polynomials. We have

$$\begin{aligned}
\nabla B^d_{ijk} &= \nabla \left(\frac{d!}{i!j!k!} b_1^i b_2^j b_3^k \right), \\
&= d\left(B^{d-1}_{i-1,j,k} \nabla b_1 + B^{d-1}_{i,j-1,k} \nabla b_2 + B^{d-1}_{i,j,k-1} \nabla b_3 \right).
\end{aligned}$$

Moreover the basic definition of barycentric coordinates leads us to the formula

$$\nabla b_k := \frac{|e_k|}{2|T|}\mathbf{n}_k, \; k = 1, 2, 3,$$

where $\mathbf{n}_k$ is the unit inward normal to the edge e_k opposite to the vertex v_k of triangle $T := \langle v_1, v_2, v_3 \rangle$ and $|e_k|$ denotes the length of the edge e_k.

3.1.5 Smoothness Conditions

Before our study of smooth spline spaces we give conditions for a smooth join of two polynomials on neighbouring triangles. These conditions help us glue different polynomial pieces up with desired smoothness. An interesting fact is that these algebraic smoothness conditions can easily be expressed, interpreted and analyzed geometrically. For example given two neighbouring triangles $T := \langle v_1, v_2, v_3 \rangle$ and $\tilde{T} := \langle v_4, v_3, v_2 \rangle$ sharing the edge $e := \langle v_2, v_3 \rangle$ and two polynomials

$$p_1 = \sum_{i+j+k=d} c_{ijk} B^d_{ijk} \text{ and } p_2 = \sum_{r+s+t=d} \tilde{c}_{rst} \tilde{B}^d_{rst}, \tag{3.1.4}$$

over T and $\tilde{T}$ respectively then p_1 and p_2 joins with C^r smoothness, $r = 0, \ldots, d-1$, if and only if for each $n = 0, \ldots, r$,

$$\tilde{c}_{njk} = \sum_{\nu+\mu+\kappa=n} c_{\nu,k+\mu,j+\kappa} B^n_{\nu\mu\kappa}(v_4), \; j+k = d-n. \tag{3.1.5}$$

Let us express these smoothness conditions for $r = 0, 1$, more explicitly. It is easy to see that p_1 and p_2 join continuously ($r = 0$) along e if their BB coefficients over e coincide, i.e. if

$$\tilde{c}_{0jk} = c_{0kj}, \; j+k = d, \tag{3.1.6}$$

see Figure 3.1. Furthermore, the condition for C^1 smoothness across e is that (3.1.6) holds along with

$$\tilde{c}_{1jk} = b_1 c_{1,k,j} + b_2 c_{0,k+1,j} + b_3 c_{0,k,j+1}, \; j+k = d-1, \tag{3.1.7}$$

where (b_1, b_2, b_3) are barycentric coordinates of v_4 relative to T, see Figure 3.2.

As a consequence we note that the coefficients $\tilde{c}_{ijk}$ can be computed using smoothness conditions from c_{ijk}'s involved in the conditions. In fact if we know all coefficients of p then (3.1.5) can be used to compute the coefficients of $\tilde{p}$ corresponding to domain points in the first r rows parallel to the edge e and [43, Lemma 2.29] shows that these computations are a stable process.

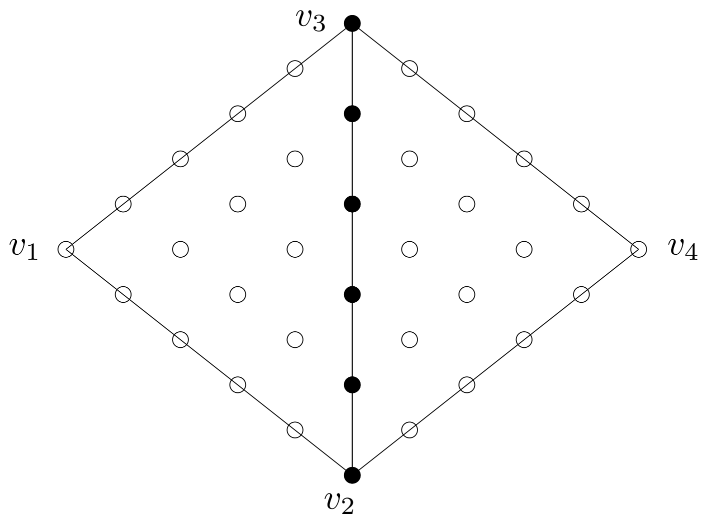

Figure 3.1: Black dots are Bézier coefficients involved in C^0 smoothness conditions across the edge $\langle v_2, v_3 \rangle$. $(d = 5)$

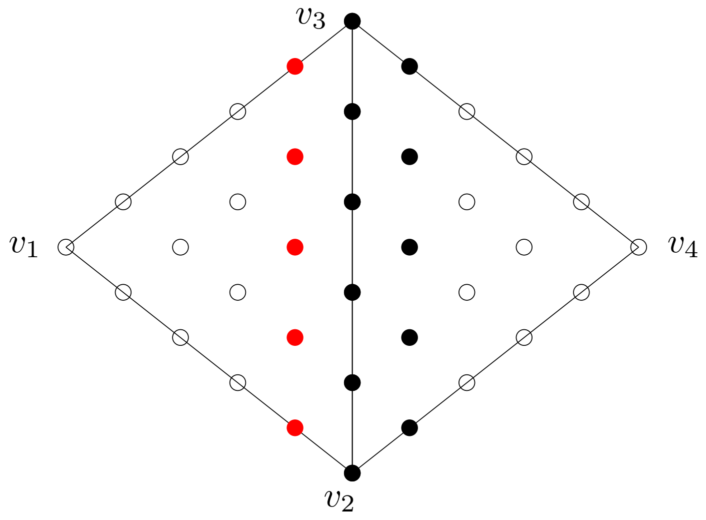

Figure 3.2: Bézier coefficients involved in C^1 smoothness conditions across the edge $\langle v_2, v_3 \rangle$. The coefficients $\tilde{c}_{1jk},\ j + k = d - 1$, are marked as red dots while coefficients on R.H.S of (3.1.7) are marked as black dots. $(d = 5)$

3.1.6 Degree Raising

Since a polynomial p of degree d can be considered as a polynomial of degree $d+r$, $r > 0$, thus p written in BB-form (3.1.1) can be expressed in terms of a Bernstein basis of degree $d+r$. In other words we can compute the coefficients $c_{ijk}^{[d+r]}$ from known coefficients $c_{ijk}^{[d]} = c_{ijk}$ [43, Section 2.15] to write p in the form

$$p = \sum_{i+j+k=d+r} c_{ijk}^{[d+r]} B_{ijk}^{d+r}. \tag{3.1.8}$$

For the sake of simplicity let $r = 1$. Then by [43, Theorem 2.39] we have

$$c_{ijk}^{[d+1]} := \frac{ic_{i-1,j,k} + jc_{i,j-1,k} + kc_{i,j,k-1}}{d+1}, \; i+j+k = d+1. \tag{3.1.9}$$

Obviously to compute $c_{ijk}^{[d+r]}$ the process can be repeated r times. From (3.1.9) we see that

$$c_{ijk}^{[d+1]} \leq \max\left\{c_{i-1,j,k}, c_{i,j-1,k}, c_{i,j,k-1}\right\}, \; i+j+k = d+1.$$

and as a consequence this inequality tells us that the process of degree raising is a stable process. Also see Remark 3.1.3.

Remark 3.1.3. Since degree raising can be used to create a sequence of control surfaces and the fact that this sequence of control surfaces converges uniformly to corresponding polynomial surface also indicates that the process of degree raising is a stable process [43, Theorem 3.23]).

Remark 3.1.4. Note that the degree raising formulas can also be used over a subdomain of T if required. For example if for given coefficients of a polynomial of degree d in $D_2(v)$, where v is one of the vertices of T, then the corresponding coefficients for higher degree in $D_2(v)$ can be computed by using (3.1.9) only over a subdomain $D_2(v)$.

3.1.7 Bernstein-Bézier Finite Elements and Spline Spaces

Recall that $\triangle$ is a triangulation of a bounded domain $\Omega \subset \mathbb{R}^2$ as defined in Definition 1.1.1, $D_{d,\triangle} := \bigcup_{T\in\triangle} D_{d,T}$ is a set of domain points associated with $\triangle$ where $D_{d,T}$ is as defined in (3.1.2) and $S_d^r(\triangle)$ is the space of bivariate splines of degree d and smoothness r as defined in (2.2.15). Now (3.1.1) confirms that every $s \in S_d^r(\triangle)$ can be written, over T, as

$$s|_T = \sum_{\xi \in D_{d,T}} c_\xi B_\xi^{T,d},$$

where $B_\xi^{T,d}$ are BB-basis polynomials of degree d associated with triangle T in $\triangle$. Note that, in view of the smoothness condition (3.1.6), continuity of s implies that the BB-vectors of $s|_T$ and $s|_{\tilde{T}}$ agree on domain points on the edge shared by triangles T and $\tilde{T}$. We now introduce some additional notation. We refer to the set

$$R_n(v) := \{\xi_{ijk} \in D_{d,\triangle} : i = d-n\}, \quad 0 \le n \le d,$$

of domain points as the *ring* of radius n around the vertex v and refer to the set

$$D_n(v) := \bigcup_{m=0}^{n} R_m(v)$$

as the *disk* of radius n around the vertex v.

3.1.8 Minimal Determining Set

A key concept for dealing with spline spaces, using Bernstein-Bézier techniques, is that of a minimal determining set defined as follows.

Definition 3.1.5. *A set $M \subset D_{d,\triangle}$ is a* determining set *for a linear space $S \subset S_d^r(\triangle)$ if*

$$s \in S \text{ and } c_\xi = 0 \quad \forall \xi \in M \quad \Rightarrow \quad s = 0,$$

and M is a minimal determining set *(MDS) for the space S if there is no smaller determining set.*

Moreover, the property of MDS M that $\dim S := \#\{M\}$ makes it more interesting and shows that there is a one-to-one correspondence between points in M and degrees of freedom for space S. In fact we can use an MDS to construct a basis for the corresponding space. Before we go in detail about such a construction we need to define some more properties an MDS M might possess. First let

$$\Gamma_\eta := \{\xi \in M \ : \ c_\eta \text{ depends on } c_\xi\},$$

where we say that c_η depends on c_ξ, $\xi \in M$, if the value of c_η is changed when we change the value of c_ξ.

Definition 3.1.6. *A minimal determining set M for a space S is said to be local if there exists ℓ not depending on $\triangle$ such that*

$$\Gamma_\eta \subset \operatorname{star}^\ell(T_\eta) \quad \forall \eta \in D_{d,\triangle} \backslash M,$$

where T_η is a triangle containing η.

Definition 3.1.7. *A minimal determining set M for a space S is called stable if there exists a constant K which may depend only on d, ℓ and the smallest angle $\theta_\triangle$ in the triangulation $\triangle$ such that*

$$|c_\eta| \le K \max_{\xi \in \Gamma_\eta} |c_\xi| \quad \forall \eta \in D_{d,\triangle} \backslash M. \tag{3.1.10}$$

Given a stable local minimal determining set M for space $S \subset S_d^r(\triangle)$, if we assign values to the coefficients $\{c_\xi\}_{\xi \in M}$, then the remaining coefficients c_η, $\eta \in D_{d,\triangle} \backslash M$ can be computed using the smoothness conditions. Thus a stable local MDS M can be used to construct a stable local basis $\{s_\xi\}_{\xi \in M}$ for S [43, Section 5.8] defined as

$$B := \{s_\xi : \ c_\xi = 1, \text{ and } c_\eta = 0 \text{ for all } \eta \in M \setminus \{\xi\}\}. \tag{3.1.11}$$

This basis is usually named as an M-basis for S.

Remark 3.1.8. *An* M-basis, obtained using (3.1.11), satisfies the standard conditions of stable local basis in a sense defined in Section 2.2.1.

3.1.9 Stable Splitting

A stable splitting of an M-basis B is achieved by an appropriate splitting of the MDS, which leads to the following definition.

Definition 3.1.9. *Assume that the space $S \subset S_d^r(\triangle)$ has a stable local MDS M and let*

$$S_0 := \{s \in S \ : \ s|_{\partial\Omega} = 0\}\,. \tag{3.1.12}$$

The MDS M is said to admit a stable splitting *if M is the disjoint union of two subsets $M_0, M_b \subset M$ such that*

$$S_0 = \{s \in S : c_\xi = 0 \ \forall \xi \in M_b\} \tag{3.1.13}$$

and M_0 and M_b are stable local MDS for the spaces S_0 and S_b, respectively, where

$$S_b := \{s \in S : c_\xi = 0 \ \forall \xi \in M_0\}\,. \tag{3.1.14}$$

Note that if M is a stable local MDS, and $M = M_0 \cup M_b$ is a disjoint union, then it is a stable splitting as soon as (3.1.13) holds. Indeed, assume (3.1.13) is correct. If $s \in S_0$, then its coefficients related to M_b are zero, and similarly if $s \in S_b$ then its coefficients related to M_0 are zero. Hence computing s from the coefficients corresponding to points in M_0 (respectively, M_b) is equivalent to computing from M, and so M_0 and M_b are determining sets for S_0 and S_b, respectively. They are minimal determining sets because otherwise M would not be minimal. Obviously stability and locality properties of M_0 and M_b are also inherited from M. If M admits a stable splitting, then $S = S_0 + S_b$ and it is easy to see that

$$\{s_\xi\}_{\xi \in M} = \{s_\xi\}_{\xi \in M_0} \cup \{s_\xi\}_{\xi \in M_b}$$

is a stable splitting of the stable local basis $\{s_\xi\}_{\xi \in M}$ with $\{s_\xi\}_{\xi \in M_0}$ and $\{s_\xi\}_{\xi \in M_b}$ as basis for S_0 and S_b respectively.

3.2 Stable Splitting for Argyris Finite Element

Argyris space is in fact a superspline subspace of $S_5^1(\triangle)$. Superspline subspaces have actually an enhanced smoothness at certain vertices of $\triangle$. Generally the superspline subspaces $S_d^{r,\rho}(\triangle)$, $r \le \rho \le d$, of $S_d^r(\triangle)$ are defined as

$$S_d^{r,\rho}(\triangle) = \{\, s \in S_d^r(\triangle) : s \in C^\rho(v)\ \forall v \in V \}\,, \tag{3.2.15}$$

where V is the set of all vertices of $\triangle$. Then Argyris finite element space is obtained with $d = 5$, $r = 1$ and $\rho = 2$ in (3.2.15). Before we describe a minimal determining set for Argyris space we need to introduce some notation as follows. Let E be the set of all edges of $\triangle$. For each $v \in V$, let T_v be any one of the triangles sharing the vertex v and let $M_v := D_2(v) \cap T_v$. For each edge e of the triangulation $\triangle$, let $T_e := \langle v_1, v_2, v_3 \rangle$ be one of the triangles sharing the edge $e := \langle v_2, v_3 \rangle$ and let $M_e := \left\{\xi_{122}^{T_e}\right\}$. Then from [43, Theorem 6.1] we have

Theorem 3.2.1. $\dim S_5^{1,2}(\triangle) = 6\#\{V\} + \#\{E\}$ *and*

$$M = \bigcup_{v \in V} M_v \cup \bigcup_{e \in E} M_e \tag{3.2.16}$$

is a stable local minimal determining set for $S_5^{1,2}(\triangle)$.

An example of minimal determining set for Argyris space is given in Figure 3.3 (left).

3.2.1 Modified Argyris Space

We now modify the Argyris space to achieve the stable splitting. A similar construction is discussed in term of nodal basis functions in [20]. We will explain in Section 3.2.3 why this modification is required. Let us denote the modified Argyris space by $\tilde{S}$, where

$$\tilde{S} := \left\{ s \in S_5^1(\triangle) : s \in C^2(v), \text{ for all } \textit{interior} \text{ vertices } v \text{ of } \triangle \right\}. \tag{3.2.17}$$

Note that the modification is removing C^2 smoothness only at boundary vertices. Let us now distinguish between boundary vertices and interior vertices by using V_I and V_B for the sets of interior and boundary vertices respectively. And let E_I and E_B denote interior and boundary edges respectively, such that

$$V = V_I \cup V_B, \quad E = E_I \cup E_B.$$

We describe a minimal determining set $\tilde{M}$ for this modified space $\tilde{S}$. Since we have modified the space only at the boundary vertices, so the points in M related to interior vertices and related to all edges, will belong to $\tilde{M}$. That is,

$$\left(\bigcup_{v \in V_I} M_v \cup \bigcup_{e \in E} M_e \right) \subset \tilde{M}.$$

However, we will have to modify the sets corresponding to the boundary vertices $v \in V_B$. First of all, we require that each T_v, $v \in V_B$, is a triangle sharing an edge with the boundary of Ω (we call it a *boundary triangle*). Furthermore, we add some more points to M_v, $v \in V_B$, as follows. Let us denote all edges of $\triangle$ emanating from a vertex $v \in V_B$, in counter-clockwise order, by

$$E_v = \{e_1, e_2, \cdots, e_n\}.$$

Then clearly $e_1, e_n \in E_B$, and the triangle T_v is formed by either e_1, e_2 or e_{n-1}, e_n. For each e_i, let ξ_i be the (unique) domain point in $R_2(v) \cap e_i$, $i = 1, \ldots, n$. We set

$$\tilde{M}_v := M_v \cup \{\xi_1, \xi_2, \cdots, \xi_n\}.$$

Theorem 3.2.2. $\dim \tilde{S} = 6\#\{V_I\} + \#\{E\} + \sum_{v \in V_B} (4 + \#E_v)$ *and*

$$\tilde{M} := \bigcup_{v \in V_I} M_v \cup \bigcup_{e \in E} M_e \cup \bigcup_{v \in V_B} \tilde{M}_v. \tag{3.2.18}$$

is stable local MDS for modified Argyris space $\tilde{S}$.

Proof. We set the coefficients $\{c_\xi\}_{\xi \in \tilde{M}}$ for any spline $s \in \tilde{S}$ to arbitrary values and show that all other coefficients, i.e. $\{c_\xi\}_{\xi \in D_{5,\triangle} \setminus \tilde{M}}$, of s can be determined consistently.

Now first note that for each $v \in V_I$ and for each $e \in E$ the points in M_v and M_e are the same as for Argyris space. So we only need to prove that for each $v \in V_B$ the set $\tilde{M}_v$ is an MDS on $D_2(v)$. To this end, for each $v \in V_B$, we set the coefficients of s corresponding to points in $\tilde{M}_v$ and see that, in view of C^1 smoothness conditions, all coefficients corresponding to domain points in $D_2(v)$ can be determined consistently. Thus by [43, Theorem 5.15] $\tilde{M}$ is minimal determining set for the space $\tilde{S}$. Observe that $\tilde{M}$ is a stable MDS. Indeed, for each $v \in V_I$ and all edges $e \in E$ the stability follows from [43, Lemma 2.29]. And for each $v \in V_B$ the set $\tilde{M}_v$ is a stable MDS for S^1_5 on $D_2(v)$ by [43, Theorem 11.7]. Standard arguments show that $\tilde{M}$ is local. ■

An example of minimal determining set for modified Argyris space is given in Figure 3.3 (right).

3.2.2 Stable Splitting

Now we show how to determine a stable splitting $\tilde{M} = \tilde{M}_0 \cup \tilde{M}_b$ of the MDS $\tilde{M}$ for modified Argyris space $\tilde{S}$.

It is already understood that all those points of $\tilde{M}$ which are on the boundary will be in $\tilde{M}_b$ and those points lying in M_v, $v \in V_I$, and M_e along with the points

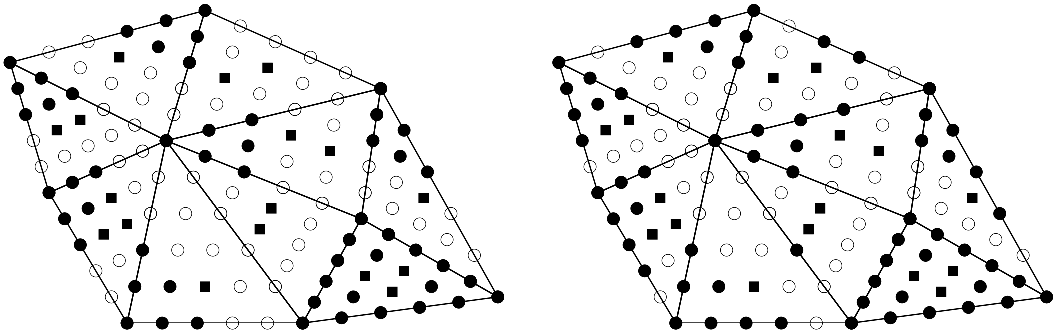

Figure 3.3: Minimal determining sets for the Argyris space (*left*) and for the modified Argyris space (*right*). The points in the sets M_v, $\tilde{M}_v$ are marked by black dots, and those in M_e by black squares.

in $R_2(v)$, $v \in V_B$, but not on either e_1 or e_n, will be in $\tilde{M}_0$. Consider, for each $v \in V_B$, the remaining point which lies in $R_1(v)$, $v \in V_B$, but not on the boundary edges. We denote this point by ξ_v. Whether ξ_v belongs to $\tilde{M}_0$ or $\tilde{M}_b = \tilde{M} \setminus \tilde{M}_0$ depends on the geometry of the boundary edges e_1 and e_n, as follows.

- If e_1 and e_n are non-collinear, then $\xi_v \in \tilde{M}_b$.
- If e_1 and e_n are collinear, then $\xi_v \in \tilde{M}_0$.

Indeed, in the non-collinear case the coefficient corresponding to ξ_v is zero for all $s \in \tilde{S}_0$, whereas in the collinear case it can be chosen freely. The Figures 3.4 and 3.5 show points in $\tilde{M}_0$ and $\tilde{M}_b$ for the boundary vertex with collinear and non-collinear edges respectively.

Theorem 3.2.3. *$\tilde{M} = \tilde{M}_0 \cup \tilde{M}_b$ is a stable splitting of MDS $\tilde{M}$.*

Proof. If $s \in \tilde{S}_0$, then all its Bézier coefficients corresponding to domain points on the boundary are zero since $s|_{\partial\Omega} = 0$. For $v \in V_B$, where the boundary edges are non-collinear, the C^1 smoothness implies that the gradient at v is also

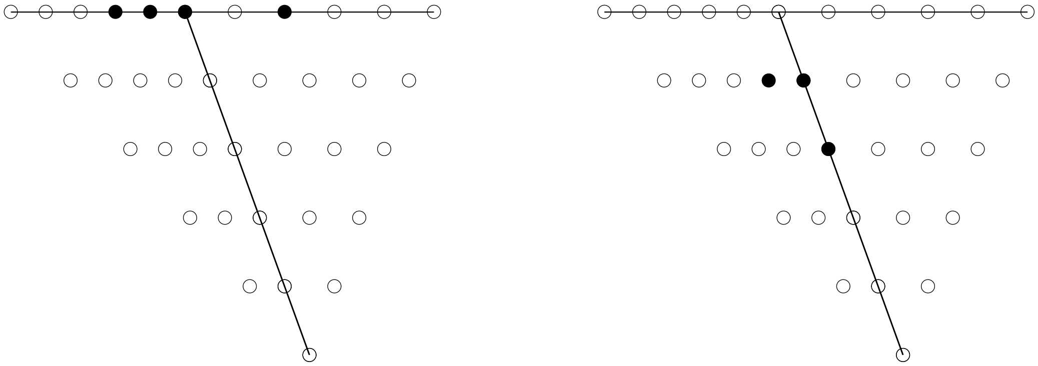

Figure 3.4: Splitting of points in $\tilde{M}_v$, $v \in V_B$ for modified Argyris space with collinear boundary edges. *Left*: $\tilde{M}_v \cap \tilde{M}_b$, *right*: $\tilde{M}_v \cap \tilde{M}_0$.

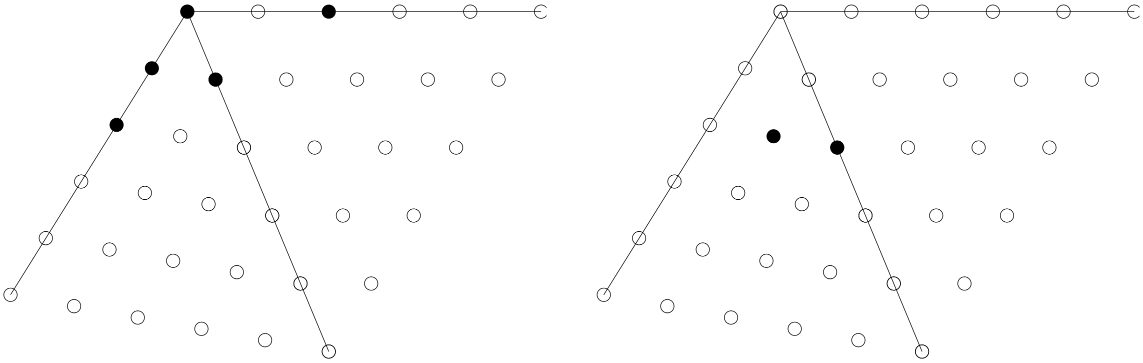

Figure 3.5: Splitting of points in $\tilde{M}_v$, $v \in V_B$ for modified Argyris space with noncollinear boundary edges. *Left*: $\tilde{M}_v \cap \tilde{M}_b$, *right*: $\tilde{M}_v \cap \tilde{M}_0$.

zero, and hence the coefficient of s at ξ_v is also zero. This shows that $\tilde{S}_0 \subset \{s \in \tilde{S} : c_\xi = 0 \ \forall \xi \in \tilde{M}_b\}$. Conversely, assume $s \in \tilde{S}$ and $c_\xi = 0$ for all $\xi \in \tilde{M}_b$. Let $v \in V_B$ and $E_v = \{e_1, e_2, \cdots, e_n\}$ as before. Without loss of generality assume that $D_2(v) \cap e_1 \subset \tilde{M}_v$ and $R_2(v) \cap e_n \subset \tilde{M}_v$. Therefore $c_\xi = 0$ at all these points. However, due to the C^1 smoothness $c_\xi = 0$ also for the domain point in $R_1(v) \cap e_n$, both in the collinear and non-collinear case. This shows that $c_\xi = 0$ for all domain points on the boundary of Ω and hence $s|_{\partial\Omega} = 0$. Thus, $\tilde{S}_0 = \{s \in \tilde{S} : c_\xi = 0 \ \forall \xi \in \tilde{M}_b\}$, which completes the proof, see the discussion following Definition 3.1.9. ■

3.2.3 Why Modification in Argyris Space is Required

We now prove that modification is needed in Argyris space at the boundary vertices to achieve a stable splitting.

We first consider the Argyris space $S_5^{1,2}(\triangle)$ with M in Theorem 3.2.1 being its MDS, and show that no splitting $M = M_0 \cup M_b$ is possible. For the sake of simplicity consider a boundary vertex v with two triangles attached such that the boundary edges are non-collinear. On the contrary, assume that such a splitting has been found. Let $T := \langle v_1, v_2, v_3\rangle$ and $\tilde{T} := \langle v_4, v_3, v_2\rangle$ be two triangles in $\triangle$ with v_3 as a boundary vertex and assume that the edges $\langle v_3, v_4\rangle$ and $\langle v_3, v_1\rangle$ are boundary edges. Consider the set

$$M_{v_3} := D_2(v_3) \cap T = \{\xi_{005}, \xi_{014}, \xi_{023}, \xi_{104}, \xi_{113}, \xi_{203}\} \subset M,$$

see the Figure 3.6, and let

$$s|_T = \sum_{i+j+k=5} c_{ijk} B^5_{ijk}, \quad s|_{\tilde{T}} = \sum_{i+j+k=5} \tilde{c}_{ijk} \tilde{B}^5_{ijk},$$

where B^5_{ijk} and $\tilde{B}^5_{ijk}$ are Bernstein basis polynomials associated with T and $\tilde{T}$ respectively. In the case that the edges $\langle v_3, v_4\rangle$ and $\langle v_3, v_1\rangle$ are non-collinear, the points $\{\xi_{005}, \xi_{014}, \xi_{104}, \xi_{203}\}$ must be in M_b, because $s \in S$ has zero coefficients at these points. We show that $\{\xi_{113}, \xi_{023}\} \not\subset M_0$. Let (b_1, b_2, b_3) be barycentric coordinates of v_4 relative to T. Then by a C^2 smoothness condition, see [43, Theorem 2.28], across the edge $e := \langle v_3, v_2 \rangle$ we can write

$$\tilde{c}_{230} = b_1^2 c_{203} + 2b_1 b_2 c_{113} + 2b_2 b_3 c_{014} + b_2^2 c_{023} + 2b_1 b_3 c_{104} + b_3^2 c_{005},$$

and because $\tilde{c}_{230} = c_{203} = c_{014} = c_{104} = c_{005} = 0$,

$$0 = 2b_1 c_{113} + b_2 c_{023},$$

which shows that c_{113} and c_{023} are linearly dependent so that ξ_{113}, ξ_{023} cannot be both in M_0. Moreover, we cannot shift one of these points to M_b because there is a spline $s \in S_0$ such that

$$c_{113}, c_{023} \neq 0,$$

e.g. s with $c_{113} = b_2$ and $c_{023} = -2b_1$. Note that $b_2 \neq 0$ if the boundary edges are non-collinear.

Moreover, we prove that no other MDS admits a stable splitting, either.

