Universidad de Concepción
Dirección de Postgrado
Facultad de Ciencias Físicas y Matemáticas
Programa de Doctorado en Ciencias Aplicadas
con Mención en Ingeniería Matemática

MÉTODOS DE GALERKIN DISCONTINUO PARA PROBLEMAS NO LINEALES EN FÍSICA DE PLASMAS

DISCONTINUOUS GALERKIN METHODS FOR NON-LINEAR PROBLEMS IN PLASMA PHYSICS

Tesis para optar al grado de Doctor en Ciencias
Aplicadas con mención en Ingeniería Matemática

Nestor Abel Sánchez Goycochea
Concepción-chile
2021

Profesor Guía: Manuel Solano Palma
CI^{2}MA y Departamento de Ingeniería Matemática
Universidad de Concepción, Chile

Cotutor: Tonatiuh Sánchez-Vizuet
Department of Mathematics
The University of Arizona, United States

Discontinuous Galerkin methods for non-linear problems in plasma physics

Nestor Abel Sánchez Goycochea

Directores de Tesis: Manuel Solano, Universidad de Concepción, Chile.
Tonatiuh Sánchez-Vizuet, The University of Arizona, United States.

Director de Programa: Raimund Bürger, Universidad de Concepción, Chile.

Comisión evaluadora

Prof. Gerardo Hernández Dueñas , Universidad Nacional Autónoma de México, México .

Prof. Lise-Marie Imbert-Gérard, The University of Arizona, United States.

Prof. Weifeng Qui, City University of Hong Kong, China.

Comisión examinadora

Firma: _____
Prof. Gabriel N. Gatica , Universidad de Concepción , Chile.

Firma: _____
Prof. Gerardo Hernández Dueñas , Universidad Nacional Autónoma de México, México .

Firma: _____
Prof. Lise-Marie Imbert-Gérard, The University of Arizona, United States.

Firma: _____
Prof. Ricardo Oyarzúa Vargas , Universidad del Bío Bío , Chile.

Calificación: _____

Concepción, Septiembre de 2021

Abstract

The goal of this thesis is to develop hybridizable discontinuous Galerkin-type discretizations applied to non-linear elliptic problems from plasma physics. The complexity of these problems lies in the non-linearity of the unknowns and their source terms, as well as the fact that the equations are posed in non-polygonal domains. To deal with the curved boundaries, a high-order transfer technique is applied for the boundary data.

First, we present the *a priori* and *a posteriori* error analysis of a high order hybridizable discontinuous Galerkin method (HDG) applied to a semi-linear elliptic problem, raised in a non-polygonal domain Ω. In this case, the non-linearity appears in the source term. We approximate Ω by a polygonal subdomain Ω_h and guarantee optimal convergence under mild assumptions related to the non-linear source term and the distance between the boundaries of the polygonal subdomain Ω_h and the original domain Ω. In addition, we use a local nonlinear post-processing of the scalar unknown to guarantee an additional order of convergence. Finally, we provide a reliable and locally efficient *a posteriori* error estimator that includes the approximation error between the original and artificial boundary data.

Then, we extend the above analysis to a class of nonlinear elliptic boundary value problems posed on curved domains, where both the source term and the diffusion coefficient are nonlinear. The non-linearity of the diffusion coefficient can be presented by means of a scalar function or a vector function. Therefore, we divide the analysis into two cases: In the first one, we consider that the non-linear diffusion coefficient depends on the solution, while in the second case, this coefficient depends on the gradient of the solution. We also show that under minor assumptions about the source term and the computational domain, the discrete systems, for both cases, are well defined. In addition, we provide *a priori* error estimates that show that the discrete solution will have an optimal order of convergence as long as the distance between the curved boundary and the computational boundary remains the same order of magnitude as the mesh parameter.

Finally, we propose a formulation that combines the HDG method with boundary element method (BEM) used for a more general problem from plasma physics. In this situation, the location of the plasma is unknown and it is necessary to solve the equilibrium condition in the half-plane to determine both the flow and the confinement region. The BEM method is ideal for working in unbounded domains, since the FEM approach would need an infinite number of elements to cover the domain.

Resumen

El objetivo de esta tesis es desarrollar discretizaciones de tipo Galerkin discontinuo hibridizable aplicados a problemas elípticos no lineales de la física de plasmas. La complejidad de éstos problemas radica en la no linealidad de las incógnitas y sus términos fuentes , así como en el hecho de que las ecuaciones se plantean en dominios no poligonales. Para lidiar con las fronteras curvas se aplica una técnica de transferencia de alto orden para los datos de frontera.

Primero, presentamos el análisis de error *a priori* y *a posteriori* de un método de Galerkin discontinuo hibridizable de alto orden (HDG) aplicado a un problema elíptico semi-lineal planteado en un dominio no poligonal Ω. En éste caso la no linealidad aparece en el término fuente. Aproximamos Ω por un subdominio poligonal Ω_h y garantizamos convergencia óptima bajo suposiciones menores relacionadas al término fuente no lineal y la distancia entre las fronteras del subdominio poligonal Ω_h y el dominio original Ω. Además, usamos un posprocesamiento local no lineal de la incógnita escalar para garantizar un orden adicional de convergencia. Finalmente, proporcionamos un estimador de error *a posteriori* confiable y localmente eficiente que incluye el error de aproximación entre los datos de la frontera original y artifical.

Luego, extendemos el análisis anterior para una clase de problemas de valores de frontera elípticos no lineales planteados en dominios curvos, donde tanto el término fuente como el coeficiente de difusión son no lineales. La no linealidad del coeficiente de difusión puede ser presentada mediante una función escalar o una función vectorial. Por lo tanto, dividimos éste análisis en dos casos: En el primero, consideramos que el coeficiente de difusión no lineal depende de la solución, mientras que para el segundo caso, dicho coeficiente depende del gradiente de la solución. Mostramos también que bajo hipótesis no restrictivas sobre el término fuente y el dominio computacional, los sistemas discretos, para ambos casos, están bien definidos. Además, proporcionamos estimaciones de error *a priori* que muestran que la solución discreta tendrá un orden óptimo de convergencia siempre que la distancia entre la frontera curva y la frontera computacional permanezca con el mismo orden de magnitud que el parámetro de la malla.

Finalmente, proponemos una formulación que combina el método HDG con método de elementos de frontera (conocido com BEM por sus siglas en inglés) utilizados para un problema más general proveniente de física de plasmas. En ésta situación se desconoce la ubicación del plasma y es necesario resolver la condición de equilibrio en el semiplano para determinar tanto el flujo como la región de confinamiento. El método BEM es ideal para trabajar en dominios no acotados, ya que el enfoque FEM necesitaría un número infinito de elementos para cubrir el dominio.

Agradecimientos

Quiero comenzar expresando mi total agradecimiento a mis padres, Regulo y Jesús Amelia; quienes sacrificaron muchas cosas para darme la mejor educación y de quienes me siento muy orgulloso y les dedico con mucho amor este gran logro. A mis hermanos, Alfonso y Deysi; y mis sobrinos Jian yJohan; quienes me apoyaron emocionalmente desde el momento en que sali a estudiar al extranjero. Agradezco de manera especial a las dos personas que se han convertido en mi principal fuente de inspiración: mi esposa Juana, por el amor que me brinda, su paciencia y comprensión durante todo este tiempo del doctorado; a mi hijo, Néstor Rodrigo, por quién me motivo cada día y lo feliz que me siento de estar a su lado.

Agradezco a mi director de tesis, el profesor Manuel Solano, por su gran calidad humana, por haberse adaptado a mis horarios, por su consejos, su paciencia y hacerme sentir siempre en confianza. Mi total admiración por su profesionalismo y sus conocimientos brindados, sin duda me han hecho crecer tanto en mi desarrollo profesional como personal. Agradezco por estar siempre disponible, por apoyarme, comprenderme y trabajar a mi ritmo, que a pesar de convertirse en padre y tener menos tiempo disponible hizo todo lo posible en concluir de la mejor manera este trabajo. Gracias por todo ese apoyo y dedicación.

A mi co-director de tesis, el profesor Tonatuh Sánchez–Vizuet, que sin conocerme mucho, confió en mi desde el primer momento. Por su disponibilidad y profesionalismo durante todas la reuniones de trabajo. Por ese apoyo incondicional que me brindó durante mi estadia en New York junto a su esposa Lise-Marie y que hizo más fácil mi adaptación, demostrando así su gran calidad humana. Por preocuparse no sólo por mi vida profesional sino también en lo personal. Por todos sus consejos, que sin duda influyeron en poder tomar muy buenas decisiones. Mi sincera admiración en todo lo que hace y la dedicación que le puso en este trabajo.

También quiero agradecer a mis profesores del programa: Gabriel N. Gatica, Raimund Bürger yMauricio Sepúlveda, por todas sus enseñanzas y conocimientos brindados. A los profesores Ricardo Oyarzúa y Luis Gatica, quienes me motivaron a iniciar mis estudios de doctorado.

Agradezco a la Comisión Evaluadora por el tiempo dedicado a revisar esta tesis y por la retroalimentación recibida. También agradezco a los integrantes de la Comisión Examinadora por la disponibilidad para ser jurados de la defensa de tesis.

De igual forma, agradezco la buena disponibilidad del personal administrativo del CI^{2}MA, del Departamento de Ingeniería Matemática, y de la Dirección de Postgrado de la Universidad de Concepción: Lorena Carrasco, Cecilia Leiva, Paola Castro, Jorge Muñoz e Iván Tobar. Un agradecimiento especial al profesor Rodolfo Rodriguez, por su gestión y apoyo durante el tiempo que fue director del Programa.

A mis amigos de Doctorado: Paul, Willian, Iván , Paulo, Rafael, Sergio, Cristian, Rodrigo, Daniel, Victor, Bryan, Patrick, Elvis, Cinthya, Joaquín, Juan Paulo y Romel. Un agradecimiento especial a mis compañeros y amigos de generación: Yissedt, Adrian y Yolanda, con quienes inicié esta etapa y con los cuales guardo gratos recuerdos. A mis compatriotas y alumnos de otros programas: Ángel, Alex, Nolbert, Carlos, Hobby, Heli, Miguel, Nathalie, Yocelyn, con quienes compartí muy buenos momentos e hicieron más fácil mi estancia en Chile.

Agradezco a las instituciones y proyectos que han financiado mis estudios e investigación: Beca de doctorado nacional de la Comisión Nacional de Ciencia y Tecnología (CONICYT), hoy Agencia Nacional de Investigación y Desarrollo (ANID), por el financiamiento mediante el Programa Formación de Capital Humano Avanzado (PFCHA/DOCTORADO NACIONAL/2019-21191566). A los proyectos Fondecyt N° 1160320 y N° 1200569, al Proyecto AFB 170001 del Programa PIA: Concurso Apoyo a Centros Científicos y Tecnológicos de Excelencia con Financiamiento Basal gestionado por el profesor Manuel Solano y el proyecto financiado por el Departamento de Energía de los Estados Unidos a través del proyecto N° DE-FG02-86ER53233, bajo la gestión del profesor Tonatiuh Sánchez–Vizuet. A la dirección de Postgrado de la Universidad de Concepción y al Departamento de Ingeniería Matemática (DIM), por haber financiado mi estadía en el doctorado, y al Centro de Investigación en Ingeniería Matemáica (CI^2MA) por haberme brindado las comodidades de una oficina en un ambiente adecuado necesario para llevar a cabo con éxito la etapa de doctorado.

A todos les dedico este logro.

Nestor Abel Sánchez Goycochea

Contents

List of Tables

List of Figures

Introduction

Interest in the study of plasma physics has grown greatly in recent decades, this is understandable, since plasma is present in 99% of the known universe. In 1927, Irving Langmuir—Nobel Prize winner in Chemistry—first introduced the term "plasma" to refer to an ionized gas, which is formed by subjecting the gas to very high temperatures to such a point that its atoms collide with each other and its electrons are eliminated [40,51]. One of the first applications where plasma was used was made in 1960 to perform styrene polymerization. Later, Holländer *et al.*, studied the industrial processing of polymers by low pressure plasmas [83]. Other applications are in the study of the optical and electrical properties of plasma polymerized silicones [31], the chemical treatment to clean surfaces [28], in odontology [79], etc. In nature, plasma appears in lightning bolts, Northern lights and sun flames ; the Earth itself is contained within a thin plasma called the solar wind and is surrounded by a dense plasma known as the Ionosphere, [55].

A particular application of plasma physics occurs in the nuclear reaction that produces two nuclei of light atoms that merge together to form a heavier nucleus. During this process, known as nuclear fusion, large amounts of energy are released in the form of electromagnetic radiation. In a controlled setting, this type of reaction is possible, for example, in a family of axially symmetrical reactors known as "Tokamak" whose design dates back to the late 1960's by soviet scientists [4,84]. This type of reactor uses magnetic fields generated by an external coil matrix that allows the confinement of the plasma until the reaction occurs, thus avoiding damage to the reactor walls. Due to the cylindrical symmetry of these devices, the magnetic field generated by this type of reactor can be described in terms of two scalar functions: the poloidal flux u and the toroidal field g. The equilibrium between the magnetic pressure and hydrodynamical pressure, p, is described through a differential equation expressed in the following free boundary problem.

$$-\nabla\cdot\left(\frac{1}{\mu\, x}\nabla u\right)=\begin{cases} F(u) & \text{in } \Omega_P(u) \\ I_i & \text{in } \Omega_{C_i} \\ 0 & \text{elsewhere} \end{cases}, \qquad (*)$$

where μ is the magnetic permeability; F is a function that contains the toroidal field g and the hydrodynamic pressure p; I_i are the values of the currents that flow through the external coils located in the domains C_i; and Ω_P is, *a priori* unknown, the region where the plasma is confined.

The first row of equation (*) is known as the Grad–Shafranov equation (or Grad–Shafranov–Schluter equation) [41, 56, 78]. In this work, we focus on the analysis of a more general form of equation (*). We will study separately, the different situations that derive from said equation.

The region where the plasma is confined is a domain enclosed by a level set of the solution, which is a smooth piecewise curve. Due to the non-polygonal nature of this domain, the geometric complexity represents an additional problem when using a discretization scheme. Standard Galerkin methods devised to solve partial differential equations in curved domains Ω (Ω_P in (*)) do not guarantee

high-order convergence when Ω approaches a close domain Ω_h—due to the presence of singularities or high gradients of the continuous solution, which arise from domains with re-entrant corners or boundary layers. Here Ω_h is a polygonal domain that approximates Ω. To deal with the lack of precision in convergence, in the standard literature it is suggested to use "fitted" methods or high order approximations to the boundary. In particular, we can mention isoparametric finite elements (see for example, [53]) where the mesh of Ω_h fits the domain Ω under an explicit parameterization of its boundary $\Gamma = \partial\Omega$. This type of elements can be implemented efficiently without much difficulty. However, they lose precision if the domains are evolutionary (domains that involve time). On the contrary, the "unfitted" methods build Ω_h, putting Ω in a uniform mesh *background* trying to make Ω_h as independent of Γ as possible. One of the first studies where the boundary value is corrected was done by Bramble–Hilbert [6], where the correction is based on what was proposed by Nitsche [63] together with the method of polygonal domain approximation made by Thomée in [80]. Other methods that follow the "unfitted" approach are: the CutFEM method [9–12], or the immersed boundary methods [54, 66]. There are other numerical techniques to improve the higher order precision and which are used in this type of domains [52, 61].

Returning to the problem presented in the equation (*) and taking into account that the region where the plasma remains confined is a non-polygonal domain, we propose a hybridizable discontinuous Galerkin (HDG) discretization. One of the first contributions for Dirichlet boundary value problems proposed in the context of HDG method was made in [21] and improved in [18]. Our approach is to pose the discrete problem in a polygonal subdomain $\Omega_h \subset \Omega$ and transfer the boundary data prescribed on Γ to the computational boundary $\Gamma_h := \partial\Omega_h$ using a family of segments called *transfer paths*. This technique was proposed for one-dimensional problems in [17] and involves a line integral of the numerical flow.

In **Chapter** 2, we consider the magnetic permeability μ as constant and only the source term F depends on u, resulting in a semilinear elliptic problem posed in a non-polygonal domain Ω. We approximate Ω by a polygonal subdomain Ω_h using transfer techniques and propose a high-order hybridizable discontinuous Galerkin method. We show optimal convergence under minor assumptions on the non-linear source term, and the distance between the boundaries of the polygonal subdomain Ω_h and the original domain Ω. Furthermore, we propose a non-linear local post-processing of the unknown scalar function u providing an additional order of convergence. A reliable and locally efficient *a posteriori* error estimator that takes into account the error in the approximation of the boundary data is also provided . This first contribution was accepted in

[71] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, *A priori and a posteriori error analysis of an unfitted HDG method for semi-linear elliptic problems in curved domains.* **Numerische Mathematik,** 148 (2021), pp. 919–958.

As we mentioned earlier, in the presence of ferroelectric materials, permeability is affected by the magnetic field $\mathbf{B}$ which is proportional to the gradient of u taking the form $\mu = \mu(\nabla u)$ and leads to a quasi-linear equation that requires more detailed treatment. Recently, there have been some theoretical studies of the HDG method applied to quasilinear problems [27, 37, 38], however, these efforts are limited to polygonal domains. Furthermore, the first reference does not consider non-linearities where the diffusion coefficient depends on ∇u, while in [37, 38], the authors analyzed an

augmented HDG discretization for a strictly quasi-linear problem which arises from a non-linear Stokes flow under an approach based on a non-linear version of the Babuška–Brezzi theory. As we will show, our analysis will be valid for both quasi-linear and semi-linear problems, without requiring an augmented formulation, and the domain can be piecewise smooth.

In the event that an iron component appears, the permeability becomes a function dependent on the magnitude of the magnetic field, that is:

$$\mu = \begin{cases} \mu_0 & \text{in vacuum} \\ \mu(|\nabla u|^2/x^2) & \text{in iron} \end{cases},$$

where μ_0 is the magnetic permeability of the vacuum, this case is analyzed in Chapter 3. More precisely, in **Chapter 3** we studied HDG discretizations for a class of nonlinear elliptic boundary value problems posed on curved domains where both the source term and the diffusion coefficient are not linear. Here, we focus on situations where the source term is independent of u and μ takes one of the forms given in (3.1c). Then, we proceed to study separately the HDG discretizations for the case where the diffusion coefficient depends only on u (Section 3.2), and the case where the coefficient depends on ∇u (Section 3.3). We also show that, under proper assumptions about the source term and the computational domain, the discrete systems are well posed. In addition, we provide *a priori* error estimates that guarantee that the discrete solution has an optimal order of convergence as long as the distance between the curved boundary and the computational boundary are of the same order of magnitude as the mesh parameter. The preprint version of this work is:

[70] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, *Error analysis of an unfitted HDG method for a class of non-linear elliptic problems* (Submitted). Preprint 2021-13, Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción, Chile, Preprint available at **https://www.ci2ma.udec.cl/ publicaciones/prepublicaciones/prepublicacion.php?id=451**.

Lastly, in **Chapter 4** we study the coupled problem given in the equation (*) combining a finite element method with a boundary element method. This formulation is geared towards the solution of a variant of the problem known as *the free boundary problem* [32]. In this situation the location of the plasma is unknown and it is necessary to solve the equilibrium condition in the semi-plane to determine both u and the confinement region. We propose an unfitted discretization scheme that couples HDG with the boundary element method (BEM) for the solution to a non-linear problem posed in an unbounded domain. The transfer of information between the non-touching grids is done via the method of transfer paths and the coupling is done using Costabel's symmetric approach [26]. The unfitted computational domain breaks the symmetry of the scheme and we are able to show that under a suitable local proximity condition on the grids, the influence of the perturbation vanishes as the mesh parameter tends to zero. Finally we show that if the sources have small Lipschitz constant, the nonlinear discrete problem is well posed. This work will result in two separate articles that are in preparation:

[68] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, Afternote to "*Coupling at a distance*": convergence analysis and *a priori* error estimates. (*Dedicated to the memory of Francisco–Javier Sayas*). In preparation.

