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

Oleg Davydov* Abid Saeed†

April 16, 2013

Abstract

We present an implementation of Böhmer's finite element method for fully nonlinear elliptic partial differential equations on convex polygonal domains, based on a modified Argyris element and Bernstein-Bézier techniques. Our numerical experiments for several test problems, involving the classical Monge-Ampère equation and an unconditionally elliptic equation, confirm the convergence and error bounds predicted by Böhmer's theoretical results.

1 Introduction

Numerical solution of fully nonlinear elliptic partial differential equations is a topic of intensive research and great practical interest, see [7]. The motivation behind this interest is the presence of these equations in different fields of science and engineering including differential geometry [2], fluid mechanics [18] and optimal transportation [5].

Several numerical methods have been recently proposed in the literature for the fully nonlinear equations, in particular finite difference [13, 20] and finite element type methods [6, 14, 15, 8, 3, 4, 19]. The most of these methods are restricted to the equations of Monge-Ampère type, such as the Monge-Ampère equation, the equation of Gaussian curvature and Pucci's equation.

Böhmer's method [6, 7] allows solving the Dirichlet problem for any fully nonlinear elliptic equations of second order. It is based on a finite element discretisation of the linearised elliptic equations, with C^1 finite element spaces

*Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK, oleg.davydov@strath.ac.uk

†Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK, abid.saeed@strath.ac.uk

that admit a stable splitting into the subspace satisfying zero boundary conditions and its complement. This method does not require any variational formulation of the fully nonlinear equation. Full theoretical justification of the method is given in [6, 7], including a proof of convergence and error bounds. However, no numerical results have been provided.

This paper presents the first implementation of Böhmer's method based on a modified Argyris finite element space with a stable splitting developed in [11, 12]. Our construction makes use of the Bernstein-Bézier techniques for piecewise polynomials [17, 21]. These techniques are widespread in geometric modelling, and currently gain more attention in the finite element method because of a number of desirable properties, in particular optimal complexity of the element system matrix assembly [1]. Note that C^1 conforming discretisations based on a variational formulation of either the Monge-Ampère equation or a fouth order quasilinear equation resulting from a singular perturbation of the Monge-Ampère equation have been recently explored in [3, 4, 15]. Neither of these discretisations is equivalent to Böhmer's method. In particular, the *spline element method* of [3, 4] uses Lagrange multipliers to enforce inter-element smoothness and boundary conditions by solving a saddle point problem, rather than relying on a stable splitting of a C^1 finite element space.

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 [6, 7].

The paper is organised as follows. Section 2 is devoted to the formulation of Böhmer's method. In Section 3 we recall our construction of the modified Argyris space [12] and provide the details of the numerical implementation of a stable local basis admitting a stable splitting. In Section 4 we discuss 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 5 is devoted to the numerical experiments.

2 Böhmer's Method for Elliptic Equations

2.1 Fully Nonlinear Elliptic Operators

Let Ω be a bounded domain in $\mathbb{R}^n$ and 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 operator G.

The operator G is said to be *elliptic* in a subset $\widetilde{\Gamma} \subset \widetilde{\Omega} \times \Gamma$ if the matrix $[\frac{\partial \widetilde{G}}{\partial r_{ij}}(w)]_{i,j=1}^n$ is well defined and positive definite for all $w \in \widetilde{\Gamma}$ [7, 16]. If $\widetilde{G}$ is a linear function of (z, p, r) for each fixed x, then G is a linear differential operator. Under suitable restrictions on $\widetilde{G}$, classes of quasilinear and semilinear differential operators are obtained [7, p. 80], but in general G may be fully nonlinear.

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

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

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. Note that $G'(\hat{u})$ is the Fréchet derivative of G if G is Fréchet differentiable at $\hat{u}$.

