Finite element method with local damage of the mesh

Michel Duprez, Vanessa Lleras, Alexei Lozinski

Finite element method with local damage of the mesh *

Michel Duprez†, Vanessa Lleras‡ and Alexei Lozinski§

March 14, 2019

Abstract

We consider the finite element method on locally damaged meshes allowing for some distorted cells which are isolated from one another. In the case of the Poisson equation and piecewise linear Lagrange finite elements, we show that the usual *a priori* error estimates remain valid on such meshes. We also propose an alternative finite element scheme which is optimally convergent and, moreover, well conditioned, *i.e.* the conditioning number of the associated finite element matrix is of the same order as that of a standard finite element method on a regular mesh of comparable size.

1 Introduction

We are interested in the finite element method on meshes containing some isolated degenerate cells. The meshes of this type can be encountered in bio-mechanical applications, where the objects with very complicated geometry (as a human face) should be meshed, and the mesh generators or mesh morphing techniques are not always able to satisfy the usual regularity constraints (see *e.g.* [8, p.3]). Our work is a preliminary study in which we propose a suitable finite element approximation in such situations without requiring to reconstruct a high quality mesh everywhere. We restrict ourselves to the simplest model: the Poisson equation with Dirichlet boundary conditions

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where Ω is a bounded polygonal (resp. polyhedral) domain in $\mathbb{R}^n$, $n = 2$ (resp. $n = 3$), $\partial\Omega$ is its boundary, and $f \in L^2(\Omega)$ is a given function. We only consider the standard piecewise linear continuous finite elements on a simplicial mesh without hanging nodes. The formal (quite usual) definitions of the exact and approximated solutions to (1) in the appropriate functional spaces are given in the beginning of Section 2.

The first goal of the present work is to highlight that we can recover the optimal convergence of the finite element method even if the mesh contains several isolated almost degenerate simplexes. More precisely, we shall assume that the majority of the simplexes in the mesh are regular in the usual Ciarlet sense [11] but there are some distorted simplexes that are typically adjacent to regular mesh cells and well separated from one another by layers of regular cells. The formal assumptions will be given in the beginning of Section 2. To prove the optimal convergence of the standard finite element method, we shall construct a modification of the nodal interpolation operator replacing the standard interpolating polynomial on a degenerate cell by another one obtained by averaging the interpolated function on a patch of cells surrounding the degenerate one.

*The authors acknowledge the support of Région Bourgogne Franche-Comté "Convention Région 2015C-4991. Modèles mathématiques et méthodes numériques pour l'élasticité non-linéaire".

†Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France. e-mail: mduprez@math.cnrs.fr

‡IMAG, Univ Montpellier, CNRS, Montpellier, France. vanessa.lleras@umontpellier.fr

§Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université Bourgogne Franche-Comté, 16, route de Gray, 25030 Besançon Cedex, France. e-mail: alexei.lozinski@univ-fcomte.fr

Although the standard finite element method turns out optimally convergent on the locally damaged meshes, as outlined above, it can suffer from bad conditioning of the stiffness matrix. Indeed, the gradient operator can have an arbitrary large norm on the space of piecewise polynomial functions on a mesh containing very elongated cells even if all the cells are of approximately the same diameter h. In the present paper, we inspire ourselves by [10, 13, 9] to propose an alternative finite element discretization in which we avoid excessively high gradient jumps either by redefining the approximation on bad elements by an extension of the polynomial on a good adjacent element, or by an interior penalty stabilization. We are able to prove that such a scheme is optimally convergent and well conditioned, *i.e.* its conditioning is of the same order as that of a standard finite element method on a usual regular mesh of comparable size, provided the number of degenerate cells remains uniformly bounded.

The present article is a contribution to the already rich literature studying the influence of the mesh cell geometry on the convergence of finite element approximations. The optimal H^1-convergence has been proved in [22] for second order elliptic equation and in [21] for linear elasticity equations under the *minimum angle condition* in 2D: there exists $\alpha_0 \in (0, \pi)$ such that for each considered mesh $\mathcal{T}_h$ and any mesh cell $K \in \mathcal{T}_h$,

$$0 < \alpha_0 \leq \alpha_K, \tag{2}$$

where α_K is the minimum angle of K. In [6, 5], this condition was generalized to the higher dimensions. If we denote by h_K the diameter of K and ρ_K the diameter of the largest ball contained in K, then (2) is equivalent to already mentioned *Ciarlet condition* [11]: there exists c_0 such that for each considered mesh $\mathcal{T}_h$ and any mesh cell $K \in \mathcal{T}_h$,

$$h_K/\rho_K \leq c_0. \tag{3}$$

The conditions above were further relaxed in several ways. Three groups (see [3, 4, 14]) have proposed independently in 1976 a weaker assumption called the *maximum angle condition*: there exists $\beta_0 \in (0, \pi)$ such that for each considered mesh $\mathcal{T}_h$ and any mesh cell $K \in \mathcal{T}_h$,

$$\beta_K \leq \beta_0 < \pi, \tag{4}$$

where β_K is the maximum angle of K. The first condition (2) implies the second (4). The second condition was generalized for higher dimensions in [15, 18].

Furthermore, it is shown in [12] that even the maximum angle condition may be not necessary. More precisely, if a degenerate triangulation is included in a non-degenerate one, then optimal convergence rates hold true. The convergence on appropriate anisotropic meshes is studied in [2]. A sufficient condition for convergence (not necessarily of optimal order) was derived in [16] under the name of the *circumradius condition*: $\max_K R_K \to 0$ as $h \to 0$ where R_K is the circumradius of the mesh cell K. Both the maximum angle and circumradius conditions for $O(h^\alpha)$ convergence are generalized in [17]. It is proved that the triangulations can contain many elements violating these conditions as long as they are isolated in a certain sense. However, one cannot hope for an optimal convergence on completely arbitrary meshes: an example of a heavily distorted mesh family stemming from [3] has been recently analyzed in [20] showing rigorously that the finite element method may fail to converge at all. The present paper propose yet another choice of assumptions on the mesh in the spirit of, but different from [12] and [17], guaranteeing the optimal convergence.

The rest of the paper is organized as follows: in Section 2, we prove that one can allow degenerate cells, which violate Condition (2) or (4), if they are isolated in some sense. We shall establish in Subsection 2.1 the optimal L^2- and H^1-convergence of the standard finite elements method on such meshes. We also recall, in Subsection 2.2, the well known fact that the presence of degenerate cells may induce a large conditioning number of the stiffness matrix. We then propose in Section 3 a modified finite element method that preserves the optimal convergence while ensuring a good conditioning. We conclude with some numerical illustrations in Section 4.

2 Approximation by linear finite elements under local mesh damage assumption

Let us first recall the notions of the weak and approximated solutions to System (1). We call a *weak solution* in $V := H_0^1(\Omega)$ to System (1) a function $u \in V$ such that

$$a(u, v) = l(v) \text{ for all } v \in V, \tag{5}$$

where the bilinear form a and the linear form l are defined for all $u, v \in V$ by

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v \, dx \qquad \text{and} \qquad l(v) := \int_\Omega f v \, dx.$$

It is well known that System (1) admits a unique weak solution thanks to Lax-Milgram lemma.

Consider now a simplicial mesh $\mathcal{T}_h$ on Ω without hanging nodes. This means that $\bar{\Omega} = \cup_{K \in \mathcal{T}_h} K$ with each mesh cell $K \in \mathcal{T}_h$ being a simplex (triangle in 2D, tetrahedron in 3D) and every two mesh cells $K_1, K_2 \in \mathcal{T}_h$ being either disjoint or sharing a vertex, an edge, or a face (in 3D). We recall that ρ_K denotes the diameter of the largest ball contained in a mesh cell K. Moreover, h_ω will denote the diameter of any bounded domain ω and we set $h = \max_{K \in \mathcal{T}_h} h_K$. As mentioned in the Introduction, we will assume that the cells of mesh $\mathcal{T}_h$ satisfy Ciarlet Condition (3) up to some isolated cells.

Assumption 1. *We suppose that there exist $M \in \mathbb{N}$ and $c_0, c_1, c_2, c_3 > 0$ such that the following holds for each considered mesh $\mathcal{T}_h$:*
The mesh contains I degenerate cells $K_1^{deg}, ..., K_I^{deg}$ ($I \geq 0$) violating Ciarlet Condition (3), i.e. for $i \in \{1, ..., I\}$

$$K_i^{deg} \in \mathcal{T}_h \text{ and } h_{K_i^{deg}} / \rho_{K_i^{deg}} > c_0.$$

There exist patches $\mathcal{P}_j$ for $j \in \{1, ..., J\}$, i.e. some unions of mesh cells, star-shaped with respect to a ball of diameter $\rho_{\mathcal{P}_j}$ such that

$$h_{\mathcal{P}_j} / \rho_{\mathcal{P}_j} \leq c_1 \text{ and } h_{\mathcal{P}_j} < c_2 h.$$

We denote by $\widetilde{\mathcal{P}}_j \supset \mathcal{P}_j$ the larger patch composed of mesh cells sharing at least a vertex with $\mathcal{P}_j$. Then

- *The patches $\widetilde{\mathcal{P}}_j$ are mutually disjoint, i.e. $\widetilde{\mathcal{P}}_j$ and $\widetilde{\mathcal{P}}_k$ have no common cells for $j \neq k$.*
- *The number of cells in each $\widetilde{\mathcal{P}}_j \backslash \mathcal{P}_j$ is bounded by a constant M.*

The intersection of boundaries $\partial\mathcal{P}_j$ and $\partial\Omega$ is either empty, or reduced to a point, or to a line segment of length $\geq c_3 h_{\mathcal{P}_j}$, or (in 3D) to a polygon containing a circle of radius $\geq c_3 h_{\mathcal{P}_j}$.

We assume that each degenerate cell K_i^{deg} is included in a patch $\mathcal{P}_j$ (a patch $\mathcal{P}_j$ can contain several degenerate cells). As a consequence, all the cells outside of the patches $\{\mathcal{P}_j\}_{j=1,...,J}$ are non-degenerate.

Notational warning. In what follows, the letter C will stand for constants which depend only on the generalized mesh regularity in the sense of Assumption 1 (unless stated otherwise). This means that C can depend on c_0, c_1, c_2, c_3, and M, but otherwise independent from the choice of mesh $\mathcal{T}_h$. As usual, the value of C can change from one line to another.

An example of patches $\mathcal{P}_i$ and $\widetilde{\mathcal{P}}_i$ is given in Fig. 1. We illustrate there a typical situation of a degenerate triangle K_i^{deg} (dotted in red) adjacent to a regular triangle K_i^{nd} (dotted in grey). The patch $\mathcal{P}_i$ is then formed of these two triangles K_i^{deg} and K_i^{nd}. It is obviously star-shaped with respect to a ball (for example, the largest ball inscribed in K_i^{nd}). Its chunky parameter $h_{\mathcal{P}_i} / \rho_{\mathcal{P}_i}$ is close to that of surrounding regular triangles. We emphasise that Assumption 1 allows for more general configurations, for example, a patch can contain several degenerate cells and does not necessarily contain a non-degenerate cell.