Theorem 3.2.4. *No MDS for the Argyris space can be stably split on arbitrary triangulations.*

Proof. Assume that the triangulation $\triangle$ is such that there is a boundary vertex v with two triangles T and $\tilde{T}$ attached, and the boundary edges are non-collinear at v, as in the above proof. Let M be some MDS for Argyris space.

From the dimension argument we know that there must be exactly six points in $M \cap D_2(v)$. For the non-collinear boundary edges, no points on boundary edges or in $R_1(v)$ can be in M_0 because, all the corresponding coefficients of splines in S_0 are zero. So the only candidates for M_0 are the points in $R_2(v)$ not on boundary edges. Now we discuss the relation between the coefficients $\tilde{c}_{131}, c_{113}, c_{023}$ of $s \in S_0$

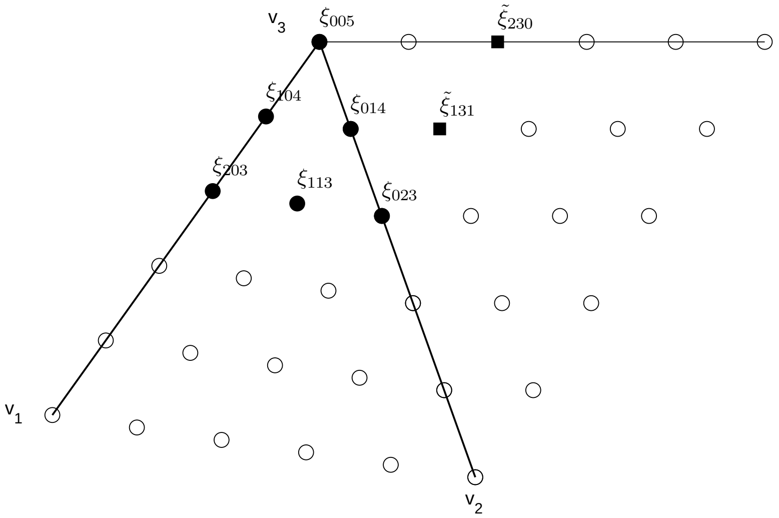

Figure 3.6: The black dots are MDS points in $M_{v_3}, v_3 \in V_B$, for Argyris space. The two domain points marked by black squares are involved in the smoothness conditions discussed in the proof of Theorem 3.2.4.

at these points. By using C^1 and C^2 condition across the common edge of T and $\tilde{T}$ we get

$$\begin{aligned}\tilde{c}_{131} &= b_1 c_{113} + b_2 c_{023}\\ 0 &= 2b_1 c_{113} + b_2 c_{023}\end{aligned}$$

By subtracting these equations we can write

$$\tilde{c}_{131} = -b_1 c_{113}$$

Hence the three coefficients cannot be set arbitrarily. Only one of them can be chosen freely, which cannot be either $\tilde{c}_{131}$ or c_{113}. Indeed, let us choose e.g. c_{113} arbitrarily, then from the above equations we obtain

$$c_{023} = \frac{-2b_1 c_{113}}{b_2}$$

and hence $c_{023} \to \infty$ for $b_2 \to 0$ as the boundary edges get collinear. This would be unstable as the minimum angles in $T, \tilde{T}$ do not degenerate.

Thus ξ_{023} is the only point to be in M_0. It is easy to see that M_b must contain $\xi_{203}, \tilde{\xi}_{230}$ and three points in $D_1(v)$. Consider the basis spline s in S_b corresponding to $\tilde{\xi}_{230}$. Then its coefficient satisfy

$$\tilde{c}_{230} = 1, \quad c_{203} = c_{023} = 0, \quad c_\xi = 0, \quad \xi \in D_1(v)$$

Now again using C^1 and C^2 smoothness conditions we find

$$\tilde{c}_{230} = 2b_1 b_2 c_{113} \;\text{ or }\; c_{113} = \frac{1}{2b_1 b_2},$$

which is unbounded for $b_2 \to 0$ as the boundary gets flat. ■

Remark 3.2.5. If a boundary vertex v has exactly two triangles attached and the boundary edges are not collinear at v, then stable splitting of an MDS is impossible for any spline space S where each spline is C^2 continuous at v. Indeed, this follows by the arguments in the proof of Theorem 3.2.4. In fact, it is easy to see that the set $D_2(v) \cap T$ as MDS for S on $D_2(v)$ cannot be split stably for a boundary vertex with any number of triangles attached.

3.2.4 Numerical Comparison of Stability of MDS for Argyris and Modified Argyris Spaces

To endorse Theorem 3.2.4 with numerical support we present numerical results of comparison of stability of the minimal determining sets for Argyris space and modified Argyris space in this section. We implement the second MDS for Argyris space suggested in Theorem 3.2.4, see Figure 3.8, let us call it M_A, and an MDS $\tilde{M}$ for modified Argyris space. We look and compare for the maximum of the absolute values of BB coefficients of basis functions corresponding to respective MDS's for both spaces to see the stability of $\tilde{M}$ and M_A over a sequence of triangulations where boundary edges gets more and more collinear. For example this happens when one needs to approximate a circle by a polygonal domain where very refined meshes result in triangulations with near collinear boundary edges. To illustrate this we define a polygonal domain Ω^h with boundary vertices on a circle with an initial mesh $\triangle^h$, see Figure 3.7, where h is the maximum length of edges in $\triangle^h$. We get a sequence of refined meshes $\triangle^h$ by joining mid-points of edges where, for boundary triangles, we take mid-point on circular arc for boundary edges. Let $\tilde{B}$ and B_A be bases for modified Argyris and Argyris spaces corresponding to their MDS's $\tilde{M}$ and M_A respectively obtained using (3.1.11).

We need to introduce some more notation. Let $\{T_\kappa\}_{\kappa=1}^{N_t}$ be the set of triangles in $\triangle^h$ with some fixed ordering. Recall that any spline $s \in S_d^0(\triangle^h)$ restricted to T_κ can be written in the form

$$s|_{T_\kappa} = \sum_{i+j+k=5} c_{ijk} B_{ijk}^d,$$

where c_{ijk} are Bézier coefficients of s on T_κ. Let $\mathbf{C}_{T_\kappa}$, $\kappa = 1, \cdots, N_t$, denote the row vector of these coefficients c_{ijk} of s on T_κ, ordered lexicographically and let $\mathbf{V}(s)$ be a row vector of all $\mathbf{C}_{T_\kappa}$'s, $\kappa = 1, \cdots, N_t$, for a spline s,

$$\mathbf{V}(s) = \left[\mathbf{C}_{T_1}, \mathbf{C}_{T_2}, \cdots, \mathbf{C}_{T_{N_t}}\right]. \tag{3.2.19}$$

Furthermore let

$$\begin{aligned} c_1 &= \max_{\xi \in \tilde{M}} \|V(s_\xi)\|_\infty \\ c_2 &= \max_{\xi \in M_A} \|V(s_\xi)\|_\infty. \end{aligned} \tag{3.2.20}$$

We plot c_1 and c_2 against n_r where n_r denotes number of refinements as shown in Figure 3.9. We see that c_2 is growing exponentially with refinements,(i.e. as

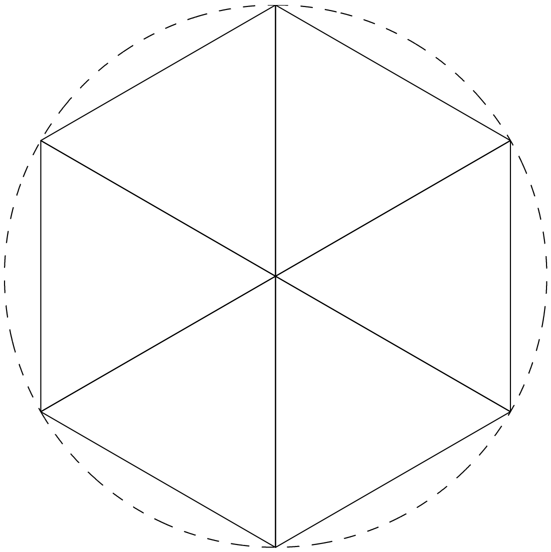

Figure 3.7: Initial triangulation $\triangle^h$ of a circle.

boundary edges get near collinear), and c_1 remains bounded for all meshes. Violation of the inequality (3.1.10) indicates the instability of MDS M_A .

3.3 C^1 Macro-element Spaces

Now we discuss the possibility of stable splitting of minimal determining sets of some of the C^1 macro-element spaces.

3.3.1 Stable Splitting of Clough-Tocher Macro-element Space

Given a triangulation $\triangle$ of a domain Ω, let $\triangle_{CT}$ be corresponding Clough-Tocher refinement of $\triangle$, where each triangle is split into three subtriangles, see Figure 3.10.

Consider the stable local MDS M given in [43, Theorem 6.5] for C^1 Clough-

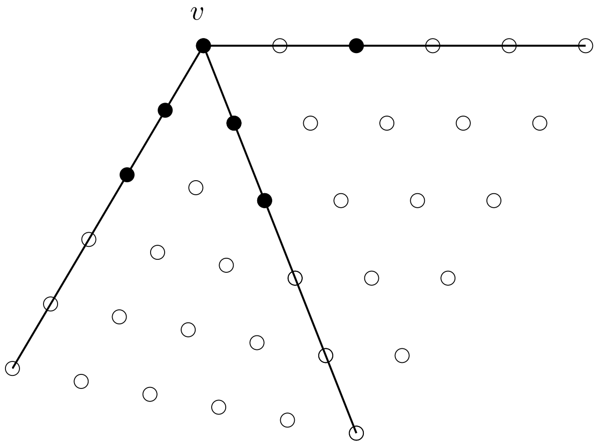

Figure 3.8: MDS points for $D_2(v)$, $v \in V_B$ for Argyris space.

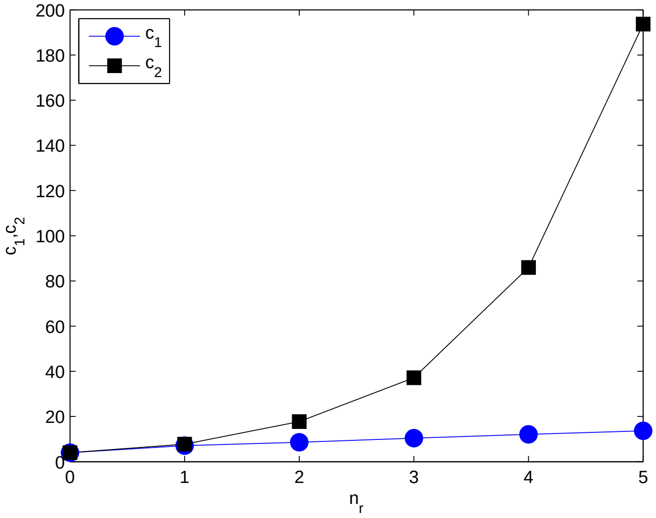

Figure 3.9: Comparison of stability of MDS M_A and $\tilde{M}$. c_1 and c_2 are defined in (3.2.20).

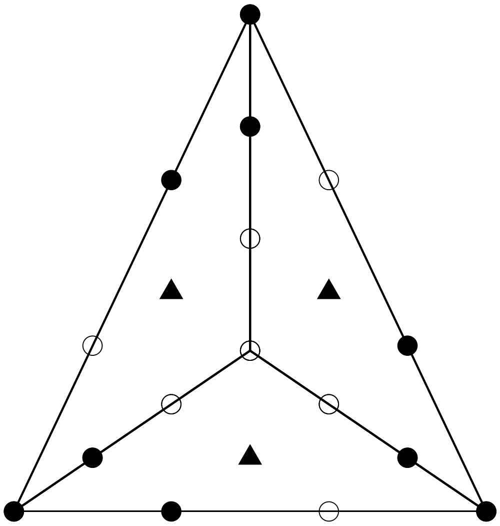

Figure 3.10: A typical Clough-Tocher refinement of one triangle with points in M_v marked as black dots and points in M_e marked as black triangles.

Tocher macro-element space $S_3^1(\triangle_{CT})$ as

$$M = \bigcup_{v \in V} M_v \cup \bigcup_{e \in E} M_e, \tag{3.3.21}$$

where $M_v := D_1(v) \cap T_v$ and $M_e := \left\{\xi_{111}^{T_e}\right\}$, and T_v and T_e are triangles in $\triangle_{CT}$. Denote by V and E the sets of vertices and edges in $\triangle$, respectively. Let

$$S_0 := \left\{ s \in S_3^1(\triangle_{CT}) \;:\; s|_{\partial\Omega} = 0 \right\}.$$

Let V_I and V_B be the sets of interior and boundary vertices of $\triangle$, respectively. We assume that T_v is a boundary triangle for each M_v, $v \in V_B$. Then stable splitting for M is possible as follows. Let

$$\left(\bigcup_{v \in V_I} M_v \cup \bigcup_{e \in E} M_e \right) \subset M_0. \tag{3.3.22}$$

However, M_0 may contain some more points from M_v, $v \in V_B$. Note that, for boundary vertices v, two points in M_v are always on the boundary and one is not.

These two boundary points are in M_b but the point ξ_v in M_v, which is not on the boundary, belongs to either M_0 or M_b depending on the geometry of boundary edges attached to v in the same way as the point ξ_v in Section 3.2.2. This point will be in M_0 for those boundary vertices where boundary edges are collinear. Otherwise it will be in M_b. Then we arrive at the following result.

Theorem 3.3.1. *$M := M_0 \cup M_b$ is a stable splitting of a minimal determining set M for Clough-Tocher macro-element space.*

Proof. Let $s \in S_3^1(\triangle_{CT})$ and $c_\xi = 0$ for all $\xi \in M_b$. Consider $E_v = \{e_1, e_2, \cdots, e_n\}$ for $v \in V_B$ then $D_1(v) \cap e_1 \subset M_v$ and $c_\xi = 0$ for $\xi \in D_1(v) \cap e_1$ by assumption. Moreover the C^1 smoothness condition means c_ξ vanishes for $\xi \in R_1(v) \cap e_n$ as well. Hence $c_\xi = 0$ for domain points on $\partial\Omega$ which results in $s|_{\partial\Omega} = 0$. Thus

$$\{s \in S_3^1(\triangle_{CT}) \,:\, c_\xi = 0 \;\forall \xi \in M_b\} \subset S_0.$$

Conversely, let $s \in S_0$; we need to show that all Bézier coefficients c_ξ for $\xi \in M_b$ vanish. Since $s|_{\partial\Omega} = 0$ therefore Bézier coefficients of s corresponding to domain points on the boundary are zero. For $v \in V_B$, where boundary edges are non-collinear, the gradient of s at v is also zero, due to C^1 smoothness of s at v, thus the coefficient of s at ξ_v is zero as well. This implies that

$$S_0 \subset \{s \in S_3^1(\triangle_{CT}) \,:\, c_\xi = 0 \;\forall \xi \in M_b\}.$$

Stability and locality follows as M is a stable local MDS for $S_3^1(\triangle_{CT})$. ■

3.3.2 Powell-Sabin Macro-element Space

Now for a given triangulation $\triangle$ of a domain Ω, let $\triangle_{PS}$ be the corresponding Powell-Sabin refinement [43, Definition 4.18], see the Figure 3.11. For each $v \in V$, let T_v be some triangle of $\triangle_{PS}$ attached to v, and $M_v := D_1(v) \cap T_v$. Then

$$M = \bigcup_{v \in V} M_v \tag{3.3.23}$$

is a stable local minimal determining set for Powell-Sabin space $S_2^1(\triangle_{PS})$ [43, Theorem 6.9]. Now similarly if

$$S_0 := \left\{ s \in S_2^1(\triangle_{PS}) \ : \ s|_{\partial\Omega} = 0 \right\}$$

and we take T_v to be a boundary triangle for M_v, $v \in V_B$, then we split M given in (3.3.23) for $S_2^1(\triangle_{PS})$ in M_0 and M_b as follows:

Let

$$\left(\bigcup_{v \in V_I} M_v \right) \subset M_0 \text{ and } (M_v \cap e_1) \subset M_b \; \forall v \in V_B, \tag{3.3.24}$$

where $E_v = \{e_1, e_2, \cdots, e_n\}$ is a set as defined in Section 3.2.1. The point ξ_v in $R_1(v) \cap e_2$ belongs to M_0 or M_b based on discussion made in Section 3.2.2. Then we formulate the following theorem.

Theorem 3.3.2. *$M := M_0 \cup M_b$ is a stable splitting of a minimal determinig set M for Powell-Sabin macro-element space.*

Proof. Follow the same arguments as in the proof of Theorem 3.3.1. ■

3.3.3 Powell-Sabin-12 Macro-element Space

Let $\triangle_{PS12}$ be the Powell-Sabin-12 refinement [43, Definition 4.21] of a given triangulation $\triangle$ of a domain Ω, see Figure 3.12. For each e of $\triangle$, let u_e be the midpoint of e and let v_T be the incenter of a triangle T in $\triangle$ attached to e. Let $\xi_e := \frac{v_T + u_e}{2}$ and $M_e := \{\xi_e\}$. For each vertex $v \in V$, let T_v be a triangle of $\triangle_{PS12}$ attached to v, and let $M_v := D_1(v) \cap T_v$. Then the set

$$M = \bigcup_{v \in V} M_v \cup \bigcup_{e \in E} M_e \tag{3.3.25}$$

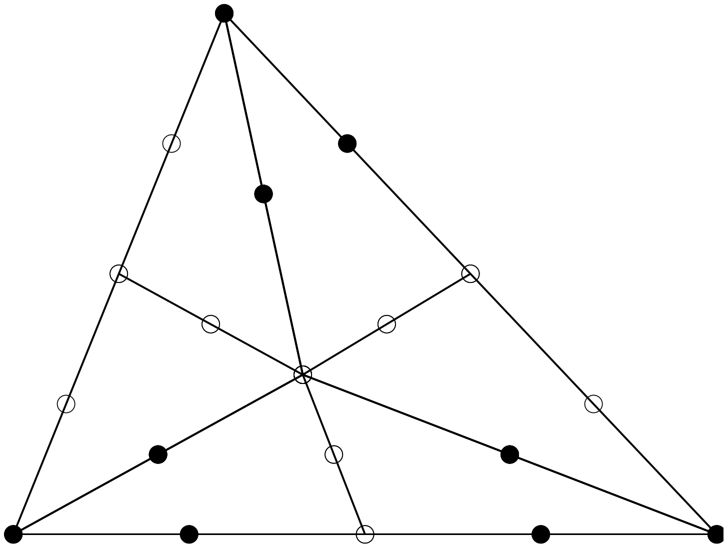

Figure 3.11: Powell-Sabin refinement of one triangle with points in M_v marked as black dots.

is a stable local MDS for the space $S_2^1(\triangle_{PS12})$ [43, Theorem 6.13]. Now let

$$S_0 := \left\{ s \in S_2^1(\triangle_{PS12}) \ : \ s|_{\partial\Omega} = 0 \right\}.$$

Again, we assume that T_v is a boundary triangle of $\triangle_{PS12}$ for any boundary vertex v. Let us split M into M_0 and M_b by the same method as for the Clough-Tocher elements. Then we conclude with the following theorem.

Theorem 3.3.3. *$M := M_0 \cup M_b$ is a stable splitting of a minimal determinig set M for $S_2^1(\triangle_{PS12})$.*

Proof. Follow the same arguments as in the proof of Theorem 3.3.1. ■

3.3.4 Quadrilateral Macro-element Space

Let $\diamondsuit$ be a strictly convex quadrangulation of a polygonal domain Ω and let $\triangle_Q$ be triangulation obtained by drawing in the diagonals of each quadrilateral of $\diamondsuit$.

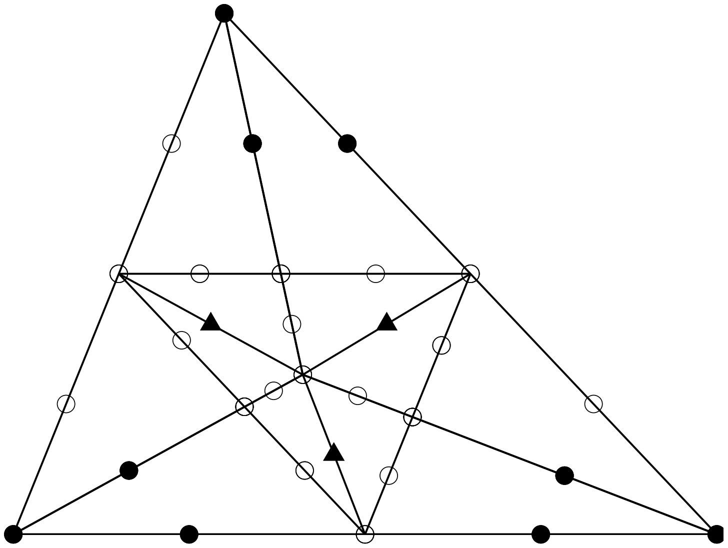

Figure 3.12: A Powell-Sabin-12 refinement of one triangle with points in M_v marked as black dots and points in M_e marked as black triangles.

Let V and E be the sets of vertices and edges of $\diamondsuit$. Here we will discuss the cubic spline space $S_3^1(\triangle_Q)$. Again let $M_v := D_1(v) \cap T_v$, for each $v \in V$, where T_v is a triangle in $\triangle_Q$ attached to v, and T_v is a boundary triangle in the case of a boundary vertex v. For each $e \in E$, let T_e be some triangle in $\triangle_Q$ containing e and let $M_e := \left\{\xi_{111}^{T_e}\right\}$. Then

$$M = \bigcup_{v \in V} M_v \cup \bigcup_{e \in E} M_e \tag{3.3.26}$$

is a stable local MDS for the space $S_3^1(\triangle_Q)$ [43, Theorem 6.17]. Again the splitting of M for $S_3^1(\triangle_Q)$ in M_0 and M_b can be achieved by following similar arguments as for the other C^1 macro-elements discussed above which results in the theorem formulated below.

Theorem 3.3.4. *$M := M_0 \cup M_b$ is a stable splitting of a minimal determining set M for $S_3^1(\triangle_Q)$.*

Proof. Follow the same lines as in the proof of Theorem 3.3.1. ■

Note that in [43, Section 6.5] the above triangle T_v is chosen such that it has the largest shape ratio $\mathrm{diam}(T)/\rho(T)$ among all triangles attached to v. This allows stable MDS even in the presence of small angles in $\triangle_Q$ if the smallest angle in $\diamondsuit$ is separated from zero. However, this choice of T_v might be unsuitable for stable splitting if v is a boundary vertex because we need T_v to be a boundary triangle whereas the shape ratio might be larger for some interior triangle attached to v. Therefore, our construction of stable splitting is valid only if $\triangle_Q$ satisfies the minimum angle condition.

Chapter 4

Numerical Solution of Fully Nonlinear Elliptic Equations by Böhmer's Method : Numerical Results

4.1 Introduction

In this chapter we present a first implementation of Böhmer's finite element method for fully nonlinear elliptic partial differential equations on convex polygonal domains, based on a modified Argyris element discussed in the previous chapter. Bernstein-Bézier techniques are our main tools in this implementation. Our numerical experiments for several test problems, including the classical Monge-Ampère equation and an unconditionally elliptic equation, confirm the convergence and error bounds predicted by Böhmer's theoretical results.

Recall that Böhmer's method [10, 11] permits solution of the Dirichlet problem for any fully nonlinear elliptic equations of second order. It is based on a finite element discretization of the linearised elliptic equations, with C^1 finite element spaces that admit a stable splitting into the subspace satisfying zero boundary conditions and its complement. Therefore we use a modified Argyris space $\tilde{S}$ with splitting $\tilde{M} := \tilde{M}_0 \cup \tilde{M}_b$, see Theorem 3.2.3, instead of classical Argyris space. Because we proved in last chapter that Argyris space does not admit a stable splitting of the basis functions. Full theoretical justification of the method can be seen in [10, 11], including a proof of convergence/stability and error bounds. However, no numerical results have been provided.

Our numerical experiments include several standard test problems for the Monge-Ampère equation on a square, an example for a non-rectangular convex polygonal domain, and an unconditionally elliptic equation. The numerical results confirm the theoretical error bounds given in [10, 11].

The chapter is organised as follows. In Section 4.2 we provide details of the implementation of Böhmer's method, including the assembly of the system matrix for the linearised elliptic equations arising in each step of Newton iteration. Finally, Section 4.3 is devoted to the numerical experiments.

As we will see in the next section, a key step in the implementation of the finite element stiffness matrices using Bernstein-Bézier techniques is the computation of the Bézier coefficients of the basis splines $\{s_\xi\}_{\xi \in M}$ corresponding to an MDS M. We therefore conclude this section by providing Algorithm 1 that gives all details of this computation for the basis splines of the modified Argyris space.

Algorithm for computing the Bernstein-Bézier coefficients

The algorithm to compute all Bézier coefficients of $s \in S$, the Argyris space, for given coefficients $\{c_\eta : \ \eta \in M\}$ can be extracted from the proof of [43, Theorem 6.1]. The algorithm for modified Argyris space $\tilde{S}$, using MDS $\tilde{M}$, is similar except at boundary vertices. We formulate Algorithm 1 to compute Bézier coefficients of basis splines $\{s_\xi\}_{\xi\in\tilde{M}}$ given that

$$c_\xi = 1, \xi \in \tilde{M} \text{ and } c_\eta = 0, \eta \in \tilde{M}\backslash\{\xi\}.$$

4.2 Implementation of Böhmer's Method

In this section we describe in detail our implementation of Böhmer's method using Bernstein-Bézier techniques. We study the numerical approximation of Dirichlet problem (2.2.13)-(2.2.14) for a fully nonlinear equation of second order.

Discretization

Recall that $\triangle^h$ is a quasi-uniform triangulation of a convex polygonal domain $\Omega \subset \mathbb{R}^2$. As discussed in Section 2.2.2, solving the nonlinear problem (2.2.13)-(2.2.14) by Böhmer's method amounts to running a Newton-Kantorovich iteration scheme, on each level of triangulation, to get a sequence $\left\{u_k^h\right\}_{k\in\mathbb{Z}_+}$ of approximations of $\hat{u}$ generated by

$$u_{k+1}^h = u_k^h - w^h, \quad k = 0, 1, \ldots, \tag{4.2.1}$$

where $w^h \in S_0^h$ is the solution of the linear elliptic problem: Find $w^h \in S_0^h$ such that

$$(G'(u_k^h)w^h, v^h)_{L^2(\Omega)} = (G(u_k^h), v^h)_{L^2(\Omega)} \; \forall v^h \in S_0^h, \tag{4.2.2}$$

where G' is the linearisation (2.1.2) of the nonlinear operator G. We solve this linear equation for w^h, for a given u_k^h, by using the standard Galerkin finite element method with the modified Argyris space $\tilde{S}^h$ on $\triangle^h$ as an approximating space, with the stable splitting $\tilde{S}^h = \tilde{S}_0^h + \tilde{S}_b^h$ according to Theorem 3.2.3.

Algorithm 1 Compute Bézier coefficients of a basis spline s_ξ, $\xi \in \tilde{M}$.

```
Require: Given $\xi$, initialize $\{c_\eta : \eta \in D_{5,\triangle}\}$ by zeros and set $c_\xi = 1$.
Ensure: Compute $c_\eta \ \forall \eta \in D_\triangle \backslash \tilde{M}$.
 1. if $\xi \in M_v,\ v \in V_I$ then
 2.     Find triangles $\{T_\kappa\}_{\kappa=1}^{k}$ attached to vertex $v$, arranged in anti-clockwise order, with $T_1 := T_v$.
 3.     Move anti-clockwise by computing $c_\nu, \nu \in D_2(v) \cap T_{\kappa+1}$ from known coefficients $c_\eta, \eta \in D_2(v) \cap T_\kappa,\ \kappa = 1, \cdots, k-1$, using $C^1$ and $C^2$ smoothness conditions [43, Lemma 2.30] and see Section 3.1.5.
 4.     For each edge $e \in E_v$. Let the edge $e := \langle v, v_1 \rangle$ be shared by triangles $T_e := \langle v, v_2, v_1 \rangle$ and $\tilde{T}_e := \langle v_3, v, v_1 \rangle$. Since $c_{302}$ is known, for $\xi_{302} \in D_2(v)$, we compute $c_{122}^{\tilde{T}_e}$ using $C^1$ smoothness condition over $e$
            $$c_{122}^{\tilde{T}_e} = b_1 c_{302},$$
        where $(b_1, b_2, b_3)$ are barycentric coordinates of $v_3$ w.r.to $T_e$.
 5. else if $\xi \in \tilde{M}_v,\ v \in V_B$ then
 6.     Do as in 2) by choosing $T_1 := T_v$ be one of the boundary triangles attached to $v$.
 7.     Compute $c_\nu, \nu \in \{D_2(v) \cap T_{\kappa+1}\} \backslash \tilde{M}_v$ from known coefficients $c_\eta, \eta \in D_2(v) \cap T_\kappa,\ \kappa = 1, \cdots, k-1$, again using $C^1$ smoothness conditions, see Section 3.1.5.
 8.     Do as in 4) only for $e \in E_v \backslash E_B$.
 9. else if $\xi \in M_e,\ e \in E_I$ then
10.     Let the edge $e := \langle v_1, v_2 \rangle$ is shared by triangles $T_e := \langle v_1, v_4, v_2 \rangle$ and $\tilde{T}_e := \langle v_3, v_1, v_2 \rangle$. Then $\xi := \xi_{212}^{T_e}$ and we compute $c_{122}^{\tilde{T}_e}$ with the help of $c_{212}^{T_e} = 1$ using $C^1$ smoothness condition over $e$ given by $c_{122}^{\tilde{T}_e} = b_3$, where $(b_1, b_2, b_3)$ are barycentric coordinates of $v_3$ w.r.to $T_e$.
11. end if
```

After a standard transformation to the weak form, (4.2.2) is translated into the following problem: Find $w^h \in \tilde{S}^h_0$ such that for all $v^h \in \tilde{S}^h_0$,

$$\int_\Omega \nabla w^h \cdot A \nabla v^h dx + \int_\Omega v^h b \cdot \nabla w^h dx + \int_\Omega c w^h v^h dx = \int_\Omega f v^h dx, \tag{4.2.3}$$

where $A = \left[\frac{\partial \widetilde{G}}{\partial r_{ij}}(u^h_k)\right]^2_{i,j=1}$, $b = \left[\frac{\partial \widetilde{G}}{\partial p_i}(u^h_k)\right]^2_{i=1}$, $f = G(u^h_k)$ and $c = \frac{\partial \widetilde{G}}{\partial z}(u^h_k)$.