Many nonlinear elliptic operators and corresponding equations $G(u) = 0$ are important for applications. A standard example of a fully nonlinear equation is the *Monge-Ampère equation* on $\Omega \subset \mathbb{R}^2$, given by

$$G_{\mathrm{MA}}(u) := \det(\nabla^2 u) - f(x) = 0, \quad f(x) > 0 \text{ for } x \in \Omega.$$

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

$$G(u) = 0, \quad x \in \Omega, \tag{2}$$

$$u = \phi, \quad x \in \partial\Omega, \tag{3}$$

where ϕ is a continuous function defined on $\partial\Omega$. Under certain assumptions, including the exterior sphere condition for $\partial\Omega$ and sufficient smoothness of $\widetilde{G}$, this problem has a unique solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ [16, Theorem 17.17]. Note that the Monge-Ampère operator G_{MA} is elliptic in subsets $\widetilde{\Gamma}$ satisfying

$$\widetilde{\Gamma} \subset \widetilde{\Omega} \times \mathbb{R} \times \mathbb{R}^n \times \{r \in \mathbb{R}^{n \times n} : r \text{ is positive definite}\}.$$

Therefore there exists a unique *convex* solution of $G_{\mathrm{MA}}(u) = 0$, whereas it is known that the Monge-Ampère equation has another, concave solution [9, Chapter 4].

2.2 Spline Spaces and Stable Splitting

As usual in the finite element method, the discretisation of the Dirichlet problem is done with the help of spaces of piecewise polynomial functions (splines). Let $\triangle$ be a triangulation of a polyhedral domain $\Omega \subset \mathbb{R}^n$, that 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

$$S_d^r(\triangle) = \{s \in C^r(\Omega) : s|_T \in P_d \text{ for all simplices } T \text{ in } \triangle\}, \tag{4}$$

where $d > r \geq 0$ and P_d is the space of polynomials of total degree d in n variables. Recall that the *star* of a vertex v of $\triangle$, denoted by $\operatorname{star}(v) = \operatorname{star}^1(v)$, is the union of all simplices $T \in \triangle$ attached to v. We define $\operatorname{star}^j(v)$, $j \geq 2$, inductively as the union of the stars of all vertices of $\triangle$ contained in $\operatorname{star}^{j-1}(v)$, and $\operatorname{star}(T)$ as the union of the stars of all vertices of the simplex T.

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 space with basis $s_1, \ldots, s_N$ and dual linear functionals $\lambda_1, \ldots, \lambda_N : S^h \to \mathbb{R}$ 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 $\triangle^h$, (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 $s \in S^h$, see [10, 11] and [7, Section 4.2.6].

To handle the Dirichlet boundary condition (3), the following subspace of S^h is important:

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

We say that S^h admits a *stable splitting*

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

if there is a stable local basis $\{s_1, \ldots, s_N\}$ for S^h that 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 the subspaces 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 [11, 12] (see also [7, Section 4.2.6]) how the stable splitting can be achieved for a modified Argyris finite element space.

2.3 Böhmer's Method

Let $u = \hat{u}$ be the solution of (2)–(3). According to [6, 7], 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

$$(G(\hat{u}^h), v^h)_{L^2(\Omega)} = 0 \quad \forall v^h \in S_0^h, \quad \text{and} \tag{5}$$

$$(\hat{u}^h, v_b^h)_{L^2(\partial\Omega)} = (\phi, v_b^h)_{L^2(\partial\Omega)} \quad \forall v_b^h \in S_b^h, \tag{6}$$

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 (5)–(6) is equivalent to a system of algebraic equations with respect to the coefficients of $\hat{u}^h$ in a basis of S^h.

Theorem 1 ([7, Theorem 5.2]) *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 [7, Section 5.2.3]. Assume that G is continuously differentiable in the neighbourhood of an isolated solution $\hat{u}$ of (2)–(3), such that $\hat{u} \in D(G) \cap H^\ell(\Omega)$, $\ell > 2$, 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 (5)–(6) has a unique solution $\hat{u}^h \in S^h$ as soon as the maximum edge length h is sufficiently small. Moreover,*

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

Note that, in particular, all conditions of Theorem 1 are satisfied by the Monge-Ampère operators, where $D(G_{\mathrm{MA}}) = C^2(\Omega)$, see [7, Example 3.26].