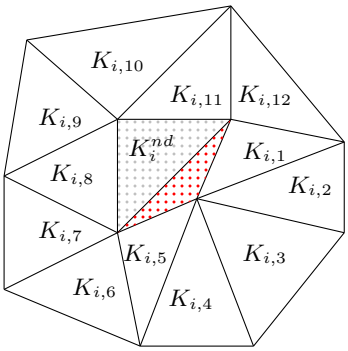

Figure 1: Example of configuration: patch $\mathcal{P}_i$ (dotted), patch $\widetilde{\mathcal{P}}_i$ (all the cells), non-degenerate cell K_i^{nd} (gray) of the patch $\mathcal{P}_i$.

We now set the finite element space on mesh $\mathcal{T}_h$ and the finite element approximation to System (1). Let

$$V_h := \{v_h \in V : v_{h|K} \in \mathbb{P}_1(K)\ \forall K \in \mathcal{T}_h\},$$

where $\mathbb{P}_1(K)$ is the space of polynomials of degree ≤ 1 on cell K. Consider the following finite element approximation to System (5): find $u_h \in V_h$ such that:

$$a(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h. \tag{6}$$

2.1 *A priori* error estimate

In what follows, $|\cdot|_{i,A}$ and $\|\cdot\|_{i,A}$ denote the semi-norm and the norm associated to $H^i(A)$.

Theorem 2. *Let $u \in V$ and $u_h \in V_h$ be the solutions to System (5) and System (6), respectively. Then, under Assumption 1,*

$$|u - u_h|_{1,\Omega} \leq Ch|u|_{2,\Omega}. \tag{7}$$

Moreover, if Ω is convex,

$$\|u - u_h\|_{0,\Omega} \leq Ch^2|u|_{2,\Omega}. \tag{8}$$

The proof of this theorem is completely standard (*cf.* [11, 7]) provided one has constructed an interpolant to V_h satisfying the optimal error estimates. We thus go directly to the construction of such an interpolation operator which we shall call $\widetilde{\mathcal{I}}_h$ and properly introduce in Definition 1. The necessary properties of this operator will be established in the Proposition 1. We start with some technical lemmas.

Lemma 1. *Under Assumption 1, for any $v \in H^2(\Omega) \cap H_0^1(\Omega)$ on any patch $\mathcal{P}_i$ there exists a polynomial $Q_h^i(v)$ on $\mathcal{P}_i$ of degree ≤ 1 vanishing on $\partial\mathcal{P}_i \cap \partial\Omega$ such that*

$$|v - Q_h^i(v)|_{1,\mathcal{P}_i} \leq Ch_{\mathcal{P}_i}|v|_{2,\mathcal{P}_i}, \quad \|v - Q_h^i(v)\|_{0,\mathcal{P}_i} \leq Ch_{\mathcal{P}_i}^2|v|_{2,\mathcal{P}_i}, \quad \|v - Q_h^i(v)\|_{L^\infty(\mathcal{P}_i)} \leq Ch_{\mathcal{P}_i}^{2-n/2}|v|_{2,\mathcal{P}_i}. \tag{9}$$

Proof. We consider first the case of the patch $\mathcal{P}_i$ lying completely inside Ω. We take then $Q_h^i(v)$ on $\mathcal{P}_i$ as the Taylor polynomial Q^2v, *cf.* Definition (4.1.3) from [7], averaged over the ball of diameter $\rho_{\mathcal{P}_i}$ mentioned in Assumption 1. The estimates (9) for $Q_h^i(v) = Q^2v$ are thus given by Proposition (4.3.2) and Bramble-Hilbert Lemma (4.3.8) from [7].

We now turn to the case when the boundary $\partial\mathcal{P}_i$ intersects $\partial\Omega$ in only one point, say x. The polynomial $Q_h^i(v)$ should vanish at x so that we correct Q^2v by subtracting from it its value at point. We set thus $Q_h^i(v) = Q^2v - c_h$ where $c_h = Q^2v(x)$. Since $v(x) = 0$, we have by the above mentioned properties of Q^2v

$$|c_h| = |Q^2v(x) - v(x)| \leq \|Q^2v - v\|_{L^\infty(\mathcal{P}_i)} \leq Ch^{2-n/2}|v|_{2,\mathcal{P}_i}$$

which entails

$$\|Q_h^i v - v\|_{0,\mathcal{P}_i} \leq \|Q^2 v - v\|_{0,\mathcal{P}_i} + |c_h||\mathcal{P}_i|^{1/2} \leq Ch^2 |v|_{2,\mathcal{P}_i}$$

and

$$\|Q_h^i v - v\|_{L^\infty(\mathcal{P}_i)} \leq \|Q^2 v - v\|_{L^\infty(\mathcal{P}_i)} + |c_h| \leq Ch^{2-n/2} |v|_{2,\mathcal{P}_i} .$$

The H^1 semi-norm of the error is not affected by the constant c_h, so that the announced estimate for $|Q_h^i v - v|_{1,\mathcal{P}_i}$ is also valid.

The last case to consider is when $\partial\mathcal{P}_i$ has a non-empty intersection with a side, say Γ, of $\partial\Omega$, which is not reduced to one point. We introduce the polynomial c_h of degree ≤ 1 that coincides with $Q^2 v$ on $\partial\mathcal{P}_i \cap \Gamma$ and does not vary in the directions perpendicular to $\Gamma \cap \partial\mathcal{P}_i$. Setting $Q_h^i(v) = Q^2 v - c_h$ we see immediately that $Q_h^i(v)$ vanishes on Γ. Let $\Pi_\Gamma \mathcal{P}_i$ be the projection of $\mathcal{P}_i$ on Γ (or, in the 3D case when the intersection $\partial\mathcal{P}_i \cap \Gamma$ is reduced to a segment, the projection of $\mathcal{P}_i$ on the line containing this segment). Thanks to our geometrical assumptions and the fact that v vanishes on $\partial\mathcal{P}_i \cap \Gamma$,

$$\|c_h\|_{L^\infty(\mathcal{P}_i)} = \|c_h\|_{L^\infty(\Pi_\Gamma \mathcal{P}_i)} \leq C\|c_h\|_{L^\infty(\partial\mathcal{P}_i \cap \Gamma)} = C\|Q^2 v - v\|_{L^\infty(\partial\mathcal{P}_i \cap \Gamma)} \leq Ch_{\mathcal{P}_i}^{2-n/2} |v|_{2,\mathcal{P}_i} .$$

Note that the first inequality is valid thanks to the hypothesis on the intersection between $\mathcal{P}_i$ and Γ given in Assumption 1 as proven in Lemma 9. We can thus prove the desired estimates for $\|Q_h^i v - v\|_{0,\mathcal{P}_i}$ and $\|Q_h^i v - v\|_{L^\infty(\mathcal{P}_i)}$ as in the previous case. Finally, by an inverse inequality (proven in this context in Lemma 10),

$$\|\nabla c_h\|_{L^\infty(\mathcal{P}_i)} \leq \frac{C}{h_{\mathcal{P}_i}} \|c_h\|_{L^\infty(\partial\mathcal{P}_i \cap \Gamma)} \leq Ch_{\mathcal{P}_i}^{1-n/2} |v|_{2,\mathcal{P}_i}$$

so that

$$|Q_h^i v - v|_{1,\mathcal{P}_i} \leq |Q^2 v - v|_{1,\mathcal{P}_i} + \|\nabla c_h\|_{L^\infty(\mathcal{P}_i)} |\mathcal{P}_i|^{1/2} \leq Ch|v|_{2,\mathcal{P}_i} .$$

□

We also recall the usual interpolation error estimates on regular cells for the standard Lagrange interpolation operator $\mathcal{I}_h$ to the space of piecewise linear functions, *cf.* [11, 7].

Lemma 2. *Under Assumption 1, we have on each mesh cell $K \in \mathcal{T}_h$ outside of patches $\mathcal{P}_i$*

$$|v - \mathcal{I}_h(v)|_{1,K} \leq Ch|v|_{2,K}, \quad \|v - \mathcal{I}_h(v)\|_{0,K} \leq Ch^2 |v|_{2,K}, \quad \|v - \mathcal{I}_h(v)\|_{L^\infty(K)} \leq Ch^{2-n/2} |v|_{2,K},$$

for any $v \in H^2(K)$.

Remark 1. In Assumption 1, we can replace the Ciarlet condition by the maximum angle condition (4) on the cells $K \in \mathcal{T}_h$ outside of patches $\widetilde{\mathcal{P}}_i$. Indeed, the inequalities of Lemma 2 remain valid in this case.

Definition 1. For all $v \in H^2(\Omega) \cap H_0^1(\Omega)$, let $\widetilde{\mathcal{I}}_h(v)$ be the function in V_h that coincides with $Q_h^i(v)$ from Lemma 1 on each patch $\mathcal{P}_i$, and with the standard Lagrange interpolation $\mathcal{I}_h v$ on all the cells $K \in \mathcal{T}_h$ out of the extended patches $\widetilde{\mathcal{P}}_i$, *i.e.* $\widetilde{\mathcal{I}}_h(v)(x) = v(x)$ at all the mesh nodes $x \in \bar{\Omega} \setminus \cup_{i \in \{1,\dots,I\}} \widetilde{\mathcal{P}}_i$.

Note that $\widetilde{\mathcal{I}}_h(v)$ is uniquely defined also on the mesh cells from $\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$, $i = 1, \dots, I$ although they are not explicitly mentioned above. Indeed, all the vertices of such cells are shared either with a patch $\mathcal{P}_i$ or with a regular cell from $\bar{\Omega} \setminus \cup_{i \in \{1,\dots,I\}} \widetilde{\mathcal{P}}_i$. Since the values of $\widetilde{\mathcal{I}}_h(v)$ are given at all these nodes by the definition above, the piecewise linear function $\widetilde{\mathcal{I}}_h(v)$ is well defined everywhere.

We now prove the global interpolation estimates for the interpolation operator $\widetilde{\mathcal{I}}_h$.

Proposition 1. *Under Assumption 1, we have for all $v \in H^2(\Omega) \cup H_0^1(\Omega)$*

$$|v - \widetilde{\mathcal{I}}_h(v)|_{1,\Omega} \leq Ch|v|_{2,\Omega}, \quad \|v - \widetilde{\mathcal{I}}_h(v)\|_{0,\Omega} \leq Ch^2 |v|_{2,\Omega}.$$

Proof. The contributions to the interpolation errors on the patches $\mathcal{P}_i$ and on the mesh cells outside the patches $\widetilde{\mathcal{P}}_i$ (where the interpolators $\widetilde{\mathcal{I}}_h$ and $\mathcal{I}_h$ coincide) are already covered by Lemmas 1 and 2. It remains to bound the error on mesh cells in $\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$.