[69] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, Analysis of a coupled *HDG-BEM* formulation for non-linear elliptic problems with curved interfaces. In preparation.

In the opening **Chapter 1**, we introduce basic notations on Sobolev spaces and introduce the idea of extended domains and *transfer paths*. In this chapter we will also describe some geometric hypotheses about the computational domain that will be useful for the solvability of the problem. We also present a dual problem and the HDG projection which will both be necessary for the *a priori* error estimates, and the Clément and Oswald interpolants used in the error estimates *a posteriori*. In the general case, studied in this work, the diffusion coefficient will be denoted by κ and the Grad–Shafranov equation is recovered by making $\kappa = 1/(\mu\, x)$.

Introducción

El interés por el estudio de física de plasmas ha crecido de gran manera en las últimas décadas, ésto es entendible, pues el plasma está presente en el 99 % del universo conocido . En 1927, Irving Langmuir—ganador del Premio Nobel en Química—introdujo por primera vez el término "plasma" para referirse a un gas ionizado, el cual se forma al someter el gas a temperaturas muy elevadas a tal punto que sus átomos chocan entre sí y sus electrones son eliminados [40, 51]. Una de las primeras aplicaciones donde se usó plasma, fue hecha en 1960 para realizar polimerización de estireno. Más adelante, Holländer y colaboradores, estudiaron el procesamiento industrial de polímeros por plasmas de baja presión [83]. Otras aplicaciones se dan en el estudio de las propiedades ópticas y eléctricas de las siliconas polimerizadas por plasma [31], el tratamiento químico para realizar limpiezas de superficies [28], en la odontología [79], etc. En la naturaleza, el plasma aparece en los relámpagos, las auroras boreales y las llamas del sol; la Tierra misma se encuentra dentro de un plasma delgado llamado viento solar y está rodeada por un plasma denso conocido como Ionósfera, [55].

Una aplicación particular de física de plasmas se produce en la reacción nuclear que producen dos núcleos de átomos ligeros que se fusionan para formar un núcleo más pesado. Durante éste proceso, conocido como fusión nuclear, se liberan grandes cantidades de energía en forma de radiación electromagnética. En un entorno controlado, este tipo de reacción es posible, por ejemplo, en una familia de reactores axialmente simétricos conocidos como "Tokamak" y cuyo diseño se remonta a finales de la década de 1960 realizada por científicos soviéticos [4, 84]. Éste tipo de reactores utilizan campos magnéticos generados por una matriz externa de bobinas y permiten el confinamiento del plasma hasta que la reacción se produzca, evitando así daños en las paredes del reactor. Debido a la simetría cilíndrica de éstos dispositivos, el campo magnético generado por éste tipo de reactores pueden describirse en términos de dos funciones escalares: el flujo poloidal u y el campo toroidal g. El equilibrio ente la presión magnética y la presión hidrodinámica, p, es descrita a través de una ecuación diferencial expresada en el siguiente problema de frontera libre.

$$-\nabla\cdot\left(\frac{1}{\mu\, x}\nabla u\right)=\begin{cases} F(u) & \text{en } \Omega_P \\ I_i & \text{en } \Omega_{C_i} \\ 0 & \text{en otras partes}\end{cases}, \qquad (*)$$

donde, μ es la permeabilidad magnética; F es una función que contiene al campo toroidal g y la presión hidrodinámica p; I_i son los valores de las corrientes que atraviezan las bobinas externas localizadas en los dominios C_i; y Ω_P es, *a priori* desconocida, la región donde el plasma es confinado.

La primera fila de la ecuación (*) es conocida como ecuación de Grad–Shafranov (o ecuación de Grad–Shafranov–Schlüter) [41, 56, 78]. En éste trabajo, nos enfocamos en el análisis de una manera general la forma de la ecuación (*). Estudiaremos de manera separada, las diferentes situaciones que se derivan de dicha ecuación.

La región donde el plasma es confinado es un dominio encerrado por el conjunto de nivel cero

de la solución, el cual es una curva suave por partes. Debido a la naturaleza no poligonal de este dominio, la complejidad geométrica representa un problema adicional cuando se usa un esquema de discretización. Los métodos estándar de Galerkin ideados para resolver ecuaciones diferenciales parciales en dominios curvos Ω (Ω_P en (*)) no garantizan una convergencia de orden alto cuando Ω se aproxima a un dominio cercano Ω_h—debido a la presencia de singularidades o altos gradientes de la solución continua, que surgen de dominios con esquinas re-entrantes o capas límite——. Aquí Ω_h es un dominio poligonal que aproxima a Ω. Para lidiar con la falta de precisión en la convergencia, se sugieren emplear métodos "fitted" o aproximaciones de alto orden para la frontera. En particular, podemos mencionar a los elementos finitos isoparamétricos (ver por ejemplo, [53]) donde la malla de Ω_h se ajusta al dominio Ω bajo una parametrización explícita de su frontera $\Gamma = \partial\Omega$. Éste tipo de elementos pueden implementarse sin mucha dificultad de manera eficiente. Sin embargo, pierden precisión si los dominios son evolutivos (dominios que involucran tiempo). Por el contrario, los métodos "unfitted" construyen Ω_h, poniendo Ω en una malla unifrome *background* tratando de que Ω_h sea tan independiente de Γ como sea posible. Una de los primeros estudios donde se corrije el valor de frontera fue hecho por Bramble–Hilbert [6], donde la corrección se basa en lo propuesto por Nitsche [63] junto con el método de aproximación de dominio poligonal hecho por Thomée en [80]. Otros métodos que siguen el enfoque "unfitted"son: el método CutFEM [9–12], o los métodos immersed boundary [54,66]. Existen otras técnicas numéricas para mejorar la precisión de orden superior y que son usados en éste tipo de dominios [52, 61].

Volviendo al problema presentado en la ecuación (*) y teniendo en cuenta que la región donde el plasma permanece confinado es un dominio no poligonal, proponemos una discretización Galerkin Discontinuo Hibridizable (HDG). Una de las primeras contribuciones para problemas con valores de forntera Dirichlet propestas en el contexto del método HDG fue hecha en [21] y mejorada en [18]. Nuestro enfoque consiste en plantear el problema discreto en un subdominio poligonal $\Omega_h \subset \Omega$ y transferir los datos de frontera Γ a la frontera computacional $\Gamma_h := \partial\Omega_h$ usando una familia de segmentos, llamados *caminos de transferencia*. Ésta técnica fue propuesta para problemas unidimensionales en [17] e involucra una integral de línea del flujo numérico.

En el **Capítulo 2** consideramos la permeabilidad magnética μ como constante y sólo el término fuente F depende de u, resultando en un problema elíptico semilineal planteado en un dominio no poligonal. Aproximamos Ω mediante un subdominio poligonal Ω_h usando técnicas de tranferencia y proponemos un método de Galerkin discontinuo hibridizable. Mostramos convergencia óptima bajo supusiciones pequeñas, dadas sobre el término fuente no lineal y la distancia entre las fronteras del subdominio poligonal Ω_h y el dominio original Ω. Además, proponemos un posprocesamiento local no lineal de la función escalar desconocida u proporcionando un orden adicional de convergencia. Probamos también, un estimador de error *a posteriori* confiable y localmente eficiente que toma en cuenta el error en la aproximación de los datos de la frontera. Ésta primera contribución está aceptada en :

[71] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, *A priori and a posteriori error analysis of an unfitted HDG method for semi-linear elliptic problems in curved domains*. **Numerische Mathematik,** 148 (2021), pp. 919–958.

Como mencionamos anteriormente, en presencia de materiales ferroeléctricos, la permeabilidad se ve afectada por el campo magnético $\mathbf{B}$ que es proporcional al gradiente de u tomando la forma $\mu = \mu(\nabla u)$, lo que lleva a una ecuación cuasi-lineal que requiere un tratamiento más detallado. Recientemente se han realizado algunos estudios teóricos del método HDG aplicado a problemas cuasilineales [27,37,38], sin embargo, éstos esfuerzos se limitan a los dominios poligonales. Además, la primera referencia no considera las no linealidades donde el coeficiente de difusión depende de ∇u, mientras que en [37,38], los autores analizaron una discretización HDG aumentada para un problema estrictamente cuasi lineal que surge a partir de un flujo de Stokes no lineal bajo un enfoque basado en una versión no lineal de la teoría de Babuška–Brezzi. Como mostraremos, nuestro análisis será válido tanto para problemas cuasi-lineales como semi-lineales, sin requerir una formulación aumentada y el dominio puede ser suave por partes.

En el caso de que aparezca un componente de hierro, la permeabilidad pasa a ser una función dependiente de la magnitud del campo magnético, es decir

$$\mu = \begin{cases} \mu_0 & \text{en el vacío} \\ \mu(|\nabla u|^2/x^2) & \text{en el hierro} \end{cases},$$

donde μ_0 es la permeabilidad magnética del vacío, este caso es analizado en el **Capítulo 3**. Más precisamente, en el **Capítulo 3** estudiamos discretizaciones HDG para problemas elípticon con valores de forntera no lineal planteados sobre dominios curvos, donde tanto el término fuente como el coeficiente de difusión son no lineales. Aquí nos enfocamos en situaciones donde el término fuente es independiente de u y μ toma una de las formas dadas en (3.1c). Luego, procedemos a estudiar por separado las discretizaciones HDG para el caso donde el coeficiente de difusión depende sólo de u (Sección 3.2), y el caso donde el coeficiente depende de ∇u (Sección 3.3). También mostramos que, bajo supusiciones adecuadas sobre el término fuente y el dominio computacional, los sistemas discretos están bien planteados. Además, proporcionamos estimaciones de error *a priori* que garantizan que la solución discreta tenga un orden de convergencia óptimo siempre que la distancia entre la frontera curva y la frontera computacional sean del mismo orden de magnitud que el parámetro de la malla. La versión preprint de este trabajo es:

[70] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, *Error analysis of an unfitted HDG method for a class of non-linear elliptic problems.*. Preprint 2021-13, Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción, Chile, Preprint available at **https://www.ci2ma.udec.cl/ publicaciones/prepublicaciones/prepublicacion.php?id=451**.

Por último, en el **Capítulo 4**, estudiamos el problema acoplado dado en (*) combinando un método de elementos finitos con un método de elementos de contorno. Ésta formulación está orientada a la solución de una variante del problema conocido como *problema de frontera libre* [32]. En ésta situación, se desconoce la ubicación del plasma y es necesario resolver la condición de equilibrio en el semiplano para determinar tanto u como la región de confinamiento. Proponemos un esquema de discretización no ajustado que combina HDG con el método de elementos de frontera (BEM) para la solución de un problema no lineal planteado en un dominio no acotado. La transferencia de información entre las mallas que no se tocan se realiza mediante el método de los caminos de transferencia y el acoplamiento

se realiza utilizando el enfoque simétrico de Costabel [26]. El dominio computacional no ajustado rompe la simetría del esquema, pero podemos demostrar que bajo una condición de proximidad local en las mallas, la influencia de la perturbación se desvanece cuando el parámetro de la malla tiende a cero. Finalmente, mostramos que si las fuentes tienen una constante de Lipschitz pequeña, el problema discreto no lineal está bien planteado. Éste trabajo resultará en dos publicaciones que se encuentran en preparación:

[68] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, Afternote to "*Coupling at a distance*": convergence analysis and *a priori* error estimates. (*Dedicated to the memory of Francisco–Javier Sayas*). In preparation.

[69] Nestor Sánchez, Tonatiuh Sánchez-Vizuet and Manuel E. Solano, Analysis of a coupled *HDG-BEM* formulation for non-linear elliptic problems with curved interfaces. In preparation.

En el inicio del **Capítulo 1**, introducimos notaciones básicas sobre espacios de Sobolev y las ideas de dominios extendidios y *caminos de trnasferencia*. En éste capítulo describimos también algunas hipótesis geométricas acerca del dominio computacional que será usado para la solubilidad del problema. Presentamos también un problema dual y las proyecciones HDG, necesarias para las estimaciones de error *a priori*, y los interpolantes de Clément y Oswald usados en las estimaciones de error *a posteriori*. En el caso general que se considera en éste trabajo, el coeficiente de difusión será denotado por κ y la ecuación de Grad–Shafranov es recuperada haciendo $\kappa = 1/(\mu x)$

CHAPTER 1

Preliminaries

In this chapter we introduce basic notations on Sobolev spaces and introduce the idea of extended domains and transfer paths. In this chapter we will also describe some geometric hypotheses about the computational domain that will be useful for the solubility of the problem. We also present a dual problem and the HDG projection necessary for the *a priori* error estimates. and the Clément and Oswald interpolants used in the error estimates *a posteriori*.

1.1 Computational domains and admissible triangulations

Given a domain $\Omega \subseteq \mathbb{R}^d$ with $d \in \{2, 3\}$, we will define a family of polygonal subdomains and admissible triangulations approximating Ω where we will ultimately pose our discretization. First, consider a family of simply connected domains $\{\Omega_\alpha\}_{\alpha>0}$ such that, for every α the following conditions hold: (1) $\Omega_\alpha \subseteq \Omega$, (2) the boundary $\Gamma_\alpha := \partial\Omega_\alpha$ is a polygon, and (3) for every $\epsilon > 0$ there are infinitely many indices α such that $\lambda(\Omega \backslash \Omega_\alpha) < \epsilon$. In the preceding expression $\lambda(\cdot)$ denotes the Lebesgue measure. These conditions ensure that the family of subdomains $\{\Omega_\alpha\}_{\alpha>0}$ will exhaust Ω.

Having built the family $\{\Omega_\alpha\}_{\alpha>0}$ satisfying all the conditions above, the next step is to define a family of admissible simplicial triangulations $\{\mathcal{T}_h\}_{h>0}$. To be considered admissible, a triangulation $\mathcal{T}_h$ must be such that: (1) $\mathcal{T}_h$ is a triangulation for at least one $\Omega_h \in \{\Omega_\alpha\}_{\alpha>0}$ (we will identify both $\mathcal{T}_h$ and the respective domain Ω_h with the same subscript h, adding copies of Ω_h to $\{\Omega_\alpha\}_{\alpha>0}$ if necessary to account for different triangulations of the same domain) (2) it is shape regular, meaning that there exists $\beta > 0$ such that for all elements $T \in \mathcal{T}_h$ and all $h > 0$, $h_T/\rho_T \leq \beta$, where h_T is the diameter of T and ρ_T is the diameter of the largest ball contained in T, and (3) for every $T \in \mathcal{T}_h$ such that $T \cap \Gamma_h \neq \varnothing$, the maximum distance between $\boldsymbol{x} \in T \cap \Gamma_h$ and $\boldsymbol{y} \in \Gamma$ is of the same order of magnitude as the element diameter h_T. More precisely, if $d_{loc} := \max\{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in T \cap \Gamma_h \text{ and } \boldsymbol{y} \in \Gamma\}$ then $d_{loc} = \mathcal{O}(h_T)$. This last requirement, which will be referred to as the *local proximity condition* and is depicted schematically in Figure 1.1, is of key importance for the transfer process that will be defined later on.

For every element T in a particular triangulation, we will denote by $\boldsymbol{n}_T$ the outward unit normal vector to T, or simply $\boldsymbol{n}$ instead of $\boldsymbol{n}_T$ whenever the context prevents any confusion. As it is conven-

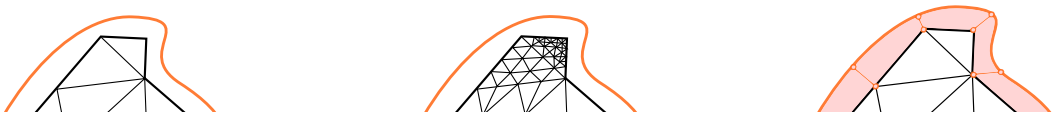

Figure 1.1: *Left and center: The proximity condition ensures that the distance between the computational and the physical boundaries remains always of the same order of magnitude as the local element diameter. The schematic shows close ups to the boundary of two triangulations of the same domain: an admissible triangulation satisfying the local proximity condition (left), and inadmissible one that violates the local proximity condition (center). Right: An admissible triangulation and one possible arrangement of extension patches T^e_{ext} (shaded in the figure) defined on the region $\Omega \setminus \Omega_h$.*

tional, we will denote the mesh parameter as $h := \max_{T\in\mathcal{T}_h} h_T$, which will be assumed to be smaller than one for the sake of simplicity. We will denote by e any face of a simplex and its length by h_e. Moreover, we will talk about an *interior face* e if there are two elements T^+ and T^- in the triangulation $\mathcal{T}_h$ such that $e = \partial T^+ \cap \partial T^-$. The set of all interior faces will be denoted by $\mathcal{E}_h^\circ$. In a similar manner, we will talk about a *boundary face* e if there is an element $T \in \mathcal{T}_h$ such that $e = \partial T \cap \Gamma_h$; the set of boundary faces will be denoted by $\mathcal{E}_h^\partial$. Note that with these definitions, the entirety of the faces of the triangulation denoted by $\mathcal{E}_h$ (often referred to as the *skeleton* of the mesh) can then be decomposed as $\mathcal{E}_h = \mathcal{E}_h^\circ \cup \mathcal{E}_h^\partial$.

Henceforth, we will be working with functions that in general will not be continuous across mesh elements. For a scalar-valued function we will use the symbol $[\![w]\!] := w^+ - w^-$ to refer to its jump across any given interior face. At the boundary faces, the jump will be defined as $[\![w]\!] := w - \varphi_h$, where φ_h is the approximation of the boundary data at Γ_h that will be defined later. In the case of vector-valued functions $\boldsymbol{v}$, we will be interested in the discontinuity of its normal component across interior faces, which will be denoted by $[\![\boldsymbol{v}]\!] := \boldsymbol{v}^+ \cdot \boldsymbol{n}^+ + \boldsymbol{v}^- \cdot \boldsymbol{n}^-$.

A remark on the local proximity condition and mesh refinement: The local proximity condition limits the minimum size that the elements near the boundary of an admissible triangulation can attain. Therefore, mesh refinement in this context must be understood as moving through a sequence of computational domains in the set $\{\Omega_h\}_{h>0}$ and their corresponding admissible triangulations in $\{\mathcal{T}_h\}_{h>0}$ as the parameter $h \to 0$. As it will be shown later, the error estimates will not depend on the particular domain Ω_h or triangulation $\mathcal{T}_h$ as long as the three requirements on the mesh stated above are satisfied. Possible ways of building sequences of admissible triangulations and computational domains have been detailed in [21] for uniform meshes and in [74] for adaptively refined triangulations.

1.2 The extended domain

Having defined the family of polygonal subdomains and admissible triangulations on which the discretization will be performend, we will now proceed to detail the process through which the boundary information will be transferred from the boundary into the computational domain. In order to do that we will have to tessellate the region enclosed between the two boundaries Γ and Γ_h as follows.

