A Bivariate Spline Method for Second Order Elliptic Equations in Non-divergence Form

Ming-Jun Lai & Chunmei Wang

Journal of Scientific Computing

ISSN 0885-7474

J Sci Comput
DOI 10.1007/s10915-017-0562-0

Springer

J Sci Comput
DOI 10.1007/s10915-017-0562-0

A Bivariate Spline Method for Second Order Elliptic Equations in Non-divergence Form

Ming-Jun Lai[1] · **Chunmei Wang**[2]

Received: 5 February 2017 / Revised: 19 August 2017 / Accepted: 13 September 2017

Abstract A bivariate spline method is developed to numerically solve second order elliptic partial differential equations in non-divergence form. The existence, uniqueness, stability as well as approximation properties of the discretized solution will be established by using the well-known Ladyzhenskaya–Babuska–Brezzi condition. Bivariate splines, discontinuous splines with smoothness constraints are used to implement the method. Computational results based on splines of various degrees are presented to demonstrate the effectiveness and efficiency of our method.

Keywords Primal-dual · Discontinuous Galerkin · Finite element methods · Spline approximation · Cordes condition

Mathematics Subject Classification 65N30 · 65N12 · 35J15 · 35D35

1 Introduction

We are interested in developing an efficient numerical method for solving second order elliptic equations in non-divergence form. To this end, consider the model problem: Find $u = u(x)$ satisfying

The research of Ming-Jun Lai was partially supported by Simons collaboration Grant 280646 and the National Science Foundation Award DMS-1521537. The research of Chunmei Wang was partially supported by National Science Foundation Awards DMS-1522586 and DMS-1648171.

✉ Chunmei Wang
c_w280@txstate.edu

Ming-Jun Lai
mjlai@uga.edu

[1] Department of Mathematics, University of Georgia, Athens, GA 30602, USA

[2] Department of Mathematics, Texas State University, San Marcos, TX 78666, USA

Published online: 20 September 2017

$$\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 u + cu = f, \quad \text{in } \Omega, \tag{1.1}$$

$$u = 0, \quad \text{on } \partial\Omega, \tag{1.2}$$

where Ω is an open bounded domain in $\mathbb{R}^2$ with a Lipschitz continuous boundary $\partial\Omega$, ∂_{ij}^2 is the second order partial derivative operator with respect to x_i and x_j for $i, j = 1, 2$, and the function $f \in L^2(\Omega)$. Assume that the tensor $a(\mathbf{x}) = \{a_{ij}(\mathbf{x})\}_{2\times 2}$ is symmetric positive definite and uniformly bounded over Ω, the coefficient $c(\mathbf{x})$ is non-positive and uniformly bounded over Ω. In addition, we assume that the coefficients $a_{ij}(\mathbf{x})$ are essentially bounded so that the second order model problem (1.1) cannot be rewritten in a divergence form. Thus, the problem of designing stable and convergent numerical methods for (1.1) is subtle and currently an active area of research. See [23–26,29], and references therein.

For convenience, we shall assume that the model problem (1.1) has a unique strong solution $u \in H^2(\Omega)$ satisfying the H^2 regularity:

$$\|u\|_{H^2(\Omega)} \le C\|f\|_{L^2(\Omega)} \tag{1.3}$$

for a positive constant C. For example, when Ω is bounded with $C^{1,1}$ smoothness boundary, the Calderon–Zygmund theory (see e.g. [14, Theorem 9.15]) ensures that the solution to (1.1) has a unique solution and satisfies (1.3) if $a(\mathbf{x})$ is continuous over $\bar{\Omega}$ and $c \in L^\infty(\Omega)$. For another example, when $a(\mathbf{x})$ is only $L^\infty(\Omega)$, the Cordes condition can ensure the existence and uniqueness of strong solution if the domain Ω is convex with C^2 boundary (cf. Theorem 1.2.1 in [20]), where the coefficient tensor $a(\mathbf{x})$ is said to satisfy the Cordes condition if

$$\frac{\sum_{i,j=1}^{2} a_{ij}^2}{\left(\sum_{i=1}^{2} a_{ii}\right)^2} \le \frac{1}{n-1+\epsilon}, \quad \text{in} \quad \Omega \subset \mathbb{R}^n \tag{1.4}$$

for a positive number $\epsilon \in (0, 1]$. This Cordes condition is reasonable in $\mathbb{R}^2$ in the sense that when the coefficient tensor $a(\mathbf{x})$ satisfies the standard uniform ellipticity condition, i.e., there exist two positive numbers λ_1 and λ_2 such that

$$\lambda_1 \xi^\top \xi \le \xi^\top a(\mathbf{x})\xi \le \lambda_2 \xi^\top \xi, \quad \forall \xi \in \mathbb{R}^2, \mathbf{x} \in \Omega, \tag{1.5}$$

then the Cordes condition holds true in $\mathbb{R}^2$ (cf. [20]).

Furthermore, the assumption that the underlying domain Ω is convex is not necessary to ensure the H^2 regularity of the solution to the Dirichlet problem of Poisson equations. Based on the main result in [1], a bounded Lipschitz domain Ω satisfying an uniform outerball condition implies the H^2 regularity. Here, a domain satisfies an uniform outerball condition if there exists a positive number $r > 0$ such that every point $\mathbf{x}$ on the boundary $\partial\Omega$, there exists a ball of radius r touched at $\mathbf{x}$ which lies outside of Ω. Clearly, any convex domain satisfies an uniform outerball condition with the radius $r = \infty$. Also, any convex domain is a Lipschitz domain (cf. Corollary 9.1.2 in [2]). Thus, the result in [1] includes convex domains as a special case. The uniform outerball condition is also called semi-convex in [21]. Thus, when Ω is Lipschitz and semi-convex, there exists a strong solution $u \in H^2(\Omega)$ of (1.1) satisfying the H^2 regularity (1.3) if the PDE coefficient tensor $a(\mathbf{x})$ satisfies the Cordes condition.

Next as each function a_{ij} in the coefficient tensor $a(\mathbf{x})$ is in $L^\infty(\Omega)$, we assume in this paper that a_{ij} can be decomposed into finitely many pieces such that over each piece a_{ij} is a continuous function. Such an assumption is reasonable as often seen in practice. Under this

assumption, although using a polygonal partition may find the decomposition of a_{ij} more conveniently, we shall use a triangulation to decompose Ω in this paper to demonstrate the numerical performance. If a polygonal mesh is indeed used, the polygonal splines constructed in [12] should be used.

Recently, Smears and Süli [24] used the well-known Lax–Milgram theorem to establish the weak solution to (1.1) with $c \equiv 0$. They employed the Cordes condition to define a nonsymmetric bilinear form for their weak solution. By testing the Laplace of piecewise polynomials of degree k over a triangulation or polygonal partition, they compute their numerical solution. In fact, the solution is a strong solution due to the regularity (1.3). In [29], C. Wang and J. Wang used a primal-dual weak Galerkin finite element method to convert (1.1) with $c \equiv 0$ into a constrained minimization problem. The bilinear form associated with (1.1) was shown to satisfy the Ladyzhenskaya–Babuska–Brezzi condition by using the regularity assumption (1.3). Thus, the weak formulation associated with (1.1) is well-posed. The convergence and convergence rates of these two numerical methods were established in [24] and [29] together with numerical evidence of convergence over non-convex domains.

In this paper, we provide another efficient computational method for numerical solution of (1.1). More precisely, we propose a bivariate spline method based on the minimization of the jumps of functions across edges and the boundary condition to solve the constrained minimization similar to the one in [29]. Bivariate splines in $S_k^r(\triangle)$ of smoothness $r \geq 0$ and degree $k > r$ over triangulation $\triangle$ can be written in terms of $S_k^{-1}(\triangle)$, the space of discontinuous piecewise polynomial functions. Each polynomial over a triangle in $\triangle$ is written in Bernstein–Bézier polynomial form (cf. [18]). The smoothness constraints across an interior edge e of triangulation $\triangle$ are written in terms of the coefficients of polynomials over the two triangles sharing the common edge e. In particular, smoothness conditions of any order across interior edges have been implemented in MATLAB which can be simply used. This is an improvement over the internal penalties in the DG method in [24] and stabilizers in the weak Galerkin method in [29]. Bivariate splines have been used for numerical solutions of various types of PDE. See [4,5,15,16,19,22], and etc. They can be very convenient for numerical solutions of this type of PDE. See an extensive numerical evidence in §6.

Note that in [24], an hp-version discontinuous Galerkin finite element method was used. The method yielded an optimal order of convergence regarding to the mesh size h, i.e. $k-1$ for polynomial degree $k = 2, 3, 4, 5$. We use the C^1 spline function for the same PDE with discontinuous coefficients as in [24] and provide an evidence that the convergence rate of the root mean square error (RMSE) of $|u - S_u|_{H^2(\Omega)}$ using bivariate spline method is also $k-1$ for $k = 2, 3, 4, 5$ when $c \equiv 0$. In fact, our spline method produces more accurate results than that in [24] and [29]. One of the reasons is that our spline method more flexible in the sense that we can employ various spline spaces for primal and dual variables. That is, in the primal and dual formulation, when using $X_h = S_k^1(\triangle)$ for the primal variable, we can use $M_h = S_k^{-1}(\triangle)$ for the dual variable instead of $S_{k-2}^{-1}(\triangle)$ and $S_{k-1}^{-1}(\triangle)$ as in [29]. Such a choice can produce more accurate results although not for all the cases. (See Sect. 6 for detail.) In addition, we can use higher degree splines very easily by inputting a large degree in our MATLAB code. The flexibility of using bivariate splines of various degrees make our method more convenient to increase the accuracy of solutions. When $c \neq 0$, we have the similar convergence behavior. In particular, the convergence rate of $u - S_u$ in $H^2(\Omega)$ semi-norm is still $k-1$.

The paper is organized as follows: We first start with an explanation of the primal-dual discontinuous Galerkin method to solve (1.1) in the next section. Mainly, we establish some basic properties such as the existence, uniqueness, stability of the method in Sect. 3. Then

in Sect. 4 we present an error analysis of the numerical solution. Next we reformulate the primal-dual discontinuous Galerkin algorithm based on the bivariate spline functions which were implemented in [5]. Extensive numerical results are reported in Sect. 6. We start with a PDE with smooth coefficients and test on a smooth solution to demonstrate that the bivariate spline method works very well. Then we solve some PDE with discontinuous coefficients and nonsmooth solutions. For comparison purpose, we use the PDE in (1.1) with $c \equiv 0$ in [24] and [29]. Although these PDEs have discontinuous coefficients and non-smooth solutions, our spline method is able to approximate the solution very well. Therefore, the bivariate spline method is effective and efficient.

2 A Primal-Dual Discontinuous Galerkin Scheme

Our model problem seeks for a function $u \in H^2(\Omega)$ satisfying $u|_{\partial\Omega} = 0$ and

$$\left(\sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 u + cu, w \right) = (f, w), \quad \forall w \in L^2(\Omega), \tag{2.1}$$

where $(\cdot, \cdot)$ is the standard L^2 projection defined on the domain Ω.

Let $\mathcal{T}_h$ be a polygonal finite element partition of the domain $\Omega \subset \mathbb{R}^2$. Denote by $\mathcal{E}_h$ the set of all edges in $\mathcal{T}_h$ and $\mathcal{E}_h^0 = \mathcal{E}_h \setminus \partial\Omega$ the set of all interior edges. Assume that $\mathcal{T}_h$ satisfies the shape regularity conditions described in [7,30]. Denote by h_T the diameter of $T \in \mathcal{T}_h$ and $h = \max_{T\in\mathcal{T}_h} h_T$ the mesh size of the partition $\mathcal{T}_h$. Let $k \geq 0$ be an integer. Let $P_k(T)$ be the space of polynomials of degree no more than k on the element $T \in \mathcal{T}_h$.

