Numerische Mathematik (2019) 143:583–604
https://doi.org/10.1007/s00211-019-01067-1

Numerische Mathematik

A primal discontinuous Galerkin method with static condensation on very general meshes

Alexei Lozinski[1]

Received: 19 March 2018 / Revised: 15 July 2019 / Published online: 27 July 2019

Abstract
We propose an efficient variant of a primal discontinuous Galerkin method with interior penalty for the second order elliptic equations on very general meshes (polytopes with eventually curved boundaries). Efficiency, especially when higher order polynomials are used, is achieved by static condensation, i.e. a local elimination of certain degrees of freedom cell by cell. This alters the original method in a way that preserves the optimal error estimates. Numerical experiments confirm that the solutions produced by the new method are indeed very close to that produced by the classical one.

Mathematics Subject Classification 65N12 · 65N30

1 Introduction

The recent years have seen the emergence (or the revival) of several numerical methods capable to solve approximately elliptic partial differential equations using general polygonal/polyhedral meshes. This is witnessed for example by the book [4]. The methods reviewed in this book (if we restrict our attention only to finite element type methods using piecewise polynomial approximation spaces in one form or another) include interior penalty discontinuous Galerkin (DG) methods [2,8], hybridizable discontinuous Galerkin (HDG) methods ([10], introduced in [12]), the Virtual Element (VE) method ([7], introduced in [5,6]), the Hybrid High-Order (HHO) method ([16], introduced in [14,15]). One can add to this list the weak Galerkin finite element [22], which is similar to HDG. The relations between HHO and HDG methods were exhibited in [11].

Among the above, the primal interior penalty DG methods are the most classical. In the symmetric form, also referred to as SIP—symmetric interior penalty, this

✉ Alexei Lozinski
alexei.lozinski@univ-fcomte.fr

1 Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, Univ. Bourgogne Franche-Comté, 16 route de Gray, 25030 Besançon Cedex, France

method dates back to [3,24] and is now presented and thoroughly studied in several monographs, for example [13,21]. It is well suited to the discretization on very general meshes because its approximation space is populated by polynomials of degree, say $\leq k$, on each mesh cell without any constraints linking the polynomials on two adjacent cells. It leaves thus a lot of freedom on the choice of the mesh cells which can be not only polytopes but also virtually any geometrical shapes. It is generally admitted however that the SIP method is too expensive especially when higher order polynomials are employed. Indeed, its cost, i.e. the dimension of the approximation space, is the product of the number of mesh cells and the dimension of the space of polynomials of degree $\leq k$. The cost on a given mesh is thus proportional to k^2 in 2D (resp. k^3 in 3D). This should be contrasted with the cost of HDG or HHO methods which is proportional to k in 2D (resp. k^2 in 3D).

The goal of the present article is to modify the SIP method so that its cost is reduced to that of HDG or HHO methods. In doing so, we inspire ourselves from the static condensation procedure for the standard continuous Galerkin (CG) finite element methods. It is indeed well known that the dimension of the approximation space in CG is proportional to k^2 in 2D on a given mesh, but the degrees of freedom interior to each mesh cell can be locally eliminated which leaves a global problem of the size proportional to k (these numbers are changed to, respectively, k^3 and k^2 in 3D). Although the notion of interior degrees of freedom does not make sense in the DG context, we shall be able to select, on each mesh cell, a subspace of the approximating polynomials that can be used to construct a local approximation through the solution of a local problem. The remaining degrees of freedom will then be used in a global problem. We shall thus achieve a significant reduction of the global problem size in the DG SIP-like method, similarly to that achieved in CG by static condensation. The resulting DG method, which can be refereed to as scSIP (static condensation SIP), will not produce exactly the same approximation as the original SIP method. We shall prove however that these two solutions satisfy the same optimal a priori error bounds in H^1 and L^2 norms. Moreover, they turn out to be very close to each other in our numerical experiments.

We treat here only the diffusion equation with variable, but sufficiently smooth, coefficients. The extension to other problems, such as convection–reaction–diffusion, linear elasticity, Stokes, as well as to other DG variants (IIP, NIP) seems relatively straight-forward. Our assumptions on the mesh allow for cells of general shape, not necessarily the polytopes.

The article is organized as follows: in the next section, we present the idea of our method starting by the description of the governing equations. We then recall the static condensation for the classical CG FEM. Our variants of DG FEM (SIP and scSIP) are first introduced in Sect. 2.2. The convergence proofs are in Sect. 3. They are done assuming some properties of the discontinuous FE spaces and the underlying mesh. In Sect. 4, we give an example of the hypotheses on the mesh under which the necessary properties of the FE spaces can be established. Finally, some implementation details and numerical illustrations are presented in Sect. 5.

2 Description of the problem and static condensation for FEM (CG and DG cases)

We consider the second-order elliptic problem

$$\mathscr{L}u = f \quad \text{in } \Omega, \quad u = g \quad \text{on } \partial\Omega \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, is a bounded Lipschitz domain, $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ are given functions. The differential operator $\mathscr{L}$ is defined by

$$\mathscr{L}u = -\partial_i(A_{ij}(x)\partial_j u)$$

with ∂_i denoting the partial derivative in the direction x_i, $i = 1, \ldots, d$ and assuming the summation over i, j. The coefficients A_{ij} are supposed to form a positive definite matrix $A = A(x)$ for any $x \in \Omega$ which is sufficiently smooth with respect to x so that

$$\alpha|\xi|^2 \le \xi^T A(x)\xi \le \beta|\xi|^2, \quad \forall \xi \in \mathbb{R}^d, x \in \Omega \tag{2}$$

and

$$|\nabla A_{ij}(x)| \le M, \quad \forall x \in \Omega, i, j = 1, \ldots, d \tag{3}$$

with some constants $\beta \ge \alpha > 0$, $M > 0$.

2.1 Static condensation for CG FEM

To present our idea, we start by recalling the idea of static condensation, going back to [19], as applied to the usual CG finite element method for problem (1). Let us assume for the moment (in this subsection only) that Ω is a polygon (polyhedron) and introduce a conforming mesh $\mathscr{T}_h$ on Ω consisting of triangles (tetrahedra). Assuming for simplicity $g = 0$, the usual continuous $\mathbb{P}_k$ finite element discretization of (1) is then written: find $u_h \in W_h$ such that

$$a(u_h, v_h) := \int_\Omega A\nabla u_h \cdot \nabla v_h = \int_\Omega f v_h, \quad \forall v_h \in W_h \tag{4}$$

where W_h is the space of continuous piecewise polynomial functions (polynomials of degree $\le k$ on each mesh cell $T \in \mathscr{T}_h$ for some $k \ge 1$) vanishing on $\partial\Omega$. The size of this problem, i.e. the dimension of W_h, is of order k^2 on a given mesh in 2D (resp. k^3 in 3D). To reduce this cost, one can decompose the space W_h as follows

$$W_h = W_h^{loc} \oplus W_h'$$

where the subspace W_h^{loc} consists of functions of W_h that vanish on the boundaries of all the mesh cells $T \in \mathscr{T}_h$, and W_h' is the complement of W_h^{loc}, orthogonal with respect to the bilinear form a. Decomposing $u_h = u_h^{loc} + u_h'$ with $u_h^{loc} \in W_h^{loc}$ and $u_h' \in W_h'$ we see that (4) is split into two independent problems

$$u_h^{loc} \in W_h^{loc} : \qquad a(u_h^{loc}, v_h^{loc}) = \int_\Omega f v_h^{loc}, \quad \forall v_h^{loc} \in W_h^{loc} \tag{5}$$

$$u_h' \in W_h' : \qquad a(u_h', v_h') = \int_\Omega f v_h', \quad \forall v_h' \in W_h' \tag{6}$$

The first problem above is further split into a collection of mutually independent problems on every mesh cell $T \in \mathscr{T}_h$:

$$\text{Find } u_h^{loc,T} \text{ such that } \int_T A\nabla u_h^{loc,T} \cdot \nabla v_h^{loc,T} = \int_\Omega f v_h^{loc,T}, \quad \forall v_h^{loc,T} \in W_h^{loc,T} \tag{7}$$

where $W_h^{loc,T}$ is the restriction of W_h^{loc} on T, i.e. the set of all polynomials of degree $\leq k$ vanishing on ∂T. The cost of solution of these local problems is negligible and we thus get very cheaply $u_h^{loc}|_T = u_h^{loc,T}$. Note also that Problem (7) can be recast as

$$\pi_T \mathscr{L}(u_h^{loc}|_T) = \pi_T f, \quad \forall T \in \mathscr{T}_h \tag{8}$$

where π_T is the projection to $W_h^{loc,T}$, orthogonal in $L^2(T)$.