Given a triangulation $\mathcal{T}_h$ of the computational domain Ω_h and a boundary face $e \in \mathcal{E}_h^\partial$, we will

denote by T^e the unique element of $\mathcal{T}_h$ such that $e \cap \overline{T^e} = e$. To every point $\boldsymbol{x} \in e$, we will associate a point $\overline{\boldsymbol{x}} \in \Gamma$ and set $l(\boldsymbol{x}) = |\boldsymbol{x} - \overline{\boldsymbol{x}}|$. We will define the *extension patch* T^e_{ext} as

$$T^e_{ext} := \{\boldsymbol{x} + s\boldsymbol{t} : 0 \leq s \leq l(\boldsymbol{x}), \boldsymbol{x} \in e\},$$

where $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is the unit vector anchored at $\boldsymbol{x}$ and pointing in the direction of $\overline{\boldsymbol{x}}$. With this notation, the line segment connecting $\boldsymbol{x}$ to $\overline{\boldsymbol{x}}$ can be parameterized by

$$\sigma_{\boldsymbol{t}}(\boldsymbol{x}) := \{\boldsymbol{x} + s\boldsymbol{t} : s \in [0, l(\boldsymbol{x})]\}.$$

The point $\overline{\boldsymbol{x}} \in \Gamma$ and therefore the vector $\boldsymbol{t}(\boldsymbol{x})$ can be specified in several ways. Here, we will consider that the point has been determined in such a way that

$$\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{n} \text{ for all } \boldsymbol{x} \in e. \tag{1.1}$$

This assumption is made with the sole purpose of making the analysis simpler, it can in fact be relaxed to the existence of a constant a_0 such that $0 < a_0 \leq \boldsymbol{t}(\boldsymbol{x}) \cdot \boldsymbol{n}$ for every $\boldsymbol{x}$ belonging to a boundary edge. The numerical method described here is remarkably robust with respect to the method used to choose $\boldsymbol{t}(\boldsymbol{x})$. Previously, the direction had been determined using the algorithm proposed by [21], which assigns $\overline{\boldsymbol{x}}$ in such a way that the three following conditions are satisfied: (1) $\overline{\boldsymbol{x}}$ is unique, (2) any two different line segments $\sigma_{\boldsymbol{t}}$ do not intersect each other inside T^e_{ext}, and (3) the segments $\sigma_{\boldsymbol{t}}$ do not intersect the interior of Ω_h. As it is proven in the aforementioned reference, these three conditions guarantee that the union of T^e_{ext} completely covers $\Omega^c_h := \Omega \setminus \overline{\Omega_h}$. We would like to point out that the algorithm developed in [21] for the two-dimensional case, always produces a family of connecting segments satisfying the aforementioned conditions, independently of how complicated the boundary is, since it makes use of a background mesh where the domain Ω is immerse. The same ideas can be extended to three dimensions. On the other hand, the first condition is not essential and is only required to simplify the analysis and the computational implementation. If it is not satisfied, Ω^c_h will be composed by overlapping extension patches. Then, we can consider the average of the extrapolated polynomials that share an overlapped region. An alternate method was used and tested numerically in [73,74], where $\boldsymbol{t}(\boldsymbol{x})$ was determined using a weighted average of the normal vectors from neighboring boundary edges.

For an extended patch T^e_{ext} and a mesh element $T^e \in \mathcal{T}_h$ sharing a boundary face $e \in \mathcal{E}^\partial_h$, we denote by $h^\perp_e$ (resp. $H^\perp_e$) the largest distance between a point inside T_e (resp. T^e_{ext}) and the plane determined by the face e. The ratio between these two distances will be denoted by $r_e := H^\perp_e / h^\perp_e$ and the maximum such ratio taken over all the boundary edges will be denoted by $R_{\mathcal{T}_h} := \max_{e \in \mathcal{E}^\partial_h} r_e$. We will refer to this quantity as the *proximity parameter* of a geometric discretization.

The proximity parameter $R_{\mathcal{T}_h}$ plays a key role in many of the error and convergence estimates in this work, therefore a few remarks on its properties are in order. By definition, for any fixed geometric discretization pair $(\Omega_h, \mathcal{T}_h)$, the quantity $R_{\mathcal{T}_h}$ is constant, however its value may change between any two given pairs $(\Omega_{h_1}, \mathcal{T}_{h_1})$ and $(\Omega_{h_2}, \mathcal{T}_{h_2})$. Therefore, if a family of geometric discretizations is labeled by the parameter h, then the dependence of $R_{\mathcal{T}_h}$ with respect to the geometric discretization may be

expressed succinctly with the notation

$$R_h := \max_{e \in \mathcal{E}_h^\partial} r_e, \tag{1.2}$$

which we shall prefer over the alternate $R_{\mathcal{T}_h}$ whenever the dependence on h needs to be made explicit. We stress the fact that, in the definition above, *varying h must be understood as varying the discretization pair* $(\Omega_h, \mathcal{T}_h)$. It follows from the local proximity condition discussed in the previous section that, for any admissible family of discretizations $(\Omega_h, \mathcal{T}_h)$, The proximity parameter must satisfy

$$0 \leq R_h \leq Ch^p \quad \text{for } p > 0, \tag{1.3}$$

which implies that for an admissible family of discretizations $\{(\Omega_h, \mathcal{T}_h)\}_{h \geq 0}$, the proximity parameter vanishes at least with order h^p, but may be identically zero or vanish as a larger power of h.

We will also define the class of non-trivial vector-valued polynomials of degree at most k defined in and across both patches as

$$\mathcal{V}^k := \left\{ \boldsymbol{p} \in [\mathbb{P}_k(T^e_{ext} \cup T^e)]^2 \,:\, \boldsymbol{p} \cdot \boldsymbol{n}_e \neq \boldsymbol{0} \right\}.$$

We can then introduce, for all those elements with a non-empty intersection with the computational boundary, the element-wise constants

$$C^e_{ext} := \frac{1}{\sqrt{r_e}} \sup_{\boldsymbol{\chi} \in \mathcal{V}^k} \frac{\|\boldsymbol{\chi} \cdot \boldsymbol{n}_e\|_{T^e_{ext}}}{\|\boldsymbol{\chi} \cdot \boldsymbol{n}_e\|_{T^e}} \quad \text{and} \quad C^e_{inv} := h_e^\perp \sup_{\boldsymbol{\chi} \in \mathcal{V}^k} \frac{\|\nabla \boldsymbol{\chi} \cdot \boldsymbol{n}_e\|_{T^e}}{\|\boldsymbol{\chi} \cdot \boldsymbol{n}_e\|_{T^e}}, \tag{1.4}$$

where, in abuse of notation, $\boldsymbol{n}_e$ is a constant vector field defined in $T^e_{ext} \cup T^e \cup e$ that coincides with the unit exterior normal vector associated to the face e and pointing in the direction of the extension patch. Above, the norms $\|\cdot\|_{T^e_{ext}}$ and $\|\cdot\|_{T^e}$ are the standard L^2 norms supported on the extension patch T^e_{ext} and its neighboring element T^e respectively. In [18], these constants were bounded in terms of the polynomial degree of the approximation, k, and the regularity constant of the mesh, β, as

$$C^e_{ext} \leq C_1 (k+1)^2 (3\beta + 2)^k \qquad \text{and} \qquad C^e_{inv} \leq C_2 k^2, \tag{1.5}$$

where C_1 and C_2 depend only on the mesh regularity. As we will see below, these constants will determine the magnitude of the proximity constant and therefore the maximum admissible gap between the computational and physical boundaries.

1.3 The transfer paths

Having established the requirements for an admissible triangulation, and defined the extension patches T^e_{ext} in such a way that for each of them there corresponds a single $T^e \in \mathcal{T}_h$, we can define a way to extend polynomial functions from the computational domain into Ω^c_h. This extension process will enable us to transfer the boundary condition from Γ into the computational boundary Γ_h. Let $p : T^e \to \mathbb{R}$ be a polynomial function and T^e_{ext} an extension patch associated to T^e. We will define the

extension $\boldsymbol{E}_h(p)$ of p to T^e_{ext} by extrapolation as follows:

$$\begin{aligned} E_h: \quad \mathbb{P}_k(T^e) &\longrightarrow \mathbb{P}_k(T^e \cup T^e_{ext}) \\ p(\boldsymbol{y})\ \forall\, \boldsymbol{y} \in T^e &\longmapsto p(\boldsymbol{y})\ \forall\, \boldsymbol{y} \in T^e \cup T^e_{ext}. \end{aligned}$$

Where, to keep notation simple, a polynomial function p should be understood as its extrapolation $E_h(p)$ whenever an evaluation outside of Ω_h is required, which should be clear from the context. For vector-valued polynomial functions, the extension is defined similarly component by component.

1.4 Sobolev space notation

To denote spaces of functions we will make use of the standard notation and terminology from Sobolev space theory. Let $\mathcal{O}$ be a domain in $\mathbb{R}^d$, and Σ be either a Lipschitz curve (if $d = 2$) or surface (if $d = 3$); for scalar-valued functions and non zero real numbers s, we will use the spaces $H^s(\mathcal{O})$ and $H^s(\Sigma)$ with their usual definition, whereas for the case $s = 0$ we will write simply $L^2(\mathcal{O})$ and $L^2(\Sigma)$. The spaces of vector-valued functions will be denoted in bold face, therefore $\boldsymbol{H}^s(\mathcal{O}) := [H^s(\mathcal{O})]^d$ and $\boldsymbol{H}^s(\Sigma) := [H^s(\Sigma)]^d$.

The L^2 inner products for both scalar and vector-valued functions on volumes and surfaces will be denoted by $(\cdot,\cdot)_{\mathcal{O}}$ and $\langle\cdot,\cdot\rangle_{\Sigma}$ respectively. The associated norms will be denoted by $\|\cdot\|_{s,\mathcal{O}}$ and $\|\cdot\|_{s,\Sigma}$ and simply $\|\cdot\|_{\mathcal{O}}$ for the case $s = 0$. As is common, will we write $|\cdot|_{s,\mathcal{O}}$ for the $\boldsymbol{H}^s$ and H^s-semi norms.

We denote by $\gamma : H^1(\mathcal{O}) \to L^2(\partial\mathcal{O})$ the trace operator and set $H^{1/2}(\partial\mathcal{O}) := \gamma(H^1(\mathcal{O}))$, the space of traces of $H^1(\mathcal{O})$-functions, endowed with the norm

$$\|\mu\|_{1/2,\partial\mathcal{O}} := \inf\left\{\|v\|_{1,\mathcal{O}} : v \in H^1(\mathcal{O}) \,\text{such that}\, \gamma(v) = \mu\right\}.$$

The dual space of $H^{1/2}(\partial\mathcal{O})$ will be denoted by $H^{-1/2}(\partial\mathcal{O})$ and we endow it with the induced norm $\|\cdot\|_{-1/2,\partial\mathcal{O}}$.

Given a triangulation $\mathcal{T}_h$ we will define the following mesh-dependent inner products over elements and edges

$$(\cdot,\cdot)_{\mathcal{T}_h} := \sum_{T\in\mathcal{T}_h} (\cdot,\cdot)_T, \qquad \langle\cdot,\cdot\rangle_{\partial\mathcal{T}_h} := \sum_{T\in\mathcal{T}_h} \langle\cdot,\cdot\rangle_{\partial T} \quad \text{and} \quad \langle\cdot,\cdot\rangle_{\Gamma_h} := \sum_{e\in\mathcal{E}_h^\partial} \langle\cdot,\cdot\rangle_e.$$

These inner products induce mesh-dependent norms that will be denoted, respectively, by

$$\|\cdot\|_{\Omega_h} := \left(\sum_{T\in\mathcal{T}_h} \|\cdot\|_T^2\right)^{1/2}, \qquad \|\cdot\|_{\partial\mathcal{T}_h} := \left(\sum_{T\in\mathcal{T}_h} \|\cdot\|_{\partial T}^2\right)^{1/2} \quad \text{and} \quad \|\cdot\|_{\Gamma_h} := \left(\sum_{e\in\mathcal{E}_h^\partial} \|\cdot\|_e^2\right)^{1/2}.$$

In the forthcoming analysis, the expression $a \lesssim b$ should be understood as meaning $a \le Cb$ where C is a positive constant independent of h.

For the discrete formulations that will be introduced in the next sections, we will make use of the

following finite dimensional spaces of piece-wise polynomial functions

$$\boldsymbol{V}_h := \{\boldsymbol{v} \in \boldsymbol{L}^2(\mathcal{T}_h) : \boldsymbol{v}|_T \in [\mathbb{P}_k(T)]^d,\ \forall\ T \in \mathcal{T}_h\}, \tag{1.6a}$$

$$W_h := \{w \in L^2(\mathcal{T}_h) : w|_T \in \mathbb{P}_k(T),\ \forall\ T \in \mathcal{T}_h\}, \tag{1.6b}$$

$$M_h := \{\mu \in L^2(\mathcal{E}_h) : \mu|_T \in \mathbb{P}_k(e),\ \forall\ e \in \mathcal{E}_h\}, \tag{1.6c}$$

where, $\mathbb{P}_k(T)$ denotes the space of polynomials of degree at most k defined in $T \in \mathcal{T}_h$. Similarly, $\mathbb{P}_k(e)$ denotes the space of polynomials of degree at most k defined over a face $e \in \mathcal{E}_h$.

1.5 Dual Problem

We present an auxiliary problem that generalizes the result of Lemma 3.3 in [18] to our semi-linear case. We will consider that, given $\Theta \in L^2(\Omega)$, the solution to the auxiliary problem

$$\kappa^{-1}\boldsymbol{\phi} + \nabla\psi = 0 \qquad \text{in } \Omega, \tag{1.7a}$$

$$\nabla \cdot \boldsymbol{\phi} = \Theta \qquad \text{in } \Omega, \tag{1.7b}$$

$$\psi = 0 \qquad \text{on } \partial\Omega, \tag{1.7c}$$

satisfies the regularity estimate

$$\|\boldsymbol{\phi}\|_{H^1(\Omega)} + \|\psi\|_{H^2(\Omega)} \leq C_{reg}\|\Theta\|_\Omega. \tag{1.8}$$

1.6 HDG projection

In order to make this manuscript self-contained, in this section we provide previous results that will help us to analyze our discrete scheme. First of all we recall the HDG projection operators introduced by [16]. Given constants $l_u, l_{\boldsymbol{q}} \in [0, k]$ and a pair of functions $(\boldsymbol{q}, u) \in H^{1+l_q}(T) \times H^{1+l_u}(T)$, we denote by $\boldsymbol{\Pi}(\boldsymbol{q}, u) := (\boldsymbol{\Pi}_{\mathrm{v}}\boldsymbol{q}, \Pi_{\mathrm{w}}u)$ the projection over $\boldsymbol{V}_h \times W_h$ defined as the unique element-wise solutions of

$$(\boldsymbol{\Pi}_{\mathrm{v}}\boldsymbol{q}, \boldsymbol{v})_T = (\boldsymbol{q}, \boldsymbol{v})_T \qquad \forall\ \boldsymbol{v} \in [\mathbb{P}_{k-1}(T)]^d, \tag{1.9a}$$

$$(\Pi_{\mathrm{w}}u, w)_T = (u, w)_T \qquad \forall\ w \in \mathbb{P}_{k-1}(T), \tag{1.9b}$$

$$\langle \boldsymbol{\Pi}_{\mathrm{v}}\boldsymbol{q} \cdot \boldsymbol{n} + \tau\Pi_{\mathrm{w}}u, \mu\rangle_F = \langle \boldsymbol{q} \cdot \boldsymbol{n} + \tau u, \mu\rangle_F \qquad \forall\ \mu \in \mathbb{P}_k(F), \tag{1.9c}$$

for every element $T \in \mathcal{T}_h$, and $F \in \partial T$. The L^2 projection into M_h will be denoted as P_M. If the stabilization function is chosen so that $\tau_T^{\max} := \max \tau|_{\partial T} > 0$, then by [16] there is a constant $C > 0$ independent of T and τ such that

$$\|\boldsymbol{\Pi}_{\mathrm{v}}\boldsymbol{q} - \boldsymbol{q}\|_T \leq Ch_T^{l_{\boldsymbol{q}}+1}|\boldsymbol{q}|_{\boldsymbol{H}^{l_{\boldsymbol{q}}+1}(T)} + Ch_T^{l_u+1}\tau_T^*|u|_{H^{l_u+1}(T)}, \tag{1.10a}$$

$$\|\Pi_{\mathrm{w}}u - u\|_T \leq Ch_T^{l_u+1}|u|_{H^{l_u+1}(T)} + C\frac{h_T^{l_{\boldsymbol{q}}+1}}{\tau_T^{\max}}|\nabla \cdot \boldsymbol{q}|_{H^{l_{\boldsymbol{q}}}(T)}. \tag{1.10b}$$

Here $\tau_T^* := \max \tau|_{\partial T \setminus F^*}$ and F^* is a face of T at which $\tau|_{\partial T}$ is maximum. As is customary, the symbol $|\cdot|_{H^s}$ is to be understood as the Sobolev semi norm of order $s \in \mathbb{R}$.

1.7 Auxiliary estimates

The following results were used throughout the text.

Lemma 1.1. *Consider $\boldsymbol{x} \in \Gamma_h$ and any smooth enough function $\boldsymbol{v}$ defined in $T^e \cup T^e_{ext}$, and define*

$$\delta_{\boldsymbol{v}}(\boldsymbol{x}) := \frac{1}{l(\boldsymbol{x})} \int_0^{l(\boldsymbol{x})} [\boldsymbol{v}(\boldsymbol{x} + \boldsymbol{n}s) - \boldsymbol{v}(\boldsymbol{x})] \cdot \boldsymbol{n}\, ds. \tag{1.11}$$

The following estimates hold for each $e \in \mathcal{E}_h^\partial$:

$$\|l^{1/2}\, \delta_{\boldsymbol{v}}\|_e \leq \frac{1}{\sqrt{3}}\, r_e^{3/2}\, C^e_{ext}\, C^e_{inv}\, \|\boldsymbol{v}\|_{T^e} \qquad \forall\ \boldsymbol{v} \in [\mathbb{P}_k(T)]^d, \tag{1.12a}$$

$$\|l^{1/2}\, \delta_{\boldsymbol{v}}\|_e \leq \frac{1}{\sqrt{3}}\, r_e\, \|h^\perp \partial_n \boldsymbol{v} \cdot \boldsymbol{n}\|_{T^e_{ext}} \qquad \forall\ \boldsymbol{v} \in [H^1(T)]^d, \tag{1.12b}$$

$$\|l^{1/2}\, \delta_{\boldsymbol{v}}\|_\infty \leq \frac{1}{\sqrt{3}}\, r_e \sup_{\boldsymbol{x} \in e} \|h_e^\perp\, \partial_n \boldsymbol{v} \cdot \boldsymbol{n}\|_{l(\boldsymbol{x})} \qquad \forall\ \boldsymbol{v} \in [H^1(T)]^d. \tag{1.12c}$$

Proof. See [18, Lemma 5.2]. ∎

The following lemma was needed to bound the terms in the decomposition of $\mathbb{T}^u$ carried out in Lemma 2.5.

Lemma 1.2. *[18, Lemma 5.5] Suppose Assumption (2.8d) and the elliptic regularity inequality (1.8) hold. Then,*

$$\|(h^\perp)^{-1/2}(Id_M - P_M)\psi\|_{\Gamma_h} \lesssim h\|\Theta\|_\Omega, \tag{1.13a}$$

$$\|l^{1/2}(Id_M - P_M)\partial_n\psi\|_{\Gamma_h} \lesssim R^{1/2}h\|\Theta\|_\Omega, \tag{1.13b}$$

$$\|l^{-3/2}(\psi + l\partial_n\psi)\|_{\Gamma_h} \lesssim \|\Theta\|_\Omega, \tag{1.13c}$$

$$\|l^{-1}\psi\|_{\Gamma_h} \lesssim \|\Theta\|_\Omega. \tag{1.13d}$$

The result below is used when deducing the bound for the term of the estimator involving the jump in the flux.

Lemma 1.3. *Let $e \in \mathcal{E}_h^\partial$ and $\boldsymbol{v} \in \boldsymbol{H}(\boldsymbol{div}; T^e)$. It holds*

$$\|E_{T^e}(\boldsymbol{v})\|^2_{T^e_{ext}} \lesssim r_e^2\, \|\boldsymbol{v}\|^2_{T^e} + r_e^2\, h_T^2 \|\nabla \cdot \boldsymbol{v}\|^2_{T^e}. \tag{1.14a}$$

Proof. We employ a scaling argument. Let $\Phi : T^e \to \widehat{T}$ be the affine mapping from T^e to the reference

element $\widehat{T}$ and set $\widehat{T^e_{ext}} := \Phi^{-1}(T^e_{ext})$. We have

$$\|E_{T^e}(v)\|^2_{T^e_{ext}} = 2|T^e_{ext}|\|\widehat{E}(\widehat{v})\|^2_{\widehat{T^e_{ext}}} \lesssim |T^e_{ext}|\|\widehat{v}\|^2_{H(\mathbf{div};\widehat{T})} = |T^e_{ext}|\left(\|\widehat{v}\|^2_{\widehat{T}} + \|\widehat{\nabla}\cdot\widehat{v}\|^2_{\widehat{T}}\right)$$
$$\lesssim |T^e_{ext}|\left(\frac{1}{|T^e|}\|v\|^2_{T^e} + \|\nabla\cdot v\|^2_{T^e}\right).$$

Thus, considering that $|T^e_{ext}| \lesssim (H^\perp_e)^2 = R^2_e\,(h^\perp_e)^2 \le r^2_e\,h^2_T$, and $|T^e| \lesssim h^2_T$, the inequality (1.14a) can be deduced. ■

The following result pertaining bubble functions is useful when addressing the local efficiency of the error estimator.