For any given integer $k \geq 2$, we define the finite element spaces composed of piecewise polynomials of degree k and $k-2$, respectively; i.e.,

$$\begin{aligned} X_h &= \{u : u|_T \in P_k(T), \quad \forall T \in \mathcal{T}_h\}, \\ M_h &= \{u : u|_T \in P_{k-2}(T), \quad \forall T \in \mathcal{T}_h\}. \end{aligned}$$

Denote by $[\![v]\!]$ the jump of v on an edge $e \in \mathcal{E}_h$; i.e.,

$$[\![v]\!] = \begin{cases} v|_{T_1} - v|_{T_2}, & e = (\partial T_1 \cap \partial T_2) \subset \mathcal{E}_h^0, \\ v, & e \subset \partial\Omega, \end{cases} \tag{2.2}$$

where $v|_{T_i}$ denotes the value of v as seen from the element T_i, $i = 1, 2$. The order of T_1 and T_2 is non-essential in (2.2) as long as the difference is taken in a consistent way in all the formulas. Analogously, one may define the jump of the gradient of u on an edge $e \in \mathcal{E}_h$, denoted by $[\![\nabla u]\!]$.

For any $v \in X_h$, the quadratic functional $J(v)$ is given by

$$J(v) = \tfrac{1}{2} \sum_{e\in\mathcal{E}_h} h_T^{-3} \langle [\![v]\!], [\![v]\!] \rangle_e + \tfrac{1}{2} \sum_{e\in\mathcal{E}_h^0} h_T^{-1} \langle [\![\nabla v]\!], [\![\nabla v]\!] \rangle_e. \tag{2.3}$$

It is clear that $J(v) = 0$ if and only if $v \in C^1(\Omega) \cap X_h$ with the homogeneous Dirichlet boundary data $v = 0$ on $\partial\Omega$.

We introduce a bilinear form

$$b_h(v, q) = \sum_{T\in\mathcal{T}_h} \left(\sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 v + cv, q \right)_T, \quad \forall v \in X_h, \ \forall q \in M_h. \tag{2.4}$$

The numerical solution of the model problem (1.1) and (1.2) can be characterized a constrained minimization problem as follows: *Find* $u_h \in X_h$ such that

$$u_h = \text{argmin}_{v \in X_h,\ b_h(v,q)=(f,q),\ \forall q \in M_h} J(v). \tag{2.5}$$

By introducing the following bilinear form

$$s_h(u,v) = \sum_{e \in \mathcal{E}_h} h_T^{-3} \langle [\![u]\!], [\![v]\!] \rangle_e + \sum_{e \in \mathcal{E}_h^0} h_T^{-1} \langle [\![\nabla v]\!], [\![\nabla v]\!] \rangle_e, \quad \forall u, v \in X_h, \tag{2.6}$$

the constrained minimization problem (2.5) has an Euler–Lagrange formulation that gives rise to a system of linear equations by taking the Fréchet derivative. The Euler–Lagrange formulation for the constrained minimization algorithm (2.5) gives the following numerical scheme.

Algorithm 2.1 *(Primal-Dual Discontinuous Galerkin FEM)* A numerical approximation of the second order elliptic problem (1.1) and (1.2) seeks to find $(u_h; \lambda_h) \in X_h \times M_h$ satisfying

$$s_h(u_h, v) + b_h(v, \lambda_h) = 0, \quad \forall v \in X_h, \tag{2.7}$$

$$b_h(u_h, q) = (f, q), \quad \forall q \in M_h. \tag{2.8}$$

3 Existence, Uniqueness and Stability

In this section, we will derive the existence, uniqueness, and stability for the solution $(u_h; \lambda_h)$ of the primal-dual discontinuous Galerkin scheme (2.7) and (2.8).

For each element $T \in \mathcal{T}_h$, let B_T be the largest disk inside of T centered at c_0 with radius r and $F_{k,B_T}(f)$ be the averaged Taylor polynomial of degree k for $f \in L^1(T)$ (see page 4 of [18] for details). Note that the averaged Taylor polynomial $F_{k,B_T}(f)$ satisfies (cf. Lemma 1.5 in [18])

$$\partial_{ij}^2 F_{k,B_T}(f) = F_{k-2,B_T}(\partial_{ij}^2 f) \tag{3.1}$$

if $\partial_{ij}^2 f \in L^1(T)$. Let $P_{X_h}(f)$ and $P_{M_h}(f)$ be interpolations/projections of f onto the spaces X_h and M_h defined by $P_{X_h}(f)|_T = F_{k,B_T}(f)$ and $P_{M_h}(f)|_T = F_{k-2,B_T}(f)$ on each element $T \in \mathcal{T}_h$, respectively. Using (3.1) gives rise to

$$\partial_{ij}^2 P_{X_h}(f) = P_{M_h}(\partial_{ij}^2 f), \tag{3.2}$$

on each element $T \in \mathcal{T}_h$,

Lemma 3.1 [18] *The interpolant operators P_{X_h} and P_{M_h} are bounded in $L^2(\Omega)$. In other words, for any $f \in L^2(\Omega)$ we have*

$$\|P_{X_h}(f)\| \le C\|f\|, \tag{3.3}$$

$$\|P_{M_h}(f)\| \le C\|f\|, \tag{3.4}$$

where $\|\cdot\|$ denotes the L^2 norm defined on the domain Ω, C is a constant depending only on the shape parameter $\theta_{\mathcal{T}_h} = \max_{T \in \mathcal{T}_h} \frac{h_T}{\rho_T}$, ρ_T is the radius of the largest inscribed circle of T.

Recall that $\mathcal{T}_h$ is a shape-regular finite element partition of the domain Ω. For any $T \in \mathcal{T}_h$ and $\phi \in H^1(T)$, the following trace inequality holds true:

$$\|\phi\|_{\partial T}^2 \le C\left(h_T^{-1}\|\phi\|_T^2 + h_T\|\nabla\phi\|_T^2\right). \tag{3.5}$$

Deonte by Q_{k-2} the L^2 projection onto the finite element space M_h. We introduce a semi-norm in the finite element space X_h, denoted by $|||\cdot|||$; i.e.,

$$|||v||| = \left(\sum_{T\in\mathcal{T}_h} \| Q_{k-2} \left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right) \|_T^2 + s_h(v,v) \right)^{\frac{1}{2}}, \quad v \in X_h. \tag{3.6}$$

The following result shows that $|||\cdot|||$ defined in (3.6) is indeed a norm on X_h when the meshsize h is sufficiently small.

Lemma 3.2 *Assume that the H^2 regularity (1.3) holds true for the model problem (1.1) and (1.2), and that the coefficient tensor $a(x) = \{a_{ij}(x)\}_{2\times 2}$ and $c(x)$ are uniformly piecewise continuous in Ω with respect to the finite element partition $\mathcal{T}_h$. Then, there exists an $h_0 > 0$ such that $|||\cdot|||$ in (3.6) defines a norm on X_h when the meshsize h is sufficiently small such that $h \le h_0$.*

Proof It suffices to verify the positivity property for $|||\cdot|||$. To this end, note that for any $v \in X_h$ satisfying $|||v||| = 0$ we have $s_h(v,v) = 0$. It follows that $[\![v]\!] = 0$ on each edge $e \in \mathcal{E}_h$ and $[\![\nabla u]\!] = 0$ on each interior edge $e \in \mathcal{E}_h^0$. Hence, $v \in C^1(\Omega)$ and $v = 0$ on $\partial\Omega$. In addition, on each element $T \in \mathcal{T}_h$, we have

$$Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right) = 0.$$

Thus,

$$\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv = (I - Q_{k-2}) \left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right) := F.$$

Using the H^2-regularity assumption (1.3), there exists a constant C such that

$$\|v\|_2 \le C\|F\|. \tag{3.7}$$

Note that $a_{ij}(x)$ and $c(x)$ are uniformly piecewise continuous in Ω with respect to the finite element partition $\mathcal{T}_h$. Let $\bar{a}_{ij}$ and $\bar{c}$ be the average of a_{ij} and c on each element $T \in \mathcal{T}_h$. Then, for any $\varepsilon > 0$, there exists a $h_0 > 0$ such that

$$\|a_{ij} - \bar{a}_{ij}\|_{L^\infty(\Omega)} \le \varepsilon, \quad \|c - \bar{c}\|_{L^\infty(\Omega)} \le \varepsilon,$$

if the meshsize h is sufficiently small such that $h \le h_0$. Denote by $\bar{c}$ and $\bar{v}$ the average of c and v on each element $T \in \mathcal{T}_h$, respectively. It follows from the linearity of the projection Q_{k-2} that

$$\begin{aligned}
\|F\| &\le \sum_{i,j=1}^{2} |a_{ij} - \bar{a}_{ij}| \|\partial_{ij}^2 v\| + \sum_{i,j=1}^{2} \left\| Q_{k-2}((a_{ij} - \bar{a}_{ij})\partial_{ij}^2 v) \right\| \\
&\quad + \|(I - Q_{k-2})(cv - \bar{c}\bar{v})\| \\
&\le C\varepsilon\|v\|_2 + \|cv - \bar{c}\bar{v}\| \le C\varepsilon\|v\|_2 + \|(c-\bar{c})v + \bar{c}(v - \bar{v})\| \\
&\le C\varepsilon\|v\|_2 + C\varepsilon\|v\| + Ch\|v\|_1 \le C\varepsilon\|v\|_2 + Ch\|v\|_2,
\end{aligned}$$

where $\|\cdot\|_2$ is the H^2 norm defined on the domain Ω, and we have used the boundedness of the L^2 projection Q_{k-2}, which, combined with (3.7), gives

$$\|v\|_2 \le C(\varepsilon + h)\|v\|_2.$$

This yields that $v = 0$ as long as ε is sufficiently small such that $C\varepsilon < 1$, which can be easily achieved by adjusting the parameter h_0. This completes the proof of the lemma. □

We are now in a position to establish an *inf-sup* condition for the bilinear form $b_h(\cdot, \cdot)$.

Lemma 3.3 (inf-sup condition) *Under the assumptions of Lemma 3.2, for any $q \in M_h$, there exists a $v_q \in X_h$ such that*

$$b_h(v_q, q) \geq \beta \|q\|^2, \tag{3.8}$$

$$|||v_q||| \leq C\|q\|, \tag{3.9}$$

provided that the meshsize h is sufficiently small.

Proof Consider an auxiliary problem that seeks $w \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying

$$\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 w + cw = q, \quad \text{in } \Omega. \tag{3.10}$$

From the regularity assumption (1.3), it is easy to know that the problem (3.10) has one and only one solution, and furthermore, the solution satisfies the H^2 regularity property; i.e.,

$$\|w\|_2 \leq C\|q\|. \tag{3.11}$$

By letting $v_q = P_{X_h}(w)$, from (3.2) we obtain

$$\partial_{ij}^2 v_q = \partial_{ij}^2 P_{X_h}(w) = P_{M_h}(\partial_{ij}^2 w).$$

Letting $\bar{a}_{ij}$ be the average of a_{ij} over $T \in \mathcal{T}_h$, we arrive at

$$\begin{aligned}
&\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v_q + cv_q \\
&= \sum_{i,j=1}^{2} \left\{(a_{ij} - \bar{a}_{ij})P_{M_h}(\partial_{ij}^2 w) + P_{M_h}(\bar{a}_{ij}\partial_{ij}^2 w)\right\} + (c - \bar{c})P_{X_h}(w) + P_{X_h}(\bar{c}w) \\
&= \sum_{i,j=1}^{2} \left\{(a_{ij} - \bar{a}_{ij})P_{M_h}(\partial_{ij}^2 w) + P_{M_h}((\bar{a}_{ij} - a_{ij})\partial_{ij}^2 w) + P_{M_h}(a_{ij}\partial_{ij}^2 w)\right\} \\
&\quad + (c - \bar{c})P_{X_h}(w) + P_{X_h}((\bar{c} - c)w) + P_{X_h}(cw) \\
&= \sum_{i,j=1}^{2} \left\{(a_{ij} - \bar{a}_{ij})P_{M_h}(\partial_{ij}^2 w) + P_{M_h}((\bar{a}_{ij} - a_{ij})\partial_{ij}^2 w)\right\} + (c - \bar{c})P_{X_h}(w) \\
&\quad + P_{X_h}((\bar{c} - c)w) + P_{M_h}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 w + cw\right) + P_{X_h}(cw) - P_{M_h}(cw) \\
&= E_T + q + P_{X_h}(cw) - P_{M_h}(cw).
\end{aligned}$$

where we have used (3.10) and $P_{M_h} q = q$. Here, $E_T = \sum_{i,j=1}^{2}\{(a_{ij} - \bar{a}_{ij})P_{M_h}(\partial_{ij}^2 w) + P_{M_h}((\bar{a}_{ij} - a_{ij})\partial_{ij}^2 w)\} + (c - \bar{c})P_{X_h}(w) + P_{X_h}((\bar{c} - c)w)$.