Let $K \in \mathcal{T}_h$ and $K \subset \widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$. By the triangle inequality and Lemma 2,

$$|v - \widetilde{\mathcal{I}}_h(v)|_{1,K} \leq |v - \mathcal{I}_h(v)|_{1,K} + |\mathcal{I}_h(v) - \widetilde{\mathcal{I}}_h(v)|_{1,K} \leq Ch|v|_{2,K} + |r_h|_{1,K},$$

where we have denoted $r_h := \mathcal{I}_h(v) - \widetilde{\mathcal{I}}_h(v)$. By a homogeneity argument and the equivalence of norms on finite dimensional space, we see easily

$$|r_h|_{1,K} \leq Ch^{n/2-1}\|r_h\|_{L^\infty(K)}.$$

Recalling that r_h is a polynomial of degree ≤ 1 vanishing at the vertices of K on $\partial\widetilde{\mathcal{P}}_i$, the other vertices belonging to $\partial\mathcal{P}_i$, we conclude

$$\|r_h\|_{L^\infty(K)} \leq \|r_h\|_{L^\infty(\partial K \cap \partial\mathcal{P}_i)} \leq \|v - \mathcal{I}_h(v)\|_{L^\infty(K)} + \|v - \widetilde{\mathcal{I}}_h(v)\|_{L^\infty(\mathcal{P}_i)} \leq Ch^{2-n/2}(|v|_{2,K} + |v|_{2,\mathcal{P}_i}).$$

Putting the estimates above together yields

$$|v - \widetilde{\mathcal{I}}_h(v)|_{1,K} \leq Ch(|v|_{2,K} + |v|_{2,\mathcal{P}_i}). \tag{10}$$

Similarly,

$$\|v - \widetilde{\mathcal{I}}_h(v)\|_{0,K} \leq Ch^2(|v|_{2,K} + |v|_{2,\mathcal{P}_i}). \tag{11}$$

Taking the square on both sides of (10) and (11), summing them over all the mesh cells $K \subset \widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$, $i = 1, \ldots, I$ (recall that the number of such cells on each patch is bounded by a predefined constant M), adding the estimates from lemma 1 on the patches $\mathcal{P}_i$ and those of Lemma 2 on the mesh cells outside the patches $\widetilde{\mathcal{P}}_i$ gives the desired result. □

2.2 Poor conditioning of the system matrix

In this section, we shall recall the well known fact that the presence of degenerate cells can induce an arbitrary large conditioning number of the associated finite element matrix. In the following proposition, we consider a particular example of a mesh satisfying Assumption 1 and give an estimator for the conditioning number. This result should be contrasted with the "normal" conditioning number of order $1/h^2$ on a quasi-uniform mesh.

Proposition 2. *Suppose that the mesh $\mathcal{T}_h$ satisfies Assumption 1 and contains a degenerate cell K^{deg} such that*

$$\rho_{K^{deg}} = \varepsilon, \quad h_{K^{deg}} \geqslant C_1 h. \tag{12}$$

Then the conditioning number $\boldsymbol{\kappa}(\boldsymbol{A}) := \|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|_2$ of the matrix $\boldsymbol{A}$ associated to the bilinear form a in V_h satisfies

$$\boldsymbol{\kappa}(\boldsymbol{A}) \geqslant \frac{C}{h\varepsilon}$$

for sufficiently small h, with C depending only on C_1 and Ω. Here, $\|\cdot\|_2$ stands for the matrix norm associated to the vector 2-norm.

Proof. Denote by N the dimension of V_h. Consider ϕ_h the basis function of V_h equal to 1 at the node of K^{deg} opposite to the largest edge (face) of K^{deg}, vanishing at all the other nodes, and $\boldsymbol{\phi} \in \mathbb{R}^N$ the vector representing ϕ_h in the basis of hat functions. Then, denoting by $|\cdot|_2$ the vector 2-norm on $\mathbb{R}^N$ and by $(\cdot,\cdot)$ the associated inner product,

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{u}\in\mathbb{R}^N} \frac{(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{u})}{|\boldsymbol{u}|_2^2} \geqslant (\boldsymbol{A}\boldsymbol{\phi}, \boldsymbol{\phi}) = a(\phi_h, \phi_h) = |\phi_h|_{1,\Omega}^2 \geqslant |\phi_h|_{1,K^{deg}}^2.$$

By (12), the gradient of ϕ_h is of order $1/\varepsilon$ on K^{deg}, and the area of K^{deg} is of order εh^{n-1}. Thus,

$$\|\boldsymbol{A}\|_2 \geqslant |\phi_h|^2_{1,K^{deg}} \geqslant C\frac{h^{n-1}}{\varepsilon}.$$

Now take any $\psi \in H^2(\Omega) \cap H^1_0(\Omega)$, $\psi \neq 0$ and let $\boldsymbol{\psi} \in \mathbb{R}^N$ be the vector associated to $\widetilde{\mathcal{I}}_h\psi$. Then

$$\|\boldsymbol{A}^{-1}\|_2 = \sup_{\boldsymbol{u}\in\mathbb{R}^N} \frac{|\boldsymbol{u}|^2_2}{(\boldsymbol{A}\boldsymbol{u},\boldsymbol{u})} \geqslant \frac{|\boldsymbol{\psi}|^2_2}{(\boldsymbol{A}\boldsymbol{\psi},\boldsymbol{\psi})} = \frac{|\boldsymbol{\psi}|^2_2}{a(\widetilde{\mathcal{I}}_h\psi,\widetilde{\mathcal{I}}_h\psi)} \geqslant \frac{C}{h^n}\frac{\|\widetilde{\mathcal{I}}_h\psi\|^2_{0,\Omega}}{|\widetilde{\mathcal{I}}_h\psi|^2_{1,\Omega}}.$$

We have used here the bound

$$\|v_h\|^2_{0,\Omega} \leq Ch^n|\boldsymbol{v}|^2_2$$

valid for any $v_h \in V_h$ and the corresponding vector $\boldsymbol{v}$ since all the mesh cells are of diameter $\leq h$. Proposition 1 implies

$$\begin{cases} \|\widetilde{\mathcal{I}}_h\psi\|_{0,\Omega} \geqslant \|\psi\|_{0,\Omega} - Ch^2|\psi|_{2,\Omega} \geqslant C\|\psi\|_{0,\Omega}, \\ |\widetilde{\mathcal{I}}_h\psi|_{1,\Omega} \leq |\psi|_{1,\Omega} + Ch|\psi|_{2,\Omega} \leq C|\psi|_{1,\Omega}, \end{cases}$$

for h small enough. So that

$$\|\boldsymbol{A}^{-1}\|_2 \geqslant \frac{C}{h^n}.$$

This gives the desired result. □

3 A well conditionned alternative finite element scheme

In this section, we build an alternative finite element method for which the optimal convergence rates (7) and (8) hold true and the conditioning number of the finite element matrix is of order C/h^2 if all the mesh cells are of diameter $\sim h$. We start by the observation that such a method could be based on a subspace $\widetilde{V}_h \subset V_h$ which is the image of interpolation operator $\widetilde{\mathcal{I}}_h$, *i.e.*

$$\widetilde{V}_h := \{v_h \in V_h : [\nabla v_h]_{|F} = 0 \text{ for all } F \in \mathcal{F}_i, i \in \{1, ..., I\}\},$$

where $\mathcal{F}_i$ is the set of interior edges (faces) of the patch $\mathcal{P}_i$ and $[\cdot]_{|F}$ represents the jump on F. In view of our interpolation estimates, the problem of finding $\widetilde{u}_h \in \widetilde{V}_h$ such that

$$a(\widetilde{u}_h, \widetilde{v}_h) = l(\widetilde{v}_h) \text{ for all } \widetilde{v}_h \in \widetilde{V}_h$$

would produce an approximate solution with optimal error. Moreover, it is easy to see that the matrix would be well-conditioned since the space $\widetilde{V}_h$ ignores the degenerate cells. Such a method is only of theoretical interest because one cannot easily construct a basis for $\widetilde{V}_h$ using available finite element libraries. In what follows, we use this problem rather as an inspiration in constructing an implementable finite element scheme.

In doing so, we shall impose further restrictions on the mesh:

Assumption 3. *The mesh satisfies Assumption 1. Moreover, there exists $c_4 > 0$ and $I_{\max}, M' \in \mathbb{N}$ such that for each considered mesh $\mathcal{T}_h$:*

- *The number of patches I is bounded by some $I_{\max}$.*
- *Each patch $\mathcal{P}_i$ contains a non-degenerate cell K_i^{nd}, i.e. such that $h_{\mathcal{P}_i}/\rho_{K_i^{nd}} \leq c_4$.*
- *The number of cells in each $\mathcal{P}_i$ is bounded by a constant M'.*

In what follows, the constants C will be allowed to depend on the additional parameters in Assumption 3, *i.e.* c_4, $I_{\max}$, and M'.

We shall need the following modification of the previously defined interpolation operator $\widetilde{\mathcal{I}}_h$, which makes sense under Assumption 3, *cf.* also Fig. 1, and will be incorporated explicitly into our modified finite element scheme.

Definition 2. For all $v \in H^2(\Omega) \cap H_0^1(\Omega)$, let $\widehat{\mathcal{I}}_h(v)$ be the function in V_h that coincides with the standard Lagrange interpolation $\mathcal{I}_h v$ on all the cells $K \in \mathcal{T}_h$ out of the extended patches $\widetilde{\mathcal{P}}_i$, and is given on each patch $\mathcal{P}_i$, not touching the boundary $\partial\Omega$, by

$$\widehat{\mathcal{I}}_h(v)|_{\mathcal{P}_i} := \text{Ext}\left(\mathcal{I}_h(v)|_{K_i^{nd}}\right), \tag{13}$$

where $\mathcal{I}_h$ stands again for the standard Lagrange interpolation operator on K_i^{nd}, and Ext stands for the extension of a polynomial from $K_i^{nd} \subset \mathcal{P}_i$ to the whole $\mathcal{P}_i$ without changing the coefficients of the polynomial. If the patch $\mathcal{P}_i$ touches $\partial\Omega$, then $\widehat{\mathcal{I}}_h(v)$ is also based there on formula (13), corrected as in Lemma 1.

REMARK 2. The new interpolation operator $\widehat{\mathcal{I}}_h(v)$ satisfies the same optimal estimates as that for the old operator $\widetilde{\mathcal{I}}_h(v)$ which are given in Proposition 1, the proof of which is based on Lemma 1. To prove that Lemma 1 remains valid for $\widehat{\mathcal{I}}_h$, *i.e.* redefining in (9) the original Q_h^i by $Q_h^i := \widehat{\mathcal{I}}_h(v)|_{\mathcal{P}_i}$ as in (13), we refer to Theorem (4.4.4) and Corollary (4.4.7) from [7]. Following their proofs, one can see that the only thing to check is the boundedness of operator Ext in (13) as a linear map on the space of polynomials of degree ≤ 1 equipped with the norm of $L^\infty(K_i^{nd})$ to $L^\infty(\mathcal{P}_i)$. This, in turn, follows easily from our geometrical Assumptions 1, 3.

REMARK 3. Our construction of the interpolation operator $\widehat{\mathcal{I}}_h$ is very similar to that of [17]. The assumptions on the mesh in [17] are much more refined than ours and are intended to be close to necessary ones. Our proofs, on the other hand, are significantly simpler. Moreover, we are able to treat 3D meshes, which is not the case in [17].
We note that our assumptions could be somewhat relaxed (at the expense of readability of the present paper) so that some mesh configurations from [17], not allowed by our Assumption 1, could be recovered. For example, in Fig. 1, we do not really need to include the cells $K_{i,6}, ..., K_{i,12}$ in the patch $\widetilde{\mathcal{P}}_i$ if we define the interpolation by extension from the non-degenerate triangles as in (13). Indeed, $\widehat{\mathcal{I}}_h$ is different from the standard Lagrange interpolation only on a part of the patch $\widetilde{\mathcal{P}}_i$, *i.e.* on the degenerate triangle and on $K_{i,1}, ..., K_{i,5}$. Our construction of the optimal interpolation operator would thus remain valid if Assumption 1 were modified so that $\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$ only contained the "essential" cells ($K_{i,1}, ..., K_{i,5}$ in Fig. 1). In particular, $\widehat{\mathcal{I}}_h$ provides an optimal interpolation operator on the band of heavily stretched cells as represented at Fig. 13 in [17].
As in [17] (cf. Definition 25 and Lemma 31), we can also consider some clusterings of degenerate cells (separated from one another by non-degenerate cells), but we should then stick to our first construction of the interpolation operator $\widetilde{\mathcal{I}}_h(v)$ from Definition 1. In this situation, our Assumption 1 is even more general than that of [17] since we do not need to suppose that each patch $\mathcal{P}_i$ contain a non-degenerate cell.