On the other hand, Problem (6) remains global but its size is only proportional to k in 2D (resp. k^2 in 3D) which is much smaller than that of the original problem (4). Indeed, the degrees of freedom are associated to the standard interpolation points of $\mathbb{P}^k$ finite elements on the edges of the mesh. Note also that a basis for W_h' can be constructed solving cheap local problems of the type

$$\pi_T \mathscr{L}(v_h'|_T) = 0, \quad \forall T \in \mathscr{T}_h \tag{9}$$

with appropriate boundary conditions on ∂T insuring the continuity of functions in W_h'.

2.2 DG FEM: SIP and scSIP methods

We turn now to the main subject of this paper: the DG methods. We now let Ω be a bounded domain of general shape, and $\mathscr{T}_h$ be a splitting of Ω into a collection of non-overlapping subdomains (again of general shape, the precise definitions and assumptions on the mesh will be given Sects. 3 and 4). Let V_h denote the space of discontinuous piecewise polynomial functions of degree $\leq k$ on each mesh cell $T \in \mathscr{T}_h$ for some $k \geq 2$[1]:

$$V_h = \{v \in L^2(\Omega) : v|_T \in \mathbb{P}_k(T), \forall T \in \mathscr{T}_h\} \tag{10}$$

[1] The usual SIP DG method makes perfect sense also for piecewise linear polynomials ($k = 1$). We restrict ourselves however to $k \geq 2$ since the forthcoming modification of the method allowing for the static condensation is pertinent to this case only.

The SIP (symmetric interior penalty) method is then written as: find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = L_h(v_h), \quad \forall v_h \in V_h \tag{11}$$

with the bilinear form a_h and the linear form L_h defined by

$$\begin{aligned} a_h(u, v) = &\sum_{T \in \mathscr{T}_h} \int_T A\nabla u \cdot \nabla v - \sum_{E \in \mathscr{E}_h} \int_E (\{A\nabla u \cdot n\}[v] + \{A\nabla v \cdot n\}[u]) \\ &+ \sum_{E \in \mathscr{E}_h} \frac{\gamma}{h_E} \int_E [u][v] \end{aligned} \tag{12}$$

and

$$L_h(v) = \sum_{T \in \mathscr{T}_h} \int_T fv + \sum_{E \in \mathscr{E}_h^b} \int_E g\left(\frac{\gamma}{h_E} v - A\nabla v \cdot n\right) \tag{13}$$

where $\mathscr{E}_h$ is the set of all the edges/faces of the mesh, $\mathscr{E}_h^b \subset \mathscr{E}_h$ regroups the edges/faces on the boundary $\partial\Omega$, n, $[\cdot]$ and $\{\cdot\}$ denote the unit normal, the jump and the mean over $E \in \mathscr{E}_h$. More precisely, for any internal facet E shared by two mesh cells T_1^E and T_2^E, we choose n as the unit vector, normal to E and looking from T_1^E to T_2^E. We then define for any function v which is H^1 on both T_1^E and T_2^E but discontinuous across E

$$[v]|_E := v|_{T_1^E} - v|_{T_2^E}, \quad \{A\nabla v \cdot n\}|_E := \frac{1}{2} A\left(\nabla v|_{T_1^E} + \nabla v|_{T_2^E}\right) \cdot n$$

On a boundary edge $E \in \mathscr{E}_h^b$, n is the unit normal looking outward Ω and $[v] = v$, $\{A\nabla v \cdot n\} = A\nabla v \cdot n$. The parameter h_E in (12)–(13) is the local length scale of the mesh near the facet E which will be properly defined in Lemma 1.[2] Finally, γ is the interior penalty parameter which should be chosen sufficiently large.

Unlike the case of continuous finite elements, Problem (11) does not allow directly for a static condensation. However, we can construct a modification of (11) that mimics the characterization of local and global components of the solution by the projectors on local polynomial spaces (8)–(9). These spaces are now defined simply as

$$V_h^{loc,T} = \mathbb{P}^{k-2}(T)$$

We also let $\pi_{T,k-2}$ to be the projection to $V_h^{loc,T}$, orthogonal in $L^2(T)$, and propose the following scheme:

[2] The usual choice $h_E = \mathrm{diam}(E)$ is not appropriate on general meshes since some of the facets can be of much smaller diameter than that of the adjacent cell.

– Compute $u_h^{loc} \in V_h$ by solving

$$\pi_{T,k-2}\mathscr{L}(u_h^{loc}|_T) = \pi_{T,k-2} f, \quad \forall T \in \mathscr{T}_h \tag{14}$$

I.e. find $u_h^{loc}|_T \in \mathbb{P}^k(T)$ on all mesh cells $T \in \mathscr{T}_h$ such that

$$\int_T \mathscr{L}(u_h^{loc}|_T) q_T = \int_T f q_T, \quad \forall q_T \in \mathbb{P}^{k-2}(T)$$

– Define the subspace of V_h

$$V_h' = \left\{ v_h' \in V_h : \pi_{T,k-2}\mathscr{L}(v_h'|_T) = 0, \quad \forall T \in \mathscr{T}_h \right\} \tag{15}$$

I.e. the subspace of functions $v_h' \in V_h$ such that

$$\int_T \mathscr{L}(v_h'|_T) q_T = 0, \quad \forall q_T \in \mathbb{P}^{k-2}(T) \tag{16}$$

on all mesh cells $T \in \mathscr{T}_h$.

– Compute $u_h' \in V_h'$ such that

$$a_h(u_h', v_h') = L_h(v_h') - a_h(u_h^{loc}, v_h'), \quad \forall v_h' \in V_h' \tag{17}$$

– Set

$$u_h = u_h^{loc} + u_h'. \tag{18}$$

We shall show that the local problem (14) admits an infinity of solutions. We can choose any of these solutions on each mesh cell to form u_h^{loc}. Nevertheless, the final result u_h given by (18) is unique, cf. Lemma 3.

Note that the dimension of the "global" space V_h' on a given mesh is of order k in 2D (k^2 in 3D) so that global problem (17) is much cheaper than (11) for large k. We have thus asymptotically the same costs for the global problems as for CG FEM with static condensation. There is though a fundamental difference between static condensation approaches in CG and DG cases from the implementation point of view: the basis functions for the global space W_h' in the CG case are known a priori, whereas those for the space V_h' in the DG case should be constructed as solutions to local problems (16), cf. the discussion of the implementation issues in Sect. 5.1. Note however that one can get rid of problems (16) in the special case of a constant coefficient matrix A in (1), cf. Remark 1. Indeed, (16) is reduced in this case to $\mathscr{L}(v_h'|_T) = A_{ij}\partial_i\partial_j(v_h'|_T) = 0$ on any cell $T \in \mathscr{T}_h$. The structure of V_h' does not thus vary from one cell to another and a basis for V_h' can be chosen a priori on all the cells.

The local projection step (14) is not necessarily consistent with the original formulation (11) so that the solution u_h given by (14)–(18) is different from that of (11). We shall prove however that SIP and scSIP approximations satisfy the same a priori error bounds. Moreover, they turn out to be very close to each other in numerical experiments.

3 Well posedness of SIP and scSIP methods and a priori error estimates

Let us now be more precise about the hypotheses on the mesh. Recall that $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, is a Lipschitz bounded domain and $\mathscr{T}_h$ is a general (not necessarily polygonal or polyhedral) mesh on Ω. We mean by this that $\mathscr{T}_h$ is a decomposition of Ω into mutually disjoint cells $\bar{\Omega} = \cup_{T \in \mathscr{T}_h} \bar{T}$ so that each cell T is a Lipschitz subdomain of Ω and for every $T_1, T_2 \in \mathscr{T}_h$ we have either $T_1 = T_2$ or $T_1 \cap T_2 = \varnothing$ (the cells $T_1 \in \mathscr{T}_h$ are treated here as open sets). We also introduce the sets of internal and boundary edges/faces as respectively

$$\mathscr{E}_h^i = \{E = \bar{T}_1 \cap \bar{T}_2 \text{ for some } T_1, T_2 \in \mathscr{T}_h\}$$
$$\mathscr{E}_h^b = \{E = \bar{T} \cap \partial\Omega \text{ for some } T \in \mathscr{T}_h\}$$

and denote by $\mathscr{E}_h := \mathscr{E}_h^i \cup \mathscr{E}_h^b$ the union of all the edges/faces.

