An Arbitrarily High Order Unfitted Finite Element Method for Elliptic Interface Problems with Automatic Mesh Generation*

Zhiming Chen† Yong Liu‡

Abstract. We consider the reliable implementation of high-order unfitted finite element methods on Cartesian meshes with hanging nodes for elliptic interface problems. We construct a reliable algorithm to merge small interface elements with their surrounding elements to automatically generate the finite element mesh whose elements are large with respect to both domains. We propose new basis functions for the interface elements to control the growth of the condition number of the stiffness matrix in terms of the finite element approximation order, the number of elements of the mesh, and the interface deviation which quantifies the mesh resolution of the geometry of the interface. Numerical examples are presented to illustrate the competitive performance of the method.

Key words. Cell merging algorithm; unfitted finite element method; condition number

AMS classification. 65N50, 65N30

1 Introduction

Interface problems arise from diverse physical and engineering applications in which the coefficients of the governing partial differential equations are discontinuous across material interfaces that separate the physical domains. The body-fitted finite element methods resolve the geometry of the interface by requiring the vertices of the finite element mesh located on the interfaces [3, 23, 19]. For domains with complex geometry, the construction of body-fitted shape regular finite element meshes may be difficult and time-consuming, which is the main driving force of the study of unfitted finite element methods. In this paper we will show that the shape regular body-fitted mesh can indeed be constructed for any shaped smooth interface based on our new merging cell algorithm (see remarks below Theorem 3.1). We emphasize, however, that even when the body-fitted shape regular mesh is available, the construction of high-order finite element methods still requires substantial new ideas including, for example, the isoparametric finite element method [24, 34] or unfitted finite element methods which are the focus of this paper. We remark that the shape regularity

*This work is supported in part by China National Key Technologies R&D Program under the grant 2019YFA0709602 and China Natural Science Foundation under the grant 118311061,12288201.

†LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences and School of Mathematical Science, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: zmchen@lsec.cc.ac.cn

‡LSEC, Institute of Computational Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: yongliu@lsec.cc.ac.cn

assumption of the finite element mesh is not only fundamental in the mathematical theory of finite element methods (see, e.g., [24]) but also essential in controlling the condition number of the finite element stiffness matrix for elliptic equations (see, e.g. [10]).

Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain which is divided by a C^2-smooth interface Γ into two nonintersecting subdomains $\Omega_1 \subset \bar{\Omega}_1 \subset \Omega$, $\Omega_2 = \Omega \setminus \bar{\Omega}_1$, see Fig.1.1. We consider the following elliptic interface problem

$$-\operatorname{div}(a\nabla u) = f \quad \text{in } \Omega_1 \cup \Omega_2, \tag{1.1}$$

$$[\![u]\!]_\Gamma = 0,\ [\![a\nabla u \cdot n]\!]_\Gamma = 0 \quad \text{on } \Gamma, \quad u = g \quad \text{on } \partial\Omega, \tag{1.2}$$

where $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$, n is the unit outer normal to Ω_1, and $[\![v]\!] := v|_{\Omega_1} - v|_{\Omega_2}$ stands for the jump of a function v across the interface Γ. We assume that the coefficient $a(x)$ is positive and piecewise constant, namely, $a = a_1\chi_{\Omega_1} + a_2\chi_{\Omega_2}$, $a_1, a_2 > 0$, where χ_{Ω_i} denotes the characteristic function of Ω_i, $i = 1, 2$.

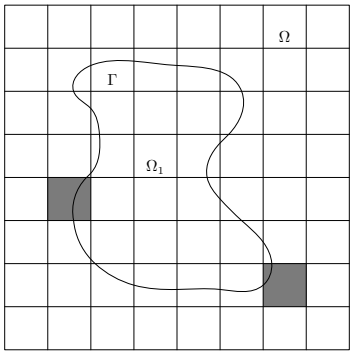


Figure 1.1: The setting of the elliptic interface problem and the unfitted mesh.

Unfitted finite element methods in the discontinuous Galerkin (DG) framework have attracted considerable interests in the literature in the last twenty years starting from the seminal work [31] in which an unfitted finite element method is proposed for elliptic interface problems. The method is defined on a fixed background mesh and uses different finite element functions in different cut cells which is the intersection of the mesh elements and physical domains. The jump condition on the interface is enforced by penalties which extends an earlier idea of Nitsche [41]. This unfitted finite element method can also be viewed as the interior penalty discontinuous Galerkin method (see, e.g., [2]) defined on meshes allowing curve-shaped elements. The main difficulty in using the unfitted finite element methods is the so-called *small cut cell problem*: the cut cells can be arbitrarily small and anisotropic, which can make the stiffness matrix extremely ill-conditioned, especially for high-order finite element methods [44, 8]. For other approaches to design unfitted discretization methods by constructing special finite element bases on interface elements or finite difference stencils along the interface, we refer to the immersed boundary method [43], the immersed interface method [35, 36], or the immersed finite element method [37, 22].

There are two approaches in the literature to attack the small cut cell problem. One is by appropriate techniques of stabilization [16, 17, 47, 38, 48, 30]. Among them, for example, the method of ghost penalty [16, 17, 30] adds additional penalties on the jumps of derivatives across sides or facets of interface elements. The other approach is by merging the small cut

cells with neighboring large elements [33, 32, 9, 21, 15] so that the merged macro-elements have enough support. While the DG formulation is still used in [33, 32, 21], the aggregated unfitted finite element method in [9] relies on the construction of stable extension operators so that the finite element space is still C^0. We refer to recent works [13, 18, 7, 8] for further information about ghost penalty and the aggregated unfitted finite element method.

In [21] an adaptive high-order unfitted finite element method is proposed for elliptic interface problems in which the *hp* a priori and a posteriori error estimates are derived based on novel *hp* domain inverse estimates and the concept of interface deviation. The interface deviation is a measure to quantify the mesh resolution of the geometry of the interface. We remark that the study on *hp* inverse estimates on curved domains is not only of mathematical interests, it is also essential to understand and control the exponential growth on the finite element approximation order p of the condition number of the stiffness matrix of the unfitted finite element method in this paper.

The macro-elements, which are the union of small interface elements and their surrounding elements, are assumed to be rectangular in [21]. This assumption is different from those in [33, 32, 9], see Fig.1.2. The macro-elements in [33, 32, 9] need not to be of rectangular shape, which makes the implementation simpler but the crucial inverse estimates on extended elements in [33, 32] or the stability of the extension operators [9] are shown without considering the dependence on the finite element approximation order p. The assumption that the macro-elements should be rectangular in [21] raises the question of how to construct the merging algorithm in practical applications.

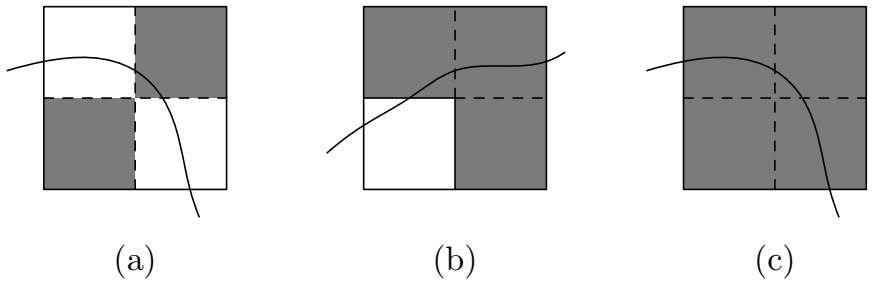


Figure 1.2: Three different ways of generating macro-elements which are marked in dark. The left, middle, and right figures illustrate the macro-elements used in [33, 15], [32, 9], and [21], respectively.

The first objective of this paper is to propose a reliable algorithm to merge small interface elements with their surrounding elements to generate the macro-elements. The algorithm is based on the concept of admissible chain of interface elements, the classification of patterns for merging elements, and appropriate ordering in generating macro-elements from the patterns so that the reliability of the algorithm in the sense that it terminates in finite number of steps can be proved. This algorithm also leads to a reliable algorithm of automatically generating 2D shape regular body-fitted finite element meshes for arbitrarily shaped smooth interfaces. To the authors' best knowledge, this algorithm introduces a new way to generate body-fitted finite element meshes and may be of independent interest.

The second objective of the paper is to study the condition number of the stiffness

matrix of high-order unfitted finite element methods which are known to be of the order $O(h^{-2})$ in the literature [16, 33, 32, 9, 6] on quasi-uniform meshes with the mesh size h. For high order methods, it is known [44] that the condition number of the stiffness matrix may grow exponentially with the finite element approximation order p in terms of the measure of cut cells. This indicates that the geometry of the cut cells is essential in controlling the condition number of the stiffness matrix.

In this paper, we will take the basis functions of the spectral element, that is, the Lagrangian interpolation functions at the Gauss-Lobatto points on elements not intersecting with the interface. For the interface elements, extra care must be taken as the basis of the spectral element on K is ill-conditioned on the subsets $K_i = K \cap \Omega_i$, $i = 1, 2$, which is similar to the observation in [27, P.346] for Legendre polynomials. Here we choose the L^2-orthogonal functions on some special polygons inside K_i, $i = 1, 2$, as the basis functions for the interface elements K. We show that the condition number of the stiffness matrix is bounded by $\Theta^2(p^3(N - N^\Gamma) + p^4 N^\Gamma)$ up to a logarithmic factor, where N is the number of total elements, N^Γ is the number of interface elements, and Θ depends on the interface deviation and p. This bound is optimal and indicates that the mesh has to sufficiently resolve the geometry of the interface to control the condition number of the stiffness matrix.

The results of this paper allow for extensions in several directions. Firstly, for the ease of exposition, we consider in this paper the case when the domain Ω is a union of rectangles and the interface is smooth. The extension to the general domains with smooth boundary is straightforward. Secondly, the case when the interface is piecewise smooth will be pursued in our forthcoming work by combining the ideas in [21] on large elements and interface deviation for interfaces with singularities with the merging algorithm developed in this paper. Thirdly, the theoretical results in this paper and in [21] including the hp domain inverse estimates and the concept of the interface deviation can be extended to study three-dimensional interface problems. The merging algorithm in the three-dimensional case is more challenging. Nevertheless, we believe that with the new insights gained in this paper for the two-dimensional case, reliable algorithms for constructing cubic macro-elements can be achieved in future. Finally, we remark that our argument to analyze and control the condition number of the stiffness matrix is fairly general, it can be used in other unfitted finite element methods including three-dimensional cases.

The layout of the paper is as follows. In section 2 we introduce our unfitted finite element method. In section 3 we construct the merging algorithm to generate the induced mesh. In section 4 we prove the discrete Poincaré inequality and the hp estimate for the condition number of the stiffness matrix. In section 5 we present several numerical examples to confirm our theoretical results.

2 The unfitted finite element method

Let $\Omega \subset \mathbb{R}^2$ be a domain which is a union of rectangles and $\mathcal{T}$ a Cartesian finite element mesh of Ω with possible hanging nodes. This allows us to locally refine the mesh near the interface to resolve the geometry to save the computational costs away from the interface. The elements of the mesh are (open) rectangles whose sides are parallel to the coordinate axes. We assume that the interface intersects the boundary of K twice at different sides (including the end points).

For any element K, let h_K stand for its diameter. Denote $\mathcal{T}^\Gamma := \{K \in \mathcal{T} : K \cap \Gamma \neq \emptyset\}$ the set of interface elements. We recall the definition of large element in Chen et al [21, Definition 2.1].

Definiton 2.1. *(Large element) For $i = 1, 2$, an element $K \in \mathcal{T}$ is called a large element with respect to Ω_i if $K \subset \Omega_i$ or $K \in \mathcal{T}^\Gamma$ for which there exists a constant $\delta_0 \in (0, 1/2)$ such that $|e \cap \Omega_i| \geq \delta_0 |e|$ for each side e of K having nonempty intersection with Ω_i. Specially, K is called a large element if $K \in \mathcal{T}^\Gamma$ is large with respect to both Ω_1 and Ω_2. Otherwise, K is called a small element.*

Note that it is possible that $K \in \mathcal{T}^\Gamma$ may not be a large element. The following assumption in [21] is inspired by Johansson and Larson [33] in which a fictitious boundary method is considered.

Assumption (H1) For each $K \in \mathcal{T}^\Gamma$, there exists a rectangular macro-element $M(K)$ which is a union of K and its surrounding element (or elements) such that $M(K)$ is a large element. We assume $h_{M(K)} \leq C_0 h_K$ for some constant $C_0 > 0$.

In section 3 we will construct a merging algorithm to find the macro-element for each small element in an admissible chain of interface elements. This indicates that the assumption (H1) can always be satisfied by using the algorithm. In the following, we will always set $M(K) = K$ if $K \in \mathcal{T}^\Gamma$ is a large element. Then, the induced mesh of $\mathcal{T}$ is defined as

$$\mathcal{M} = \{M(K) : K \in \mathcal{T}^\Gamma\} \cup \{K \in \mathcal{T} : K \not\subset M(K') \text{ for some } K' \in \mathcal{T}^\Gamma\}.$$

We will write $\mathcal{M} = \text{Induced}(\mathcal{T})$. Note that $\mathcal{M}$ is also a Cartesian mesh of Ω in the sense that either $M(K) \cap M(K') = \emptyset$ or $M(K) = M(K')$ for any two different elements $K, K' \in \mathcal{T}$. All elements in $\mathcal{M}$ are large elements.

For any $K \in \mathcal{M}^\Gamma := \{K \in \mathcal{M} : K \cap \Gamma \neq \emptyset\}$, denote $K_i = K \cap \Omega_i$, $i = 1, 2$, $\Gamma_K = \Gamma \cap K$, and Γ_K^h the open line segment connecting the two intersection points of Γ and ∂K. Γ_K^h divides the element K into two polygons K_1^h and K_2^h which are the polygonal approximation of K_1 and K_2, respectively. An important property of K being a large element is that K_i^h, $i = 1, 2$, is a *strongly* shape regular polygon in the sense that it is the union of shape regular triangles in the sense of Ciarlet [24]. We remark that there are different definitions of shape regular polygons in the literature, see, e.g., Ming and Shi [40] and Brenner and Sung [14].

The following concept of interface deviation is introduced in [21].

Definiton 2.2. *For any $K \in \mathcal{M}^\Gamma$, the interface deviation η_K is defined as $\eta_K = \max(\eta_K^1, \eta_K^2)$, where for $i = 1, 2$, if $A_K^i \in \Omega_i$ is the vertex of K which has the maximum distance to Γ_K^h among all vertices of K in Ω_i,*

$$\eta_K^i = \frac{\text{dist}_{\text{H}}(\Gamma_K, \Gamma_K^h)}{\text{dist}(A_K^i, \Gamma_K^h)}.$$

Here $\text{dist}_{\text{H}}(\Gamma_1, \Gamma_2) = \max_{x \in \Gamma_1}(\min_{y \in \Gamma_2} |x - y|)$ *and* $\text{dist}(A, \Gamma_1) = \min_{y \in \Gamma_1} |A - y|$.

The interface deviation is a measure on how well the mesh resolves the geometry of the interface. We will show in section 4 that this concept also links to the control of the condition number of the stiffness matrix.

It is known that if Γ_K is C^2-smooth, $\mathrm{dist}_{\mathrm{H}}(\Gamma_K, \Gamma_K^h) \le Ch_K^2$ (see, e.g., Feistauer [29, §3.3.2]) and thus $\eta_K \le Ch_K$ for some constant C independent of h_K. Therefore, the interface deviation can be made arbitrarily small by locally refining the mesh near the interface. When the interface Γ is Lipschitz and piecewise C^2-smooth, the definition of the large element and interface deviation has to be modified in the elements containing the singular points of the interface, see [21] for the details.

For any integer $p \ge 1$ and $K \in \mathcal{M}$, denote $Q_p(K)$ the set of polynomials in K which is of degree p in each variable. The following hp domain inverse estimate is proved in [21, Lemma 2.4].

Lemma 2.1. *Let Δ be a triangle with vertices $A = (a_1, a_2)^T$, $B = (0,0)^T$, $C = (c_1, 0)^T$, where $a_2, c_1 > 0$. Let $\delta \in (0, a_2)$ and $\Delta_\delta = \{x \in \Delta : \mathrm{dist}(x, BC) > \delta\}$. Then we have*

$$\|v\|_{L^2(\Delta)} \le \mathsf{T}\left(\frac{1+\delta a_2^{-1}}{1-\delta a_2^{-1}}\right)^{2p+3/2} \|v\|_{L^2(\Delta_\delta)} \quad \forall v \in Q_p(\Delta),$$

where $\mathsf{T}(t) = t + \sqrt{t^2-1}$ $\forall t \ge 1$.