3.1 An alternative scheme

We denote by a_ω the restriction of a on a subset ω of Ω, and by $(\cdot,\cdot)_\omega$ the inner product in $L^2(\omega)$. Consider the bilinear form a_h defined for all $u_h,\ v_h \in V_h$ by

$$a_h(u_h, v_h) := a_{\Omega_h^{nd}}(u_h, v_h) + \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h v_h) + \sum_i \frac{1}{h_{\mathcal{P}_i}^2}((\text{Id} - \widehat{\mathcal{I}}_h)u_h, (\text{Id} - \widehat{\mathcal{I}}_h)v_h)_{\mathcal{P}_i}, \tag{14}$$

where $\Omega_h^{nd} := \Omega \backslash (\cup_i \overline{\mathcal{P}_i})$ and the interpolation operator $\widehat{\mathcal{I}}_h$ is defined by (13), *i.e.* u_h is not used directly inside the patches in the second term of a_h, but rather it is extended from a non-degenerate cell inside each patch. The third term in a_h will serve, loosely speaking, to penalize the eventual gap between the approximate solution u_h and the optimal subspace $\widetilde{V}_h$, which is here quantified by the difference between u_h and its interpolation $\widehat{\mathcal{I}}_h u_h$ in $\widetilde{V}_h$.

We now introduce the following method approximating System (5): find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h. \tag{15}$$

The idea of using the polynomial extension from "good" to "bad" mesh cells in the scheme (15) is borrowed from [13]. We shall also see that the scheme can be recast in a form using the interior penalization on the mesh facets between "good" and "bad" cells, as in the ghost penalty method [9].

3.2 *A priori* estimate

The approximation of System (5) by (15) induces a quasi-optimal convergence rate:

Theorem 4 (*A priori* estimate)**.** *Let $u \in V$ and $u_h \in V_h$ be the solutions to System* (5) *and System* (15)*, respectively. Then, under Assumption 3, we have for any $\varepsilon > 0$*

$$|u - \Pi_h u_h|_{1,\Omega} := |u - u_h|_{1,\Omega_h^{nd}} + \sum_i |u - \widehat{\mathcal{I}}_h u_h|_{1,\mathcal{P}_i} \leq \begin{cases} \frac{C}{\varepsilon} h^{1-\varepsilon} \|u\|_{2,\Omega} & \text{if } n = 2, \\ Ch\|u\|_{2,\Omega} & \text{if } n = 3, \end{cases} \tag{16}$$

where $\Pi_h u_h$ is equal to u_h on Ω_h^{nd} and $\widehat{\mathcal{I}}_h u_h$ on $\mathcal{P}_i$. Moreover, if Ω is convex,

$$\|u - u_h\|_{0,\Omega} \leq C_\varepsilon \begin{cases} \frac{C}{\varepsilon} h^{2-\varepsilon} \|u\|_{2,\Omega} & \text{if } n = 2, \\ Ch^2\|u\|_{2,\Omega} & \text{if } n = 3. \end{cases}$$

Before proving Theorem 4, we first give some auxiliary results. Note that the optimal error order (h for the H^1-norm and h^2 for the L^2-norm can be recovered also in the 2D case, assuming more regularity on u, cf. Remark 4.

Lemma 3 (Galerkin orthogonality)**.** *Consider u and u_h the solution to Systems* (5) *and* (15)*. Then*

$$a_{\Omega_h^{nd}}(u_h - u, v_h) - \sum_i a_{\mathcal{P}_i}(u, v_h) + \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h v_h) + \sum_i \frac{1}{h_{\mathcal{P}_i}^2}((Id - \widehat{\mathcal{I}}_h)u_h, (Id - \widehat{\mathcal{I}}_h)v_h)_{0,\mathcal{P}_i} = 0,$$

for all $v_h \in V_h$.

The proof of Lemma 3 is immediate.

We shall need the norm $|||\cdot|||$ defined for all $v_h \in V_h$ by

$$|||v_h||| := a_h(v_h, v_h)^{1/2} = \left(|v_h|_{1,\Omega_h^{nd}}^2 + \sum_i |\widehat{\mathcal{I}}_h v_h|_{1,\mathcal{P}_i}^2 + \sum_i \frac{1}{h_{\mathcal{P}_i}^2} \|v_h - \widehat{\mathcal{I}}_h v_h\|_{0,\mathcal{P}_i}^2 \right)^{\frac{1}{2}}.$$

Note for the future use that this norm is also well defined on $V \cap H^2(\Omega)$.

Lemma 4. *Under Assumption 3, for all $v_h \in V_h$, it holds*

$$\sum_i |v_h - \widehat{\mathcal{I}}_h v_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2 \leq C \,|||\, v_h \,|||^2 .$$

Proof. It suffices to prove for each patch

$$\sup \frac{|v_h - \widehat{\mathcal{I}}_h v_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2}{|v_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2 + |\widehat{\mathcal{I}}_h v_h|_{1,\mathcal{P}_i}^2 + \frac{1}{h_{\mathcal{P}_i}^2}\|(\mathrm{Id} - \widehat{\mathcal{I}}_h)v_h\|_{0,\mathcal{P}_i}^2} \leq C, \tag{17}$$

where the supremum is taken over all the configurations of $\mathcal{P}_i$ and $\widetilde{\mathcal{P}}_i$ allowed by Assumption 3 and over all the continuous piecewise linear functions v_h on the extended patch $\widetilde{\mathcal{P}}_i$ such that $|v_h - \widehat{\mathcal{I}}_h v_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \neq 0$ *i.e.* $v_h - \widehat{\mathcal{I}}_h v_h \neq const$ on $\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i$. First of all, we need to verify that the expression to maximize in (17) is well defined on such v_h, *i.e.* no division by zero occurs. In other words, we need to prove that if the denominator vanishes, *i.e.*

$$|v_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2 + |\widehat{\mathcal{I}}_h v_h|_{1,\mathcal{P}_i}^2 + \frac{1}{h_{\mathcal{P}_i}^2}\|(\mathrm{Id} - \widehat{\mathcal{I}}_h)v_h\|_{0,\mathcal{P}_i}^2 = 0$$

then $v_h - \widehat{\mathcal{I}}_h v_h$ is constant on $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$. This verification goes as follows: since $|v_h|_{1,\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i} = 0$ and $|\widehat{\mathcal{I}}_h v_h|_{1,\mathcal{P}_i} = 0$

$$v_h = C_i \text{ in } \widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i, \quad \widehat{\mathcal{I}}_h v_h = D_i \text{ in } \mathcal{P}_i$$

with some constants C_i and D_i. Since $\|(\mathrm{Id} - \widehat{\mathcal{I}}_h) v_h\|_{0,\mathcal{P}_i} = 0$, we deduce that

$$v_h = \widehat{\mathcal{I}}_h v_h = C_i = D_i \text{ in } \mathcal{P}_i$$

so that $(\mathrm{Id} - \widehat{\mathcal{I}}_h) v_h = 0$ on $\partial \mathcal{P}_i$. This entails $(\mathrm{Id} - \widehat{\mathcal{I}}_h) v_h = 0$ in $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$ by construction of $\widehat{\mathcal{I}}_h$.

The finiteness of the supremum in (17) now follows from the fact that both $v_h|_{\widetilde{\mathcal{P}}_i}$ and the geometry of the patch $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$ are governed by a finite number of parameters, which can be assumed to live in a bounded closed set thanks to homogeneity arguments. More specifically, when maximizing the expression in (17), we can safely assume the following:

- The number of cells in each $\widetilde{\mathcal{P}}_i$ is equal to some $m \in \mathbb{N}$ (in fact, this number is bounded by a constant M from Assumption 1 so that one can perform the maximization for, successively, $m = 1, 2, \ldots, M$ and then take the maximum of supremums obtained for each m).
- $|v_h - \widehat{\mathcal{I}}_h v_h|_{1,\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i} = 1$. Indeed, the expression to maximize in (17) is invariant under the transformation $v_h \mapsto \alpha v_h$ for any $\alpha \neq 0$.
- The diameter of $\widetilde{\mathcal{P}}_i$ is equal to 1 and its barycenter is at the origin of the coordinate system. Indeed, the expression to maximize in (17) is invariant under the coordinate transformations $x \mapsto hx$ (homothety with the coefficient $h \neq 0$) and $x \mapsto x + a$ (shift by a vector a).

Imposing the constraints above to the maximization in (17) we see that the maximum is sought here over a bounded set in a finite dimensional space, *i.e.* the space of parameters that give both $v_h|_{\widetilde{\mathcal{P}}_i}$ and the geometry of the patch $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$. This set is also closed since a converging sequence of patch geometries satisfying Assumption 1 tends to a patch geometry also satisfying this Assumption (neither the simplices in $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$, nor the patch $\mathcal{P}_i$ can degenerate in the limit to something other than a simplex or a patch, since all of those geometrical objects should always contain some balls whose radius cannot go to zero). Moreover, the expression to maximize in (17) depends continuously on $v_h|_{\widetilde{\mathcal{P}}_i}$ and the geometry of the patch $\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i$ (no division by zero). We conclude thus that the supremum in (17) is attained and it is thus finite. □

Lemma 5. *Under Assumption 3, we have for any $\varepsilon > 0$*

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i} \right)^{\frac{1}{2}} \leq \begin{cases} \frac{C}{\varepsilon} h^{1-\varepsilon} \|u\|_{2,\Omega} & \text{if } n = 2, \\ Ch \|u\|_{2,\Omega} & \text{if } n = 3, \end{cases}$$

for all $u \in H^2(\Omega)$.