Let B_T, for any $T \in \mathscr{T}_h$, denote the smallest ball containing T, and B_T^{in} denote the largest ball inscribed in T. Set $h_T = \operatorname{diam}(T)$ and $h = \max_{T \in \mathscr{T}_h} h_T$. From now on, we assume that mesh $\mathscr{T}_h$ is

- **Shape regular**: there is a mesh-independent parameter $\rho_1 > 1$ such that, for all $T \in \mathscr{T}_h$,

$$R_T \leq \rho_1 r_T \tag{19}$$

  where r_T is the radius of B_T^{in} and R_T is the radius of B_T. This also implies $h_T \leq 2\rho_1 r_T$ and $R_T \leq \rho_1 h_T$.

Choose an integer $k \geq 2$ and recall the discontinuous FE space (10). We assume that V_h has two following properties (and we shall prove in Sect. 4 that these properties hold under some additional assumptions on the mesh):

- **Optimal interpolation**: there exists an operator $I_h : H^{k+1}(\Omega) \to V_h$ such that for any $v \in H^{k+1}(\Omega)$

$$\left(\sum_{T \in \mathscr{T}_h} \left(|v - I_h v|_{H^1(T)}^2 + \frac{1}{h_T^2} \|v - I_h v\|_{L^2(T)}^2 + h_T^2 |v - I_h v|_{H^2(T)}^2 \right.\right.$$
$$\left.\left. + h_T \|\nabla v - \nabla I_h v\|_{L^2(\partial T)}^2 + \frac{1}{h_T} \|v - I_h v\|_{L^2(\partial T)}^2 \right) \right)^{\frac{1}{2}}$$

$$\leq C \left(\sum_{T \in \mathscr{T}_h} h_T^{2k} |v|_{H^{k+1}(T)} \right)^{\frac{1}{2}} \tag{20}$$

– **Inverse inequalities**: for any $v_h \in V_h$ and any $T \in \mathscr{T}_h$

$$\|v_h\|_{L^2(\partial T)} \leq \frac{C}{\sqrt{h_T}} \|v_h\|_{L^2(T)} \qquad \|\nabla v_h\|_{L^2(\partial T)} \leq \frac{C}{\sqrt{h_T}} \|\nabla v_h\|_{L^2(T)} \tag{21}$$

and

$$|v_h|_{H^2(T)} \leq \frac{C}{h_T} |v_h|_{H^1(T)} \tag{22}$$

We can now study the well posedness and establish optimal a priori error estimates for the classical SIP method (11).

Lemma 1 *Under the above assumptions on the mesh and on V_h, setting*

$$\begin{aligned} h_E &= 2\left(\tfrac{1}{h_{T_1}} + \tfrac{1}{h_{T_2}}\right)^{-1} && \text{for any } E \in \mathscr{E}_h^i \text{ with } E = \partial T_1 \cap \partial T_2, \\ h_E &= h_T && \text{for any } E \in \mathscr{E}_h^b \text{ with } E = \partial T \cap \partial \Omega, \end{aligned} \tag{23}$$

and choosing γ large enough, $\gamma \geq \gamma_0$, the bilinear form a_h defined by (12) is coercive, i.e.

$$a_h(v_h, v_h) \geq c \,|||\, v_h |||^2, \quad \forall v_h \in V_h \tag{24}$$

with some $c > 0$ and the triple norm defined by

$$|||\, v |||^2 = \sum_{T \in \mathscr{T}_h} \left(|v|^2_{H^1(T)} + \frac{1}{h_T} \|[v]\|^2_{L^2(\partial T)} \right) \tag{25}$$

The constants c, γ_0 depend only on the parameters in the assumptions on the mesh and on V_h, as well as on α, β in (2).

We skip the proof of this well known result. We stress however that our definitions of the length scale h_E and of the triple norm (25) may be slightly different from those available in the literature. In particular, we avoid to use the diameter of a facet E (or any other geometrical information on E) to define h_E. This choice enables us to establish straightforwardly the coercivity of a_h with respect to the triple norm, which does not see the separate mesh facets either (only the whole boundaries of the mesh cells are present there). More elaborate choices for the interior penalty parameters are proposed in [8].

Lemma 1 implies that problem (11) of the SIP method is well posed. Moreover, we have the following error estimate, the proof of which is also skipped (actually, it goes along the same lines as that of our forthcoming Theorem 2).

Theorem 1 *Assume that the solution u to (1) is in $H^{k+1}(\Omega)$. Under the assumptions of Lemma 1, there exists the unique solution u_h to (11) and it satisfies*

$$|u-u_h|_{H^1(\mathscr{T}_h)} \le Ch^k|u|_{H^{k+1}(\Omega)}$$

where $H^1(\mathscr{T}_h)$ is the broken H^1 space on the mesh $\mathscr{T}_h$ and $|\cdot|_{H^1(\mathscr{T}_h)} := \left(\sum_{T\in\mathscr{T}_h} |\cdot|^2_{H^1(T)}\right)^{\frac{1}{2}}$. If, moreover, the elliptic regularity property holds for (1), then

$$\|u-u_h\|_{L^2(\Omega)} \le C|u|_{H^{k+1}(\Omega)}h^{k+1}$$

We turn now to the study of the scSIP method (14)–(17) and start by the following technical lemma.

Lemma 2 *There exists $h_0 > 0$ such that for all $T \in \mathscr{T}_h$ with $h_T \le h_0$ and for all $q_T \in \mathbb{P}_{k-2}(T)$ one can find $u_T \in \mathbb{P}_k(T)$ such that*

$$\int_T q_T(\mathscr{L}u_T) \ge \frac{1}{2}\|q_T\|^2_{L^2(T)} \tag{26}$$

and

$$|u_T|^2_{H^1(T)} + \frac{1}{h_T}\|u_T\|^2_{L^2(\partial T)} \le Ch_T^2\|q_T\|^2_{L^2(T)} \tag{27}$$

The constants h_0 and C depend only on the regularity of the mesh and on α, β and M in (2) and (3). One can put $h_0 = +\infty$ if the coefficient matrix A is constant on T.

Proof Let χ_T be the polynomial of degree 2 vanishing on ∂B_T^{in}, i.e.

$$\chi_T(x) = \left(\sum_{i=1}^d (x_i - x_i^0)^2 - r_T^2\right)$$

where $x^0 = (x_1^0, \ldots, x_d^0)$ is the center of B_T^{in} and r_T is its radius. Set $A_{ij}^0 = A_{ij}(x^0)$ and $\mathscr{L}^0 = -\partial_i A_{ij}^0 \partial_j$. Consider the linear map

$$Q : \mathbb{P}_{k-2}(T) \to \mathbb{P}_{k-2}(T)$$

defined by

$$Q(v) = \mathscr{L}^0(\chi_T v)$$

The kernel of Q is $\{0\}$. Indeed, if $Q(v) = 0$ then $w := \chi_T v$ is the solution to

$$\mathscr{L}^0 w = 0 \quad \text{in } B_T^{in}, \quad w = 0 \quad \text{on } \partial B_T^{in}$$

so that $w = 0$ as a solution to an elliptic problem with vanishing right-hand side and boundary conditions. Since Q is a linear map on the finite dimensional space $\mathbb{P}_{k-2}(T)$, this means that Q is one-to-one.

Take any $q_T \in \mathbb{P}_{k-2}(T)$ and let $u_T = \chi_T v_T$ with $v_T \in \mathbb{P}_{k-2}(T)$ such that $Q(v_T) = q_T$. We have thus constructed $u_T \in \mathbb{P}_k(T)$ such that $\mathscr{L}^0 u_T = q_T$. This immediately proves (26) in the case of an operator $\mathscr{L} = \mathscr{L}^0$ with constant coefficients. Moreover, by scaling,

$$|u_T|_{W^{2,\infty}(B_T)} + \frac{1}{h_T}|u_T|_{W^{1,\infty}(B_T)} + \frac{1}{h_T^2}\|u_T\|_{L^\infty(B_T)} \le \frac{C}{h_T^{d/2}}\|q_T\|_{L^2(B_T^{in})} \tag{28}$$

with a constant C depending only on α, β and the ratio R_T/r_T. Thus,

$$|u_T|_{H^1(T)} \le |T|^{1/2}|u_T|_{W^{1,\infty}(B_T)} \le Ch_T\|q_T\|_{L^2(T)}$$

which proves the estimate in $H^1(T)$ norm in (27). Similarly, $\|u_T\|_{L^2(T)} \le Ch_T^2\|q_T\|_{L^2(T)}$ and the estimate in $L^2(\partial T)$ norm in (27) follows by the trace inverse inequality.