The proof of this lemma makes use of the following one-dimensional domain inverse estimate in [21, Lemma 2.3]

$$\|g\|^2_{L^2(I_\lambda \setminus \bar{I})} \le \frac{1}{2}\left[(\lambda + \sqrt{\lambda^2-1})^{2p+1} - 1\right] \|g\|^2_{L^2(I)} \quad \forall g \in Q_p(I_\lambda), \tag{2.1}$$

where $I = (-1,1), I_\lambda = (-\lambda, \lambda)$, $\lambda > 1$, and $Q_p(I_\lambda)$ is the set of polynomials of order p in I_λ. We remark that the growing factor $(\lambda + \sqrt{\lambda^2-1})^{2p+1}$ is sharp which is attained by the Chebyshev polynomials whose explicit expression is $C_n(t) = \frac{1}{2}[(t+\sqrt{t^2-1})^n + (t - \sqrt{t^2-1})^n]$, $n \ge 0$, see DeVore and Lorentz [26, P.76].

Let $\delta_K := \mathrm{dist}_{\mathrm{H}}(\Gamma_K, \Gamma_K^h)$, We also define two polygons $K_i^{h-\delta_K}$, $i = 1, 2$, as follows. Let $\Gamma_{K_i}^{h-\delta_K} \subset K_i$ be the line segment which is parallel to Γ_K^h and its distance to Γ_K^h is δ_K. Let $K_i^{h-\delta_K}$ be the polygon bounded by sides of K and $\Gamma_{K_i}^{h-\delta_K}$.

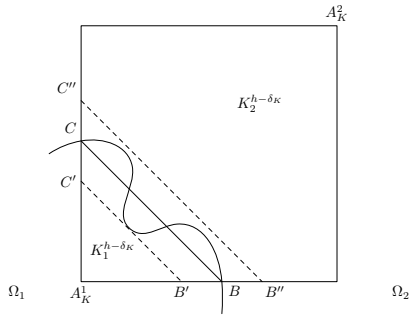


Figure 2.1: The figure used in the proof of Lemma 2.2.

Lemma 2.2. *Let $K \in \mathcal{M}^\Gamma$ and $\eta_K \le 1/2$, Then for $i = 1, 2$, we have*

$$\|v_i\|_{L^2(K_i^{h-\delta_K})} \le \|v_i\|_{L^2(K_i)} \le C\mathsf{T}\left(\frac{1+3\eta_K}{1-\eta_K}\right)^{2p+3/2} \|v_i\|_{L^2(K_i^{h-\delta_K})} \quad \forall v \in Q_p(K), \tag{2.2}$$

where the constant C is independent of h_K, p, and η_K.

Proof. The left inequality (2.2) is trivial since $K_i^{h-\delta_K} \subset K_i$. Here we prove the right inequality in (2.2) when Γ intersects ∂K at neighboring sides. The other cases can be proved similarly.

We use the notation in Fig.2.1 in which $B'C'$, $B''C''$ are parallel to Γ_K^h and the distances of $B'C', B''C''$ to Γ_K^h are δ_K. Then $K_1^{h-\delta_K} = \Delta A_K^1 B'C'$ and $K_2^{h-\delta_K}$ is the polygon bounded by sides of K and $B''C''$. Let $d_i = \mathrm{dist}(A_K^i, \Gamma_K^h)$, $i = 1, 2$. By definition, the interface deviation $\eta_K \geq \delta_K/d_i$, $i = 1, 2$. By Lemma 2.1, for any $v \in Q_p(K)$,

$$\begin{aligned} \|v\|_{L^2(K_1)} \leq \|v\|_{L^2(\Delta A_K^1 B''C'')} &\leq \mathsf{T}\left(\frac{1+2\delta/(d_1+\delta)}{1-2\delta/(d_1+\delta)}\right)^{2p+3/2} \|v\|_{L^2(\Delta A_K^1 B'C')} \\ &\leq \mathsf{T}\left(\frac{1+3\eta_K}{1-\eta_K}\right)^{2p+3/2} \|v\|_{L^2(K_1^{h-\delta_K})}. \end{aligned}$$

The case for K_2 can be proved similarly. This completes the proof. □

The numerical results in Example 1 in section 5 indicate that the bound in Lemma 2.2 is sharp. Now for any $K \in \mathcal{M}$, we denote

$$a_K = \begin{cases} \frac{a_1+a_2}{2} & \text{if } K \in \mathcal{M}^\Gamma, \\ a_i & \text{if } K \in \Omega_i. \end{cases}, \quad \Theta_K = \begin{cases} \mathsf{T}\left(\frac{1+3\eta_K}{1-\eta_K}\right)^{4p+3} & \text{if } K \in \mathcal{M}^\Gamma, \\ 1 & \text{otherwise.} \end{cases}$$

Based on the concept of interface deviation, the following *hp* inverse estimates on curved domains are proved in [21, Lemma 2.8, (2.12)].

Lemma 2.3. *Let $K \in \mathcal{M}^\Gamma$ and $\eta_K \leq 1/2$, Then for $i = 1, 2$, we have*

$$\begin{aligned} &\|\nabla v\|_{L^2(K_i)} \leq Cp^2 h_K^{-1} \Theta_K^{1/2} \|v\|_{L^2(K_i)} \quad \forall v \in Q_p(K), \\ &\|v\|_{L^2(\partial K_i)} \leq Cp h_K^{-1/2} \Theta_K^{1/2} \|v\|_{L^2(K_i)} \quad \forall v \in Q_p(K), \end{aligned}$$

where the constant C is independent of h_K, p, and η_K.

We remark that *hp* inverse estimates on star-shaped curve elements are studied in Massjung [38], Wu and Xiao [47], and Cangiani et al [20] which can be viewed as different forms of assumption on the mesh to resolve the geometry. Lemma 2.3 does not require the locally star-shaped assumption on the interface and is robust with respect to small variations of the interface as long as the interface deviation is the same.

Notice that if $\eta_K \leq \frac{1}{p(p+1)}$, for $s = \frac{1+3\eta_K}{1-\eta_K} = 1+\gamma_K$, where $\gamma_K = \frac{4\eta_K}{1-\eta_K} \leq 4p^{-2}$, we have $\mathsf{T}(s) = s + \sqrt{s^2-1} = 1+\rho_K$ with $\rho_K = \gamma_K + \sqrt{\gamma_K^2 + 2\gamma_K} \leq p^{-1}(4p^{-1} + \sqrt{16p^{-2}+8})$. Thus $\Theta_K = e^{(4p+3)\ln(\mathsf{T}(s))} \leq e^{(4p+3)\rho_K} \leq C$ for some constant C independent of p and η_K. This motivates us to make the following assumption in the remainder of this paper which can be easily satisfied for C^2-smooth interfaces if the mesh is locally refined near the interface.

Assumption (H2) For any $K \in \mathcal{M}^\Gamma$, $\eta_K \leq \frac{1}{p(p+1)}$.

Now we introduce some notation for DG methods. Let $\mathcal{E} = \mathcal{E}^{\mathrm{side}} \cup \mathcal{E}^\Gamma \cup \mathcal{E}^{\mathrm{bdy}}$, where $\mathcal{E}^{\mathrm{side}} = \{e = \partial K \cap \partial K' : K, K' \in \mathcal{M}\}$, $\mathcal{E}^\Gamma = \{\Gamma_K : K \in \mathcal{M}\}$, and $\mathcal{E}^{\mathrm{bdy}} = \{e = \partial K \cap \partial\Omega :$

$K \in \mathcal{M}\}$. For $i = 1, 2$, denote by $\mathcal{M}_i = \{K \in \mathcal{M} : K \cap \Omega_i \neq \emptyset\}$. Then $\Omega_i \subset \Omega_i^h = \cup\{K : K \in \mathcal{M}_i\}$. We denote $\mathcal{E}_i^{\text{side}}$ the set of all sides of $\mathcal{M}_i$ interior to Ω_i^h, that is, not on the boundary $\partial\Omega_i^h$. Finally, we set $\bar{\mathcal{E}} = \mathcal{E}_1^{\text{side}} \cup \mathcal{E}_2^{\text{side}} \cup \mathcal{E}^{\Gamma} \cup \mathcal{E}^{\text{bdy}}$.

For any $e \in \mathcal{E}$, we fix a unit normal vector n_e of e with the convention that n_e is the unit outer normal to $\partial\Omega$ if $e \in \mathcal{E}^{\text{bdy}}$ and n_e is the unit outer normal to $\partial\Omega_1$ if $e \in \mathcal{E}^{\Gamma}$. For any $v \in H^1(\mathcal{M}) := \{v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2} : v_i|_K \in H^1(K), K \in \mathcal{M}, i = 1, 2\}$, we define the jump of v across e as

$$[\![v]\!]_e := v_- - v_+ \quad \forall e \in \mathcal{E}^{\text{side}} \cup \mathcal{E}^{\Gamma}, \qquad [\![v]\!]_e := v_- \quad \forall e \in \mathcal{E}^{\text{bdy}},$$

where $v_\pm$ is the trace of v on e in the $\pm n_e$ direction. We define the normal vector function $n \in L^\infty(\mathcal{E})$ by $n|_e = n_e \quad \forall e \in \mathcal{E}$.

For any subset $\widehat{\mathcal{M}} \subset \mathcal{M}$ and $\hat{\mathcal{E}} \subset \bar{\mathcal{E}}$, we use the notation

$$(u, v)_{\widehat{\mathcal{M}}} := \sum_{K \in \widehat{\mathcal{M}}} (u, v)_K, \quad \langle u, v\rangle_{\hat{\mathcal{E}}} := \sum_{e \subset \hat{\mathcal{E}}} \langle u, v\rangle_e,$$

where $(u, v)_K$ is the inner product of $L^2(K)$ and $\langle u, v\rangle_e$ is the inner product of $L^2(e)$.

The unfitted finite element method is based on the idea of "doubling of unknowns" in Hansbo and Hansbo [31]. We define the unfitted finite element space as

$$\mathbb{X}_p(\mathcal{M}) = \{v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2} : v_i|_K \in Q_p(K), K \in \mathcal{M}, i = 1, 2\}.$$

For any $v \in H^1(\mathcal{M})$, we denote $\nabla_h v|_K := \nabla v_1\chi_{K_1} + \nabla v_2\chi_{K_2}$, where χ_{K_i} is the characteristic function of K_i, $i = 1, 2$. For any $v \in H^1(\mathcal{M}), g \in L^2(\partial\Omega)$, we define the liftings $\mathsf{L}(v) \in [\mathbb{X}_p(\mathcal{M})]^2$, $\mathsf{L}_1(g) \in [\mathbb{X}_p(\mathcal{M})]^2$ such that

$$(w, \mathsf{L}(v))_{\mathcal{M}} = \langle w^- \cdot n, [\![v]\!]\rangle_{\mathcal{E}}, \qquad (w, \mathsf{L}_1(g))_{\mathcal{M}} = \langle w \cdot n, g\rangle_{\mathcal{E}^{\text{bdy}}} \quad \forall w \in [\mathbb{X}_p(\mathcal{M})]^2 \tag{2.3}$$

Our unfitted finite element method is to find $U \in \mathbb{X}_p(\mathcal{M})$ such that

$$a_h(U, v) = F_h(v) \qquad \forall v \in \mathbb{X}_p(\mathcal{M}), \tag{2.4}$$

where the bilinear form $a_h : H^1(\mathcal{M}) \times H^1(\mathcal{M}) \to \mathbb{R}$, and the functional $F_h : H^1(\mathcal{M}) \to \mathbb{R}$ are given by

$$a_h(v, w) = (a(\nabla_h v - \mathsf{L}(v)), \nabla_h w - \mathsf{L}(w))_{\mathcal{M}} + \langle \alpha[\![v]\!], [\![w]\!]\rangle_{\bar{\mathcal{E}}} + \langle p^{-2}h\nabla_T[\![v]\!], \nabla_T[\![w]\!]\rangle_{\mathcal{E}^{\Gamma}}, \tag{2.5}$$

$$F_h(v) = (f, v)_{\mathcal{M}} - (a\mathsf{L}_1(g), \nabla_h v - \mathsf{L}(v))_{\mathcal{M}} + \langle \alpha g, v\rangle_{\mathcal{E}^{\text{bdy}}}, \tag{2.6}$$

where ∇_T is the surface gradient on Γ. For any $v = v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2}, w = w_1\chi_{\Omega_1} + w_2\chi_{\Omega_2} \in H^1(\mathcal{M})$,

$$\langle \alpha[\![v]\!], [\![w]\!]\rangle_{\bar{\mathcal{E}}} := \sum_{i=1}^{2} \langle \alpha[\![v_i]\!], [\![w_i]\!]\rangle_{\mathcal{E}_i^{\text{side}}} + \langle \alpha[\![v]\!], [\![w]\!]\rangle_{\mathcal{E}^{\Gamma} \cup \mathcal{E}^{\text{bdy}}}.$$

The interface penalty function $\alpha \in L^\infty(\mathcal{E})$ is

$$\alpha|_e = \alpha_0 a_e \Theta_e h_e^{-1} p^2 \quad \forall e \in \mathcal{E}, \tag{2.7}$$

where $\alpha_0 > 0$ is a fixed constant, $a_e = \max\{a_K : e \cap \bar{K} \neq \emptyset\}\ \forall e \in \mathcal{E}$, $\Theta_e = \max\{\Theta_K : e \cap \bar{K} \neq \emptyset\}\ \forall e \in \mathcal{E}$, and the mesh function $h|_e = (h_K + h_{K'})/2$ if $e = \partial K \cap \partial K' \in \mathcal{E}^{\rm side}$ and $h|_e = h_K$ if $e = K \cap \Gamma \in \mathcal{E}^{\Gamma}$ or $e = \partial K \cap \partial\Omega \in \mathcal{E}^{\rm bdy}$.

We remark that our unfitted finite element method (2.4) is the so-called local discontinuous Galerkin (LDG) method in Cockburn and Shu [25] which is different from the interior penalty discontinuous Galerkin (IPDG) method used in [31]. We choose the LDG method because the penalty constant α_0 in (2.7) can be any fixed constant, while the corresponding penalty constant in the IPDG method has to be sufficiently large to ensure the stability. We refer to Arnold et al [2] for a review of different DG methods for elliptic equations.

Notice that the last term $\langle p^{-2}h\nabla_T[\![v]\!], \nabla_T[\![w]\!]\rangle_{\mathcal{E}^\Gamma}$ in the bilinear form (2.5) is not present in [21]. It is included in this paper in order to show the discrete Poincaré inequality for unfitted finite element functions in Lemma 4.2 which is crucial for us to study the condition number of the stiffness matrix. We also remark that $\langle p^{-2}h\nabla_T[\![v]\!], \nabla_T[\![w]\!]\rangle_{\mathcal{E}^\Gamma}$ penalizes the tangential gradient of the finite element solution, not the normal flux of the solution as in Burman and Hansbo [16], Xiao and Wu [47].

For any $v \in H^2(\mathcal{M})$, we introduce the DG norm

$$\|v\|_{\rm DG}^2 := \|a^{1/2}\nabla v\|_{\mathcal{M}}^2 + \|\alpha^{1/2}[\![v]\!]\|_{\bar{\mathcal{E}}}^2 + \|p^{-1}h^{1/2}\nabla_T[\![v]\!]\|_{\mathcal{E}^\Gamma}^2,$$

where $\|\alpha^{1/2}[\![v]\!]\|_{\bar{\mathcal{E}}}^2 = \langle\alpha[\![v]\!], [\![v]\!]\rangle_{\bar{\mathcal{E}}}$ and $\|p^{-1}h^{1/2}\nabla_T[\![v]\!]\|_{\mathcal{E}^\Gamma}^2 = \langle p^{-2}h\nabla_T[\![v]\!], \nabla_T[\![v]\!]\rangle_{\mathcal{E}^\Gamma}$. By Lemma 2.3, it is easy to show that

$$a_h(v,v) \le C\|v\|_{\rm DG}^2 \quad \forall v \in X_p(\mathcal{M}).$$

Moreover, by [21, Theorem 2.1] we know that

$$a_h(v,v) \ge c_{\rm stab}\|v\|_{\rm DG}^2 \quad \forall v \in \mathbb{X}_p(\mathcal{M}),$$

where $c_{\rm stab} > 0$ is a constant independent of the mesh sizes, p, and the interface deviations η_K for all $K \in \mathcal{M}^\Gamma$.