With the above chosen v_q as $P_{X_h}(w)$, we have

$$\begin{aligned} b_h(v_q, q) &= \sum_{T\in\mathcal{T}_h} \left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 P_{X_h}(w) + cP_{X_h}(w), q \right)_T \\ &= \sum_{T\in\mathcal{T}_h} (E_T, q)_T + \|q\|^2 + \sum_{T\in\mathcal{T}_h} (P_{X_h}(cw) - P_{M_h}(cw), q)_T. \end{aligned} \tag{3.12}$$

Note that the coefficient tensor $a(x) = \{a_{ij}\}_{2\times 2}$ and $c(x)$ are uniformly piecewise continuous over $\mathcal{T}_h$. Thus, for any given sufficiently small $\varepsilon > 0$, we have $\|a_{ij} - \bar{a}_{ij}\|_{L^\infty(\Omega)} \le \varepsilon$ and $\|c-\bar{c}\|_{L^\infty(\Omega)} \le \varepsilon$ for sufficiently small meshsize h. It then follows from the Cauchy–Schwarz inequality, (3.3) and (3.4), and the H^2 regularity property (3.11) that

$$\begin{aligned} &\left| \sum_{T\in\mathcal{T}_h} (E_T, q)_T \right| \\ &\le C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \sum_{i,j=1}^{2} \|P_{M_h}(\partial_{ij}^2 w)\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} + C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \sum_{i,j=1}^{2} \|\partial_{ij}^2 w\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} \\ &\quad + C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \|P_{X_h} w\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} + C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \|w\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} \\ &\le C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \sum_{i,j=1}^{2} \|\partial_{ij}^2 w\|_T^2 \right)^{\frac{1}{2}} \|q\| + C\varepsilon \left(\sum_{T\in\mathcal{T}_h} \|w\|_T^2 \right)^{\frac{1}{2}} \|q\| \\ &\le C\varepsilon \|w\|_2 \|q\| \le C\varepsilon \|q\|^2, \end{aligned}$$

and

$$\begin{aligned} &\left| \sum_{T\in\mathcal{T}_h} (P_{X_h}(cw) - P_{M_h}(cw), q)_T \right| \\ &\le \left(\sum_{T\in\mathcal{T}_h} \|P_{X_h}(cw - \bar{c}\bar{w}) - P_{M_h}(cw - \bar{c}\bar{w})\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} \\ &\le \left(\sum_{T\in\mathcal{T}_h} \|cw - \bar{c}\bar{w}\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} \\ &\le \left(\sum_{T\in\mathcal{T}_h} \|(c-\bar{c})w + \bar{c}(w-\bar{w})\|_T^2 \right)^{\frac{1}{2}} \|q\| \\ &\le (C\varepsilon\|w\| + Ch\|w\|_1)\|q\| \le C(\varepsilon + h)\|w\|_2\|q\| \le C(\varepsilon + h)\|q\|^2, \end{aligned}$$

where $\bar{c}$ and $\bar{w}$ are the average of c and w on each element $T \in \mathcal{T}_h$, respectively, C is a generic constant independent of $\mathcal{T}_h$. Substituting the above estimate into (3.12) yields

$$b_h(v_q, q) \ge (1 - C(2\varepsilon + h))\|q\|^2,$$

which leads to the estimate (3.8) when the meshsize h is sufficiently small.

It remains to derive the estimate (3.9). To this end, recall that

$$|||v_q|||^2 = \sum_{T\in\mathcal{T}_h} \left\| Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v_q + cv_q\right)\right\|_T^2 + s_h(v_q, v_q). \tag{3.13}$$

Letting $v_q = P_{X_h}(w)$, the first term on the right-hand side of (3.13) can be bounded by using (3.2), (3.3) and (3.11) as follows:

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_h} \left\| Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v_q + cv_q\right)\right\|_T^2 \\
&= \sum_{T\in\mathcal{T}_h} \left\| Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 P_{X_h}(w) + cP_{X_h}(w)\right)\right\|_T^2 \\
&\le \sum_{T\in\mathcal{T}_h}\sum_{i,j=1}^{2} \|a_{ij}P_{M_h}(\partial_{ij}^2 w)\|_T + \|cP_{X_h}(w)\|_T^2 \\
&\le C\sum_{i,j=1}^{2} \|a_{ij}\|_{L^\infty(\Omega)}^2 \sum_{T\in\mathcal{T}_h} \|\partial_{ij}^2 w\|_T^2 + C\|c\|_{L^\infty(\Omega)}^2 \sum_{T\in\mathcal{T}_h} \|w\|_T^2 \\
&\le C\|q\|^2.
\end{aligned} \tag{3.14}$$

As to the term $s_h(v_q, v_q)$ in (3.13), note that it is defined by (2.6) using the jump of v_q on each edge $e \in \mathcal{E}_h$ plus the jump of ∇v_q on each interior edge $e \in \mathcal{E}_h^0$. For an interior edge $e \in \mathcal{E}_h^0$ shared by two elements T_1 and T_2, we have

$$\begin{aligned}
[\![v_q]\!]|_e &= v_q|_{T_1\cap e} - v_q|_{T_2\cap e} \\
&= P_{X_h}(w)|_{T_1\cap e} - P_{X_h}(w)|_{T_2\cap e} \\
&= (P_{X_h}(w)|_{T_1\cap e} - w|_e) + (w|_e - P_{X_h}(w)|_{T_2\cap e}).
\end{aligned}$$

It follows that

$$\langle [\![v_q]\!], [\![v_q]\!]\rangle_e \le 2\|P_{X_h}(w)|_{T_1\cap e} - w|_e\|_e^2 + 2\|P_{X_h}(w)|_{T_2\cap e} - w|_e\|_e^2. \tag{3.15}$$

Using the trace inequality (3.5), we have

$$\|P_{X_h}(w)|_{T_1\cap e} - w|_e\|_e^2 \le Ch_T^{-1}\|P_{X_h}(w) - w\|_{T_1}^2 + Ch_T\|\nabla(P_{X_h}(w) - w)\|_{T_1}^2.$$

Analogously, the following holds true

$$\|P_{X_h}(w)|_{T_2\cap e} - w|_e\|_e^2 \le Ch_T^{-1}\|P_{X_h}(w) - w\|_{T_2}^2 + Ch_T\|\nabla(P_{X_h}(w) - w)\|_{T_2}^2.$$

Substituting the last two inequalities into (3.15) yields

$$\langle [\![v_q]\!], [\![v_q]\!]\rangle_e \le C\sum_{i=1}^{2}\left(h_T^{-1}\|P_{X_h}(w) - w\|_{T_i}^2 + Ch_T\|\nabla(P_{X_h}(w) - w)\|_{T_i}^2\right). \tag{3.16}$$

For boundary edge $e \subset \partial\Omega$, from $w|_{e\subset\partial\Omega} = 0$ we have

$$[\![v_q]\!]|_e = v_q|_e = P_{X_h}(w)|_e - w|_e.$$

Thus, the estimate (3.16) remains to hold true. Summing (3.16) over all the edges yields

$$\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle [\![v_q]\!], [\![v_q]\!]\rangle_e \le C\sum_{T\in\mathcal{T}_h}\left(h_T^{-4}\|P_{X_h}(w)-w\|_T^2 + Ch_T^{-2}\|\nabla(P_{X_h}(w)-w)\|_T^2\right)$$
$$\le C\|w\|_2^2 \tag{3.17}$$

where we have used the estimate (4.3) with $m = 1$ and $s = 0, 1$ in the last inequality. Combining (3.17) with the regularity estimate (3.11) gives rise to

$$\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle [\![v_q]\!], [\![v_q]\!]\rangle_e \le C\|q\|^2. \tag{3.18}$$

A similar argument can be applied to yield the following estimate

$$\sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle [\![\nabla v_q]\!], [\![\nabla v_q]\!]\rangle_e \le C\|q\|^2. \tag{3.19}$$

We emphasize that the summation in (3.19) is taken over all the interior edges so that no boundary value for ∇w is needed in the derivation of the estimate (3.19). Combining (3.18) and (3.19) with $s_h(v_q, v_q)$ yields

$$s_h(v_q, v_q) \le C\|q\|^2,$$

which, together with (3.14), completes the derivation of the estimate (3.9). □

Lemma 3.4 (Boundedness) *The following inequalities hold true:*

$$|s_h(u, v)| \le |\!|\!|u|\!|\!|\,|\!|\!|v|\!|\!|, \quad \forall u, v \in X_h,$$
$$|b_h(v, q)| \le C|\!|\!|v|\!|\!|\,\|q\|, \quad \forall v \in X_h, q \in M_h.$$

Proof It follows from the definition of $s_h(\cdot,\cdot)$, $|\!|\!|\cdot|\!|\!|$ and Cauchy–Schwarz inequality that for any $u, v \in X_h$, we have

$$|s_h(u, v)| = \left|\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle [\![u]\!], [\![v]\!]\rangle_e + \sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle [\![\nabla u]\!], [\![\nabla v]\!]\rangle_e\right|$$
$$\le \left(\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle [\![u]\!], [\![u]\!]\rangle_e\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle [\![v]\!], [\![v]\!]\rangle_e\right)^{\frac{1}{2}}$$
$$+\left(\sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle [\![\nabla u]\!], [\![\nabla u]\!]\rangle_e\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle [\![\nabla v]\!], [\![\nabla v]\!]\rangle_e\right)^{\frac{1}{2}}$$
$$\le s_h(u, u)^{\frac{1}{2}} s_h(v, v)^{\frac{1}{2}} \le |\!|\!|u|\!|\!|\,|\!|\!|v|\!|\!|.$$

Next from the definition of $b_h(\cdot,\cdot)$, $|\!|\!|\cdot|\!|\!|$, and Cauchy–Schwarz inequality that for any $v \in X_h$, $q \in M_h$, we have

$$
\begin{aligned}
|b_h(v,q)| &= \left| \sum_{T\in\mathcal{T}_h} \left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv, q \right)_T \right| \\
&= \left| \sum_{T\in\mathcal{T}_h} (Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right), q)_T \right| \\
&\le \left(\sum_{T\in\mathcal{T}_h} \left\| Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right) \right\|_T^2 \right)^{\frac{1}{2}} \left(\sum_{T\in\mathcal{T}_h} \|q\|_T^2 \right)^{\frac{1}{2}} \le |||v|||\|q\|.
\end{aligned}
$$

These complete the proof. □

Define the subspace of X_h as follows:

$$\Xi_h = \{v \in X_h : \quad b_h(v,q) = 0, \quad \forall q \in M_h\}.$$

Lemma 3.5 (Coercivity on the Kernel) *There exists a constant α, such that*

$$s_h(v,v) \ge \alpha |||v|||^2, \quad \forall v \in \Xi_h.$$

Proof For any $v \in \Xi_h$, we have

$$b_h(v,q) = 0, \quad \forall q \in M_h.$$

It follows from the definition of $b(\cdot,\cdot)$ in (2.4) that

$$0 = b_h(v,q) = \sum_{T\in\mathcal{T}_h} \left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv, q \right)_T = \sum_{T\in\mathcal{T}_h} \left(Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right), q \right)_T,$$

which yields

$$Q_{k-2}\left(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv \right) = 0,$$

on each $T \in \mathcal{T}_h$ by letting $q = Q_{k-2}(\sum_{i,j=1}^{2} a_{ij}\partial_{ij}^2 v + cv)$. This implies $s_h(v,v) = |||v|||^2$, which completes the proof with $\alpha = 1$. □

Using the abstract theory for the saddle-point problem developed by Babuska [6] and Brezzi [8], we arrive at the following theorem based on Lemmas 3.3–3.5.