It remains to prove (26) in the case of operator $\mathscr{L}$ with variable coefficients. To this end, we use the estimates in (28) as follows

$$\begin{aligned}
\int_T q_T \mathscr{L} u_T &= \int_T q_T \mathscr{L}^0 u_T + \int_T q_T \partial_i((A_{ij} - A_{ij}^0)\partial_j u_T) \\
&\ge \|q_T\|_{L^2(T)}^2 - \|q_T\|_{L^2(T)}|T|^{1/2}\Big[\max_{x\in T}|A(x) - A_0||u_T|_{W^{2,\infty}(T)} \\
&\quad + \max_{x\in T}|\nabla A(x)||u_T|_{W^{1,\infty}(T)}\Big] \\
&\ge \|q_T\|_{L^2(T)}^2 - \|q_T\|_{L^2(T)}|T|^{1/2}h_T \max_{x\in T}|\nabla A(x)|\frac{C}{h_T^{d/2}}\|q_T\|_{L^2(B_T^{in})} \\
&\ge \|q_T\|_{L^2(T)}^2 - Ch_T\|q_T\|_{L^2(T)}^2 \ge \frac{1}{2}\|q_T\|_{L^2(T)}
\end{aligned}$$

for sufficiently small h_T. □

Corollary 1 *Introduce the bilinear form*

$$b_h(q, v) = \sum_{T\in\mathscr{T}_h} h_T^2 \int_T q\mathscr{L}u$$

and the space

$$M_h = \{v \in L^2(\Omega) : v|_T \in \mathbb{P}_{k-2}(T), \forall T \in \mathscr{T}_h\}$$

Equip the space V_h with the triple norm (25) and the space M_h with

$$\|q\|_h = \left(\sum_{T \in \mathscr{T}_h} h_T^2 \|q\|_{L^2(T)}^2 \right)^{1/2}$$

The bilinear form b_h satisfies the inf-sup condition

$$\inf_{q_h \in M_h} \sup_{v_h \in V_h} \frac{b_h(q_h, v_h)}{\|q_h\|_h \; |||v_h|||} \geq \delta \tag{29}$$

with a mesh-independent constant $\delta > 0$. Moreover, b_h is continuous on $M_h \times V_h$ with a mesh-independent continuity bound.

Proof Take any $q_h \in M_h$, denote $q_T = q_h|_T$, construct u_T as in Lemma 2 and introduce $u_h \in V_h$ by $u_h|_T = u_T$ on all $T \in \mathscr{T}_h$. This yields using (26) and (27),

$$\frac{b_h(q_h, u_h)}{|||u_h|||} \geq \frac{\sum_{T \in \mathscr{T}_h} \frac{h_T^2}{2} \|q_h\|_{L^2(T)}^2}{\left(\sum_{T \in \mathscr{T}_h} C h_T^2 \|q_h\|_{L^2(T)}^2 \right)^{1/2}} = \frac{1}{2\sqrt{C}} \|q_h\|_h$$

which is equivalent to (29) with $\delta = \frac{1}{2\sqrt{C}}$. Finally, the continuity of b_h is easily seen from the inverse inequality (22). □

Lemma 2 implies that operator $\pi_{T,k-2}\mathscr{L}$ appearing in (14) is surjective from $\mathbb{P}^k(T)$ to $\mathbb{P}^{k-2}(T)$ so that (14) has indeed a solution at least on sufficiently refined meshes. The existence of a solution to (17) follows from the coercivity of a_h. Thus, scheme (14)–(18) produces some $u_h \in V_h$. In order to establish the error estimates for this u_h, we reinterpret its definition as a saddle point problem.

Lemma 3 *The problem of finding $u_h \in V_h$ and $p_h \in M_h$ such that*

$$a_h(u_h, v_h) + b_h(p_h, v_h) = L_h(v_h), \qquad \forall v_h \in V_h \tag{30}$$

$$b_h(q_h, u_h) = \sum_{T \in \mathscr{T}_h} h_T^2 \int_T f q_h, \qquad \forall q_h \in M_h \tag{31}$$

has a unique solution. Moreover, u_h given by (30)–(31) *coincides with u_h given by* (14)–(18).[3] *This implies that u_h produced by the scheme* (14)–(18) *is unique.*

Proof The existence and uniqueness of the solution to (30)–(31) follows from the standard theory of saddle point problems, cf. for example Corollary 4.1 from [18], thanks to the coercivity of a_h (Lemma 1) and to the inf-sup property on b_h (Corollary 1).

[3] More precisely, all solutions u_h of (14)–(18) may be accompanied by $p_h \in M_h$ so that the resulting couples (u_h, p_h) also solve (30)–(31). Since the solution to (30)–(31) is unique, the inverse statement is also true: u_h given by (30)–(31) is also a solution to (14)–(18).

In order to explore its relation with $u_h = u_h^{loc} + u_h'$ from (14)–(18), we note

$$b_h(q_h, u_h^{loc}) = \sum_{T \in \mathcal{T}_h} h_T^2 \int_T f q_h, \quad \forall q_h \in M_h$$

and $b_h(q_h, u_h') = 0$ for all $q_h \in M_h$ since $u_h' \in V_h'$. We obtain thus

$$b_h(q_h, u_h^{loc} + u_h') = \sum_{T \in \mathcal{T}_h} h_T^2 \int_T f q_h, \quad \forall q_h \in M_h \tag{32}$$

Equation (17) can be rewritten as

$$L_h(v_h') - a_h(u_h^{loc} + u_h', v_h') = 0, \quad \forall v_h' \in V_h'$$

This, together with the fact that V_h' is precisely the kernel of the bilinear form b_h, i.e. $V_h' = \{v_h \in V_h : b(q_h, v_h) = 0, \ \forall q_h \in M_h\}$, means that there exists $\tilde{p}_h \in M_h$ such that

$$b_h(\tilde{p}_h, v_h) = L_h(v_h) - a_h(u_h^{loc} + u_h', v_h), \quad \forall v_h \in V_h$$

cf. Lemma 4.1 from [18]. The last equation can be rewritten as

$$a_h(u_h^{loc} + u_h', v_h) + b_h(\tilde{p}_h, v_h) = L_h(v_h), \quad \forall v_h \in V_h \tag{33}$$

Comparing (33)–(32) on one hand with (30)–(31) on the other hand, we identify u_h with $u_h^{loc} + u_h'$ and p_h with $\tilde{p}_h$ □

Theorem 2 *Assume that the solution u to (1) is in $H^{k+1}(\Omega)$. Under the assumptions of Lemma 1 and h sufficiently small, the scSIP method (14)–(18) produces the unique solution $u_h \in V_h$, which satisfies*

$$|u - u_h|_{H^1(\mathcal{T}_h)} \le C h^k |u|_{H^{k+1}(\Omega)} \tag{34}$$

If, moreover, the elliptic regularity property holds for (1), then

$$\|u - u_h\|_{L^2(\Omega)} \le C |u|_{H^{k+1}(\Omega)} h^{k+1} \tag{35}$$

Proof We shall use the saddle point reformulation (30)–(31). This discretization is consistent. Indeed setting $p = 0$ we have

$$\begin{aligned} a_h(u, v_h) + b_h(p, v_h) &= L_h(v_h), \quad \forall v_h \in V_h \\ b_h(q_h, u) &= \sum_{T \in \mathcal{T}_h} h_T^2 \int_T f q_h, \quad \forall q_h \in M_h \end{aligned}$$

Thus, by the standard approximation theory for saddle point problems, cf. for example Proposition 2.36 from [17], recalling the coercivity of a_h (Lemma 1) and the inf-sup property on b_h (Corollary 1), we get

$$\begin{aligned}
&|||\, u_h - I_h u \,||| + \|p_h\|_h \\
&\le C \sup_{\substack{(v_h, q_h) \in V_h \times M_h \\ |||v_h\,||| + \|q_h\|_h = 1}} (a_h(u_h - I_h u, v_h) + b_h(p_h, v_h) + b_h(q_h, u_h - I_h u)) \\
&= C \sup_{\substack{(v_h, q_h) \in V_h \times M_h \\ |||v_h\,||| + \|q_h\|_h = 1}} (a_h(u - I_h u, v_h) + b_h(q_h, u - I_h u)) \\
&\le C \left(|||u - I_h u\,|||_a^2 + \sum_{T \in \mathscr{T}_h} h_T^2 \|\mathscr{L}(u - I_h u)\|_{L^2(T)}^2 \right)^{\frac{1}{2}}
\end{aligned}$$

with the augmented triple norm $|||\cdot|||_a$ defined by

$$|||v|||_a^2 := |||v\,|||^2 + \sum_{E \in \mathscr{E}_h} h_E \|\{A\nabla v \cdot n\}\|_{L^2(E)}^2$$

Applying the interpolation estimates (20) and the triangle inequality gives

$$|||\, u - I_h u \,|||_a + \|p_h\|_h \le C h^k |u|_{H^{k+1}} \tag{36}$$

This implies in particular (34).