The nonlinear problem (5)–(6) can be solved iteratively by a Newton method as suggested in [6], 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

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

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 [6, 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 $\hat{u}$.

3 C^1 Finite Elements with Stable Splitting

3.1 Bernstein-Bézier Techniques

This section is devoted to the key concepts of the Bernstein-Bézier techniques we rely upon in our implementation of the finite element spaces suitable for Böhmer's method. A comprehensive treatment of these techniques can be found in [17]. We restrict to the case of two variables.

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain and $\triangle$ a triangulation of Ω. For a given $d \geq 1$, let $D_{d,\triangle} := \bigcup_{T\in\triangle} D_{d,T}$ be the set of *domain points*, where

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

for each triangle $T := \langle v_1, v_2, v_3\rangle$ in $\triangle$. We will use the following terminology for certain subsets of $D_{d,T}$. We refer to the set

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

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.

Recall that every $v \in \mathbb{R}^2$ can be uniquely represented in the form

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

where the components of the triplet (b_1, b_2, b_3) are called the *barycentric coordinates* of v relative to the triangle $T := \langle v_1, v_2, v_3\rangle$. Barycentric coordinates are linear functions of v, and the functions

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

are the *Bernstein-Bézier basis polynomials* of degree d associated with triangle T. Every polynomial p of total degree d can be written uniquely as

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

where c_{ijk} are the *Bézier coefficients* of p. For each $s \in S^0_d(\triangle)$ and $\xi = \xi_{ijk} \in D_{d,\triangle}$ we denote by c_ξ the coefficient c_{ijk} of the restriction of s to any triangle

$T \in \triangle$ containing ξ. (Because of the continuity of s the coefficient c_ξ does not depend on the particular choice of such triangle.)

A key concept for dealing with spline spaces in Bernstein-Bézier form is that of a minimal determining set. The set $M \subset D_{d,\triangle}$ is a *determining set* for a linear space $S \subset S_d^0(\triangle)$ if

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

and M is a *minimal determining set* (MDS) for the space S if there is no smaller determining set. Then $\dim S$ equals the cardinality $\#\{M\}$ of M.

Usually subspaces $S \subset S_d^0(\triangle)$ are defined with the help of certain *smoothness conditions* which can be explicitly written down as linear equations involving the coefficients c_ξ, $\xi \in D_{d,\triangle}$. For a given minimal determining set M for S, if we assign values to the coefficients $\{c_\xi\}_{\xi \in M}$, then the remaining coefficients c_η, $\eta \in D_{d,\triangle} \backslash M$ of a spline $s \in S$ can be computed using the smoothness conditions. Hence, an MDS M can be used to construct the *M-basis* $\{s_\xi\}_{\xi \in M}$ for S by requiring that the Bézier coefficients c_η, $\eta \in M$, of s_ξ satisfy $c_\xi = 1$ and $c_\eta = 0$ for all $\eta \in M \setminus \{\xi\}$.

We now introduce the concept of a stable and local MDS, which applies to algorithms of constructing an MDS for any triangulation of a given family, for example for all triangulations in two variables with a given lower bound on the minimum angle of the triangles. Let

$$\Gamma_\eta := \{\xi \in M : \, c_\eta \text{ depends on } c_\xi\}, \qquad \eta \in D_{d,\triangle} \backslash M,$$

where we say that c_η depends on c_ξ, $\xi \in M$, if the value of c_η for a spline $s \in S$ is changed when we change the value of c_ξ. This simply means that the coefficient c_η of the basis spline s_ξ is not zero. A minimal determining set M for a space S is said to be *local* if there exists an absolute integer constant 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. Moreover, M 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

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

If M is a stable local MDS, then the corresponding M-basis of S is stable and local in the sense of Section 2.2. A stable splitting of this basis can often be achieved by an appropriate splitting of the MDS, which leads to the following definition.