Theorem 3.6 *The primal-dual discontinuous Galerkin finite element method* (2.7)–(2.8) *has a unique solution $(u_h;\lambda_h) \in X_h \times M_h$, provided that the meshsize $h < h_0$ holds true for a sufficiently small but fixed parameter $h_0 > 0$. Moreover, there exists a constant C such that the solution $(u_h;\lambda_h)$ satisfies*

$$|||u_h||| + \|\lambda_h\| \le C\|f\|. \tag{3.20}$$

4 Error Estimates

Let $(u_h;\lambda_h) \in X_h \times M_h$ be the approximate solution of the model problem (1.1) arsing from primal-dual discontinuous Galerkin finite element method (2.7) and (2.8). Note that $\lambda = 0$

is the exact solution of the trival dual problem $b_h(v, \lambda) = 0$ for all $v \in H^2(\Omega)$. Define the errors functions by

$$e_h = u_h - P_{X_h} u, \quad \epsilon_h = \lambda_h - P_{M_h} \lambda.$$

Lemma 4.1 *The error functions e_h and ϵ_h satisfy the following equations:*

$$s_h(e_h, v) + b_h(v, \epsilon_h) = -s_h(P_{X_h} u, v), \quad \forall v \in X_h, \tag{4.1}$$

$$b_h(e_h, p) = l_u(p), \quad \forall p \in M_h, \tag{4.2}$$

where $l_u(p) = \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} (a_{ij}(I - P_{M_h})\partial_{ij}^2 u, p)_T + \sum_{T \in \mathcal{T}_h} (c(I - P_{X_h})u, p)_T$.

Proof By subtracting $s_h(P_{X_h} u, v)$ from both sides of (2.7), we obtain

$$s_h(u_h - P_{X_h} u, v) + b_h(v, \lambda_h - 0) = -s_h(P_{X_h} u, v), \quad \forall v \in X_h,$$

which completes the proof of (4.1).

Substracting $b_h(P_{X_h} u, p)$ from both sides of (2.8), it follows from (3.2) and (1.1) that

$$\begin{aligned}
&b_h(u_h, p) - b_h(P_{X_h} u, p) \\
&= (f, p) - b_h(P_{X_h} u, p) \\
&= (f, p) - \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} \left(a_{ij} \partial_{ij}^2 (P_{X_h} u) + c P_{X_h} u, p \right)_T \\
&= (f, p) - \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} \left(a_{ij} P_{M_h} (\partial_{ij}^2 u) + c P_{X_h} u, p \right)_T \\
&= (f, p) - \sum_{T \in \mathcal{T}_h} \left(\sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 u + cu, p \right)_T - \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} \left(a_{ij} (P_{M_h} - I) \partial_{ij}^2 u, p \right)_T \\
&\quad - \sum_{T \in \mathcal{T}_h} (c(P_{X_h} - I)u, p)_T \\
&= (f, p) - (f, p) - \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} \left(a_{ij} (P_{M_h} - I) \partial_{ij}^2 u, p \right)_T - \sum_{T \in \mathcal{T}_h} (c(P_{X_h} - I)u, p)_T \\
&= \sum_{T \in \mathcal{T}_h} \sum_{i,j=1}^{2} \left(a_{ij} (I - P_{M_h}) \partial_{ij}^2 u, p \right)_T + \sum_{T \in \mathcal{T}_h} (c(I - P_{X_h})u, p)_T,
\end{aligned}$$

which completes the proof of (4.2). □

The Eqs. (4.1) and (4.2) are called error equations for the primal-dual discontinuous Galerkin finite element scheme. This is a saddle point system for which Brezzi's Theorem can be employed for the analysis of stability.

Lemma 4.2 [7,30] *Let $\mathcal{T}_h$ be a finite element partition of Ω satisfying the shape regular assumption given in* [7,30]. *Then, for any $0 \le s \le 2$ and $1 \le m \le k$, one has*

$$\sum_{T \in \mathcal{T}_h} h_T^{2s} \|u - P_{X_h} u\|_{s,T}^2 \le C h^{2(m+1)} \|u\|_{m+1}^2, \tag{4.3}$$

$$\sum_{T \in \mathcal{T}_h} h_T^{2s} \|u - P_{M_h} u\|_{s,T}^2 \le C h^{2(m-1)} \|u\|_{m-1}^2. \tag{4.4}$$

Theorem 4.3 *Assume that the coefficient tensor $a(x) = \{a_{ij}(x)\}_{2\times 2}$ and $c(x)$ are uniformly piecewise continuous in Ω with respect to the finite element partition $\mathcal{T}_h$. Let u and $(u_h; \lambda_h) \in X_h \times M_h$ be the solutions of (1.1) and (2.7) and (2.8), respectively. Assume that the exact solution u of (1.1) is sufficiently regular such that $u \in H^{k+1}(\Omega)$. There exists a constant C such that*

$$|||u_h - P_{X_h}u||| + \|\lambda_h - P_{M_h}\lambda\| \le Ch^{k-1}\|u\|_{k+1},$$

provided that the meshsize $h < h_0$ holds true for a sufficiently small, but fixed $h_0 > 0$.

Proof It follows from Lemmas 3.3–3.5 that the Brezzi's stability conditions are satisfied for the saddle point problem (4.1) and (4.2). Thus, there exists a constant C such that

$$|||e_h||| + \|\epsilon_h\| \le C\left(\sup_{v\in X_h, v\neq 0}\frac{|-s_h(P_{X_h}u, v)|}{|||v|||} + \sup_{p\in M_h, p\neq 0}\frac{|l_u(p)|}{\|p\|}\right). \tag{4.5}$$

Recall that

$$\begin{aligned}&\sup_{v\in X_h, v\neq 0}\frac{|-s_h(P_{X_h}u, v)|}{|||v|||}\\ &\le \sup_{v\in X_h, v\neq 0}\frac{|\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle[\![P_{X_h}u]\!], [\![v]\!]\rangle_e| + |\sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle[\![\nabla P_{X_h}u]\!], [\![\nabla v]\!]\rangle_e|}{|||v|||}\end{aligned} \tag{4.6}$$

As to the first term of the right-hand side of (4.6), from Cauchy–Schwarz inequality, trace inequality (3.5) and (4.3), we have

$$\begin{aligned}\left|\sum_{e\in\mathcal{E}_h} h_T^{-3}\langle[\![P_{X_h}u]\!], [\![v]\!]\rangle_e\right| &\le C\left(\sum_{e\in\mathcal{E}_h} h_T^{-3}\|[\![P_{X_h}u]\!]\|_e^2\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h} h_T^{-3}\|[\![v]\!]\|_e^2\right)^{\frac{1}{2}}\\ &\le C\left(\sum_{e\in\mathcal{E}_h} h_T^{-3}(\|[\![P_{X_h}u]\!] - [\![u]\!]\|_e^2 + \|[\![u]\!]\|_e^2)\right)^{\frac{1}{2}}|||v|||\\ &\le C\left(\sum_{T\in\mathcal{T}_h} h_T^{-4}\|[\![P_{X_h}u - u]\!]\|_T^2 + h_T^{-2}\|[\![P_{X_h}u - u]\!]\|_{1,T}^2\right)^{\frac{1}{2}}|||v|||\\ &\le Ch^{k-1}\|u\|_{k+1}|||v|||,\end{aligned} \tag{4.7}$$

where we used $[\![u]\!] = 0$ as $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Similarly, we have

$$\left|\sum_{e\in\mathcal{E}_h^0} h_T^{-1}\langle[\![\nabla P_{X_h}u]\!], [\![\nabla v]\!]\rangle_e\right| \le Ch^{k-1}\|u\|_{k+1}|||v|||. \tag{4.8}$$

Substituting (4.7) and (4.8) into (4.6), we have

$$\sup_{v\in X_h, v\neq 0}\frac{|-s_h(P_{X_h}u, v)|}{|||v|||} \le Ch^{k-1}\|u\|_{k+1}. \tag{4.9}$$

From Cauchy–Schwarz inequality and (4.4), we obtain

$$\begin{aligned}
&\sup_{p\in M_h, p\neq 0} \frac{|l_u(p)|}{\|p\|} \\
&= \sup_{p\in M_h, p\neq 0} \frac{\left|\sum_{T\in\mathcal{T}_h}\sum_{i,j=1}^{2}(a_{ij}(I-P_{M_h})\partial_{ij}^2 u, p)_T\right|}{\|p\|} + \sup_{p\in M_h, p\neq 0} \frac{\left|\sum_{T\in\mathcal{T}_h}(c(I-P_{X_h})u, p)_T\right|}{\|p\|} \\
&\leq \sup_{p\in M_h, p\neq 0} \frac{\left|\|a_{ij}\|_{L^\infty(\Omega)}\left(\sum_{T\in\mathcal{T}_h}\sum_{i,j=1}^{2}\|(I-P_{M_h})\partial_{ij}^2 u\|_T^2\right)^{\frac{1}{2}}\left(\sum_{T\in\mathcal{T}_h}\|p\|_T^2\right)^{\frac{1}{2}}\right|}{\|p\|} \\
&\quad + \sup_{p\in M_h, p\neq 0} \frac{\left|\|c\|_{L^\infty(\Omega)}\left(\sum_{T\in\mathcal{T}_h}\|(I-P_{X_h})u\|_T^2\right)^{\frac{1}{2}}\left(\sum_{T\in\mathcal{T}_h}\|p\|_T^2\right)^{\frac{1}{2}}\right|}{\|p\|} \\
&\leq Ch^{k-1}\|u\|_{k+1} + Ch^{k+1}\|u\|_{k+1} \\
&\leq Ch^{k-1}\|u\|_{k+1}. \qquad (4.10)
\end{aligned}$$

Substituting (4.9) and (4.10) into (4.5) completes the proof. □

5 Bivariate Spline Implementation of Algorithm 2.1

We shall use a discontinuous spline space X_h of degree k over a finite element partition $\mathcal{T}_h$ for the primal variable and use another discontinuous spline space M_h of degree k_1, e.g. $k_1 = k-2$ over $\mathcal{T}_h$ for dual variable. When $\mathcal{T}_h$ is a triangulation, these are spline spaces which have been thoroughly studied in [5] and [18]. In this paper, let us explain how to use these spline functions for numerical solution of the second order elliptic PDE (1.1). When $\mathcal{T}_h$ is a triangulation, spline functions use the Bernstein–Bézier representation as explained in [18]. That is, the prime-dual discontinuous Galerkin FEM method discussed in the previous sections can be reformulated by using the Bernstein–Bézier representation. The representation has several nice properties (cf. [18]): (1) the basis functions form a partition of unity, (2) the basis functions are nonnegative, and (3) the basis functions have explicit formulas for their derivatives, integration, their inner product, and triple product integration.