To prove the L^2 error estimate, we consider the auxiliary problem for $z \in H^2(\Omega)$

$$\mathscr{L} z = u - u_h \quad \text{in } \Omega, \quad z = 0 \quad \text{on } \partial\Omega$$

Then, for all $v \in H^1(\mathscr{T}_h)$ and $q \in L^2(\Omega)$,

$$a_h(v, z) + b_h(q, z) = \int_\Omega (u - u_h) v + \sum_{T \in \mathscr{T}_h} h_T^2 \int_T (u - u_h) q$$

Setting $v = u - u_h$ and $q = p - p_h$ (with $p = 0$) and using Galerkin orthogonality yields

$$\begin{aligned}
&\|u - u_h\|_{L^2(\Omega)}^2 + \sum_{T \in \mathscr{T}_h} h_T^2 \int_T (u - u_h)(p - p_h) = a_h(u - u_h, z) + b_h(p - p_h, z) \\
&= a_h(u - u_h, z - z_h) + b_h(p - p_h, z - z_h)
\end{aligned}$$

for any $z_h \in V_h$. Thus, taking $z_h = I_h z$ and applying the interpolation estimates,

$$
\begin{aligned}
\|u-u_h\|^2_{L^2(\Omega)} \le & \; |a_h(u-u_h, z-z_h)| + h\|p-p_h\|_h \\
& \times \left(\sum_{T\in\mathscr{T}_h} \|\mathscr{L}z - \mathscr{L}z_h\|^2_{L^2(T)} \right)^{1/2} \\
& + h\|p-p_h\|_h \|u-u_h\|_{L^2(\Omega)} \\
\le & \; Ch \;|||\; u - I_h u \;|||_a \; |z|_{H^2(\Omega)} \\
& + Ch\|p_h\|_h (|z|_{H^2(\Omega)} + \|u-u_h\|_{L^2(\Omega)})
\end{aligned}
$$

Recalling $|z|_{H^2(\Omega)} \le C\|u-u_h\|_{L^2(\Omega)}$ and (36) yields (35). □

4 An example of assumptions on the mesh that guarantee the interpolation and inverse estimates

In this section, we adopt the following assumptions on the mesh.

M1: $\mathscr{T}_h$ is shape regular in the sense (19) with a parameter $\rho_1 > 1$.

M2: $\mathscr{T}_h$ is locally quasi-uniform in the following sense: for any two mesh cells $T, T' \in \mathscr{T}_h$ such that $B_{T'} \cap B_T \neq \varnothing$ there holds

$$\frac{1}{\rho_2} h_{T'} \le h_T \le \rho_2 h_{T'}$$

with a parameter $\rho_2 > 1$.

M3: The cell boundaries are not too wiggly: for all $T \in \mathscr{T}_h$

$$|\partial T| \le \rho_3 h_T^{d-1}$$

with a parameter $\rho_3 > 0$.

We shall show that these assumptions allow us to construct an interpolation operator I_h to the discontinuous finite element space (10) for $k \ge 2$ and to prove the interpolation error estimate (20) and the inverse estimates (21)–(22).

First of all, the assumptions that the mesh is shape regular and locally quasi-uniform entail the following

Lemma 4 *Define, for any $x \in \mathbb{R}^d$,*

$$N_{ball}(x) = \#\{T \in \mathscr{T}_h : x \in B_T\}$$

with # standing for the "number of". Under assumptions M1 and M2, there holds

$$N_{ball}(x) \le N_{int}, \quad \forall x \in \mathbb{R}^d$$

with a constant N_{int} depending only on ρ_1 and ρ_2.

Proof Take any $x \in \mathbb{R}^d$ with $N_{ball}(x) > 0$ (otherwise, for x with $N_{ball}(x) = 0$, there is nothing to prove). We now choose arbitrarily $T' \in \mathscr{T}_h$ such that $x \in B_{T'}$, set $h_x = h_{T'}$, and then consider all the mesh cells T such that $x \in B_T$. By Assumption M2, $h_T \le \rho_2 h_x$ for any such T. Hence, by Assumption M1, $R_T \le \rho_1\rho_2 h_x$, so that T is inside the ball B_x of radius $2\rho_1\rho_2 h_x$ centered at x. Recall that T contains an inscribed ball of radius $r_T \ge \frac{R_T}{\rho_1} \ge \frac{h_T}{2\rho_1} \ge \frac{h_x}{2\rho_1\rho_2}$. If there are several such cells T, then their respective inscribed balls B_T^{in} do not intersect each other and they are all inside B_x. Thus, their number satisfies the bound

$$N_{ball}(x) \le \frac{|B_x|}{\min\limits_{T\in\mathscr{T}_h: x\in B_T} |B_T^{in}|} \le \frac{(2\rho_1\rho_2 h_x)^d}{\left(\frac{h_x}{2\rho_1\rho_2}\right)^d} = (2\rho_1\rho_2)^{2d}$$

as announced. □

Recall that V_h is the discontinuous FE space on $\mathscr{T}_h$ of degree $k \ge 2$, cf. (10).

Lemma 5 (Local interpolation estimate) *Take any $T \in \mathscr{T}_h$. Let π_h denote the $L^2(B_T)$-orthogonal projection to the space of polynomials, i.e. given $v \in L^2(B_T)$, $v_h = \pi_h v$ is a polynomial of degree $\le k$ such that*

$$\int_{B_T} v_h \varphi_h = \int_{B_T} v\varphi_h \quad \forall \varphi_h \in \mathbb{P}^k(T)$$

Under Assumptions M1 and M3, we have then for any $v \in H^{k+1}(B_T)$

$$|v - v_h|_{H^1(T)} + \frac{1}{h_T}\|v - v_h\|_{L^2(T)} + h_T|v - v_h|_{H^2(T)} + \sqrt{h_T}\|\nabla(v - v_h)\|_{L^2(\partial T)} + \frac{1}{\sqrt{h_T}}\|v - v_h\|_{L^2(\partial T)} \le Ch_T^k|v|_{H^{k+1}(B_T)}$$

with a constant $C > 0$ depending only on ρ_1 and ρ_3.

Proof Since $H^{k+1}(B_T)$ is embedded into $L^\infty(B_T)$, Deny-Lions lemma together with a scaling argument (cf. Theorem 15.3 from [9]) entail

$$\|v - v_h\|_{L^\infty(B_T)} \le Ch_T^{k+1-d/2}|v|_{H^{k+1}(B_T)}$$

Hence,

$$\|v - v_h\|_{L^2(T)} \le |T|^{\frac{1}{2}}\|v - v_h\|_{L^\infty(B_T)} \le Ch_T^{k+1}|v|_{H^{k+1}(B_T)}$$

and, in view of the hypothesis $|\partial T| \le \rho_3 h_T^{d-1}$,

$$\|v - v_h\|_{L^2(\partial T)} \le (\rho_3 h_T^{d-1})^{\frac{1}{2}}\|v - v_h\|_{L^\infty(B_T)} \le Ch_T^{k+1/2}|v|_{H^{k+1}(B_T)}$$

The estimates for $|v - v_h|_{H^1(T)}$ and $\|\nabla v - \nabla v_h\|_{L^2(\partial T)}$ are proven in the same way starting from

$$\|\nabla v - \nabla v_h\|_{L^\infty(B_T)} \le C h_T^{k-d/2} |v|_{H^{k+1}(B_T)}$$

This is valid since $H^{k+1}(B_T)$ is embedded into $W^{1,\infty}(B_T)$ for $k \ge 2$.