Lemma 1.4. *[82, Lemma 3.3.] Let $B_T := \Pi_{i=1}^{d+1}\lambda_i$ be the element–bubble function associated to $T \in \mathcal{T}_h$, where $\{\lambda_i\}_{i=1}^{d+1}$ are the barycentric coordinates of T, and $B_e := \Pi_{\substack{i=1\\ i\neq j}}^{d+1}\lambda_i$ be the face–bubble function associated to $e \subset \partial T$, where λ_j vanishes on e. Then, the following estimates hold*

$$\begin{array}{llllll} \|\boldsymbol{v}\|^2_T \lesssim & (\boldsymbol{v}, B_T\boldsymbol{v})_T, & \|B_T\boldsymbol{v}\|_T \lesssim & \|\boldsymbol{v}\|_T, & \|B_T\boldsymbol{v}\|_{1,T} \lesssim & h_T^{-1}\,\|\boldsymbol{v}\|_T, \\ \|\boldsymbol{\mu}\|^2_e \lesssim & (\boldsymbol{\mu}, B_e\boldsymbol{\mu})_e, & \|B_e\boldsymbol{\mu}\|_{\Delta_e} \lesssim & h_e^{1/2}\,\|\boldsymbol{\mu}\|_e, & \|B_e\boldsymbol{\mu}\|_{1,\Delta_e}, \lesssim & h_e^{-1/2}\,\|\boldsymbol{\mu}\|_e, \end{array} \tag{1.15}$$

for all $\boldsymbol{v} \in [\mathbb{P}_k(T)]^d,\ T \in \mathcal{T}_h$ and for each $\boldsymbol{\mu} \in [\mathbb{P}_k(e)]^d,\ e \in \mathcal{E}_h$.

1.8 Clément and Oswald interpolants

The following two interpolants are useful in the arguments leading to the reliability of the estimator. They allow to control the behavior of functions with piecewise H^1 regularity by representatives belonging to the global $H^1_0(\Omega)$ space.

First, in the next lemma, we state the approximation properties of the Clément interpolation operator $\mathcal{C}_h : L^2(\Omega_h) \to W^{1,c}_h \cap H^1_0(\Omega)$, introduced in [13] as

$$\mathcal{C}_h w := \sum_{z\in\mathcal{N}_h}\left(\frac{1}{|\Omega_z|}\int_{\Omega_z} w\,dx\right)\phi_z,$$

where ϕ_z is the $\mathbb{P}_1$ nodal basic functions associated to the interior vertex z, $\Omega_z = supp\ \phi_z$, and $W^{1,c}_h := \{w \in C(\Omega) : w|_T \in \mathbb{P}_1(T), T \in \mathcal{T}_h\}$.

Lemma 1.5. *[82, Lemma 3.2] For any $T \in \mathcal{T}_h,\ e \in \mathcal{E}^\circ_h$ and $0 \le m \le 1$, the following estimates hold, for all $w \in H^1_0(\Omega)$*

$$\|\mathcal{C}_h w\|_{m,\Omega} \lesssim \|w\|_{m,\Omega}, \quad \|w - \mathcal{C}_h w\|_{0,T} \lesssim h_T\|w\|_{1,\Delta_T}, \quad \|w - \mathcal{C}_h w\|_{0,e} \lesssim h_e^{1/2}\|w\|_{1,\Delta_e},$$

where $\Delta_T := \{T' \in \mathcal{T}_h : \overline{T} \cap \overline{T'} \neq \emptyset\}$ and $\Delta_e = \{T' \in \mathcal{T}_h : \overline{T'} \cap \overline{e} \neq \emptyset\}$.

We define now the space

$$W_h^* := \{w \in L^2(\mathcal{T}_h) : w|_T \in \mathbb{P}_{k+1}(T),\ \forall T \in \mathcal{T}_h\},$$

and shows that an element w of W_h^* can be approximated by a continuous function $\widetilde{w} \in W_h^*$, sometimes referred to as *Oswald interpolant*, and that the approximation error can be controlled by the size of the inter-element jumps of w.

Lemma 1.6. *[50, Theorem 2.2.] For any $w_h \in W_h^*$ and any multi-index with $|\alpha| = 0, 1$, the following approximation results holds: Let u_D be the restriction to Γ_h of a function in $W_h^* \cap H^1(\Omega_h)$. then there exists a function $\widetilde{w}_h \in W_h^* \cap H^1(\Omega_h)$ satisfying $\widetilde{w}_h|_\Gamma = u_D$, and*

$$\sum_{T\in\mathcal{T}_h} \|D^\alpha(w_h - \widetilde{w}_h)\|_T^2 \leq C_O \left(\sum_{e\in\mathcal{E}_h^\circ} h_e^{1-2|\alpha|} \|[\![w_h]\!]\|_e^2 + \sum_{e\in\mathcal{E}_h^\partial} h_e^{1-2|\alpha|} \|u_D - w_h\|_e^2 \right),$$

above, C_O is a positive constant independent of the mesh size.

CHAPTER 2

A priori and *a posteriori* error analysis of an unfitted HDG method for semi-linear elliptic problems

In this chapter we present *a priori* and *a posteriori* error analysis of a high order hybridizable discontinuous Galerkin (HDG) method applied to a semi-linear elliptic problem posed on a piecewise curved, non polygonal domain. We approximate Ω by a polygonal subdomain Ω_h and propose an HDG discretization, which is shown to be optimal under mild assumptions related to the non-linear source term and the distance between the boundaries of the polygonal subdomain Ω_h and the true domain Ω. Moreover, a local non-linear post-processing of the scalar unknown is proposed and shown to provide an additional order of convergence. A reliable and locally efficient *a posteriori* error estimator that takes into account the error in the approximation of the boundary data of Ω_h is also provided.

2.1 Introduction

In this chapter, we carry out *a priori* and *a posteriori* error analyses of a hybridizable discontinuous Galerkin (HDG) method [14] applied to semi-linear elliptic problems of the form

$$-\nabla \cdot (\kappa \nabla u) = \mathcal{F}(u) \qquad \text{in } \Omega, \tag{2.1a}$$

$$u = g \qquad \text{on } \Gamma := \partial\Omega, \tag{2.1b}$$

where the domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is not necessarily polygonal/polyhedral, κ is a positive function in Ω, $\mathcal{F}$ is a source term that depends on the solution u and g is the Dirichlet boundary data on Γ. To avoid the trivial solution, we will assume that if the boundary conditions are homogeneous, the source term will not vanish for $u = 0$.

In the present study, the source term $\mathcal{F}$ will be assumed to be Lipschitz-continuous in Ω, i.e, there exists $L_\Omega > 0$ such that

$$\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\|_\Omega \leq L_\Omega \|u_1 - u_2\|_\Omega \qquad \forall\, u_1, u_2 \in L^2(\Omega). \tag{2.2}$$

In addition, we assume that there exist positive constants $\underline{\kappa}$ and $\overline{\kappa}$ such that

$$\underline{\kappa} \leq \kappa(\boldsymbol{x}) \leq \overline{\kappa} \quad \forall\, \boldsymbol{x} \in \Omega.$$

An HDG discretization requires us to formulate the problem in mixed from through the introduction of the flux $\boldsymbol{q} := -\kappa \nabla u$ as an additional unknown. This choice makes it possible to write (2.1) as the equivalent first order system

$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega, \tag{2.3a}$$
$$\nabla \cdot \boldsymbol{q} = \mathcal{F}(u) \qquad \text{in } \Omega, \tag{2.3b}$$
$$u = g \qquad \text{on } \partial\Omega. \tag{2.3c}$$

HDG schemes, as many other discretization methods, are based on a triangulation of the domain. In our case, Ω has a piecewise curved boundary which complicates the use high order methods, since the boundary must be properly interpolated by "curved" triangles or tetrahedra in order to preserve high order convergence. An alternative is to approximate Ω by a polygonal/polyhedral subdomain $\Omega_h \subset \Omega$, that can be easily discretized by a uniform triangulation of size $h > 0$. Then, the system (2.3) can be restricted to Ω_h:

$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega_h, \tag{2.4a}$$
$$\nabla \cdot \boldsymbol{q} = \mathcal{F}(u) \qquad \text{in } \Omega_h, \tag{2.4b}$$
$$u = \varphi \qquad \text{on } \Gamma_h := \partial\Omega_h, \tag{2.4c}$$

where the unknown φ is the Dirichlet data on the computational boundary Γ_h. A clever way to determine φ was proposed for one dimension in [17] and then extended to higher dimensions by [21]. The method consists of using the definition of the flux to transfer the Dirichlet data from Γ to Γ_h along segments called *transferring paths*. In fact, given $\boldsymbol{x} \in \Gamma_h$ and $\overline{\boldsymbol{x}} \in \Gamma$, one can integrate (2.3a) along a segment of length $l(\boldsymbol{x})$ with unit tangent vector $\boldsymbol{t}(\boldsymbol{x})$ connecting them to obtain the following representation for φ:

$$\varphi(\boldsymbol{x}) = g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} (\kappa^{-1}\, \boldsymbol{q})(\boldsymbol{x} + \boldsymbol{t}(\boldsymbol{x})s) \cdot \boldsymbol{t}(\boldsymbol{x}) ds. \tag{2.5}$$

Above, we have considered that $u(\overline{\boldsymbol{x}}) = g(\overline{\boldsymbol{x}})$. At the end of Section 1.3 we will describe a way to pick $\overline{\boldsymbol{x}}$ in such a way that the transfer will preserve the order of approximation of the underlying discretization. Notice that the assumption (2.2) implies that $\mathcal{F}$ is also Lipschitz continuous in Ω_h with constant $L \leq L_\Omega$; this observation will be useful in the analysis to follow.

In previous works the authors had applied this transfer technique in combination with an iterative HDG discretization to deal with the nonlinear system (2.4) arising from the Grad-Shafranov equation [73] and explored an h-adaptive HDG scheme for the solution of the problem [74]. The adaptive strategy was powered by a residual-based error estimator first proposed by Cockburn and Zhang [23], albeit for polygonal domains—therefore not requiring the transfer of the boundary data— and linear problems. The goal of this work is to provide a rigorous justification for the numerical results obtained previously by the authors when applying these techniques for semi-linear problems in curved geometries. The present communication is mainly theoretical and we refer the reader interested on

numerical experiments to [73, 74] where plenty of experiments are provided within the context of plasma equilibrium. The results presented here, however, are not limited to plasma applications and remain valid for general semi-linear elliptic equations .

2.2 The HDG method

The HDG scheme associated to (2.3) reads: Find $(\boldsymbol{q}_h, u_h, \widehat{u}_h) \in \boldsymbol{V}_h \times W_h \times M_h$, such that

$$(\kappa^{-1}\boldsymbol{q}_h, \boldsymbol{v})_{\mathcal{T}_h} - (u_h, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} + \langle \widehat{u}_h, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{2.6a}$$

$$-(\boldsymbol{q}_h, \nabla w)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = (\mathcal{F}(u_h), w)_{\mathcal{T}_h}, \tag{2.6b}$$

$$\langle \widehat{u}_h, \mu\rangle_{\Gamma_h} = \langle \varphi_h, \mu\rangle_{\Gamma_h}, \tag{2.6c}$$

$$\langle \widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}, \mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} = 0, \tag{2.6d}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$. Here

$$\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n} := \boldsymbol{q}_h\cdot\boldsymbol{n} + \tau(u_h - \widehat{u}_h) \quad \text{on} \quad \partial\mathcal{T}_h, \tag{2.6e}$$

with τ being a positive stabilization function, whose maximum will be denoted by $\overline{\tau}$, and the approximate boundary condition, motivated by (2.5), is given by

$$\varphi_h(\boldsymbol{x}) := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} (\kappa^{-1}\,\boldsymbol{q}_h)(\boldsymbol{x} + \boldsymbol{t}(\boldsymbol{x})s)\cdot\boldsymbol{t}(\boldsymbol{x})ds, \quad \text{for} \quad \boldsymbol{x} \in \Gamma_h. \tag{2.6f}$$

Note that the function κ is defined by (2.3) in the full domain Ω, while the flux $\boldsymbol{q}_h$ is extended from Ω_h to Ω as defined in Section 1.2.

2.3 Well-posedness

In this section we employ a Banach fixed-point argument to ensure the well-posedness of (2.6). To that end, we define the operator $\mathcal{J} : W_h \to W_h$ that maps ζ to the second component of the triplet $(\boldsymbol{q}, u, \widehat{u}) \in \boldsymbol{V}_h \times W_h \times M_h$ satisfying the linearized HDG system (2.6) where the source has been evaluated at ζ, namely

$$(\kappa^{-1}\,\boldsymbol{q}, \boldsymbol{v})_{\mathcal{T}_h} - (u, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} + \langle \widehat{u}, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{2.7a}$$

$$-(\boldsymbol{q}, \nabla w)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = (\mathcal{F}(\zeta), w)_{\mathcal{T}_h}, \tag{2.7b}$$

$$\langle \widehat{u}, \mu\rangle_{\Gamma_h} = \langle \varphi_{\boldsymbol{q}}, \mu\rangle_{\Gamma_h}, \tag{2.7c}$$

$$\langle \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, \mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} = 0, \tag{2.7d}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$, where $\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} := \boldsymbol{q}\cdot\boldsymbol{n} + \tau(u - \widehat{u})$ and

$$\varphi_{\boldsymbol{q}}(\boldsymbol{x}) := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} (\kappa^{-1}\,\boldsymbol{q})(\boldsymbol{x} + \boldsymbol{t}(\boldsymbol{x})s)\cdot\boldsymbol{t}(\boldsymbol{x})ds, \tag{2.7e}$$

for $\boldsymbol{x} \in \Gamma_h$. We stress the difference between the non-linear mapping $\mathcal{J}$, which maps *arguments* of the source $\mathcal{F}(\cdot)$ to solutions of the corresponding HDG system, and the linear solution operator $\mathcal{S} : W_h \to W_h$ that maps the source term $\mathcal{F}$ itself to the solution of the corresponding HDG system.

In [18] it was shown that the constants C^e_{ext} and C^e_{inv} (defined in Section 1.2) are independent of h, but depend on the polynomial degree; in particular, the supremum appearing in the definition of C^e_{inv} is proportional to $(h^\perp_e)^{-1}$. With these definitions in place, we can now state the following set of geometric assumptions on the boundary faces of the triangulation.

Assumptions. For each $e \in \mathcal{E}^\partial_h$ we will require the following to hold:

$$\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{n} \text{ for all } \boldsymbol{x} \in e, \tag{2.8a}$$

$$r_e \leq C, \tag{2.8b}$$

$$\overline{\tau}\, H^\perp_e\, \underline{\kappa}^{-1} \leq \frac{1}{3}, \tag{2.8c}$$

$$\overline{\kappa}\, \underline{\kappa}^{-1}\, r_e^3\, (C^e_{ext}\, C^e_{inv})^2 \leq 1. \tag{2.8d}$$

Before proceeding, let us comment on this set of assumptions. As mentioned at the end of Section 1.2, the vector $\boldsymbol{t}(\boldsymbol{x})$ does not necessarily have to be normal to the face e. Therefore, the results presented in what follows still hold if (2.8a) is not satisfied as long as the difference $1 - \boldsymbol{t}(\boldsymbol{x}) \cdot \boldsymbol{n}$ is positive and small enough. However, this assumption helps us to facilitate the presentation of the ideas behind the proofs. On the other hand, (2.8b) imposes the geometric constraint that the family of triangulations $\{\mathcal{T}_h\}$ should be such that the distance between the computational boundary Γ_h and the true boundary Γ remains locally of the same order of magnitude as the face mesh parameter h_e. Moreover, it guarantees that as long as $H^\perp_e > 0$, then $H^\perp_e \sim h^\perp_e$; if $H^\perp_e = 0$ for some e, then no transfer of boundary conditions is needed on that particular face—as this would only happen if $e \cap \Omega = e$. Given that the stabilization parameter τ is of order one, (2.8c) states that the minimum value of the diffusion coefficient κ imposes a restriction on how far apart Γ_h and Γ are allowed to be. Due to the proportionality guaranteed by (2.8b), then (2.8c) will hold whenever the mesh size—and therefore the distance between the boundaries—is small enough. Assumption (2.8d) is the most demanding of all. By requiring r_e to be small enough compared to the product $\overline{\kappa}\,\underline{\kappa}^{-1}$, the condition effectively limits the range of values of κ that the method is able to resolve for a given, fixed, distance between the computational and physical boundaries, as measured by $H^\perp_e$. Not surprisingly, the closer to zero the diffusion coefficient gets, the smaller $H^\perp_e$ must be with respect to the mesh size near the computational boundary.

The main result of this section, Theorem 2.1, is that under suitable assumptions $\mathcal{J}$ is a contraction and therefore the solution of (2.6) can be obtained by applying it iteratively. The proof of this fact is almost a straightforward consequence of the linearity of the solution map and of a key stability bound established in Lemma 2.1 that estimates the norm of u in terms of those of the sources and the boundary conditions. However, the proof of the latter follows from a lengthy series of estimates. In order to prioritize clarity of exposition, we first present this estimate without proving it and show how the main result follows from it. The technical details of the proof of Lemma 2.1 are then presented afterwards.

Lemma 2.1. *Suppose that Assumptions* (2.8a) *throughout* (2.8d) *and the elliptic regularity of the*

auxiliary problem (1.7) *are satisfied. Then, there exists* $\widetilde{c} > 0$*, independent of* h *such that*

$$\|u\|_{\Omega_h} \leq 4 \max\{\widetilde{c}^2\, h, 1\} \|\mathcal{F}(\zeta)\|_{\Omega_h} + 2\,\widetilde{c}(\sqrt{3}+1)\, h^{1/2} \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}, \tag{2.9}$$

where $\overline{\mathrm{g}}(\boldsymbol{x}) = \mathrm{g}(\overline{\boldsymbol{x}}(\boldsymbol{x}))\ \forall\, \boldsymbol{x} \in \Gamma_h$*.*

Thanks to this estimate the main result, from which well-posedness of the problem follows, can be proved in a very compact way, as we now demonstrate.

Theorem 2.1. *If Assumptions* (2.8a) *throughout* (2.8d) *and the elliptic regularity of the auxiliary problem* (1.7) *hold, then* $\mathcal{J}$ *is well-defined. Furthermore, if we assume* $4\,L \max\{\widetilde{c}^2\, h, 1\} < 1$*, then* $\mathcal{J}$ *is a contraction operator.*

Proof. The system in (2.7) is linear and has a unique solution under the set of assumptions (2.8) (see [18]), therefore the operator $\mathcal{J}$ is well-defined.

Let $\zeta_1, \zeta_2 \in W_h$ and consider $u_1 = \mathcal{J}(\zeta_1)$ and $u_2 = \mathcal{J}(\zeta_2)$. Then, u_1 and u_2 are the second component of the solution of (2.7) with right hand sides $\mathcal{F}(\zeta_1)$ and $\mathcal{F}(\zeta_2)$, respectively. Hence, the difference $u_1 - u_2$ satisfies equations (2.7), with source term $\mathcal{F}(\zeta_1) - \mathcal{F}(\zeta_2)$ and homogeneous boundary conditions on Γ. By the stability estimate in Lemma 2.1 and Lipschitz continuity assumption, we have

$$\|\mathcal{J}(\zeta_1) - \mathcal{J}(\zeta_2)\|_{\Omega_h} = \|u_1 - u_2\|_{\Omega_h} \leq 4 \max\{\widetilde{c}^2\, h, 1\} \|\mathcal{F}(\zeta_1) - \mathcal{F}(\zeta_2)\|_{\Omega_h} \leq 4\, L\, \max\{\widetilde{c}^2\, h, 1\} \|\zeta_1 - \zeta_2\|_{\Omega_h}.$$

The result follows due to $4\,L \max\{\widetilde{c}^2\, h, 1\} < 1$. ■

As a consequence of the above result, system (2.6) subject to the hypotheses of the theorem has a unique solution that depends continuously on the problem data.