Definition 2 *Assume that the space $S \subset S_d^0(\triangle)$ has a stable local MDS M and let*

$$S_0 := \{s \in S \ : \ s|_{\partial\Omega} = 0\}. \tag{7}$$

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*

$$S_0 = \{s \in S : c_\xi = 0 \ \forall \xi \in M_b\} \tag{8}$$

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

$$S_b := \{s \in S : c_\xi = 0 \ \forall \xi \in M_0\}. \tag{9}$$

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 (8) holds.

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 $\{s_\xi\}_{\xi\in M}$.

3.2 Modified Argyris Space

Recall that the *superspline* spaces $S_d^{r,\rho}(\triangle)$, $r \le \rho \le d$, of $S_d^r(\triangle)$ are defined as

$$S_d^{r,\rho}(\triangle) = \{\ s \in S_d^r(\triangle) : s \in C^\rho(v) \ \forall v \in V\}, \tag{10}$$

where V is the set of all vertices of $\triangle$.

The *Argyris finite element space* is obtained by choosing $d = 5$, $r = 1$ and $\rho = 2$ in (10). Now for each $v \in V$, let T_v be 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 $M_e := \{\xi_{122}^{T_e}\}$. Then from [17, Theorem 6.1] we have

Theorem 3 *The dimension of the Argyris finite element space is given by* $\dim S_5^{1,2}(\triangle) = 6\#\{V\} + \#\{E\}$, *and*

$$M = \bigcup_{v\in V} M_v \cup \bigcup_{e\in E} M_e \tag{11}$$

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

The *modified Argyris space* $\tilde{S}$ [11, 12] is given by

$$\tilde{S} := \left\{s \in S_5^1(\triangle) : s \in C^2(v), \text{ for all } \textit{interior} \text{ vertices } v \text{ of } \triangle\right\}. \tag{12}$$

We now introduce some further notation that will help us to describe a stable local MDS for $\tilde{S}$. Let V_I and V_B denote the sets of interior and boundary vertices of $\triangle$, respectively, and let E_I and E_B represent the sets of interior and boundary edges, such that $V = V_I \cup V_B$, $E = E_I \cup E_B$. Let, furthermore, $E_v = \{e_1, e_2, \cdots, e_n\}$ denote all edges of $\triangle$ emanating from a vertex $v \in V$, in counterclockwise order. For each e_i, let ξ_i be the (unique) domain point in $R_2(v) \cap e_i$, $i = 1, \ldots, n$. For each $v \in V$, we choose a triangle T_v as above, where we assume in addition that T_v shares an edge with the boundary of Ω if $v \in V_B$. We define M_v and M_e as above, and set

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

Theorem 4 ([12, Theorem 4]) *The dimension of the modified Argyris space $\tilde{S}$ is given by* $\dim \tilde{S} = 6\#V_I + \#E + \sum_{v \in V_B} (4 + \#E_v)$, *and*

$$\tilde{M} := \bigcup_{v \in V_I} M_v \cup \bigcup_{e \in E} M_e \cup \bigcup_{v \in V_B} \tilde{M}_v. \tag{13}$$

is a stable local MDS for $\tilde{S}$.

We now split M into two disjoint subsets $\tilde{M}_0$ and $\tilde{M}_b$ as follows. Let

$$\left(\bigcup_{v \in V_I} M_v \cup \bigcup_{e \in E} M_e \right) \subset \tilde{M}_0, \tag{14}$$

and let all points of $\tilde{M}$ lying on the boundary be in $\tilde{M}_b$. Also let

$$\{R_2(v) \cap \tilde{M}_v\} \backslash \{e_1, e_n\} \in \tilde{M}_0, \text{ for each } v \in V_B.$$

Now only one point in $R_1(v) \cap \tilde{M}_v$, for each $v \in V_B$, is not assigned to either $\tilde{M}_0$ or $\tilde{M}_b$. We denote this point by ξ_v. Where ξ_v belongs depends on the geometry of the boundary edges attached to v.