If $(s_1, \dots, s_{N_0})$ is a basis of $\tilde{S}^h_0$, then, as usual in the finite element method, (4.2.3) results in the linear system

$$(\mathcal{S} + \mathcal{B}^t + \mathcal{M})a = \mathcal{L}, \tag{4.2.4}$$

where a is the vector of the coefficients in the expansion $w^h = \sum_{i=1}^{N_0} a_i s_i$, and $\mathcal{S}$, $\mathcal{B}$, $\mathcal{M}$ and $\mathcal{L}$ are the stiffness, convection and mass matrices and the load vector, respectively, with the entries, for $i, j = 1, \dots, N_0$, defined as

$$\mathcal{S}_{ij} = \int_\Omega \nabla s_i \cdot A \nabla s_j dx, \; \mathcal{B}_{ij} = \int_\Omega s_j b \cdot \nabla s_i dx, \; \mathcal{M}_{ij} = \int_\Omega c s_i s_j dx, \; \mathcal{L}_i = \int_\Omega f s_i dx.$$

It is worth emphasising that we do not use these formulae directly to compute the system matrices. Before we describe how we compute them let us define a transformation matrix $\mathcal{T}$ required for this.

Transformation Matrix

Let $\{T_\kappa\}_{\kappa=1}^{N_t}$ be the triangles in $\triangle^h$ with some fixed ordering. To define transformation matrix we make use of vector $V(s)$, defined in (3.2.19), for any spline $s \in \tilde{S}^h$. If we construct a matrix by taking these vectors $\mathbf{V}(s_i)$ for the basis splines $s_1, \dots, s_{N_0}$ as its rows, then this matrix is our desired *transformation matrix* $\mathcal{T}$,

$$\mathcal{T} = [\mathbf{V}(s_1)^t, \dots, \mathbf{V}(s_{N_0})^t]^t, \tag{4.2.5}$$

where ^{t} denotes a transpose of a matrix. Let $\tilde{S}_5(\triangle^h)$ denote the space of all discontinuous quintic splines over the same triangulation $\triangle^h$. Clearly, $\mathcal{T}^t$ represents the transformation that maps the vector $\{c_\xi\}_{\xi\in\tilde{M}}$ corresponding to $s \in \tilde{S}_0^h$ onto the array of the coefficients of s in the basis of the space $\tilde{S}_5(\triangle^h)$ defined by the quintic Bernstein basis polynomials B_{ijk}^5 on all triangles.

Now consider the block matrices $\hat{\mathcal{S}} = \operatorname{diag}\left(\hat{\mathcal{S}}_{T_\kappa}, T_\kappa \in \triangle^h\right)$, $\hat{\mathcal{B}} = \operatorname{diag}\left(\hat{\mathcal{B}}_{T_\kappa}, T_\kappa \in \triangle^h\right)$ and $\hat{\mathcal{M}} = \operatorname{diag}\left(\hat{\mathcal{M}}_{T_\kappa}, T_\kappa \in \triangle^h\right)$ with blocks defined by

$$\hat{\mathcal{S}}_{T_\kappa} = \int_{T_\kappa} \nabla B_{ijk}^5 \cdot A \nabla B_{rst}^5 dx,\ \hat{\mathcal{B}}_{T_\kappa} = \int_{T_\kappa} B_{ijk}^5 b \cdot \nabla B_{rst}^5 dx,\ \hat{\mathcal{M}}_{T_\kappa} = \int_{T_\kappa} c B_{ijk}^5 B_{rst}^5 dx. \tag{4.2.6}$$

Then we can compute the system matrices in (4.2.4) by using the relations

$$\mathcal{S} = \mathcal{T}\hat{\mathcal{S}}\mathcal{T}^t,\ \mathcal{B} = \mathcal{T}\hat{\mathcal{B}}\mathcal{T}^t,\ \mathcal{M} = \mathcal{T}\hat{\mathcal{M}}\mathcal{T}^t. \tag{4.2.7}$$

Note that this method of computing the system matrices is particularly efficient as it is shown in [1] that the matrices $\hat{\mathcal{S}}$, $\hat{\mathcal{B}}$ and $\hat{\mathcal{M}}$ can be computed in optimal complexity (constant cost per entry) even for high polynomial orders, and the matrix $\mathcal{T}$ is sparse because the basis splines are locally supported.

Boundary Conditions

As discussed in Section 2.2.2, in order to impose the non-homogeneous boundary conditions we require that the initial guess $u_0^h \in \tilde{S}^h$, satisfy the following condition

$$(u_0^h, v_b^h)_{L^2(\partial\Omega)} = (\phi, v_b^h)_{L^2(\partial\Omega)} \quad \forall v_b^h \in \tilde{S}_b^h.$$

Now if $(s_1, \ldots, s_{N_0}, s_{N_0+1}, \ldots, s_N)$ is the $\tilde{M}$-basis for the space $\tilde{S}^h$ and $(s_{N_0+1}, \ldots, s_N)$ is a basis for $\tilde{S}_b^h$, then the above boundary condition, following the usual procedure, is reduced to the matrix equation

$$\mathcal{M}_b \mathcal{C}_b = \mathcal{L}_b,$$

where

$$\mathcal{M}_b = \left[\int_{\partial\Omega} s_i s_j ds\right]_{i,j=N_0+1}^{N} \text{ and } \mathcal{L}_b = \left[\int_{\partial\Omega} \phi s_i ds\right]_{i=N_0+1}^{N}.$$

It is important to mention that $s_i|_e,\ e \in E$, are univariate polynomials and they keep the univariate BB-form [43, Remark 2.4]. Moreover, there is an explicit formula for integration of product of polynomials in BB-form given by

$$\int_e s_i s_j ds = \frac{|e|}{11} \sum_{\substack{\alpha=0\\ \beta=0}}^{5} c_\alpha c'_\beta \frac{\binom{5}{\alpha}\binom{5}{\beta}}{\binom{10}{\alpha+\beta}},$$

where $|e|$ is the length of e,

$$s_i|_e = \sum_{\alpha=0}^{5} c_\alpha B_\alpha^5 \text{ and } s_j|_e = \sum_{\beta=0}^{5} c'_\beta B_\beta^5,$$

with $B_\alpha^5 = \binom{5}{\alpha} t^\alpha (1-t)^{5-\alpha},\ \alpha = 0, \ldots, 5$, being the univariate quintic Bernstein polynomials on the edge e. Thanks to BB-form that helps us compute entries for $\mathcal{M}_b$ exactly but the presence of the function ϕ in the integrals for $\mathcal{L}_b$ forces us to use an appropriate quadrature rule. For this we see that

$$\int_e \phi s_i ds = \int_e \phi \sum_{\alpha=5} c_\alpha B_\alpha^5 ds = \sum_{\alpha=5} c_\alpha \int_e \phi B_\alpha^5 ds. \tag{4.2.8}$$

Thus, computing the entries for $\mathcal{L}_b$ is reduced to approximating the *Bernstein-Bézier moments* $\mu_\alpha^5(\phi) = \int_e \phi B_\alpha^5 ds$ of ϕ using an appropriate quadrature rule [1]. We use the Gauss-Legendre 6-point rule to approximate the moments $\mu_\alpha^5(\phi)$ which returns the exact solution for polynomials of order up to 11. Note that, unlike using C^0 elements, here some degrees of freedom for $\tilde{S}_b^h$ lie inside the domain Ω, see Theorem 3.2.3. Thus it would be difficult to impose the boundary conditions merely by interpolating the function ϕ at the points corresponding to the degrees of freedom lying on the boundary.

4.3 Numerical Results

This section is devoted to the numerical results for several fully nonlinear problems, involving the Monge-Ampère equation and an unconditionally elliptic problem considered in [42]. The numerics for these problems confirm the convergence and the theoretical error bounds of Theorem 2.2.2.

4.3.1 The Monge-Ampère Equation

The Dirichlet problem for the *Monge-Ampère equation* is given by

$$\begin{aligned} G_{MA}(u) = \det(\nabla^2 u) - g(x) = & \ \ 0, \quad x \in \Omega \\ u = & \ \ \phi, \ x \in \partial\Omega \end{aligned} \tag{4.3.9}$$

where g and ϕ are given functions with $g > 0$ on Ω required to keep the problem elliptic. The weak formulation (4.2.3) of the linearised problem in this case is to find $w^h \in S_0^h$ such that

$$\int_\Omega \nabla w^h \cdot A \nabla v^h dx = \int_\Omega f v^h dx, \text{ for all } v^h \in S_0^h, \tag{4.3.10}$$

with $A = \mathrm{cof}(\nabla^2 u_k^h)$ as $b = 0$, $c = 0$ and $f = G_{MA}(u_k^h) = \det(\nabla^2 u_k^h) - g(x)$, where $\mathrm{cof}(M)$ denotes the cofactor of a 2×2 matrix M. As a result we are left with the stiffness matrix and load vector to solve the linear system

$$\mathcal{SC} = \mathcal{L},$$

for the unknown vector of Bézier coefficients $\mathcal{C}$.

As the Monge-Ampère equation is elliptic only for convex functions, we need the initial guess to be convex as well. In [24, Remark 2.1] it has been shown that (4.3.9) and the Poisson-Dirichlet problem

$$\begin{aligned} \Delta u = & \ \ 2\sqrt{g}, \ x \in \Omega \\ u = & \ \ \ \phi, \ x \in \partial\Omega \end{aligned} \tag{4.3.11}$$

are closely related. Therefore we use the approximation solution of the Poisson-Dirichlet problem (4.3.11) as an initial guess for the Newton scheme (4.2.1). The initial guess obtained this way performs very well in our experiments. However, we get much faster convergence of the Newton method by using this initial guess only on the initial mesh, whereas on the refined meshes we take a quasi-interpolant [43, Section 5.7] of the solution from the previous level as an initial guess. We call this a *multilevel approach*.

The first three and the fifth test problems are standard benchmark problems for (4.3.9) over $\Omega = (0,1)^2$ considered in many papers on the numerical solution of the Monge-Ampère equation. In this case $\triangle^h$ is the uniform triangulation obtained by first dividing the domain into squares of side length h and then drawing in the diagonals parallel to the line $x_2 = x_1$. In the fourth test problem a non-rectangular domain is considered.

1. As the *first test problem* we solve (4.3.9) for the data

$$\begin{aligned} g(x) &= (1+|x|^2)e^{|x|^2}, \text{ in } \Omega, \\ \phi(x) &= e^{\frac{1}{2}|x|^2} \quad \forall x \in \partial\Omega, \end{aligned}$$

   where $|x| = \sqrt{x_1^2 + x_2^2}$. With this data the exact solution to the problem is $u(x) = e^{\frac{1}{2}|x|^2} \in C^\infty(\overline{\Omega})$. The numerical results are presented in Table 4.1. They confirm the convergence rate $O(h^4)$ in the H^2-norm predicted by Theorem 2.2.2, where $\ell = 6$ as we are using polynomials of degree 5. Moreover, as expected, we observe the convergence rates of $O(h^6)$ and $O(h^5)$ in the L^2 and H^1 norms, respectively. The first row of the table shows the errors for the initial guess. In addition to the errors, Table 4.1 presents the number of Newton iterations (N) on each mesh, the L^2-norm of the residuals $r := \|G(u_k^h)\|_{L^2(\Omega)}$, and the size $\|p\|_{L^2(\Omega)}$ of the L_2-projection p of $G(u_k^h)$ on the space $\tilde{S}_0^h$. The projection p is found as a solution of the system $\mathcal{M}\mathcal{C}_p = \mathcal{L}$, where $\mathcal{M}$ is the mass matrix, $\mathcal{L}$ is a load vector and $\mathcal{C}_p$ is the vector of coefficients of the expansion of p in the $\tilde{M}_0$-basis. The size of the projection measures how well the approximate solution u_k^h solves the problem (2.2.16). We observe that the number of Newton iterations is extremely small thanks to the fact that the initial guess is chosen by the multilevel approach. The size of the residual is close to the H^2-norm error, as one can expect, and the size of the projection is close to the unit round-off initially, and gets larger on further refinement levels, obviously due to growing condition numbers of the system matrices.

Table 4.1: Errors of approximate solution and rate of convergence for the first test problem, N denotes the number of Newton's iterations, $r := \|G(u_k^h)\|_{L^2(\Omega)}$ is the size of the residual, and $\|p\|_{L^2(\Omega)}$ is the size of the L_2-projection of $G(u_k^h)$ on $\tilde{S}_0^h$.

h	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	5.78e-3		3.25e-2		2.66e-1			9.64e-1	
1	1.17e-4		1.03e-3		1.74e-2		2	5.15e-2	2.30e-15
1/2	4.77e-6	4.6	7.75e-5	3.7	2.25e-3	3.0	1	5.14e-3	1.74e-14
1/4	1.92e-7	4.6	7.04e-6	3.5	3.32e-4	2.8	1	8.28e-4	9.44e-14
1/8	2.42e-9	6.3	1.65e-7	5.4	1.58e-5	4.4	1	3.93e-5	3.89e-13
1/16	4.31e-11	5.8	6.61e-9	4.6	1.20e-6	3.7	1	3.56e-6	1.79e-12
1/32	6.60e-13	6.0	1.95e-10	5.1	7.45e-8	4.0	1	2.04e-7	7.38e-12
1/64	1.14e-14	5.9	7.28e-12	4.7	6.06e-9	3.6	1	1.66e-8	2.83e-11
1/128	8.16e-15	0.5	2.96e-13	4.6	3.73e-10	4.0	1	1.07e-9	1.06e-10

2. *Second test problem* is defined by

$$\begin{aligned} g(x) &= \frac{R^2}{(R^2-|x|^2)^2} \;\; \forall x \in \Omega, \;\text{ with } R \geq \sqrt{2}, \\ \phi(x) &= -\sqrt{R^2-|x|^2} \quad \forall x \in \partial\Omega, \end{aligned}$$

in (4.3.9). The exact solution is $u(x) = -\sqrt{R^2 - |x|^2}$. The function $g(x)$ has a singularity at $R = \sqrt{2}$ and $u \in W_p^1(\Omega)$, $1 \le p < 4$ for this value of R, lacking H^2-regularity. The method diverges for $R = \sqrt{2}$, in line with Böhmer's theory that guarantees convergence only if the solution is in H^2. But for $R > \sqrt{2}$ we have $u \in C^\infty(\overline{\Omega})$ and again, in Table 4.2 and Table 4.3 for two different values of R, the results show the same behaviour as in the first problem. The tables confirm that the further the value of R is away from singularity the faster convergence is achieved. Note that in this experiment much higher accuracy is attained as compared to the results in [24] for the same test problem.

Table 4.2: Errors of approximate solution and rate of convergence for the second test problem with $R = \sqrt{2} + .1$. The meaning of the last three columns is the same as in Table 4.1.

h	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	2.00e-3		1.67e-2		2.69e-1			1.02e0	
1	2.34e-3		1.25e-2		2.15e-1		2	5.91e-1	1.92e-15
1/2	1.70e-4	3.8	1.57e-3	3.0	7.32e-2	1.6	2	1.68e-1	7.89e-15
1/4	6.01e-6	4.8	1.58e-4	3.3	1.75e-2	2.1	2	3.80e-2	2.92e-14
1/8	1.72e-7	5.1	1.31e-5	3.6	3.17e-3	2.5	1	6.61e-3	1.34e-13
1/16	3.92e-9	5.4	8.10e-7	4.0	4.05e-4	3.0	1	8.44e-4	5.04e-13
1/32	1.02e-10	5.3	3.71e-8	4.4	3.53e-5	3.5	1	7.23e-5	2.07e-12
1/64	1.93e-12	5.7	1.41e-9	4.7	2.80e-6	3.7	1	5.49e-6	8.45e-12

Table 4.3: Errors of approximate solution and rate of convergence for the second test problem with $R = \sqrt{2} + 2$.

h	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	1.34e-5		7.38e-5		6.07e-4			1.95e-4	
1	7.66e-7		5.89e-6		8.20e-5		2	2.64e-5	1.84e-15
1/2	1.28e-8	5.9	2.50e-7	4.6	7.85e-6	3.4	1	2.49e-6	7.68e-15
1/4	4.33e-10	4.9	1.72e-8	3.9	8.65e-7	3.2	1	2.58e-7	2.97e-14
1/8	6.66e-12	6.0	4.94e-10	5.1	9.78e-8	4.1	1	1.46e-8	1.49e-13
1/16	1.10e-13	5.9	1.75e-11	4.8	3.36e-9	3.8	1	1.00e-9	5.71e-13
1/32	7.67e-15	3.6	5.53e-13	4.9	2.12e-10	3.9	1	6.17e-11	2.25e-12

3. *Third test problem* is defined by

$$\begin{aligned} g(x) &= \frac{1}{|x|} \; \forall x \in \Omega, \\ \phi(x) &= \frac{(2|x|)^{\frac{3}{2}}}{3} \; \forall x \in \partial\Omega. \end{aligned}$$

   for the Monge-Ampère problem (4.3.9). The difference to the previous problems is that the exact solution $u(x) = \dfrac{(2|x|)^{\frac{3}{2}}}{3}$ is not infinitely differentiable, even $u \notin C^2(\Omega)$. However, as mentioned in [24], $u \in W^2_p$, for $1 \leq p < 4$. It shows that $u \in H^\gamma(\Omega)$ for some $\gamma > 2$. The results in Table 4.4 shows the convergence of $O(h^{\frac{5}{2}})$ in L_2-norm which is an indication that $u \in H^\gamma(\Omega)$ with γ approaching $\frac{5}{2}$.

4. *Fourth test problem.* This problem is different from the others because we consider a non-rectangular domain Ω, as Böhmer's method is applicable to

Table 4.4: Errors of approximate solution and rate of convergence for third test problem.

h	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	6.08e-3		2.99e-2		4.02e-1			2.23e-2	
1	8.18e-4		1.21e-2		3.87e-1		2	2.28e-2	1.23e-16
1/2	1.95e-4	2.1	4.55e-3	1.4	2.77e-1	0.48	2	1.13e-2	3.53e-16
1/4	6.76e-5	1.5	1.73e-3	1.4	1.95e-1	0.50	2	1.49e-2	1.54e-15
1/8	1.65e-5	2.0	6.40e-4	1.4	1.36e-1	0.51	2	3.68e-2	5.53e-15
1/16	3.46e-6	2.3	2.30e-4	1.5	9.44e-2	0.53	2	8.47e-2	2.50e-14
1/32	6.75e-7	2.4	8.08e-5	1.5	6.33e-2	0.57	2	1.82e-1	9.76e-14

any convex polygonal domain. Let Ω be bounded by the straight lines

$$x_1 = \pm 0.75,\ x_2 = \pm 0.75,\ \text{and } |x_2| - |x_1| = 1,$$

see Figure 4.1 (left), which also includes the initial triangulation. We generate a sequence of meshes by the uniform refinement, where each triangle is split into 4 similar subtriangles. This test problem is for (4.3.9) with the same data as in first test problem. Again we choose an initial guess by the multilevel approach and use a solution of (4.3.11) on the first level. The numerics again show the same rate of convergence as for the rectangular domains, see Table 4.5. The graph of approximate solution u^h on the last level of triangulation is visualised in Figure 4.1 (right).

5. *Fifth test problem.* Here we consider a homogeneous Dirichlet problem for (4.3.9) with $g = 1$ over $\Omega = [0, 1]^2$. This test problem is interesting because it does not have a smooth classical solution. Clearly, Theorem 2.2.2 does

Table 4.5: Errors of approximate solution and rate of convergence for the fourth test problem.

Levels	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	9.30e-4		3.96e-3		3.58e-2			4.39e-2	
1st	5.01e-7		8.39e-6		3.94e-4		2	5.83e-4	1.56e-14
2nd	1.18e-8	5.4	3.45e-7	4.6	2.87e-5	3.8	1	3.97e-5	6.47e-14
3rd	2.11e-10	5.8	1.11e-8	4.9	1.91e-6	3.9	1	2.67e-6	2.76e-13
4th	3.54e-12	5.9	3.36e-10	5.1	1.33e-7	3.8	1	1.85e-7	1.12e-12
5th	4.36e-14	6.3	1.12e-11	4.9	8.79e-9	3.9	1	1.20e-8	4.65e-12
6th	4.85e-14	-0.2	5.00e-13	4.5	5.69e-10	3.9	1	8.12e-10	1.82e-11

not apply in this case. Nevertheless, we applied the algorithm and noticed the convergence of the Newton method on coarse levels, until $h = \frac{1}{4}$, but when we moved to more refined meshes we did not see convergence any more even if we used the multilevel approach. Let $RF = G(u_k^h)$ be the residual function where u_k^h is the approximate solution. The cross section of RF along the straight line $x_2 = x_1$ is depicted in Figure 4.2 that shows the strong singularity at the corners. It can also be seen that the method tries to converge inside of the domain away from the singularity. Also see test 3 in Chapter 6 where we consider the same problem over domains with smooth boundaries. The approximate solution u^h and its contour plot on a mesh with $h = \frac{1}{4}$ is visualized in Figure 4.3.

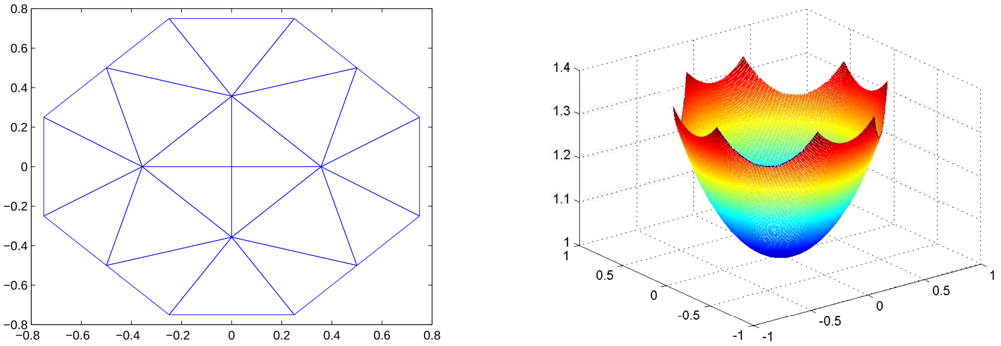

Figure 4.1: Non-rectangular domain Ω for fourth test problem with initial triangulation (*left*) and approximate solution u^h on the last level of triangulation (*right*).

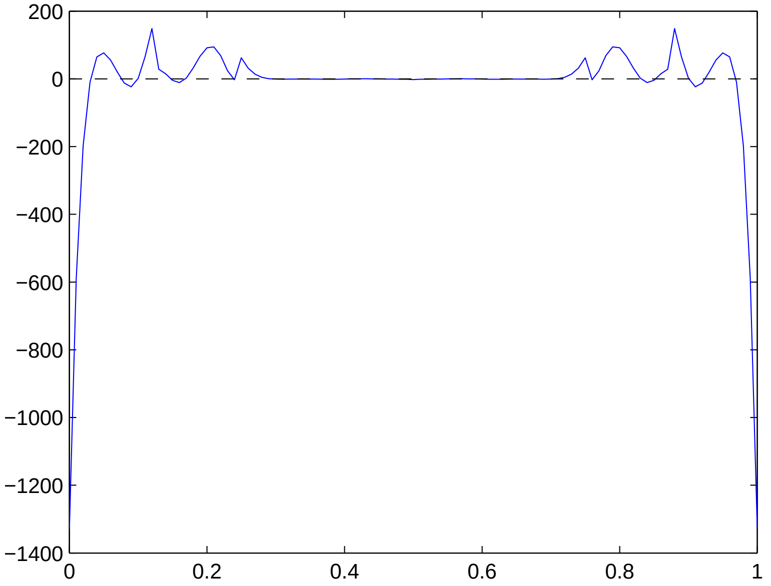

Figure 4.2: Cross section of $RF = G(u_k^{\frac{h}{4}})$ along the straight line $x_2 = x_1$.

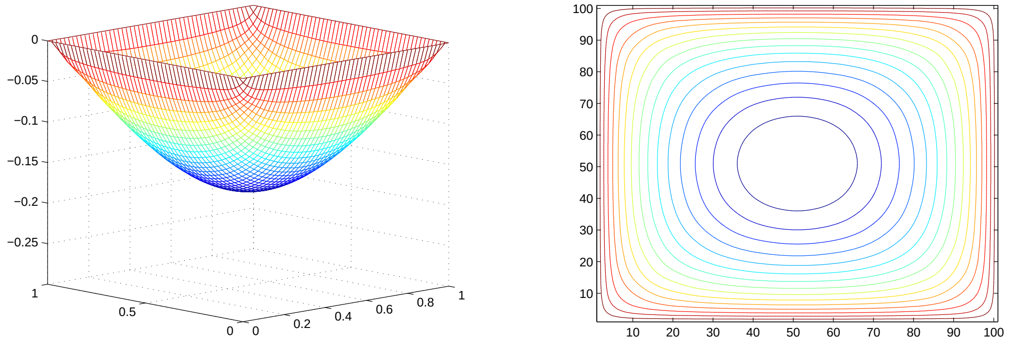

Figure 4.3: Approximate solution u^h of test 5 and its contour plot, $h = \frac{1}{4}$

4.3.2 Second Example

Consider the problem suggested in [42]

$$\begin{aligned} G_2(u) = u_{11}^3 + u_{22}^3 + u_{11} + u_{22} - g(x) = &\quad 0, \quad x \in \Omega \\ u = &\quad \phi, \; x \in \partial\Omega \end{aligned} \tag{4.3.12}$$

where $u_{ii} = (\partial^i)^2 u$, $i = 1, 2$. This problem is unconditionally elliptic, i.e. the operator G_2 is elliptic for any function $u \in D(G_2) = C^2(\Omega)$. Note that Condition H of [11] is satisfied in this example. The last of our test problems is for (4.3.12) in the domain $\Omega = [-1, 1]^2$, with the data given by

$$\begin{aligned} g(x) &= ((4x_1^2 - 2)^3 + (4x_2^2 - 2)^3)e^{-3|x|^2} + (4|x|^2 - 4)e^{-|x|^2}, \; \forall x \in \Omega, \\ \phi(x) &= e^{-|x|^2} \; \forall x \in \partial\Omega. \end{aligned}$$

The matrix A in this case is

$$A = \begin{bmatrix} 3u_{11}^2 + 1 & 0 \\ 0 & 3u_{22}^2 + 1 \end{bmatrix}$$

and $b = 0$, $c = 0$. Note that A is strictly positive definite for any function u. The triangulations $\triangle^h$ with side length h are generated the same way as for $\Omega = [0, 1]^2$ in Section 4.3.

To find an initial guess for the Newton method on the initial triangulation $\triangle^2$ we use the approximate solution of the Laplace-Dirichlet problem

$$\begin{aligned} \Delta u = &\quad 0, \quad x \in \Omega, \\ u = &\quad \phi,\ x \in \partial\Omega, \end{aligned} \tag{4.3.13}$$

whereas on the subsequent refinement levels we use the multilevel approach as described in Section 4.3. Note that the method was divergent with initial guess generated by (4.3.13) for $h \leq \frac{1}{2}$.

The numerical results are presented in Table 4.6. We see a very slow convergence of Newton's iterations in this example, compare N in Tables 4.1–4.6. The theoretical convergence rate of Böhmer's method is, however, as expected as we see that $\|u - u^h\|_{H^2(\Omega)} = O(h^4)$. We also observe the difference in the behaviour of $\|p\|_{L^2(\Omega)}$, which seems to indicate that Newton method does not find a solution of (2.2.16). This phenomenon requires further investigation.