Theorem 2.1. *Let the solution of the problem* (1.1)-(1.2) $u \in H^k(\Omega_1 \cup \Omega_2)$, $k \ge 2$, *and* $U \in \mathbb{X}_p(\mathcal{M})$ *be the solution of the problem* (2.4). *Then we have*

$$\|u - U\|_{\rm DG} \le C\Theta^{1/2}\frac{h^{\min(p+1,k)-1}}{p^{k-3/2}}\|u\|_{H^k(\Omega_1\cup\Omega_2)},$$

where $h = \max_{K\in\mathcal{M}} h_K$, $\Theta = \max_{K\in\mathcal{M}}\Theta_K$, *and the constant* C *is independent of the mesh sizes,* p, *and the interface deviations* η_K *for all* $K \in \mathcal{M}^\Gamma$.

Proof. For the sake of completeness, we sketch a proof by using the argument in e.g., Perugia and Schötzau [42], Wu and Xiao [47]. For $i = 1, 2$, let $\tilde{u}_i \in H^k(\mathbb{R}^2)$ be the Stein extension (cf., e.g., Adams and Fournier [1, Theorem 5.14]) of $u_i = u|_{\Omega_i} \in H^k(\Omega_i)$, which is available for any Lipschitz domains, such that $\|\tilde{u}_i\|_{H^k(\mathbb{R}^2)} \le C\|u_i\|_{H^k(\Omega_i)}$. Let $u_I = I_{hp}(\tilde{u}_1)\chi_{\Omega_1} + I_{hp}(\tilde{u}_2)\chi_{\Omega_2}$, where $I_{hp} : H^1(\mathcal{M}) \to \mathbb{V}_p(\mathcal{M}) = \Pi_{K\in\mathcal{M}}Q_p(K)$ is the interpolation operator defined in Babuška and Suri [5, Lemma 4.5]. For any $K \in \mathcal{M}$, it satisfies that for any $0 \le j \le k$,

$$\|w - I_{hp}(w)\|_{H^j(K)} \le C\frac{h_K^{\min(p+1,k)-j}}{p^{k-j}}\|v\|_{H^k(K)} \quad \forall v \in H^k(K), \tag{2.8}$$

where the constant C is independent of h_K, p, but may depend on k. By the multiplicative trace inequality, we have

$$\|w\|_{L^2(\partial K)} \le Ch_K^{-1/2}\|w\|_{L^2(K)} + C\|w\|_{L^2(K)}^{1/2}\|\nabla w\|_{L^2(K)}^{1/2} \quad \forall w \in H^1(K).$$

For any $K \in \mathcal{M}^\Gamma$, by Xiao et al [48, Lemma 3.1], [21, Lemma 2.6], we have that for $i = 1, 2$,

$$\|w\|_{L^2(\Gamma_K)} \le C\|w\|_{L^2(K_i)}^{1/2}\|\nabla w\|_{L^2(K_i)}^{1/2} + \|w\|_{L^2(\partial K_i \setminus \bar{\Gamma}_K)} \quad \forall w \in H^1(K). \tag{2.9}$$

Thus we obtain by using (2.8) that for any $K \in \mathcal{M}$, $j = 0, 1$,

$$\|w - I_{hp}(w)\|_{H^j(\partial K_i)} \le C\frac{h^{\min(p+1,k)-j-1/2}}{p^{k-j-1/2}}\|w\|_{H^k(K)} \quad \forall w \in H^k(K).$$

This implies easily that

$$\|u - u_I\|_{\mathrm{DG}} \le C\Theta^{1/2}\frac{h^{\min(p+1,k)-1}}{p^{k-3/2}}\|u\|_{H^k(\Omega_1 \cup \Omega_2)}. \tag{2.10}$$

On the other hand, since $a_h(u, v) = F_h(v) \quad \forall v \in \mathbb{X}_p(\mathcal{M})$, we use (2.4) to conclude that

$$\begin{aligned} \|u_I - U\|_{\mathrm{DG}}^2 \le c_{\mathrm{stab}}^{-1} a_h(u_I - U, u_I - U) &= c_{\mathrm{stab}}^{-1} a_h(u_I - u, u_I - U) \\ &\le C\|u_I - u\|_{\mathrm{DG}}\|u_I - U\|_{\mathrm{DG}}. \end{aligned}$$

This completes the proof by (2.10) and the triangle inequality. □

To conclude this section, we remark that the same a posteriori error estimate in [21, Theorem 3.1] also holds for the solution $U \in \mathbb{X}_p(\mathcal{M})$ in (2.4). Here we omit the details.

3 The merging algorithm

In this section, we construct a merging algorithm for the admissible chain of interface elements so that each small interface element in the chain is included in some macro-element which is a large element. We first introduce the concept of admissible chain in §3.1 and five types of patterns of merging small interface elements with their surrounding elements in §3.2. We propose our merging algorithm and prove its reliability in §3.3.

3.1 The admissible chain of interface elements

A chain of interface elements $\mathfrak{C} = \{G_1 \to G_2 \to \cdots \to G_n\}$ orderly consists of n interface elements $G_i \in \mathcal{T}^\Gamma$, $i = 1, \cdots, n$, such that $\bar{\Gamma}_{G_i} \cup \bar{\Gamma}_{G_{i+1}}$ is a continuous curve, $1 \le i \le n-1$. We call n the length of $\mathfrak{C}$ and denote $\mathfrak{C}\{i\} = G_i$, $i = 1, \cdots, n$.

For any element $K \in \mathcal{T}$, we call $N(K) \in \mathcal{T}$ a neighboring element of K if K and $N(K)$ share a common side, and $D(K) \in \mathcal{T}$ a diagonal element of K if K and $D(K)$ only share one common vertex. Set $\mathcal{S}(K)_0 = \{K\}$, and for $j \ge 1$, denote $\mathcal{S}(K)_j = \{K'' \in \mathcal{T} : \exists\, K' \in \mathcal{S}(K)_{j-1}$ such that $\bar{K}'' \cap \bar{K}' \ne \emptyset\}$, that is, $\mathcal{S}(K)_j$ is the set of all k-th layer elements surrounding K, $0 \le k \le j$. Obviously, $\mathcal{S}(K)_0 \subset \mathcal{S}(K)_1 \subset \cdots \subset \mathcal{S}(K)_j$ for any $j \ge 1$.

Definiton 3.1. *A chain of interface elements $\mathfrak{C}$ is called admissible if the following rules are satisfied.*

1. *For any $K \in \mathfrak{C}$, all elements in $\mathcal{S}(K)_2$ have the same size as that of K.*
2. *If $K \in \mathfrak{C}$ has a side e such that $\bar{e} \subset \Omega_i$, then e must be a side of some neighboring element $N(K) \subset \Omega_i$, $i = 1, 2$.*
3. *Any elements $K \in \mathcal{T} \backslash \mathcal{T}^\Gamma$ can be neighboring at most two elements in $\mathfrak{C}$.*
4. *For any $K \subset \Omega_i$, the interface elements in $\mathcal{S}(K)_j$, $j = 1, 2$, must be connected in the sense that the interior of the closed set $\cup\{\bar{G} : G \in \mathcal{S}(K)_j \cap \mathcal{E}^\Gamma\}$ is a connected domain.*

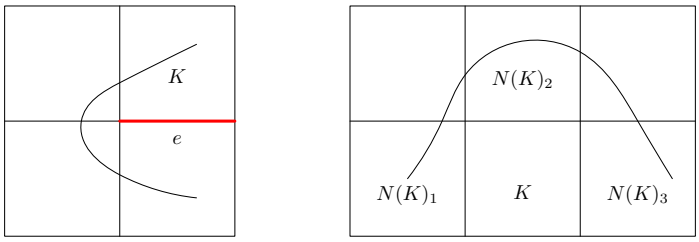


Figure 3.1: The patch of elements not allowed by Rule 2 (left) and Rule 3 (right) in Definition 3.1.

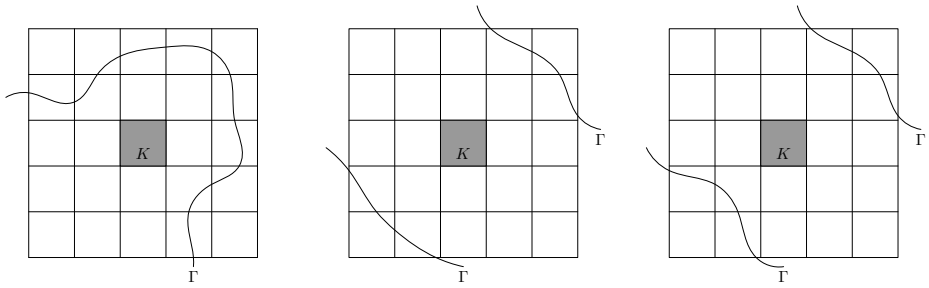


Figure 3.2: The patch of elements not allowed by Rule 4 in Definition 3.1.

We remark that the four rules of the admissible chains can be easily satisfied if the mesh is well refined near the interface. The purpose of Rules 2 and 3 is to exclude the situations illustrated in Fig.3.1, in which refinements are required to resolve the geometry of the interface. By the Rule 4, the three cases illustrated in Fig.3.2 are not allowed since the interface elements in $\mathcal{S}(K)_1$ in the left figure and in $\mathcal{S}(K)_2$ in the middle and right figures are not connected, where K is the dark element. We notice that the interface elements in $\mathcal{S}(K)_2$ in the left figure of Fig.3.2 is however connected.

3.2 The patterns

Since the interface intersects the boundary of K twice at different sides (including the end points), the interface intersects any element only in four possible ways as shown in

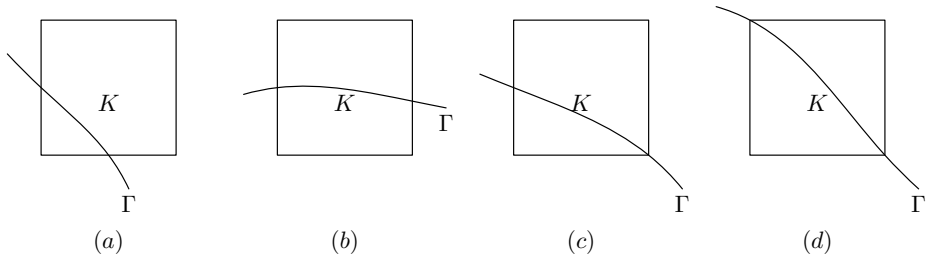


Figure 3.3: Different types of interface elements. The type 2 elements include elements illustrated in (b) and (c).

Fig.3.3. We denote $\mathcal{T}_1$ the set of interface elements shown in Fig.3.3(a), $\mathcal{T}_2$ the set of interface elements shown in Fig.3.3(b) and (c), and $\mathcal{T}_3$ the set of interface elements shown in Fig.3.3(d). By Definition 2.1, each element in $\mathcal{T}_3$ is a large element. Thus we only need to consider the merging of type $\mathcal{T}_1$ and $\mathcal{T}_2$ elements.

A pattern is a set of interface elements and their neighboring and diagonal elements whose union consists of a macro-element. We introduce five types of patterns according to the combination of different types of interface elements, which will be used in our merging algorithm for the admissible chain of interface elements. In the following, for any $K \in \mathcal{T}$, $h_i(K)$ stands for its length of the side of K which is parallel to the x_i-axis, $i = 1, 2$.

Pattern 1: $K \in \mathcal{T}_1$ has two neighboring elements $N(K)_1, N(K)_2 \in \mathcal{T}_2$, see Fig.3.4. e_1 and e_2 are respectively the thick part of the sides of $N(K)_1$ and $N(K)_2$ in the figure. We use Algorithm 1 to obtain the macro-elements $M(K)$, $M(N(K)_1)$, and $M(N(K)_2)$. Here for any closed set $T \subset \mathbb{R}^2$, T° stands for the interior of T.

Algorithm 1: Pattern 1

Input: $(N(K)_1, K, N(K)_2)$
Output: $(M(N(K)_1), M(K), M(N(K)_2))$
if K, $N(K)_1$, and $N(K)_2$ are large elements **then**
$M(N(K)_1) = N(K)_1$, $M(K) = K$, $M(N(K)_2) = N(K)_2$;
else
if $|e_1|/h_2(K) \geq 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)})^\circ$;
else if $|e_1|/h_2(K) \geq 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup \overline{G}_4 \cup \overline{G}_5)^\circ$;
else if $|e_1|/h_2(K) < 2\delta_0$ and $|e_2|/h_1(K) \geq 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup \overline{G}_1 \cup \overline{G}_2)^\circ$;
else if $|e_1|/h_2(K) < 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup (\cup_{j=1}^5 \overline{G}_j))^\circ$.
end

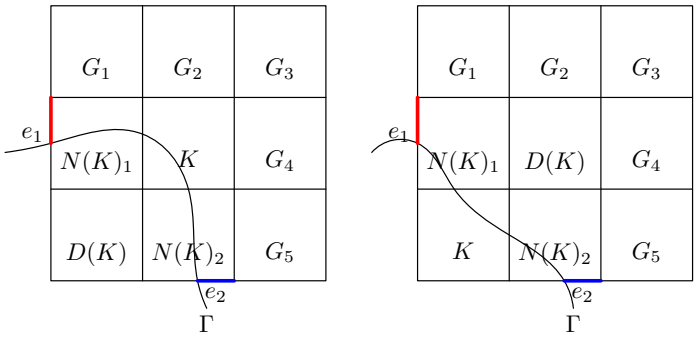


Figure 3.4: Illustration of type 1 (left) and type 2 (right) patterns.

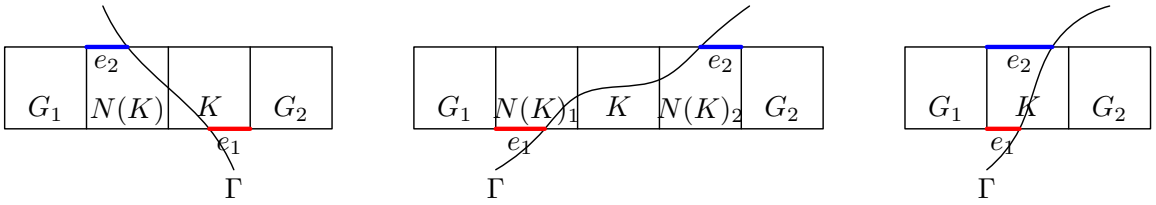


Figure 3.5: Illustration of type 3 (left), type 4 (middle) and type 5 (right) patterns.

end

Lemma 3.1. *Let $\delta_0 \in (0, 1/3]$. The macro-elements $M(K)$, $M(N(K)_1)$, $M(N(K)_2)$ of the output of Algorithm 1 are large elements.*

Proof. We only prove $M(K)$ is a large element when $|e_1|/h_2(K) < 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$. The other cases can be proved analogously. Since $\delta_0 \in (0, 1/3]$, we have

$$\frac{|e_1| + h_2(K)}{3\,h_2(K)} \geq \frac{1}{3} \geq \delta_0, \quad \frac{2h_2(K) - |e_1|}{3\,h_2(K)} \geq \frac{1}{3} \geq \delta_0.$$

Similar inequalities hold for $|e_2|$. Thus $|e \cap \Omega_i| \geq \delta_0 |e|$ for each side e of $M(K)$ having nonempty intersection with Ω_i, $i = 1, 2$. This implies that $M(K)$ is a large element. □

Pattern 2: $K \in \mathcal{T}_1$ has two neighboring elements $N(K)_1, N(K)_2 \in \mathcal{T}_1$, see Fig.3.4. e_1 and e_2 are respectively the thick part of the side of $N(K)_1$ and $N(K)_2$ in the figure. We use Algorithm 2 to obtain $M(K)$, $M(N(K)_1)$, and $M(N(K)_2)$.

Algorithm 2: Pattern 2

Input: $(N(K)_1, K, N(K)_2)$
Output: $(M(N(K)_1), M(K), M(N(K)_2))$
if $K, N(K)_1$, and $N(K)_2$ are large elements **then**
  let $M(N(K)_1) = N(K)_1$, $M(K) = K$, $M(N(K)_2) = N(K)_2$;
else
  if $|e_1|/h_2(K) \geq 2\delta_0$ and $|e_2|/h_1(K) \geq 2\delta_0$ **then**

let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)})^\circ$;
else if $|e_1|/h_2(K) \geq 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup \overline{G}_4 \cup \overline{G}_5)^\circ$;
else if $|e_1|/h_2(K) < 2\delta_0$ and $|e_2|/h_1(K) \geq 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup \overline{G}_1 \cup \overline{G}_2)^\circ$;
else if $|e_1|/h_2(K) < 2\delta_0$ and $|e_2|/h_1(K) < 2\delta_0$ **then**
let $M(K) = M(N(K)_1) = M(N(K)_2) = (\overline{K} \cup \overline{N(K)}_1 \cup \overline{N(K)}_2 \cup \overline{D(K)} \cup (\cup_{j=1}^5 \overline{G}_j))^\circ$.
end
end

Pattern 3: $K \in \mathcal{T}_1$ has one neighboring element $N(K) \in \mathcal{T}_1$, see Fig. 3.5. e_1 and e_2 are respectively the thick part of the side of K and $N(K)$ in the figure. We use Algorithm 3 to obtain $M(K)$, $M(N(K))$.