Proof. Let $\mathcal{P} = \cup_{i=1}^I (\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i)$. Assumption 3 implies $|\mathcal{P}| \leqslant Ch^n$ (recall that the number of patches I is assumed uniformly bounded). Let us consider first the case $n = 3$. Using Hölder inequality with exponents 3 and $\frac{3}{2}$, we have

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i} \right)^{\frac{1}{2}} = \left(\int_{\mathcal{P}} |\nabla u|^2 \right)^{\frac{1}{2}} \leq \|\nabla u\|_{L^6(\mathcal{P})} |\mathcal{P}|^{\frac{1}{3}} \leq Ch \|\nabla u\|_{L^6(\Omega)}.$$

We conclude thanks to the well known Sobolev embedding $H^1(\Omega) \to L^6(\Omega)$

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i \backslash \mathcal{P}_i} \right)^{\frac{1}{2}} \leq Ch \|\nabla u\|_{1,\Omega}.$$

We now turn to the case $n = 2$. Using Hölder inequality with exponents $\frac{q}{2}$ and $\frac{q/2}{q/2-1}$ for any $q > 2$, we have

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i}\right)^{\frac{1}{2}} = \left(\int_{\mathcal{P}} |\nabla u|^2\right)^{\frac{1}{2}} \leq \|\nabla u\|_{L^q(\mathcal{P})}|\mathcal{P}|^{\frac{q-2}{2q}} \leq Ch^{\frac{q-2}{q}}\|\nabla u\|_{L^q(\Omega)}. \tag{18}$$

We now use Sobolev embedding $H^1(\Omega) \to L^q(\Omega)$ with the explicit dependence on q (cf. [19, Cor. 1.57])

$$\|v\|_{L^q(\Omega)} \leqslant C'\frac{q}{2}\|v\|_{1,\Omega}(C''|\Omega|)^{\frac{1}{q}}, \quad \forall v \in H^1(\Omega) \tag{19}$$

with constants $C', C'' > 0$ depending only on Ω. Note that (19) is slightly different from the actual statement in [19] and was adapted here as follows:

- The constant C in [19, Cor. 1.57] is not explicitly written, but it is easily restored from the proof. It is indeed equal to $\frac{q}{2}$ in our setting $p = n = 2$, $q > 2$.
- The proof of [19, Cor. 1.57] is done for the functions vanishing on the boundary. The only purpose of this assumption there is to extend v by 0 outside Ω resulting in a function in the same Sobolev space over the whole $\mathbb{R}^n$. In our case of $v \in H^1(\Omega)$ we can use instead an extension operator $H^1(\Omega) \to H^1(\mathbb{R}^2)$ assuming that the extended functions are compactly supported in a bounded set $\tilde{\Omega} \supset \Omega$ with $|\tilde{\Omega}| \leq C''|\Omega|$. We thus recover (19) with a constant C' which is the norm of the extension operator.

Finally, setting $v = \nabla u$ in (19), we get

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i}\right)^{\frac{1}{2}} \leq C\frac{q}{2}h^{\frac{q-2}{q}}\|\nabla u\|_{1,\Omega}$$

hence the result setting $\varepsilon = 2/q$. □

Remark 4. In the 2D case, assuming $\nabla u \in L^\infty(\Omega)$, the estimate (18) can be improved as

$$\left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i}\right)^{\frac{1}{2}} \leq \|\nabla u\|_{L^\infty(\mathcal{P})}|\mathcal{P}|^{\frac{1}{2}} \leq Ch\|\nabla u\|_{L^\infty(\Omega)}$$

Thus, following the proof of Theorem 4 below, we obtain

$$\begin{aligned} |u - \Pi_h u_h|_{1,\Omega} &\leq Ch(|u|_{2,\Omega} + \|\nabla u\|_{L^\infty(\Omega)}), \\ \|u - u_h\|_{0,\Omega} &\leq Ch^2(|u|_{2,\Omega} + \|\nabla u\|_{L^\infty(\Omega)}) \qquad (\text{ if } \Omega \text{ is convex}). \end{aligned}$$

Proof of Theorem 4. Let $e_h := \widehat{\mathcal{I}}_h u - u_h$. We remark that

$$\begin{aligned} |||e_h|||^2 &= a_h(e_h, e_h) \\ &= a_{\Omega_h^{nd}}(\widehat{\mathcal{I}}_h u - u_h, e_h) + \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u - \widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h e_h) \\ &\qquad + \sum_i \frac{1}{h^2_{\mathcal{P}_i}}((\widehat{\mathcal{I}}_h - \mathrm{Id})u_h, (\mathrm{Id} - \widehat{\mathcal{I}}_h)e_h)_{\mathcal{P}_i}. \end{aligned}$$

Lemma 3 leads to

$$\begin{aligned} |||e_h|||^2 &= a_{\Omega_h^{nd}}(\widehat{\mathcal{I}}_h u - u, e_h) + \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u, \widehat{\mathcal{I}}_h e_h) - \sum_i a_{\mathcal{P}_i}(u, e_h) \\ &= a_{\Omega_h^{nd}}(\widehat{\mathcal{I}}_h u - u, e_h) + \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u - u, \widehat{\mathcal{I}}_h e_h) + \sum_i a_{\mathcal{P}_i}(u, \widehat{\mathcal{I}}_h e_h - e_h). \end{aligned} \tag{20}$$

We now estimate each term in the right-hand side. Using Proposition 1 for the interpolation operator $\widehat{I}_h$, cf. Remark 2, it holds

$$\begin{aligned} a_{\Omega_h^{nd}}(\widehat{\mathcal{I}}_h u - u, e_h) &\leq |\widehat{\mathcal{I}}_h u - u|_{1,\Omega} |e_h|_{1,\Omega_h^{nd}} \\ &\leq Ch|u|_{2,\Omega} \,|||\, e_h |||\end{aligned} \tag{21}$$

and

$$\begin{aligned} \sum_i a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u - u, \widehat{\mathcal{I}}_h e_h) &\leq \sum_i |\widehat{\mathcal{I}}_h u - u|_{1,\mathcal{P}_i} |\widehat{\mathcal{I}}_h e_h|_{1,\mathcal{P}_i} \\ &\leq Ch|u|_{2,\Omega} \,|||\, e_h \,|||\,. \end{aligned} \tag{22}$$

Concerning the third term, it holds

$$\begin{aligned} \sum_i a_{\mathcal{P}_i}(u, \widehat{\mathcal{I}}_h e_h - e_h) &= \sum_i (\partial_n u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\partial\mathcal{P}_i} - \sum_i (\Delta u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\mathcal{P}_i} \\ &= \sum_i (\Delta u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} - \sum_i (\Delta u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\mathcal{P}_i} + \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(u, \widehat{\mathcal{I}}_h e_h - e_h). \end{aligned}$$

Since $h_{\widetilde{\mathcal{P}}_i} \leq Ch$, we obtain the Poincaré type inequality

$$\|\widehat{\mathcal{I}}_h e_h - e_h\|_{0,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \leq Ch \|\widehat{\mathcal{I}}_h e_h - e_h\|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}. \tag{23}$$

By Cauchy-Schwarz inequality, inequality (23), Lemma 4 it holds

$$\sum_i (\Delta u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \leq \left(\sum_i |u|^2_{2,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \right)^{1/2} \left(\sum_i \|\widehat{\mathcal{I}}_h e_h - e_h\|^2_{0,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \right)^{1/2} \leq Ch|u|_{2,\Omega} \,|||\, e_h \,|||\,. \tag{24}$$

Again, using Cauchy-Schwarz inequality, we obtain

$$\sum_i (\Delta u, \widehat{\mathcal{I}}_h e_h - e_h)_{0,\mathcal{P}_i} \leq \left(\sum_i |u|^2_{2,\mathcal{P}_i} \right)^{1/2} \left(\sum_i \|\widehat{\mathcal{I}}_h e_h - e_h\|^2_{0,\mathcal{P}_i} \right)^{1/2} \leq Ch|u|_{2,\Omega} \,|||\, e_h \,|||\,. \tag{25}$$

By Lemmas 4 and 5, for each $\varepsilon > 0$

$$\sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(u, \widehat{\mathcal{I}}_h e_h - e_h) \leq \left(\sum_i |u|^2_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \right)^{\frac{1}{2}} \left(\sum_i |\widehat{\mathcal{I}}_h e_h - e_h|^2_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i} \right)^{\frac{1}{2}} \leq C_\varepsilon h^{1-\varepsilon} \|u\|_{2,\Omega} \,|||\, e_h |||, \tag{26}$$

where $C_\varepsilon = C/\varepsilon$ if $n = 2$ and $C_\varepsilon = C$ if $n = 3$. Thus, Proposition 1 for the interpolation operator $\widehat{I}_h$ and (20), (21), (22), (24), (25), (26) lead to

$$|||u - u_h||| \leq |||u - \widehat{\mathcal{I}}_h u \,||| + |||\, \widehat{\mathcal{I}}_h u - u_h ||| \leq C_\varepsilon h^{1-\varepsilon} |u|_{2,\Omega}.$$

Consider now the solution $w \in V$ to

$$a(w, v) = (u - \widehat{\mathcal{I}}_h u_h, v), \quad \forall v \in V.$$

Observe

$$a(\widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h w) = a_h(u_h, \widehat{\mathcal{I}}_h w) - \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(u_h - \widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h w) = (f, \widehat{\mathcal{I}}_h w) - \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(u_h - \widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h w)$$

so that

$$a(u - \widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h w) = \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(u_h - \widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h w) = \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(e_h - \widehat{\mathcal{I}}_h e_h, \widehat{\mathcal{I}}_h w)$$

with $e_h{=}u_h - \widehat{\mathcal{I}}_h u$. Thus,

$$\begin{aligned}\|u - \widehat{\mathcal{I}}_h u_h\|_{0,\Omega}^2 &= a(u - \widehat{\mathcal{I}}_h u_h, w - \widehat{\mathcal{I}}_h w) + \sum_i a_{\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}(e_h - \widehat{\mathcal{I}}_h e_h, \widehat{\mathcal{I}}_h w) \\ &\leq Ch|u - \widehat{\mathcal{I}}_h u_h|_{1,\Omega}|w|_{2,\Omega} + \left(\sum_i |e_h - \widehat{\mathcal{I}}_h e_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}} \left(\sum_i |\widehat{\mathcal{I}}_h w|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}},\end{aligned} \tag{27}$$

where we have used the interpolation estimate. Using Lemma 3 and (above), we obtain

$$\left(\sum_i |e_h - \widehat{\mathcal{I}}_h e_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}} \leq C \,|||\, e_h|||| \leq C_\varepsilon h^{1-\varepsilon}|u|_{2,\Omega}$$

and

$$\begin{aligned}|u - \widehat{\mathcal{I}}_h u_h|_{1,\Omega} &\leq C\left(|u - u_h|_{1,\Omega^{nd}}^2 + |u_h - \widehat{\mathcal{I}}_h u_h|_{1,\Omega^{nd}}^2 + \sum_i |u - \widehat{\mathcal{I}}_h u_h|_{1,\mathcal{P}_i}^2\right)^{\frac{1}{2}} \\ &\leq C_\varepsilon h^{1-\varepsilon}|u|_{2,\Omega} + \left(\sum_i |e_h - \widehat{\mathcal{I}}_h e_h|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}} \leq C_\varepsilon h^{1-\varepsilon}|u|_{2,\Omega}.\end{aligned}$$

We also have by regularity of elliptic problem in a convex polygon (polyhedron)

$$|w|_{2,\Omega} \leq C\|u - \widehat{\mathcal{I}}_h u_h\|_{0,\Omega}$$

and by Lemma 5

$$\begin{aligned}\left(\sum_i |\widehat{\mathcal{I}}_h w|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}} \leq |w - \widehat{\mathcal{I}}_h w|_{1,\Omega} + \left(\sum_i |w|_{1,\widetilde{\mathcal{P}}_i \setminus \mathcal{P}_i}^2\right)^{\frac{1}{2}} &\leq Ch|w|_{2,\Omega} + C_\varepsilon h^{1-\varepsilon}\|w\|_{2,\Omega} \\ &\leq C_\varepsilon h^{1-\varepsilon}\|u - \widehat{\mathcal{I}}_h u_h\|_{0,\Omega}.\end{aligned}$$

Substituting into (27) gives

$$\|u - \widehat{\mathcal{I}}_h u_h\|_{0,\Omega}^2 \leq C_\varepsilon h^{2-2\varepsilon}|u|_{2,\Omega}\|u - \widehat{\mathcal{I}}_h u_h\|_{0,\Omega}$$

which in combination with the triangle inequality and the estimate for $\|u_h - \widehat{\mathcal{I}}_h u_h\|_{0,\mathcal{P}_i}$ contained in the estimate for $|||u - u_h|||$ gives the announced L^2-error estimate.