Table 4.6: Errors of approximate solution and rate of convergence for the sixth test problem.

h	L^2-error	rate	H^1-error	rate	H^2-error	rate	N	r	$\|p\|_{L^2(\Omega)}$
initial	3.32e-1		7.21e-1		1.62e0			5.81e0	
2	2.85e-2		1.56e-1		9.72e-1		17	6.43e0	1.01e0
1	5.12e-4	5.8	4.33e-3	5.2	6.09e-2	4.0	10	1.45e-1	1.03e-3
1/2	1.76e-5	4.9	2.48e-4	4.1	5.72e-3	3.4	12	2.19e-2	2.16e-5
1/4	2.21e-7	6.3	5.52e-6	5.5	2.70e-4	4.4	11	1.27e-3	1.07e-7
1/8	3.07e-9	6.2	1.63e-7	5.1	1.58e-5	4.1	10	9.66e-5	8.29e-10
1/16	5.37e-11	5.8	4.95e-9	5.0	9.50e-7	4.1	12	5.68e-6	7.63e-12
1/32	8.21e-13	6.0	1.56e-10	4.9	6.03e-8	3.9	12	3.50e-7	8.01e-12
1/64	7.40e-14	3.5	4.86e-12	5.0	3.72e-9	4.0	9	2.16e-8	3.15e-11

Chapter 5

H^1 Polynomial Finite Element Method for Domains Enclosed by Piecewise Conics

5.1 Introduction

Let $\Omega \subset \mathbb{R}^2$ be a bounded curvilinear polygon with $\Gamma = \partial\Omega = \bigcup_{j=1}^{m} \overline{\Gamma_j}$, where each Γ_j is an open arc of an algebraic curve of at most second order (*i.e.* either a straight line or a conic). Let $Z = \{z_1, \dots, z_m\}$ be the set of the endpoints of all arcs numbered counter-clockwise such that z_j, z_{j+1} are the endpoints of Γ_j, $j = 1, \dots, m$ (we set $z_{j+m} = z_j$). Furthermore, for each j we denote by ω_j the internal angle between the tangents τ_j^+ and τ_j^- to Γ_j and Γ_{j-1}, respectively, at z_j. We assume that $0 < \omega_j \leq 2\pi$, see Figure 5.1 and set $\omega := \min\{\omega_j \,:\, 1 \leq j \leq m\}$.

The purpose of this chapter is to develop an H^1 conforming finite element method with polynomial shape functions suitable for solving second order elliptic problems for curvilinear polygons of the above type.

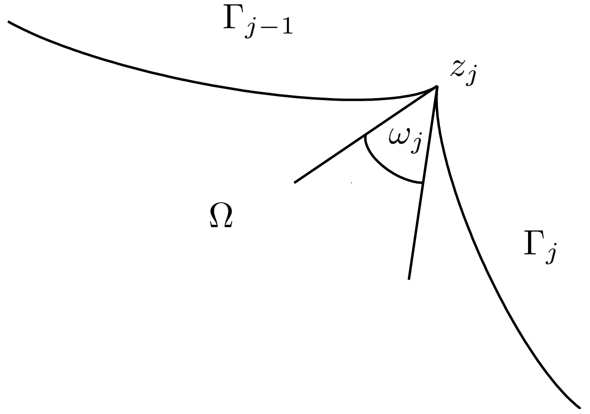

Figure 5.1: Definition of ω_j.

Let us now formulate some of the problems. For simplicity we restrict ourselves to elliptic problems with Dirichlet boundary conditions and consider in detail

1) the case when the corresponding variational problem

$$\begin{aligned} &u \in H^1(\Omega), \\ &a(u,v) = (f,v), \qquad \text{all } v \in H_0^1(\Omega), \end{aligned} \tag{5.1.1}$$

with *bounded* and *coercive* bilinear form $a(\cdot,\cdot)$ and such that the *regularity condition*

$$\|u\|_{H^2(\Omega)} \le C_R \|f\|_{L^2(\Omega)}, \tag{5.1.2}$$

holds.

2) The *membrane eigenvalue problem*,

$$\begin{aligned} &\lambda \in \mathbb{R}, \quad \exists\, u \in H_0^1(\Omega), \quad u \neq 0, \\ &-\Delta u = \lambda u \quad \text{in } \Omega, \qquad u|_\Gamma = 0, \end{aligned} \tag{5.1.3}$$

which also has a variational formulation

$$\begin{aligned} &\lambda \in \mathbb{R}, \quad \exists\, u \in H_0^1(\Omega), \quad u \neq 0, \\ &\textstyle\int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv, \qquad \text{all } v \in H_0^1(\Omega). \end{aligned} \tag{5.1.4}$$

The chapter is organized as follows. Section 5.2 is reserved for the full detailed description of construction of spaces along with integration technique over triangles with a curved side. In Section 5.3 we describe how a conic given in rational Bézier form can be transformed to BB-form. We formulate theorems, in Section 5.4, regarding error bounds for our method. The implementation of the method for general second order elliptic problems in two dimensions is discussed in Section 5.5, while Section 5.6 is devoted to numerical results of some experiments for some classical elliptic problems, followed by a discussion of the possible extensions of the method to deal with non-homogeneous boundary conditions in Section 5.7.

5.2 Finite Element Spaces

Let $\triangle = \triangle_0 \cup \triangle_1$ be a *triangulation* of Ω, *i.e.* a subdivision of Ω into triangles, where each triangle $T \in \triangle_1$ either has one (and only one) edge replaced with a curved segment of the boundary (a so called *pie-shaped* triangle), or has a common edge with a pie-shaped triangle (we call these *buffer* triangles), while the remaining triangles $T \in \triangle_0$ have all straight edges, see Figure 5.2. To be more clear, let

$$\triangle_1 = \triangle_{Bf} \cup \triangle_P,$$

where $\triangle_{Bf}$ and $\triangle_P$ contains buffer and pie-shaped triangles respectively. Also let T^* denote the triangle associated with $T \in \triangle_P$ obtained by joining its boundary vertices by a straight line, see Figure 5.8. Buffer triangles are used to maintain global continuity as they "digest" everything that comes from neighbours. As usual, we assume that no vertex of a triangle lies in the interior of an edge of another triangle.

Suppose q_j is a bivariate polynomial such that $\Gamma_j \subset \{z \in \mathbb{R}^2 \ : \ q_j(z) = 0\}$, $j = 1, \dots, m$. We assume that $q_j \in P_1$ or $q_j \in P_2$ depending on whether Γ_j is a

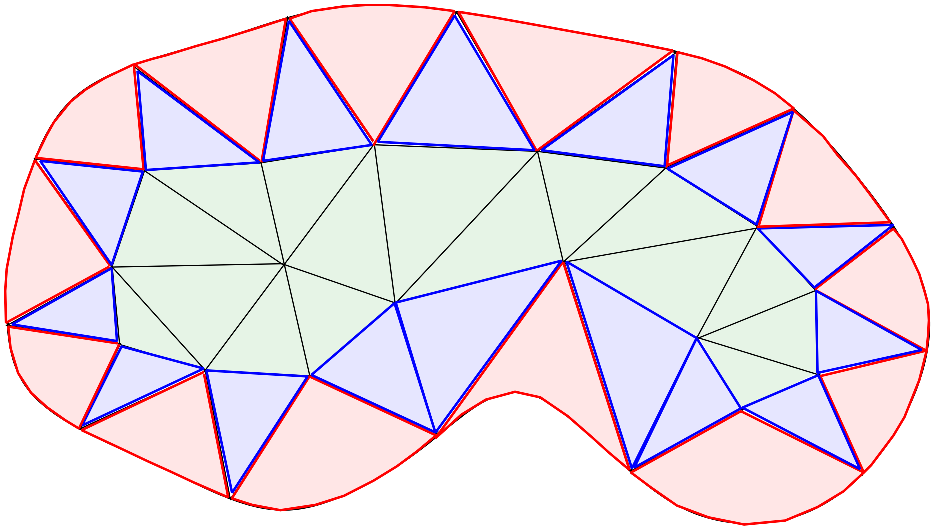

Figure 5.2: A triangulation of curved domain with buffer triangles(blue), pie-shaped triangles(pink) and ordinary triangles(green).

straight interval or a genuine conic arc, where P_d denotes the set of all algebraic polynomials in two variables of total degree at most d. It is worth noting that if $q_j \in P_1$ the boundary triangle with Γ_j as boundary edge belongs to $\triangle_0$ and we use standard Bernstein-Bézier elements on it. A boundary triangle belongs to $\triangle_P$ when $q_j \in P_2 \backslash P_1$.

Furthermore, let, V and E denote the set of all vertices and edges of $\triangle$ respectively. For each $v \in V$, star(v) is the union of all triangles in $\triangle$ attached to v. We also denote by θ the smallest angle of the triangles $T \in \triangle$, where the angle between an interior edge and a boundary segment is understood in the tangential sense.

We assume that $\triangle$ satisfies the following *conditions*:

(a) $Z = \{z_1, \ldots, z_M\} \subseteq V$.

(b) No interior edge has both endpoints on the boundary.

(c) If $q_j/q_{j-1} \neq$ const, if $q_j \in P_2 \backslash P_1$ or $q_{j-1} \in P_2 \backslash P_1$ then there is at least one triangle $T \in \triangle_{Bf}$ attached to z_j.

(d) For each $T \in \triangle_P$, with its curved side on Γ_j and its third (interior) vertex v,

$$q_j(z) \neq 0, \quad \forall z \in T \setminus \Gamma_j, \tag{5.2.5}$$

$$q_j(v) = 1, \tag{5.2.6}$$

$$q_j(z) \leq A \quad \forall z \in T, \tag{5.2.7}$$

for some constant A.

Note that (a)–(c) can always be achieved by slightly modifying a given triangulation near the boundary, while (d) is obtained by re-scaling q_j, if necessary, assuming that the triangulation is fine enough.

Let $d \geq 1$. We set

$$S_d(\triangle) := \{s \in C^0(\Omega) \ : \ s|_T \in P^{d+i}, \ T \in \triangle_i, \ i = 0, 1\} \subset H^1(\Omega),$$
$$S_{d,0}(\triangle) := S_d(\triangle) \cap H_0^1(\Omega),$$

when there is no ambiguity, we simply write S_d, $S_{d,0}$.

Let $V_I(E_I)$ and $V_B(E_B)$ be sets of interior and boundary vertices(edges) of the triangulation $\triangle$ of Ω. Note that $E_B := \{\Gamma_j \ : \ j = 1, \ldots, m\}$.

We describe the above space $S_d(\triangle)$, that possess a property of stable splitting $S_d := S_{d,0} + S_{d,b}$ by constructing a minimal determining set (MDS) for it. We start by first constructing an MDS for the space $S_{d,0}$ and then we will extend it to the full space S_d. The main idea is to factorize polynomials over boundary elements. Over each $T \in \triangle_P$, with curved side as Γ_j, we consider the polynomials $P_{d-1}q_j \subset P_{d+1}$ that satisfy the homogeneous Dirichlet conditions exactly. Recall that by Bézout's theorem every polynomial that vanishes on Γ_j is divisible by q_j.

Since the BB polynomials B^{d-1}_{ijk}, $i+j+k=d-1$ w.r.t. T^* form a basis for P_{d-1} it is obvious that the set $\left\{B^{d-1}_{ijk}q_j,\ i+j+k=d-1\right\}$ is a basis for $P_{d-1}q_j$.

Recall that $D_{d,\triangle_0}$ is the set of domain points associated with the subtriangulation $\triangle_0$ of $\triangle$. The main technicalities come while constructing an MDS on $\triangle_{Bf}$ and $\triangle_P$. For each $T := \langle v_1, v_2, v_3\rangle$ in $\triangle_{Bf}$ let $D^0_{d+1,T}$ be the set of interior domain points over T *i.e.*

$$D^0_{d+1,T} := \left\{\frac{iv_1+jv_2+kv_3}{d+1}\ :\ i+j+k=d+1,\ i,j,k\geq 1\right\},$$

see Figure 5.6. Finally, for each $T\in\triangle_P$, let $D^*_{d-1,T}$ be domain points over T^* for degree $d-1$. However, the meaning of the dual functionals associated with domain points $\xi\in D^*_{d-1,T}$ is non-standard. For $p\in P_{d+1}$ vanishing on the curved boundary of T it is the coefficients c_ξ in the expansion

$$p = q\sum_{\xi\in D^*_{d-1,T}} c_\xi B^{d-1}_\xi,$$

where $q=0$ is the curved boundary edge of T. For example see Figure 5.4 depicting the points $\xi\in D^*_{d-1,T}$ for $d=5$ where as Figure 5.5 shows the Bézier net for $p\in P_{d+1}$.

Then we have the following result.

Theorem 5.2.1. *Let*

$$M_0 := D_{d,\triangle_0\backslash\partial\Omega}\cup\bigcup_{T\in\triangle_{Bf}} D^0_{d+1,T}\cup\bigcup_{T\in\triangle_P} D^*_{d-1,T}. \tag{5.2.8}$$

Then M_0 is a stable local minimal determining set for the space $S_{d,0}$.

Proof. We set the coefficients $\{c_\xi\ :\ \xi\in M_0\}$ for any spline $s\in S_{d,0}$ and show that all coefficients of s can be computed from them consistently.

As $\triangle = \triangle_0\cup\triangle_{Bf}\cup\triangle_P$, let us consider the patches of s over $\triangle_0$, $\triangle_{Bf}$ and $\triangle_P$ separately for the sake of convenience. We will glue these patches up at the end of the proof.

First the set $D_{d,\triangle_0 \backslash \partial\Omega}$ being the set of domain points over $\triangle_0$ is a minimal determining set for $S_{d,0}(\triangle_0)$ by [43, Theorem 5.5] and hence $\{c_\xi : \xi \in D_{d,\triangle_0 \backslash \partial\Omega}\}$ uniquely determines $s|_{\triangle_0}$ satisfying $s|_{\partial\Omega} = 0$. In particularl BB-coefficients of $s|_{\triangle_0}$ at domain points on the boundary are zero.

Now consider $\triangle_P$. Let $T \in \triangle_P$ (with T^* being an associated triangle with straight sides) and let $q \in P_2$ be a conic representing the curved side of T. Since $s|_T = pq \in P_{d+1}$ for some $p \in P_{d-1}$ and the coefficients $\left\{c_\xi : \xi \in D^*_{d-1,T}\right\}$ uniquely describe $p \in P_{d-1}$, the product pq uniquely determines the patch $s|_T$, $T \in \triangle_P$. Note that $D^*_{d-1,T} \cap D_{d,\triangle_0 \backslash \partial\Omega} = \{v\}$, where v is the interior vertex of T. Since $q(v) = 1$, we have $s(v) = p(v)$, so that c_v is uniquely defined independently of whether we treat v as element of $D^*_{d-1,T}$ or $D_{d,\triangle_0 \backslash \partial\Omega}$.

For buffer triangles it is easy to see that the coefficients $\left\{c_\xi : \xi \in D^0_{d+1,T}\right\}$ give us the interior part of the Bézier net of the patch $s|_T$, $T \in \triangle_{Bf}$. Now we describe how we determine the coefficients for $s|_T$ corresponding to domain points on edges of T. This is actually how we use buffer triangles for global C^0 smoothness among pie-shaped triangles and interior triangles. To this end first let us consider the set $E_{I,Bf} \subset E$ of edges shared by an interior and a buffer triangle. Let $e \in E_{I,Bf}$ with T_I and T_{Bf} being, respectively, the triangles in $\triangle_0$ and $\triangle_{Bf}$ attached to e. Then based on our construction we have $s|_{T_I} \in P_d$ and $s|_{T_{Bf}} \in P_{d+1}$. Let

$$s|_{T_I} = s_I \text{ and } s|_{T_{Bf}} = s_{Bf}.$$

Now we have set the coefficients $\{c_\xi : \xi \in D_{d,T_I} \cap e\}$ to arbitrary values for $s_I|_e$, and coefficients $\left\{c_\xi : \xi \in D_{d+1,T_{Bf}} \cap e\right\}$ for $s_{Bf}|_e$ can be determined by raising the degree of polynomial $s_I|_e$ of degree d by 1 in conjunction with C^0 smoothness conditions (3.1.6) over the edge e.

Now let the set $E_{P,Bf} \subset E$ be the set of edges shared by pie-shaped and buffer triangles. Let T_P and T_{Bf} be pie-shaped and buffer triangles, respectively, attached to the edge $e \in E_{P,Bf}$ with

$$s|_{T_P} = s_P \text{ and } s|_{T_{Bf}} = s_{Bf}.$$

Since $s_P, s_{Bf} \in P_{d+1}$ thus in view of C^0 smoothness conditions (3.1.6) over the edge e the Bézier net for $s_{Bf}|_e$ can be determined from the Bézier net of $s_P|_e$ already computed. This completes the Bézier net for the spline $s \in S_{d,0}$ consistently.

We now show that M_0 is stable in the sense of Definition 3.1.7. Indeed this follows as all Bézier coefficients $\{c_\eta : \eta \notin M_0\}$ for $s \in S_{d,0}$ can be computed by product of polynomials and degree raising in conjuction with C^0 smoothness conditions, which are stable processes, see Section 3.1.3, Section 3.1.6 and [43, Lemma 2.29].

We now prove that M_0 is local as defined in Definition 3.1.6. Let c_η be the Bézier coefficient of a spline s with $\eta \notin M_0$ but $\eta \in T_\eta$. Then it is easy to see that Γ_η is always contained in star(T_η), which results in the locality of M with $l = 1$ in the sense of Definition 3.1.6. ■

5.2.1 Integration over Curved Elements

To develop a finite element method, that does not use any kind of nonlinear mapping to transform curved triangles into reference triangle, we obviously need a quadrature rule to integrate any function over pie-shaped triangles directly. For this purpose we make use of a Gauss-Legendre quadrature rule to approximate integrals over pie-shaped triangles. Let us describe it in detail. Let $T := \langle v_1, v_2, v_3 \rangle$ be a pie-shaped triangle with $v_3 \in V_I$ and let a conic arc $\widehat{v_1 v_2}$ be given by $q = 0$. Also let

$$T = T^* \cup R, \text{ if } T \text{ is convex,} \tag{5.2.9}$$

$$T = T^* \setminus R, \text{ otherwise,} \tag{5.2.10}$$

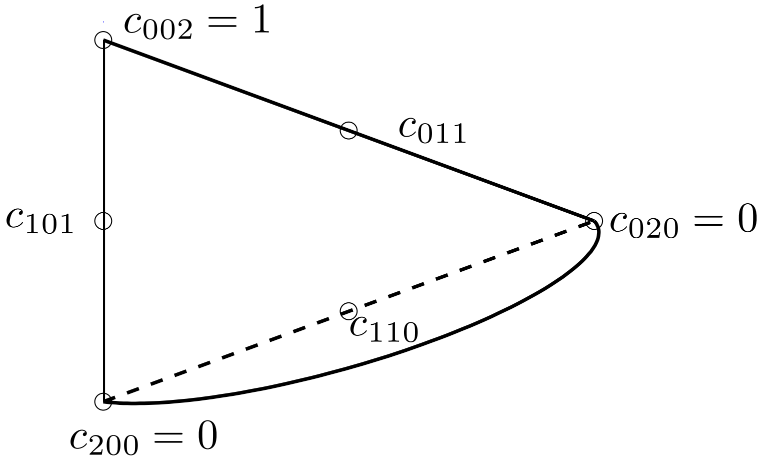

Figure 5.3: Bézier net of q over a pie-shaped triangle.

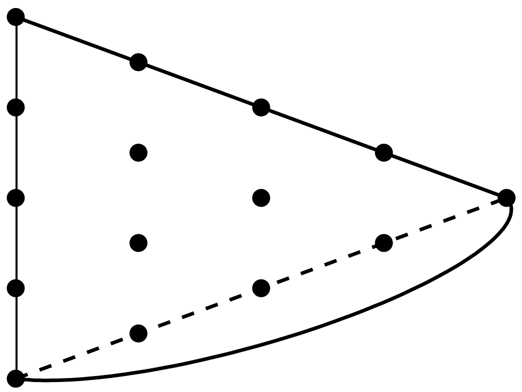

Figure 5.4: The points $\xi \in D^*_{d-1,T}$ for $d = 5$ over a pie-shaped triangle.

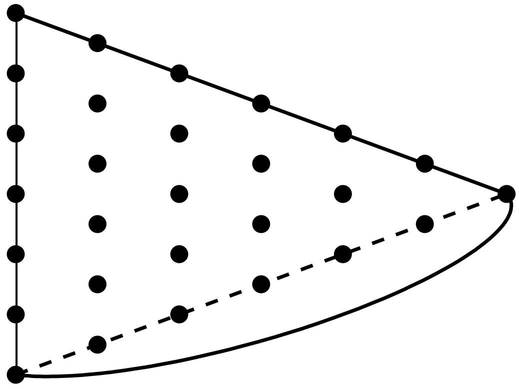

Figure 5.5: Bézier net for polynomial $p \in P_{d+1}$ for $d = 5$.

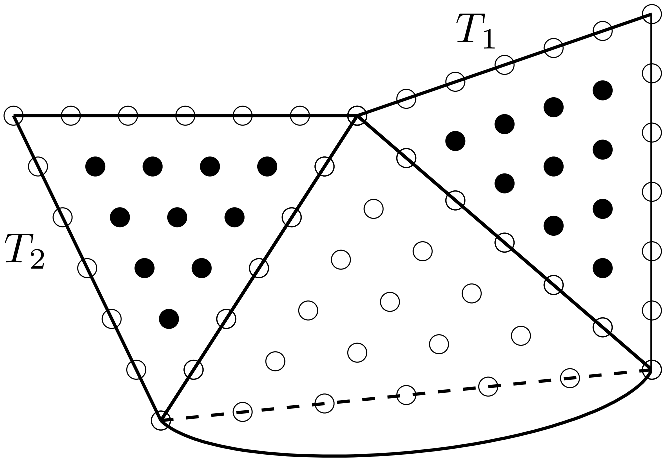

Figure 5.6: The points $\xi \in \{D^0_{d+1,T_1} \cup D^0_{d+1,T_2}\}$ are marked as black dots for $d = 5$, where $T_1, T_2 \in \triangle_{Bf}$.

where $T^* := \langle v_1, v_2, v_3 \rangle$ is the triangle with all straight edges obtained by joining v_1 and v_2 by line segment $\overline{v_2v_3}$ and R denotes the curved region bounded by $\widehat{v_1v_2}$ and $\overline{v_1v_2}$ as shown in Figure 5.8 and Figure 5.9. Then, obviously, for any function $f \in L_1(T \cup R)$ (to make quadrature well-defined in the case when the boundary is nonconvex we need to extend the integrand so that it is defined at quadrature points) we have

$$\begin{aligned} \int_T f &= \int_{T^*} f + \int_R f, \text{ if } T \text{ is convex,} && (5.2.11) \\ \int_T f &= \int_{T^*} f - \int_R f, \text{ otherwise,} && (5.2.12) \end{aligned}$$

Now the first integrals on R.H.S. of (5.2.11) and (5.2.12) is integration over the straight triangle T^* which can be evaluated using any suitable quadrature rule, as usual. The bottleneck is to approximate the second integral over the curved region R to the desired accuracy.

To approximate the double integral over R we consider two orthogonal directions i.e. the directions parallel and perpendicular to the straight edge $\overline{v_1v_2}$. We assume that the perpendicular line at any point of $\overline{v_1v_2}$ crosses $\widehat{v_1v_2}$ only once. This is always the case if the triangulation of Ω is sufficiently fine. For the sake of simplicity let the line segment $\overline{v_1v_2}$ lie on the x-axis as shown in Figure 5.7. Let $x_2 = \psi(x_1) \geq 0$, $a \leq x_1 \leq b$ be the equation of the conic arc in explicit form. Obviously $\psi(x_1)$ can be computed from the given implicit equation $q = 0$ of the conic arc. Then

$$\int_R f = \int_a^b \int_0^{\psi(x_1)} f(x_1, x_2) dx = \int_a^b g(x_1) dx_1, \text{ where } g(x_1) = \int_0^{\psi(x_1)} f(x_1, x_2) dx_2.$$

We use Gauss-Legendre quadrature rule of order n first for the integral $\int_a^b g(x_1)dx_1$

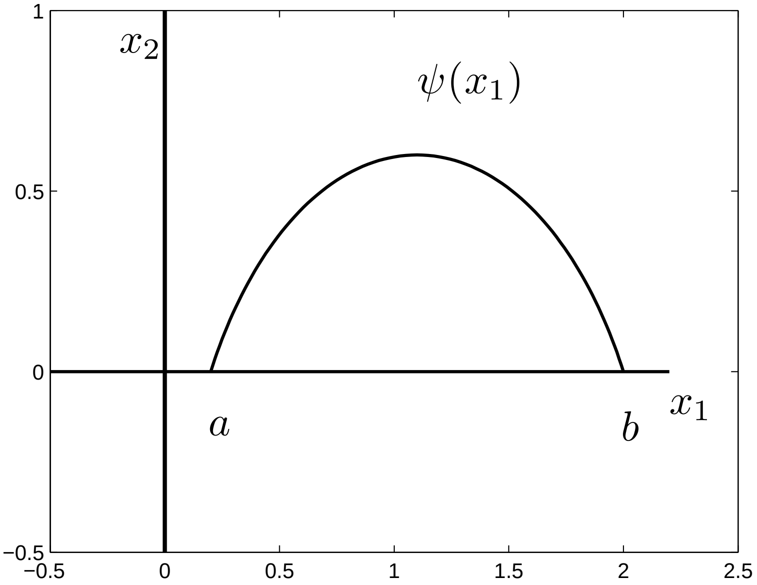

Figure 5.7: A function $\psi(x_1)$ representing a conic.

and then for each of the integrals $\int_0^{\psi(x_1^i)} f(x_1^i, x_2)dx_2$ to get

$$\begin{aligned} \int_R f &\approx \sum_{i=1}^m w_i g(x_1^i) = \sum_{i=1}^m w_i \int_0^{\psi(x_1^i)} f(x_1^i, x_2)dx_2 && (5.2.13) \\ &\approx \sum_{i=1}^m w_i \left(\sum_{j=1}^m w_{ij} f(x_1^{ij}, x_2^j) \right) && (5.2.14) \\ &= \sum_{i,j=1}^m w_i w_{ij} f(x_1^{ij}, x_2^j), && (5.2.15) \end{aligned}$$

where x_1^i, x_2^i are nodes and w_i are weights of the Gauss-Legendre quadrature rule for $[a, b]$ while x_1^{ij} and w_{ij} are nodes and weights of the Gauss-Legendre quadrature rule for intervals $[0, \psi(x_1^i)]$.

Remark 5.2.2. In case $f \in P_{2d-2}$ then $g(x_1) = F(x_1, \psi(x_1)) - F(x_1, 0)$, where

$F \in P_{2d-1}$ is an anti-derivative of f. Then

$$\int_R f \approx \sum_{i=1}^{m} w_i g(x_1^i) \tag{5.2.16}$$

$$\approx \sum_{i=1}^{m} w_i (F(x_1^i, \psi(x_1^i)) - F(x_1^i, 0)), \tag{5.2.17}$$

where x_1^i and w_i are nodes and weights of the Gauss-Legendre quadrature rule for intervals $[0, \psi(x_1^i)]$.

Since the accuracy of the quadrature scheme used to approximate the entries of element matrices affects the asymptotic convergence of FEM, a common rule is to use a quadrature scheme with sufficient order so that the error produced from the quadrature scheme does not dominate the approximation error of the finite element method. It is shown in [7, 5] that this is achieved by using the Gauss quadrature rule of order $d+1$, at least, when we use finite elements of order d for solving second order partial differential equations. We follow the same by using the Gauss-Legendre rule of order $d+2$ because we use shape functions of order $d+1$ on curved elements. Numerical experiments, discussed in Section 5.6, confirm that order $d+2$ is sufficient to get optimal asymptotic convergence of FEM.

5.3 The Conics

Here we briefly discuss the conics, their rational Bézier representation and conversion of conic written in parametric form to BB-form. It is well known that rational Bézier curves of degree two can be used to represent conic sections exactly [36].

Given three control points P_0, P_1 and P_2, the quadratic rational Bézier curve can be described by

$$B(t) = \frac{P_0 B_0^2(t) + \beta P_1 B_1^2(t) + P_2 B_2^2(t)}{B_0^2(t) + \beta B_1^2(t) + B_2^2(t)}, \ 0 \le t \le 1,$$

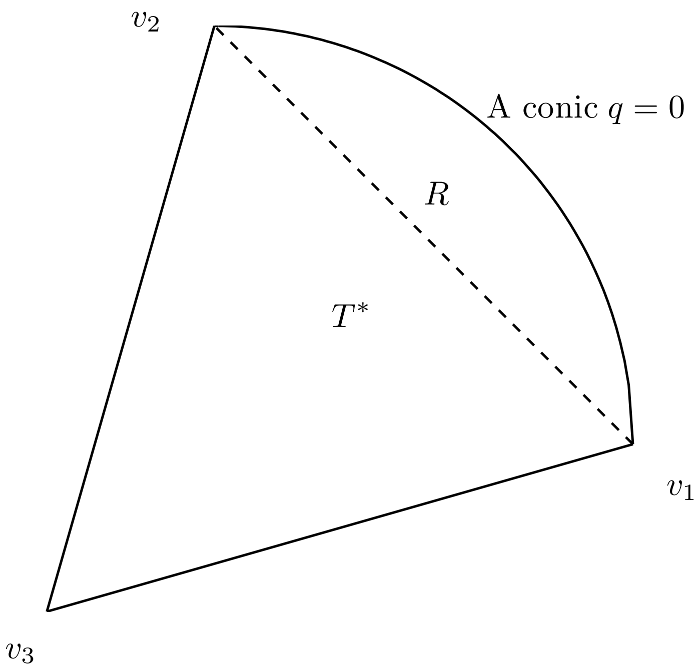

Figure 5.8: A pie-shaped triangle with a convex curved side, associated triangle T^* with straight sides and curved region R.