Algorithm 3: Pattern 3

Input: $(K, N(K))$
Output: $(M(K), M(N(K)))$
if K and $N(K)$ are both large elements **then**
let $M(K) = K$, $M(N(K)) = N(K)$
else
if $|e_1|/h_1(K) \geq 2\delta_0$ and $|e_2|/h_1(K) \geq 2\delta_0$ **then**
let $M(K) = M(N(K)) = (\overline{K} \cup \overline{N(K)})^\circ$;
else if $|e_1|/h_1(K) \geq 3\delta_0$ **then**
let $M(K) = M(N(K)) = (\overline{K} \cup \overline{N(K)} \cup \overline{G}_1)^\circ$;
else if $|e_2|/h_1(K) \geq 3\delta_0$ **then**
let $M(K) = M(N(K)) = (\overline{K} \cup \overline{N(K)} \cup \overline{G}_2)^\circ$;
else
let $M(K) = M(N(K)) = (\overline{K} \cup \overline{N(K)} \cup \overline{G}_1 \cup \overline{G}_2)^\circ$.
end
end

Pattern 4: $K \in \mathcal{T}_2$ has two neighboring elements $N(K)_1, N(K)_2 \in \mathcal{T}_1$, see Fig. 3.5. e_1 and e_2 are respectively the thick part of the side of $N(K)_1$ and $N(K)_2$ in the figure. We use Algorithm 4 to obtain $M(K)$, $M(N(K)_1)$, $M(N(K)_2)$.

Algorithm 4: Pattern 4

Input: $(N(K)_1, K, N(K)_2)$
Output: $(M(N(K)_1), M(K), M(N(K)_2))$
if K, $N(K)_1$, and $N(K)_2$ are all large elements **then**

let $M(K) = K$, $M(N(K)_1) = N(K)_1$, $M(N(K)_2) = N(K)_2$;
else
if $|e_1|/h_1(K) \geq 3\delta_0$ and $|e_2|/h_1(K) \geq 3\delta_0$ **then**
let $M(N(K)_1) = M(K) = M(N(K)_2) = (\overline{N(K)}_1 \cup \overline{K} \cup \overline{N(K)}_2)^\circ$;
else if $|e_2|/h_1(K) \geq 4\delta_0$ **then**
let $M(N(K)_1) = M(K) = M(N(K)_2) = (\overline{N(K)}_1 \cup \overline{K} \cup \overline{N(K)}_2 \cup \overline{G}_1)^\circ$;
else if $|e_1|/h_1(K) \geq 4\delta_0$ **then**
let $M(N(K)_1) = M(K) = M(N(K)_2) = (\overline{N(K)}_1 \cup \overline{K} \cup \overline{N(K)}_2 \cup \overline{G}_2)^\circ$;
else
let $M(N(K)_1) = M(K) = M(N(K)_2) = (\overline{N(K)}_1 \cup \overline{K} \cup \overline{N(K)}_2 \cup \overline{G}_1 \cup \overline{G}_2)^\circ$.
end
end

Pattern 5: $K \in \mathcal{T}_2$, see Figure 3.5. e_1 and e_2 are respectively the thick part of the sides of K in the figure. We use Algorithm 5 to obtain $M(K)$.

Algorithm 5: Pattern 5

Input: K
Output: $M(K)$
if K is a large element **then**
let $M(K) = K$;
else
if $|e_1|/h_1(K) < 1 - 2\delta_0$ and $|e_2|/h_1(K) < 1 - 2\delta_0$ **then**
let $M(K) = (\overline{K} \cup \overline{G}_1)^\circ$;
else if $|e_1|/h_1(K) \geq 2\delta_0$ and $|e_2|/h_1(K) \geq 2\delta_0$ **then**
let $M(K) = (\overline{K} \cup \overline{G}_2)^\circ$;
else
let $M(K) = (\overline{K} \cup \overline{G}_1 \cup \overline{G}_2)^\circ$.
end
end

The following lemma can be proved by the same argument as that in Lemma 3.1. Here we omit the details.

Lemma 3.2. *The output macro-elements of Algorithm 2, Algorithm 3, Algorithm 4, and Algorithm 5 are large elements if* $\delta_0 \in (0, 1/3]$, $\delta_0 \in (0, 1/4]$, $\delta_0 \in (0, 1/5]$, *and* $\delta_0 \in (0, 1/3]$, *respectively.*

To conclude this subsection, we make the following observations which can be easily checked from the construction of the patterns.

Remark 3.1. *Only elements in* $\{\mathcal{S}(K)_2 : K \in \mathcal{T}^\Gamma\}$ *can be possibly merged with small interface elements. The elements two layers away from the interface will not be touched in the merging algorithm.*

Remark 3.2. *An element $G \in \mathcal{T}_2$ is merged with some element $K \in \mathcal{T}_1$ if and only if there exists an element $G' \in \mathcal{T}_2$ such that G, K, G' form a pattern of type 1 or there exists an element $G' \in \mathcal{T}_1$ such that G, K, G' form a pattern of type 4.*

Remark 3.3. *An element $G \subset \Omega_i$, $i = 1, 2$, is merged with some element $K \in \mathcal{T}_1$ such that K and G has only one common vertex, then G, K, and two neighboring elements of K are in the same pattern of type 1 or type 2.*

Remark 3.4. *If an element $G \subset \Omega_i$, $i = 1, 2$, is merged with some element $K \in \mathcal{T}^{\Gamma}$ such that K and G has only one common vertex, then K and G are in the same pattern of type 1 or type 2. If K and G are in the same pattern of type 2, then G can be any one of the elements $G_2, G_4, D(K)$ which has only one common vertex with some interface element in Fig.3.4 (right). Since the interface elements in $\mathcal{S}(G)_1$ must be connected by the rule 4 of the admissible chain, G cannot have any neighboring element in $\mathcal{T}_2$. Thus, if G has a neighboring element $N(G) \in \mathcal{T}_2$ which is neighboring to K, then $K, N(G), G$ are in the same pattern of type 1, which implies, in particular, that $N(G)$ is merged with K.*

3.3 The merging algorithm

Let $\mathfrak{C}$ be an admissible chain of interface elements. The following algorithm constructs a locally induced mesh from $\mathfrak{C}$ which consists of the large interface elements of $\mathfrak{C}$ and macro-elements including all small elements of $\mathfrak{C}$ so that the elements in the induced mesh are all large elements.

Algorithm 6: The merging algorithm for the admissible chain of interface elements

---

Input: The admissible chain $\mathfrak{C}$
Output: The induced mesh Induced($\mathfrak{C}$)
1° Find all subchains $\mathfrak{S}$ of length $n \geq 2$ of $\mathfrak{C}$ such that $\mathfrak{S}\{i\} \in \mathcal{T}_1$, $i = 1, \ldots, n$;
if $n = 2k + 1$ is odd **then**
  for $i = 1, 2, \ldots, k - 1$ **do**
    call the Algorithm 3 with the input $(\mathfrak{S}\{2i\}, \mathfrak{S}\{2i + 1\})$;
  end
call the Algorithm 2 with the input $(\mathfrak{S}\{2k - 1\}, \mathfrak{S}\{2k\}, \mathfrak{S}\{2k + 1\})$
else if $n = 2k$ is even **then**
  for $i = 1, 2, \ldots, k$ **do**
    call the Algorithm 3 with the input $(\mathfrak{S}\{2i - 1\}, \mathfrak{S}\{2i\})$;
  end
end
2° Find all subchains $\mathfrak{S}$ of length $n = 3$ in the remaining interface elements such that $\mathfrak{S}\{1\} \in \mathcal{T}_1$, $\mathfrak{S}\{2\} \in \mathcal{T}_2$, $\mathfrak{S}\{3\} \in \mathcal{T}_1$;
call the Algorithm 4 with the input $(\mathfrak{S}\{1\}, \mathfrak{S}\{2\}, \mathfrak{S}\{3\})$;
3° Find all subchains $\mathfrak{S}$ of length $n = 3$ in the remaining interface elements such that $\mathfrak{S}\{1\} \in \mathcal{T}_2$, $\mathfrak{S}\{2\} \in T_1$, $\mathfrak{S}\{3\} \in \mathcal{T}_2$;
call the Algorithm 1 with the input $(\mathfrak{S}\{1\}, \mathfrak{S}\{2\}, \mathfrak{S}\{3\})$;
4° Find all elements $K \in \mathcal{T}_2$ in the remaining interface elements;

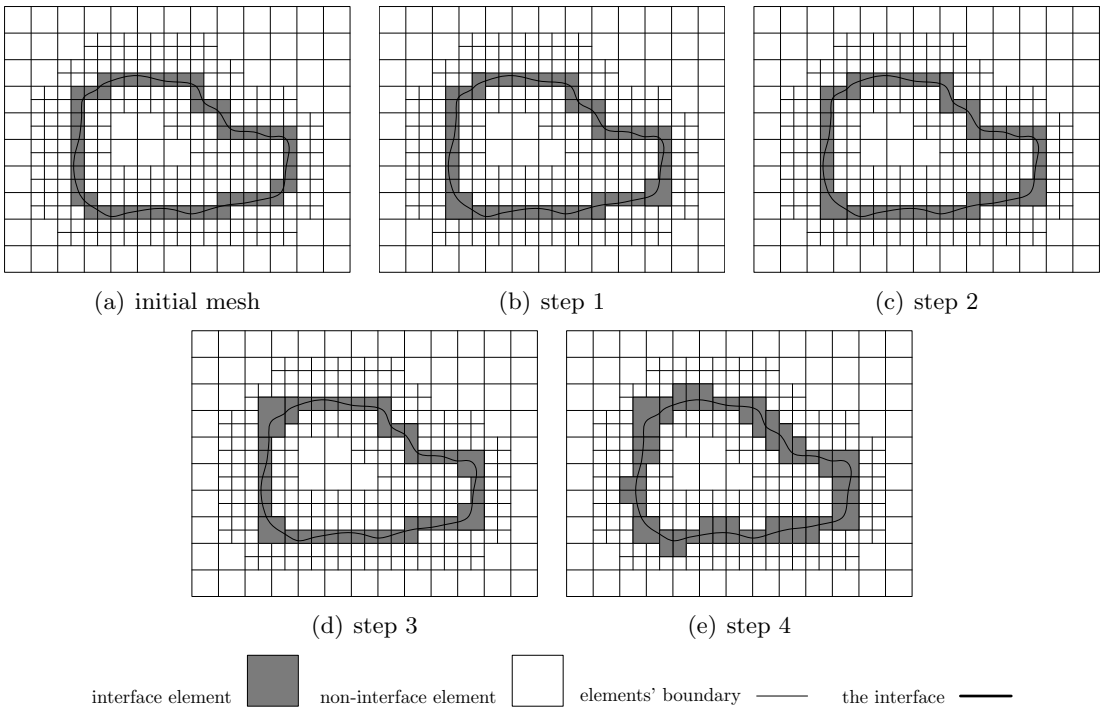


Figure 3.6: Illustration of the merging algorithm of the admissible chain of interface elements

call the Algorithm 5 with the input K.

Figure 3.6 illustrates each step in Algorithm 6 starting from an admissible chain of interface elements. The black thin lines represent the boundaries of the elements. We remove the lines which are shared by adjacent elements in steps 1° to 4°, meaning that two adjacent elements have been merged in the same macro-element.

We notice that for any $K \in \mathcal{T}\backslash\mathcal{T}^\Gamma$, the Rule 4 of the admissible chain requires that the interface elements in $\mathcal{S}(K)_1$ must be connected. These interface elements may belong to different patterns. The following lemma shows that the interface elements in $\mathcal{S}(K)_1$ belonging to the union of different patterns must be connected if K belongs to these patterns. The proof indicates that the order of merging different types of patterns in Algorithm 6 is crucial. The lemma will be used in our proof of the reliability of Algorithm 6.

Lemma 3.3. *Let $\mathfrak{C}$ be an admissible chain of interface elements of length $n \geq 2$. If $K \in \mathcal{T}\backslash\mathcal{T}^\Gamma$ is merged with interface elements in $\mathcal{S}(K)_1$ which belong to two different patterns $\mathcal{P}_1$ and $\mathcal{P}_2$ by Algorithm 6, then the interface elements in $(\mathcal{P}_1 \cup \mathcal{P}_2) \cap \mathcal{S}(K)_1$ are connected.*

Proof. Denote $\mathcal{P}_j^\Gamma := \mathcal{P}_j \cap (\mathcal{S}(K)_1 \cap \mathcal{T}^\Gamma)$, $j = 1, 2$, the interface elements of $\mathcal{P}_j$ in $\mathcal{S}(K)_1$. Let $\mathrm{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = \min_{D_1 \in \mathcal{P}_1^\Gamma, D_2 \in \mathcal{P}_2^\Gamma} \mathrm{dist}(D_1, D_2)$, where $\mathrm{dist}(D_1, D_2)$ is the minimum number of non-interface elements connecting D_1, D_2 in $\mathcal{S}(K)_1\backslash\mathcal{S}(K)_0$. Clearly, $0 \leq \mathrm{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) \leq 3$ and $\mathrm{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = 0$ implies the interface elements in $(\mathcal{P}_1 \cup \mathcal{P}_2) \cap \mathcal{S}(K)_1$ are connected. We now show that $\mathrm{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) \neq 0$ is impossible in three steps.

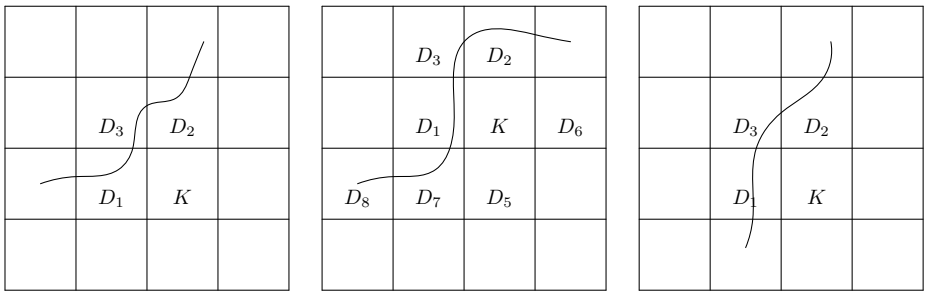


Figure 3.7: The element K and $D_1, D_2 \in N(K)$.

1° $\text{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = 1$. Let $D_1 \in \mathcal{P}_1^\Gamma$, $D_2 \in \mathcal{P}_2^\Gamma$, $\text{dist}(D_1, D_2) = 1$. By $S(K)_1 \cap \mathcal{T}^\Gamma$ is connected and the Rule 3 of the admissible chain, we know that D_3, which is the neighboring element to D_1 and D_2, is in $\mathcal{T}^\Gamma$ and D_3 can be either neighboring to K or diagonal to K.

(1) If D_3 is diagonal to K, by the Rule 2 of the admissible chain, $D_3 \in \mathcal{T}_1$, see Fig. 3.7. Firstly assume $D_1, D_2 \in \mathcal{T}_1$, see Fig. 3.7 (a), then there are three elements $D_1 \to D_3 \to D_2$ forming a chain and all elements belong to $\mathcal{T}_1$. By our merging Algorithm, D_1, D_3 or D_2, D_3 will form a pattern of type 3 which will be merged by Algorithm 3 before they can be merged with other interface elements forming a pattern of type 2. But these type 3 patterns will not use K, which contradicts to the assumption that K is merged with D_1 and D_2.