□

3.3 Conditioning of the system matrix

We are now going to prove that the conditioning number of the finite element matrix associated to the bilinear form a_h of the alternative scheme does not deteriorate in the presence of degenerate cells: it is of order $1/h^2$ if the mesh is quasi-uniform in a sense specified below.

Proposition 3 (Conditioning)**.** *Suppose that Assumption 3 holds. Then, the conditioning number $\kappa(\boldsymbol{A})$ of the matrix $\boldsymbol{A}$ associated to the bilinear form a_h in V_h satisfies*

$$\kappa(\boldsymbol{A}) \leq Ch^{-2}.$$

Before proving Proposition 3, we first introduce some auxiliary results:

Lemma 6 (Coercivity of a_h)**.** *Under the assumptions of Proposition 3, it holds for all $v_h \in V_h$*

$$a_h(v_h, v_h) \geqslant C\|v_h\|_{0,\Omega}^2.$$

Proof. Let $v_h \in V_h$. Observe, using triangle and Poincaré inequalities,

$$\begin{aligned}\|v_h\|_{0,\Omega} &\leq \|\widehat{\mathcal{I}}_h v_h\|_{0,\Omega} + \|v_h - \widehat{\mathcal{I}}_h v_h\|_{0,\Omega} \\ &\leq C|\widehat{\mathcal{I}}_h v_h|_{1,\Omega} + \|v_h - \widehat{\mathcal{I}}_h v_h\|_{0,\Omega} \\ &\leq C\left(|v_h|^2_{1,\Omega_h^{nd}} + \sum_i |\widehat{\mathcal{I}}_h v_h|^2_{1,\mathcal{P}_i} + \sum_i |v_h - \widehat{\mathcal{I}}_h v_h|^2_{1,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i} + \sum_i \|v_h - \widehat{\mathcal{I}}_h v_h\|^2_{0,\widetilde{\mathcal{P}}_i}\right)^{\frac{1}{2}}.\end{aligned}$$

The arguments similar to those in the proof of Lemma 4 entail

$$\sum_i |v_h - \widehat{\mathcal{I}}_h v_h|^2_{1,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i} + \sum_i \|v_h - \widehat{\mathcal{I}}_h v_h\|^2_{0,\widetilde{\mathcal{P}}_i\setminus\mathcal{P}_i} \leq C \,|||\, v_h \,|||^2 .$$

This implies $\|v_h\|_{0,\Omega} \leq C \,|||\, v_h|||$ which is equivalent to the desired result.

□

Lemma 7 (Continuity of a_h)**.** *Under the assumptions of Proposition 3, it holds for all $u_h, v_h \in V_h$*

$$a_h(u_h, v_h) \leq \frac{C}{h^2}\|u_h\|_{0,\Omega}\|v_h\|_{0,\Omega}.$$

Proof. Let $u_h, v_h \in V_h$. Since the cells of Ω_h^{nd} and the patches $\mathcal{P}_i$ are regular, we obtain using the inverse inequality

$$a_{\Omega_h^{nd}}(u_h, v_h) \leq \frac{C}{h^2}\|u_h\|_{0,\Omega_h^{nd}}\|v_h\|_{0,\Omega_h^{nd}}$$

and

$$a_{\mathcal{P}_i}(\widehat{\mathcal{I}}_h u_h, \widehat{\mathcal{I}}_h v_h) \leq \frac{C}{h^2}\|\widehat{\mathcal{I}}_h u_h\|_{0,\mathcal{P}_i}\|\widehat{\mathcal{I}}_h v_h\|_{0,\mathcal{P}_i}.$$

Using the equivalence of the norm in finite dimensional spaces and the fact that $\mathcal{P}_i$ and K_i^{nd} are regular, for all $w_h \in V_h$, it holds

$$\|\widehat{\mathcal{I}}_h w_h\|_{0,\mathcal{P}_i} \leq C\|w_h\|_{0,K_i^{nd}}.$$

The proof of such inequality is similar to the one used in the proof of Lemma 4. We deduce that

$$\|(\mathrm{Id} - \widehat{\mathcal{I}}_h)w_h\|_{0,\mathcal{P}_i} + \|\widehat{\mathcal{I}}_h w_h\|_{0,\mathcal{P}_i} \leq C(\|w_h\|_{0,\mathcal{P}_i} + \|w_h\|_{0,K_i^{nd}}) \leq C\|w_h\|_{0,\mathcal{P}_i}$$

which leads to the conclusion. □

Proof of Proposition 3. Using Assumption 3, there exists $C_1, C_2 > 0$ such that for all $w_h \in V_h$ and $\boldsymbol{w}$ its associated vector in $\mathbb{R}^N$

$$C_1 h^{n/2}|\boldsymbol{w}|_2 \leq \|w_h\|_0 \leq C_2 h^{n/2}|\boldsymbol{w}|_2. \tag{28}$$

Indeed, denoting by $\mathcal{N}_h$ the set of nodes of $\mathcal{T}_h$, by $\mathcal{N}_h(K)$ the set of nodes of a simplex $K \in \mathcal{T}_h$, and using $\sim$ to denote the equivalence with universal constant, as in (28), we can conclude

$$\|w_h\|_0^2 \sim \sum_{K\in\mathcal{T}_h} |K| \sum_{x\in\mathcal{N}_h(K)} |w_h(x)|^2 = \sum_{x\in\mathcal{N}_h} |w_h(x)|^2|\omega_x| \sim h^n|\boldsymbol{w}|_2^2.$$

In what follows, $\boldsymbol{v} \in \mathbb{R}^N$ denotes the vector associated to $v_h \in V_h$. Inequality (28) with Lemma 7 imply

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{v}\in\mathbb{R}^N} \frac{(\boldsymbol{A}\boldsymbol{v}, \boldsymbol{v})}{|\boldsymbol{v}|_2^2} = \sup_{\boldsymbol{v}\in\mathbb{R}^N} \frac{a_h(v_h, v_h)}{|\boldsymbol{v}|_2^2} \leq Ch^n \sup_{v_h\in V_h} \frac{a_h(v_h, v_h)}{\|v_h\|_0^2} \leq Ch^{n-2}.$$

Similarly, (28) with Lemma 6 imply

$$\|\boldsymbol{A}^{-1}\|_2 = \sup_{\boldsymbol{v}\in\mathbb{R}^N} \frac{|\boldsymbol{v}|_2^2}{(\boldsymbol{A}\boldsymbol{v}, \boldsymbol{v})} = \sup_{\boldsymbol{v}\in\mathbb{R}^N} \frac{|\boldsymbol{v}|_2^2}{a_h(v_h, v_h)} \leq Ch^{-n} \sup_{v_h\in V_h} \frac{\|v_h\|_0^2}{a_h(v_h, v_h)} \leq Ch^{-n}.$$

These estimates lead to the desired result. □

3.4 An equivalent, easily implementable variational formulation with interior penalty

Since implementing the interpolation operator $\widehat{\mathcal{I}}_h$ is not necessary trivial, we rewrite in this section the bilinear form a_h given in (14) in an equivalent form, which introduces the jumps of the gradients over the interior facets. The resulting method is similar to the ghost penalty from [9].

Lemma 8. *Under Assumption 3, suppose moreover that each patch $\mathcal{P}_i$ is composed of a non-degenerate cell K_i^{nd} and a degenerate cell K_i^{deg}. Denote by F_i the facet between K_i^{nd} and K_i^{deg}, as illustrated in Fig. 2.*

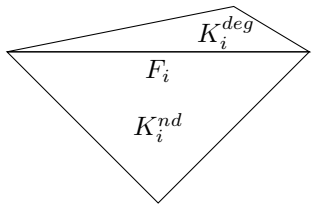

Figure 2: Example of patch $\mathcal{P}_i = K_i^{nd} \cup K_i^{deg}$.

Then, for all $u_h, v_h \in V_h$, it holds

$$a_h(u_h, v_h) = a_{\Omega_h^{nd}}(u_h, v_h) + \sum_i \frac{|\mathcal{P}_i|}{|K_i^{nd}|} a_{K_i^{nd}}(u_h, v_h) + \kappa_n \sum_i \frac{|K_i^{deg}|^3}{h_{\mathcal{P}_i}^2 |F_i|^2} [\nabla u_h]_{F_i} \cdot [\nabla v_h]_{F_i} \tag{29}$$

with $\kappa_n := \frac{2n^2}{(n+1)(n+2)}$.

Proof. Let us assume, without loss of generality, that the coordinate axes are chosen so that the y axis is orthogonal to F_i, as in Fig. 3. We also denote by h_i the height of the simplex K_i^{deg} drawn to the base F_i.

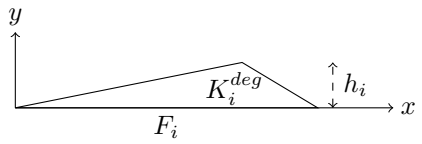

Figure 3: degenerate cell K_i^{deg}.

We first remark that, for all $u_h \in V_h$,

$$(\mathrm{Id} - \widehat{\mathcal{I}}_h) u_h = \begin{cases} [\nabla u_h]_{F_i} \, y & \text{on } K_i^{deg}, \\ 0 & \text{on } K_i^{nd}. \end{cases}$$

Hence, we deduce that

$$((\mathrm{Id} - \widehat{\mathcal{I}}_h) u_h, (\mathrm{Id} - \widehat{\mathcal{I}}_h) v_h)_{\mathcal{P}_i} = [\nabla u_h]_{F_i} \cdot [\nabla v_h]_{F_i} \int_{K_i^{deg}} y^2.$$

Moreover

$$\int_{K_i^{deg}} y^2 = \int_0^{h_i} y^2 |F_i| \left(1 - \frac{y}{h_i}\right)^{n-1} dy = |F_i| h_i^3 \frac{2}{n(n+1)(n+2)} = \frac{|K_i^{deg}|^3}{|F_i|^2} \frac{2n^2}{(n+1)(n+2)},$$

since $|K_i^{deg}| = \frac{1}{n} |F_i| h_i$. This leads to the conclusion.

□

4 Numerical simulations

In this section, we will illustrate with some numerical examples the sharpness of the *a priori* estimates of Theorem 2 and the efficiency of the method proposed in Section 3.4 to ensure the good conditioning of the matrix. The simulations of this section have been implemented using the finite element library FEniCS [1].

We consider problem (1) on the domain $\Omega := (0,1) \times (0,1)$ with the right hand side defined by $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$ so that the exact solution is given by $u(x,y) = \sin(\pi x)\sin(\pi y)$ for $(x,y) \in \Omega$. To construct the meshes $\mathcal{T}_h$ in all our numerical experiments presented below, we start from a uniform Cartesian mesh of step size h and degenerate certain cells so that for each degenerate cell K^{deg}, $h_{K^{deg}} = h$, $\rho_{K^{deg}} \sim h^2$ (more precisely, the distance between the longer side and the opposite node will be equal to h^2). In doing so, we take care that each degenerate cell be included in a patch of surrounding cells, and the patches corresponding to distinct degenerate cells do not intersect each other, cf. Fig. 4. Assumptions 1 and 3 are thus satisfied.