We now present the arguments that lead to the proof of Lemma 2.1. We start by establishing a connection between the norm of the transferred boundary conditions $\varphi_{\boldsymbol{q}}$, the magnitude of the flux $\boldsymbol{q}$ and the length of the transfer path taken. In order to do so, we will make use of a tool introduced in [18]. More Precisely, we use the auxiliary function $\delta_{\boldsymbol{v}}$ and its properties listed in the Lemma 1.1. The significance of this function is that it will allow us to separate the contributions to the boundary conditions coming from the flux, from the diffusivity, and from the length of the transfer path.

Observe that due to Assumption (2.8a) and (2.7e), we have

$$\begin{aligned} \varphi_{\boldsymbol{q}}(\boldsymbol{x}) - \mathrm{g}(\overline{\boldsymbol{x}}(\boldsymbol{x})) &= \int_0^{l(\boldsymbol{x})} \kappa^{-1}(\boldsymbol{x})\, \boldsymbol{q}(\boldsymbol{x} + \boldsymbol{n}s) \cdot \boldsymbol{n} ds \\ &= \kappa^{-1}(\boldsymbol{x}) \int_0^{l(\boldsymbol{x})} [\boldsymbol{q}(\boldsymbol{x} + \boldsymbol{n}s) - \boldsymbol{q}(\boldsymbol{x})] \cdot \boldsymbol{n} ds + l(\boldsymbol{x}) \kappa^{-1}(\boldsymbol{x}) \boldsymbol{q}(\boldsymbol{x}) \cdot \boldsymbol{n}, \end{aligned}$$

with $\boldsymbol{q} \in \boldsymbol{V}_h$, and using the definition of $\delta_{\boldsymbol{q}}$, given in (1.11), we can rewrite the above as

$$\varphi_{\boldsymbol{q}}(\boldsymbol{x}) - \mathrm{g}(\overline{\boldsymbol{x}}(\boldsymbol{x})) = \kappa^{-1}(\boldsymbol{x})\, l(\boldsymbol{x})\, \delta_{\boldsymbol{q}}(\boldsymbol{x}) + \kappa^{-1}(\boldsymbol{x})\, l(\boldsymbol{x})\, \boldsymbol{q}(\boldsymbol{x}) \cdot \boldsymbol{n}. \tag{2.10}$$

This expression, combined with the bounds that we will derive in Lemma 2.2 below, will enable us to estimate the approximate solution in terms of the sources, as will become evident in Lemma 2.3.

The following three inequalities follow readily from estimate (1.12a), assumptions (2.8d) and (2.8c), and Young's inequalities.

Lemma 2.2. *Let $\varphi_{\boldsymbol{q}}$ be the transferred boundary condition defined in (2.7e) and suppose that Assumptions S are satisfied. It holds*

$$
\begin{aligned}
|\langle \varphi_{\boldsymbol{q}}, \delta_{\boldsymbol{q}} \rangle_{\Gamma_h}| &\leq \frac{1}{6}\|\kappa^{1/2} l^{-1/2} \varphi_{\boldsymbol{q}}\|^2_{\Gamma_h} + \frac{1}{2}\|\kappa^{-1/2}\boldsymbol{q}\|^2_{\Omega_h} \\
|\langle \varphi_{\boldsymbol{q}}, \tau(u-\widehat{u}) \rangle_{\Gamma_h}| &\leq \frac{1}{6}\|\kappa^{1/2} l^{-1/2} \varphi_{\boldsymbol{q}}\|^2_{\Gamma_h} + \frac{1}{2}\|\tau^{1/2}(u-\widehat{u})\|^2_{\partial\mathcal{T}_h} \\
|\langle \varphi_{\boldsymbol{q}}, \kappa\, l^{-1}\, \overline{\mathrm{g}} \rangle_{\Gamma_h}| &\leq \frac{1}{6}\|\kappa^{1/2} l^{-1/2}\, \varphi_{\boldsymbol{q}}\|^2_{\Gamma_h} + \frac{3}{2}\|\kappa^{1/2} l^{-1/2}\, \overline{\mathrm{g}}\|^2_{\Gamma_h}
\end{aligned}
$$

. □

The expression for $\varphi_{\boldsymbol{q}}$ in (2.10) implies that $\boldsymbol{q}(\boldsymbol{x})\cdot\boldsymbol{n} = \kappa(\boldsymbol{x}) l^{-1}(\boldsymbol{x})(\varphi_{\boldsymbol{q}}(\boldsymbol{x}) - \overline{\mathrm{g}}) - \delta_{\boldsymbol{q}}(\boldsymbol{x})$. Thus, thanks to the definition of $\widehat{\boldsymbol{q}}\cdot\boldsymbol{n}$, it follows that

$$
\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} = \kappa\, l^{-1}(\varphi_{\boldsymbol{q}} - \overline{\mathrm{g}}) - \delta_{\boldsymbol{q}} + \tau(u-\widehat{u}) \qquad \text{on } \Gamma_h. \tag{2.11}
$$

The above expression can now be combined with the estimates from Lemma 2.2 to produce a bound for the norm of $(\boldsymbol{q}, u-\widehat{u}, \varphi_{\boldsymbol{q}})$ in terms of the source $\mathcal{F}(\zeta)$ and the boundary data $\overline{\mathrm{g}}$ as we will show next.

Lemma 2.3. *If Assumptions (2.8) hold, then*

$$
|\!|\!|(\boldsymbol{q}, u-\widehat{u}, \varphi_{\boldsymbol{q}})|\!|\!|^2 \leq 2\|\mathcal{F}(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + 3\|\kappa^{1/2} l^{-1/2}\overline{\mathrm{g}}\|^2_{\Gamma_h},
$$

where

$$
|\!|\!|(\boldsymbol{v}, w, \mu)|\!|\!| := \Big(\|\kappa^{-1/2}\boldsymbol{v}\|^2_{\Omega_h} + \|\tau^{1/2} w\|^2_{\partial\mathcal{T}_h} + \|\kappa^{1/2} l^{-1/2}\mu\|^2_{\Gamma_h}\Big)^{1/2}. \tag{2.12}
$$

Proof. Take $\zeta \in W_h$ and let $u = \mathcal{J}(\zeta) \in W_h$ be the corresponding solution satisfying (2.7). Integrating by parts the left hand side in (2.7b), testing (2.7) with $\boldsymbol{v} = \boldsymbol{q}$, $w = u$, and

$$
\mu := \begin{cases} -\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} & \text{on } \Gamma_h, \\ -\widehat{u} & \text{on } \partial\mathcal{T}_h \setminus \Gamma_h, \end{cases}
$$

and adding the resulting equalities, we get

$$
\|\kappa^{-1/2}\,\boldsymbol{q}\|^2_{\Omega_h} + \|\tau^{1/2}\,(u-\widehat{u})\|^2_{\partial\mathcal{T}_h} = (\mathcal{F}(\zeta), u)_{\mathcal{T}_h} - \langle \varphi_{\boldsymbol{q}}, \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}\rangle_{\Gamma_h}.
$$

Then, using the fact that $\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} = \boldsymbol{q}\cdot\boldsymbol{n} + \tau(u-\widehat{u})$ in combination with identity (2.11), we obtain

$$
|\!|\!|(\boldsymbol{q}, u-\widehat{u}, \varphi_{\boldsymbol{q}})|\!|\!|^2 \leq \|\mathcal{F}(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + |\langle \varphi_{\boldsymbol{q}}, \delta_{\boldsymbol{q}}\rangle_{\Gamma_h}| + |\langle \varphi_{\boldsymbol{q}}, \tau(u-\widehat{u})\rangle_{\Gamma_h}| + |\langle \varphi_{\boldsymbol{q}}, \kappa\, l^{-1}\,\overline{\mathrm{g}})\rangle_{\Gamma_h}|.
$$

By the Cauchy-Schwarz inequality and Lemma 2.2, we arrives at

$$\begin{aligned}|\!|\!|(\boldsymbol{q}, u-\widehat{u}, \varphi_{\boldsymbol{q}})|\!|\!|^2 \leq & \|\mathcal{F}(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + \frac{1}{2}\|\kappa^{1/2} l^{-1/2}\varphi_{\boldsymbol{q}}\|^2_{\Gamma_h} + \frac{1}{2}\|\kappa^{-1/2}\boldsymbol{q}\|^2_{\Omega_h} \\ & + \frac{1}{2}\|\tau^{1/2}(u-\widehat{u})\|^2_{\partial\mathcal{T}_h} + \frac{3}{2}\|\kappa^{1/2} l^{-1/2}\,\overline{\mathrm{g}}\|^2_{\Gamma_h},\end{aligned}$$

and the result follows. ■

Proof of Lemma 2.1. With all the previous technical results in place we are now in a position to prove the crucial result. In the arguments below, $\boldsymbol{\Pi_V}$ and Π_W are, respectively, the $\boldsymbol{V}$ and W components of the HDG projector introduced in the Section 1.6.

Proof. Given $\Theta \in L^2(\Omega)$, the solution to the auxiliary problem (1.5) and following the argument of [16, Lemma 4.1] it is possible to show that if u satisfies (2.7) then

$$(u,\Theta)_{\mathcal{T}_h} = (\kappa^{-1}\,\boldsymbol{q}, \boldsymbol{\Pi_V}\boldsymbol{\phi} - \boldsymbol{\phi})_{\mathcal{T}_h} + \langle \widehat{u}, \boldsymbol{\phi}\cdot\boldsymbol{n}\rangle_{\Gamma_h} - \langle \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h} + (\mathcal{F}(\zeta), \Pi_W\psi)_{\mathcal{T}_h}.$$

We will now use this expression to bound the norm of u. In order to simplify the exposition, we will group some terms on the right hand side of this expression and treat them separately. Hence, we decompose the above expression as

$$(u,\Theta)_{\mathcal{T}_h} = \mathbb{T}_{\boldsymbol{q}} + \mathbb{T}_{\mathcal{F}} + \mathbb{T}_u, \tag{2.13}$$

by defining

$$\mathbb{T}_{\boldsymbol{q}} := (\kappa^{-1}\,\boldsymbol{q}, \boldsymbol{\Pi_V}\boldsymbol{\phi} - \boldsymbol{\phi})_{\mathcal{T}_h}, \quad \mathbb{T}_u := \langle \widehat{u}, \boldsymbol{\phi}\cdot\boldsymbol{n}\rangle_{\Gamma_h} - \langle \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}, \quad \text{and} \quad \mathbb{T}_{\mathcal{F}} := (\mathcal{F}(\zeta), \Pi_W\psi)_{\mathcal{T}_h}.$$

The terms $\mathbb{T}_{\boldsymbol{q}}$ and $\mathbb{T}_{\mathcal{F}}$ can be bounded by an application of the estimates (1.10a) in combination with the elliptic regularity (1.8), yielding

$$|\mathbb{T}_{\boldsymbol{q}}| \leq \underline{\kappa}^{-1/2}\,\|\kappa^{-1/2}\,\boldsymbol{q}\|_{\Omega_h}\|\boldsymbol{\Pi_V}\boldsymbol{\phi} - \boldsymbol{\phi}\|_{\Omega_h} \lesssim h\,\|\kappa^{-1/2}\,\boldsymbol{q}\|_{\Omega_h}\|\Theta\|_{\Omega}, \tag{2.14}$$

and

$$|\mathbb{T}_{\mathcal{F}}| \lesssim \|\mathcal{F}(\zeta)\|_{\Omega_h}\,\|\Theta\|_{\Omega}. \tag{2.15}$$

The treatment of the term $\mathbb{T}_u$ requires more work. Denoting by Id_M the identity operator in M, considering (2.11) and equation (1.7a), $\mathbb{T}_u$ can be written as $\mathbb{T}_u = \sum_{i=1}^5 \mathbb{T}_u^i$, where

$$\begin{aligned}&\mathbb{T}_u^1 := -\langle \kappa l^{-1}\varphi_{\boldsymbol{q}}, \psi + l\partial_n\psi\rangle_{\Gamma_h}, && \mathbb{T}_u^2 := \langle \kappa\varphi_{\boldsymbol{q}}, (P_M - Id_M)\partial_n\psi\rangle_{\Gamma_h},\\ &\mathbb{T}_u^3 := \langle \delta_{\boldsymbol{q}}, \psi\rangle_{\Gamma_h}, && \mathbb{T}_u^4 := -\langle \tau(u-\widehat{u}), P_M\psi\rangle_{\Gamma_h},\\ &\mathbb{T}_u^5 := \langle \kappa\, l^{-1}\,\overline{\mathrm{g}}, \psi\rangle_{\Gamma_h}.\end{aligned}$$

We will now determine bounds for all the terms in the decomposition.

By Young's inequality and combining the fact that $l(\boldsymbol{x}) \lesssim R_h h$, $\forall\, \boldsymbol{x} \in \Gamma_h$ with the estimate (1.13c),

we have

$$|\mathbb{T}_u^1| \leq \left| \langle \kappa^{1/2}\, l\, \kappa^{1/2}\, l^{-1/2}\, \varphi_{\boldsymbol{q}}, l^{-3/2}\, (\psi + l\partial_n \psi) \rangle_{\Gamma_h} \right| \lesssim R_h\, h\, \|\kappa^{1/2}\, l^{-1/2}\, \varphi_{\boldsymbol{q}}\|_{\Gamma_h} \|\Theta\|_{\Omega}.$$

Analogously, we get

$$|\mathbb{T}_u^2| \lesssim R_h^{1/2}\, h\, \|\kappa^{1/2} l^{-1/2} \varphi_{\boldsymbol{q}}\|_{\Gamma_h} \|\Theta\|_{\Omega}.$$

To bound $\mathbb{T}_u^3$, we employ (1.12a), (1.13d), and (2.8d) yielding

$$\begin{aligned} |\mathbb{T}_u^3| &\lesssim R_h\, h)^{1/2} \|l^{1/2}\, \delta_{\boldsymbol{q}}\|_{\Gamma_h} \|l^{-1}\, \psi\|_{\Gamma_h} \\ &\lesssim (R_h\, h)^{1/2} \left(\sum_{e \in \mathcal{E}_h^{\partial}} \frac{1}{3} r_e^3\, (C_{ext}^e\, C_{inv}^e)^2 \|\boldsymbol{q}\|_{T^e}^2 \right)^{1/2} \|\Theta\|_{\Omega} \\ &\lesssim R_h^2\, h^{1/2}\, \|\kappa^{-1/2} \boldsymbol{q}\|_{\Omega_h} \|\Theta\|_{\Omega}. \end{aligned}$$

Similarly, using (1.13d) we can bound

$$|\mathbb{T}_u^4| \lesssim \overline{\tau}^{1/2}\, R_h\, h\, \|\tau^{1/2}\, (u - \widehat{u})\|_{\partial \mathcal{T}_h} \|\Theta\|_{\Omega} \quad \text{and} \quad |\mathbb{T}_u^5| \lesssim (R_h h)^{1/2} \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h} \|\Theta\|_{\Omega}.$$

Taking $\Theta = u$ in Ω_h and $\Theta = 0$ in Ω_h^c in (1.7) and considering the bounds for the terms $\mathbb{T}_u^i$, we can combine the decomposition (2.13) with the estimates (2.14) and (2.15), to obtain

$$\begin{aligned} \|u\|_{\Omega_h} \lesssim\, & h\, \|\kappa^{-1/2} \boldsymbol{q}\|_{\Omega_h} + h\, (R_h + R_h^{1/2})\, \|\kappa^{1/2}\, l^{-1/2} \varphi_{\boldsymbol{q}}\|_{\Gamma_h} + R_h^2\, h^{1/2} \|\kappa^{-1/2} \boldsymbol{q}\|_{\Omega_h} \\ & + \overline{\tau}^{1/2}\, R_h\, h\, \|\tau^{1/2} (u - \widehat{u})\|_{\partial \mathcal{T}_h} + \|\mathcal{F}(\zeta)\|_{\Omega_h} + (R_h h)^{1/2} \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}. \end{aligned}$$

Now, let

$$\widetilde{c} := C \left\{ 1 + R_h^2 + R_h + R_h^{1/2} + \overline{\tau}^{1/2}\, R_h \right\}, \tag{2.16}$$

where $C > 0$ is the constant hidden in the symbol $\lesssim$. Then, since $h < 1$ by Lemma 2.3 and Young's inequality, we infer

$$\begin{aligned} \|u\|_{\Omega_h} &\leq \widetilde{c} h^{1/2} \left(\sqrt{2}\, \|\mathcal{F}(\zeta)\|_{\Omega_h}^{1/2} \|u\|_{\Omega_h}^{1/2} + \sqrt{3}\, \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h} \right) + \|\mathcal{F}(\zeta)\|_{\Omega_h} + \widetilde{c} h^{1/2}\, \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h} \\ &\leq \widetilde{c}^2\, h\, \|\mathcal{F}(\zeta)\|_{\Omega_h} + \frac{1}{2} \|u\|_{\Omega_h} + \|\mathcal{F}(\zeta)\|_{\Omega_h} + \widetilde{c}\, (\sqrt{3} + 1)\, h^{1/2}\, \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}, \end{aligned}$$

and thus

$$\|u\|_{\Omega_h} \leq 4\, \max\{\widetilde{c}^2\, h, 1\} \|\mathcal{F}(\zeta)\|_{\Omega_h} + 2\, \widetilde{c}\, (\sqrt{3} + 1) h^{1/2} \|\kappa^{1/2}\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}.$$

■

2.4 *A priori* error analysis

We now provide the *a priori* error bounds for the method. As we will see, some of the results presented in this section can be proven by using similar arguments to those of Section 2.3 and many details will be omitted. Given that the set of assumptions (2.8) is required to hold in order to ensure the well-posedness of the problem, in the present section they will be assumed as true and used for

the error analysis without explicitly stating them in the results. Similarly, the regularity assumption (1.8) will be assumed to hold.

The total approximation error has a component due to the accuracy of the discretization, and a component due entirely to the approximation properties of the discrete subspace. This is made apparent using the HDG projection defined in (1.9) and defining *the projections of the errors*

$$\boldsymbol{\varepsilon}^{\boldsymbol{q}} := \boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q} - \boldsymbol{q}_h \quad \text{and} \quad \varepsilon^u := \Pi_W u - u_h,$$

and the *error of the projections*

$$\boldsymbol{I}^{\boldsymbol{q}} := \boldsymbol{q} - \boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q} \quad \text{and} \quad I^u := u - \Pi_W u.$$

Using these quantities we can decompose the error as follows

$$\boldsymbol{q} - \boldsymbol{q}_h = \boldsymbol{\varepsilon}^{\boldsymbol{q}} + \boldsymbol{I}^{\boldsymbol{q}} \quad \text{and} \quad u - u_h = \varepsilon^u + I^u.$$

In addition, we define $\varepsilon^{\widehat{u}} := P_M u - \widehat{u}_h$, where we recall that P_M is the L^2 projection into M_h.