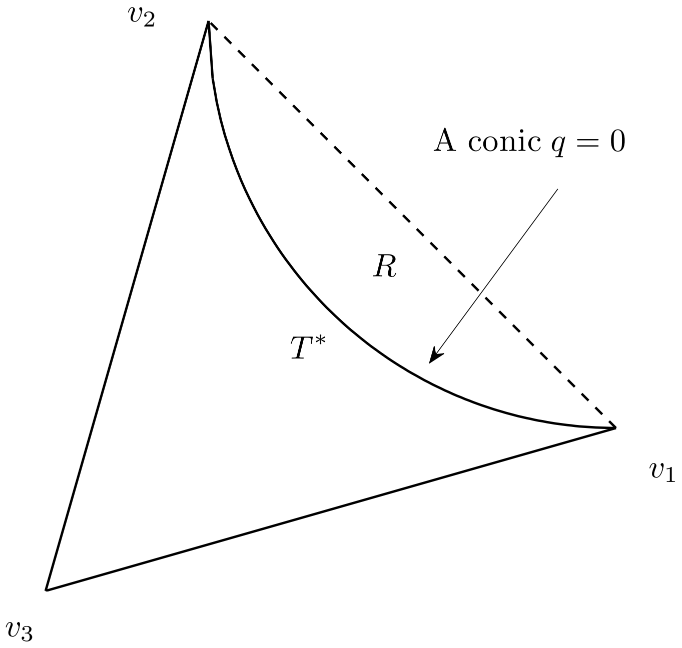

Figure 5.9: A pie-shaped triangle with a concave curved side, associated triangle T^* with straight sides and curved region R.

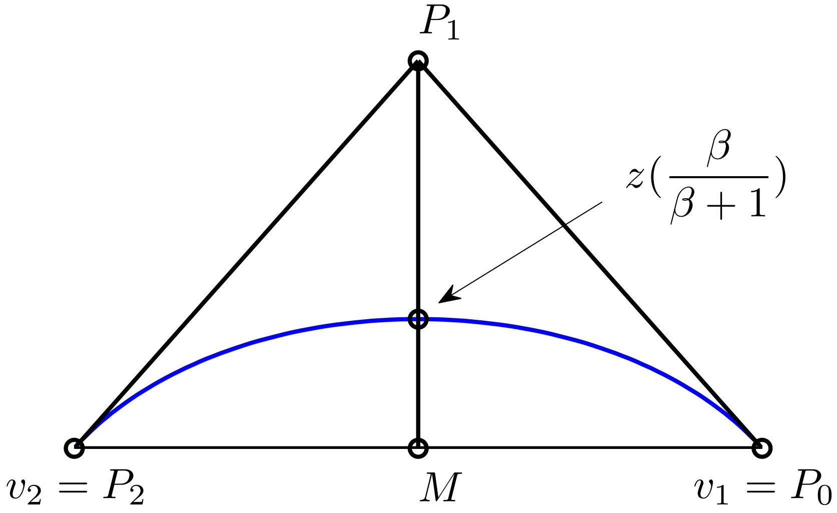

Figure 5.10: Conic as a rational Bézier curve.

where $\beta > 0$ is a weight and $B_i^2(t) = \binom{2}{i} t^i (1-t)^{2-i}$, $i = 0, 1, 2$ are quadratic Bernstein polynomials. [36, Lemma 4.5] says that rational Bézier curve $B(t)$ gives us a parabolic arc, an elliptic arc or a hyperbolic arc for $\beta = 1$, $\beta < 1$ or $\beta > 1$ respectively. Now let M be the mid-point of line segment $\overline{P_0P_2}$ and the line segment $\overline{MP_1}$ can be parametrized as

$$z(s) = M(1-s) + P_1 s, \ 0 \leq s < 1.$$

For $s = \frac{\beta}{1+\beta}$, $z(s)$ lies on the conic arc [36] , see Figure 5.10.

Let $(\beta_1, \beta_2, \beta_3)$ be barycentric coordinates of control point P_1 w.r.t. the boundary triangle $T = \langle v_1, v_2, v_3 \rangle$ with v_3 as interior vertex. Now our conic is given by the implicit equation $q = 0$ where q is a quadratic polynomial written in BB-form as

$$q = \sum_{i+j+k=2} c_{ijk} B_{ijk}^2.$$

Since $q(v_1) = 0 = q(v_2)$ we have $c_{200} = c_{020} = 0$ and since $q(v_3) = 1$ we obtain

$c_{002} = 1$, we can write

$$q = c_{110}B^2_{110} + c_{101}B^2_{101} + c_{011}B^2_{011} + B^2_{002},$$

or, more explicitly,

$$q = 2(c_{110}b_1b_2 + c_{101}b_1b_3 + c_{011}b_2b_3 + 0.5b_3^2), \tag{5.3.18}$$

see Figure 5.3.

Now as barycentric coordinates of the point M w.r.t T are $(\frac{1}{2}, \frac{1}{2}, 0)$, we get for $z(s)$

$$\begin{aligned} z(s) &= M(1-s) + P_1 s \\ &= \left(\frac{1}{2}, \frac{1}{2}, 0\right)(1-s) + (\beta_1, \beta_2, \beta_3)\, s \\ &= \left(\frac{1-s}{2} + s\beta_1, \frac{1-s}{2} + s\beta_2, s\beta_3\right) \end{aligned} \tag{5.3.19}$$

Now $q = 0$ is conic curve in the x_1, x_2 plane but $q(x_1, x_2)$ can be considered as a surface in 3-D. Thus using (5.3.19) in (5.3.18) gives us a parametrized curve $q(z(s))$ in space lying on this surface by restricting it to line segment $\overline{MP_1}$. Then obviously this curve has a root for parameter $s = \frac{\beta}{1+\beta}$ for given β, which results in the equation

$$\begin{aligned} & c_{110}(\frac{1-s}{2} + s\beta_1)(\frac{1-s}{2} + s\beta_2) + c_{101}(\frac{1-s}{2} + s\beta_1)(s\beta_3) + \\ & c_{011}(\frac{1-s}{2} + s\beta_2)(s\beta_3) + 0.5(s\beta_3)^2 = 0. \end{aligned} \tag{5.3.20}$$

To get all Bézier coefficients for q we need two more equations. For this we make use of the fact that tangents to the conic $q = 0$ at v_1 and v_2 are parallel to P_0P_1 and P_2P_1 respectively.

Since $\beta_1(P_0) = 1 = \beta_2(P_2)$ and $\beta_2(P_0) = \beta_3(P_0) = \beta_1(P_2) = \beta_3(P_2) = 0$, we get $(\beta_1 - 1, \beta_2, \beta_3)$ and $(\beta_1, \beta_2 - 1, \beta_3)$ as directional coordinates of P_0P_1 and P_2P_1

respectively. Note that $\beta_3 > 0(< 0)$ if P_1 lies inside T (outside T) and $\beta_3 = 0$ if P_1 lies on line P_0P_2.

Now

$$D_{P_0P_1}q(v_1) = 0,$$

where $D_{P_0P_1}q(v_1)$ is directional derivative of q in the direction of P_0P_1. Then using [43, Theorem 2.12] we have

$$\begin{aligned}
\sum_{i+j+k=1} c^{(1)}_{ijk}B^1_{ijk}(v_1) &= 0 \\
c^{(1)}_{100} &= 0 \\
(\beta_1 - 1)c_{200} + \beta_2 c_{110} + \beta_3 c_{101} &= 0 \\
\beta_2 c_{110} + \beta_3 c_{101} &= 0. \qquad (5.3.21)
\end{aligned}$$

Similarly $D_{P_2P_1}q(v_2) = 0$ leads us to the equation

$$\beta_1 c_{110} + \beta_3 c_{011} = 0. \qquad (5.3.22)$$

thus from (5.3.20)–(5.3.22) we get a system

$$Ac = L$$

where

$$A = \begin{bmatrix} \beta_1 & 0 & \beta_3 \\ \beta_2 & \beta_3 & 0 \\ (\frac{1-s}{2} + s\beta_1)(\frac{1-s}{2} + s\beta_2) & (\frac{1-s}{2} + s\beta_1)(s\beta_3) & (\frac{1-s}{2} + s\beta_2)(s\beta_3) \end{bmatrix},$$

$c = [c_{110}\ c_{101}\ c_{011}]^t$ and $L = [0\ 0\ -0.5(s\beta_3)^2]^t$. Solving the system allows us to write q in BB-form (5.3.18).

5.4 Error Bounds

In this section we provide some typical routine error bounds in the context of approximation theory and finite element techniques. The following theorem speaks about the approximating order of the space $S_{d,0}(\triangle)$.

Theorem 5.4.1. *[23, Theorem 5.1] Let $d \geq 1$ and $1 \leq m \leq d+1$. For any $u \in H^m(\Omega) \cap H_0^1(\Omega)$,*

$$\inf_{s \in S_{d,0}(\triangle)} \|u - s\|_{L^2(\Omega)} \leq C_1 h^m \|u\|_{H^m(\Omega)}, \tag{5.4.23}$$

$$\inf_{s \in S_{d,0}(\triangle)} \|u - s\|_{H^1(\Omega)} \leq C_2 h^{m-1} \|u\|_{H^m(\Omega)}, \tag{5.4.24}$$

where h is the maximal diameter of the triangles in $\triangle$, and C_1, C_2 are constants depending only on d, ω, A and θ.

Now the discretized version of (5.1.1) can be formulated as follows

$$\begin{aligned} &\text{Find } \tilde{u} \in S_{d,0}(\triangle), \text{ such that,} \\ &a(\tilde{u}, s) = (f, s), \qquad \text{for all } s \in S_{d,0}(\triangle). \end{aligned} \tag{5.4.25}$$

It is well known that this problem has a unique solution $\tilde{u}$ by the Lax-Milgram Theorem [12] for a coercive and bounded bilinear form $a(\cdot, \cdot)$.

As a consequence of Theorem 5.4.1, we obtain the following error estimate for finite element approximations in $S_{d,0}(\triangle)$.

Theorem 5.4.2. *Suppose the variational problem (5.1.1) is coercive and regular in the sense of (5.1.2). Then for any $2 \leq m \leq d+1$, the unique solution $\tilde{u}$ of (5.4.25) satisfies*

$$\|u - \tilde{u}\|_{L^2(\Omega)} \leq C_1 h^m \|u\|_{H^m(\Omega)}, \tag{5.4.26}$$

$$\|u - \tilde{u}\|_{H^1(\Omega)} \leq C_2 h^{m-1} \|u\|_{H^m(\Omega)}, \tag{5.4.27}$$

where h is the maximal diameter of the triangles in $\triangle$, and C_1, C_2 are constants depending only on d, ω, A and θ.

Proof. Follows from the Céa Lemma in view of Theorem 5.4.4 and Theorem 5.4.8 in [12]. ■

Note that the regularity condition (5.1.2) holds for the domains considered in this work if $\omega_j \leq \pi$, $j = 1, \ldots, m$, see *e.g.* [51, p. 158].

Let us now turn towards the eigenvalue problem (5.1.3) whose corresponding finite dimensional i.e. discritized problem is given as

$$\begin{aligned} &\lambda \in \mathbb{R}, \quad \exists\, \tilde{u} \in S_{d,0}(\triangle), \quad \tilde{u} \neq 0, \\ &\textstyle\int_\Omega \nabla \tilde{u} \cdot \nabla s = \lambda \int_\Omega \tilde{u} s, \qquad \text{for all } s \in S_{d,0}(\triangle). \end{aligned} \tag{5.4.28}$$

The following result follows from Theorem 5.4.1 and [6, Theorem 3.1].

Theorem 5.4.3. *Suppose that $\omega_j \leq \pi$, $j = 1, \ldots, m$. Let*

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \cdots$$

and

$$\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_N$$

be the eigenvalues of the problems (5.1.4) and (5.4.28), respectively, with $N := \dim S_{d,0}(\triangle)$. Then

$$|\lambda_n - \tilde{\lambda}_n| \leq C_n h^{2d}, \qquad n = 1, \ldots, N, \tag{5.4.29}$$

where h is the maximal diameter of the triangles in $\triangle$, and the constants C_n depend only on n, d, ω, A and θ.

5.5 Implementation of the FEM

In this section we briefly discuss the implementation of our FEM for solving second order elliptic problems coupled with Dirichlet boundary conditions, over domains with piecewise smooth boundary. This is done in a similar way to the solution of linearized problems in Section 4.2 of Chapter 3.

Recall that we confine ourself to problems where the bilinear form (not necessarily symmetric) is coercive and bounded, and where the solution satisfies a regularity condition (5.1.2). Consider a variational form, for a general second order linear operator, defined by

$$a(u,v) = \int_{\Omega} (\nabla u \cdot A \nabla v + v\mathbf{b} \cdot \nabla u + cuv)dx, \tag{5.5.30}$$

where $A = A(x)$, $\mathbf{b} = \mathbf{b}(x)$ and $c = c(x)$ are bounded functions on $\Omega \subseteq \mathbb{R}^2$ and $u, v \in H^1(\Omega)$. Under certain assumptions on A, $\mathbf{b}$ and c [12, Theorem 2.9.4] we know that there exist a unique solution to the variational problem

$$\text{Find } u \in H^1(\Omega) \text{ such that } a(u,v) = \langle f, v \rangle \ \forall v \in H_0^1(\Omega). \tag{5.5.31}$$

Discretization

Assuming $\triangle^h$ a triangulation of domain Ω, we use a standard Galerkin discretization of (5.5.31) based on elements in $S_d(\triangle^h)$. Thus the discretized version of (5.5.31) can be formulated as

$$\text{Find } u^h \in S_d(\triangle^h) \text{ such that } a(u^h, v^h) = \left\langle f, v^h \right\rangle \ \forall v^h \in S_{d,0}(\triangle^h).$$

[12, Theorem 2.9.4] guarantees existence of a unique solution for this problem. Given M a stable local MDS of $S_d(\triangle^h)$, we use the M-basis for the space $S_d(\triangle^h)$ dual to M. Let $\{s_1, \ldots, s_N\}$ be the M-basis for $S_d(\triangle^h)$ then the usual procedure to solve the discretized problem leads to the following element matrices $\mathcal{S}$ (the stiffness matrix), $\mathcal{B}$ (the convection matrix) and $\mathcal{M}$ (the mass matrix) whose entries are given by

$$\mathcal{S}_{ij} = \int_{\Omega} \nabla s_i \cdot A \nabla s_j dx, \ \mathcal{B}_{ij} = \int_{\Omega} s_j \mathbf{b} \cdot \nabla s_i dx, \ \mathcal{M}_{ij} = \int_{\Omega} c s_i s_j dx,$$

for all $i, j = 1, \ldots, N$, while an entry for load vector $\mathcal{L}$ is given as $\mathcal{L}_i = \int_\Omega f s_i dx$.

As in Section 4.2, we again use transformation matrix $\mathcal{T}$ to compute element matrices using relation (4.2.7). Note that the blocks of the block diagonal matrices $\hat{\mathcal{S}}$, $\hat{\mathcal{B}}$ and $\hat{\mathcal{M}}$ are of different sizes as we are using polynomials of different degrees for triangles in $\triangle_1^h$. These blocks, for $T \in \triangle_m^h$ as $\triangle^h = \triangle_0^h + \triangle_1^h$, are defined as

$$\hat{\mathcal{S}}_T = \int_T \nabla B_{ijk}^{d+m} \cdot A \nabla B_{rst}^{d+m} dx, \tag{5.5.32}$$

$$\hat{\mathcal{B}}_T = \int_T B_{ijk}^{d+m} \mathbf{b} \cdot \nabla B_{rst}^{d+m} dx, \tag{5.5.33}$$

$$\hat{\mathcal{M}}_T = \int_T c B_{ijk}^{d+m} B_{rst}^{d+m} dx. \tag{5.5.34}$$

We get the block diagonal matrices as follows

$$\hat{\mathcal{S}} = \text{diag}\left(\hat{\mathcal{S}}_{T_\kappa}, T_\kappa \in \triangle^h\right), \tag{5.5.35}$$

$$\hat{\mathcal{B}} = \text{diag}\left(\hat{\mathcal{B}}_{T_\kappa}, T_\kappa \in \triangle^h\right), \tag{5.5.36}$$

$$\hat{\mathcal{M}} = \text{diag}\left(\hat{\mathcal{M}}_{T_\kappa}, T_\kappa \in \triangle^h\right). \tag{5.5.37}$$

Hence we assemble system matrices using the same relation (4.2.7) given by

$$\mathcal{S} = \mathcal{T}\hat{\mathcal{S}}\mathcal{T}^t, \ \mathcal{B} = \mathcal{T}\hat{\mathcal{B}}\mathcal{T}^t, \ \mathcal{M} = \mathcal{T}\hat{\mathcal{M}}\mathcal{T}^t. \tag{5.5.38}$$

5.6 Numerical Experiments

To see the performance of our FEM we present numerical results, as obtained, of implementation of the method in this section. We consider three of the classical elliptic test models including the membrane eigenvalue problem and Poisson's problem over different domains with curved boundaries. We consider both h and p refinements and compare our results to the state-of-the-art software COMSOL Multiphysics. COMSOL uses the standard isoparametric approach to deal with curved domains. We use version 4.2a of COMSOL which allows us to use elements of order up to 5 in 2-D. To see the performance of the method we consider different domains in test problems including the smoothest domain *i.e.* a circular disk and a domain bounded by linear and quadratic pieces with C^0 boundary.

The numerics confirm the theoretical rate of convergence given in Theorems 5.4.2 and 5.4.3.

Example 1 : Circular Membrane Problem

The free vibrations of a homogeneous membrane are governed by the equation

$$\Delta u + \lambda u = 0, \quad x \in \Omega. \tag{5.6.39}$$

In addition, if the membrane is fixed along its boundary then (5.6.39) is coupled with

$$u = 0, \quad x \in \partial\Omega. \tag{5.6.40}$$

(5.6.39) and (5.6.40) actually comprises a problem of finding eigenvalues and eigenfunctions of the Laplacian completed by homogeneous Dirichlet boundary condition.

We consider $\Omega \subseteq \mathbb{R}^2$ to be a unit circular disk and approximate the smallest few eigenvalues for the circular membrane. The exact solution to this problem is known [39]. The eigenvalues of the circular membrane are given by

$$\lambda_{m,n} = (j_{m,n})^2, \quad m = 0, 1, \ldots, \quad n = 1, 2, \ldots,$$

where $j_{m,n}$ is the nth root of the mth Bessel function J_m of the first kind.

The weak variational formulation corresponding to (5.6.39) and (5.6.40), for $\lambda \in \mathbb{R}$ and $u \neq 0$, is

$$\text{Find } u \in H_0^1(\Omega), \int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv, \ \forall \ v \in H_0^1(\Omega). \tag{5.6.41}$$

We discretize this problem in the approximating space $S_{d,0}$, for $\lambda \in \mathbb{R}$ and $u^h \neq 0$, to get

$$\text{Find } u^h \in S_{d,0}(\Omega), \int_\Omega \nabla u^h \cdot \nabla v^h = \lambda \int_\Omega u^h v^h, \; \forall \; v^h \in S_{d,0}(\Omega). \tag{5.6.42}$$

Hence if $\{s_1, \ldots, s_N\}$ is an M_0-basis for the space $S_{d,0}(\Omega)$ then as usual (5.6.42) boils down to matrix equation of the form

$$\mathcal{S} = \lambda \mathcal{M},$$

where $\mathcal{S}$ and $\mathcal{M}$ are stiffness and mass matrices. We solve this system using the MATLAB command

$$[V, \lambda] = \text{eig}(\mathcal{S}, \mathcal{M}).$$

To ensure a fair comparison of the numerical results with COMSOL, we import the mesh from COMSOL to MATLAB and run our code on it. The initial mesh, as visualized in Figure 5.12, is the same in all tests for different degrees while solving problem (5.6.42). We get a sequence of meshes $\triangle^h$ by uniform refinement whereby each triangle is subdivided into four triangles, on each level, by joining the midpoints of every edge. For a pie shaped triangle with Γ_j being the corresponding curved boundary edge we take the midpoint of the curved boundary edge Γ_j, see Figure 5.11.

In Figures 5.13–5.16 we plot absolute errors for approximating the few smallest eigenvalues using our implementation and using COMSOL for degree $d = 3$ and $d = 5$. It can be seen that, comparing to COMSOL, our method approximates the 1st two eigenvalues significantly better showing its effectiveness since in practical applications usually the principal eigenvalue needs to be accurately approximated as it plays an important role in many processes. Though for the other eigenvalues the results are comparable for $d = 3$ but for higher degree $d = 5$ accuracy achieved using our method is better.

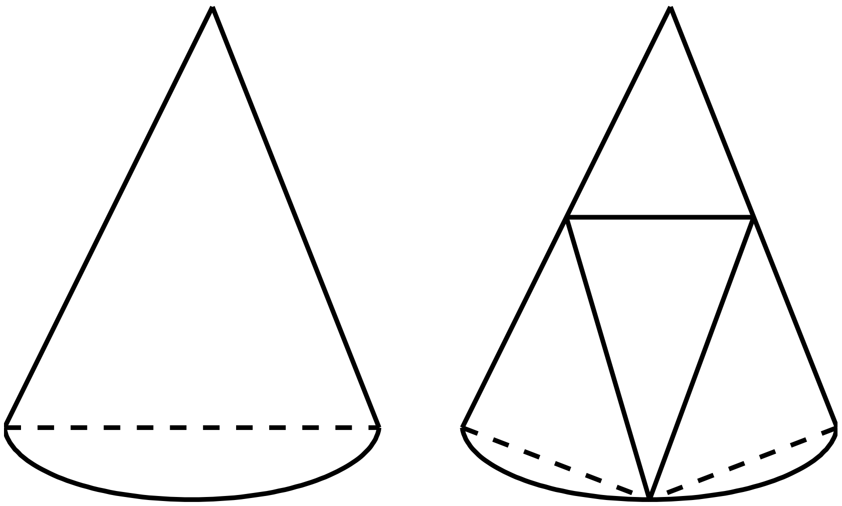

Figure 5.11: Refinement on a pie shaped triangle.

Since this problem is an example of a smooth problem thus to get higher order accuracy we can use polynomials of higher orders. Figure 5.17 depicts the errors for the smallest 15 eigenvalues for order $d = 9$ decays with the expected rate.

We also looked for p refinements for this problem on initial mesh. We plot absolute errors for 1st, 7th and 15th eigenvalue in Figure 5.18 for comparing results with COMSOL. COMSOL could only go to quintic. Figure 5.19 illustrate errors for the 15 smallest eigenvalues. Results show the expected exponential order of convergence for p method.

Example 2 : Poisson's Problem

In our second example we consider a different curved domain bounded by a boundary with non-constant curvature. The elliptic domain is one of such example. Let us consider the most frequently used model *i.e.* Poisson's equation coupled with

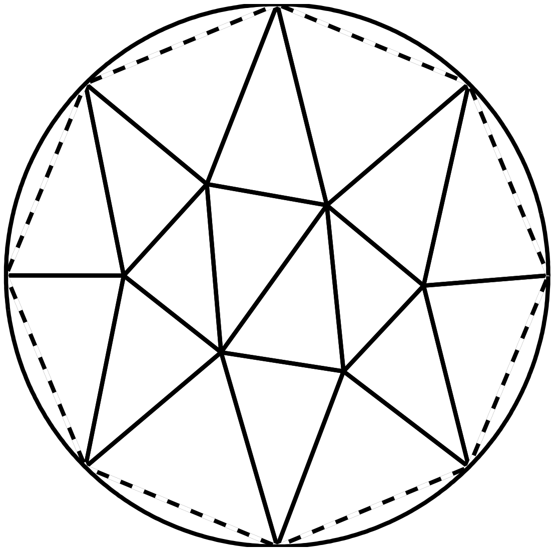

Figure 5.12: Initial mesh of circle for eigenvalue problem imported from COMSOL.

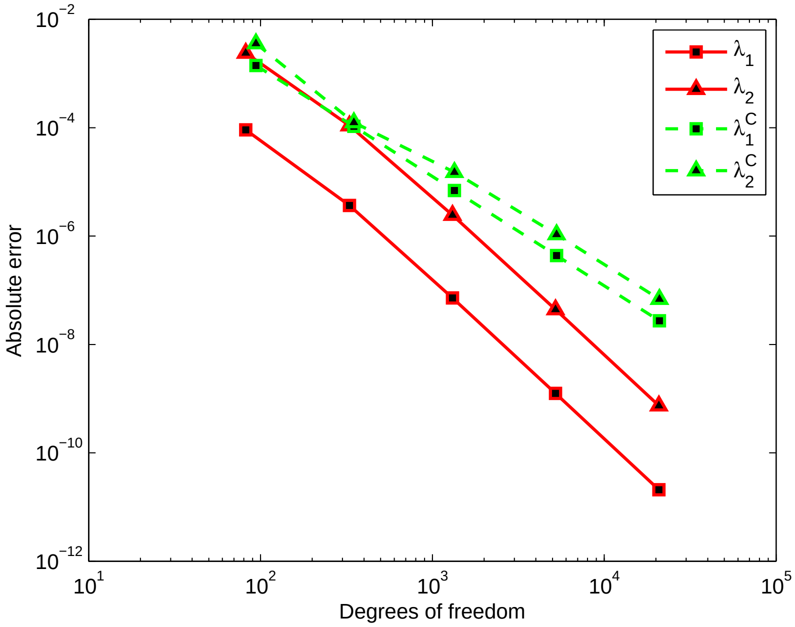

Figure 5.13: Absolute errors to 1st and 2nd eigenvalues from our method indicated by (λ_1, λ_2) and from COMSOL for $d = 3$ $(\lambda_1^C,\lambda_2^C)$.

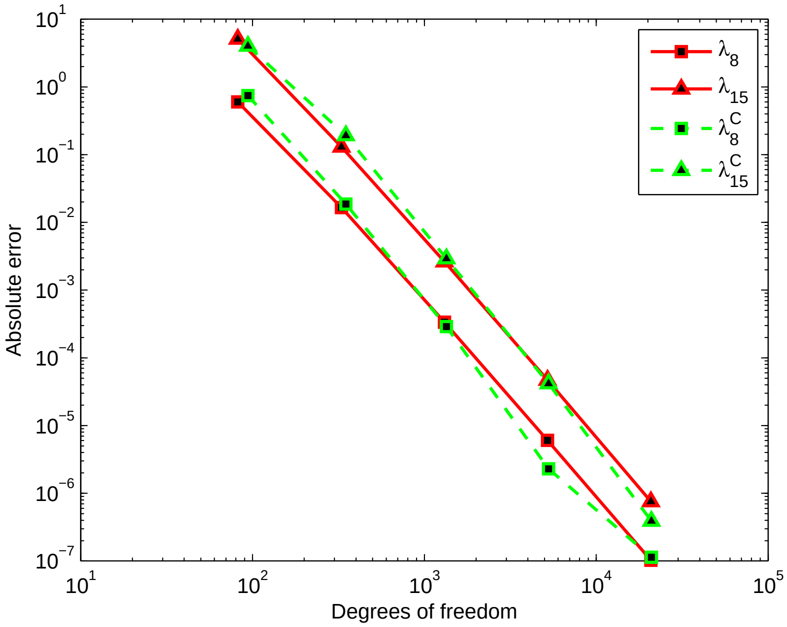

Figure 5.14: Absolute errors to 8th and 15th eigenvalues from our method and from COMSOL for $d = 3$.

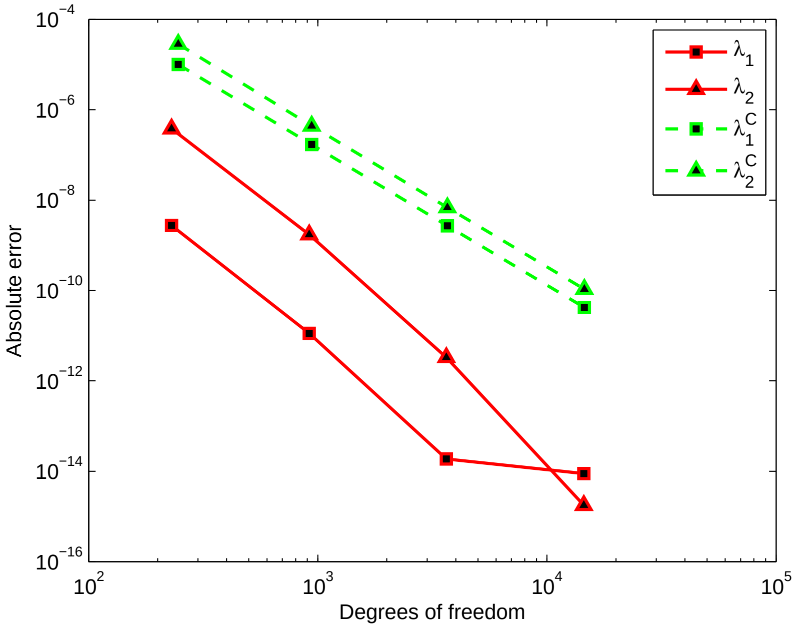

Figure 5.15: Absolute errors to 1st and 2nd eigenvalues from our method and from COMSOL for $d = 5$.