We report in Fig. 5 the numerical results obtained on a series of meshes with decreasing h, taking 10 degenerate cells (as described above) for every h. We use here the standard scheme (6) to produce the approximated solution u_h. The L^2 and H^1 absolute errors between u and u_h are given on the left in Fig. 5. The optimal convergence rates are indeed observed, as predicted by Theorem 2. However, the conditioning number of the associated finite element matrix is much bigger than $1/h^2$, which would be expected on a quasi-uniform mesh with step h. This is illustrated by Fig. 5, right. The estimate on the conditioning number from Proposition 2 is recovered, *i.e.* $\kappa(A) \sim 1/(h\varepsilon) \sim 1/h^3$, since $\varepsilon = \rho_{K^{deg}} \sim h^2$.

We now turn to the alternative scheme (15). We have implemented it using the reformulation (29). The results are reported in Fig. 6 using the same meshes containing 10 degenerate cells as above. The errors are reported on the left. We recall that Theorem 4 predicts the optimal convergence in the H^1 norm only if the approximate solution u_h is post-processed on the degenerate cells, by replacing the actual polynomial giving u_h on such a cell by the extension $\Pi_h u_h$ of u_h from the attached regular cell, cf. the definition of $|u - \Pi_h u_h|_{1,\Omega}$ in (16). Numerical experiments confirm the optimal H^1 convergence of the post-processed solution and also the necessity of such a post-processing. Indeed, the error with respect to the non-processed approximate solution $|u - u_h|_{1,\Omega}$ is not of optimal order h. It is also much bigger than $|u - \Pi_h u_h|_{1,\Omega}$. We also note that the optimal L^2 convergence is recovered without any post-processing, as predicted by Theorem 4. We recall that the introduction of the alternative scheme (15) was motivated by the desire to obtain less ill-conditioned matrices. The results in Fig. 6 (right) confirm that conditioning number for this scheme is indeed no longer affected by the presence of degenerate cells, in accordance with Proposition 3.

To illustrate the efficiency of the alternative scheme, we present in Figure 7 the number of iterations of the Conjugate Gradient algorithm (CG) [1] either without a preconditioner, or preconditioning by Block Jacobi iteration combined with the Incomplete LU factorization (bjacobi + ilu), on meshes containing 10 degenerate cells. In the case of the standard scheme, the number of iterations is increasing when ρ_{K_i}/h gets very small and the algorithm does not converge for ρ_{K_i}/h smaller than 10^{-8} in the case without preconditioner (10^{-13} in the case with preconditioner).

We recall that the theory of Section 3 concerning the alternative scheme (15) is developed under Assumption 3 supposing, in particular, that the number of degenerate cells is uniformly bounded. In the numerical experiments reported in Figs. 8 and 9, we wish to check if such an assumption is indeed necessary. We consider to this end a sequence of meshes constructed as above, but containing an increasing number of degenerate cells, cf. Fig. 8. We consider namely the densest packing of the degenerate cells allowed by Assumption 1 (the non-intersection of the surrounding patches), which gives approximately 5.5% of degenerate cells. Otherwise, the procedure for degenerating the cells is as above, in particular, $\rho_{K_i^{deg}} \approx h^2$. The results are presented in Fig. 9 both for the standard scheme on the left, and the alternative scheme (15) on the right. We first remark that the standard scheme remains optimally convergence in L^2 and H^1, in accordance with Theorem 2. On the contrary, the alternative scheme (15) does not converge. This observation highlights the sharpness of the results given in Theorem 4.

[1] The other simulations of the present work use Unsymmetric MultiFrontal sparse LU factorization (UMFPACK) as linear solver.

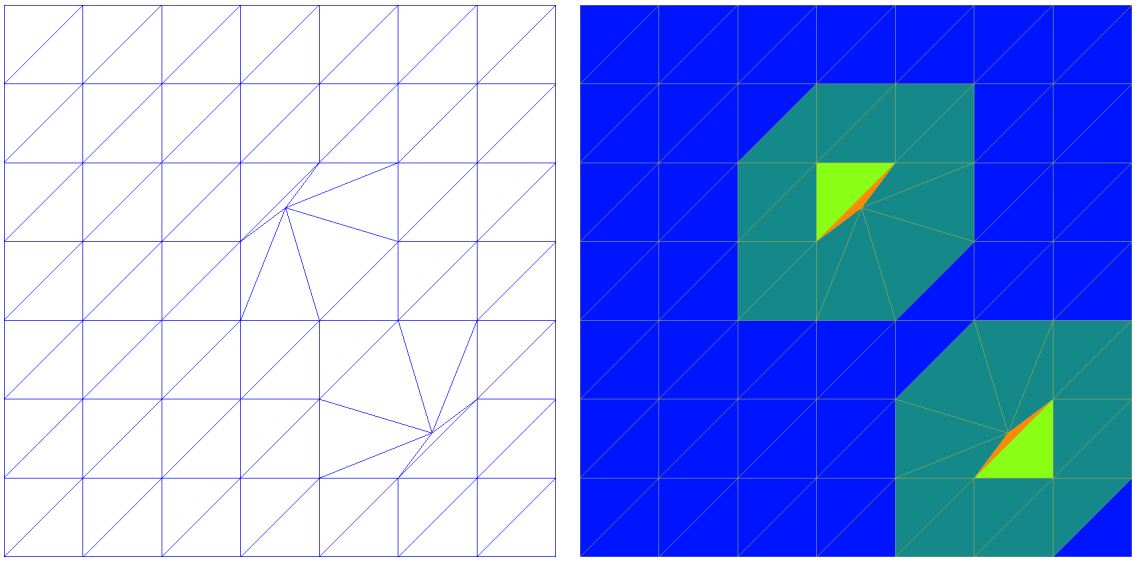

Figure 4: Example of a uniform mesh with 2 arbitrarily chosen degenerate cells K_i^{deg}, $\rho_{K_i^{deg}} \sim h^2$ (left). On the right, the degenerate cells are painted in red, the adjacent regular cells in light green, and the surrounding patches in dark green.

We conclude that both the standard scheme and the alternative scheme are optimally convergent and give very similar results if the number of degenerate cells remains bounded. As expected, the alternative scheme produces better-conditioned matrices. Moreover, if a conjugate gradient algorithm is used to solve the linear systems, it converges more rapidly when the discretization is obtained by the alternative scheme than by the standard one.

A proof of inverse inequalities in Lemma 2.3

Lemma 9. *Let $\mathcal{P}$ and $\mathcal{Q}$ be two bounded polytopes in $\mathbb{R}^n (n = 1, 2, 3)$ such that $\mathcal{P} \subset \mathcal{Q}$ and $\mathcal{P}$ contains a ball of radius $\geqslant c\,\mathrm{diam}(\mathcal{Q})$. There is a positive constant $C = C(c, n)$ such that for any polynomial v_h of degree ≤ 1*

$$\|v_h\|_{L^\infty(\mathcal{Q})} \leq C \|v_h\|_{L^\infty(\mathcal{P})}.$$

Proof. Denote by $h = \mathrm{diam}(\mathcal{Q})$ and let B_{in}, B_{out} be two balls such that $B_{\mathrm{in}} \subset \mathcal{P} \subset \mathcal{Q} \subset B_{\mathrm{out}}$. These balls can be chosen so that $\rho(B_{\mathrm{in}}) \geqslant ch$ and $\rho(B_{\mathrm{out}}) = 2h$ ($\rho(B)$ here denotes the diameter of the ball B). Let us first prove for any polynomial v_h of degree ≤ 1

$$\|v_h\|_{L^\infty(B_{\mathrm{out}})} \leq C \|v_h\|_{L^\infty(B_{\mathrm{in}})} \tag{30}$$

with a constant $C > 0$ (here and everywhere in this Appendix C denotes constants depending only on c and d). By the change of variables $x \mapsto (x - O)/2h$ where O is the center of B_{out}, this ball is transformed into the ball B_1 of radius 1 centered at the origin, and B_{in} into some ball $\widetilde{B} \subset B_1$ of radius $\geqslant c/2$, so that (30) is equivalent to

$$\|v_h\|_{L^\infty(B_1)} \leq C \|v_h\|_{L^\infty(\tilde{B})}, \quad \forall v_h \in \mathbb{P}_1.$$

Considering all the possible positions of the inscribed ball, the last inequality is a consequence of

$$\|v_h\|_{L^\infty(B_1)} \leq C \min_{\tilde{B} \subset B_1, \rho(\tilde{B}) \geq c/4} \|v_h\|_{L^\infty(\tilde{B}))}, \quad \forall v_h \in \mathbb{P}_1,$$

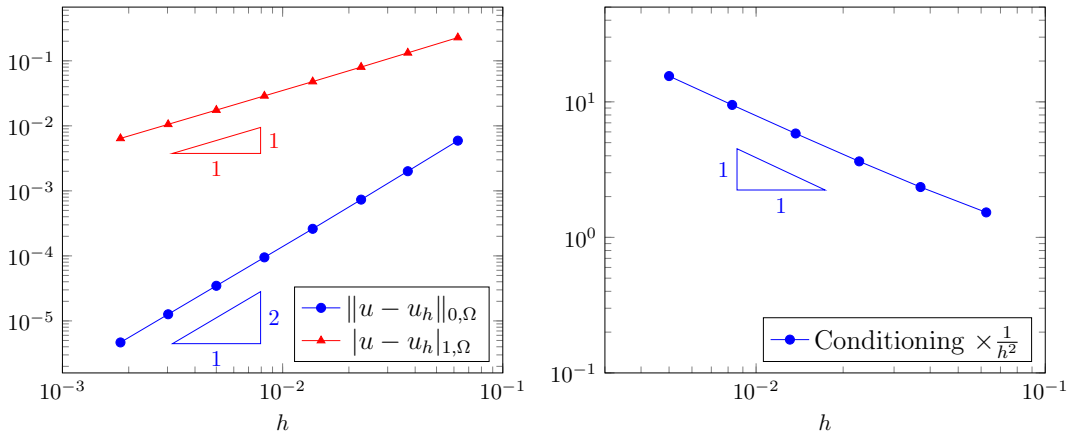

Figure 5: Errors (left) and conditioning (right) for the standard finite element scheme (6) on a sequence of meshes containing 10 degenerate cells with $\rho_{K_i^{deg}} \sim h^2$.

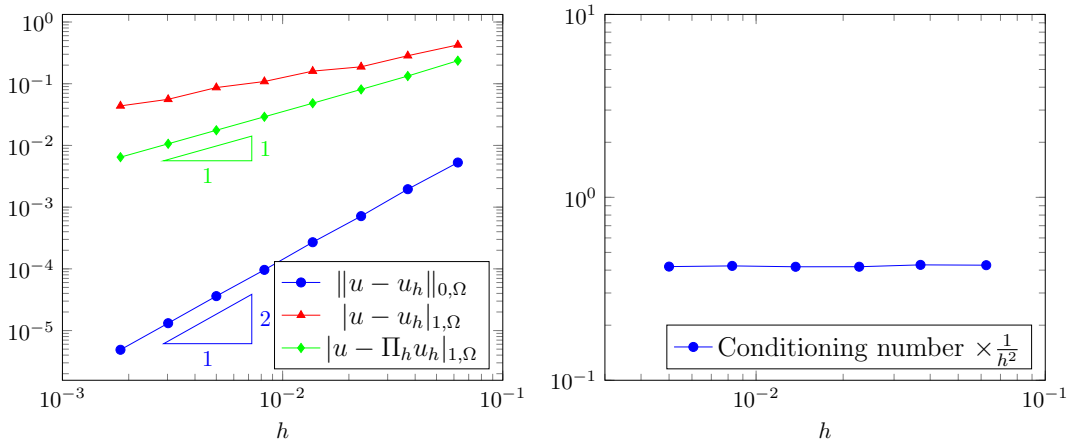

Figure 6: Errors (left) and conditioning number (right) for the alternative finite element scheme (15) on a sequence of meshes containing 10 degenerate cells with $\rho_{K_i^{deg}} \sim h^2$. The H^1 norm is calculated both using the approximate solution u_h directly and extending it to the degenerate cells from the adjacent regular cells, as in (16).