In the remaining of the paper, we use both $u \in X_h$ and its coefficient vector $\mathbf{u}$ in terms of Bernstein–Bézier representation to write a discontinuous spline function u. Similarly, we use both $q \in M_h$ and its coefficient vector $\mathbf{q}$. Most importantly, for any function $u \in X_h$, u is a piecewise polynomial function of degree k over $\mathcal{T}_h$, the jump function $[\![u]\!]$ over an interior edge e of $\mathcal{T}_h$ can be rewritten by using the smoothness conditions between the coefficients of two polynomial pieces $u|_{T_1}$ and $u|_{T_2}$ on their common edge e for triangles $T_1, T_2 \in \mathcal{T}_h$ which share e. See [11] and [18]. The smoothness conditions are linear and all these conditions over each interior edge can be expressed together by using $H\mathbf{u} = 0$ as explained in [5], where H is a rectangular and sparse matrix and $\mathbf{u}$ is the coefficient vector of u.

On the boundary of Ω, u has to satisfy the Dirichlet boundary condition which can be approximated by using a standard polynomial interpolation method, i.e., $u(\mathbf{x})|_e = g(\mathbf{x})$ for $k+1$ distinct points $\mathbf{x} \in e$, where e is a boundary edge of $\mathcal{T}_h$. As u is a polynomial on e, the interpolation condition $u(\mathbf{x})|_e = g(\mathbf{x})$ can be expressed by linear equations in terms of its coefficients. We put these linear equations for all boundary edges together and express them by $B\mathbf{u} = \mathbf{g}$, where B is a rectangular and sparse matrix and $\mathbf{g}$ is a vector consisting of the values of g at the $k+1$ equally-spaced points over e for all boundary edges $e \in \triangle$.

The PDE equation in (2.1) can be discretized by using Bernstein–Bézier representation as follows. We first approximate the right-hand side f by discontinuous spline functions in $S_f \in M_h$. For example, we may choose S_f to be the piecewise polynomial function which interpolates f at the domain points on T of degree k_1 for all triangle $T \in \mathcal{T}_h$, under the assumption that f is a continuous function. For another example, we choose $S_f \in M_h$ such that for each triangle $T \in \mathcal{T}_h$,

$$\int_T f q dx dy = \int_T S_f q dx dy, \quad \forall q \in \mathcal{P}_{k_1}, \tag{5.1}$$

where $\mathcal{P}_{k_1}$ is the standard polynomial space of total degree k_1. It is easy to know that the problem (5.1) has a unique solution of $S_f|_T$. Thus, $S_f \in \mathcal{M}_h$ is well-defined. In fact, we have the following properties

$$\|S_f\| \leq \|f\| \text{ and } \|S_f - f\| = \min_{s \in M_h} \|s - f\|. \tag{5.2}$$

Indeed, we have $\int_T |S_f|^2 dx dy = \int_T f S_f dx dy$ for all $T \in \mathcal{T}_h$ and use Cauchy–Schwarz inequality to have the inequality in (5.2). The equality in (5.2) can be seen from the solution of the least squares problem in (5.1).

We compute the inner product integration on the right-hand of (2.1) exactly by using Theorem 2.34 in [18] and a triple inner product formula. That is, we have

$$\int_\Omega f q dx dy = \int_\Omega S_f q dx dy = \langle M\mathbf{f}, \mathbf{q} \rangle,$$

where $\mathbf{f}$ is the coefficient vector of S_f, M is called the mass matrix which is a blockly diagonal matrix and $\mathbf{q}$ is the coefficient vector of q.

Similarly, we approximate the coefficients a_{ij} by discontinuous spline functions in another discontinuous spline space $S_{ij} \in L_h = S_1^{-1}(\mathcal{T}_h)$ of degree 1, say piecewise linear interpolation of a_{ij}.

$$\int_T a_{ij} \partial_{ij}^2 u q dx dy \approx \int_T S_{i,j} \partial_{ij}^2 u q dx dy, \quad \forall u \in \mathcal{P}_k, q \in \mathcal{P}_{k-2}. \tag{5.3}$$

Once we have S_{ij}, we compute triple product integration on the left-hand side of (2.1). That is, $\int_T S_{ij} \partial_{ij}^2 u q dx dy$ has an exact formula in terms of the coefficients of S_{ij}, u, and q. Thus we have

$$\int_\Omega \sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 u q dx dy \approx \int_\Omega \sum_{i,j=1}^{2} S_{ij} \partial_{ij}^2 u q dx dy = \langle K\mathbf{u}, \mathbf{q} \rangle,$$

where K is the stiffness matrix related to the PDE (1.1).

In order to have an equality in the above formula, we now use the standard L^2 projection P_{M_h} which is defined by $P_{M_h}(v) \in M_h$ such that

$$\langle P_{M_h}(v), q \rangle = \langle v, q \rangle, \forall q \in M_h. \tag{5.4}$$

Thus, we have

$$\int_\Omega \sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 u q dx dy = \left\langle P\left(\sum_{i,j=1}^{2} a_{ij} \partial_{ij}^2 u \right), q \right\rangle = \int_\Omega \sum_{i,j=1}^{2} P_{M_h}(a_{ij} \partial_{ij}^2 u) q dx dy.$$

Since the the projection is linear, we can write

$$\int_\Omega \sum_{i,j=1}^{2} P_{M_h}(a_{ij}\partial_{ij}^2 u) q dx dy = \langle K\mathbf{u}, \mathbf{q}\rangle,$$

for a blockly diagonal matrix K and for all $q \in M_h$. In this way, we obtain a discretized PDE equation: $\langle K\mathbf{u}, \mathbf{q}\rangle = \langle M\mathbf{f}, \mathbf{q}\rangle$ for all $\mathbf{q} \in \mathbb{R}^{d(\mathcal{M}_h)}$ or a linear system:

$$K\mathbf{u} = M\mathbf{f}. \tag{5.5}$$

Note that both M and K can be computed in parallel.

In terms of the Berstein-Bézier representation, the bilinear forms in (2.6) and (2.4) can be rewritten as

$$s(u, v) = h^2\langle H\mathbf{u}, H\mathbf{v}\rangle + h^2\langle B\mathbf{u}, B\mathbf{v}\rangle, \quad \forall u, v \in X_h, \tag{5.6}$$

and

$$b(u, q) = \langle K\mathbf{u}, \mathbf{q}\rangle, \quad \forall u \in X_h, q \in M_h. \tag{5.7}$$

With the above preparation, Algorithm 2.1 can be recast as follows.

Let us consider the following minimization problem for (2.1): Find $\mathbf{u}$ satisfying

$$\min \frac{h^2}{2}(\|H\mathbf{u}\|^2 + \|B\mathbf{u} - \mathbf{g}\|^2), \quad \text{subject to } K\mathbf{u} = M\mathbf{f}. \tag{5.8}$$

Note that the boundary condition is imposed by minimizing the error in an least-squares sense so that the boundary conditions do not need to be strictly enforced.

This minimization problem (5.8) can be reformulated by using Lagrange multiplier method as follows: let

$$L(\mathbf{u}, \lambda) = \frac{h^2}{2}(\|H\mathbf{u}\|^2 + \|B\mathbf{u} - \mathbf{g}\|^2) + \lambda^\top(K\mathbf{u} - M\mathbf{f}), \tag{5.9}$$

where λ is a Lagrange multiplier. Thus, the minimizer $\mathbf{u}^*$ of (5.8) satisfies (5.10). Hence, we have

Algorithm 5.1 *(The Primal-Dual Bivariate Spline Method)* Find a vector pair $(\mathbf{u}^*, \lambda^*) \in \mathbb{R}^{d(X_h)} \times \mathbb{R}^{d(M_h)}$ satisfying

$$\begin{cases} h^2\langle H\mathbf{u}^*, H\mathbf{d}\rangle + h^2\langle B\mathbf{u}, B\mathbf{d}\rangle + \langle \lambda^*, K\mathbf{d}\rangle & = h^2\langle \mathbf{g}, B\mathbf{d}\rangle, \quad \forall \mathbf{d} \in \mathbb{R}^{d(X_h)}, \\ \langle \mathbf{q}, K\mathbf{u}^*\rangle & = \langle \mathbf{q}, M\mathbf{f}\rangle, \quad \forall \mathbf{q} \in \mathbb{R}^{d(M_h)}, \end{cases} \tag{5.10}$$

where $d(X_h)$ is the dimension of X_h and $d(M_h)$ is the dimension of M_h. In fact, $d(X_h) = (k+1)(k+2)N(\mathcal{T}_h)/2$ and $d(M_h) = (k_1+1)(k_1+2)N(\mathcal{T}_h)/2$ with $N(\mathcal{T}_h)$ being the number of triangles in $\mathcal{T}_h$. We shall denote by $u_h \in X_h$ the spline solution with coefficient vector $\mathbf{u}^*$ and similarly, $\lambda_h \in M_h$ with coefficient vector λ^*.

This Algorithm 5.1 will be implemented and numerically experimented in this paper. We will have a flexibility to choose X_h and M_h. In [29], the researchers used $k_1 = k - 2$ and $k_1 = k - 1$. We shall experiment various choices of k_1 and report our numerical results in the next section.

6 Numerical Results Based on Minimization (5.8)

We have implemented Algorithm 5.1 in MATLAB based on the spline function implementation method discussed in [5] which is completely different from the spline functions implemented in [22].

We shall use $S_d^{-1}(\triangle)$ for $d \geq 1$ over a triangulation $\triangle$ and let $S_u \in S_d^{-1}(\triangle)$ be the spline solution with the coefficient vector $\mathbf{c}(u)$ which is the minimizer of (5.8) and report the root mean squared error (RMSE) of $u - S_u$, $\nabla(u - S_u) = (\frac{\partial}{\partial x}(u - S_u), \frac{\partial}{\partial y}(u - S_u))$ and $\nabla^2(u - S_u) = (\frac{\partial^2}{\partial x^2}(u - S_u), \frac{\partial^2}{\partial x \partial y}(u - S_u), \frac{\partial^2}{\partial y^2}(u - S_u))$ based on their values over equally-spaced points, e.g. 1001×1001 grid points located over Ω. More precisely, we report the RMSE of $\nabla(u - S_u)$ which is the average of the RMSE of $\frac{\partial}{\partial x}(u - S_u)$ and the RMSE of $\frac{\partial}{\partial y}(u - S_u)$. Similar for the RMSE of $\nabla^2(u - S_u)$. We shall also present the rates of convergence of RMSE between refinement levels.

The remaining of this section is divided into three subsections. In the first subsection, we present numerical results based on the PDE with smooth coefficients and $c \equiv 0$. We also use smooth solutions to test our spline method. One of purposes is to demonstrate that our MATLAB implementation is correct and is able to produce excellent numerical solution. Another purpose is to compare with the numerical results in [29]. We shall show that the higher order splines produce a much better approximation than using the lower order weak-Galerkin method in [29].

In the next two subsections, we mainly present numerical results from the second order elliptic PDE with discontinuous coefficients and nonsmooth solution which were studied in [24]). Our numerical experiments show that by choosing $X_h = M_h$, the bivariate spline method, i.e. Algorithm 5.1 give a better approximation than the numerical results in [24].

Finally we show some spline solutions for PDE in (1.1) with nonzero function c for smooth and nonsmooth exact solutions. Numerical results are similar to the case when $c \equiv 0$.

6.1 The Case with Smooth Coefficients

In the following examples, we shall use spline spaces $S_d^{-1}(\triangle_\ell)$ of various degrees $d = 2, 3, 4, 5, 6, 7, 8 \ldots$ to solve the PDE of interest, where $\triangle_0$ is a standard triangulation of Ω and $\triangle_\ell$ is the uniform refinement of $\triangle_{\ell-1}$ for $\ell = 1, 2, 3, 4$.