- If boundary edges attached to v are non-collinear, then $\xi_v \in \tilde{M}_b$.
- If boundary edges attached to v are collinear, then $\xi_v \in \tilde{M}_0$.

Now we are in position to formulate the theorem about stable splitting for the modified Argyris space.

Theorem 5 ([12]) *$\tilde{M} = \tilde{M}_0 \cup \tilde{M}_b$ is the stable splitting of the stable local MDS $\tilde{M}$ for $\tilde{S}$.*

Stable splitting of an MDS is impossible for the standard Argyris space in general, as the following result shows.

Theorem 6 ([12]) *No MDS for the Argyris space can be stably split on arbitrary triangulations.*

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.

4 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)-(3) for a fully nonlinear equation of second order.

Discretisation

Recall that $\triangle^h$ is a quasi-uniform triangulation of a convex polygonal domain $\Omega \subset \mathbb{R}^2$. As discussed in Section 2, solving the nonlinear problem (2)-(3) by Böhmer's method amounts to running a Newton-Kantorovich iteration scheme to get a sequence $\left\{u_k^h\right\}_{k\in\mathbb{Z}_+}$ of approximations of $\hat{u}$ generated by

$$u_{k+1}^h = u_k^h - w^h, \quad k = 0, 1, \ldots, \tag{17}$$

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

$$(G'(u_k^h)w^h, v^h)_{L^2(\Omega)} = (G(u_k^h), v^h)_{L^2(\Omega)} \; \forall v^h \in S_0^h, \tag{18}$$

where G' is the linearisation (1) of the nonlinear operator G. We solve this linear equation 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 5.

After a standard transformation to the weak form, (18) is translated into the following problem: Find $w^h \in \tilde{S}_0^h$ such that for all $v^h \in \tilde{S}_0^h$,

$$\int_\Omega \nabla w^h \cdot A \nabla v^h dx + \int_\Omega v^h b \cdot \nabla w^h dx + \int_\Omega c w^h v^h dx = \int_\Omega f v^h dx, \tag{19}$$

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

If $(s_1, \ldots, s_{N_0})$ is a basis of $\tilde{S}_0^h$, then, as usual in the finite element method, (19) results in the linear system

$$(\mathcal{S} + \mathcal{B}^t + \mathcal{M})a = \mathcal{L} \tag{20}$$

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

Require: Given ξ, initialize $\{c_\eta : \eta \in D_{5,\triangle}\}$ by zeros and set $c_\xi = 1$. Recall that T_e is triangle sharing the edge $e \in E$. Let $\tilde{T}_e$ be the other triangle sharing the edge e if $e \in E_I$.

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 [17, Lemma 2.30]. We write these smoothness conditions explicitly. Let $T_1 := \langle v, v_2, v_1 \rangle$ and $T_2 := \langle v_3, v, v_1 \rangle$ are two of the triangles attached to vertex v then we compute $c'_\nu, \nu \in D_2(v) \cap T_2$ from known coefficients $c_\eta, \eta \in D_2(v) \cap T_1$ as follows

$$
\begin{aligned}
c'_{131} &= b_1 c_{401} + b_2 c_{311} + b_3 c_{302}, && (15)\\
c'_{140} &= b_1 c_{500} + b_2 c_{410} + b_3 c_{401}, && (16)\\
c'_{230} &= b_1^2 c_{500} + 2 b_1 b_2 c_{410} + b_2^2 c_{320} + 2 b_2 b_3 c_{311} + b_3^2 c_{302} + 2 b_3 b_1 c_{401},
\end{aligned}
$$