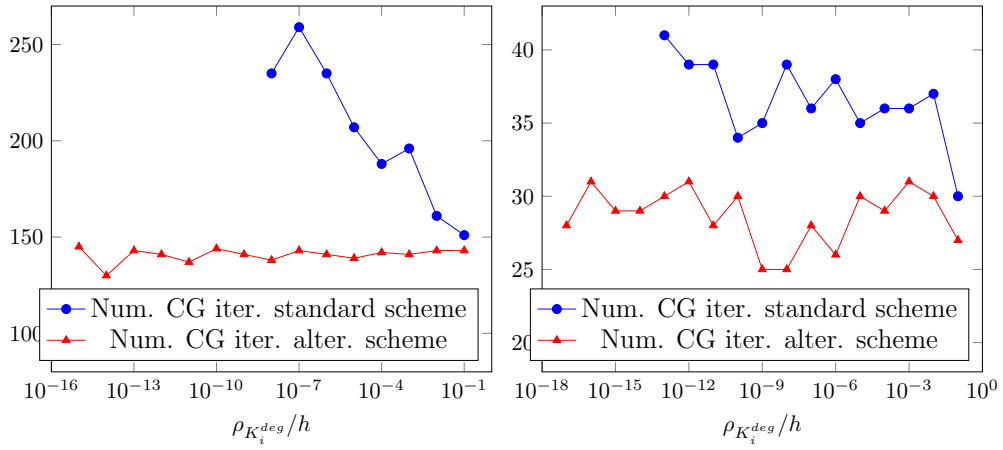

Figure 7: The number of iterations of the Conjugate Gradient algorithm (CG) for the standard finite element scheme (6) and the alternative finite element scheme (15) on a sequence of meshes with $h = 10^{-2}$ and containing 10 degenerate cells. Left: CG without preconditioner; Right: CG with (bjacobi + ilu) preconditioner. In both cases, the algorithms were assumed to converge when the absolute tolerance of 10^{-15} or the relative tolerance of 10^{-6} was achieved. The absence of certain points on the blue curve indicates that the algorithm did not converge on the corresponding meshes.

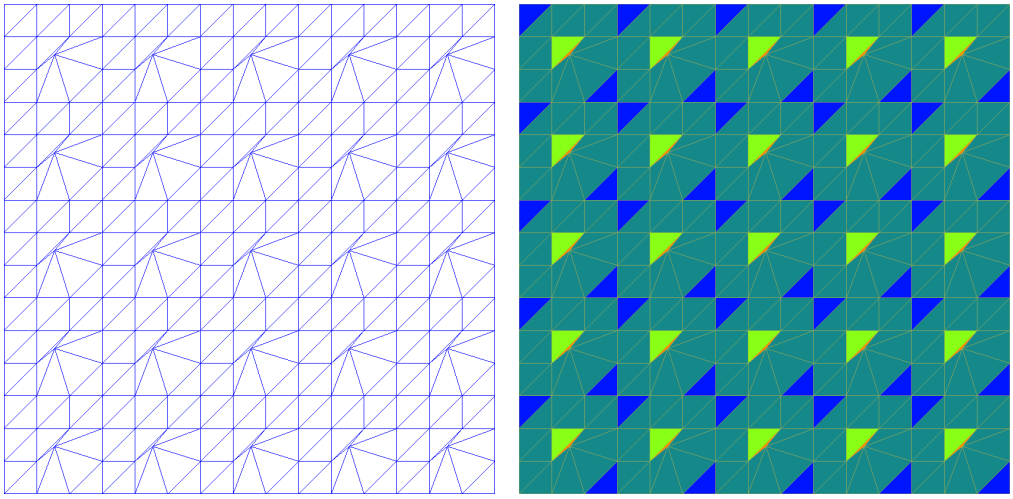

Figure 8: Example of a mesh with densely packed degenerate cells ($\approx 5.5\%$ of degenerate cells). Left: the mesh. Right: the disjoint patches surrounding the degenerate cells.

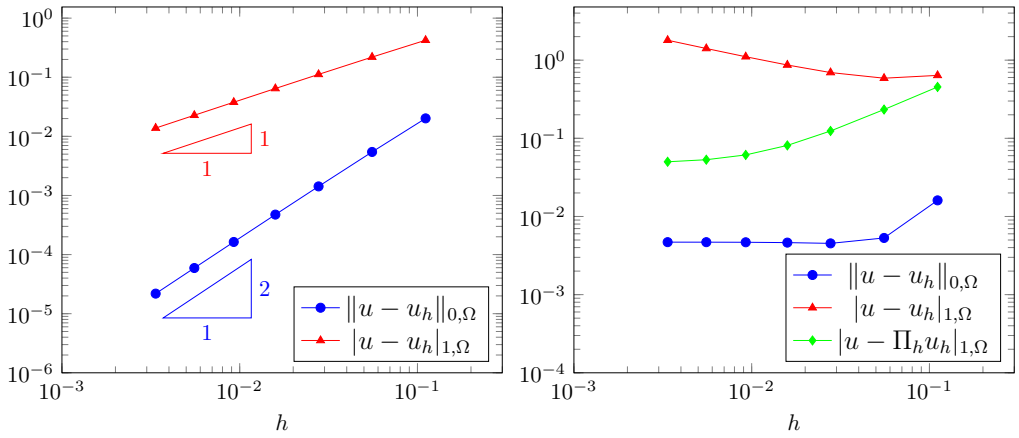

Figure 9: Errors on the meshes containing $\approx 5.5\%$ of degenerate cells. Left: the standard scheme (6). Right: alternative scheme (15).

which is true by equivalence of norms on the finite dimensional space $\mathbb{P}_1$. This proves (30) which entails in turn

$$\|v_h\|_{L^\infty(\mathcal{Q})} \leq C\|v_h\|_{L^\infty(B_{\text{out}})} \leq C\|v_h\|_{L^\infty(B_{\text{in}})} \leq C\|v_h\|_{L^\infty(\mathcal{P})}.$$

□

Lemma 10. *Let $\mathcal{P}$ be a bounded polytope in $\mathbb{R}^n (n = 1, 2, 3)$ of diameter $h_\mathcal{P}$ containing a ball of radius $\geqslant ch_\mathcal{P}$. There is a positive constant $C = C(c, n)$ such that for any polynomial v_h of degree ≤ 1*

$$\|\nabla v_h\|_{L^\infty(\mathcal{P})} \leq \frac{C}{h_\mathcal{P}}\|v_h\|_{L^\infty(\mathcal{P})}.$$

Proof. Let B_{in}, B_{out} be two balls such that $B_{\text{in}} \subset \mathcal{P} \subset B_{\text{out}}$ and $\rho(B_{\text{in}}) \geqslant ch$, $\rho(B_{\text{out}}) = 2h$. Similar to the proof of the preceding Lemma, we have for any polynomial v_h of degree ≤ 1

$$\|\nabla v_h\|_{L^\infty(B_{\text{out}})} \leq \frac{C}{h_\mathcal{P}}\|v_h\|_{L^\infty(B_{\text{in}})}.$$

This entails

$$\|\nabla v_h\|_{L^\infty(\mathcal{P})} \leq C\|\nabla v_h\|_{L^\infty(B_{\text{out}})} \leq \frac{C}{h_\mathcal{P}}\|v_h\|_{L^\infty(B_{\text{in}})} \leq \frac{C}{h_\mathcal{P}}\|v_h\|_{L^\infty(\mathcal{P})}.$$

□

Acknowledgements

The authors are thankful to Marek Bucki (TexiSense) and Franz Chouly (Université de Bourgogne Franche-Comté) for inspiring discussions which were at the origin of this project and for constant support during its realization.

References

[1] L. Anders, M. Kent-Andre, G. N. Wells, and al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[2] T. Apel. *Anisotropic finite elements: local estimates and applications*. Advances in Numerical Mathematics. B. G. Teubner, Stuttgart, 1999.

[3] I. Babuška and A. K. Aziz. On the angle condition in the finite element method. *SIAM J. Numer. Anal.*, 13(2):214–226, 1976.

[4] R. E. Barnhill and J. A. Gregory. Sard kernel theorems on triangular domains with application to finite element error bounds. *Numer. Math.*, 25(3):215–229, 1975/76.

[5] J. Brandts, S. Korotov, and M. Křížek. On the equivalence of regularity criteria for triangular and tetrahedral finite element partitions. *Comput. Math. Appl.*, 55(10):2227–2233, 2008.

[6] J. Brandts, S. Korotov, and M. Křížek. Generalization of the Zlámal condition for simplicial finite elements in $\mathbb{R}^d$. *Appl. Math.*, 56(4):417–424, 2011.

[7] S. Brenner and R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[8] M. Bucki, C. Lobos, and Y. Payan. A fast and robust patient specific finite element mesh registration technique: application to 60 clinical cases. *Medical image analysis*, 14(3):303–317, 2010.

[9] E. Burman. Ghost penalty. *C. R. Math. Acad. Sci. Paris*, 348(21-22):1217–1220, 2010.

[10] E. Burman, S. Claus, P. Hansbo, M. G. Larson, and A. Massing. CutFEM: discretizing geometry and partial differential equations. *Internat. J. Numer. Methods Engrg.*, 104(7):472–501, 2015.

[11] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[12] A. Hannukainen, S. Korotov, and M. Křížek. The maximum angle condition is not necessary for convergence of the finite element method. *Numer. Math.*, 120(1):79–88, 2012.

[13] J. Haslinger and Y. Renard. A new fictitious domain approach inspired by the extended finite element method. *SIAM J. Numer. Anal.*, 47(2):1474–1499, 2009.

[14] P. Jamet. Estimations d'erreur pour des éléments finis droits presque dégénérés. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér.*, 10(R-1):43–60, 1976.

[15] P. Jamet. Estimation of the interpolation error for quadrilateral finite elements which can degenerate into triangles. *SIAM J. Numer. Anal.*, 14(5):925–930, 1977.

[16] K. Kobayashi and T. Tsuchiya. On the circumradius condition for piecewise linear triangular elements. *Jpn. J. Ind. Appl. Math.*, 32(1):65–76, 2015.

[17] V. Kučera. On necessary and sufficient conditions for finite element convergence. *arXiv preprint arXiv:1601.02942*, 2016.

[18] M. Křížek. On the maximum angle condition for linear tetrahedral elements. *SIAM J. Numer. Anal.*, 29(2):513–520, 1992.

[19] J. Malý and W. Ziemer. *Fine regularity of solutions of elliptic partial differential equations*, volume 51 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1997.

[20] P. Oswald. Divergence of FEM: Babuška-Aziz triangulations revisited. *Appl. Math.*, 60(5):473–484, 2015.

[21] A. Zenísek. Convergence of the finite element method for boundary value problems of a system of elliptic equations. *Apl. Mat.*, 14:355–377, 1969.

[22] M. Zlámal. On the finite element method. *Numer. Math.*, 12:394–409, 1968.