Example 6.1 We begin with a 2nd order elliptic equation with constant coefficients and smooth solution $u = \sin(x)\sin(y)$ which satisfies the following partial differential equation:

$$3\frac{\partial^2}{\partial x^2}u + 2\frac{\partial^2}{\partial x \partial y}u + 2\frac{\partial^2}{\partial y^2}u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2, \tag{6.1}$$

where Ω is a standard square domain $[0, 1]^2$ (cf. [29]). We use $X_h = S_d^{-1}(\triangle_\ell)$ and $M_h = S_{d-2}^{-1}(\triangle_\ell)$ with $h = |\triangle_\ell|$. We use a triangulation $\triangle_0$ which consists of 2 triangles and then uniformly refine $\triangle_0$ repeatedly to obtain $\triangle_\ell$, $\ell = 1, 2, 3, 4, 5$.

Table 1 may be compared with Table 8.1 in [29]. First of all, we recall that there is a superconvergence in L^2 norm approximation in Table 8.1 in [29]. That is, the convergence rate in [29] is about 4 although they only use piecewise polynomials of degree 2. So far there is no mathematical theory to guarantee this superconvergence. Note that the computation of

Table 1 The RMSE of spline solutions using $X_h = S_2^{-1}(\triangle_\ell)$ and $M_h = S_0^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4, 5$ of PDE (6.1)

$\|\triangle\|$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	2.052453e−03	0.00	1.564506e−02	0.00	1.163198e−01	0.00
0.3536	7.574788e−04	1.44	4.728042e−03	1.72	6.078911e−02	0.94
0.1768	2.779251e−04	1.45	1.397469e−03	1.76	3.022752e−02	1.01
0.0884	8.156301e−05	1.77	3.809472e−04	1.88	1.489634e−02	1.03
0.0442	2.161249e−05	1.92	9.836874e−05	1.95	7.401834e−03	1.01

their convergence is based on node points of the underlying triangulation, that is, 6 points per triangle for all triangles in $\mathcal{T}_h$ for each $h > 0$. In our Table 1, the convergence is measured in the RMSE based on 1001×1001 equally-spaced points over Ω and our convergence rate is about 2 for $M_h = S_0^{-1}(\triangle_\ell)$. Nevertheless, our convergence of $\nabla(u_h - u)$ is better than that in Table 8.1 in [29]. Also, we are able to show the convergence in the second order derivatives of $u - u_h$, i.e. the semi-norm $|u - u_h|_{H^2(\Omega)}$.

In the next few tables, we use $X_h = S_k^{-1}(\triangle_\ell)$ and $M_h = S_{k_1}^{-1}(\triangle_\ell)$ with $k_1 \geq 1$. Then the order of convergence will increase if $k_1 = k$. This is an advantage of our numerical algorithm over the numerical method in [29]. For $k = 3$ and $k_1 = 1$, we have numerical results in Table 2.

To increase the convergence rates for $u - u_h$ and $\nabla(u - u_h)$, we use $k_1 = k$ which can be easily adjusted in our MATLAB code. As we can see from Table 3. The convergence and convergence rates are much better than Tables 1 and 2.

Similarly, we can use $k = 4$ and $k_1 = 4$. The numerical results are given in Tables 4, 5 and show that the convergence rate is more than $k = 4$.

Note that in the last row of Table 5, the rate of convergence in L_2 norm is 5.02 which is lower than 5.92. This is because the iterative solution of the linear system achieves the machine precision for this test function using MATLAB. Indeed, if we use $u = \sin(2\pi x)\sin(2\pi y)$ which is slightly harder to approximate than $u = \sin(x)\sin(y)$, the rate of convergence will be around 6. See the rates of convergence in the RMSE of the spline solution shown in Table 6, where the rate is 5.74.

We have tested other solutions (e.g. $u = 1/(1 + x^2 + y^2)$, $u = \sin(\pi x)\sin(\pi y)$, $u = \sin(\pi(x^2 + y^2))$ and etc.. Numerical results are similar to Tables 6, 7, and 8. We can see that the rate of convergence in L_2 norm is optimal for $d \geq 5$ and for sufficiently smooth solutions. That is, the optimal convergence rate is reached when using splines in $S_d^1(\triangle)$ with $d \geq 5$.

Finally, our algorithm is efficient in the following sense: each table above (Tables 5, 6, 7, 8) is generated within 180 seconds based on a desktop computer of 16GB in RAM with Intel Processor i7-3770CPU @3.4GHz speed. For Tables 1, 2, 3, and 4, it takes 550 seconds to generate. Major time is spent on the evaluation of 1001×1001 spline values.

6.2 The Case with Discontinuous Coefficients and Nonsmooth Solution

In this subsection, we shall demonstrate that our method works well for those PDE with discontinuous coefficients which can not be converted into its divergence form. Higher order splines can produce very accurate solutions even the solution is only $C^1(\Omega)$. We shall use two examples studied in [24] each of which has discontinuous PDE coefficients and compare with their results to demonstrate the advantage of our bivariate spline method.

Table 2 The RMSE of spline solutions using $X_h = S_3^{-1}(\triangle_\ell)$ and $M_h = S_1^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4, 5$ of PDE (6.1)

$\lvert\triangle\rvert$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	1.549234e−03	0.00	5.551342e−03	0.00	2.571257e−02	0.00
0.3536	3.614335e−04	2.10	1.266889e−03	2.13	6.506533e−03	1.99
0.1768	8.995656e−05	2.01	3.098134e−04	2.03	1.627964e−03	2.00
0.0884	2.255287e−05	2.00	7.741892e−05	2.00	4.087224e−04	1.99
0.0442	5.639105e−06	2.00	1.935553e−05	2.00	1.026039e−04	1.99

Table 3 The RMSE of spline solutions using $X_h = S_3^{-1}(\triangle_\ell)$ and $M_h = S_3^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4, 5$ of PDE (6.1)

$\lvert\triangle\rvert$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	1.544907e−04	0.00	1.004675e−03	0.00	9.443382e−03	0.00
0.3536	1.044383e−05	3.89	1.351050e−04	2.89	2.474539e−03	1.94
0.1768	8.189057e−07	3.67	1.757983e−05	2.94	6.360542e−04	1.97
0.0884	8.172475e−08	3.32	2.226705e−06	2.98	1.612220e−04	1.98
0.0442	8.968295e−09	3.19	2.803368e−07	2.99	4.053880e−05	1.99

Table 4 The RMSE of spline solutions using $X_h = S_4^{-1}(\triangle_\ell)$ and $M_h = S_4^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4, 5$ of PDE (6.1)

$\lvert\triangle\rvert$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	7.146215e−06	0.00	8.190007e−05	0.00	1.185424e−03	0.00
0.3536	2.645725e−07	4.76	5.224157e−06	3.97	1.449168e−04	3.03
0.1768	1.316127e−08	4.33	3.160371e−07	4.05	1.685747e−05	3.10
0.0884	6.399775e−10	4.36	1.937981e−08	4.03	1.987492e−06	3.08
0.0442	2.456211e−11	4.70	1.200460e−09	4.01	2.409873e−07	3.04

Table 5 The RMSE of spline solutions using $X_h = S_5^{-1}(\triangle_\ell)$ and $M_h = S_5^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4, 5$ of PDE (6.1)

$\lvert\triangle\rvert$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	2.760695e−07	0.00	3.427271e−06	0.00	5.952484e−05	0.00
0.3536	4.721134e−09	5.87	1.113495e−07	4.94	3.938359e−06	3.92
0.1768	7.777767e−11	5.92	3.351050e−09	5.05	2.373035e−07	4.05
0.0884	2.394043e−12	5.02	1.026261e−10	5.03	1.447321e−08	4.04

Example 6.2 We show the performance of our bivariate spline solutions for a PDE with discontinuous coefficients and nonsmooth exact solution $u = xy(e^{1-|x|} - 1)(e^{1-|y|} - 1)$ which satisfies

$$2\frac{\partial^2}{\partial x^2}u + 2\text{sign}(x)\text{sign}(y)\frac{\partial^2}{\partial x \partial y}u + 2\frac{\partial^2}{\partial y^2}u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2 \tag{6.2}$$

Table 6 The RMSE of spline solutions using $X_h = M_h = S_5^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4$ of PDE (6.1) with $u = \sin(2\pi x)\sin(2\pi y)$.

$\|\triangle\|$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	2.390050e−02	0.00	2.699640e−01	0.00	4.628174e+00	0.00
0.3536	4.997698e−04	5.58	1.076435e−02	4.66	3.787099e−01	3.63
0.1768	8.812568e−06	5.83	3.225226e−04	5.06	2.356171e−02	3.99
0.0884	1.648941e−07	5.74	8.620885e−06	5.22	1.260638e−03	4.20

Table 7 The RMSE of spline solutions using $X_h = M_h = S_6^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4$ of PDE (6.1) with $u = \sin(2\pi x)\sin(2\pi y)$.

$\|\triangle\|$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	1.862502e−03	0.00	2.436280e−02	0.00	4.404080e−01	0.00
0.3536	5.460275e−05	5.09	1.350238e−03	4.18	5.202210e−02	3.08
0.1768	5.354973e−07	6.67	2.432368e−05	5.79	1.842914e−03	4.80
0.0884	3.836807e−09	7.12	4.105804e−07	5.89	6.417771e−05	4.84

Table 8 The RMSE of spline solutions using $X_h = M_h = S_7^{-1}(\triangle_\ell)$ for $\ell = 1, 2, 3, 4$ of PDE (6.1) with $u = \sin(2\pi x)\sin(2\pi y)$.

$\|\triangle\|$	$u - S_u$	Rate	$\nabla(u - S_u)$	Rate	$\nabla^2(u - S_u)$	Rate
0.7071	1.167121e−03	0.00	2.022185e−02	0.00	5.174476e−01	0.00
0.3536	4.520586e−06	8.01	1.575837e−04	7.01	8.136845e−03	6.00
0.1768	2.063180e−08	7.78	1.352130e−06	6.87	1.347347e−04	5.92
0.0884	9.814292e−11	7.72	1.032652e−08	7.03	1.947362e−06	6.10

where $u = 0$ on the boundary of $\Omega = [-1, 1] \times [-1, 1]$ as in [24]. As the discontinuity of one of the PDE coefficients are straight lines, we took these lines into consideration when partitioning the underlying domain as seen in Fig. 1. Note that the solution is in $H^2(\Omega)$, but not continuously twice differentiable. We shall use $X_h = S_d^{-1}(\triangle_\ell)$ and $M_h = S_d^{-1}(\triangle_\ell)$ with $\triangle_\ell$ shown in Fig. 1.

Instead of showing the convergence rates of $|u - u_h|_{H^2(\Omega)}$ in a loglog graph for various $d = 2, 3, 4, 5$ as in [24], we present a loglog graph of the root mean squared error (RMSE) of $(|D_x^2(u - u_h)| + |D_x D_y(u - u_h)| + |D_y^2(u - u_h)|)/3$ based on 333×333 equally-spaced points over $\Omega = [-1, 1] \times [-1, 1]$.

The graph in Fig. 2 can be compared with the one in Fig. 2 in [24]. The comparison shows that the accuracy of our spline method is much better. One of the advantages of our method is to be able to use $M_h = S_k^{-1}(\triangle_\ell)$ for various $k > 0$. Our experiments show that the accuracy are getting better from $k = k-2, k-1, k$, but not significantly better for $k = k+1$ or larger.