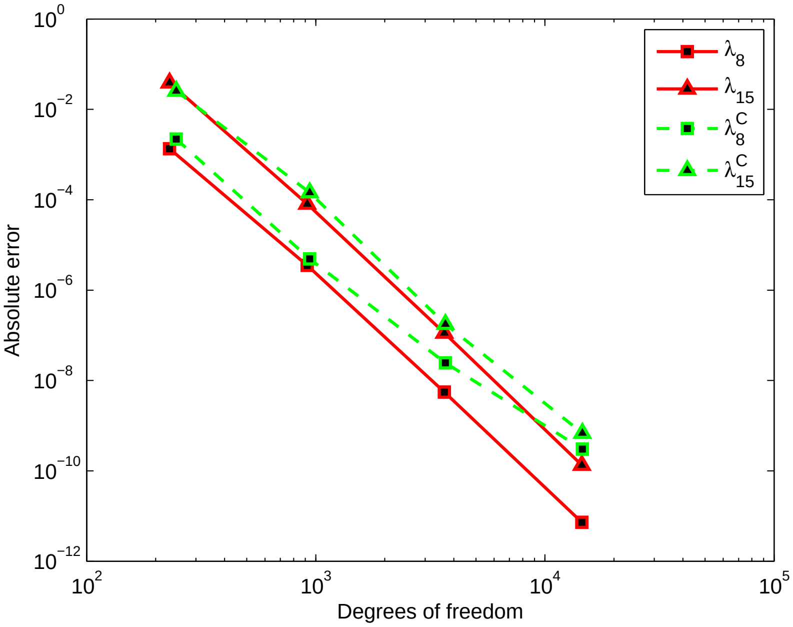

Figure 5.16: Absolute errors to 8th and 15th eigenvalues from our method and from COMSOL for $d = 5$.

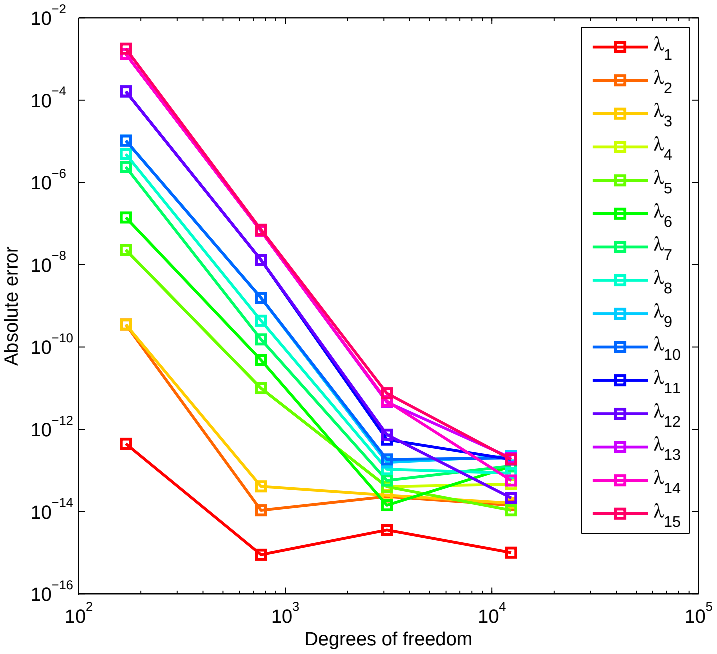

Figure 5.17: Absolute errors for smallest 15 eigenvalues using our method for $d = 9$.

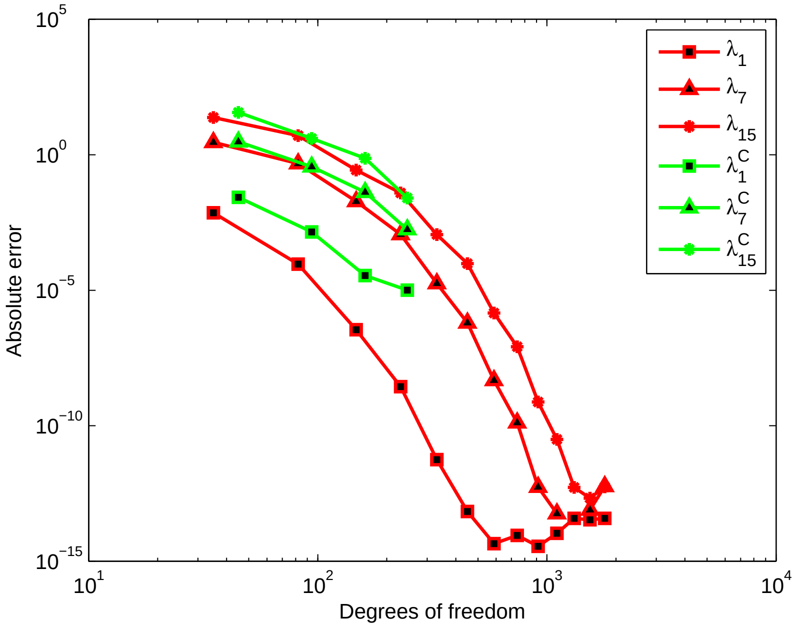

Figure 5.18: Absolute errors for 1st, 7th and 15th eigenvalues using p method over initial mesh shown in Figure 5.12.

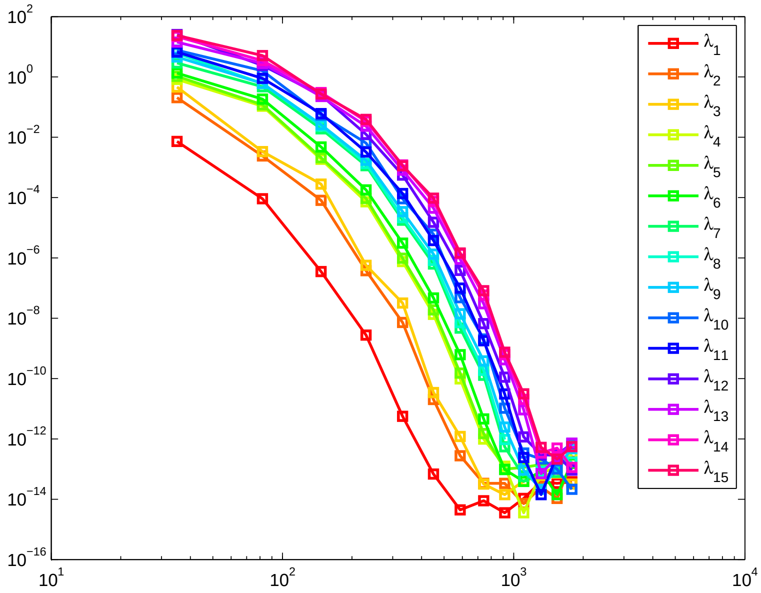

Figure 5.19: Absolute errors for smallest 15 eigenvalues using p method over initial mesh shown in Figure 5.12.

homogeneous Dirichlet data over an elliptic domain $\Omega \subset \mathbb{R}^2$ bounded by the ellipse $x_1^2 + 6.25x_2^2 = 1$. The model as usual can be formulated by

$$\begin{aligned} \Delta u &= f \text{ in } \Omega \qquad (5.6.43) \\ u &= 0 \text{ on } \partial\Omega. \qquad (5.6.44) \end{aligned}$$

We choose f such that the model possess the exact solution given as $u = e^{0.5(x_1^2+6.25x_2^2)} - e^{0.5}$, which can be used for precise error analysis. We follow the same procedure to get a sequence of meshes as in the first test problem starting with initial mesh imported from COMSOL. Error plots for $\|u-u^h\|_{L_2(\Omega)}$ and $\|u-u^h\|_{H^1(\Omega)}$ are depicted in Figure 5.20 and in Figure 5.21, respectively, for $d = 2, 3, 4, 5$, where u^h is the approximate solution to the problem. Green colour is used to indicate errors from COMSOL in these plots. The results show that both methods behave similarly for all different orders and confirm the theoretical estimates of Theorem 5.4.2. We also consider the p method for the example on third level of triangulation and visualize errors in Figure 5.22 that again show the expected exponential decay of errors.

Example 3

Here we consider a domain with C^0 boundary bounded by linear and quadratic boundary segments. Let Ω be a domain bounded by two straight lines $x_2 = \pm 2$ and parabolas $x_1 = \pm(x_2^2 - 6)$. We design a homogeneous Poisson's model over Ω such that it has the exact solution $u = (x_2^2 - 4)(x_1^2 - (x_2^2 - 6)^2)/100$. We consider elements of different orders and again compare our results with COMSOL. Figure 5.23 and Figure 5.24 illustrate the L_2 and H^1 errors for degrees $d = 2, 3, 4, 5$. Again the numbers show the robust behaviour of our method for different orders while confirming the error bounds of Theorem 5.4.2.

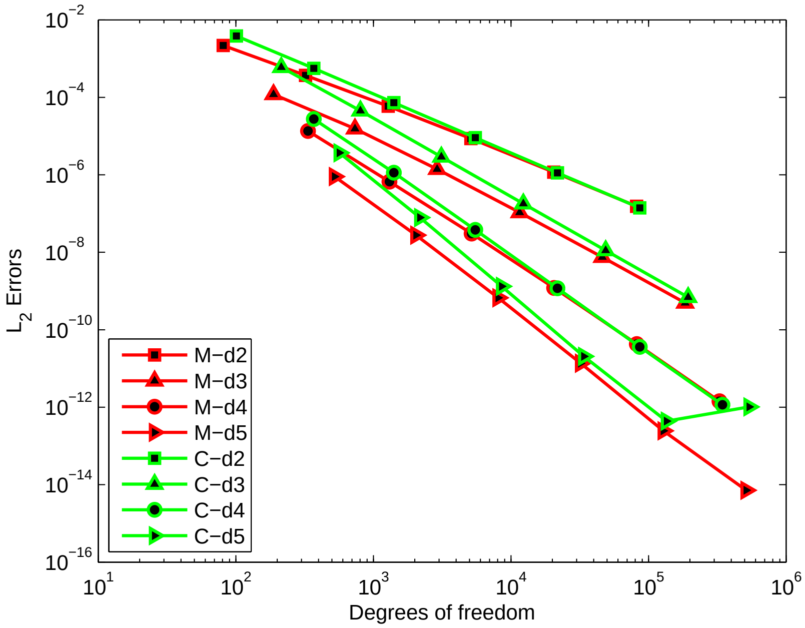

Figure 5.20: L_2 errors for example 2 using our method and COMSOL.

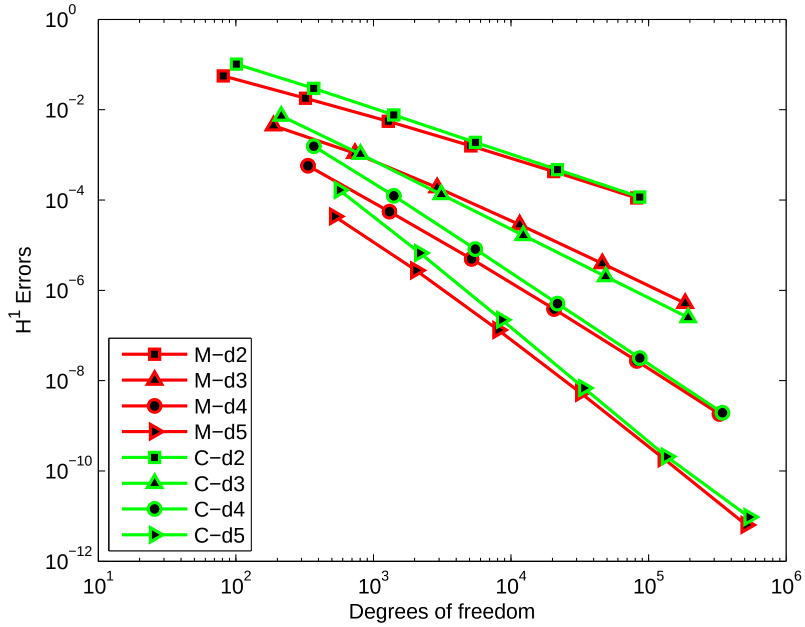

Figure 5.21: H^1 errors for example 2 using our method and COMSOL.

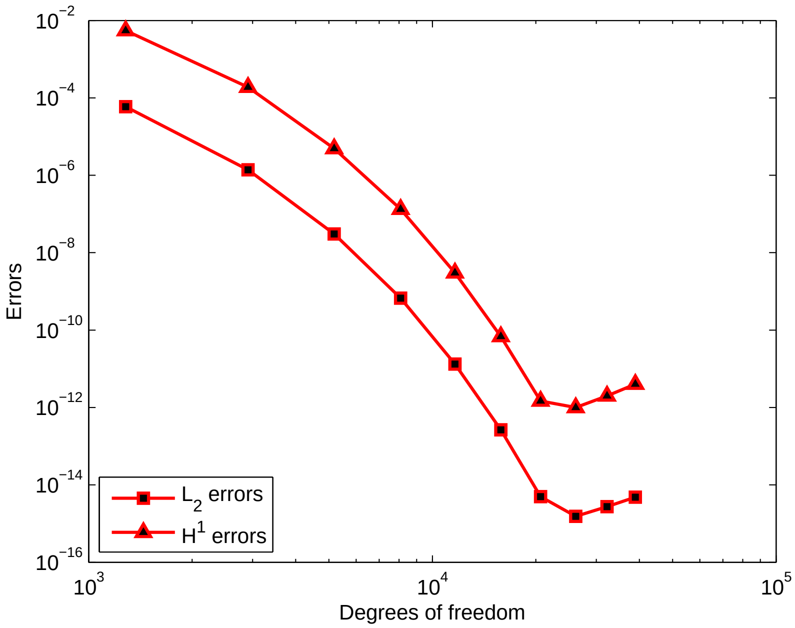

Figure 5.22: L_2 and H^1 errors for example 2 using p method.

Remark 5.6.1. Being polynomial of degree 6 the solution u in example 3 lies in the approximating space for degree $d \geq 6$ therefore we do not consider p refinement for this example.

5.7 Non-homogeneous Boundary Conditions

Until now we discussed the construction of the space $S_{d,0}$ that satisfies homogeneous Dirichlet conditions. We want to construct a space $S_{d,b}$ such that

$$S_{d,0} + S_{d,b},$$

is suitable for solving non-homogeneous problems. We summarize all of our attempts we make in this regard.

Now we need to look to add more shape functions to $P_{d-1}q$, over a pie-shaped triangle, that will deal with non-homogeneous conditions. We looked for the fol-

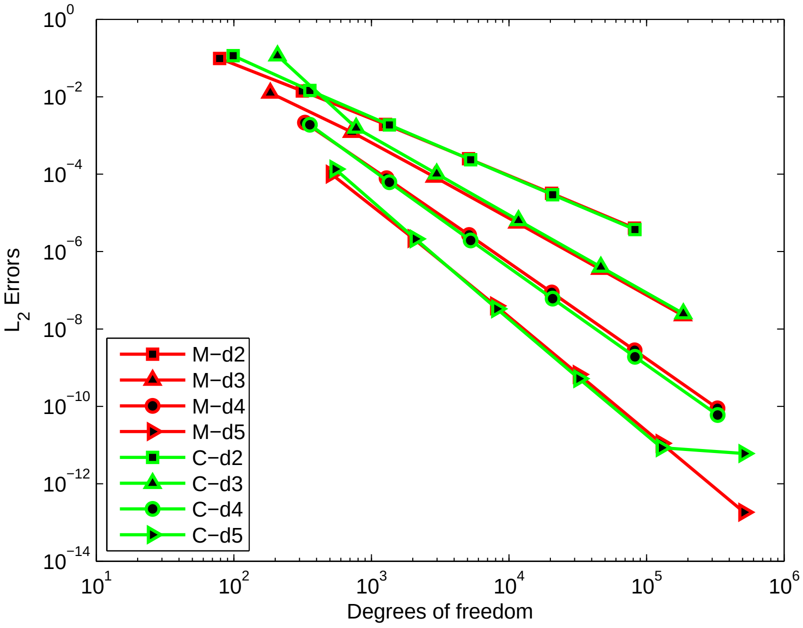

Figure 5.23: L_2-errors for example 3 using our method and COMSOL.

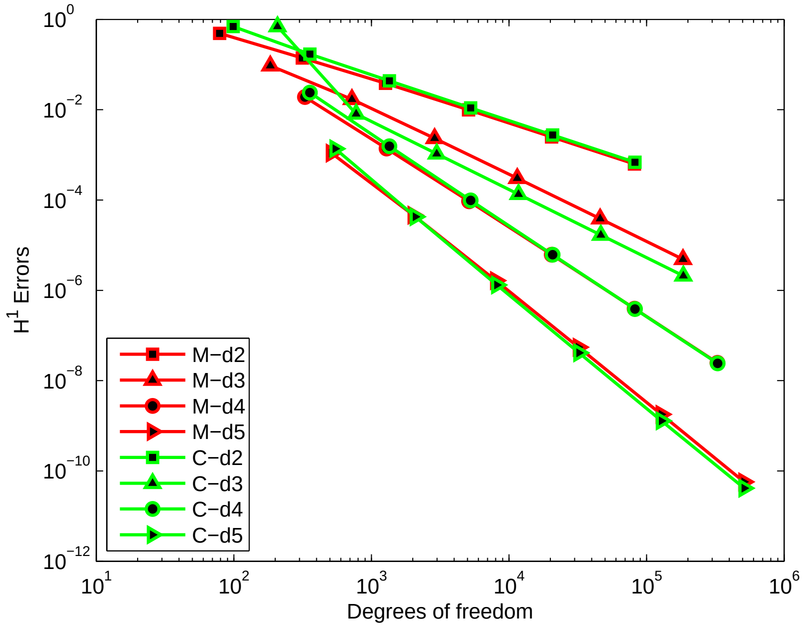

Figure 5.24: H^1-errors for example 3 using our method and COMSOL.

lowing options.

1. The first and natural option is to consider the polynomials B^d_{0jk}, $j+k=d$, over the triangle T^* associated with the pie-shaped triangle T with 1st vertex as interior vertex. We need to raise the degree of polynomials B^d_{0jk}, $j+k=d$, by 1, using the relation (3.1.9), just to make them compatible with $P_{d-1}q$.

   Let $\widehat{P}_{d,b} = \operatorname{span}\{B^d_{0jk},\ j+k=d\}$, then it is easy to see that

   $$\dim\left(P_{d-1}q+\widehat{P}_{d,b}\right) = \binom{d+2}{2} = \dim P_d,$$

   see Remark 5.7.2. Note that $P_d \not\subset P_{d-1}q+\widehat{P}_{d,b}$.

   If $u=g$ on the curved boundary $q=0$ of T, then to approximate the boundary conditions we solve the following interpolation problem

   $$\sum_{j+k=d} B^d_{0jk}(z_i)x_i = g(z_i),\ i=0,\ldots,d, \tag{5.7.45}$$

   for the x_i, where z_i for $i=0,\ldots,d$ are interpolation points on boundary curve $q=0$ obtained by perpendicular projection (the direction perpendicular to the straight boundary edge of T^*) of Lobatto interpolation points of order d lying on a straight boundary edge of T^*. The problem (5.7.45) is solvable for any choice of interpolation points, see Remark 5.7.3. Let us call the method "Method 1" when we use the interpolation problem (5.7.45) to solve any non-homogeneous problem.

   The space $\widehat{S}_{d,b}$ is spanned by 1) the functions B^d_{0jk}, $j+k=d$, $j,k>0$, for each triangle with a side on the boundary, extended by zero otherwise, as well as by 2) piecewise polynomials s_v, for each boundary vertex v, defined on each triangle attached to v as the Bernstein polynomial $B^{T,d}_v$ corresponding to the domain point v if $T \in \triangle_0 \cup \triangle_{Bf}$, or as $B^{T^*,d}_v$ if $T \in \triangle_P$, and zero on all other triangles of $\triangle$. Note that $S_{d,0}+\widehat{S}_{d,b} \neq S_d$.

To describe the MDS for space $\widehat{S}_{d,b}$ for Method 1 we proceed as follows. Let for each $T := \langle v_1, v_2, v_3\rangle$ in $\triangle_P$ with $v_1 \in V_I$

$$M_T^1 := \left\{ \frac{jv_2 + kv_3}{d} \;:\; j + k = d \right\},$$

then the set $\widehat{M_b} := \bigcup_{T\in\triangle_P} M_T^1$ is, obviously, a stable MDS for $\widehat{S}_{d,b}$ and we arrive at the following result.

Theorem 5.7.1. *$M := M_0 \cup \widehat{M_b}$ is a stable local MDS for the space $S_{d,0} + \widehat{S}_{d,b}$. Moreover, $M_0 \cup \widehat{M_b}$ is a stable splitting of M.*

2. In this part we look for a complement $S_{d,b}$ of $S_{d,0}$, to get the full space $S_d = S_{d,0} + S_{d,b}$. As we are already using polynomials of degree $d+1$ on pie-shaped triangles we have to add such shape functions to $P_{d-1}q$ that ensure the reproduction of all polynomials of degree $d+1$. The polynomials

$$\left\{ B_{ijk}^{d+1},\ i + j + k = d + 1,\ i = 0, 1 \right\}$$

over T^* help us in this regard because

$$P_{d-1}q + \operatorname{span}\left\{ B_{ijk}^{d+1},\ i + j + k = d + 1,\ i = 0, 1 \right\} = P_{d+1},$$

(for proof see Remark 5.7.2). We therefore define for each $T := \langle v_1, v_2, v_3\rangle$ in $\triangle_P$ with $v_1 \in V_I$,

$$M_T^2 := \left\{ \frac{iv_1 + jv_2 + kv_3}{d+1} \;:\; i + j + k = d + 1,\, i = 0, 1 \right\},$$

and set $M_b := \bigcup_{T\in\triangle_P} M_T^2$. Then $M = M_0 \cup M_b$ is easily seen to be an MDS for S_d, and $S_{d,b}$ is defined by (3.1.14).

In this case we can approximate the boundary conditions by solving the following interpolation problem:

$$\sum_{\substack{i+j+k=d+1\\ i=0,1}} B_{ijk}^{d+1}(z_r)x_r = g(z_r),\ r = 0, \ldots, 2d + 2, \tag{5.7.46}$$

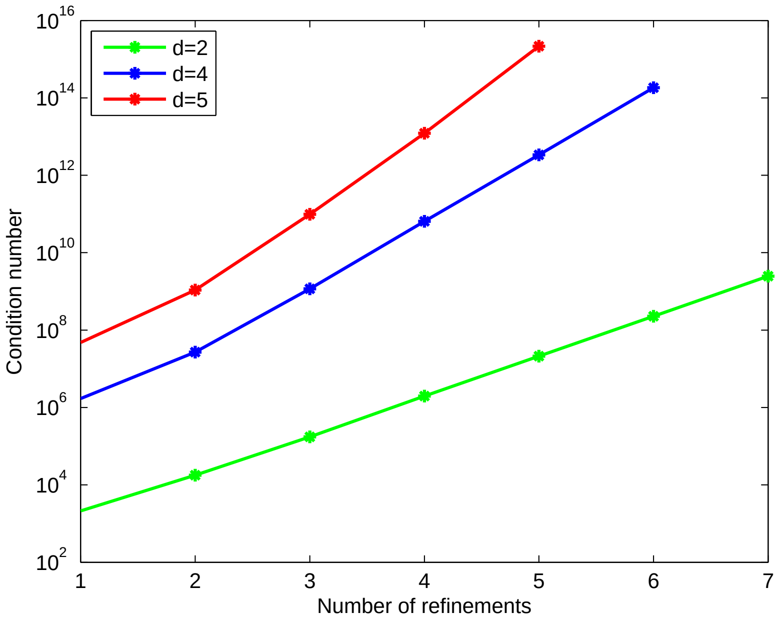

Figure 5.25: Condition numbers of the interpolation matrix obtained from (5.7.46) for different degrees.

for the x_r, where again z_r's are interpolation points on a curve side $q = 0$ obtained in a way as mentioned above.

The solvability of (5.7.46) follows from [17, Proposition 3.2]. However, its condition number grows rapidly both with h and p refinements, see Figure 5.25. This can be explained by the fact that the polynomials B^{d+1}_{1jk} are zero on the straight side v_2v_3 of T^*. This makes interpolation on curved side instable for refined meshes when the curved side $q = 0$ gets closer to the straight one.

3. Due to the instability of interpolation problem (5.7.46) we consider a different technique to approximate boundary conditions while using the same space $S_{d,b}$ as before. Let $T \in \triangle_P$. Given a function f defined on T^*, we obtain its approximation $Q(f) \in \text{span}\{B^{d+1}_{\xi} : \xi \in M^2_T\}$ with small error over the

curved side $q = 0$ of T in the following way. We first interpolate f by a polynomial $p_{d+1} \in P_{d+1}$ at the domain points D_{d+1,T^*}. Then a unique polynomial $p_{d-1} \in P_{d-1}$ is found such that

$$p_{d+1} - p_{d-1}q \in \operatorname{span}\{B_\xi^{d+1} : \xi \in M_T^2\}.$$

Clearly, the B-coefficients of p_{d-1} can be obtained by solving the linear system resulting from the conditions that all B-coefficients of $p_{d+1} - p_{d-1}q$ corresponding to domain points in $D_{d+1,T^*} \setminus M_T^2$ vanish. We set $Q(f) := p_{d+1} - p_{d-1}q$. Note that $Q(f)$ interpolates f at the boundary vertices v_2, v_3 of T. Hence, by applying the above procedure to all pie-shaped triangles and using the buffer triangles in the usual way we obtain an approximating function in $S_{d,b}$.

If u is an exact solution to the BVP with $u|_{\partial\Omega} = g$ then we get a function f in two different ways. One we define $f = u|_T$ and we refer to this method as "Method 2" in the sequel. Second we define f as a constant projection of boundary data g such that if $x \in T$ then $f(x) = g(\widehat{x})$ where $\widehat{x}$ lies on curved boundary $q = 0$ of T and is obtained by projection of x on boundary in a direction perpendicular to straight boundary side of T^*. We call this approach "Method 3" while presenting numerical results for solving boundary value problems below, whereas the method when we use interpolation problem (5.7.46) to approximate the boundary conditions is referred to as "Method 4".

Let us now consider a non-homogeneous Poisson problem over an elliptic domain Ω bounded by $x_1^2 + 6.25x_2^2 = 1$ with $u = \sin 2(x_1 + x_2) + \sin 2(x_1 - x_2)$ as an exact solution to compare performance of different approaches discussed above. We also solve the problem using COMSOL and illustrate the errors in Figure 5.26 and Figure 5.27 for $d = 5$. Poor performance of Method 1 is obvious and we think it is due to the lack of its ability to reproduce all polynomials of degree d on pie shaped triangles. Instability of Method 4 can be seen clearly. Method 2 seems to perform best but obviously it is not practical since it relies on the knowledge of the exact solution on the pie shaped triangles. The good performance of Method 2 seems to indicate that $S_{d,b}$ possesses a stable basis. Method 3 works well for this example but it is not robust. We consider another non-homogeneous Poisson problem over domain bounded by straight lines $x_2 = \pm 1$ and parabolas $x_1 = \pm(x_2^2 - 3)$ with exact solution $u = e^{-\left(\frac{|x_1|}{10}\right)^2}$ to compare Method 2 and Method 3. Figure 5.28 depicts the L_2 and H^1 errors for this problem. It shows that, though, convergence order is optimal for Method 3 but it is lagging behind in accuracy comparing to Method 2.

Remark 5.7.2. Without any loss of generality we assume that the interior vertex v_1 of $T \in \triangle_P$ lies at the origin. Then $\left\{B^d_{0jk},\ j+k=d\right\}$ are homogeneous polynomials of degree d so that their span $H_d := \text{span}\left\{B^d_{0jk},\ j+k=d\right\}$ is the space of all homogeneous polynomials of degree d.

Let $qp_{d-1} + h_d \in P_{d-1}q + H_d$ such that $qp_{d-1} + h_d = 0$, where $h_d \in H_d$ and $p_{d-1} \in P_{d-1}$. Then q and p_{d-1} can be written as $q = c + h_1 + h_2$ and $p_{d-1} = \widehat{c} + \widehat{h}_1 + \ldots + \widehat{h}_{d-1}$, where h_i and $\widehat{h}_i$ are both homogeneous polynomials of degree i. The condition $q(v_1) \neq 0$, as assumed, results in $c \neq 0$. Now consider

$$\begin{aligned} qp_{d-1} &= (c + h_1 + h_2)\left(\widehat{c} + \widehat{h}_1 + \ldots + \widehat{h}_{d-1}\right) \\ &= c\widehat{c} + \left(c\widehat{h}_1 + \widehat{c}h_1\right) + \left(c\widehat{h}_2 + h_1\widehat{h}_1 + \widehat{c}h_2\right) + \ldots. \end{aligned}$$

Since $qp_{d-1} = -h_d \in H_d$ implies that all terms for qp_{d-1} with degree less than d are zero. Hence $c\widehat{c} = 0 \Rightarrow \widehat{c} = 0$ as $c \neq 0$. Similarly $c\widehat{h}_1 + \widehat{c}h_1 = 0$ results

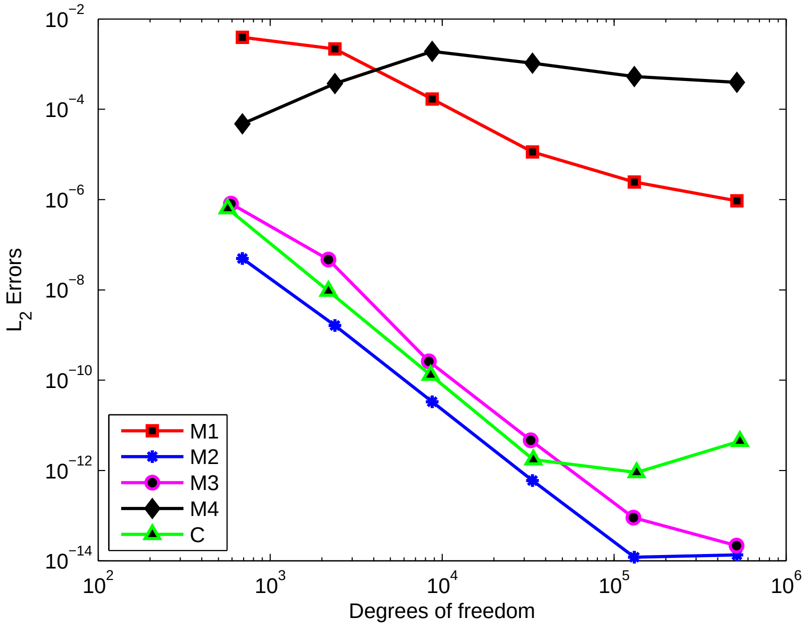

Figure 5.26: L_2 errors obtained using different methods for non-homogeneous problem for $d = 5$.