Secondly assume $D_1, D_2 \in \mathcal{T}_2$, see Fig. 3.7 (b). By the Rule 3 of the admissible chain, $D_5, D_6 \in \mathcal{T} \backslash \mathcal{T}^\Gamma$. If K is merged with its neighboring element $D_1 \in \mathcal{T}_2$ then D_1 and K will be in a pattern of type 1 or 5. When $\mathcal{P}_1$ is a pattern of type 1, since D_3 is not merged with K, then $D_7 \in \mathcal{T}_1$, $D_8 \in \mathcal{T}_2$, D_1, D_7, D_8 can form a pattern of type 1 and merged with K. However, in this case, D_7, D_1, D_3 will form a pattern of type 4 which will be merged by Algorithm 4 before D_1, D_7, D_8 are merged by Algorithm 1 in the second step of our merging Algorithm. Thus $\mathcal{P}_1$ cannot be of type 1. Similarly, $\mathcal{P}_2$ also cannot be of type 1. The remaining case is that $\mathcal{P}_1$ and $\mathcal{P}_2$ are both patterns of type 5. But this case is also impossible because D_1, D_3, D_2 will form a pattern of type 1 which will be merged by Algorithm 1 before D_1, D_2 can possibly be merged with K by Algorithm 5 in the third step of our merging Algorithm.

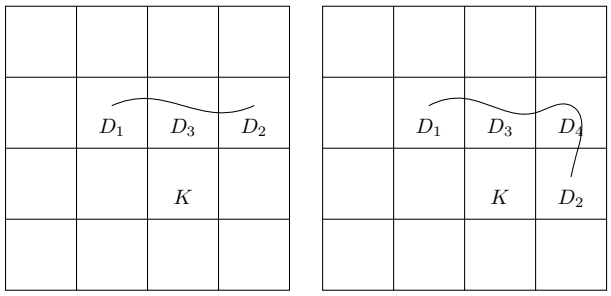
Finally assume $D_1 \in \mathcal{T}_2$, $D_2 \in \mathcal{T}_1$, see Fig. 3.7 (c). In this case D_3 and D_2 will be in a pattern of type 2 or 3 which will be merged by Algorithm 2 or 3 in the first step of our merging Algorithm. In both cases, they will not use the element K, which contradicts to the assumption that K is merged with D_2.

(2) If D_3 is neighboring to D_1, D_2, again since $\mathcal{S}(K)_1 \cap \mathcal{T}^\Gamma$ is connected and by the Rule 3 of the admissible chain, $D_3 \in \mathcal{T}_2$, see Fig.3.8 (a). Since K is merged with its diagonal elements D_1, and K has a neighboring element $D_3 \in \mathcal{T}_2$, by Remark 3.4, D_3 is merged with K. This contradicts to $\text{dist}(\mathcal{P}_1, \mathcal{P}_2) = 1$.

2° $\text{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = 2$, see Fig. 3.8 (b). Since $S(K)_1 \cap \mathcal{T}^\Gamma$ is connected, we have $D_3 \in \mathcal{T}_2$, $D_4 \in \mathcal{T}_1$. Again by Remark 3.4, D_3 is merged with K, which contradicts to $\text{dist}(\mathcal{P}_1, \mathcal{P}_2) = 2$.

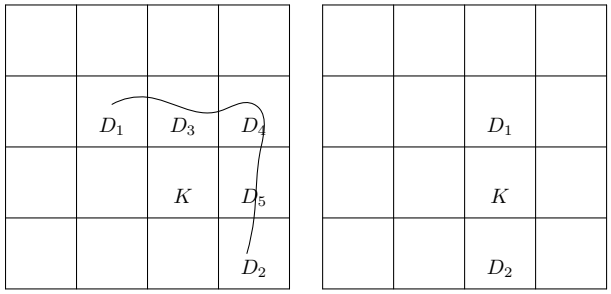
4° $\text{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = 3$, see Fig. 3.9. There are two possibilities:

(1) D_1, D_2 are diagonal to K, see Fig. 3.9 (a). Then $D_3 \in \mathcal{T}_2$, $D_4 \in \mathcal{T}_1$, and $D_5 \in \mathcal{T}_2$. By Remark 3.4, D_3, D_5 are merged with K, which contradicts to $\text{dist}(\mathcal{P}_1^\Gamma, \mathcal{P}_2^\Gamma) = 3$.



(a) (b)

Figure 3.8: The element K and $D_1, D_2 \in D(K)$ (left); The element K and $\text{dist}(D_1, D_2) = 2$ (right).



(a) (b)

Figure 3.9: The element K and $D_1, D_2 \in D(K)$, $\text{dist}(D_1, D_2) = 3$ (left); The element K and $D_1, D_2 \in N(K)$, $\text{dist}(D_1, D_2) = 3$ (right).

(2) D_1, D_2 are neighboring to K, see Fig. 3.9 (b). This case is also impossible because $S(K)_1 \cap \mathcal{T}^\Gamma$ is connected, then it will lead to K has 3 neighboring elements in $\mathcal{T}^\Gamma$ which contradicts to the Rule 3 of the admissible chain. This completes the proof. □

We attach any chain of interface elements $\mathfrak{C}$ of length $n \geq 1$ an accompany chain $\mathfrak{N}(\mathfrak{C}) = \{N_1 \to N_2 \to \cdots \to N_n\}$ with $N_i = 1$ or 2 according to $\mathfrak{C}\{i\} \in \mathcal{T}_1$ or $\mathcal{T}_2$, $i = 1, \cdots, n$. The following theorem shows the reliability of the merging algorithm.

Theorem 3.1. *Let $\delta_0 \in (0, 1/5]$. For any admissible chain of interface elements $\mathfrak{C}$ with length $n \geq 2$, if $\mathfrak{C}(1), \mathfrak{C}(n) \in \mathcal{T}_2$ or $\mathfrak{C}(1) = \mathfrak{C}(n)$, then Algorithm 6 terminates in finite number of steps with input $\mathfrak{C}$. All elements of the locally induced mesh* Induced($\mathfrak{C}$) *are large elements.*

Proof. By the step 1° of the algorithm, any two consecutive elements of type $\mathcal{T}_1$ are merged. Thus in the remaining elements of the chain, the type $\mathcal{T}_1$ elements must be interlaced if they are present. The step 2° merges all remaining elements in the chain which consists of a subchain of length 3 of the type $1 \to 2 \to 1$. The remaining type $\mathcal{T}_1$ elements in the chain of length 3 can appear only in the form $2 \to 1 \to 2$ which are merged by the step 3°. Thus the first three steps of the algorithm merge all elements in $\mathcal{T}_1$. Here we have used the assumption that the first and last elements in $\mathfrak{C}$ both belong to $\mathcal{T}_2$ or the first

and last elements are the same interface elements. The left type $\mathcal{T}_2$ elements are treated in the step 4° of the algorithm. The elements in $\mathcal{T}_3$ are all large elements and thus need not be merged. This shows that Algorithm 6 will merge all interface elements in the chain to output a locally induced mesh Induced($\mathfrak{C}$) which consists of the large elements of $\mathfrak{C}$ and the macro-elements containing all small elements of the chain $\mathfrak{C}$. By Lemmas 3.1-3.2, the elements in Induced($\mathfrak{C}$) are all large elements since $\delta_0 \in (0, 1/5\,]$.

It remains to show that the non-interface elements of the mesh $\mathcal{T}$ will not be used twice in the merging Algorithm 6 to guarantee the success of the algorithm. Let $K \in \mathcal{T}\backslash\mathcal{T}^\Gamma$. We first assume K is merged with interface elements in $\mathcal{S}(K)_1$ which belong to two patterns $\mathcal{P}_1$ and $\mathcal{P}_2$. By Lemma 3.3, the elements in $(\mathcal{P}_1 \cup \mathcal{P}_2) \cap \mathcal{S}(K)_1$ must be connected. Assume $D_1 \in \mathcal{P}_1$, $D_2 \in \mathcal{P}_2$ are connected, then one of D_1 and D_2 must be diagonal to K. Without loss of generality, we assume $D_1 = D(K)$. If $D_1 \in \mathcal{T}_2$, then by Remark 3.2 the element D' neighboring K and D_1 must be in $\mathcal{T}_1$ so that K, D_1', D_1 form a pattern of type 1. Thus by Rule 2 of the admissible chain, D_2, as an interface element, cannot be neighboring D_1, see Fig.3.10 (top left). This is a contradiction. Therefore, D_1 can only be of type $\mathcal{T}_1$. There are three possibilities illustrated in Fig.3.10.

(1) In the case of Fig.3.10 (top right), Rule 2 implies D_2 must be in $\mathcal{T}_2$. By Remark 3.2, K, D_1, and D_2 form a pattern of type 1, which contradicts to the assumption that D_1, D_2 belong to different patterns.

(2) In the case of Fig.3.10 (bottom left), Rule 2 implies D_2 cannot be neighboring D_1.

(3) In the case of Fig.3.10 (bottom middle), D_1 has only one common vertex with K. By Remark 3.3 the neighboring elements of D_1, D_1', D_1'' both must be of type $\mathcal{T}_1$ or $\mathcal{T}_2$ and K, D_1, D_1', D_1'' form a pattern of type 1 or 2. If K, D_1, D_1', D_1'' form a pattern of type 1, then this case belongs to (1). If K, D_1, D_1', D_1'' form a pattern of type 2, D_2 must be equal to one of D_1' or D_1'', which contradicts that D_1, D_2 are in different patterns.

In conclusion, K cannot be merged with two interface elements in $\mathcal{S}(K)_1$ belonging to different patterns. It remains to show that K cannot be merged with different interface elements in $\mathcal{S}(K)_2$ belonging to two patterns $\mathcal{P}_1, \mathcal{P}_2$ of which one pattern, e.g., $\mathcal{P}_1$, consists of only interface elements in $\mathcal{S}(K)_2\backslash\mathcal{S}(K)_1$. By the construction of the patterns in §3.2, $\mathcal{P}_1$ must be a pattern of type 2 and all interface elements D, D', D'' in $\mathcal{P}_1$ are in the second layer of elements surrounding K, see Fig.3.10 (bottom right). In this case, we know by the Rule 4 of the admissible chain that K cannot be merged with elements in $\mathcal{S}(K)_2$ other than D, D', D'', that is, K cannot be merged with interface elements belonging to the second pattern $\mathcal{P}_2$. This completes the proof. □

To conclude this section, we show that the merging Algorithm 6 leads to a reliable algorithm to automatically construct a body-fitted shape regular mesh for arbitrarily shaped smooth interface. We start from a conforming uniform mesh $\mathcal{T}_0$ of the domain Ω. We refine the interface elements of $\mathcal{T}_0$ by quad refinements and their surrounding elements to generate a Cartesian mesh $\mathcal{T}$ with hanging nodes such that all interface elements of $\mathcal{T}$ form an admissible chain $\mathfrak{C}$. This is possible because the interface Γ is C^2-smooth. Now we use Algorithm 6 to obtain an induced mesh $\mathcal{M} = \text{Induced}(\mathfrak{C})$. Since each interface element $K \in \mathcal{M}^\Gamma$ is a large element, K_i^h, $i = 1, 2$, is strongly shape regular in the sense that it is the union of shape regular triangles which we denote as T_K^{ij}, $1 \le j \le m_K$. Then the mesh

$$\widetilde{\mathcal{M}} = \{T_K^{ij} : i = 1, 2,\ j = 1 \cdots, m_K, K \in \mathcal{M}^\Gamma\} \cup \{K : K \in \mathcal{M}\backslash\mathcal{M}^\Gamma\}$$

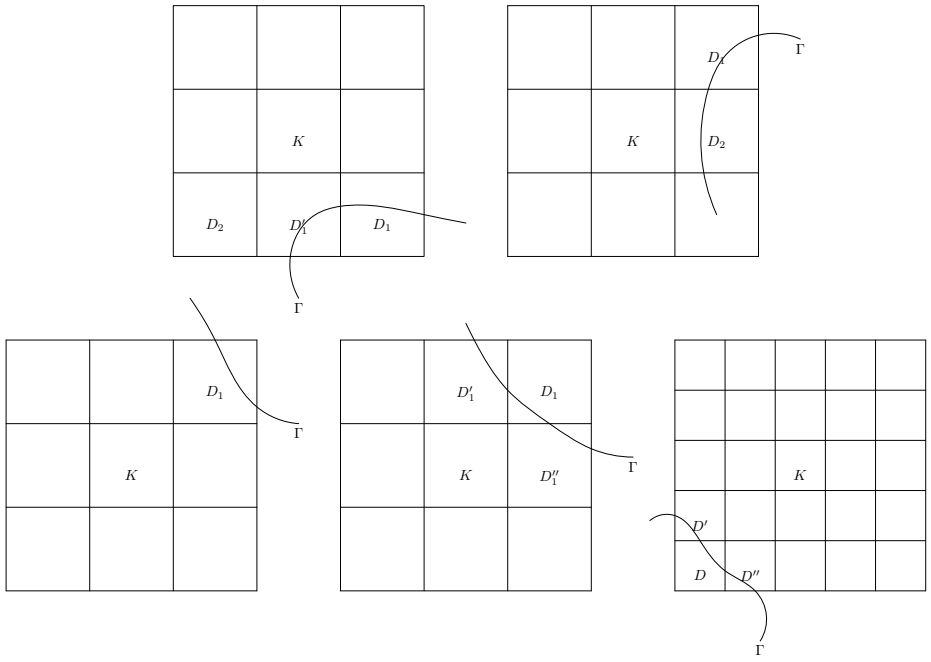


Figure 3.10: The element K and $D_1 \in \mathcal{S}(K)_1$ is a type $\mathcal{T}_2$ element (top left). The element K and $D_1 \in \mathcal{S}(K)_1$ is a type $\mathcal{T}_1$ element (top right, bottom left, and bottom middle). The element K and D, D', D'' in $\mathcal{S}(K)_2 \backslash \mathcal{S}(K)_1$ (bottom right).

is a triangular-rectangular mixed finite element mesh of the domain Ω. $\{T_K^{ij} : i = 1, 2,\ j = 1 \cdots, m_K, K \in \mathcal{M}^\Gamma\}$ is a body-fitted shape regular triangular mesh that covers the interface and $\{K : K \in \mathcal{M} \backslash \mathcal{M}^\Gamma\}$ consists of a rectangular mesh whose elements are similar to the elements of the initial mesh $\mathcal{T}_0$. Fig.3.11 shows a mixed mesh constructed from the unfitted finite element mesh in Fig.3.6(e).

4 The condition number of the stiffness matrix

In this section we study the condition number of the stiffness matrix of the unfitted finite element method defined in (2.4). Since we allow the Cartesian mesh $\mathcal{T}$ having hanging nodes, which is a nonconforming mesh in the classical sense, we recall an important concept of the K-mesh in Babuška and Miller [4]. It is introduced to control the undesirable excessive local refinements so that the local mesh sizes around each vertex of the elements are comparable. This concept is further developed in Bonito and Nochetto [12] as the control of the local level of incompatibility of the nonconforming meshes.

Let $\mathcal{N}^0$ be the set of conforming nodes of the mesh $\mathcal{T}$. A conforming node of $\mathcal{T}$ is a vertex of the elements in $\mathcal{T}$ which either locates on the boundary $\partial\Omega$ or is shared by the four elements to which it belongs. For each conforming node P, we define $\psi_P \in \mathbb{X}_1(\mathcal{T}) \cap H^1(\Omega)$, which is bilinear in each element and satisfies $\psi_P(Q) = \delta_{PQ}$ for any $Q \in \mathcal{N}^0$. Here δ_{PQ} is the Kronecker delta. It is proved in [4] that $\{\psi_P : P \in \mathcal{N}^0\}$ consists of a basis of $\mathbb{X}_1(\mathcal{T}) \cap H^1(\Omega)$ and satisfies the property of the partition of unity $\sum_{P \in \mathcal{N}^0} \psi_P = 1$. In the

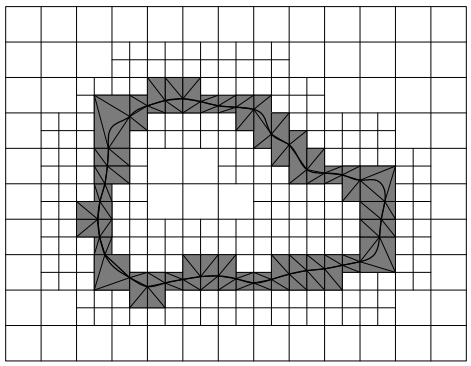


Figure 3.11: Illustration of a mixed triangular-rectangular body-fitted shape regular finite element mesh.

rest of the paper, we impose the following assumption on the finite element mesh $\mathcal{T}$ which is called the K-mesh in [4].

Assumption (H3) There exists a constant $C > 0$ uniform on the level of discretization of $\mathcal{T}$ such that for any conforming node $P \in \mathcal{N}^0$,

$$\operatorname{diam}(\operatorname{supp}(\psi_P)) \le C \min_{K \in \mathcal{T}_P} h_K,$$

where $\mathcal{T}_P := \{K \in \mathcal{T},\ K \subset \operatorname{supp}(\psi_P)\}$.