In order to compare with the numerical method in [29], i.e. to compare Tables 8.5 and 8.6 in [29], we present a similar Table 9 which contains the root mean square error of $|u - u_h|$, $(|D_x(u - u_h)| + |D_y(u - u_h)|)/2$, as well as $(|D_x^2(u - u_h)| + |D_x D_y(u - u_h)| + |D_y^2(u - u_h)|)/3$ which are based on 333×333 equally-spaced points over $[-1, 1] \times [-1, 1]$. We can see that the accuracy of our spline solution in L^2 norm in Table 9 are better than those

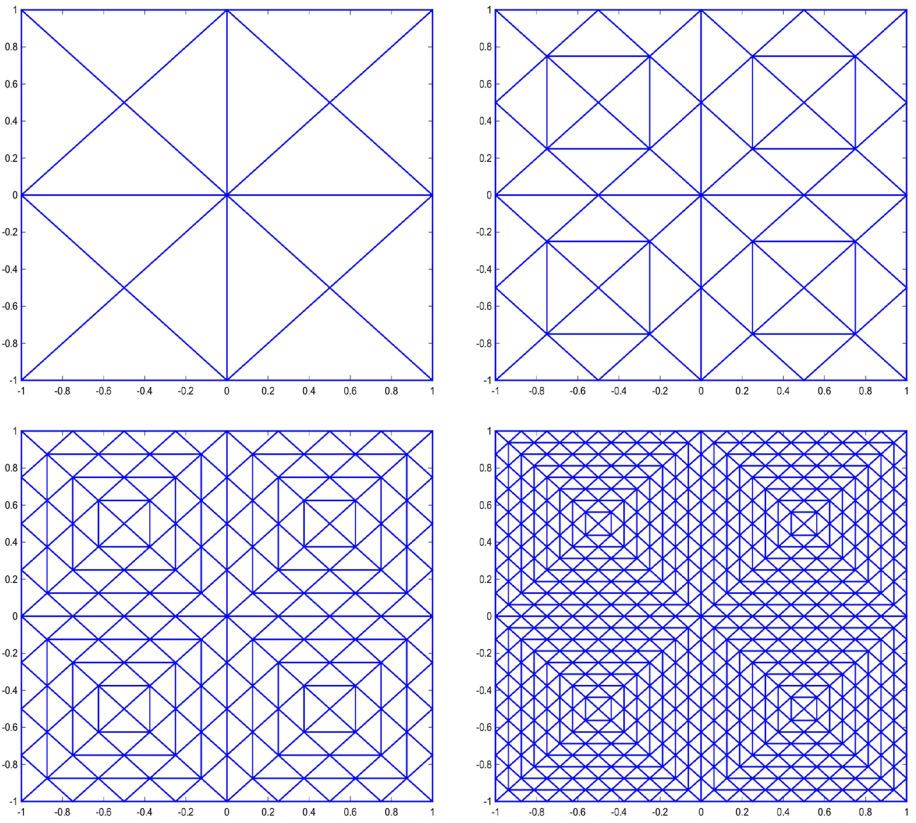

Fig. 1 Triangulations $\triangle_\ell$, $\ell = 0, 1, 2, 3$

in Table 8.5 and are similar to those in Table 8.6 in [29]. The accuracy of our spline solution in H^1 semi-norm is much better than those in Tables 8.5 and 8.6 in [29]. Higher accurate solutions are obtained when splines of higher degrees are used which can be conveniently realized by simply adjusting the degree input parameter in our MATLAB code (Tables 10, 11).

In Table 12, we note that the RMSE$(u - S_u)$ gets deteriorated in the last refinement which indicates that the machine accuracy is achieved and the result could not be improved although the RMSE$(\nabla^2(u - S_u))$ still improves at the expected convergence rate.

Furthermore, when using the degree of splines $d \geq 6$, such a deterioration of iterations continues as the accuracy of the spline coefficient vectors could not be achieved less than $1e$-15 and thus, the RMSE$(u - S_u)$ could not be better than $1e$-12. In order to show the rate of convergence when $d \geq 6$, the computation needs a triple or quadruple precision which will be left for a future study.

Example 6.3 In this example, we study the numerical solution to the following

$$\left(1 + \frac{x^2}{x^2 + y^2}\right) \frac{\partial^2}{\partial x^2} u + \frac{2xy}{x^2 + y^2} \frac{\partial^2}{\partial x \partial y} u + \left(1 + \frac{y^2}{x^2 + y^2}\right) \frac{\partial^2}{\partial y^2} u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2 \tag{6.3}$$

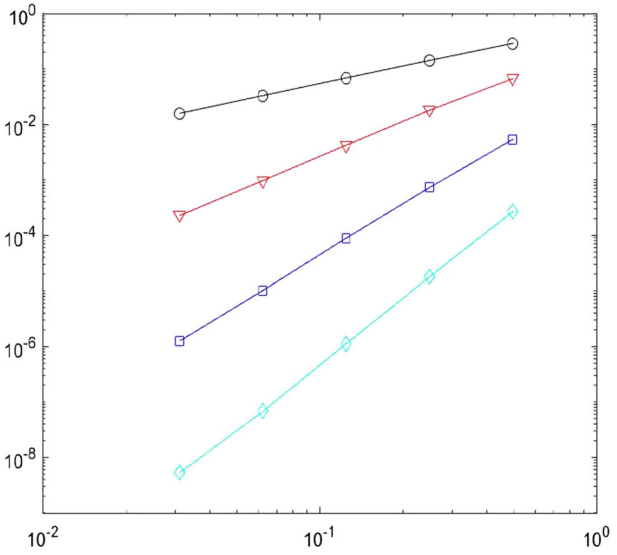

Fig. 2 loglog graph of RMSE $(|D_x^2(u-u_h)|+|D_xD_y(u-u_h)|+|D_y^2(u-u_h)|)/3$ vs the sizes of triangulations 0.5, 0.25, 0.125, 0.0625, 0.03125 for degree $d = 2, 3, 4, 5$ (from the top to the bottom)

Table 9 The RMSE of spline solutions using the pair $X_h = S_2^{-1}(\triangle_\ell)$, $M_h = S_2^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 0, 1, 2, 3, 4$ of PDE (6.2) based on uniform triangulations in Fig. 1

$h = \|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.5000	2.613916e−02	0.00	5.752513e−02	0.00	2.904282e−01	0.00
0.2500	7.300841e−03	1.84	1.676006e−02	1.78	1.428850e−01	1.02
0.1250	1.818047e−03	2.01	4.506334e−03	1.89	6.909645e−02	1.05
0.0625	4.456150e−04	2.03	1.167414e−03	1.95	3.315449e−02	1.06
0.0313	1.099735e−04	2.02	3.004970e−04	1.96	1.585805e−02	1.06

Table 10 The RMSE of spline solutions using the pair $X_h = S_3^{-1}(\triangle_\ell)$, $M_h = S_3^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 0, 1, 2, 3, 4$ of PDE (6.2) based on uniform triangulations in Fig. 1

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.5000	1.449887e−03	0.00	6.243927e−03	0.00	6.723271e−02	0.00
0.2500	9.612402e−05	3.91	7.599219e−04	3.04	1.805996e−02	1.90
0.1250	1.862840e−05	2.37	8.851518e−05	3.10	4.242236e−03	2.09
0.0625	3.991714e−06	2.22	1.242605e−05	2.83	9.748695e−04	2.13
0.0313	7.386041e−07	2.43	2.046330e−06	2.60	2.276661e−04	2.10

where $\Omega = (0, 1)^2$, $u = (x^2 + y^2)^{\alpha/2}$. Note that the middle coefficient $\frac{2xy}{x^2+y^2}$ fails to be continuous at one corner of Ω. This PDE has been studied in [14] and [20] to explain the possibility of ill-posedness of the problem. In [24] and [29], two numerical methods find a good approximation of the solution. We shall apply our spline method to find approximations

Table 11 The RMSE of spline solutions using the pair $X_h = S_4^{-1}(\triangle_\ell)$, $M_h = S_4^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 0, 1, 2, 3, 4$ of PDE (6.2) based on uniform triangulations in Fig. 1

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.5000	2.009743e−05	0.00	2.523989e−04	0.00	5.413090e−03	0.00
0.2500	8.960082e−07	4.49	1.869987e−05	3.75	7.291647e−04	2.89
0.1250	7.946654e−08	3.50	1.149021e−06	4.02	8.933381e−05	3.03
0.0625	6.918610e−09	3.52	6.792969e−08	4.08	1.026479e−05	3.12
0.0313	7.869320e−10	3.14	4.272762e−09	3.99	1.257353e−06	3.04

Table 12 The RMSE of spline solutions using the pair $X_h = S_5^{-1}(\triangle_\ell)$, $M_h = S_5^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 0, 1, 2, 3, 4$ of PDE (6.2) based on uniform triangulations in Fig. 1

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.5000	5.917644e−07	0.00	1.033305e−05	0.00	2.729830e−04	0.00
0.2500	1.359283e−08	5.44	3.622443e−07	4.83	1.821993e−05	3.90
0.1250	5.050760e−10	4.75	1.146421e−08	4.98	1.140314e−06	4.00
0.0625	3.618925e−11	3.80	3.519986e−10	5.03	6.848577e−08	4.06
0.0313	1.530946e−10	−2.08	2.946980e−10	0.26	4.240589e−09	4.01

Table 13 The RMSE of spline solutions using the pair $X_h = S_2^{-1}(\triangle_\ell)$, $M_h = S_0^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 0, 1, 2, 3, 4$ of PDE (6.3) based on uniform refinements of a simple triangulation

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.5000	4.800003e−03	0.00	2.362686e−02	0.00	2.256637e−01	0.00
0.2500	1.883197e−03	1.35	9.919694e−03	1.25	1.497819e−01	0.58
0.1250	6.785580e−04	1.47	3.798937e−03	1.38	9.673684e−02	0.62
0.0625	2.521239e−04	1.43	1.426283e−03	1.41	6.021246e−02	0.67
0.0313	1.018909e−04	1.31	5.472435e−04	1.38	3.582212e−02	0.74

of the exact solution when $\alpha = 1.6$ as in the previous literature. First, we use standard uniform refinements of a simple triangulation of Ω by adding two diagonals.

We can see that although the accuracy of our spline solution in L^2 norm in Table 13 are not as good as those in Tables 8.7 and 8.8 in [29], the accuracy of our spline solutions in H^1 semi-norm is better.

In [24], Smears and Süli provided a numerical method to be able to achieve the convergence in an exponential decay fashion by designing a sequence of quadrilateral partitions. In this paper, we provide a simple approach to improve the numerical solution of the PDE (6.3) by starting with a special triangulation in Fig. 3 since the solution at one of the corners of Ω is singular and then uniformly refine it to obtain a sequence of triangulations. Over such a sequence of triangulation, numerical results from our spline method in Table 14 are much better than those in [29], and better than the one [24] in H^2 semi-norm although we use much more elements and degrees of freedom.