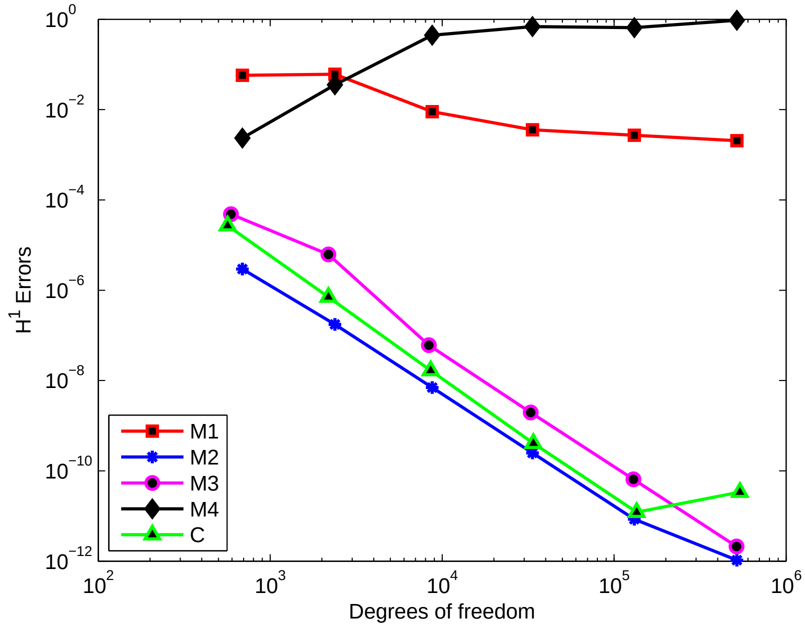

Figure 5.27: H^1 errors obtained using different methods for non-homogeneous problem for $d = 5$.

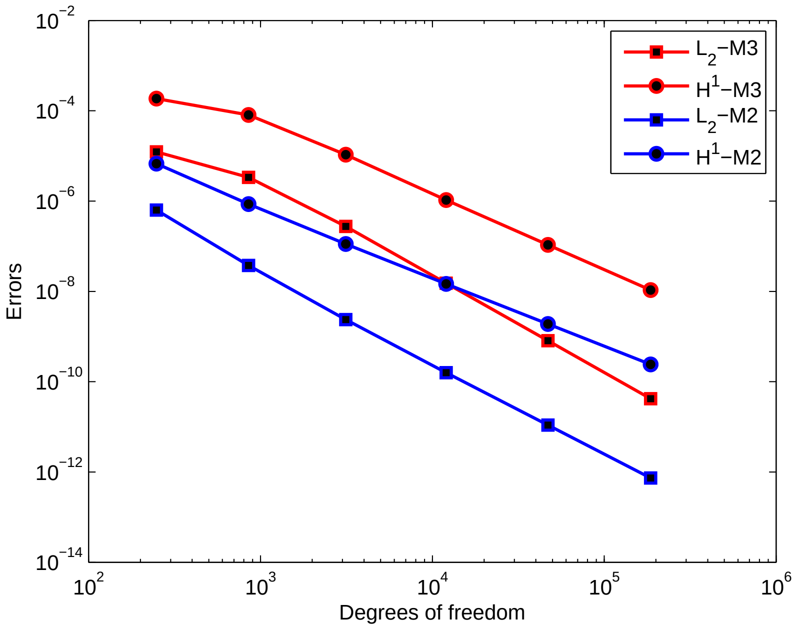

Figure 5.28: Errors obtained using Method 2 and Method 3 for non-homogeneous problem for $d = 3$.

in $\widehat{h}_1 = 0$. Proceeding in the same way ends with $\widehat{h}_i = 0, \forall i$. So $p_{d-1} = 0$ and as consequence $h_d = 0$. This shows that $\{P_{d-1}q + H_d\}$ is a direct sum hence the functions $\left\{B^{d-1}_{ijk}q, B^d_{0st} : i+j+k = d-1,\ s+t = d\right\}$ are linearly independent. Also note that

$$\#\left\{B^{d-1}_{ijk}q, B^d_{0st} : i+j+k = d-1,\ s+t = d\right\} = \binom{d+1}{2} + d + 1 = \binom{d+2}{2},$$

which is equal to the dimension of P_d.

Remark 5.7.3. We show that (5.7.45) is solvable for any choice of interpolation points z_i as soon as $[v_1, z_i] \subset T$ for all i. (Clearly, this condition is satisfied if the triangulation $\triangle$ is sufficiently fine.) Let ℓ_i denote a linear polynomial whose zero line contains the segment $[v_1, z_i]$, $i = 0, \dots, d$. Then the functions

$$\tilde{\ell}_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{d} \frac{\ell_j(x)}{\ell_j(z_i)}, \quad i = 0, \dots, d,$$

belong to $H_d = \text{span}\left\{B^d_{0jk},\ j+k=d\right\}$ and form the Lagrange basis for the interpolation problem (5.7.45), which shows its solvability.

Remark 5.7.4. We show that

$$P_{d-1}q + \text{span}\left\{B^{d+1}_{ijk},\ i+j+k=d+1,\ i=0,1\right\} = P_{d+1}.$$

Let $H_d := \text{span}\left\{B^d_{0jk},\ j+k=d\right\}$ and $H_{d+1} := \text{span}\left\{B^{d+1}_{0jk},\ j+k=d+1\right\}$. Now we see that $\text{span}\left\{B^{d+1}_{ijk},\ i+j+k=d+1, i=0,1\right\}$ is same as $H_d + H_{d+1}$. Let $qp_{d-1} + h_d + h_{d+1} \in P_{d-1}q + H_d + H_{d+1}$ such that $qp_{d-1} + h_d + h_{d+1} = 0$ then arguing in the same way as in Remark 5.7.2 implies that the set

$$\left\{B^{d-1}_{ijk}q\ :\ i+j+k=d-1\right\} \cup \left\{B^d_{0jk},\ j+k=d\right\} \cup \left\{B^{d+1}_{0jk},\ j+k=d+1\right\}$$

is linearly independent. Hence

$$\dim\left(P_{d-1}q + \text{span}\left\{B^{d+1}_{ijk},\ i+j+k=d+1, i=0,1\right\}\right) = \binom{d+3}{2} = \dim P_{d+1}.$$

Since $P_{d-1}q + H_d + H_{d+1} \subset P_{d+1}$, we conclude that $P_{d-1}q + H_d + H_{d+1} = P_{d+1}$.

Chapter 6

H^2 Polynomial Finite Elements for Curved Domains bounded by Piecewise Conics

6.1 Introduction

The purpose of this chapter is the construction of C^1 finite element space required to be used in Böhmer's method for the numerical solution of fully nonlinear second order elliptic partial differential equations over curvilinear polygons. In fact, it is an extension of C^0 polynomial finite elements, discussed in previous chapter, for domains bounded by curved boundary.

The chapter is organised as follows. Section 6.2 is to describe the full detail of construction of C^1 finite element space for curved domain, while the application of this construction in implementing Böhmer's method is discussed in Section 6.3. In the last Section we illustrate the numerical results of solving several test problems for the Monge-Ampère equation using Böhmer's method.

6.2 Construction

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain bounded by $\Gamma = \partial\Omega = \bigcup_{j=1}^{m} \overline{\Gamma_j}$, where each Γ_j is either a line segment or a conic (quadratic polynomial curve) and let $\triangle = \triangle_0 \cup \triangle_1$ be triangulation of Ω, where $\triangle_1 = \triangle_P \cup \triangle_{Bf}$ contains the buffer and pie-shaped triangles satisfying the same conditions as in Section 5.2. Again buffer triangles play the same role of maintaining the global C^1 smoothness in the interface between the patches over pie-shaped and interior triangles. We want to construct a C^1 space over Ω that resembles a modified Argyris space discussed in Chapter 3 in a sense that it has an enhanced smoothness only at the interior vertices. We use the same idea to construct these spaces as we used for the construction of C^0 curved elements *i.e.* on pie-shaped triangles we use shape functions in $P_{d-1}q \subset P_{d+1}$, P_{d+1} being the space of bivariate polynomials of degree $d+1$. It is well known that we need the degree to be quintic, at least, to have C^1 smoothness of elements (in the case of non-macro elements). For the sake of simplicity we stick to $d = 5$. The construction for higher order is the same. Thus, if

V_I is the set of interior vertices of $\triangle$,

V_B is the set of boundary vertices of $\triangle$,

$V_B^1 \subset V_B$ is the set of those boundary vertices where the tangents τ_j^+ and τ_j^- are parallel or when $\omega_j = \pi$,

E_I is the set of interior edges of $\triangle$,

E_B is the set of boundary edges of $\triangle$,

$E_{I,Bf} \subset E_I$ is the set of interior edges that are shared by a buffer and an interior triangle,

$E_{P,Bf} \subset E_I$ is the set of interior edges that are shared by a buffer and a pie-shaped triangle, then we define the space as follows

$$
\begin{aligned}
S_5^1 &:= \left\{s \in C^1(\Omega)\ :\ s \in C^2(v),\ v \in V_I \text{ and } s|_T \in P_{5+i},\ T \in \triangle_i\right\}, \qquad (6.2.1)\\
S_0 &:= \left\{s \in S_5^1\ :\ s|_{\partial\Omega} = 0\right\}, \qquad (6.2.2)
\end{aligned}
$$

where P_{5+i} is space of bivariate polynomials of degree $5+i$. Before we outline in detail the construction of a minimal determining set for the space S_0 let us introduce some more notation. For each $v \in V_I$, let $M_v := D_2(v) \cup T_v$, where T_v is one of the triangles sharing the vertex v. In the case v is also shared by a pie shaped triangle we choose $T_v \in \triangle_0$. For each edge $e \in E_I \backslash E_{P,Bf}$, let $T_e := \langle v_1, v_2, v_3\rangle$ be one of the triangles sharing the edge $e := \langle v_2, v_3\rangle$ and let $M_e := \{\xi_{122}^{T_e}\}$. In the case $e \in E_{I,Bf}$ we consider $T_e \in \triangle_0$. For each $T := \langle v_1, v_2, v_3\rangle$ in $\triangle_{Bf}$ with $v_1 \in V_B$ let $M_T^{Bf} := \{\xi_{411}, \xi_{222}\}$, (see Figure 6.2), while for each pie-shaped triangle $T := \langle v_1, v_2, v_3\rangle$, with $v_1 \in V_I$, let $M_T^P := \{\xi_{130}, \xi_{121}, \xi_{112}, \xi_{103}, \xi_{022}\} \subset D_{4,T}^*$, where $D_{4,T}^*$ is the set of domain points as defined Section 5.2, also see Figure 6.1 where the points in M_T^P are marked as black squares. Finally, for each vertex v in V_B^1 let $M_v^P := \{\xi_v\}$, where $\xi_v \in D_{4,T}^*$ is the domain point lying at v for $T \in \triangle_P$. Since $s|_{T_P} = s_p = pq \in P_6$ for some $p \in P_4$, where q is representing the curved edge of T_P, we will use the following notation for Bézier coefficients of s_p, p and q over T_P,

$$s_p = \sum_{i+j+k=6} c_{ijk} B_{ijk}^6,\ p = \sum_{i+j+k=4} p_{ijk} B_{ijk}^4 \text{ and } q = \sum_{i+j+k=2} q_{ijk} B_{ijk}^2, \qquad (6.2.3)$$

also see Remark 6.2.1.

Remark 6.2.1. Let $T := \langle v_1, v_2, v_3\rangle \in \triangle_P$ with $v_1 \in V_I$ and since $s|_T = s_p = pq \in P_6$ for some $p \in P_4$, where q is representing the curved edge of T, using the notation in (6.2.3), we have

$$pq = \left(\sum_{i+j+k=4} p_{ijk} B_{ijk}^4\right)\left(\sum_{i+j+k=2} q_{ijk} B_{ijk}^2\right) = \sum_{i+j+k=6} c_{ijk} B_{ijk}^6. \qquad (6.2.4)$$

Since q is known, (6.2.4) can be used to compute the coefficients c_{ijk}'s provided p_{rst}'s are known and vice versa. Note that (6.2.4) can also be used over a subdomain. For example let $\{c_\xi \,:\, \xi \in D_2(v_1) \cap T^* \subset D_{6,T}\}$ be known, then we get the following system by comparing the coefficients of B^6_ξ, $\xi \in D_2(v_1) \cap T^*$, in (6.2.4) given by

$$QX = C, \tag{6.2.5}$$

where

$$Q = \begin{bmatrix} q_{200} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3}q_{110} & \frac{8}{15}q_{200} & 0 & 0 & 0 & 0 \\ \frac{1}{3}q_{101} & 0 & \frac{8}{15}q_{200} & 0 & 0 & 0 \\ 0 & \frac{1}{5}q_{110} & 0 & \frac{2}{5}q_{200} & 0 & 0 \\ \frac{1}{15}q_{011} & \frac{2}{15}q_{101} & \frac{2}{15}q_{110} & 0 & \frac{4}{15}q_{200} & 0 \\ 0 & 0 & \frac{1}{5}q_{101} & 0 & 0 & \frac{2}{5}q_{200} \end{bmatrix} \text{ and } C = \begin{bmatrix} c_{600} \\ c_{510} \\ c_{501} \\ c_{420} \\ c_{411} \\ c_{402} \end{bmatrix}$$

with unknown vector $X = [p_{400}\ p_{310}\ p_{301}\ p_{220}\ p_{211}\ p_{202}]^t$. Since $q_{200} = 1$, it is easy to see that the system (6.2.5) is uniquely solvable and stable. Figure 6.1 depicts the domain points for the Bézier net of p where the domain points corresponding to unknown coefficients in X are marked as circles.

Then we arrive at the following result.

Theorem 6.2.2. *The set*

$$M_0 := \bigcup_{v\in V_I} M_v \cup \bigcup_{e\in E_I\setminus E_{P,Bf}} M_e \cup \bigcup_{T\in\triangle_{Bf}} M_T^{Bf} \cup \bigcup_{T\in\triangle_P} M_T^P \cup \bigcup_{v\in V_B^1} M_v^P, \tag{6.2.6}$$

is a stable local minimal determining set for the space S_0.

Proof. We set coefficients $\{c_\xi \,:\, \xi \in M_0\}$ for any spline $s \in S_0$ and show that all other coefficients of s can be determined from them consistently. We discuss different cases separately for the sake of simplicity and show how we compute the full Bézier net for $s \in S_0$.

Case I : For each $v \in V_I$ such that v is not shared by any pie shaped triangle then points in M_v are same as for Argyris space thus all coefficients $\{c_\eta : \eta \in D_2(v) \backslash M_v\}$ are consistently determined by [43, Lemma 5.10], in view of C^2 smoothness conditions, by first setting the coefficients $\{c_\xi : \xi \in M_v\}$.

Case II : Let $v \in V_I$ be such that v is shared by some $T \in \triangle_P$. In this case we consider $T_v \in \triangle_0$ and set the coefficients $\{c_\xi : \xi \in M_v\}$ of $s \in S_0$ to arbitrary values. Let $T_P \in \triangle_P$ and $T_{Bf} \in \triangle_{Bf}$ be among the triangles attached to v. Then we determine the rest of the coefficients of s over $D_2(v)$ for quintic polynomials over all triangles attached to v in a similar way as in Case I. But recall that $s|_{T_{Bf}} \in P_6$ and $s|_{T_P} = s_p = pq \in P_6$ for some $p \in P_4$, where q is representing the curved edge of T_P. Hence the computed coefficients at $\eta \in D_2(v) \cap T_{Bf} \subset D_{5,T_{Bf}}$ and $\eta \in D_2(v) \cap T_P \subset D_{5,T_P}$ are not the coefficients of s. Therefore we denote them by $\tilde{c}_\eta$. Hence for T_{Bf} we compute the coefficients $\{c_\xi : \xi \in D_2(v) \cap T_{Bf} \subset D_{6,T_{Bf}}\}$ for s from already known $\{\tilde{c}_\eta : \eta \in D_2(v) \cap T_{Bf} \subset D_{5,T_{Bf}}\}$ by using the degree raising formulas with $r = 1$ over a subdomain $D_2(v) \cap T_{Bf}$, see Remark 3.1.4.

Now we turn to a pie shaped triangle T_P. For T_P we need to go one more step apart from degree raising over $D_2(v)$ as done for T_{Bf}. In fact we need to determine the coefficients $\{p_\xi : \xi \in D_2(v) \cap T_P \subset D^*_{4,T_P}\}$ for polynomial $p \in P_4$. To this end we first determine the coefficients $\{c_\xi : \xi \in D_2(v) \cap T_P \subset D_{6,T_P}\}$ for s from already known $\{\tilde{c}_\eta : \eta \in D_2(v) \cap T_P \subset D_{5,T_P}\}$ by using the degree raising formulas with $r = 1$ over a subdomain $D_2(v) \cap T_P$. Then we solve the system (6.2.5) for the unknown coefficients $\{p_\xi : \xi \in D_2(v) \cap T_P \subset D^*_{4,T_P}\}$, see Remark 6.2.1.

Case III : For each $e \in E_I \backslash \{E_{I,Bf} \cup E_{P,Bf}\}$ the point in M_e is the same as for Argyris space so we restrict our discussion to each $e \in E_{I,Bf} \subset E_I$. Let $e := \langle v_2, v_3 \rangle$ be the edge shared by $T_e := \langle v_1, v_2, v_3 \rangle \in \triangle_0$ and $\tilde{T}_e := \langle v_4, v_3, v_2 \rangle \in \triangle_{Bf}$. We set $c^{T_e}_{122}$ of s to an arbitrary value. In view of Case I and II the full Bézier net of $s|_{T_e}$ of degree 5 has been determined thus we determine $c^{\tilde{T}_e}_{132}$ and $c^{\tilde{T}_e}_{123}$ by first raising the degree of $s|_{T_e}$ by 1 and then apply the C^1 smoothness conditions across the edge e.

Case IV : For $T := \langle v_1, v_2, v_3 \rangle \in \triangle_{Bf}$ with $v_1 \in V_B$ we set the coefficients c^T_{411} and c^T_{222} of $s \in S_0$. Let $e := \langle v_1, v_3 \rangle \in E_{P,Bf}$ and let $T_2 := \langle v_3, v_4, v_1 \rangle \in \triangle_P$ share e with T. Now we show how we determine the coefficient p_{013} for the polynomial p in $s|_{T_2} = pq$. For this it is easy to see that C^1 smoothness conditions across e gives us the equation

$$c^{T_2^*}_{114} = b_1 c^T_{501} + b_2 c^T_{411} + b_3 c^T_{402}, \tag{6.2.7}$$

where (b_1, b_2, b_3) are barycentric coordinates of v_4 w.r.t. T and T_2^* is straight triangle associated with T_2. Note that the coefficients of p corresponding to points in $M^P_{T_2}$ and $M^P_{v_1}$ (if $v_1 \in V^1_B$) are already set to arbitrary values. They determine $c^{T_2^*}_{204}$ and $c^{T_2^*}_{105}$ for $s|_{T_2^*} = pq$ and in view of C^0 smoothness conditions we have

$$c^T_{402} = c^{T_2^*}_{204} \text{ and } c^T_{501} = c^{T_2^*}_{105}.$$

Moreover, comparison of coefficients of B^6_{114} on both sides of (6.2.4) results in the equation

$$15 c^{T_2^*}_{114} = q_{110} p_{004} + 4 q_{101} p_{013} + 4 q_{011} p_{103}. \tag{6.2.8}$$

Thus we first compute $c^{T_2^*}_{114}$ using (6.2.7) and then we use (6.2.8) to determine p_{013} where the coefficients p_{103} and p_{004} are already prescribed in $M^P_{T_2}$ and $M^P_{v_1}$ respectively. Similarly, p_{031} is computed using $M^P_{T_2}$ and $M^P_{v_4}$ and the coefficients of s on the buffer triangle sharing $e := \langle v_1, v_2 \rangle \in E_{P,Bf}$ with T_2, see Figure 6.1. This completes the full Bézier net $\{p_\xi : \xi \in D^*_{4,T_2}\}$ of p for $s|_{T_2} = pq$. The Bézier net of $s|_{T_2}$ can be obtained by multiplying p with q. Now the coefficients c^T_{312} and c^T_{213} can be determined using C^1 smoothness conditions across e.

To prove that M_0 is local in the sense of Definition 3.1.6 let c_η be the Bézier coefficient of a spline s with $\eta \notin M_0$ but $\eta \in T_\eta$. Then it is easy to see that Γ_η is always contained in star(T_η), which results in the locality of M with $l = 1$.

We now show that M_0 is stable as defined in the Definition 3.1.7. Since all Bézier coefficients $\{c_\eta : \eta \notin M_0\}$ for $s \in S_0$ can be computed by the processes of product of polynomials, degree raising of polynomials and solution of system (6.2.5) in conjunction with C^1 smoothness conditions, which are stable processes, see Section 3.1.3, Section 3.1.6, Remark 6.2.1 and [43, Lemma 2.29]. Hence M_0 is a stable MDS for the space S_0. ■

An example of an MDS for the space S_0 over a circular disk is depicted in Figure 6.3, where the points in the sets $\bigcup_{v\in V_I} M_v$, $\bigcup_{e\in E_I\setminus E_{P,Bf}} M_e$, $\bigcup_{v\in V_B^1} M_v^P$, $\bigcup_{T\in\triangle_{Bf}} M_T^{Bf}$ and $\bigcup_{T\in\triangle_P} M_T^P$ are marked as black dots, diamonds, triangles, downward pointing triangles and squares respectively. Note that $V_B^1 = V_B$ for this example.

6.3 Implementation of Böhmer's Method using C^1 Curved Elements and Numerical Results

To judge the performance of our construction of C^1 elements for curved domains we implement Böhmer's method for several fully nonlinear equations using S_0 as

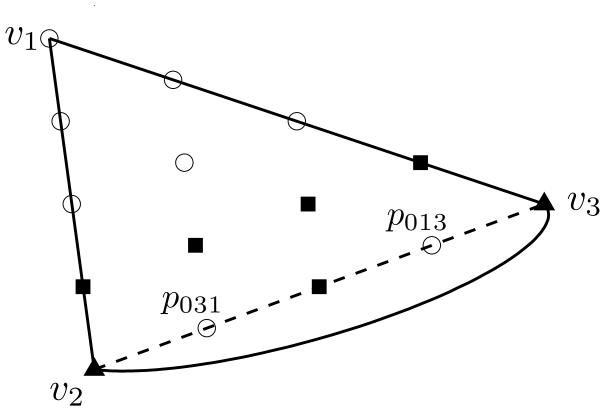

Figure 6.1: The points for the Bézier net of polynomial p over a pie shaped triangle. Points of M_T^P are marked as black squares, whereas points of $M_{v_2}^P \cup M_{v_3}^P$, in case $v_2, v_3 \in V_B^1$, are marked as triangles.

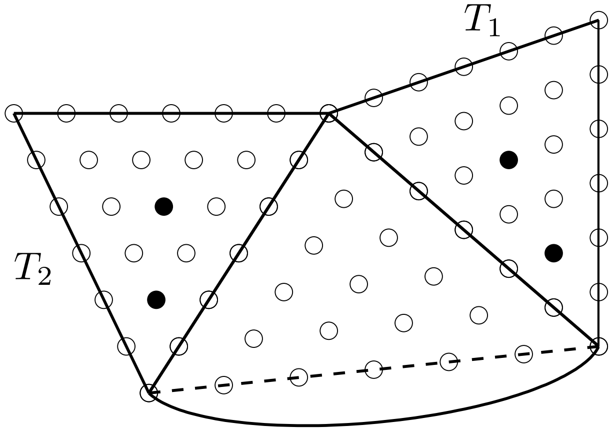

Figure 6.2: The MDS points in the set $M_{T_1}^{Bf} \cup M_{T_2}^{Bf}$ are marked as black dots.

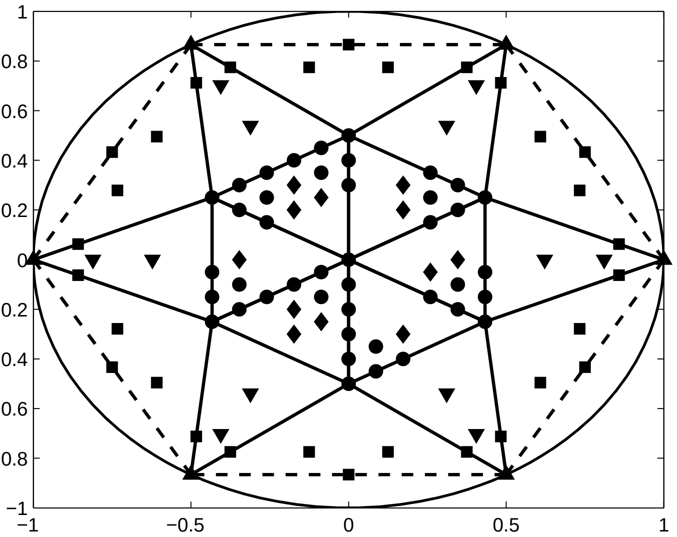

Figure 6.3: Example of MDS for the space S_0 over a circular domain Ω.

an approximation space. We study the numerical approximation of the Dirichlet problem (2.2.13)-(2.2.14) for a fully nonlinear equation of second order.

In fact we focus on the prototypical and most interesting Monge-Ampère equation. Recall that the Dirichlet problem for the Monge-Ampère equation can be formulated by

$$\begin{aligned} G(u) = \det(\nabla^2 u) - g(x) = &\ 0, \quad x \in \Omega \\ u = &\ \phi, \ x \in \partial\Omega \end{aligned} \tag{6.3.9}$$

where $g \in L_2(\Omega)$ and $\phi \in L_2(\partial\Omega)$ are given functions with $g > 0$ on Ω required to keep the Monge-Ampère operator elliptic and note that, for the Monge-Am̀pere equation, we have an additional condition of convexity of u for the sake of uniqueness.

We follow the same lines for the implementation of Böhmer's method as outlined, in detail, in Section 4.2. Therefore we keep ourselves very brief in recalling some of the details of the implementation.

Let $\triangle^h = \triangle_0^h + \triangle_1^h$ be a quasi-uniform triangulation of curved domain $\Omega \subset \mathbb{R}^2$.

We get a sequence of meshes $\triangle^h$ by uniform refinement whereby each triangle is subdivided into four triangles, on each level, by joining the midpoints of every edge. For a pie shaped triangle with Γ_j being the corresponding curved boundary edge we take midpoint of the curved boundary edge Γ_j, see Figure 5.11. To solve (6.3.9) by Böhmer's method is, in fact, to perform the Newton-Kantorovich iterative scheme, to get a sequence of $\{u_k^h\}_{\mathbb{Z}_+}$ of approximations of $\hat{u}$, generated by (4.2.1). The difference is that we use the approximating space S_0^h, defined in (6.2.2), with an M_0-basis $\{s_1, s_2, \ldots, s_N\}$ where M_0 is MDS for the space S_0^h as proved in Theorem 6.2.2. We obtain the basis $\{s_1, s_2, \ldots, s_N\}$ for S_0^h using (3.1.11). Again we use the relations (4.2.7) to assemble the element matrices for which we obtain the transformation matrix $\mathcal{T}$ using relation (4.2.5) with basis functions $\{s_1, s_2, \ldots, s_N\}$, where the required block diagonal matrices are obtained from (5.5.32)-(5.5.34) with degree $d = 5$. We consider $\phi = 0$ in all of our test problems.

6.3.1 Numerical Results

1. We consider the *first test problem* over a closed unit circular disk centred at the origin and choose data function g such that $u = e^{0.5(x_1^2+x_2^2)} - e^{0.5}$ is a classical solution of (6.3.9) with $\phi = 0$. The numerics for the problem are compiled in Table 6.1. Since the solution is infinitely smooth we see the convergence rate for the Böhmer's method approaching $O(h^4)$ in the H^2-norm as expected, while the behaviour of errors in the L^2 and H^1-norms is also near optimal. k denotes the number of Newton's iterations to get to the best solution on corresponding level of triangulation. Note that we use the approximate solution of the Poisson problem (4.3.11) as an initial guess for the Newton method only on the first level. On the next levels we again use the multilevel approach as we used for polygonal domains in Chapter

Table 6.1: Errors of approximate solution and rate of convergence for the 1st test problem over a circular domain.