   where (b_1, b_2, b_3) are barycentric coordinates of v_3 w.r.to T_1.
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 $\tilde{T}_e := \langle v_3, v, v_1 \rangle$. We compute $c_{122}^{\tilde{T}_e}$ using C^1 smoothness condition over e [17, Lemma 2.30]

$$c_{122}^{\tilde{T}_e} = b_1 c_{302},$$

   where (b_1, b_2, b_3) are barycentric coordinates of v_3 w.r.to T_e and c_{302} is known for $\xi_{302} \in D_2(v)$.
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 $c_\nu, \nu \in \{D_2(v) \cap T_{\kappa+1}\} \setminus \tilde{M}_v$ from known coefficients $c_\eta, \eta \in D_2(v) \cap T_\kappa, \; \kappa = 1, \cdots, k-1$, using the same C^1 smoothness conditions (15)-(16).
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**

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 $i, j = 1, \ldots, N_0$, 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. Recall that any spline $s \in \tilde{S}^h$ restricted to the triangle T_κ can be written in the form

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

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, where we use the lexicographic order as in [17, p. 23] to arrange these coefficients. That is, the triples of the indices (i, j, k) are arranged by the ordering function

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

Let $\mathbf{V}(s)$ be a row vector of all $\mathbf{C}_{T_\kappa}$'s, $\kappa = 1, \cdots, N_t$, for a spline s, ordered according to the triangles $\{T_\kappa\}_{\kappa=1}^{N_t}$,

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

Now, if we construct a matrix by taking these vectors $\mathbf{V}(s_i)$ for the basis splines $s_1, \ldots, s_{N_0}$ as its rows, then this matrix is our desired *transformation matrix* $\mathcal{T}$,

$$\mathcal{T} = [\mathbf{V}(s_1)^t, \ldots, \mathbf{V}(s_{N_0})^t]^t.$$

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 $s \in \tilde{S}^h_0$ 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^5_{ijk} on all triangles.

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

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

Then we can compute the system matrices in (20) by using the relations

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

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, 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 is reduced to the matrix equation

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

where $\mathcal{M}_b = \left[\int_{\partial\Omega} s_i s_j ds\right]_{i,j=N_0+1}^{N}$ 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 [17, Remark 2.4]. Moreover, there is an explicit formula for integration of the product of two 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.

Now consider

$$\int_e \phi s_i ds = \int_e \phi \sum_{\alpha=0}^{5} c_\alpha B_\alpha^5 ds = \sum_{\alpha=0}^{5} c_\alpha \int_e \phi B_\alpha^5 ds. \tag{21}$$

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 Gauss-Legendre 6-points 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 5. 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.

5 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 [19]. The numerics for these problems confirm the convergence and the theoretical error bounds of Theorem 1.

5.1 The Monge-Ampère equation

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

$$\begin{aligned} G_{MA}(u) = \det(\nabla^2 u) - g(x) = &\quad 0, \quad x \in \Omega \\ u = &\quad \phi,\ x \in \partial\Omega \end{aligned} \tag{22}$$

where g and ϕ are given functions with $g > 0$ on Ω required to keep the problem elliptic. The weak formulation (19) of the linearised problem in this case is to find $w^h \in S_0^h$ such that

$$\int_\Omega \nabla w^h \cdot A \nabla v^h dx = \int_\Omega f v^h dx, \text{ for all } v^h \in S_0^h, \tag{23}$$

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

$$\mathcal{S}\mathcal{C} = \mathcal{L},$$

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

As Monge-Ampère equation is elliptic only for convex functions, we need the initial guess to be convex as well. In [13, Remark 2.1] it has been shown that (22) and the Poisson-Dirichlet problem

$$\begin{aligned} \Delta u = &\quad 2\sqrt{g},\ x \in \Omega \\ u = &\quad \phi,\ x \in \partial\Omega \end{aligned} \tag{24}$$

are closely related. Therefore we use the approximation solution of the Poisson-Dirichlet problem (24) as an initial guess for the Newton scheme

(17). 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 initial mesh, whereas on the refined meshes we take a quasi-interpolant [17, Section 5.7] of the solution from 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 (22) over $\Omega = (0,1)^2$ considered in many paper 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 (22) 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 1. They confirm the convergence rate $O(h^4)$ in the H^2-norm predicted by Theorem 1, 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 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 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 (5). 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.