It is not difficult to show that $(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u, \varepsilon^{\widehat{u}})$ belongs to $\boldsymbol{V}_h \times W_h \times M_h$ and satisfies

$$(\kappa^{-1}\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \boldsymbol{v})_{\mathcal{T}_h} - (\varepsilon^u, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} + \langle \varepsilon^{\widehat{u}}, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = -(\kappa^{-1}\boldsymbol{I}^{\boldsymbol{q}}, \boldsymbol{v})_{\mathcal{T}_h}, \tag{2.17a}$$

$$-(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \nabla w)_{\mathcal{T}_h} + \langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = (\mathcal{F}(u) - \mathcal{F}(u_h), w)_{\mathcal{T}_h}, \tag{2.17b}$$

$$\langle \varepsilon^{\widehat{u}}, \mu\rangle_{\Gamma_h} = \langle \varphi - \varphi_h, \mu\rangle_{\Gamma_h}, \tag{2.17c}$$

$$\langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, \mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} = 0, \tag{2.17d}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$, where $\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n} := \boldsymbol{\varepsilon}^{\boldsymbol{q}}\cdot\boldsymbol{n} + \tau(\varepsilon^u - \varepsilon^{\widehat{u}}) = P_M(\boldsymbol{q}\cdot\boldsymbol{n}) - \widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}$. This *error equations* will help us establishing two results that will eventually lead to the proof of the convergence of the method.

To abbreviate the notation in the following arguments it will be useful to define

$$\Lambda_{\boldsymbol{q}} := \left(\|\boldsymbol{I}^{\boldsymbol{q}}\|^2_{\Omega_h} + \|h^{\perp}\,\partial_n(\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Omega_h^c} + \|(h^{\perp})^{1/2}\,\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Gamma_h} \right)^{1/2}, \tag{2.18a}$$

$$\Lambda_u := \left(\|(h^{\perp})^{1/2}\, I^u\|_{\Gamma_h} + \|I^u\|_{\Omega_h} \right)^{1/2}. \tag{2.18b}$$

With respect to these quantities we point out that, if $\boldsymbol{q} \in \boldsymbol{H}^{k+1}(\Omega)$, $u \in H^{k+1}(\Omega)$ and $\tau = \mathcal{O}(1)$ then, by scaling arguments and the properties (1.10), both $\Lambda_{\boldsymbol{q}}$ and Λ_u are of order h^{k+1}.

The first of these auxiliary lemmas establishes the convergence of the discrete flux $\boldsymbol{q}_h$, the restriction to the mesh skeleton $\widehat{u}_h$, and the transferred boundary data φ_h as a consequence of the convergence of the primary scalar variable u_h and the errors of the projections I^u and $\boldsymbol{I}^{\boldsymbol{q}}$.

Lemma 2.4. *Let $|||\cdot|||$ be the norm defined in (2.12). There exists a positive constant $C > 0$, independent of h, such that*

$$|||(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|||^2 \leq 4\,L(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h} + C\,\Lambda^2_{\boldsymbol{q}}. \tag{2.19}$$

Proof. Testing (2.17) with

$$\boldsymbol{v} := \boldsymbol{\varepsilon^q}, \quad w := \varepsilon^u \quad \text{and} \quad \mu := \begin{cases} -\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} & \text{on } \Gamma_h \\ -\varepsilon^u & \text{on } \partial\mathcal{T}_h \setminus \Gamma_h \end{cases},$$

results in

$$\|\kappa^{-1/2}\,\boldsymbol{\varepsilon^q}\|^2_{\Omega_h} + \|\tau^{1/2}\,(\varepsilon^u - \varepsilon^{\widehat{u}})\|_{\partial\mathcal{T}_h} = -(\kappa^{-1}\,\boldsymbol{I^q}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h} + (\mathcal{F}(u) - \mathcal{F}(u_h), \varepsilon^u)_{\mathcal{T}_h} - \langle \varphi - \varphi_h, \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n}\rangle_{\Gamma_h},$$

then, owing to (2.11), we readily obtain $\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} = \kappa\, l^{-1}\,(\varphi - \varphi_h) - \delta_{\boldsymbol{\varepsilon^q}} - \delta_{\boldsymbol{I^q}} - \boldsymbol{I^q} \cdot \boldsymbol{n} + \tau(\varepsilon^u - \varepsilon^{\widehat{u}})$ on Γ_h. Substituting this above, we get

$$\begin{aligned} |\!|\!|(\boldsymbol{\varepsilon^q}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|\!|\!|^2 \leq\ & |(\kappa^{-1}\,\boldsymbol{I^q}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h}| + L(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h} \\ & + |\langle \varphi - \varphi_h, \delta_{\boldsymbol{\varepsilon^q}} + \delta_{\boldsymbol{I^q}} + \boldsymbol{I^q} \cdot \boldsymbol{n} - \tau(\varepsilon^u - \varepsilon^{\widehat{u}})\rangle_{\Gamma_h}|. \end{aligned} \tag{2.20}$$

The estimates in Lemma 2.2, can be applied to the last term of (2.20) to arrive at

$$\begin{aligned} \langle \varphi - \varphi_h, \delta_{\boldsymbol{I^q}}\rangle_{\Gamma_h} &\leq \tfrac{1}{6}\,\|\kappa^{1/2}\,l^{-1/2}\,(\varphi - \varphi_h)\|^2_{\Gamma_h} + \tfrac{1}{2}\,\underline{\kappa}^{-1}\,\max_{e\in\mathcal{E}_h^\partial}\{r_e^2\}\|h^\perp \partial_n(\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Omega_h^c}, \\ \langle \varphi - \varphi_h, \delta_{\boldsymbol{\varepsilon^q}}\rangle_{\Gamma_h} &\leq \tfrac{1}{4}\|\kappa^{1/2}\,l^{-1/2}\,(\varphi - \varphi_h)\|^2_{\Gamma_h} + \tfrac{1}{3}\,\|\kappa^{-1/2}\,\boldsymbol{\varepsilon^q}\|^2_{\Omega_h}, \\ \langle \varphi - \varphi_h, \boldsymbol{I^q}\cdot\boldsymbol{n}\rangle_{\Gamma_h} &\leq \tfrac{1}{6}\,\|\kappa^{1/2}\,l^{-1/2}\,(\varphi - \varphi_h)\|^2_{\Gamma_h} + \tfrac{3}{2}\,\underline{\kappa}^{-1}\,\max_{e\in\mathcal{E}_h^\partial}\{r_e\}\,\|(h^\perp)^{1/2}\,\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Gamma_h}, \\ |\langle \varphi - \varphi_h, \tau(\varepsilon^u - \varepsilon^{\widehat{u}})\rangle_{\Gamma_h}| &\leq \tfrac{1}{6}\,\|\kappa^{1/2}\,l^{-1/2}\,(\varphi - \varphi_h)\|^2_{\Gamma_h} + \tfrac{1}{2}\,\|\tau^{1/2}\,(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial\mathcal{T}_h}. \end{aligned} \tag{2.21}$$

The estimate (2.19) is obtained with $C := 4\,\kappa^{-1}\,\max\{1, \frac{1}{2}\,R_h^2, \frac{3}{2}\,R_h\}$, applying Young's inequality to term $|(\kappa^{-1}\,\boldsymbol{I^q}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h}|$ and the estimates given in (2.21) . ■

Due to the previous result, it is enough to show the convergence of ε^u to guarantee the convergence of the method. The next step then is to estimate $\|\varepsilon^u\|_{\Omega_h}$, which we will do through a duality argument very much in the spirit of the proof of Lemma 2.1. Indeed, given $\Theta \in L^2(\Omega)$, and considering the linear auxiliary problem (1.7), but now using equations (2.17) instead of (2.7), we can decompose

$$(\varepsilon^u, \Theta)_{\mathcal{T}_h} = \mathbb{T}^{\mathcal{F}} + \mathbb{T}^{\boldsymbol{q}} + \mathbb{T}^u, \tag{2.22}$$

where

$$\begin{aligned} \mathbb{T}^{\mathcal{F}} &:= (\mathcal{F}(u) - \mathcal{F}(u_h), \Pi_W\psi)_{\mathcal{T}_h}, \\ \mathbb{T}^{\boldsymbol{q}} &:= (\kappa^{-1}(\boldsymbol{q} - \boldsymbol{q}_h), \boldsymbol{\Pi_V}\boldsymbol{\phi})_{\mathcal{T}_h} + (\boldsymbol{\varepsilon^q}, \nabla\psi)_{\mathcal{T}_h}, \\ \mathbb{T}^u &:= \langle \varepsilon^{\widehat{u}}, \boldsymbol{\phi}\cdot\boldsymbol{n}\rangle_{\Gamma_h} - \langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}. \end{aligned}$$

In order to estimate the size of ε^u we will now treat each of these terms separately. The term $\mathbb{T}^{\mathcal{F}}$ is easy to bound, since

$$|\mathbb{T}^{\mathcal{F}}| \leq L\|u - u_h\|_{\Omega_h}\|\Pi_W\psi\|_{\Omega_h} \leq L\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\Pi_W\psi\|_{\Omega_h} \lesssim L\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\Theta\|_{\Omega}. \tag{2.23}$$

Now, by adding and subtracting $(\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\phi})_{\mathcal{T}_h}$ in the definition of the term $\mathbb{T}^{\boldsymbol{q}}$, we obtain

$$\mathbb{T}^{\boldsymbol{q}} = (\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{\phi}-\boldsymbol{\phi})_{\mathcal{T}_h} + (\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\phi})_{\mathcal{T}_h} + (\boldsymbol{\varepsilon}^{\boldsymbol{q}},\nabla\psi)_{\mathcal{T}_h}.$$

However, due to (1.7a), it holds that $(\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\phi})_{\mathcal{T}_h} + (\boldsymbol{\varepsilon}^{\boldsymbol{q}},\nabla\psi)_{\mathcal{T}_h} = -(\boldsymbol{I}^{\boldsymbol{q}},\nabla\psi)_{\mathcal{T}_h}$. Let $\psi_h \in W_h$ be arbitrary. Then, by (1.9b), we have $(\boldsymbol{I}^{\boldsymbol{q}},\nabla\psi_h)=0$. Combining these last two facts we obtain

$$\mathbb{T}^{\boldsymbol{q}} = (\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{\phi}-\boldsymbol{\phi})_{\mathcal{T}_h} + (\boldsymbol{I}^{\boldsymbol{q}},\nabla(\psi-\psi_h))_{\mathcal{T}_h}.$$

Therefore, by choosing $\psi_h = \Pi_W\psi$, it follows that

$$\begin{aligned}|\mathbb{T}^{\boldsymbol{q}}| &\le \|\kappa^{-1/2}(\boldsymbol{\varepsilon}^{\boldsymbol{q}}+\boldsymbol{I}^{\boldsymbol{q}})\|_{\Omega_h}\|\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{\phi}-\boldsymbol{\phi}\|_{\Omega_h} + \|\boldsymbol{I}^{\boldsymbol{q}}\|_{\Omega_h}\|\nabla(\psi-\psi_h)\|_{\Omega_h}\\ &\lesssim h^{\min\{1,k\}}\|\kappa^{-1/2}\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|_{\Omega_h}\|\Theta\|_{\Omega} + h^{\min\{1,k\}}\|\boldsymbol{I}^{\boldsymbol{q}}\|_{\Omega_h}\|\Theta\|_{\Omega}\end{aligned} \tag{2.24}$$

where we have used the elliptic regularity for the auxiliary problem, the approximation properties (1.9) and (1.10) of the HDG projector. Finally, we can further decompose $\mathbb{T}^u := \sum_{i=1}^7 \mathbb{T}^u_i$, where:

$$\begin{aligned}
&\mathbb{T}^u_1 := -\langle\kappa l^{-1}(\varphi-\varphi_h),\psi+l\partial_n\psi\rangle_{\Gamma_h}, && \mathbb{T}^u_2 := -\langle\kappa(\varphi-\varphi_h),(Id_M-P_M)\partial_n\psi\rangle_{\Gamma_h},\\
&\mathbb{T}^u_3 := \langle\delta_{\boldsymbol{I}^q},\psi\rangle_{\Gamma_h}, && \mathbb{T}^u_4 := \langle\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n},(Id_M-P_M)\psi\rangle_{\Gamma_h},\\
&\mathbb{T}^u_5 := -\langle\tau P_M I^u,\psi\rangle_{\Gamma_h}, && \mathbb{T}^u_6 := \langle\delta_{\boldsymbol{\varepsilon}^{\boldsymbol{q}}},\psi\rangle_{\Gamma_h},\\
&\mathbb{T}^u_7 := -\langle\tau(\varepsilon^u-\varepsilon^{\widehat{u}}),P_M\psi\rangle_{\Gamma_h}.
\end{aligned}$$

Bounding separately each of the terms above it is possible to estimate $\|\varepsilon^u\|_{\mathcal{T}_h}$, as we show below.

Lemma 2.5. *Assume that the Lipschitz constant is such that L is small enough, and consider the discrete spaces to be of polynomial degree $k\ge 1$. Then,*

$$\|\varepsilon^u\|_{\Omega_h} \lesssim ((R_h h)^{1/2}(1+\overline{\tau}^{1/2})+h)|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!| + (R_h h^{1/2}+L)(\Lambda_{\boldsymbol{q}}+\Lambda_u). \tag{2.25}$$

Proof. By applying Young's inequality to each term in the decomposition of $\mathbb{T}^u$, considering the estimates in Lemma 1.2, using the fact $l(\boldsymbol{x}) \lesssim R_h h$,$\forall \boldsymbol{x}\in\Gamma_h$ and having in mind the estimates in (1.12), it is possible to deduce:

$$\begin{aligned}
&|\mathbb{T}^u_1| \lesssim \overline{\kappa}^{1/2}\,R_h\,h\,\|\kappa^{1/2}\,l^{-1/2}\,(\varphi-\varphi_h)\|_{\Gamma_h}\|\Theta\|_{\Omega}, && |\mathbb{T}^u_5| \lesssim \overline{\tau}\,R_h\,h^{1/2}\|(h^{\perp})^{1/2}I^u\|_{\Gamma_h}\|\Theta\|_{\Omega},\\
&|\mathbb{T}^u_2| \lesssim \overline{\kappa}^{1/2}\,R_h^{1/2}\,h\,\|\kappa^{1/2}\,l^{-1/2}\,(\varphi-\varphi_h)\|_{\Gamma_h}\|\Theta\|_{\Omega}, && |\mathbb{T}^u_6| \lesssim \overline{\kappa}^{1/2}\,R_h^2\,h^{1/2}\,\|\kappa^{-1/2}\,\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|_{\Omega_h}\|\Theta\|_{\Omega},\\
&|\mathbb{T}^u_3| \lesssim R_h^{3/2}\,h^{1/2}\|h^{\perp}\partial_n\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n}\|_{\Gamma_h}\|\Theta\|_{\Omega}, && |\mathbb{T}^7_u| \lesssim \overline{\tau}^{1/2}R_h h\|\tau^{1/2}(\varepsilon^u-\varepsilon^{\widehat{u}})\|_{\Gamma_h}\|\Theta\|_{\Omega}.\\
&|\mathbb{T}^u_4| \lesssim h\|(h^{\perp})^{1/2}\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n}\|_{\Gamma_h}\|\Theta\|_{\Omega},
\end{aligned}$$

Then, recalling the definition of the norm $|\!|\!|\cdot|\!|\!|$ in (2.12), and of the terms $\Lambda_{\boldsymbol{q}}$ and Λ_u in (2.18), we get

$$\begin{aligned}|\mathbb{T}^u| \;\lesssim\; &\left(\overline{\kappa}^{1/2}\,R_h h + \overline{\kappa}^{1/2}\,R_h^{1/2}\,h + \overline{\kappa}^{1/2}R_h^2\,h^{1/2} + \overline{\tau}^{1/2}R_h h\right)|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|\;\|\Theta\|_{\Omega}\\ &+\max\{R_h^{1/2},\overline{\tau}\}\,R_h h^{1/2}\,(\Lambda_u+\Lambda_{\boldsymbol{q}})\,\|\Theta\|_{\Omega} + h\Lambda_{\boldsymbol{q}}\|\Theta\|_{\Omega}.\end{aligned} \tag{2.26}$$

Finally, taking $\Theta=\varepsilon^u$ in Ω_h and $\Theta=0$ in Ω_h^c in (2.22) and using the estimates (2.23), (2.24) and (2.26), and considering assumption (2.8c), the estimate (2.25) is obtained. ∎

Combining Lemmas 2.4 and 2.5, we can bound the error in terms of the error of the projection I^u and $\boldsymbol{I}^{\boldsymbol{q}}$ as we do below.

Theorem 2.2. *Assume that* $6\,L\left((R_h\,h)^{1/2}\,(1+\overline{\tau}^{1/2})+h\right)<1$, τ *is of order one, and the discrete spaces are of polynomial degree* $k\geq 1$, *then*

$$|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|\lesssim \Lambda_{\boldsymbol{q}}+\Lambda_u, \tag{2.27a}$$

and

$$\|\varepsilon^u\|_{\Omega_h}\lesssim\left((R_h h)^{1/2}+L+h\right)(\Lambda_u+\Lambda_{\boldsymbol{q}}). \tag{2.27b}$$

Proof. It follows from Lemma 2.4 and the estimate in (2.25), that

$$\begin{aligned}&|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|^2\leq 6\,L\,\|\varepsilon^u\|^2_{\Omega_h}+2\,L\,\Lambda_u^2+C\,\Lambda_{\boldsymbol{q}}^2\\&\leq 6\,L\left((R_h\,h)^{1/2}\,(1+\overline{\tau}^{1/2})+h\right)|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|^2+\max\{6\,L(R_h\,h^{1/2}+L,C,2\,L)\}\,(\Lambda_u^2+\Lambda_{\boldsymbol{q}}^2),\end{aligned}$$

where C is the constant defined in Lemma 2.4. Then, due to $6\,L\left((R_h\,h)^{1/2}\,(1+\overline{\tau}^{1/2})+h\right)<1$, the estimate (2.27a) is fulfilled. Finally, (2.25) and (2.27a) imply (2.27b). ■

As a byproduct of the previous result, we are now in the position to establish the asymptotic convergence rate of the discretization.

Corollary 2.1. *Suppose that assumptions of Theorem 2.2 hold. If* $u\in H^{k+1}(\Omega)$ *and* $\boldsymbol{q}\in\boldsymbol{H}^{k+1}(\Omega)$, *then*

$$\|\boldsymbol{q}-\boldsymbol{q}_h\|_\Omega+\|u-u_h\|_\Omega\lesssim h^{k+1}\left(|u|_{k+1,\Omega}+|\boldsymbol{q}|_{k+1,\Omega}\right). \tag{2.28}$$

Proof. It follows from Theorem 2.2, Lemma 2.25, and the approximation properties (1.10), combined with Lemmas 3.7 and 3.8 in [18]. ■

2.5 *A posteriori* error analysis

2.5.1 Local post processing of the scalar solution

Our *a posteriori* error estimator will be obtained in terms of a local post processing u_h^*, which approximates the scalar unknown u with enhanced accuracy. We seek for u_h^* in the space

$$W_h^*:=\{w\in L^2(\mathcal{T}_h):w|_T\in\mathbb{P}_{k+1}(T),\ \forall\,T\in\mathcal{T}_h\},$$

such that, in each element $T\in\mathcal{T}_h$, satisfies:

$$(\kappa\nabla u_h^*,\nabla w)_T+(\mathcal{F}(u_h^*),w)_T=-(\boldsymbol{q}_h,\nabla w)_T+(\mathcal{F}(u_h),w)_T\qquad\forall\,w\in\mathbb{P}_{k+1}(T), \tag{2.29a}$$

$$(u_h^*,w)_T=(u_h,w)_T\qquad\forall\,w\in\mathbb{P}_0(T). \tag{2.29b}$$

In the case where $\mathcal{F}$ is independent of u, it is well known (Section 5.2 in [16]) that u_h^* is well defined and converges to u with order h^{k+2} when the solution has enough regularity. It is also known that there is a variety of choices to construct u_h^*. In fact, we could consider a simpler choice and use $(\kappa\nabla u_h^*, \nabla w)_T = -(q_h, \nabla w)_T$ instead of (2.29a). However, the term involving $\mathcal{F}$ plays a key role in deriving the error estimator.

Consider real numbers $l_u, l_{\boldsymbol{q}} \in [0, k]$ and assume that $u \in H^{l_u+2}(\mathcal{T}_h)$ and $\boldsymbol{q} \in H^{l_{\mathbf{q}}+1}(\mathcal{T}_h)$. We can sate the following result on the well posedness and convergence rate of the post-processing.