Levels(l)	L^2-error	rate	H^1-error	rate	H^2-error	rate	k
initial	1.04e-2		3.20e-2		1.85e-1		0
1	2.12e-6		3.84e-5		1.25e-3		2
2	2.98e-7	2.83	8.47e-6	2.18	3.35e-4	1.90	1
3	6.79e-9	5.46	3.87e-7	4.49	2.86e-5	3.55	1
4	1.36e-10	5.64	1.46e-8	4.69	2.12e-6	3.76	1
5	2.52e-12	5.76	5.23e-10	4.81	1.47e-7	3.86	1
6	9.51e-14	4.73	1.76e-11	4.89	9.53e-9	3.93	1

3, where we use the quasi-interpolant of the approximate solution at the previous level as initial guess. Though the Newton method converges on each level if we use the approximate solution of (4.3.11) as an initial guess but obviously the multilevel approach is efficient as we see, from the value of k, that we need only one Newton's method iteration to get to the solution on the corresponding level. The stopping criteria for Newton's iterations, on each level, is when

$$\|u_k^h - u_{k+1}^h\| < \epsilon,$$

for $\epsilon = 10^{-15}$, where u_k^h is the approximate solution at the kth Newton's iteration.

2. In this case we consider Ω to be an elliptic disk having boundary with varying curvature bounded by $x_1^2 + 6.25x_2^2 = 1$. We consider different test problems by considering $g_1 = e^{x_1}$ and, a significantly less smooth function,

$g_2 = \sin(\pi|x_1|) + 1.1$ for (6.3.9). [33, Theorem 17.22] assures that there exists a solution $u \in C^{2,\alpha}(\overline{\Omega}),\ 0 < \alpha < 1$, at least, for g_1 which is enough regularity for Böhmer's theory to be applicable. In case of g_2 the structure conditions required in [33, Theorem 17.22] are not satisfied. Nevertheless we apply the method and numerically look for the solution. Since we do not know the solution we look for the size of residual functions, denoted by R, to see the behaviour of Böhmer's method, where

$$R = \|G(u_k^{h,l})\|_{L_2(\Omega)},$$

and $u_k^{h,l}$ is the approximate solution to (6.3.9) at the kth Newton's iteration on level l. For the sake of convenience let us use the notation $\varepsilon^{h,l} = u^{h,l} - u^{h,l+1}$ in the sequel. We also compute the errors between approximate solutions of (6.3.9) on consecutive levels *i.e.* we compute $\|\varepsilon^{h,l}\|$ in L_2, H^1 and H^2-norms because

$$\frac{\|\varepsilon^{h,l}\|}{\|\varepsilon^{h,l+1}\|},$$

indicates the convergence behaviour of the method in the corresponding norm, see Remark 6.3.1. We see that Böhmer's method converges for g_2 as well. Comparing the convergence for both test problems shows the slow pace of convergence for g_2 which is obvious due to its lack of smoothness because of the presence of $|x_1|$.

3. In the *third test problem* we consider (6.3.9) with data $g = 1$ and $\phi = 0$ over domains Ω_1 and Ω_2 with $\partial\Omega_1 \in C^1$ and $\partial\Omega_2 \in C^2$, see Remark 6.3.3. [33, Theorem 17.22] demands a uniformly convex C^3 boundary, at least, for the existence of a classical solution u (*i.e.* $u \in C^{2,\alpha}(\overline{\Omega}),\ 0 < \alpha < 1$) for (6.3.9) in this case. Since the theorem provides only sufficient conditions we want to see numerically whether a smooth solution is likely to exist on domains with

Table 6.2: Errors for the 2nd test problem over an elliptic disk with $g_1 = e^{x_1}$.

l	$\|\varepsilon^{h,l}\|_{L_2}$	rate	$\|\varepsilon^{h,l}\|_{H^1}$	rate	$\|\varepsilon^{h,l}\|_{H^2}$	rate	R	k
							6.58e-1	0
1	1.02e-8		3.64e-7		2.90e-5		4.95e-6	4
2	9.59e-10	3.4	5.26e-8	2.8	6.37e-6	2.2	1.62e-6	1
3	1.32e-11	6.2	1.29e-9	5.3	3.16e-7	4.3	1.37e-7	1
4	2.25e-13	5.9	4.27e-11	4.9	2.05e-8	3.9	9.83e-9	1
5	8.79e-15	4.7	1.61e-12	4.7	1.56e-9	3.7	6.61e-10	1
6							4.33e-11	1

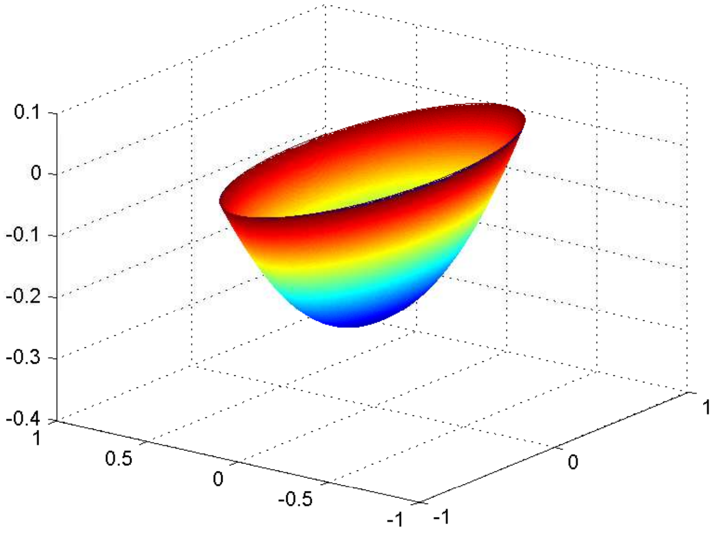

Figure 6.4: The approximate solution u^h of (6.3.9) with $\phi = 0$ and $g_2 = \sin(\pi|x_1|) + 1.1$ on 6th level of triangulation.

Table 6.3: Errors for the 2nd test problem over an elliptic disk with $g_2 = \sin(\pi|x_1|) + 1.1$.

l	$\|\varepsilon^{h,l}\|_{L_2}$	rate	$\|\varepsilon^{h,l}\|_{H^1}$	rate	$\|\varepsilon^{h,l}\|_{H^2}$	rate	R	k
							1.06e+0	0
1	2.92e-5		9.88e-4		9.48e-2		1.92e-2	3
2	5.41e-6	2.4	6.20e-5	3.9	4.44e-3	4.4	6.23e-3	2
3	1.21e-6	2.2	1.19e-5	2.4	1.40e-3	1.7	2.03e-3	1
4	6.84e-8	4.1	2.01e-6	2.6	4.90e-4	1.5	7.46e-4	1
5	1.44e-8	2.3	3.67e-7	2.5	1.47e-4	1.7	2.47e-4	1
6							9.04e-5	1

C^1 or C^2 boundary. Note that [33, Theorem 17.17] says that the solution exists which is C^2 inside the domain (in case of both Ω_1 and Ω_2). For Ω_2 we see that Böhmer's method converges with about $O(h^4)$, $O(h^3)$ and $O(h^2)$ in the L_2, H^1 and H^2-norm, respectively, see Table 6.4, which indicates that the solution should be in $H^\gamma(\Omega)$ with γ approaching 4. We again use the multilevel approach and compute the convergence rate in a same way as we did for the 2nd test problem. The domain Ω_2 is visualized in Figure 6.5 with initial triangulation where the points marked as circles on the boundary are C^2 joints (the boundary is infinitely smooth at other points). Let RF be a residual function defined by

$$RF = G(u^{h,l}) = \det(\nabla^2 u^{h,l}) - 1,$$

where $u^{h,l}$ is the approximate solution of (6.3.9) at level l. The cross section of RF, at level 6, along the straight line that passes through those boundary points where we have C^2 smoothness, is plotted in Figure 6.6. It clearly shows the mild singularities at the end points which slow down the convergence of the method. In case of Ω_1 the method does not converge but the behaviour of the method is different from that observed for the square domain in a sense that the approximate solution keeps the convex shape and also the Newton method converges on each level. Also see the fifth test problem in Section 3.3.1 and compare Figure 4.3 and Figure 6.9. Figure 6.7 depicts the cross section of the residual function RF, at level 4, along the straight line $x_2 = x_1$ that passes through the boundary points with C^1 smoothness. It shows that the singularity is stronger compared to Ω_2 that caused divergence in this case but, as expected, it is less strong compared to the square domain, see Figure 4.2. Figure 6.7 also demonstrates that the method tries to converge inside of the domain Ω_1 away from the singularity. Moreover, Figure 6.8 shows the plot of cross section of RF along the straight line $x_1 = 0$. Singularities at the end points are in fact due to straight line boundary segments for Ω_1, which was discussed in Section 2.1.2 that the Monge-Ampère equation also has singularities along straight line boundary segments.

Remark 6.3.1. Let u be an exact solution to (6.3.9). If $u^{h,l}$ denotes the approximate solution at level l then, in the case that method converges, we can assume

$$\|u - u^{h,l+1}\| \leq \gamma \|u - u^{h,l}\|,$$

for some $\gamma < 1$. The triangular inequality, then, leads us to

$$\|u - u^{h,l}\| \leq \frac{1}{1-\gamma} \|u^{h,l+1} - u^{h,l}\| = \frac{1}{1-\gamma} \|\varepsilon^{h,l}\|,$$

and hence the ratio

$$\frac{\|\varepsilon^{h,l}\|}{\|\varepsilon^{h,l+1}\|},$$

Table 6.4: Errors for the 3rd test problem over domain Ω_2.

l	$\|\varepsilon^{h,l}\|_{L_2}$	rate	$\|\varepsilon^{h,l}\|_{H^1}$	rate	$\|\varepsilon^{h,l}\|_{H^2}$	rate	R	k
							2.01e+0	0
1	1.07e-3		1.00e-2		1.34e-1		9.10e-2	2
2	4.87e-5	4.5	8.56e-4	3.5	2.20e-2	2.6	2.20e-2	1
3	3.04e-6	4.0	1.04e-4	3.0	5.30e-3	2.0	5.87e-3	1
4	2.09e-7	3.7	1.39e-5	2.9	1.38e-3	1.9	1.56e-3	1
5	1.58e-8	3.7	2.01e-6	2.8	3.80e-4	1.9	4.15e-4	1
6							1.11e-4	1

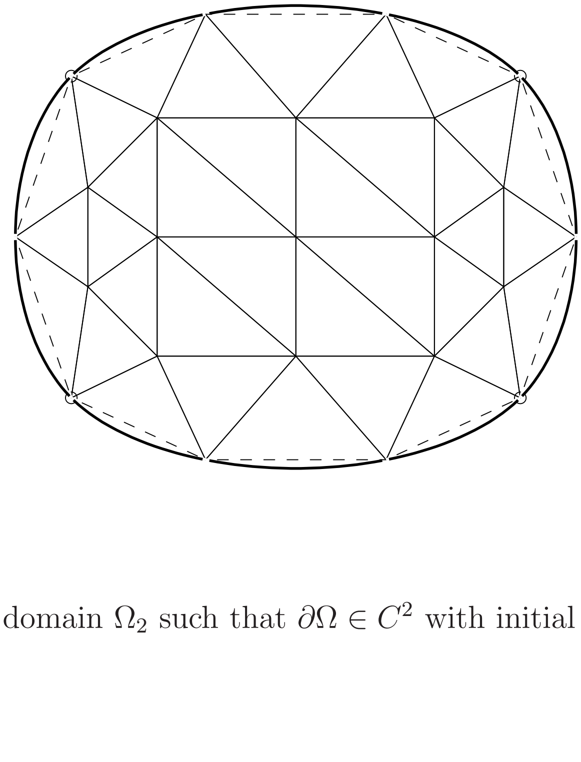

Figure 6.5: The domain Ω_2 such that $\partial\Omega \in C^2$ with initial triangulation.

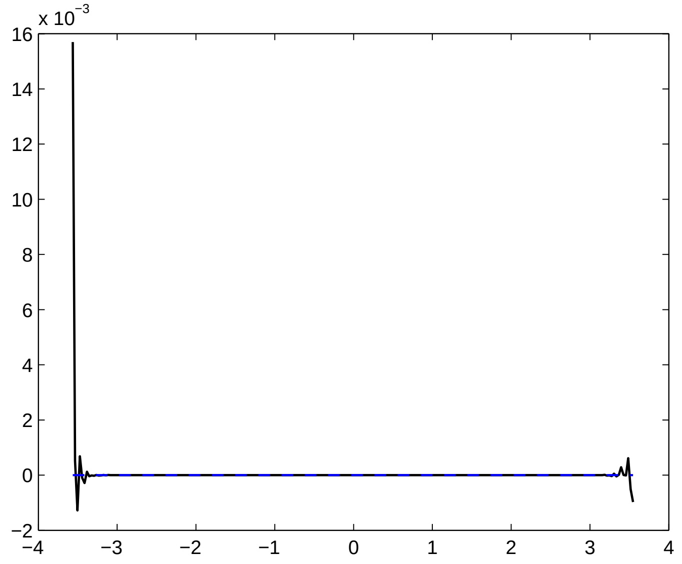

Figure 6.6: The cross section of $RF = G(u^{h,6}) = \det(\nabla^2 u^{h,6}) - 1$ along the straight line $x_2 = 0.45x_1$ for Ω_2.

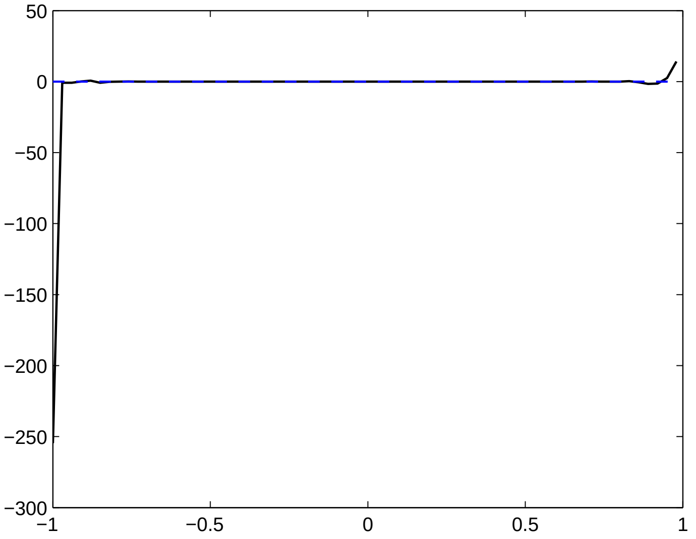

Figure 6.7: The cross section of $RF = G(u^{h,4}) = \det(\nabla^2 u^{h,4}) - 1$ along the straight line $x_2 = x_1$ for Ω_1.

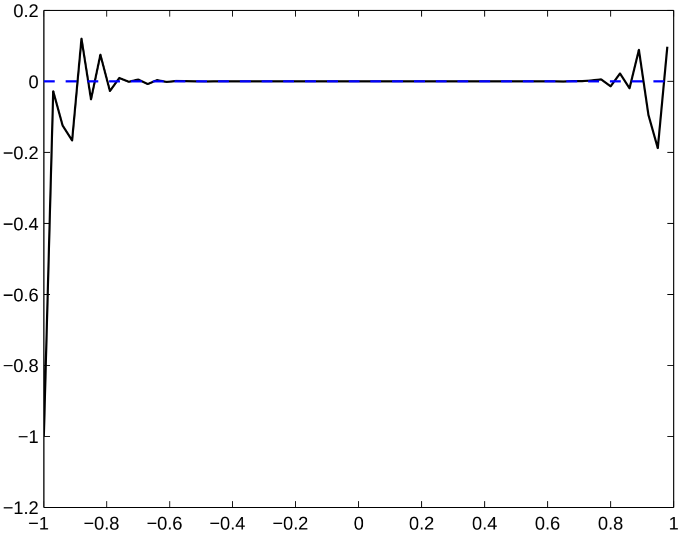

Figure 6.8: The cross section of $RF = G(u^{h,4}) = \det(\nabla^2 u^{h,4}) - 1$ along the straight line $x_1 = 0$ for Ω_1.

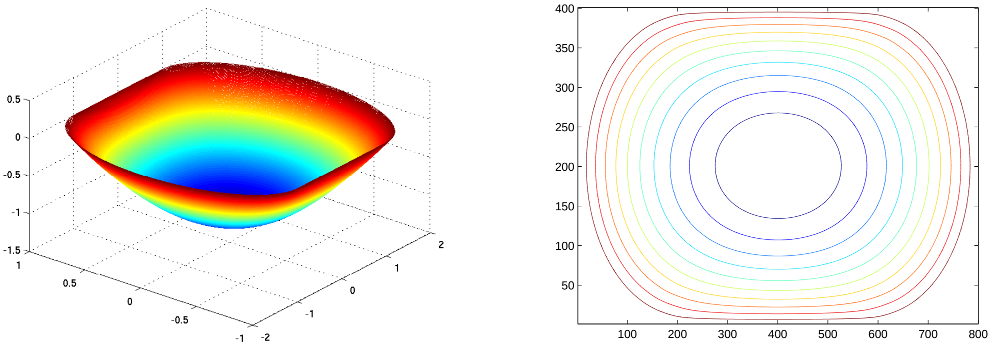

Figure 6.9: Approximate solution u^h of test 3 on Ω_1 and its contour plot on second level of triangulation.

will indicate the convergence of the method in the corresponding norm.

Remark 6.3.2. If Ω is a domain bounded by $ax_1^2+bx_2^2=1$ then the homogeneous Dirichlet problem for (6.3.9) with $g=1$ has a quadratic polynomial solution

$$u=\frac{ax_1^2+bx_2^2-1}{2\sqrt{ab}},$$

which is approximated exactly by our method.

Remark 6.3.3. The domain Ω_1 with $\partial\Omega \in C^1$ is bounded by straight lines $x_2 = \pm 1$ and semi-circles

$$x_1=\pm\left(1+\sqrt{1-x_2^2}\right),\ -1\le x_2\le 1.$$

Then it is easy to see that the points $\pm(1,1)$ and $\pm(1,-1)$ are C^1 joints on the boundary (but they are not C^2).

We design the domain Ω_2 with $\partial\Omega \in C^2$ using the concept of osculating circle to a given curve at any point that has the same curvature at that point. The domain Ω_2 is bounded by elliptic pieces

$$1.69x_1^2+16(x_2\pm 1.03)^2=27.04 \text{ with } |x_1|\le 3.56,$$

and circular pieces

$$(x_1\pm 2.53)^2+x_2^2=3.69 \text{ with } |x_2|\le 1.62.$$

Then it is easy to see that the points $\pm(3.56,1.62)$ and $\pm(3.56,-1.62)$ on the boundary are C^2 joints.

Chapter 7

Conclusions and Future Work

Conclusions

The main objective of the work done in this thesis was the construction of bivariate spline finite element spaces for curved domains that lead to new conforming finite element methods to solve second order elliptic PDEs on curved domains, keeping in mind that these spaces must possess an extra property called stable splitting so that they can also be used in Böhmer's method to solve second order fully nonlinear elliptic problems. We summarise our achievements in this regard and also report possible future directions below.

1. As our first goal we study different C^1 spaces on polygonal domains. We first return to modified Argyris space using Bernstein-Bézier techniques and show that property of stable splitting can also be formulated as splitting of a minimal determining set for the space. We also looked for lower order C^1 spaces whether they possess the property of stable splitting so that there might be other choices, to be used as approximating spaces, while using Böhmer's finite element method to solve nonlinear problems over polygonal

domains. We end up with the answer that some of the C^1 macro-element spaces, including Clough-Tocher and Powell-Sabin macro-element spaces, are also among the list as candidates for Böhmer's method.

To analyse Böhmer's method numerically we implement the method using modified Aygyris space as a solution space and solve several benchmark Dirichlet problems for fully nonlinear PDEs on polygonal domains, including the prototypical Monge-Ampère equation. Numerical results confirm the theoretical results.

2. For curved domains, bounded by piecewise conics, we construct C^0 bivariate spline spaces of arbitrary order and develop a new H^1 conforming finite element method to solve Dirichlet problems for elliptic PDEs. The method is in fact isogeometric method in a sense that the shape functions we use on pie shaped triangle also describe the exact geometry of the domain.

   The numerical experiments for several second order elliptic linear problems, including eigenvalue problem, over different curved domains show optimal results both for lower and higher order elements. Since our construction allows us to use higher order elements, if needed, we also looked for numerical results in the context of the p method. Results are again up to expectations with typical exponential decay of errors for p method.

3. Last and the most important achievement is the construction of H^2 conforming bivariate spline spaces over curved domains. The construction gave us ability to solve Dirichlet problems for second order fully nonlinear elliptic equations over curved domains using Böhmer's method.

   To see numerical results we considered several Dirichlet problems for the Monge-Ampère equation over different curved domains using Böhmer's method.

The results again endorse the proved theoretical results by Böhmer.

Future Work

1. Improvement of the methods to deal with non-homogeneous Dirichlet boundary conditions for C^0 curved elements.

   Development of such methods to C^1 curved elements.

2. Proof of results providing error bounds (both for C^0 and C^1 elements).

3. Adaptive approach in conjunction with Böhmer's method to tackle singular problems (like in Test 3 in Chapter 6).

4. Extension of the same idea to construct C^1 spaces to solve linear fourth order equations over curved domains.

5. 3-D curved C^0 elements over domains bounded by piecewise quadrics. Further generalizations to domains bounded by piecewise higher order implicit curves or surfaces.

Bibliography

[1] M. Ainsworth, G. Andriamaro, and O. Davydov, *Bernstein-Bézier finite elements of arbitrary order and optimal assembly procedures*, SIAM J. Sci. Comp., 33 (2011), 3087-3109.

[2] V. Anh-Vu, *Adaptive Hierarchical Isogeometric Finite Element Methods*, Springer-Spektrum, 2012.

[3] T. Aubin, *Nonlinear Analysis on Manifolds, Monge-Ampère equation*, Springer-Verlag, Inc Berlin, (1982).

[4] G. Awanou, *Pseudo time continuation and time marching methods for Monge-Ampère type equations*, manuscript, `http://www.math.niu.edu/~awanou/`.

[5] I. Babuška, Manil Suri, *The* P *and* H-P *Versions of the Finite Element Method, Basic Principles and Properties*, SIAM Review, 36(4), pp. 578-632, (1994).

[6] I. Babuška, and J. Osborn, *Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues*, SIAM J. Numer. Anal., 24(6) 1987.

[7] U. Banerjee, M. Suri, *The effect of numerical quadrature in the* p *version of the finite element method*, Math. Comp., 59, pp. 1-20.

[8] M. Bernadou, *Curved finite elements of class C^1: Implementation and numerical experiments. Part 1: Construction and numerical tests of the interpolation properties*, Comput. Method Appl. Mech. Engrg., 106(1-2) (1993), pp. 229-269.

[9] S. N. Bernstein, *Demonstration du théorème de Weierstrass fondée sur le calcul des probabilités,* Communications de la Société Mathématique de Kharkov (13), (1912/13) 1-2.

[10] K. Böhmer, *On finite element methods for fully nonlinear elliptic equations of second order*, SIAM J. Numer. Anal., 46 (3) (2008), 1212-1249.

[11] K. Böhmer, *Numerical Methods for Nonlinear Elliptic Differential Equations*: A Synopsis, Oxford University Press, Oxford, 2010.

[12] S. C. Brenner, and L. R. Scott, *The Mathematical Theory of Finite Element Methods,* Springer, New York, 1994.

[13] S. C. Brenner, M. Neilan, T. Gudi, L. Y. Sung, *C^0 penalty methods for the fully nonlinear Monge-Ampère equation*, Math. Comput., 80(276), 1979-1995 (2011).

[14] S. C. Brenner, M. Neilan, *Finite element approximations of the three dimensional Monge-Ampère equation*, M2AN Math. Model. Numer. Anal., 46(5), 979-1001 (2012).

[15] J. D. Benamou, Y. Brenier, *A computational fluid mechanics solution to Monge-Kantorovich mass transfer problem*, Numer. Math., (84) (2000), 375-393.

[16] L. A. Caffarelli, X. Cabré, *Fully Nonlinear Elliptic Equations*, American Mathematical Society, 1995.

[17] J. M. Carnicer, M. García-Esnaola, *Lagrange interpolation on conics and cubics*, Computer Aided Geometric Design, 19 (2002), pp. 313–326.

[18] Philippe G. Ciarlet, *The Finite Element Method for Elliptic Problems*: North-Holland, Amsterdam, 1978. Reprinted as Classics in Applied Mathematics 40, SIAM. Philadelphia, 2002.

[19] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vo.l II, Wiley Interscience, 1989.

[20] O. Davydov, *Smooth finite elements and stable splitting*, Berichte "Reihe Mathematik" der Philipps-Universität Marburg, 2007-4 (2007). An adapted version has appeared as [11, Section 4.2.6].

[21] O. Davydov and A. Saeed, *Stable splitting of bivariate splines spaces by Bernstein-Bézier methods*, in J.-D. Boissonnat et al. (Eds.): Curves and Surfaces - 2011, LNCS 6920, pp. 220–235, 2011.

[22] O. Davydov and A. Saeed, *Numerical Solution of Fully Nonlinear Elliptic equations by Böhmer's Method*, University of Strathclyde Department of Mathematics and Statistics Research Report, 2012-01.

[23] O. Davydov, G. Kostin and A. Saeed, *Polynomial Finite Element Method for Domains Enclosed by Piecewise Conics*, in preparation.

[24] E. J. Dean and R. Glowinski, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Computer Methods in Applied Mechanics and Engineering, 195 (2006), 1344-1386.

[25] E. J. Dean, R. Glowinski, *Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach*, C. R. Acad. Sci. Paris, Ser. I 336(2003)779-784.

[26] E. J. Dean, R. Glowinski, *On the numerical solution of the elliptic Monge-Ampère equation in dimension two: a least squares approach*, In Partial differential equations, Comput. Methods Appl. Sci., 16 pp. 43-63. Springer, Dordrecht, 2008.

[27] E. J. Dean, R. Glowinski, T. W. Pan, *Operator splitting methods and applications to the direct simulation of particulate flow and to the solution of the elliptic Monge-Ampère equation*, In Control and boundary ananlysis, volume 240 of Lect. Notes Pure Appl. Math., 1-27, chapman and Hall/CRC, Boca Raton, FL, 2005.

[28] Rida T. Farouki, *The Bernstein polynomial basis: a centennial retrospective*, Computer Aided Geometric Design, 29(6) (2012) pp. 379-419.

[29] G. Farin, *Curves and surfaces for CAGD: a practical guide, Fifth edition,* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[30] X. Feng, M. Neilan, *Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method*, SIAM J. Numer. Anal., 47(2) 1226-1250, 2009.

[31] X. Feng, M. Neilan, *Vanishing moment method and moment solution for fully nonlinear second order partial differential equations*, J. Sci. Comput., 38(1) 78-98, 2009.

[32] J. Foley, A. Van Dan, S. Feiner, and J. Hughes, *Computer Graphics: Principles and Practice,* Adison-Wesley Publishing Company Inc., 1996.

[33] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, (2001).

[34] C. G. Gibson, *Elementary Geometry of Algebraic curves*: An undergraduate introduction, Cambridge University Press, (1998).

[35] K. Höllig, U. Reif, J. Wipper, *Weighted extended B-spline approximation of Dirichlet problem*, SIAM J. Numer. Anal., 39 (2) (2001), 442-462.

[36] J. Hoschek, and D. Lasser, *Fundamentals of Computer Aided Geometric Design,* A K Peters Ltd, Wellesley, MA, 1993. Transleted from the 1992 German edition by Larry L. Schumaker.

[37] T. J. R. Hughes, J. A. Cottrel, Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement*, Comput. Methods Appl. Mech. Engrg., 194(2005) 4135-4195.

[38] R. Kirby, *Fast simplicial finite element algorithms using Bernstein polynomials*, Numer. Math., (2010), 1-22. 10.1007/s00211-010-0327-2.

[39] J. R. Kuttler, V. G. Sigillito, *Eigenvalues of the Laplacian in Two Dimensions*, SIAM, 26(2) (1984), 163–193.

[40] M. Lenoir, *Optimal isoparametric fintie elements and error estimates for domains involving curved boundaries*, SIAM. J. Numer, Anal., 23 (1986), 562-580.

[41] G. Loeper, F. Rapetti, *Numerical solution of the Monge-Ampère equation by a Newton's method*, C. R. Acad. Sci. Paris, Sér. I 340(4), 319-324 (2005).

[42] O. Lakkis, T. Pryer, *A non-variational finite element method for the nonlinear elliptic problems*, manuscript, `http://arxiv.org/abs/1103.2970`.

[43] M. J. Lai and L. Schumaker, *Spline Functions on Triangulations*, Cambridge University Press, (2007).

[44] F. X. Le Dimet, M. Ouberdous, *Retrieval of balanced fields: an optimal control method*, Tellus, (45A) (1993), 449-461.

[45] V. I. Olicker, L. D. Prussner, *On the numerical solution of the equation* $z_{xx}z_{yy} - z_{xy}^2 = f$ *and its discretization*, I. Numer. Math. 54(3), 271-293 (1988).

[46] A. Oberman, *Wide stencil finite difference schemes for the elliptic Monge-Ampère equations and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. Ser B 10(1), 221-238 (2008).

[47] A. Oberman, J. D. Benamou, B. D. Froese, *Two numerical methods for the elliptic Monge-Ampère equations*, MA2N Math. Model. Numer. Anal., DOI 10.1051/ma2an/2010017, 2010.

[48] R. Scott, *Finite Element Techniques for Curved Boundaries*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1973.

[49] R. Scott, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal. (12) (1975), 404–427.

[50] L. L. Schumaker, *Computing bivariate splines in scattered data fitting and the finite element method*, Numer. Algorithms, 48 (2008), 237–260.

[51] Ch. Schwab, *p- and hp-Finite Element Methods,* Clarendon Press, Oxford, 1998.