Finally, the estimate for $|v - v_h|_{H^2(T)}$ holds thanks to the embedding of $H^{k+1}(B_T)$ into $H^2(B_T)$ ($k \ge 1$). □

Lemma 6 (Global interpolation estimate) *Let $I_h : H^{k+1}(\Omega) \to V_h$ denote the operator obtained by first extending a function $v \in H^{k+1}(\Omega)$ by a function $\tilde{v} \in H^{k+1}(\mathbb{R}^d)$ and then applying the local operator π_h from Lemma 5 to $\tilde{v}$ on every $T \in \mathscr{T}_h$, i.e. $I_h v|_T := (\pi_h \tilde{v})|_T$ on any $T \in \mathscr{T}_h$. Then, under Assumptions M1–M3, (20) holds for any $v \in H^{k+1}(\Omega)$*

Proof First, extension theorem for Sobolev spaces [1] insure that there exists $\tilde{v} \in H^{k+1}(\mathbb{R}^d)$ such that

$$\tilde{v} = v \quad \text{on } \Omega \quad \text{and} \quad \|\tilde{v}\|_{H^{k+1}(\mathbb{R}^d)} \le C \|\tilde{v}\|_{H^{k+1}(\Omega)}$$

To prove (20), we sum the local interpolation estimates of Lemma 5 over all the mesh cells and then use Assumption M2 and Lemma 4:

$$\begin{aligned}
&\sum_{T \in \mathscr{T}_h} \Bigg(|v - v_h|^2_{H^1(T)} + \frac{1}{h_T^2} \|v - v_h\|^2_{L^2(T)} + h_T^2 |v - v_h|^2_{H^2(T)} \\
&\qquad + h_T \|\nabla v - \nabla v_h\|^2_{L^2(\partial T)} + \frac{1}{h_T} \|v - v_h\|^2_{L^2(\partial T)} \Bigg) \\
&\le C \sum_{T \in \mathscr{T}_h} h_T^{2k} \int_{B_T} |\nabla^{k+1} v|^2 dx \\
&\le C \int_\Omega \left(\max_{T \in \mathscr{T}_h : x \in B_T} h_T \right)^{2k} N_{ball}(x) |\nabla^{k+1} v|^2 dx \\
&\le C N_{int} \rho_2^{2k} \sum_{T' \in \mathscr{T}_h} h_{T'}^{2k} |v|^2_{H^{k+1}(T)}
\end{aligned}$$

□

Lemma 7 (Inverse inequalities) *Under assumptions M1 and M3, (21) and (22) hold for any $v_h \in V_h$ and any $T \in \mathscr{T}_h$.*

Proof Both bounds in (21) follow immediately from the following one: for any polynomial q_h of degree $\le l$ one has

$$\|q_h\|_{L^\infty(T)} \le \frac{C}{h_T^{d/2}} \|q_h\|_{L^2(T)}$$

with $C > 0$ depending only on ρ_1 and l. This follows in turn from

$$\|q_h\|_{L^\infty(B_T)} \le \frac{C}{h_T^{d/2}} \|q_h\|_{L^2(B_T^{in})}$$

where B_T^{in} is the largest ball inscribed in T. Scaling the ball B_T to a ball of radius 1 B_1 and considering all the possible positions of the inscribed ball, the last inequality can be rewritten as

$$\|q_h\|_{L^\infty(B_1)} \le C \min_{B^{in}\subset B_1,\, B^{in} \text{ a ball of radius } \ge \rho_1^{-1}} \|q_h\|_{L^2(B_T^{in})}$$

This is valid for any polynomial of degree l by equivalence of norms.

The remaining inverse inequality (22) can be proven similarly:

$$\begin{aligned} |v_h|_{H^2(T)} &\le C h_T^{d/2} |v_h|_{W^{2,\infty}(B_T)} \le C h_T^{d/2-1} |v_h|_{W^{1,\infty}(B_T)} \\ &\le \frac{C}{h_T} |v_h|_{H^1(B_T^{in})} \le \frac{C}{h_T} |v_h|_{H^1(T)} \end{aligned}$$

□

5 Implementation and numerical results

We shall illustrate the convergence of SIP and scSIP methods on polygonal meshes obtained by agglomerating the cells of a background triangular mesh. Both the mesh construction and the following calculations are done in FreeFEM++ [20]. An example of such a mesh is given in Fig. 1. To construct it, we take a positive integer n ($n = 4$ in the figure), let FreeFEM++ construct a Delaunay triangulation of $\Omega = (0, 1)^2$ with $4n$ boundary nodes on each side of the square, and finally agglomerate the triangles of this mesh into $n \times n$ cells as follows. We start by attributing the triangle containing the point

$$O_{i+jn} = \left(\frac{i-1/2}{n}, \frac{j-1/2}{n}\right), \quad i, j = 1, \ldots, n \tag{37}$$

to the cell number $i + jn$. Then, iteratively, we run over all the cells and attach yet unattributed triangles neighboring a triangle from a cell to the same cell, until all the triangles are attributed.

Some details of our implementations are given below, followed by the numerical results on two test cases.

5.1 Implementation of SIP and scSIP methods

Let us enumerate the mesh cells as $\{T_1, \ldots, T_{N_e}\}$ and introduce a basis $\{\phi_i^{(l)}\}_{i=1,\ldots,N_k}$ of $\mathbb{P}_k(T_l)$ on every cell T_l. Here and below, N_e is the number of cells in $\mathscr{T}_h$ and N_k denotes the dimension of $\mathbb{P}_k$. In our implementation, we form the basis out of monomials shifted to the "center" $O_l = (O_{l,x}, O_{l,y})$ of the cell T_l, cf. (37). I.e.

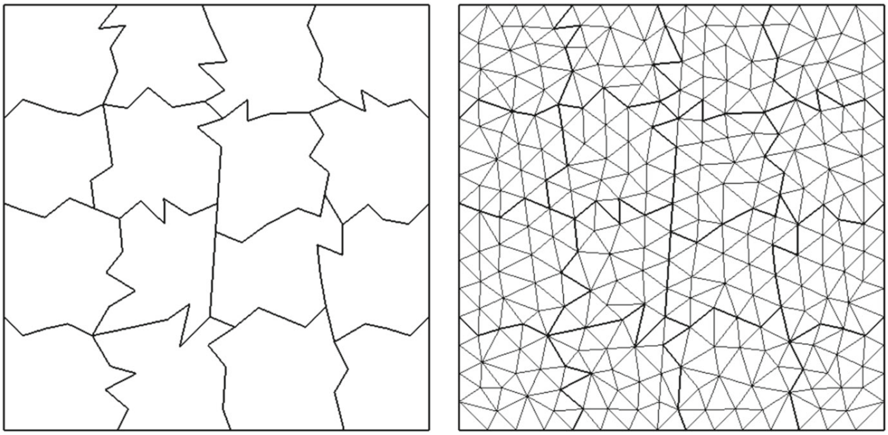

Fig. 1 On the left, a polygonal mesh consisting of 4×4 cells, which are obtained by the agglomeration of the triangles of a finer mesh seen on the right

$$\phi_i^{(l)}(x, y) = \phi_{i_1 i_2}^{(l)}(x, y) = (x - O_{l,x})^{i_1}(y - O_{l,y})^{i_2}$$

regrouping the multi-indexes (i_1, i_2), $0 \le i_1 \le k$, $0 \le i_2 \le k - i_1$ into a single index i ranging from 1 to N_k. We form then the matrices $\mathbf{A}^{(lm)}$ for every pair of cells T_l and T_m sharing some parts of their boundaries. These matrices of size $N_k \times N_k$ represent the bilinear form a_h in our bases and have the following entries

$$A_{ij}^{(lm)} = a_h(\phi_i^{(l)}, \phi_j^{(m)})$$

We also compute the right-hand side vectors $\overrightarrow{F}^{(l)} \in \mathbb{R}^{N_k}$ with $F_i^{(l)} = L_h(\phi_i^{(l)})$ on every cell T_l, put all $\overrightarrow{F}^{(l)}$ into a single vector $\overrightarrow{F}$ of size $N_{DOF} = N_e N_k$, put the matrices $\mathbf{A}^{(lm)}$ into the block matrix $\mathbf{A}$ of size $N_{DOF} \times N_{DOF}$, and finally find $\overrightarrow{U} \in \mathbb{R}^{N_{DOF}}$ as solution to

$$\mathbf{A}\overrightarrow{U} = \overrightarrow{F}$$

The vector $\overrightarrow{U}$ represents the numerical solution by the SIP method (11) in the following sense: decomposing $\overrightarrow{U}$ into the cell-by-cell components $\overrightarrow{U}^{(l)} = \{U_i^{(l)}\} \in \mathbb{R}^{N_k}$, u_h in (11) is given on each cell T_l by $u_h = \sum_{i=1}^{N_k} U_i^{(l)} \phi_i^{(l)}$.