2. *Second test problem* is defined by

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

Table 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

in (22). The exact solution is $u(x) = -\sqrt{R^2 - |x|^2}$. The function $g(x)$ has singularity at $R = \sqrt{2}$ and $u \in W_p^1(\Omega)$, $1 \leq p < 4$ 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 2 and Table 3 for two different values of R, the results show the same behaviour as in the first problem. The tables confirm that the more the value of R is away from singularity the faster convergence is achieved. Note that in this experiments much higher accuracy is attained as compared to the results in [13] for the same test problem.

3. *Third test problem* is defined by

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

in the Monge-Ampère equation (22). 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 $u \in H^s(\Omega)$, for all $s < \frac{5}{2}$, we expect convergence order $O(h^{\frac{5}{2}})$ in L_2-norm. The results, in Table 4, confirm this.

Table 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 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 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

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

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

see Figure 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 (22) with the same data as in first test problem. Again we choose initial guess by the multilevel approach and use a solution of (24) on the first level. The numerics again show the same rate of convergence as for the rectangular domains, see Table 5. The graph of approximate solution

Table 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

u^h on the last level of triangulation is visualised in Figure 1 (right).

Table 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

5. *Fifth test problem.* Here we consider a homogeneous Dirichlet problem for (22) with $g = 1$ over $\Omega = [0, 1]^2$. This test problem is interesting because it does not have a smooth classical solution. Therefore, Theorem 1 does 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. The approximate solution u^h and its contour plot on a mesh with $h = \frac{1}{4}$ is visualized in Figure 2.

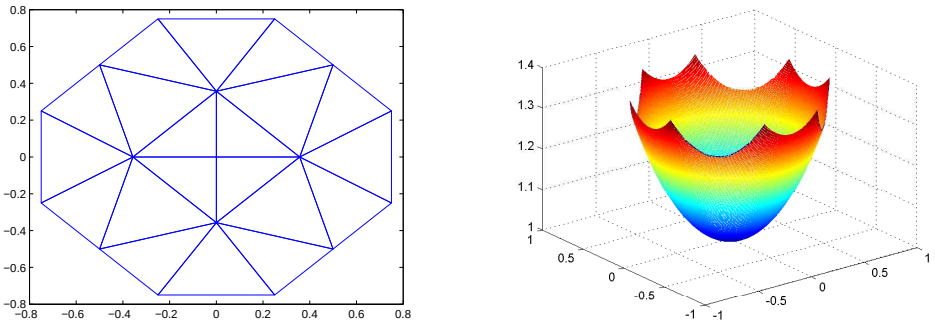

Figure 1: Non-rectangular domain Ω for fourth test problem with initial triangulation (*left*) and approximate solution u^h on the last level of triangulation (*right*).

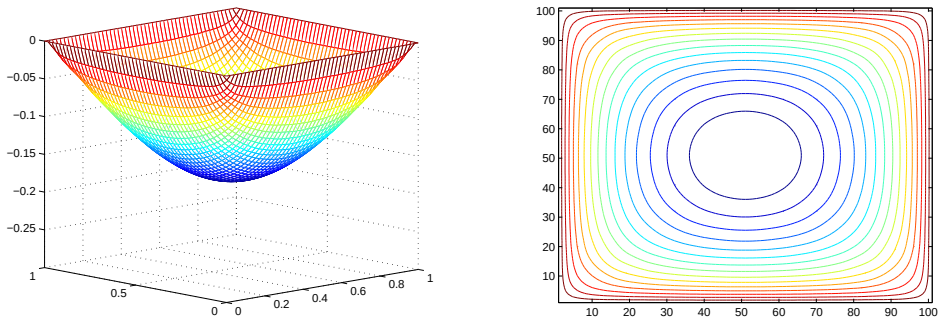

Figure 2: Approximate solution u^h of test 5 and its contour plot, $h = \frac{1}{4}$