One can find further properties of K-meshes in [4]. We refer to [12, §6] for a refinement algorithm to enforce the assumption (H3) in practical computations.

The following lemma on the continuous approximation of discontinuous piecewise polynomials on K-meshes is proved in [21, Lemma 3.2].

Lemma 4.1. *Let $\mathbb{V}_P(\mathcal{T}) = \Pi_{K\in\mathcal{T}} Q_p(K)$. There exists an interpolation operator $\pi_h : \mathbb{V}_p(\mathcal{T}) \to \mathbb{V}_p(\mathcal{T}) \cap H^1(\Omega)$ such that for any $v \in \mathbb{V}_p(\mathcal{T})$,*

$$\|v - \pi_h v\|_{L^2(K)} \le C \|p^{-1} h^{1/2} [\![v]\!]\|_{L^2(\sigma(K))},$$
$$\|\nabla(v - \pi_h v)\|_{L^2(K)} \le C \|p h^{-1/2} [\![v]\!]\|_{L^2(\sigma(K))},$$

where $\sigma(K) = \{e \in \mathcal{E}^{\text{side}} : e \subset \widetilde{\omega}(K)\}$, $\widetilde{\omega}(K)$ is a set of elements including K such that $\operatorname{diam}(\widetilde{\omega}(K)) \le C h_K$. The constant C is independent of h_K, p. Moreover, $\pi_h v \in H_0^1(\Omega)$ if $v = 0$ on $\partial\Omega$.

Since the induced mesh $\mathcal{M} = \text{Induced}(\mathcal{T})$ is obtained by merging some of the elements of $\mathcal{T}$, $\mathbb{V}_p(\mathcal{M}) \subset \mathbb{V}_p(\mathcal{T})$. Thus Lemma 4.1 is also valid for any functions $v \in \mathbb{V}_p(\mathcal{M})$. We have the following discrete Poincaré inequality.

Lemma 4.2. *For any $v \in \mathbb{X}_p(\mathcal{M})$, we have $\|v\|_{L^2(\Omega)} \le C \|v\|_{\text{DG}}$, where $C > 0$ is a constant independent of the mesh sizes, p, and the interface deviations η_K for all $K \in \mathcal{M}^\Gamma$.*

Proof. Let $v = v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2} \in \mathbb{X}_p(\mathcal{M})$. By Lemma 4.1, for $v_i \in \mathbb{V}_p(\mathcal{M}_i)$, $i = 1, 2$, there exists $\pi_h v_i \in \mathbb{V}_P(\mathcal{M}_i) \cap H^1(\Omega_i^h)$ such that

$$\|v_i - \pi_h v_i\|_{\mathcal{M}_i} \le C\|p^{-1}h^{1/2}[\![v_i]\!]\|_{\mathcal{E}_i^{side}}, \quad \|\nabla(v_i - \pi_h v_i)\|_{\mathcal{M}_i} \le C\|ph^{-1/2}[\![v_i]\!]\|_{\mathcal{E}_i^{side}}. \tag{4.1}$$

Recall that we have assumed $\bar{\Omega}_1 \subset \Omega$. Let $w_i \in H^1(\Omega_i)$, $i = 1, 2$, satisfy

$$\begin{aligned} -\Delta w_1 &= 0 \ \text{ in } \Omega_1, \quad w_1 = [\![\pi_h v]\!]_\Gamma \ \text{ on } \Gamma = \partial\Omega_1, \\ -\Delta w_2 &= 0 \ \text{ in } \Omega_2, \quad w_2 = 0 \ \text{ on } \Gamma, \quad w_2 = \pi_h v_2 \ \text{ on } \partial\Omega. \end{aligned}$$

Then $w_i \in H^1(\Omega_i)$ satisfies $\|w_1\|_{H^1(\Omega_1)} \le C\|[\![\pi_h v]\!]\|_{H^{1/2}(\Gamma)}$, $\|w_2\|_{H^1(\Omega_2)} \le C\|\pi_h v\|_{H^{1/2}(\partial\Omega)}$. From the proof of [21, Lemma 3.4] we know that

$$\|[\![\pi_h v]\!]\|^2_{H^{1/2}(\Gamma)} \le C(\|ph^{-1/2}[\![v]\!]\|_{\mathcal{E}^\Gamma \cup \mathcal{E}_1^{\rm side} \cup \mathcal{E}_2^{\rm side}} + \|p^{-1}h^{1/2}\nabla_T[\![v]\!]\|_{\mathcal{E}^\Gamma}). \tag{4.2}$$

Now we use a similare argument to bound $\|\pi_h v\|_{H^{1/2}(\partial\Omega)}$. By the the localization lemma of the $H^{1/2}$ semi-norm in Faermann [28, Lemm 2.3] and the Gagliardo-Nirenberg type estimate for the $H^{1/2}$ semi-norm, we obtain as in [21, (3.13)] that

$$\|\pi_h v_2\|^2_{H^{1/2}(\partial\Omega)} \le C \sum_{K\in\mathcal{M}_2} (\|\pi_h v_2\|_{L^2(\Sigma_K)}\|\nabla_T(\pi_h v_2)\|_{L^2(\Sigma_K)} + h_K^{-1}\|\pi_h v_2\|^2_{L^2(\Sigma_K)}), \tag{4.3}$$

where $\Sigma_K = \partial K \cap \partial\Omega$. By the hp-inverse estimate and Lemma 4.1, we obtain

$$\begin{aligned} \|\nabla_T(\pi_h v_2)\|_{L^2(\Sigma_K)} &\le \|\nabla_T v_2\|_{L^2(\Sigma_K)} + \|\nabla_T(v_2 - \pi_h v_2)\|_{L^2(\Sigma_K)} \\ &\le Cp^2 h_K^{-1}\|v_2\|_{L^2(\Sigma_K)} + Cph_K^{-1/2}\|\nabla(v_2 - \pi_h v_2)\|_{L^2(K)} \\ &\le Cp^2 h_K^{-1}\|v_2\|_{L^2(\Sigma_K)} + Cph_K^{-1/2}\|ph^{-1/2}[\![v_2]\!]\|_{L^2(\sigma(K))}. \end{aligned}$$

Similarly, one can prove $\|\pi_h v_2\|_{L^2(\Sigma_K)} \le \|v_2\|_{L^2(\Sigma_K)} + \|[\![v_2]\!]\|_{L^2(\sigma(K))}$. Recall that $[\![v_2]\!] = v_2$ on $\partial\Omega$. This implies by (4.3) that

$$\|\pi_h v_2\|_{H^{1/2}(\partial\Omega)} \le C\|ph^{-1/2}[\![v_2]\!]\|_{\mathcal{E}^{\rm bdy}\cup\mathcal{E}_2^{\rm side}}.$$

Therefore, by combining with (4.2) we have

$$\|w_1\|_{H^1(\Omega_1)} + \|w_2\|_{H^1(\Omega_2)} \le C\|v\|_{\rm DG}. \tag{4.4}$$

Let $\pi_h^c v = (\pi_h v_1 - w_1)\chi_{\Omega_1} + (\pi_h v_2 - w_2)\chi_{\Omega_2}$. Then $\pi_h^c v \in H_0^1(\Omega)$ and by using Poincaré inequality for $\pi_h^c v$, we have

$$\begin{aligned} \|v\|_{L^2(\Omega)} &\le \|v - \pi_h^c v\|_{L^2(\Omega)} + \|\pi_h^c v\|_{L^2(\Omega)} \\ &\le \sum_{i=1}^2 (\|v_i - \pi_h v_i\|_{\mathcal{M}_i} + \|w_i\|_{L^2(\Omega_i)}) + C\|\nabla\pi_h^c v\|_{L^2(\Omega)} \\ &\le \sum_{i=1}^2 (\|v_i - \pi_h v_i\|_{\mathcal{M}_i} + \|w_i\|_{L^2(\Omega_i)}) + C(\|\nabla_h(\pi_h^c v - v)\|_{\mathcal{M}} + \|\nabla_h v\|_{\mathcal{M}}) \\ &\le C\sum_{i=1}^2 (\|v_i - \pi_h v_i\|_{H^1(\mathcal{M}_i)} + \|w_i\|_{H^1(\Omega_i)}) + C\|\nabla_h v\|_{\mathcal{M}}. \end{aligned}$$

Here for $i = 1, 2$, $\|w\|^2_{H^1(\mathcal{M}_i)} = \|w\|^2_{\mathcal{M}_i} + \|\nabla_h w\|^2_{\mathcal{M}_i} \ \ \forall w \in H^1(\mathcal{M}_i)$. This completes the proof by using (4.1) and (4.4). □

Now we consider the condition number of the stiffness matrix. We start by introducing the basis functions we use in each element. If $K \in \mathcal{M}\backslash\mathcal{M}^\Gamma$ is not an interface element, we will use a set of basis functions which are Lagrangian interpolation functions corresponding to Gauss-Lobatto points. We first recall some facts about spectral method and refer to Bernardi and Maday [11] for the details.

Let $I = (-1, 1)$ and $\{L_i\}_{i=0}^p$ the set of Legendre polynomials of $Q_p(I)$ which is the set of polynomials of degree p in I. Let $\{l_i\}_{i=0}^p$ be the set of Lagrangian interpolation functions in $Q_p(I)$ corresponding to the Gauss-Lobatto points $\{\xi_i\}_{i=0}^p$ which are the zeros of $(1-\xi^2)L_p'(\xi)$ in I.

Now let $\hat{K} = I \times I$ and $\{(\xi_i, \xi_j) : 0 \le i, j \le p\}$ be the Gauss-Lobatto grid of $\hat{K}$. Any function $\hat{v} \in Q_p(\hat{K})$ can be written as $\hat{v} = \sum_{i,j=0}^p \hat{v}_{ij} l_i(\hat{x}_1) l_j(\hat{x}_2)$. The following important result is proved in Melenk [39, Proposition 2.8, Theorem 4.1].

Lemma 4.3. *There exists a constant C independent of p such that for any function $\hat{v} = \sum_{i,j=0}^p \hat{v}_{ij} l_i(\hat{x}_1) l_j(\hat{x}_2)$, there holds*

$$C^{-1}p^{-2}\sum_{i,j=0}^{p}\hat{v}_{ij}^2 \le \|\hat{v}\|_{H^1(\hat{K})}^2 \le Cp\sum_{i,j=0}^{p}\hat{v}_{ij}^2,$$

and

$$\|\hat{v}\|_{L^2(\partial\hat{K})}^2 \le Cp^{-1}\left(\sum_{i=0,p}\sum_{j=0}^{p}\hat{v}_{ij}^2 + \sum_{j=0,p}\sum_{i=0}^{p}\hat{v}_{ij}^2\right).$$

For any $K \in \mathcal{M}$, let $F_K : \hat{K} \to K$ be the one-to-one and surjective affine mapping. Denote $\phi_K^{ij} = \hat{\phi}_{ij} \circ F_K^{-1}$, where $\hat{\phi}_{ij} = l_i(\hat{x}_1) l_j(\hat{x}_2)$, $0 \le i, j \le p$. For any $v \in Q_p(K), v = \sum_{i,j=0}^p v_K^{ij}\phi_K^{ij}$, we have by Lemma 4.3 and the standard scaling argument that

$$C^{-1}p^{-2}\|\boldsymbol{V}_K\|_{\ell_2}^2 \le \|\nabla v\|_{L^2(K)}^2 + h_K^{-2}\|v\|_{L^2(K)}^2 \le Cp\|\boldsymbol{V}_K\|_{\ell_2}^2, \tag{4.5}$$

$$\|v\|_{L^2(\partial K)}^2 \le Cp^{-1}h_K\|\boldsymbol{V}_K\|_{\ell_2}^2, \tag{4.6}$$

where $\boldsymbol{V}_K = (\boldsymbol{v}_0^T, \cdots, \boldsymbol{v}_p^T)^T$, $\boldsymbol{v}_i = (v_{i_0}, \cdots, v_{i_p})^T$ is the coefficient vector corresponding to $v \in Q_p(K)$.

For the interface element $K \in \mathcal{M}^\Gamma$, we have $K_1^{h-\delta_K} \subset K_1$ and $K_2^{h-\delta_K} \subset K_2$. We also have $\hat{K}_1^{h-\delta_K} \subset \hat{K}_1$ and $\hat{K}_2^{h-\delta_K} \subset \hat{K}_2$, where $\hat{K}_i^{h-\delta_K} = F_K^{-1}(K_i^{h-\delta_K})$, $\hat{K}_i = F_K^{-1}(K_i)$, $i = 1, 2$. Let $\{\hat{\psi}_{\hat{K}_i^h}^j\}_{j=1}^{(p+1)^2}$ the L^2-orthonormal basis of $Q_p(\hat{K}_i^{h-\delta_K})$, that is, $(\hat{\psi}_{\hat{K}_i^h}^j, \hat{\psi}_{\hat{K}_i^h}^k)_{\hat{K}_i^{h-\delta_K}} = \delta_{kj}$. Denote by $\psi_{K_i^h}^j = p^{-3/2}(\hat{\psi}_{\hat{K}_i^h}^j \circ F_K^{-1})$. Then $\{\psi_{K_i^h}^j\}_{j=1}^{(p+1)^2}$ is an L^2-orthogonal basis of $Q_p(K_i^{h-\delta_K})$, that is,

$$(\psi_{K_i^h}^j, \psi_{K_i^h}^k)_{K_i^{h-\delta_K}} = p^{-3}\frac{|K|}{|\hat{K}|}\delta_{jk}. \tag{4.7}$$

The scaling constant p^{-3} in (4.7) is important for us to balance the contribution of different basis functions used in interface and non-interface elements in the estimation of the condition

number of the stiffness matrix. Now for any $v \in \mathbb{X}_p(\mathcal{M})$, $K \in \mathcal{M}^\Gamma$,

$$v|_K = \sum_{j=1}^{(p+1)^2} (v^j_{K_1} \psi^j_{K_1^h} \chi_{K_1} + v^j_{K_2} \psi^j_{K_2^h} \chi_{K_2}) := v_1 \chi_{K_1} + v_2 \chi_{K_2}. \tag{4.8}$$

Let $\boldsymbol{V}_K = (v^1_{K_1}, \cdots, v^{(p+1)^2}_{K_1}, v^1_{K_2}, \cdots, v^{(p+1)^2}_{K_2})^T$ the coefficient vector corresponding to v, then by (4.7) we have

$$\|v_1\|^2_{L^2(K_1^{h-\delta_K})} + \|v_2\|^2_{L^2(K_2^{h-\delta_K})} = p^{-3} \frac{|K|}{|\hat{K}|} \|\boldsymbol{V}_K\|^2_{\ell_2}. \tag{4.9}$$

By Lemma 2.2 we obtain

$$Cp^{-3} h_K^2 \|\boldsymbol{V}_K\|^2_{\ell_2} \le \|v\|^2_{L^2(K)} \le C\Theta_K p^{-3} h_K^2 \|\boldsymbol{V}_K\|^2_{\ell_2} \quad \forall K \in \mathcal{M}^\Gamma. \tag{4.10}$$

Now, by the construction, any function $v \in \mathbb{X}_p(\mathcal{M})$ can be written as

$$v = \sum_{K \in \mathcal{M} \setminus \mathcal{M}^\Gamma} \sum_{i,j=0}^{p} v^{ij}_K \phi^{ij}_K + \sum_{K \in \mathcal{M}^\Gamma} \sum_{j=1}^{(p+1)^2} (v^j_{K_1} \psi^j_{K_1^h} \chi_{K_1} + v^j_{K_2} \psi^j_{K_2^h} \chi_{K_2}). \tag{4.11}$$

Let $N = \#\mathcal{M}$ be the number of elements of the mesh $\mathcal{M}$, $\{G_1, \cdots, G_N\}$ the elements of $\mathcal{M}$, and $\boldsymbol{V}_{G_i}$ the coefficient vector of $v|_{G_i}$, $i = 1, \cdots, N$. We denote $\boldsymbol{V} = (\boldsymbol{V}^T_{G_1}, \cdots, \boldsymbol{V}^T_{G_N})^T$ the vector of coefficients of v. The dimension of the vector $\boldsymbol{V}$ is $N_p = (p+1)^2 N$. We write $\boldsymbol{V} = \Phi(v)$, where $\Phi : \mathbb{X}_p(\mathcal{M}) \to \mathbb{R}^{N_p}$ is the mapping between functions in $\mathbb{X}_p(\mathcal{M})$ and their coefficient vectors.