Lemma 2.6. *The local post processing u_h^* is well defined for L small enough. Moreover, if $Lh^2 < 1$ and $k \geq 1$, then*

$$\|u - u_h^*\|_{\Omega_h} \lesssim (R_h\, h)^{1/2}(h^{l_u+1}|u|_{l_u+2,\Omega_h} + h^{l_u+1}|\boldsymbol{q}|_{l_{\boldsymbol{q}}+2,\Omega_h}) + h^{l_u+2}|u|_{l_u+2,\Omega_h} + Lh^{l_u+1}|u|_{l_u+2,\Omega_h}, \tag{2.30a}$$

$$|u - u_h^*|_{1,T} \lesssim h_T^{l_u+1}|u|_{l_u+1,T} + Lh_T\|\varepsilon^u\|_{0,T} + \|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,T} + Lh_T\|u - u_h\|_{0,T}, \tag{2.30b}$$

and

$$\sum_{e\in\mathcal{E}_h^\partial} h_e^{1/2}\|[\![u_h^*]\!]\|_e \lesssim \|u - u_h^*\|_{\Omega_h}^{1/2}\left(\|u - u_h^*\|_{\Omega_h}^2 + h^2|u - u_h^*|_{1,\Omega_h}^2\right)^{1/4}. \tag{2.30c}$$

Here, R_h—which will be defined properly in the following section—is proportional to the product $h^{-1}dist(\Gamma_h, \Gamma)$. When $dist(\Gamma_h, \Gamma)$ is of order h, then R_h is of order one. This result guarantees a superconvergence of h^{k+2} if $L < h$ and R_h is of order h. If R_h is of order one, it only ensures a convergence of order $h^{k+3/2}$. However, for the linear case, [18, 21] reported numerical experiments suggesting that the order is indeed h^{k+2} even when R_h is of order one.

Proof. We will prove first that the problem (2.29) is well posed. For this, we will use a fixed point argument. Let $T \in \mathcal{T}_h$. We define the operator $S : \mathbb{P}_{k+1}(T) \to \mathbb{P}_{k+1}(T)$ as $S(\zeta) = z$, where z is the only solution of

$$(\kappa\nabla z, \nabla w)_T = -(q_h, \nabla w)_T + (\mathcal{F}(u_h), w)_T - (\mathcal{F}(\zeta), w)_T, \qquad \forall\ w \in \mathbb{P}_{k+1}(T), \tag{2.31a}$$

$$(z, w)_T = (u_h, w)_T, \qquad \forall\ w \in \mathbb{P}_0(T). \tag{2.31b}$$

Note that S is surjective because (2.31) is well-posed. We will show now that S has a unique a fixed point and in that case it is the solution of (2.29). Let $\zeta_1, \zeta_2 \in \mathbb{P}_{k+1}(T)$ such that $S(\zeta_1) = z_1$ and $S(\zeta_2) = z_2$, with z_1 and z_2 satisfying (2.31). We observe that $\zeta_1 - \zeta_2 \in \mathbb{P}_{k+1}(T)$ and

$$(\kappa\nabla(z_1 - z_2), \nabla w)_T = -(\mathcal{F}(\zeta_1) - \mathcal{F}(\zeta_2), w)_T \qquad \forall\ w \in \mathbb{P}_{k+1}(T), \tag{2.32a}$$

$$(z_1 - z_2, w)_T = 0 \qquad \forall\ w \in \mathbb{P}_0(T). \tag{2.32b}$$

Then, for $i = 1$ and 2, we set $\overline{z}_i := \dfrac{1}{|T|}\displaystyle\int_T z_i$ and noticing that $\overline{z}_1 = \overline{z}_2$ by equation (2.32b), we have

$$\|z_1 - z_2\|_T^2 = \|(z_1 - \overline{z_1}) - (z_2 - \overline{z_2})\|_T^2 \leq C_F^2\|\kappa^{1/2}\nabla(z_1 - z_2)\|_T^2,$$

where we have used the Friedrichs inequality with constant $C_F > 0$. Taking $w = z_1 - z_2$ in (2.32a), and recalling that $\mathcal{F}$ is Lipschitz continuous with constant L, we obtain

$$\|z_1 - z_2\|_T^2 \leq C_F^2(\mathcal{F}(\zeta_2) - \mathcal{F}(\zeta_1), z_1 - z_2)_T \leq C_F^2 L \|\zeta_2 - \zeta_1\|_T \|z_1 - z_2\|_T.$$

Thus, the operator S is a contraction as long as $C_F^2 L < 1$. If that is indeed the case, it has a unique fixed point.

For the inequality (2.30a), let P_0 and P_{W^*} be the L^2-projectors into the space of constants and into W_h^* respectively and decompose

$$u - u_h^* = (I - P_{W^*})u + P_0(P_{W^*}u - u_h^*) + (I - P_0)(P_{W^*}u - u_h^*). \tag{2.33}$$

We will now proceed to bound each of the terms on the right hand side of this expression separately in order to estimate the difference $u - u_h^*$. For the first term it is easy to see that

$$\|(I - P_{W^*})u\|_{0,T} \lesssim h_T^{l_u+2}|u|_{l_u+2,T}. \tag{2.34}$$

For the second term we first notice that, since W^* is a space of piecewise polynomials, the definitions of P_{W^*} and Π_W, since $k \geq 1$, imply $P_0 P_{W^*}u = P_0 u = P_0 \Pi_W u$

$$\|P_0(P_{W^*}u - u_h^*)\|_{0,T} = \|P_0(\Pi_W u - u_h)\|_{0,T} \leq \|\Pi_W u - u_h\|_{0,T} = \|\varepsilon^u\|_{0,T}. \tag{2.35}$$

In the first equality we have made use of the fact that, due the definition of u_h^* in equation (2.29b), we have $P_0 u_h^* = P_0 u_h$.

Now we move on to the third term in (2.33) and note that for every $\boldsymbol{v}$ in the space of vector valued functions with components belonging to W_h^* and $T \in \mathcal{T}$ it holds that

$$(\kappa\nabla(u - u_h^*), \boldsymbol{v})_T = (\kappa\nabla\left(P_{W^*}u - u_h^*\right), \boldsymbol{v})_T = (\kappa\nabla\left(I - P_0\right)(P_{W^*}u - u_h^*), \boldsymbol{v})_T. \tag{2.36}$$

Moreover, for the exact solutions $(u, \boldsymbol{q})$, we have $\kappa\nabla u = -\boldsymbol{q}$ so that the difference $u - u_h^*$ satisfies

$$(\kappa\nabla(u - u_h^*), \nabla w)_T = -(\boldsymbol{q} - \boldsymbol{q}_h, \nabla w)_T + (\mathcal{F}(u_h^*) - \mathcal{F}(u), w)_T - (\mathcal{F}(u_h) - \mathcal{F}(u), w)_T,$$

for every $w \in W_h^*$ and $T \in \mathcal{T}$. Letting $w := (I - P_0)(P_{W^*}u - u_h^*) \in W^*$ and ∇w be the test functions above, and using conditions (2.36) leads to

$$(\kappa\nabla w, \nabla w)_T = -(\boldsymbol{q} - \boldsymbol{q}_h, \nabla w)_T + (\mathcal{F}(u_h^*) - \mathcal{F}(u), w)_T + (\mathcal{F}(u) - \mathcal{F}(u_h), w)_T.$$

From this equation, using the the scaling argument $\|w\|_{0,T} \lesssim h_T|w|_{1,T}$ and the inverse inequality $|w|_{1,T} \lesssim h_T^{-1}\|w\|_{0,T}$ we arrive at

$$h_T^{-2}\|w\|_{0,T}^2 \lesssim \overline{\kappa}|w|_{1,T}^2 \leq \|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,T}|w|_{1,T} + L\left(\|u - u_h^*\|_{0,T} + \|u - u_h\|_{0,T}\right)\|w\|_{0,T},$$

from which we conclude that

$$\|w\|_{0,T} \lesssim h_T\|\boldsymbol{q} - \boldsymbol{q}_h\|_{0,T} + L\,h_T^2\left(\|u - u_h^*\|_{0,T} + \|u - u_h\|_{0,T}\right).$$

Recalling the decomposition (2.33), and the estimates (2.34), (2.35) we can bound the term $\|u-u_h^*\|_{0,T}$ on the right hand side yielding

$$(1-Lh_T^2)\|w\|_{0,T} \lesssim h_T\,\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,T} + L\,h_T^2\,\left(h_T^{l_u+2}|u|_{l_u+2,T} + \|\varepsilon^u\|_{0,T} + \|u-u_h\|_{0,T}\right). \tag{2.37}$$

Combining (2.37) above with (2.34) and (2.35) once more we arrive at

$$\begin{aligned}(1-\,Lh_T^2)\,\|u-u_h^*\|_{0,T} \lesssim & (1-Lh_T^2)h_T^{l_u+2}|u|_{l_u+2,T} + (1-Lh_T^2)\|\varepsilon^u\|_{0,T} + h_T\,\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,T}\\ &+ L\,h_T^2\,\left(h_T^{l_u+2}|u|_{l_u+2,T} + \|\varepsilon^u\|_{0,T} + \|u-u_h\|_{0,T}\right)\\ \lesssim & h_T^{l_u+2}|u|_{l_u+2,T} + \|\varepsilon^u\|_{0,T} + h_T\,\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,T} + L\,h_T^2\,\|u-u_h\|_{0,T}.\end{aligned}$$

So, assuming $Lh_T^2<1$ for each $T\in\mathcal{T}_h$, results

$$\|u-u_h^*\|_{0,T} \lesssim h_T^{l_u+2}|u|_{l_u+2,T} + \|\varepsilon^u\|_{0,T} + h_T\,\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,T} + Lh_T^2\|u-u_h\|_{0,T}.$$

By adding on each $T\in\mathcal{T}_h$, the estimate (2.30a) is concluded after considering the results in Theorem 2.2. Now, if we apply the inverse inequality to the estimate above, we arrive at

$$(1-Lh_T^2)|w|_{1,T} \lesssim \|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,T} + Lh_T\ \left(h_T^{l_u+2}|u|_{l_u+2,T} + \|\varepsilon^u\|_{0,T} + \|u-u_h\|_{0,T}\right).$$

Assuming again $Lh_T^2<1$ for each $T\in\mathcal{T}_h$, (2.30b) follows.

Finally, using the trace inequality, the fact that $h_e\|v\|_{0,e}^2 \lesssim \|v\|_{0,T}\left(\|v\|_{0,T}^2 + h_T^2|v|_{1,T}^2\right)^{1/2}$ for any $v\in[H^1(K)]^d$, and the estimates (2.30a) and (2.30b), we have

$$\begin{aligned}\sum_{e\in\mathcal{E}_h} h_e\|[\![u_h^*]\!]\|_{0,e}^2 \lesssim & \sum_{e\in\mathcal{E}_h}\sum_{T'\in\omega_e} h_e\|u-u_h^*|_{T'}\|_{0,e}^2\\ \lesssim & \sum_{e\in\mathcal{E}_h}\sum_{T'\in\omega_e} \|u-u_h^*\|_{0,T'}\left(\|u-u_h^*\|_{0,T'}^2 + h_{T'}^2|u-u_h^*|_{1,T'}^2\right)^{1/2},\end{aligned}$$

which implies (2.30c). ∎

2.5.2 A residual-based error estimator.

In order to prevent the proliferation of high order (with respect to h) oscillatory terms that would only make the analysis more cumbersome, we will suppose for the remainder of this sections that φ is the trace of a function in $W_h^*\cap H^1(\Omega_h)$. Let $(\mathcal{T}_h,\mathcal{E}_h^\circ,\mathcal{E}_h^\partial)$ refer to the elements, interior faces and boundary faces of the computational mesh respectively. In each element $T\in\mathcal{T}_h$, we define the following residual-type local error estimator:

$$\begin{aligned}\eta_T(\boldsymbol{q}_h,u_h^*,\varphi_h) := & \left(h_T^2\|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|_T^2 + \|\kappa^{1/2}\nabla u_h^* + \kappa^{-1/2}\boldsymbol{q}_h\|_T^2\right.\\ &\left. + \sum_{e\in\mathcal{E}^\circ\cap T}\left(h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2 + h_e^{-1}\|[\![u_h^*]\!]\|_e^2\right) + \sum_{e\in\mathcal{E}^\partial\cap T} h_e^{-1}\|(\varphi_h-u_h^*)\|_e^2\right)^{1/2},\end{aligned} \tag{2.38}$$

and introduce the data oscillation term

$$\mathrm{osc}^2(\mathcal{F},\mathcal{T}_h) := \sum_{T\in\mathcal{T}_h} h_T^2 \|\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*)\|_T^2. \tag{2.39}$$

We will show that the global error estimator, given by

$$\eta := \left(\sum_{T\in\mathcal{T}_h} \eta_T^2(\boldsymbol{q}_h, u_h^*, \varphi_h)\right)^{1/2}, \tag{2.40}$$

constitutes a reliable and efficient local *a posteriori* estimator for the error

$$\mathrm{e}_h^2 := \|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + \|u-u_h^*\|_{\Omega_h}^2 + \|h_e^{-1/2}(\varphi-\varphi_h)\|_{\Gamma_h}^2. \tag{2.41}$$

The remaining part of this section is devoted to proving one of the main contributions of this work, which is the efficiency and reliability of the local error estimator (2.38). We state the result here, and will proceed to develop the tools required for its proof. This will follow readily from Theorem 2.3 and Theorem 2.4, the proof of which is lengthy and requires a few technical lemmas.

2.5.3 Reliability and local efficiency.

If the Lipschitz constant L associated to the source term $\mathcal{F}$ and the distance between Γ_h and Γ is small enough (in a sense that will be made clear in the hypothesis of Theorem 2.3), then the error estimator η is reliable, i.e.,

$$\mathrm{e}_h^2 \lesssim \eta^2 + \mathrm{osc}^2(\mathcal{F},\mathcal{T}_h).$$

Moreover, η is locally efficient, meaning that

$$\eta_T^2(\boldsymbol{q}_h, u_h^*, \varphi_h) \lesssim \sum_{T\in\mathcal{U}_h(e)} \left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2 + \|u-u_h^*\|_T^2\right) + h_e^{-1}\|\varphi-\varphi_h\|_e^2 + \mathrm{osc}^2(\mathcal{F},\mathcal{U}_h(e)),$$

where $\mathcal{U}_h(e)$ is the set of elements that have e as an face. Namely, $\mathcal{U}_h(e) := \{T\in\mathcal{T}_h : T\cap\mathcal{E}_h = e\}$.

Before setting out to show the validity of this result, we would like to make a few remarks regarding the steps required for the proof. The efficiency of the estimator can be established by adapting some of the arguments in [24] to account for the semi-linearity of the problem and for the approximation of the boundary data due to the curved boundaries. This will be addressed at the end in Theorem 2.4. Reliability, however, requires a much lengthier argument and the proof will be divided in several steps. Lemma 2.7 establishes the connection between the residual of the HDG equation (2.6a) and the post-processed solution u_h^* that appears in the local error estimator. To show that each of the terms in the estimator are indeed upper bounds for the error in the flux, the term $\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2$ will be decomposed in four components which will be treated separately in Lemma 2.8. This shows that the error on the flux can be successfully bounded by the estimator, plus some additional terms involving the error the scalar variable u and the data in the boundary φ. Next, Lemma 2.9 shows an estimation for the error in the variable u in terms of that of the transferred boundary condition and the flux. All these results are then consolidated in Theorem 2.3, which establishes that the error can be controlled

by a combination of the estimator η and the data oscillation, that is, the reliability is finally proven.

As mentioned in Section 2.5.1, the inclusion of the nonlinear source term $\mathcal{F}$ in the definition of u_h^* helps obtaining the following result, which is important for the estimate in Lemma 2.8, that will link the post processed solution with Equation (2.6b).

Lemma 2.7. *Let $(u_h, \boldsymbol{q}_h)$ be the solutions to (2.6) and u_h^* be the post-processing of u_h given by (2.29). It holds*

$$(P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, w)_{\mathcal{T}_h} = -\langle \boldsymbol{q}_h\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} - (\kappa\nabla u_h^* + \boldsymbol{q}_h, \nabla w)_{\mathcal{T}_h}, \quad \forall\ w\in W_{1,h}^c,$$

where $W_{1,h}^c := \{w\in H_0^1(\Omega) : w|_T\in\mathbb{P}_1(T),\ \forall\ T\in\mathcal{T}_h\}$, and $\mathbb{P}_1(T)$ is the space of piecewise linear polynomials on T.

Proof. Considering $w\in W_{1,h}^c$ and integrating by parts in the equation (2.6b) we obtain, for all $T\in\mathcal{T}_h$

$$(\nabla\cdot\boldsymbol{q}_h, w)_T - \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, w\rangle_{\partial T} + \langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}, w\rangle_{\partial T} = (\mathcal{F}(u_h), w)_T.$$

Then, due to $\langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = 0$ and using (2.29), we have

$$(\nabla\cdot\boldsymbol{q}_h, w)_{\mathcal{T}_h} - \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = (\kappa\nabla u_h^*, \nabla w)_{\mathcal{T}_h} + (\mathcal{F}(u_h^*), w)_{\mathcal{T}_h} + (\boldsymbol{q}_h, \nabla w)_{\mathcal{T}_h},$$

which concludes the proof. ∎

In what follows, $\widetilde{u}_h^*\in W_h^*\cap H^1(\Omega_h)$ such that $\widetilde{u}_h^* = \varphi$ on Γ_h, will be used to denote the so-called Oswald interpolation of u_h^* defined in 1.6. Now, we apply the Lemma 1.6, with $|\alpha| = 1$, to get

$$\begin{aligned}\|\nabla(\widetilde{u}_h^* - u_h^*)\|_{\Omega_h}^2 &\le C_O\left(\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![u_h^*]\!]\|_e^2 + \sum_{e\in\mathcal{E}_h^\partial} h_e^{-1}\|\varphi - u_h^*\|_e^2\right)\\ &\le C_O\left(\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![u_h^*]\!]\|_e^2 + 2\sum_{e\in\mathcal{E}_h^\partial} h_e^{-1}\|\varphi-\varphi_h\|_e^2 + 2\sum_{e\in\mathcal{E}_h^\partial} h_e^{-1}\|\varphi_h - u_h^*\|_e^2\right),\end{aligned}$$

where $C_O > 0$ is a constant independent of h arising from the approximation properties of the Oswald's interpolant. Similarly, for $|\alpha| = 0$, we have

$$\|\widetilde{u}_h^* - u_h^*\|_{\Omega_h}^2 \le C_O\left(\sum_{e\in\mathcal{E}_h^\circ} h_e\|[\![u_h^*]\!]\|_e^2 + 2\sum_{e\in\mathcal{E}_h^\partial} h_e\|\varphi-\varphi_h\|_e^2 + 2\sum_{e\in\mathcal{E}_h^\partial} h_e\|\varphi_h - u_h^*\|_e^2\right).$$

Since for a fine enough mesh $h_e\le h_e^{-1}$, the two inequalities above can be combined into

$$\|\widetilde{u}_h^* - u_h^*\|_{\Omega_h}^2 + \|\nabla(\widetilde{u}_h^* - u_h^*)\|_{\Omega_h}^2 \le 2C_O\left(\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|[\![u_h^*]\!]\|_e^2 + \sum_{e\in\mathcal{E}_h^\partial} h_e^{-1}\|\varphi_h - u_h^*\|_e^2 + \|h_e^{-1/2}(\varphi-\varphi_h)\|_{\Gamma_h}^2\right). \tag{2.42}$$

The following three results allow us to find a preliminary estimate for each term of our error defined in (2.41). We begin rewrite $\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\mathcal{T}_h}$ in a suitable manner. Note first that using (2.3a), and adding and subtracting $\widetilde{u}_h^*$, it follows

$$\begin{aligned}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_T &= (\boldsymbol{q}-\boldsymbol{q}_h, \kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h))_T \\ &= -(\boldsymbol{q}-\boldsymbol{q}_h, \nabla(u-\widetilde{u}_h^*))_T - (\boldsymbol{q}-\boldsymbol{q}_h, \nabla\widetilde{u}_h^* - \kappa^{-1}\boldsymbol{q}_h)_T \\ &= (\nabla\cdot(\boldsymbol{q}-\boldsymbol{q}_h), u-\widetilde{u}_h^*)_T - \langle(\boldsymbol{q}-\boldsymbol{q}_h)\cdot\boldsymbol{n}, u-\widetilde{u}_h^*\rangle_{\partial T} - (\boldsymbol{q}-\boldsymbol{q}_h, \nabla\widetilde{u}_h^* - \kappa^{-1}\boldsymbol{q}_h)_T.\end{aligned}$$

Adding and subtracting $\mathcal{F}(u_h^*)$ and $P_W\mathcal{F}(u_h^*)$ in the first term above, and using (2.3b) to replace $\nabla\cdot\boldsymbol{q}$ by $\mathcal{F}(u)$ yields

$$\begin{aligned}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_T = (\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*) + P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h + \mathcal{F}(u) - \mathcal{F}(u_h^*), u-\widetilde{u}_h^*)_T \\ - \langle(\boldsymbol{q}-\boldsymbol{q}_h)\cdot\boldsymbol{n}, u-\widetilde{u}_h^*\rangle_{\partial T} - (\boldsymbol{q}-\boldsymbol{q}_h, \nabla\widetilde{u}_h^* - \kappa^{-1}\boldsymbol{q}_h)_T.\end{aligned}$$

Thus, since $\boldsymbol{q}\in H(\mathrm{div};\Omega_h)$ and $u-\widetilde{u}_h^*\in H_0^1(\Omega_h)$, we can add over the entire grid to obtain

$$\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h} := \sum_{i=1}^{4}\mathbb{T}_i, \tag{2.43}$$

where

$$\begin{aligned}\mathbb{T}_1 &:= (\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*), u-\widetilde{u}_h^*)_{\mathcal{T}_h} &,\quad \mathbb{T}_3 &:= -(\boldsymbol{q}-\boldsymbol{q}_h, \nabla\widetilde{u}_h^* - \kappa^{-1}\boldsymbol{q}_h)_{\mathcal{T}_h} \\ \mathbb{T}_2 &:= (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, u-\widetilde{u}_h^*)_{\mathcal{T}_h} + \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, u-\widetilde{u}_h^*\rangle_{\partial\mathcal{T}_h} &,\quad \mathbb{T}_4 &:= (\mathcal{F}(u) - \mathcal{F}(u_h^*), u-\widetilde{u}_h^*)_{\mathcal{T}_h}.\end{aligned}$$

In the following estimates, for a given function v, let $Q_k(v)$ be the averaged Taylor polynomial of degree k associated to v. For smooth functions this polynomial coincides with the "usual" Taylor polynomial, whereas for functions with Sobolev regularity it is defined by mollification of the weakly defined Taylor polynomial [7, Section 4.1].