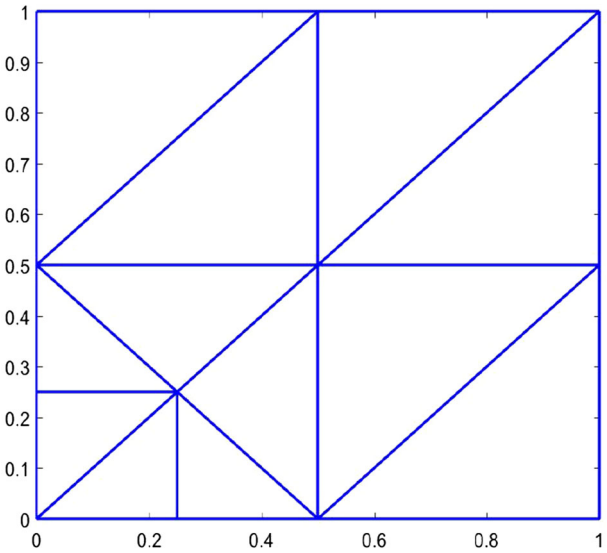

Fig. 3 A fixed triangulation $\triangle$

Table 14 The RMSE of spline solutions using the pair $X_h = S_2^{-1}(\triangle_\ell)$ and $M_h = S_0^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4, 5, 6$ of PDE (6.3) based on uniform refinements of a fixed triangulation

$\lvert\triangle\rvert$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	EMSE$(\nabla^2(u - S_u))$	Rate
0.3536	2.808940e−03	0.00	1.135469e−02	0.00	1.353720e−01	0.00
0.1768	1.263526e−03	1.15	5.183195e−03	1.13	8.234064e−02	0.72
0.0884	4.402893e−04	1.52	1.882706e−03	1.46	4.780946e−02	0.80
0.0442	1.280195e−04	1.78	5.703663e−04	1.72	2.661356e−02	0.85
0.0221	3.452513e−05	1.89	1.573659e−04	1.86	1.317950e−02	1.00
0.0110	9.078391e−06	1.93	4.181275e−05	1.91	6.370552e−03	1.04

We have tested our spline method for numerical solution of (6.3) for various degrees of splines for primal and/or dual variables. We do not report the results here due to the space limitation.

6.3 Numerical Results of PDE in (1.1) with Nonzero *c*

In this subsection, we present some numerical results from our bivariate spline method for numerical solution of the PDE in (1.1) with nonzero c. We use three examples to demonstrate that our method is effective and efficient no matter the PDE coefficients are smooth or not smooth and the solutions are smooth or not so smooth.

Example 6.4 We begin with a 2nd order elliptic equation with smooth coefficients and smooth solution $u = \sin(\pi x)\sin(\pi y)$ which satisfies the following partial differential equation:

$$3\frac{\partial^2}{\partial x^2}u + 2\frac{\partial^2}{\partial x \partial y}u + 2\frac{\partial^2}{\partial y^2}u - (1 + x^2 + y^2)u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2, \qquad (6.4)$$

Table 15 The RMSE of spline solutions using the pair $X_h = S_5^{-1}(\triangle_\ell)$, $M_h = S_3^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4$ of PDE (6.4)

$\lvert\triangle\rvert$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	RMSE$(\nabla^2(u - S_u))$	Rate
0.7071	7.148997e−04	0.00	5.688698e−03	0.00	7.513832e−02	0.00
0.3536	2.651667e−05	4.75	1.861396e−04	4.93	4.596716e−03	3.99
0.1768	1.257317e−06	4.40	6.093065e−06	4.93	2.814578e−04	4.03
0.0884	7.088746e−08	4.15	2.550376e−07	4.58	1.753997e−05	4.00

Table 16 The RMSE of spline solutions using the pair $X_h = S_5^{-1}(\triangle_\ell)$, $M_h = S_5^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4$ of PDE (6.5)

$\lvert\triangle\rvert$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	RMSE$(\nabla^2(u - S_u))$	Rate
0.7071	5.906274e−04	0.00	5.894580e−03	0.00	9.080845e−02	0.00
0.3536	1.155544e−05	5.68	1.785330e−04	5.05	5.673534e−03	3.97
0.1768	3.265019e−07	5.15	4.834318e−06	5.21	3.150799e−04	4.15
0.0884	1.568269e−08	4.38	1.463029e−07	5.05	1.866297e−05	4.07

Table 17 The RMSE of spline solutions using the pair $X_h = S_5^{-1}(\triangle_\ell)$, $M_h = S_5^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4$ of PDE (6.6)

$\lvert\triangle\rvert$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	RMSE$(\nabla^2(u - S_u))$	Rate
0.7071	2.914706e−02	0.00	2.813852e−01	0.00	4.122223e+00	0.00
0.3536	8.047145e−04	5.18	1.287751e−02	4.46	3.279903e−01	3.63
0.1768	2.963898e−05	4.76	3.942771e−04	5.02	1.893540e−02	4.06
0.0884	1.403774e−06	4.40	1.307268e−05	4.91	1.139668e−03	4.05

where Ω is a standard square domain $[-1, 1]^2$ which is split into 4 equal sub-squares and each sub-square is split into 2 triangles to form an initial triangulation $\triangle_0$. Let $\triangle_\ell$ be the ℓth uniform refinement of $\triangle_0$. See numerical results in Table 15.

Example 6.5 In this example, we use our spline method to solve the following PDE with discontinuous coefficients, but smooth solution.

$$a(x, y)\frac{\partial^2}{\partial x^2}u + b(x, y)\frac{\partial^2}{\partial x \partial y}u + c(x, y)\frac{\partial^2}{\partial y^2}u - (1+x^2+y^2)u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2 \tag{6.5}$$

where $a(x, y) = 1 + |x|$, $b(x, y) = (xy)^{1/3}$, $c(x, y) = 1 + |y|$ and Ω is a standard domain $[-1, 1]^2$. We use $u = \sin(\pi x)\sin(\pi y)$ as the exact solution. The same triangulations $\triangle_\ell$ as in Example 6.4 will be used. See Table 16 for numerical results.

Example 6.6 In this example, we show the performance of our spline solutions for a PDE with discontinuous coefficients and nonsmooth exact solution $u = xy(e^{1-|x|} - 1)(e^{1-|y|} - 1)$ which satisfies

Table 18 The RMSE of spline solutions using the pair $X_h = S_6^{-1}(\triangle_\ell)$, $M_h = S_6^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4$ of PDE (6.6)

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	RMSE$(\nabla^2(u - S_u))$	Rate
0.7071	3.494139e−06	0.00	1.538901e−05	0.00	2.054258e−04	0.00
0.3536	7.686885e−08	5.51	3.531160e−07	5.45	7.914461e−06	4.70
0.1768	1.370000e−09	5.81	6.565436e−09	5.75	2.731226e−07	4.86
0.0884	1.598556e−11	6.42	7.840289e−11	6.39	8.199009e−09	5.06

Table 19 The RMSE of spline solutions using the pair $X_h = S_7^{-1}(\triangle_\ell)$, $M_h = S_7^{-1}(\triangle_\ell)$ of spline spaces for $\ell = 1, 2, 3, 4$ of PDE (6.6)

$\|\triangle\|$	RMSE$(u - S_u)$	Rate	RMSE$(\nabla(u - S_u))$	Rate	RMSE$(\nabla^2(u - S_u))$	Rate
0.7071	6.099647e−08	0.00	4.914744e−07	0.00	1.004845e−05	0.00
0.3536	9.745659e−10	5.97	4.297639e−09	6.84	1.659379e−07	5.92
0.1768	9.091448e−12	6.74	5.183151e−11	6.37	3.038997e−09	5.78

$$2\frac{\partial^2}{\partial x^2}u + 2\text{sign}(x)\text{sign}(y)\frac{\partial^2}{\partial x \partial y}u + 2\frac{\partial^2}{\partial y^2}u - (1 + x^2 + y^2)u = f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2 \tag{6.6}$$

where $u = 0$ on the boundary of $\Omega = [-1, 1] \times [-1, 1]$ as in [24]. Note that the solution is in $H^2(\Omega)$, but not continuously twice differentiable. The same triangulations $\triangle_\ell$ as in Example 6.4 were used and $S_5^1(\triangle_\ell)$ were used to solve the PDE in (6.6). The RMSE for spline approximation to the exact solution is shown in Tables 17, 18, and 19.

References

1. Adolfsson, V.: L^2 integrability of second order derivatives for Poisson equations in nonsmooth domain. Math. Scand. **70**, 146–160 (1992)
2. Agranovich, M.S.: Sobolev Spaces, Their Generalizations and Elliptic Problems in Smooth and Lipschitz Domains, Springer Monographs in Mathematics. Springer, Cham (2015)
3. Awanou, G.: Robustness of a spline element method with constraints. J. Sci. Comput. **36**(3), 421–432 (2008)
4. Awanou, G.: Spline element method for Monge–Ampère equations. BIT **55**(3), 625–646 (2015)
5. Awanou, G., Lai, M.-J., Wenston, P.: The multivariate spline method for scattered data fitting and numerical solution of partial differential equations. In: Chen, G., Lai, M.J. (eds.) Wavelets and splines: Athens 2005, pp. 24–74. Nashboro Press, Brentwood (2006)
6. Babuska, I.: The finite element method with Lagrange multipliers. Numer. Math. **20**, 179–192 (1973)
7. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Springer, New York (1994)
8. Brezzi, F.: On the existence, uniqueness, and approximation of saddle point problems arising from Lagrange multipliers. RAIRO **8**, 129–151 (1974)
9. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, New York (1978)
10. Evens, L.: Partial Differ. Equ. American Mathematical Society, Providence (1998)
11. Farin, G.: Triangular Bernstein–Bézier patches. Comput. Aided Geom. Des. **3**(2), 83–127 (1986)
12. Floater, M., Lai, M.-J.: Polygonal spline spaces and the numerical solution of the Poisson equation. SIAM J. Numer. Anal. **54**, 797–824 (2016)
13. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Pitman, Boston (1985)
14. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (1998)

15. Gutierrez, J., Lai, M.-J., Slavov, G.: Bivariate spline solution of time dependent nonlinear PDE for a population density over irregular domains. Math. Biosci. **270**, 263–277 (2015)
16. Hu, X., Han, D., Lai, M.-J.: Bivariate splines of various degrees for numerical solution of PDE. SIAM J. Sci. Comput. **29**, 1338–1354 (2007)
17. Kellogg, O.D.: On bounded polynomials in several variables. Math. Z. **27**, 55–64 (1928)
18. Lai, M.-J., Schumaker, L.L.: Spline Functions Over Triangulations. Cambridge University Press, Cambridge (2007)
19. Lai, M.-J., Wenston, P.: Bivariate splines for fluid flows. Comput. Fluids **33**, 1047–1073 (2004)
20. Maugeri, A., Palagachev, D.K., Softova, L.G.: Elliptic and Parabolic Equations with Discontinuous Coefficients. Mathematical Research, vol. 109. Wiley-VCH Verlag, Berlin (2000)
21. Mitrea, Dorina, Mitrea, Marius, Yan, Lixin: Boundary value problems for the Laplacian in convex and semiconvex domains. J. Funct. Anal. **258**, 2507–2585 (2010)
22. Schumaker, L.L.: Spline Functions: Computational Methods. SIAM Publication, Philadelphia (2015)
23. Smears, I.: Nonoverlapping domain decomposition preconditioners for discontinuous Galerkin approximations of Hamilton–Jacobi–Bellman equations. J. Sci. Comput. (2017). doi:10.1007/s10915-017-0428-5
24. Smears, I., Süli, E.: Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordes coefficients. SIAM J. Numer. Anal. **51**(4), 2088–2106 (2013)
25. Smears, I., Süli, E.: Discontinuous Galerkin finite element approximation of Hamilton–Jacobi–Bellman equations with Cordes coefficients. SIAM J. Numer. Anal. **52**(2), 993–1016 (2014)
26. Smears, I., Süli, E.: Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordes coefficients. Numer. Math. **133**(1), 141–176 (2016)
27. Süli, E.: A brief excursion into the mathematical theory of mixed finite element methods. In: Lecture Notes. University of Oxford (2013)
28. Wang, C., Wang, J.: An efficient numerical scheme for the biharmonic equation by weak Galerkin finite element methods on polygonal or polyhedral meshes. Comput. Math. Appl. **68**(12, part B), 2314–2330 (2014)
29. Wang, C., Wang, J.: A primal–dual weak Galerkin finite element method for second order elliptic equations in non-divergence form, in revision, submitted to Math. Comput. arXiv:1510.03488v1
30. Wang, J., Ye, X.: A weak Galerkin mixed finite element method for second-order elliptic problems. Math. Comput. **83**, 2101–2126 (2014)