Turning to the scSIP method, we introduce moreover a basis of $\mathbb{P}_{k-2}(T_l)$, $\{\psi_i^{(l)}\}_{i=1,\ldots,N_{k-2}}$, and form the matrices $\mathbf{B}^{(l)}$ of size $N_{k-2} \times N_k$ on every cell T_l with the entries

$$B_{ij}^{(l)} = \int_{T_l} \psi_i^{(l)} \mathscr{L} \phi_j^{(l)} = \int_{T_l} \psi_i^{(l)} \cdot A\nabla\phi_j^{(l)} - \int_{\partial T_l} \psi_i^{(l)} n \cdot A\nabla\phi_j^{(l)}$$

These matrices will serve to compute the local contributions in (14) as well as to construct a basis of the space V_h' in (15). As mentioned earlier, the solution to (14) is not unique and one can propose several ways to compute a solution in practice. In our implementation, we have opted for a solution to (14) solving the following saddle-point problem on every cell T_l

$$\begin{aligned} \overrightarrow{u}^{(l)} + (\mathbf{B}^{(l)})^T \overrightarrow{p}^{(l)} &= 0 \\ \mathbf{B}^{(l)} \overrightarrow{u}^{(l)} &= \overrightarrow{F_\psi}^{(l)} \end{aligned} \tag{38}$$

with $\overrightarrow{F_\psi}^{(l)} = \{F_{\psi,i}^{(l)}\} \in \mathbb{R}^{N_{k-2}}$, $F_{\psi,i}^{(l)} = \int_{T_l} f \psi_i^{(l)}$. The unknowns here are $\overrightarrow{u}^{(l)} \in \mathbb{R}^{N_k}$ and $\overrightarrow{p}^{(l)} \in \mathbb{R}^{N_{k-2}}$ with $\overrightarrow{u}^{(l)}$ representing u_h^{loc} on T_l in the basis $\{\phi_i^{(l)}\}$. The saddle-point problem above is well posed thanks to Lemma 2.

A basis for V_h' from (15)–(16) can be constructed on every cell T_l using a saddle-point problem similar to (38). Indeed, V_h' on T_l is the kernel of $\mathbf{B}^{(l)}$. It is thus given by the span of vectors $\{\overrightarrow{u}^{(l,1)}, \ldots, \overrightarrow{u}^{(l,N_k)}\} \subset \mathbb{R}^{N_k}$ with $\overrightarrow{u}^{(l,s)}$ defined by

$$\begin{aligned} \overrightarrow{u}^{(l,s)} + (\mathbf{B}^{(l)})^T \overrightarrow{p}^{(l,s)} &= \overrightarrow{e}^{(s)} \\ \mathbf{B}^{(l)} \overrightarrow{u}^{(l,s)} &= 0 \end{aligned} \tag{39}$$

where $\{\overrightarrow{e}^{(1)}, \ldots, \overrightarrow{e}^{(N_k)}\}$ is the canonical basis of $\mathbb{R}^{N_k}$. In practice, we solve the problem above successively for $s = 1, 2, \ldots$ and apply the Gram-Schmidt procedure to ortho-normalize the vectors $\overrightarrow{u}^{(l,s)}$ getting rid of the vectors which turn out to be linearly dependent from the preceding ones. This provides us with a basis for $V_h'|_{T_l}$ consisting of $N_k' = N_k - N_{k-2}$ vectors (actually, the Gram-Schmidt process can be stopped once N_k' ortho-normal vectors have been found).

Remark 1 In the case when the coefficient matrix A is constant (and thus does not change from one mesh cell to another), the restriction on the functions in V_h', i.e. $\int_T q_h \mathscr{L} v_h' = 0$ for all $q_h \in \mathbb{P}_{k-2}(T)$, implies in fact $\mathscr{L} v_h' = 0$ on every cell T. This is independent from the shape of T so that the structure of V_h' is the same on all the cells, and one can keep the same basis for V_h' everywhere.

For example, in the case of Poisson equation ($\mathscr{L} = -\Delta$) in 2D, v_h supported on a cell T is in V_h' if and only if $\Delta v_h = 0$. Expanding v_h in the basis of monomials $v_h = \sum_{i_1,i_2} v_{i_1 i_2} \phi_{i_1 i_2}^{(l)}$ this gives rise to the equations

$$(i_1 + 2)(i_1 + 1) v_{i_1+2,i_2} + (i_2 + 2)(i_2 + 1) v_{i_1,i_2+2} = 0$$

for all non-negative (i_1, i_2). These equations can be easily solved to provide a basis for V_h' on all the cells.

In our implementation, to keep things simple and the code suitable for both cases of either constant A or varying A, we have used another strategy: if A is constant, we perform the Gram-Schmidt ortho-normalization on the solutions to (39) on the mesh cell number 1 only. We keep then the same basis (as expressed by the expansion coefficients in $\{\phi_i^{(l)}\}_{i=1,\ldots,N_k}$) on all the other cells T_l, $l \geq 2$.

Having constructed the basis for V_h', it remains to solve the global problem (17). We introduce to this end on every cell T_l the matrices $\mathbf{M}^{(l)}$ of size $N_k \times N_k'$ putting together the vectors representing the basis for V_h' on T_l. We form then the reduced matrices $\mathbf{A}'^{(lm)}$ of size $N_k' \times N_k'$ and the reduced right-hand side vectors $\overrightarrow{F}'^{(l)}$ out of full $\mathbf{A}^{(lm)}$ and $\overrightarrow{F}^{(l)}$ (already introduced in the description of the SIP method implementation) by, cf. (17),

$$\mathbf{A}'^{(lm)} = (\mathbf{M}^{(l)})^T \mathbf{A}^{(lm)} \mathbf{M}^{(m)} \text{ and } \overrightarrow{F}'^{(l)} = (\mathbf{M}^{(l)})^T \left(\overrightarrow{F}^{(l)} - \sum_m \mathbf{A}^{(lm)} \overrightarrow{u}^{(m)} \right)$$

with $\overrightarrow{u}^{(m)}$ representing u_h^{loc} on T_m and computed by (38). Putting the matrices $\mathbf{A}'^{(lm)}$ into the block matrix $\mathbf{A}'$ of size $N_{DOF}' \times N_{DOF}'$ with $N_{DOF}' = N_e N_k'$ and the vectors $\overrightarrow{F}'^{(l)}$ into a single vector $\overrightarrow{F}'$, we compute $\overrightarrow{U}' \in \mathbb{R}^{N_{DOF}'}$ as solution to

$$\mathbf{A}' \overrightarrow{U}' = \overrightarrow{F}'$$

The vector $\overrightarrow{U}'$ represents the solution to (17). The solution u_h by the scSIP method is finally reconstructed as follows: decomposing $\overrightarrow{U}'$ into the cell-by-cell components $\overrightarrow{U}'^{(l)} = \{U_i'^{(l)}\} \in \mathbb{R}^{N_k'}$, we recall $\overrightarrow{u}^{(l)}$ computed on each cell T_l by (38), introduce $\overrightarrow{\widetilde{U}}^{(l)} = \{\widetilde{U}_i^{(l)}\} \in \mathbb{R}^{N_k}$ as $\overrightarrow{\widetilde{U}}^{(l)} = \overrightarrow{u}^{(l)} + \mathbf{M}^{(l)} \overrightarrow{U}'^{(l)}$, and set u_h on T_l as $u_h = \sum_{i=1}^{N_k} \widetilde{U}_i^{(l)} \phi_i^{(l)}$.