5.2 Second Example

Consider the problem suggested in [19]

$$\begin{aligned} G_2(u) = u_{11}^3 + u_{22}^3 + u_{11} + u_{22} - g(x) = &\quad 0, \quad x \in \Omega \\ u = &\quad \phi,\ x \in \partial\Omega \end{aligned} \tag{25}$$

where $u_{ii} = \left(\partial^i\right)^2 u$, $i = 1, 2$. 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 [7] is satisfied in this example. The last of our test problems is for (25) 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 5.1.

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

$$\begin{aligned} \Delta u = &\quad 0, \quad x \in \Omega, \\ u = &\quad \phi,\ x \in \partial\Omega, \end{aligned} \tag{26}$$

whereas on the subsequent refinement levels we use the multilevel approach as described in Section 5.1. Note that the method was divergent with initial guess generated by (26) for $h \leq \frac{1}{2}$.

The results are presented in Table 6. They confirm the theoretical convergence rate of Böhmer's method. However, we see a very slow convergence of Newton's iterations in this example, compare N in Tables 1–6. 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 (5). This phenomenon requires further investigation.

References

[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.

Table 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

[2] T. Aubin, *Nonlinear Analysis on Manifolds, Monge-Ampère equation*, Springer-Verlag, Inc Berlin, 1982.

[3] G. Awanou, *Spline element method for the Monge-Ampere equation*, manuscript, 2010. `arXiv:1012.1775v1[math.NA]`

[4] G. Awanou, *Pseudo transient continuation and time marching methods for Monge-Ampère type equations*, manuscript, 2013. `arXiv:1301.5891v1[math.NA]`

[5] J.D. Benamou, Y. Brenier, *A computational fluid mechanics solution to Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), 375–393.

[6] K. Böhmer, *On finite element methods for fully nonlinear elliptic equations of second order*, SIAM J. Numer. Anal., 46(3) (2008), 1212–1249.

[7] K. Böhmer, *Numerical Methods for Nonlinear Elliptic Differential Equations*: A Synopsis, Oxford University Press, Oxford, 2010.

[8] S.C. Brenner, T. Gudi, M. Neilan, L.-Y. Sung, *C^0 penalty methods for the fully nonlinear Monge-Ampère equation*, Math. Comput., 80(276) (2011), 1979–1995.

[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. II, Wiley Interscience, 1989.

[10] O. Davydov, *Stable local bases for multivariate spline spaces*, J. Approx. Theory, 111 (2001), 267–297.

[11] O. Davydov, *Smooth finite elements and stable splitting*, Berichte "Reihe Mathematik" der Philipps-Universität Marburg, 2007-4 (2007). An adapted version has appeared as [7, Section 4.2.6].

[12] 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 - 7th International Conference, Avignon, France, June 24-30, 2010 LNCS 6920, Springer-Verlag, 2012, pp. 220–235.

[13] E.J. Dean, 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.

[14] 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) (2009) 1226–1250.

[15] 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.

[16] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, 2001.

[17] M. J. Lai and L. Schumaker, *Spline Functions on Triangulations*, Cambridge University Press, 2007.

[18] F.X. Le Dimet, M. Ouberdous, *Retrieval of balanced fields: an optimal control method*, Tellus, (45A) (1993), 449–461.

[19] O. Lakkis, T. Pryer, *A non-variational finite element method for the nonlinear elliptic problems*, manuscript, 2012. `arXiv:1103.2970v4[math.NA]`.

[20] 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) (2008), 221–238.

[21] L. L. Schumaker, *Computing bivariate splines in scattered data fitting and the finite element method*, Numer. Algorithms, 48 (2008), 237–260.