Let $\boldsymbol{V} = \Phi(v), \boldsymbol{W} = \Phi(w) \in \mathbb{R}^{N_p}$ for $v, w \in \mathbb{X}_p(\mathcal{M})$. Then the stiffness matrix $\mathbb{A} = (a_{ij})_{i,j=1}^{N_p}$ is defined by

$$(\mathbb{A}\boldsymbol{V}, \boldsymbol{W})_{\ell_2} = a_h(v, w).$$

Recall that $\Theta = \max_{K \in \mathcal{M}} \Theta_K$. The following theorem is the main result of this section.

Theorem 4.1. *Denote $N^\Gamma = \#\mathcal{M}^\Gamma$ the number of elements of $\mathcal{M}^\Gamma$ and $M = \min(N - N^\Gamma, N^\Gamma)$. Then the following bound of the condition number of the stiffness matrix holds*

$$\kappa(\mathbb{A}) \le C\Theta^2 (1 + |\ln(h^2_{\min} M)|) \left(p^3 (N - N^\Gamma) + p^4 N^\Gamma\right),$$

where $h_{\min} = \min_{K \in \mathcal{M}} h_K$ and the constant $C > 0$ is independent of the mesh sizes, p, and the interface deviations η_K for all $K \in \mathcal{M}^\Gamma$.

We note that $N - N^\Gamma$ is the number of non-interface elements. For elliptic equations, it is well-known that the condition number of the stiffness matrix of standard finite element methods grows linearly in terms of the number of elements (see, e.g., Bank and Scott [10]). The condition number of the stiffness matrix of the *hp* finite element method using Gauss-Lobatto shape functions is studied in [39], which in particular generalizes earlier results that the condition number grows as $O(p^3)$ of the spectral method. Thus the estimate in Theorem 4.1 is optimal in terms of the number of elements and p. Our numerical results in Example 1 of section 5 show that the bound is also sharp in terms of the growth factor Θ^2.

Proof. For any $v = v_1\chi_{\Omega_1} + v_2\chi_{\Omega_2} \in \mathbb{X}_p(\mathcal{M})$, denote $w = (\pi_h v_1)\chi_{\Omega_1} + (\pi_h v_2)\chi_{\Omega_2}$ and $\boldsymbol{W} = \Phi(w)$ the coefficient vector corresponding to w. By (4.5), (4.10) and Lemma 2.3 we know that

$$\begin{aligned}\|\boldsymbol{V}-\boldsymbol{W}\|_{\ell_2}^2 &\leq Cp^2 \sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma} (\|\nabla(v-w)\|_{L^2(K)}^2 + h_K^{-2}\|v-w\|_{L^2(K)}^2) \\ &\quad + C\sum_{K\in\mathcal{M}^\Gamma} p^3 h_K^{-2}\|v-w\|_{L^2(K)}^2 \\ &\leq Cp^2\|ph^{-1/2}[\![v]\!]\|_{\mathcal{E}_1^{\rm side}\cup\mathcal{E}_2^{\rm side}}^2.\end{aligned}$$

Thus by the triangle inequality

$$\|\boldsymbol{V}\|_{\ell_2}^2 \leq Cp^2\|ph^{-1/2}[\![v]\!]\|_{\mathcal{E}_1^{\rm side}\cup\mathcal{E}_2^{\rm side}}^2 + 2\|\boldsymbol{W}\|_{\ell_2}^2. \tag{4.12}$$

Again by (4.5), (4.10) we have

$$\|\boldsymbol{W}\|_{\ell_2}^2 \leq Cp^2 \sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma} \big(\|\nabla w\|_{L^2(K)}^2 + h_K^{-2}\|w\|_{L^2(K)}^2\big) + C\sum_{K\in\mathcal{M}^\Gamma} p^3 h_K^{-2}\|w\|_{L^2(K)}^2.$$

Now we use an argument in [10]. By Hölder inequality, for any $r \geq 2$,

$$\begin{aligned}\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma} h_K^{-2}\|w\|_{L^2(K)}^2 &\leq C\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma} h_K^{-4/r}\|w\|_{L^r(K)}^2 \\ &\leq C\left(\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma} h_K^{-4/(r-2)}\right)^{\frac{r-2}{r}} \|w\|_{L^r(\Omega_1\cup\Omega_2)}^2 \\ &\leq Ch_{\min}^{-4/r}(N-N^\Gamma)^{\frac{r-2}{r}}\|w\|_{L^r(\Omega_1\cup\Omega_2)}^2.\end{aligned}$$

Similarly,

$$\sum_{K\in\mathcal{M}^\Gamma} p^3 h_K^{-2}\|w\|_{L^2(K)}^2 \leq Cp^3 h_{\min}^{-4/r}(N^\Gamma)^{\frac{r-2}{r}}\|w\|_{L^r(\Omega_1\cup\Omega_2)}^2.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{W}\|_{\ell_2}^2 &\leq Cp^2\|\nabla w\|_{\mathcal{M}\setminus\mathcal{M}^\Gamma}^2 + C(p^2(N-N^\Gamma)+p^3N^\Gamma)h_{\min}^{-4/r}M^{-2/r}\|w\|_{L^r(\Omega_1\cup\Omega_2)}^2 \\ &\leq C(p^2(N-N^\Gamma)+p^3N^\Gamma)(h_{\min}^2M)^{-2/r}r\|w\|_{H^1(\Omega_1\cup\Omega_2)}^2,\end{aligned}$$

where we have used the embedding inequality, $\|w\|_{L^r(D)} \leq Cr^{1/2}\|w\|_{H^1(D)}$ for any $w \in H^1(D)$, $r \geq 1$, on any Lipschitz domain D. Notice that for any $\zeta > 0$, $\zeta^{-2/r} = e^{-2\ln\zeta/r} = e^{-2}$ if $r = \ln\zeta$, by taking $r = \max(2, |\ln(h_{\min}^2M)|)$ we obtain

$$\|\boldsymbol{W}\|_{\ell_2}^2 \leq C(p^2(N-N^\Gamma)+p^3N^\Gamma)(1+|\ln(h_{\min}^2M)|)\|w\|_{H^1(\Omega_1\cup\Omega_2)}^2. \tag{4.13}$$

By Lemma 2.3 and the discrete Poincaré inequality in Lemma 4.2

$$\|w\|_{H^1(\Omega_1\cup\Omega_2)}^2 \leq 2\|w-v\|_{H^1(\Omega_1\cup\Omega_2)}^2 + 2(\|\nabla_h v\|_{L^2(\Omega)}^2 + \|v\|_{L^2(\Omega)}^2)$$

$$\begin{aligned} &\leq C(\|ph^{-1/2}[\![v]\!]\|^2_{\mathcal{E}_1^{\rm side}\cup\mathcal{E}_2^{\rm side}} + \|\nabla_h v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)}) \\ &\leq Ca_h(v,v). \end{aligned}$$

This yields by (4.12)-(4.13) that

$$\|\boldsymbol{V}\|^2_{\ell_2} \leq C(p^2(N-N^\Gamma)+p^3N^\Gamma)(1+|\ln(h^2_{\min}M)|)a_h(v,v). \tag{4.14}$$

On the other hand, since $a_h(v,v)\leq C\|v\|^2_{\rm DG}$, we have

$$\begin{aligned} a_h(v,v) &\leq C\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma}\Big(\|\nabla v\|^2_{L^2(K)} + \Theta_K\|ph^{-1/2}v\|^2_{L^2(\partial K)}\Big) \\ &\quad + \sum_{K\in\mathcal{M}^\Gamma}\sum_{i=1}^{2}\Big(\|\nabla v_i\|^2_{L^2(K_i)} + \Theta_K\|ph^{-1/2}v_i\|^2_{L^2(\partial K_i)} + \|p^{-1}h^{1/2}\nabla_T v_i\|^2_{L^2(\Gamma_K)}\Big) \\ &:= \mathrm{I}+\mathrm{II}. \end{aligned} \tag{4.15}$$

By (4.5)-(4.6)

$$\mathrm{I}\leq C\Theta\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma}\Big(\|\nabla v\|^2_{L^2(K)} + \|ph^{-1/2}v\|^2_{L^2(\partial K)}\Big) \leq C\Theta p\sum_{K\in\mathcal{M}\setminus\mathcal{M}^\Gamma}\|\boldsymbol{V}_K\|^2_{\ell_2}. \tag{4.16}$$

For $K\in\mathcal{M}^\Gamma$, by Lemma 2.3 and (4.10), for $i=1,2$,

$$\|\nabla v_i\|^2_{L^2(K_i)} \leq C\Theta_K p^4 h_K^{-2}\|v_i\|^2_{L^2(K_i)} \leq C\Theta_K^2 p\|\boldsymbol{V}_K\|^2_{\ell_2}.$$

By (2.9), Lemma 2.3, the *hp* trace inequality, and inverse estimate

$$\begin{aligned} \|v_i\|^2_{L^2(\partial K_i)} &\leq C\|v_i\|_{L^2(K_i)}\|\nabla v_i\|_{L^2(K_i)} + C\|v_i\|^2_{L^2(\partial K_i^h)} \\ &\leq C\Theta_K\|v_i\|_{L^2(K_i^{h-\delta_K})}\|\nabla v_i\|_{L^2(K_i^{h-\delta_K})} + Cp^2h_K^{-1}\|v_i\|^2_{L^2(K_i^h)} \\ &\leq C\Theta_K p^2 h_K^{-1}\|v_i\|^2_{L^2(K_i^{h-\delta_K})}, \end{aligned}$$

where we used the fact $\|v_i\|^2_{L^2(K_i^h)} \leq C\Theta_K\|v_i\|^2_{L^2(K_i^{h-\delta_K})}$, which follows directly from Lemma 2.1, in the last inequality. Thus by (4.9)

$$\Theta_K\|ph^{-1/2}v_i\|^2_{L^2(\partial K_i)} \leq C\Theta_K^2 p^4 h_K^{-2}\|v_i\|^2_{L^2(K_i^{h-\delta_K})} \leq C\Theta_K^2 p\|\boldsymbol{V}_K\|^2_{\ell_2}.$$

Similarly, one can prove $\|p^{-1}h^{1/2}\nabla_T v_i\|^2_{L^2(\Gamma_K)} \leq C\Theta_K^2 p\|\boldsymbol{V}_K\|^2_{\ell_2}$. Therefore, we have

$$\mathrm{II}\leq C\Theta^2 p\sum_{K\in\mathcal{M}^\Gamma}\|\boldsymbol{V}_K\|^2_{\ell_2}. \tag{4.17}$$

Combining (4.15)-(4.17) we obtain

$$a_h(v,v)\leq C\Theta^2 p\|\boldsymbol{V}\|^2_{\ell_2}.$$

This completes the proof by using (4.14). □

To conclude this section, we remark that since $\Theta_K = \mathsf{T}(\frac{1+3\eta_K}{1-\eta_K})^{4p+3}$, Theorem 4.1 indicates that to control the condition number of the stiffness matrix, one should choose $\eta_K \ll 1$, that is, one should have the interface being well resolved by the mesh.

5 Numerical examples

In this section we provide some numerical examples to verify our theoretical results. In order to construct the orthogonal polynomials on the polygons $\hat{K}_i^{h-\delta_K}$ for the interface elements K, we adopt the Gram-Schmidt process starting from the basis functions of $Q_p(\hat{K})$ which are the Lagrange interpolation polynomials through the Gauss-Lobatto integration points on $\hat{K}$. The details can be found in Sommariva and Vianello [45]. The algorithms are implemented in MATLAB on a workstation with Intel(R) Core(TM) i9-10885H CPU 2.40GHz and 64GB memory.

Example 1. *In this example we show that the growth factor Θ^2 in the bound of the condition number of the stiffness matrix in Theorem 4.1 is sharp. For this purpose, we consider the case of one interface element. Let $K = (-2, 2)^2$ and the interface $\Gamma = \{(x(t), y(t)) \in \mathbb{R}^2 : t \in (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})\}$, where $x(t)$ and $y(t)$ are defined as follows:*

$$x(t) = \sqrt{2}\cos(\alpha + \frac{\pi}{4})t + \sqrt{2}\sin(\alpha + \frac{\pi}{4})(100t^3 - \beta t) - 1,$$
$$y(t) = -\sqrt{2}\sin(\alpha + \frac{\pi}{4})t + \sqrt{2}\cos(\alpha + \frac{\pi}{4})(100t^3 - \beta t) - 1,$$

where $\cos(\alpha) = \frac{1}{\sqrt{\mu^2+1}}$, $\sin(\alpha) = \frac{\mu}{\mu^2+1}$, $\beta = \frac{100}{\sqrt{\mu^2+1}} - \mu$ *and* $\mu = 3.8$.

The domain and the interface are shown in Fig.5.1 (left) in which $K_1^{h-\delta_K} = \Delta AE'F'$ and $K_2^{h-\delta_K}$ is the polygon with vertices F'', E'', B, C, D. The interface deviation is $\eta_K = \frac{200}{3\sqrt{3}(\mu^2+1)^2} \approx 0.16$. We first consider the condition number of the mass matrix to verify our analysis in Lemma 2.2. For $v \in \mathbb{X}_p(K)$, in the notation of (4.8), the mass matrix $\mathbb{M} \in \mathbb{R}^{(p+1)^2\times(p+1)^2}$ is defined as $(\mathbb{M}\boldsymbol{V}_K, \boldsymbol{W}_K)_{\ell_2} = (v, w)_K \quad \forall v, w \in \mathbb{X}_p(K)$. Then (4.10) implies that the condition number $\kappa(\mathbb{M}) \le C\Theta$ for some constant C independent of p and η_K. We plot $\kappa(\mathbb{M})$ vs. Θ via different degrees of polynomials with loglog scaling in Fig.5.1 (right). It is clear that the condition number of $\mathbb{M}$ grows as Θ which agrees with our theoretical bound.

We plot the curve $\kappa(\mathbb{A})$ vs. $\Theta^2 p^4$ via different degrees of polynomials with loglog scaling in Fig.5.2 (left). We observe that the condition number grows as $\Theta^2 p^4$ which confirms our analysis in Theorem 4.1. We also observe that the $\kappa(\mathbb{A})$ increases very fast with the increase of polynomial degree. One can reduce the interface deviation to reduce the $\kappa(\mathbb{A})$. We change μ to reduce η_K such that $\eta_K \le \frac{0.1}{p(p+1)}$ and plot the curve p^4 vs. $\kappa(\mathbb{A})$ in Fig.5.2 (right). We can find the $\kappa(\mathbb{A})$ is significantly reduced and the $\kappa(\mathbb{A})$ has p^4 increasing rates.

This example shows clearly the importance of reducing the interface deviation to control the condition number of the stiffness matrix. In the following we always require

$$\max_{K\in\mathcal{M}} \eta_K \le \frac{0.1}{p(p+1)}, \tag{5.1}$$

which is stronger than that in Assumption (H2). The finite element meshes in our following numerical examples are constructed as follows.

Algorithm 7: The algorithm for generating the induced mesh satisfying Assumption (H3) and (5.1)

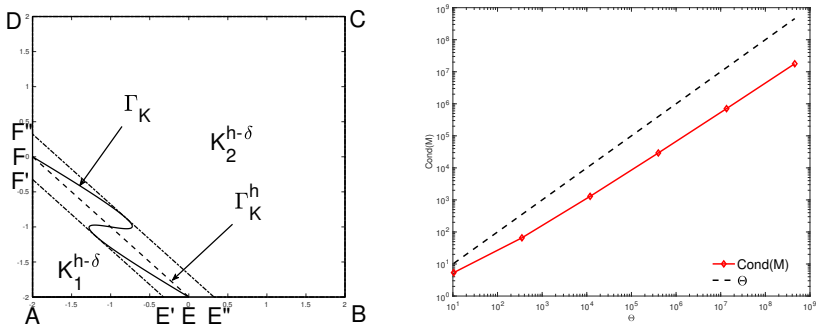


Figure 5.1: Example 1: The geometry setting of Example 1 (left) and the growth rate of the condition number of the mass matrix (right).

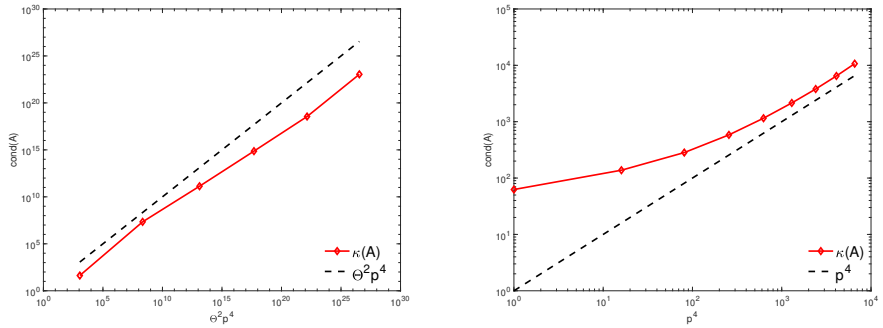


Figure 5.2: Example 1: The growth rate of the condition number of $\mathbb{A}$ with $\eta_K = 0.16$ (left) and the condition number of $\mathbb{A}$ with $\eta_K \leq \frac{0.1}{p(p+1)}$ (right).