5.2 The first test case: Poisson equation

We have considered the Poisson equation, i.e. (1) with $A = I$, on $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions $g = 0$ and the exact solution $u = \sin(\pi x) \sin(\pi y)$. We have applied SIP method (11) and scSIP method (14)–(17) to this problem on the agglomerated meshes as described in the preamble of this section. The results are presented in Fig. 2. In a slight deviation from the general notations, we set here the mesh-size as $h = 1/n$ on the $n \times n$ mesh, and $h_E = h$ on all the edges in (12). Three choices for the polynomial space degree k were investigated, namely $k = 2, 3, 4$

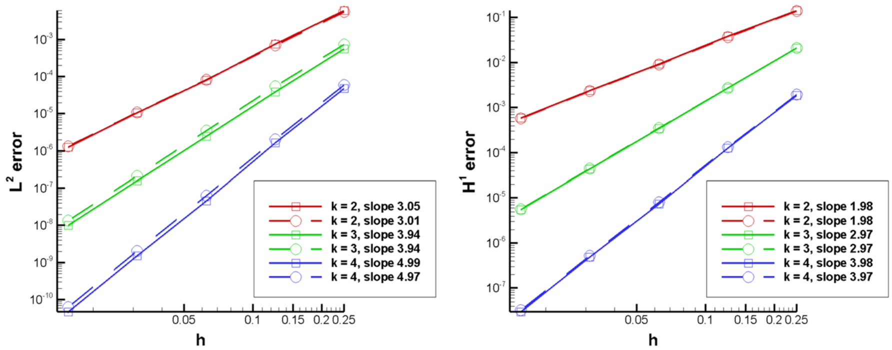

Fig. 2 The test case with Poisson equation: the error in L^2 norm and H^1 semi-norm versus mesh-size h. The solid lines with squares represent the SIP method. The dashed lines with circles represent the scSIP method

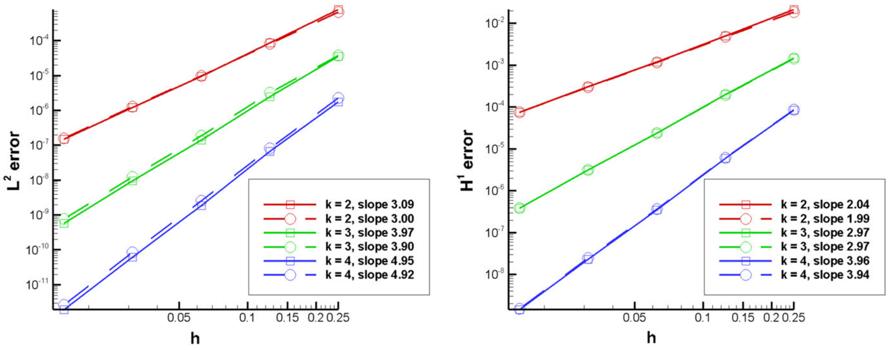

Fig. 3 Test case with the non constant coefficient matrix A: the error in L^2 norm and H^1 semi-norm versus mesh-size h. The solid lines with squares represent the SIP method. The dashed lines with circles represent the scSIP method

and the penalty parameter γ in (12) was set to $2k(k+1)$ (by a loose extrapolation to the polygonal meshes of the bound on the constant in the inverse inequality on a triangle in [23]). The numerical results confirm the theoretically expected order of convergence in both L^2 norm and H^1 semi-norm. They also demonstrate that the approximation produced by SIP and scSIP methods are very close to each other.

5.3 The second test case: non-constant coefficients *A*

We now consider problem (1) with a non-constant coefficient matrix

$$A = \begin{pmatrix} 1+x & xy \\ xy & 1+y \end{pmatrix}$$

set again on $\Omega = (0,1)^2$. The right-hand side f and non-homogeneous Dirichlet boundary conditions g are chosen so that the exact solution is given by $u = e^{xy}$. The results are presented in Fig. 3 using the same meshes and parameters h, h_E, and γ as in the first test case. We arrive at the same conclusions about the convergence of SIP and scSIP methods as before.

Acknowledgements I am grateful to Simon Lemaire for interesting discussions about HHO and msHHO methods, which were the starting point of conceiving the present article. I also thank the anonymous reviewers for very careful reading of the manuscript and for helpful remarks and suggestions.

References

1. Adams, R.A.: Sobolev Spaces. Academic Press [A Subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London (1975). Pure and Applied Mathematics, vol. 65
2. Antonietti, P.F., Cangiani, A., Collis, J., Dong, Z., Georgoulis, E.H., Giani, S., Houston, P.: Review of discontinuous Galerkin finite element methods for partial differential equations on complicated domains. In [4] (2016)

3. Arnold, D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19**(4), 742–760 (1982). https://doi.org/10.1137/0719052
4. Barrenechea, G.R., Brezzi, F., Cangiani, A., Georgoulis, E.H. (eds.): Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations. Lecture Notes in Computational Science and Engineering, vol. 114. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41640-3. Selected papers from the 101st LMS-EPSRC Symposium held at Durham University, Durham, July 8–16, 2014
5. Brezzi, F., Marini, L.D.: Virtual element methods for plate bending problems. Comput. Methods Appl. Mech. Eng. **253**, 455–462 (2013). https://doi.org/10.1016/j.cma.2012.09.012
6. Beirão da Veiga, L., Brezzi, F., Marini, L.D.: Virtual elements for linear elasticity problems. SIAM J. Numer. Anal. **51**(2), 794–812 (2013). https://doi.org/10.1137/120874746
7. Beirão da Veiga, L., Brezzi, F., Marini, L.D., Russo, A.: Virtual element implementation for general elliptic equations. In [4] (2016)
8. Cangiani, A., Georgoulis, E.H., Houston, P.: hp-version discontinuous Galerkin methods on polygonal and polyhedral meshes. Math. Models Methods Appl. Sci. **24**(10), 2009–2041 (2014). https://doi.org/10.1142/S0218202514500146
9. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Handbook of Numerical Analysis, vol. II, Handbook of Numerical Analysis, II. North-Holland, Amsterdam, pp. 17–351 (1991)
10. Cockburn, B.: Static condensation, hybridization, and the devising of the HDG methods. In [4] (2016)
11. Cockburn, B., Di Pietro, D.A., Ern, A.: Bridging the hybrid high-order and hybridizable discontinuous Galerkin methods. ESAIM Math. Model. Numer. Anal. **50**(3), 635–650 (2016). https://doi.org/10.1051/m2an/2015051
12. Cockburn, B., Gopalakrishnan, J., Lazarov, R.: Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal. **47**(2), 1319–1365 (2009). https://doi.org/10.1137/070706616
13. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 69. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-22980-0
14. Di Pietro, D.A., Ern, A.: A hybrid high-order locking-free method for linear elasticity on general meshes. Comput. Methods Appl. Mech. Eng. **283**, 1–21 (2015). https://doi.org/10.1016/j.cma.2014.09.009
15. Di Pietro, D.A., Ern, A., Lemaire, S.: An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. Comput. Methods Appl. Math. **14**(4), 461–472 (2014). https://doi.org/10.1515/cmam-2014-0018
16. Di Pietro, D.A., Ern, A., Lemaire, S.: A review of hybrid high-order methods: formulations, computational aspects, comparison with other methods. In [4] (2016)
17. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
18. Girault, V., Raviart, P.A.: Finite Element Methods for Navier–Stokes Equations. Springer Series in Computational Mathematics, vol. 5. Springer, Berlin (1986). https://doi.org/10.1007/978-3-642-61623-5. Theory and algorithms
19. Guyan, R.J.: Reduction of stiffness and mass matrices. AIAA J. **3**(2), 380 (1965)
20. Hecht, F.: New development in freefem++. J. Numer. Math. **20**(3–4), 251–265 (2012)
21. Riviere, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation. SIAM, Philadelphia (2008)
22. Wang, J., Ye, X.: A weak Galerkin finite element method for second-order elliptic problems. J. Comput. Appl. Math. **241**, 103–115 (2013). https://doi.org/10.1016/j.cam.2012.10.003
23. Warburton, T., Hesthaven, J.S.: On the constants in hp-finite element trace inverse inequalities. Comput. Methods Appl. Mech. Eng. **192**(25), 2765–2773 (2003). https://doi.org/10.1016/S0045-7825(03)00294-9
24. Wheeler, M.F.: An elliptic collocation-finite element method with interior penalties. SIAM J. Numer. Anal. **15**(1), 152–161 (1978)