Input: A uniform initial Cartesian mesh $\mathcal{T}_0$ of mesh size h

Output: The induced mesh $\mathcal{M} = \text{Induced}(\mathfrak{C})$

1° Set $\mathcal{T} = \mathcal{T}_0$;

2° Refine the elements of $\mathcal{T}$ near the interface by quad refinements to generate a Cartesian mesh (still denoted by) $\mathcal{T}$ with possible handing nodes such that all interface elements of $\mathcal{T}$ form an admissible chain $\mathfrak{C}$;

3° Call the refinement procedure in [12, §6.3] such that $\mathcal{T}$ satisfies Assumption (H3);

4° Use Algorithm 6 to generate an induced mesh $\mathcal{M} = \text{Induced}(\mathfrak{C})$;

5° If the interface elements in $\mathcal{M}$ do not satisfy (5.1), release all merged elements in $\mathfrak{C}$, go to 2°.

We remark that after step 2° in Algorithm 7, the interface elements are of the same size which is smaller than the sizes of non-interface elements. Thus when implementing the refinement procedure in [12, §6.3] in our situation, only non-interface elements are refined and consequently, the interface elements still form an admissible chain.

Example 2. *Let the interface* Γ *be the circle centered at* $(0,0)^T$ *with radius* $r_0 = 1.1$. *We*

set $\Omega = (-2,2)^2$, $\Omega_1 = \{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} < r_0\}$ *and* $\Omega_2 = \Omega \setminus \bar{\Omega}_1$. *Set* $a_1 = 10$ *and* $a_2 = 1$. *The right-hand side* f *and boundary condition* g *are computed such that the exact solution is*

$$u(x,y) = \begin{cases} e^{x^2+y^2-r_0^2} + 10r_0^2 - 1 + (x^2+y^2-r_0^2)^2 \sin(2\pi x)\sin(2\pi y) & \text{in } \Omega_1, \\ 10(x^2+y^2) + (x^2+y^2-r_0^2)^2 \sin(2\pi x)\sin(2\pi y) & \text{in } \Omega_2. \end{cases}$$

Table 1: Example 2: numerical errors $\|u-U\|_{DG}$ and orders for $p = 1,2,3,4,5$.

	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
h	error	order	error	order	error	order	error	order	error	order
1/4	1.13E+00	–	4.00E-01	–	1.20E-01	–	3.21E-02	–	2.09E-03	–
1/8	6.72E-01	0.75	1.08E-01	1.89	2.01E-02	2.58	1.55E-03	4.37	1.62E-04	3.69
1/16	3.57E-01	0.91	2.89E-02	1.90	2.49E-03	3.01	1.03E-04	3.91	5.18E-06	4.97
1/32	1.79E-01	0.99	7.32E-03	1.98	3.12E-04	3.00	6.56E-06	3.98	1.62E-07	5.00

In Table 1, we show the errors $\|u-U\|_{DG}$ and the corresponding convergence orders for $p = 1,2,3,4,5$. We clearly observe the optimal p-th order convergence and the superior performance of high order methods. Fig. 5.3 shows the induced mesh when $h = 1/4$ and the corresponding numerical solution.

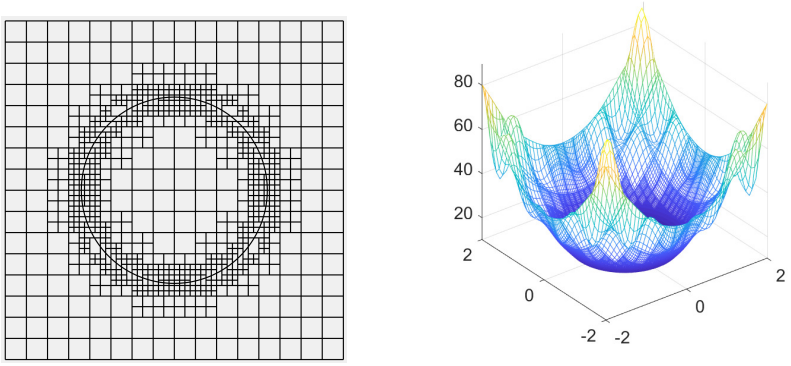


Figure 5.3: Example 2: The induced mesh of 940 elements when $h = 1/4$ (left) and the corresponding numerical solution (right).

Example 3. *In this example we consider geometrically more complex interface. Let the interface* Γ *be defined as follows:*

$$\Gamma = \{(x,y) \in \mathbb{R}^2 : r = \frac{2}{9}(3 + 4^{\sin(5\theta)})\},$$

where (r,θ) *are the polar coordinates. The domain* Ω *is divided to* Ω_1 *and* Ω_2 *by* Γ, *that is,*

$$\Omega_1 = \{(x,y) \in (-2,2)^2 : r < \frac{2}{9}(3 + 4^{\sin(5\theta)})\},$$
$$\Omega_2 = \{(x,y) \in (-2,2)^2 : r > \frac{2}{9}(3 + 4^{\sin(5\theta)})\}.$$

We set $a_1 = 10$, $a_2 = 1$, *the right-hand side* $f = 1$, *and the boundary condition* $g = 0$.

The exact solution of this example is unknown. We use the a posteriori error estimate in [21] to measure the accuracy of computation. In Table 2, we observe the optimal p-th order convergence. The induced mesh when $h = 1/4$ is shown in Fig. 5.4 which has 2654 elements. The discrete solution is depicted in Fig. 5.5.

Table 2: Example 3: A posterior error estimates and the convergence orders for $p = 1, 2, 3, 4, 5$.

	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
h	error	order	error	order	error	order	error	order	error	order
1/4	1.38E+00	–	1.08E-01	–	3.96E-02	–	2.85E-03	–	1.02E-04	–
1/8	7.31E-01	0.92	2.93E-02	1.88	5.13E-03	2.95	1.83E-04	3.96	3.35E-06	4.93
1/16	3.79E-01	0.95	8.13E-03	1.85	6.46E-04	2.99	1.15E-05	3.99	1.08E-07	4.95
1/32	1.90E-01	0.99	2.09E-03	1.96	8.12E-05	2.99	7.25E-07	3.99	3.41E-09	4.99

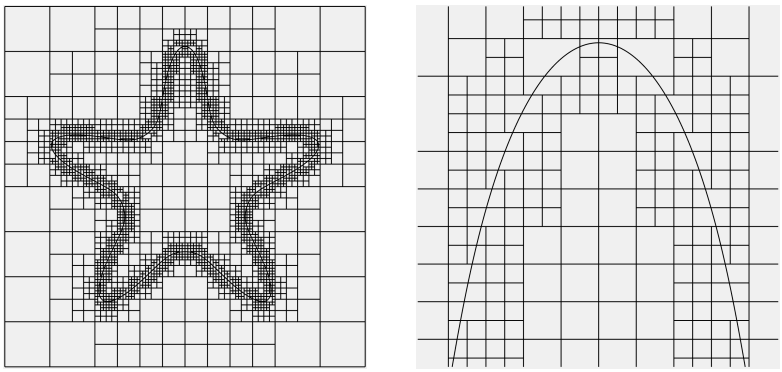


Figure 5.4: Example 3: The induced mesh of 2654 elements when $h = 1/4$ (left) and the corresponding zoomed local mesh (right).

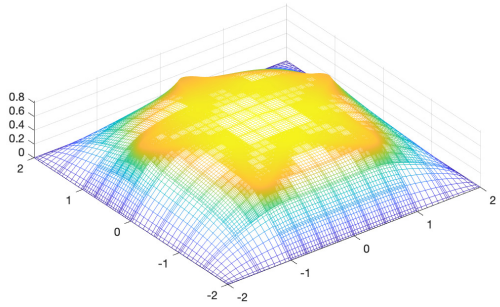


Figure 5.5: Example 3: The discrete solution on the mesh of 2654 elements.

Acknowledgement

The authors are grateful to Haijun Wu in Nanjing University for inspiring discussions.

References

[1] R.A. Adams and J.J.F. Fournier, Sobolev Spaces, second edition, Elsevier, Singapore (2009)

[2] D.N. Arnold, F. Brezzi, B. Cockburn, and L. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM J. Numer. Anal. **39**, 1749-1779 (2002)

[3] I. Babuška, The finite element method for elliptic equations with discontinuous coefficients, Computing **5**, 207-213 (1970)

[4] I. Babuška and A. Miller, A feedback finite element method with a posteriori error estimation, Part I. The finite element method and some basic properties of the a posteriori error estimator, Comput. Meth. Appl. Mech. Engrg. **61**, 1-40 (1987)

[5] I. Babuška and M. Suri, The *h-p* version of the finite element method with quasiuniform meshes, RAIRO - Model. Math. Anal. Numer. **21**, 199-238 (1987)

[6] S. Badia, J. Droniou, and L. Yemm, Conditioning of a hybrid high-order scheme on meshes with small faces, J. Sci. Comput. **92**, 71, (2022)

[7] S. Badia, E. Neiva, and F. Verdugo, Linking ghost penalty and aggregated unfitted methods, Comput. Meth. Appl. Mech. Engrg. **388**, 114232, (2022)

[8] S. Badia, E. Neiva, and F. Verdugo, Robust high-order unfitted finite elements by interpolation based discrete extension, arXiv:2201.06632v1

[9] S. Badia, F. Verdugo, and A.F. Martin, The aggregated unfitted finite element method for elliptic problems, Comput. Meth. Appl. Mech. Engrg. **336**, 533-553 (2018)

[10] R. Bank and L.R. Scott, On the conditioning of finite element equations with highly refined meshes, SIAM J. Numer. Anal. **26**, 1383-1394 (1989)

[11] C. Bernardi and Y. Maday, *Spectral Methods,* in Handbook of Numerical Analysis, Vol. 5, P.G. Cialet and J.L. Lions, eds., North-Holland, Elsevier, 209-486 (1997)

[12] A. Bonito and R.H. Nochetto, Quasi-optimal convergence rate of an adaptive discontinuous Galerkin method, SIAM J. Numer. Anal. **48**, 734-771 (2010)

[13] S. Bordas, E. Burman, M. Larson, and M. Olshanskii, eds., Geometrically Unfitted Finite Element Methods and Applications, Springer, New York (2018)

[14] S. Brenner and L. Sung, Virtual element methods on meshes with small edges or faces, Math. Models Meth. Appl. Sci. **28**, 1291-1336 (2017)

[15] E. Burman, M. Cicuttin, G. Delay, and A. Ern, An unfitted hybrid high-order method with cell agglomeration for elliptic interface problems, SIAM J. Sci. Comput. **43**, A859-A882 (2021)

[16] E. Burman and P. Hansbo, Fictitious domain finite element methods using cut elements, I. A stabilized Lagrange multiplier method, Comput. Meth. Appl. Mech. Engrg. **199**, 2680-2686 (2010)

[17] E. Burman and P. Hansbo, Fictitious domain finite element methods using cut elements, II. A stabilized Nitsche method, Appl. Numer. Math. **62**, 328-341 (2012)

[18] E. Burman, P. Hansbo, and M.G. Larson, CutFEM based on extended finite element spaces, arXiv:2101.10052v1

[19] L. Chen, H. Wei, and M. Wen, An interface-fitted mesh generator and virtual element methods for elliptic interface problems, J. Comput. Phys. **334**, 327-348 (2017)

[20] A. Cangiani, Z. Dong, and E.H. Georgoulis, *hp*-version discontinuous Galerkin methods on essentially arbitrarily-shaped elements, Math. Comp. **91**, 1-35 (2021)

[21] Z. Chen, K. Li, and X. Xiang, An adaptive high-order unfitted finite element method for elliptic interface problems, Numer. Math. **149**, 507-548 (2021)

[22] Z. Chen, Y. Xiao, and L. Zhang, The adaptive immersed interface finite element method for elliptic and Maxwell interface problems, J. Comput. Phys. **228**, 5000-5019 (2009)

[23] Z. Chen and J. Zou, Finite element methods and their convergence for elliptic and parabolic interface problems, Numer. Math. **79**, 175-202 (1998)

[24] P. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam (1978)

[25] B. Cockburn and C.-W. Shu, The local discontinuous Galerkin finite element method for time-dependent convection-diffusion systems, SIAM J. Numer. Anal. **35**, 2440-2463 (1998)

[26] R.A. DeVore and G.G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin (1993)

[27] M. Dubiner, Spectral methods on triangles and other domains, J. Sci. Comput. **6**, 345-390 (1991)

[28] B. Faermann, Localization of the Aronszaja-Slobodeckij norm and application to adaptive boundary element methods, Part I. The two-dimensional case, IMA J. Numer. Anal. **20**, 203-234 (2000)

[29] M. Feistauer, On the finite element approach of a cascade flow problem, Numer. Math. **50**, 655-684 (1987)

[30] C. Gürken, S. Sticko, and A. Massing, Stabilized cut discontinuous Galerkin method for advection-reaction problems, SIAM J. Sci. Comput. **42**, A2620-A2654 (2020)

[31] A. Hansbo and P. Hansbo, An unfitted finite element method, based on Nitsche's method, for elliptic interface problems, Comput. Meth. Appl. Mech. Engrg. **191**, 5537-5552 (2002)

[32] P. Huang, H. Wu, and Y. Xiao, An unfitted interface penalty finite element method for elliptic interface problems, Comput. Meth. Appl. Mech. Engrg. **323**, 539-436 (2017)

[33] A. Johansson and M.G. Larson, A high order discontinuous Galerkin Nitsche method for elliptic problems with fictitious boundary, Numer. Math. **123**, 607-628 (2013)

[34] C. Lehrenfeld and A. Reusken, Analysis of a high-order unfitted finite element method for elliptic interface problems, IMA J. Numer. Anal. **38**, 1351-1387 (2018)

[35] R. LeVeque and Z. Li, The immersed interface method for elliptic equations with discontinuous coefficients and singular sources, SIAM J. Numer. Anal. **31**, 1019-1044 (1994)

[36] Z. Li and K. Ito, The immersed interface method: Numerical solutions of PDEs involving interfaces and irregular domains, SIAM, Philadephia (2006)

[37] Z. Li, T. Lin, and X. Wu, New Cartesian grid methods for interface problems using finite element formulation, Numer. Math. **96**, 61-98 (2003)

[38] R. Massjung, An unfitted discontinuous Galerkin method applied to elliptic interface problems, SIAM J. Numer. Anal. **50**, 3134-3162 (2012)

[39] J.M. Melenk, On condition numbers in *hp*-FEM with Gauss-Lobatto based shape functions, J. Comput. Appl. Math. **139**, 21-48 (2002)

[40] P. Ming and Z. Shi, Quadrilateral mesh, Chin. Ann. of Math. **23B**, 235-252 (2002)

[41] J. Nitsche, Über ein Variationsprinzip zur Losung von Dirichlet-Problemen bei Verwendung von Teilraumen, die keinen Randbedingungen unterworfen sind, Abh. Math. Sem. Univ. Hamburg **36**, 9-15 (1970)

[42] I. Perugia and D. Schötzau, An *hp*-analysis of the local discontinuous Galerkin method for diffusion problems, J. Sci. Comput. **17**, 561-571 (2002)

[43] C.S. Peskin, Numerical analysis of blood flow in the heart, J. Comput. Phys. **24**, 220-252 (1997)

[44] F. de Prenter, C.V. Verhoosel, G.J. van Zwieten, and E.H. van Brummelen, Condition number analysis and prfeconditioning of the finite cell method, Comput. Meth. Appl. Mech. Engrg. **316**, 297-327 (2017)

[45] A. Sommariva and M. Vianello, Numerical hyperinterpolation over nonstandard planner regions, Math. Comput. Simul. **141**, 110-120 (2017)

[46] G. Szegö, Orthogonal Polynomials, American Mathematical Society, New York (1939)

[47] H. Wu and Y. Xiao, An unfitted *hp*-interface penalty finite element method for elliptic interface problems, J. Comput. Math. **37**, 316-339 (2010)

[48] Y. Xiao, J. Xu, and F. Wang, High-order extended finite element method for solving interface problems, Comput. Meth. Appl. Mech. Engrg. **364**, 112964 (2020)