Lemma 2.8. *There exists $\overline{C}_1 > 0$, independent of h such that*

$$\begin{aligned}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h} \le \overline{C}_1 \Bigg(osc^2(\mathcal{F},\mathcal{T}_h) + \sum_{T\in\mathcal{T}_h} h_T^2\|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|^2_T \\ + \sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla u_h^* + \kappa^{-1/2}\boldsymbol{q}_h\|^2_T + \sum_{e\in\mathcal{E}_h^\circ}\left(h_e\,\|[\![\boldsymbol{q}_h]\!]\|^2_e + h_e^{-1}\,\|[\![u_h^*]\!]\|^2_e\right) + \sum_{e\in\mathcal{E}_h^\partial} h_e^{-1}\|\varphi_h - u_h^*\|^2_e \Bigg) \\ + \overline{C}_1\,\|h_e^{-1/2}\,(\varphi-\varphi_h)\|^2_{\Gamma_h} + \overline{C}_1\,L\,\|u-u_h^*\|^2_{\Omega_h}.\end{aligned}$$

Proof. To prove the result, we will bound each of the terms $\mathbb{T}_i$ in the decomposition separately. The final result will come as a consequence of the individual estimates. In some cases we will make use of a free parameter $\epsilon_j > 0$.

Bound for $\mathbb{T}_1$. Consider $Q_0(u-\widetilde{u}_h^*)$, the zeroth order averaged Taylor polynomial associated to $u-\widetilde{u}_h^*$. Since $(\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*), Q_0(u-\widetilde{u}_h^*))_T = 0$, then by Young's inequality and the Bramble-

Hilbert lemma with constant $c > 0$, independent of h, we obtain

$$(h_T(\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*)), h_T^{-1}(u - \widetilde{u}_h^* - Q_0(u - \widetilde{u}_h^*)))_T \leq \frac{h_T^2}{4\epsilon_1}\|\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*)\|_T^2 + c\,\epsilon_1\,\|\nabla(u - \widetilde{u}_h^*)\|_T^2.$$

Using (2.4a) in the last term of the above expression to replace ∇u by $\kappa^{-1}\boldsymbol{q}$ along with adding and subtracting ∇u_h^* and $\kappa^{-1}\boldsymbol{q}_h$, we obtain

$$\|\nabla(u - \widetilde{u}_h^*)\|_T^2 \leq \frac{3}{\underline{\kappa}}\left(\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|\kappa^{1/2}\nabla u_h^* + \kappa^{-1/2}\boldsymbol{q}_h\|_T^2 + \|\kappa^{1/2}\nabla u_h^* - \kappa^{1/2}\nabla\widetilde{u}_h^*\|_T^2\right). \quad (2.44)$$

Thus

$$\begin{aligned}|\mathbb{T}_1| \leq \quad & \widehat{C}_1\epsilon_1 \sum_{T\in\mathcal{T}_h}\left(\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|\kappa^{1/2}\nabla u_h^* + \kappa^{-1/2}\boldsymbol{q}_h\|_T^2\|\kappa^{1/2}\nabla u_h^* - \kappa^{1/2}\nabla\widetilde{u}_h^*\|_T^2\right) \\ & + \frac{\widehat{C}_1}{4\,\epsilon_1}\sum_{T\in\mathcal{T}_h} h_T^2\|\mathcal{F}(u_h^*) - P_W\mathcal{F}(u_h^*)\|_T^2, \end{aligned} \quad (2.45)$$

where $\widehat{C}_1 := \max\{1, 3\,c\,\underline{\kappa}^{-1}\}$.

Bound for $\mathbb{T}_2$. We begin by rewriting $\mathbb{T}_2$ as

$$\begin{aligned}\mathbb{T}_2 = & \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, u - \widetilde{u}_h^*\rangle_{\partial\mathcal{T}_h} + (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, (u - \widetilde{u}_h^*) - \mathcal{C}_h(u - \widetilde{u}_h^*))_{\mathcal{T}_h} \\ & + (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, \mathcal{C}_h(u - \widetilde{u}_h^*))_{\mathcal{T}_h},\end{aligned}$$

where $\mathcal{C}_h$ is the Clément interpolation operator defined in Section 1.8. Rearranging terms above, using $u = \widetilde{u}_h^* = \varphi$ on Γ_h, and applying Lemma 2.7, we have

$$\begin{aligned}\mathbb{T}_2 &= \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, u - \widetilde{u}_h^*\rangle_{\partial\mathcal{T}_h} + (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, (Id_M - \mathcal{C}_h)(u - \widetilde{u}_h^*))_{\mathcal{T}_h} - \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, \mathcal{C}_h(u - \widetilde{u}_h^*)\rangle_{\partial\mathcal{T}_h} \\ &\quad - (\kappa\nabla u_h^* + \boldsymbol{q}_h, \nabla\mathcal{C}_h(u - \widetilde{u}_h^*))_{\mathcal{T}_h} \\ &= \langle\boldsymbol{q}_h\cdot\boldsymbol{n}, (Id_M - \mathcal{C}_h)(u - \widetilde{u}_h^*)\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} + (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, (Id_M - \mathcal{C}_h)(u - \widetilde{u}_h^*))_{\mathcal{T}_h} \\ &\quad - (\kappa\nabla u_h^* + \boldsymbol{q}_h, \nabla\mathcal{C}_h(u - \widetilde{u}_h^*))_{\mathcal{T}_h}\end{aligned}$$

Then, by Young's inequality,

$$\begin{aligned}|\mathbb{T}_2| \leq & \frac{1}{4\,\epsilon_2}\sum_{T\in\mathcal{T}_h} h_T^2\|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|_T^2 + \epsilon_2\sum_{T\in\mathcal{T}_h} h_T^{-2}\|(Id_M - \mathcal{C}_h)(u - \widetilde{u}_h^*)\|_T^2 \\ & + \frac{1}{4\,\epsilon_2}\sum_{e\in\mathcal{E}_h^\circ} h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2 + \epsilon_2\sum_{e\in\mathcal{E}_h^i} h_e^{-1}\|(Id_M - \mathcal{C}_h)(u - \widetilde{u}_h^*)\|_e^2 \\ & + \frac{\overline{\kappa}}{4\,\epsilon_2}\sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla u_h^* + \kappa^{-1/2}\boldsymbol{q}_h\|_T^2 + \epsilon_2\sum_{T\in\mathcal{T}_h}|\mathcal{C}_h(u - \widetilde{u}_h^*)|_{1,T}^2.\end{aligned}$$

On the other hand, the properties of Clément's interpolant—Lemma 1.5—and the Poincaré inequality with constant c_p imply that

$$\sum_{T\in\mathcal{T}_h} h_T^{-2}\|(Id_M-\mathcal{C}_h)(u-\widetilde{u}_h^*)\|_T^2 \lesssim \sum_{T\in\mathcal{T}_h}\|u-\widetilde{u}_h^*\|_{\Delta_T}^2 \leq \widehat{c}_1\, c_p \sum_{T\in\mathcal{T}_h}|u-\widetilde{u}_h^*|_{1,T}^2,$$
$$\sum_{e\in\mathcal{E}_h^\circ} h_e^{-1}\|(Id_M-\mathcal{C}_h)(u-\widetilde{u}_h^*)\|_e^2 \lesssim \sum_{e\in\mathcal{E}_h^\circ}\|u-\widetilde{u}_h^*\|_{\Delta_e}^2 \leq \widehat{c}_2\, c_p \sum_{T\in\mathcal{T}_h}|u-\widetilde{u}_h^*|_{1,T}^2,$$
$$\sum_{T\in\mathcal{T}_h}|\mathcal{C}_h(u-\widetilde{u}_h^*)|_{1,T}^2 \lesssim \sum_{T\in\mathcal{T}_h}\|u-\widetilde{u}_h^*\|_T^2 \leq \widehat{c}_3\, c_p \sum_{T\in\mathcal{T}_h}|u-\widetilde{u}_h^*|_{1,T}^2.$$

Above, the sets Δ_T and Δ_e correspond to the macro element surrounding the element T and face e respectively, i.e.

$$\Delta_T := \{T'\in\mathcal{T}_h : \overline{T}\cap\overline{T'}\neq\varnothing\} \qquad \text{and} \qquad \Delta_e = \{T'\in\mathcal{T}_h : \overline{T'}\cap\overline{e}\neq\varnothing\}.$$

Then, applying (2.44) to the right side terms of the last three inequalities, one arrives at

$$\begin{aligned}|\mathbb{T}_2| \leq\ & \frac{\widehat{C}_2}{4\epsilon_2}\left(\sum_{T\in\mathcal{T}_h} h_T^2\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|_T^2+\sum_{e\in\mathcal{E}_h^\circ} h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2\right)\\ &+\epsilon_2\widehat{C}_2\sum_{T\in\mathcal{T}_h}\left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2+\left(\frac{1}{4\,\epsilon_2}+1\right)\|\kappa^{1/2}\nabla u_h^*+\kappa^{-1/2}\boldsymbol{q}_h\|_T^2+\|\kappa^{1/2}\,\nabla(u-\widetilde{u}_h^*)\|_T\right),\end{aligned} \tag{2.46}$$

with $\widehat{C}_2=\max\{1,\overline{\kappa},3\,c_p\,\underline{\kappa}^{-1}(\widehat{c}_1+\widehat{c}_2+\widehat{c}_3)\}$.

Bound for $\mathbb{T}_3$. From Young's inequality, it follows that

$$|\mathbb{T}_3|\leq\sum_{T\in\mathcal{T}_h}\left(\frac{1}{2\,\epsilon_3}\left(\|\kappa^{1/2}\nabla u_h^*+\kappa^{-1/2}\boldsymbol{q}_h\|_T^2+\|\kappa^{1/2}\,\nabla(\widetilde{u}_h^*-u_h^*)\|_T^2\right)+\epsilon_3\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2\right). \tag{2.47}$$

Bound for $\mathbb{T}_4$. Adding and subtracting u_h^*, and using the Lipschitz continuity of $\mathcal{F}$, we have

$$|\mathbb{T}_4|\leq L\sum_{T\in\mathcal{T}_h}\left(\|u-u_h^*\|_T^2+\|u-u_h^*\|_T\|u_h^*-\widetilde{u}_h^*\|_T\right)\leq\frac{L}{2}\sum_{T\in\mathcal{T}_h}\left(3\|u-u_h^*\|_T^2+\|u_h^*-\widetilde{u}_h^*\|_T^2\right), \tag{2.48}$$

where the second inequality follows from Young's inequality.

Wrap-up. By the decomposition (2.43) and the bounds (2.45) - (2.48) obtained for the terms $\mathbb{T}_i$, we deduce that

$$\begin{aligned}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h} \leq\ & \frac{\widehat{C}_1}{4\,\epsilon_1}\sum_{T\in\mathcal{T}_h} h_T^2\|\mathcal{F}(u_h^*)-P_W\mathcal{F}(u_h^*)\|^2_{0,T}+\frac{\widehat{C}_2}{4\,\epsilon_2}\sum_{T\in\mathcal{T}_h} h_T^2\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|^2_T\\ &+\left(\widehat{C}_1\,\epsilon_1+\widehat{C}_2\,\epsilon_2+\frac{\widehat{C}_2}{4\,\epsilon_2}+\frac{1}{2\,\epsilon_3}\right)\sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla u_h^*+\kappa^{-1/2}\boldsymbol{q}_h\|^2_T+\frac{\widehat{C}_2}{4\,\epsilon_2}\sum_{e\in\mathcal{E}_h^\circ} h_e\,\|[\![\boldsymbol{q}_h]\!]\|^2_e\\ &+\left(\widehat{C}_1\,\epsilon_1+\widehat{C}_2\,\epsilon_2+\frac{1}{2\,\epsilon_3}\right)\overline{\kappa}\sum_{T\in\mathcal{T}_h}\|\nabla(u_h^*-\widetilde{u}_h^*)\|^2_T+\frac{L}{2}\sum_{T\in\mathcal{T}_h}\|u_h^*-\widetilde{u}_h^*\|^2_T+\frac{3\,L}{2}\sum_{T\in\mathcal{T}_h}\|u-u_h^*\|^2_T\\ &+(\widehat{C}_1\,\epsilon_1+\widehat{C}_2\,\epsilon_2+\epsilon_3)\sum_{T\in\mathcal{T}_h}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_T.\end{aligned}$$

Finally, considering values of ϵ_1,ϵ_2, and ϵ_3 such that $\widehat{C}_1\,\epsilon_1+\widehat{C}_2\,\epsilon_2+\epsilon_3<1/2$, and the estimate for the terms that involve $\widetilde{u}_h^*$, given in (2.42), the proof in concluded with $\overline{C}_1$ dependent only of $\widehat{C}_1$ and $\widehat{C}_2$. ■

Now, we bound the second term of the error e_h^2 (see (2.41)).

Lemma 2.9. *Under all the previous assumptions, the following bound for the error in the post processed solution holds*

$$\begin{aligned}\|u-u_h^*\|^2_{\Omega_h}\leq\ \overline{C}_2\Bigg(&\sum_{T\in\mathcal{T}_h}\|\kappa^{-1/2}\boldsymbol{q}_h+\kappa^{1/2}\nabla u_h^*\|^2_T+\sum_{e\in\mathcal{E}_h^\circ}h_e^{-1}\|[\![u_h^*]\!]\|^2_e+\sum_{e\in\mathcal{E}_h^\partial}h_e^{-1}\|\varphi_h-u_h^*\|^2_e\\ &+\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h}+\|h_e^{-1/2}\,(\varphi-\varphi_h)\|^2_{\Gamma_h}\Bigg),\end{aligned}$$

where $\overline{C}_2>0$ is a positive constant independent of h.

Proof. First, note that, since $u-\widetilde{u}_h^*\in H_0^1(\Omega_h)$, then thanks to the triangle and Poincaré inequalities with constant c_p, it follows that

$$\|u-u_h^*\|^2_{\Omega_h}\leq 2\,\|u-\widetilde{u}_h^*\|^2_{\Omega_h}+2\,\|\widetilde{u}_h^*-u_h^*\|^2_{\Omega_h}\leq 2\,c_p^2\|\nabla u-\nabla\widetilde{u}_h^*\|^2_{\Omega_h}+2\,\|\widetilde{u}_h^*-u_h^*\|^2_{\Omega_h}.$$

then, since $\boldsymbol{q}=-\kappa^{-1}\,\nabla u$ (see (2.3a)) adding $\pm\kappa^{-1}\boldsymbol{q}_h$, we get

$$\|u-u_h^*\|^2_{\Omega_h}\leq 4\,c_p^2\,\underline{\kappa}^{-1}\left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h}+\|\kappa^{-1/2}\boldsymbol{q}_h+\kappa^{1/2}\nabla\widetilde{u}_h^*\|^2_{\Omega_h}\right)+2\,\|\widetilde{u}_h^*-u_h^*\|^2_{\Omega_h}.$$

Now, adding $\pm\kappa^{-1/2}\,\nabla u_h^*$, results in

$$\begin{aligned}\|u-u_h^*\|^2_{\Omega_h}\leq\ & 4\,c_p^2\,\underline{\kappa}^{-1}\,\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|^2_{\Omega_h}+8\,c_p^2\,\underline{\kappa}^{-1}\,\|\kappa^{-1/2}\boldsymbol{q}_h+\kappa^{1/2}\nabla u_h^*\|^2_{\Omega_h}\\ &+2\,\max\{4\,c_p^2\,\underline{\kappa}^{-1}\,\overline{\kappa},1\}\left(\|\nabla(u_h^*-\widetilde{u}_h^*)\|^2_{\Omega_h}+\,\|\widetilde{u}_h^*-u_h^*\|^2_{\Omega_h}\right).\end{aligned}\tag{2.49}$$

Finally, the proof is concluded by substituting (2.42) into (2.49). ■

We conclude this part with an estimate for the last term of our error,

Lemma 2.10. *Assume that all previous assumptions are satisfied. Then, there exists a positive constant $\overline{C}_3$, independent of h such that*

$$\|h_e^{-1/2}(\varphi-\varphi_h)\|_{\Gamma_h}^2 \leq \overline{C}_3 \max_{e\in\mathcal{E}_h^\partial}\{r_e^2, r_e\,(C_{ext}^e)^2\}\left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + h^2\,L^2\,\|u-u_h^*\|_{\Omega_h}^2 + \sum_{T\in\mathcal{T}_h} h_T^2\,\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|_T^2 + osc^2(\mathcal{F},\mathcal{T}_h)\right).$$

Proof. We first notice that this term depends on what happens in the domain Ω_h^c, that is

$$\|h_e^{-1/2}(\varphi-\varphi_h)\|_{\Gamma_h}^2 \lesssim \|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h^c}^2. \tag{2.50}$$

Then, for each $T\in\mathcal{T}_h$, we have

$$\begin{aligned} h_T\|\nabla\cdot(\boldsymbol{q}-\boldsymbol{q}_h)\|_T &= h_T\|\mathcal{F}(u)-\nabla\cdot\boldsymbol{q}_h\|_T \\ &\leq h_T\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|_T + h_T\|\mathcal{F}(u)-\mathcal{F}(u_h^*)\|_T + h_T\|P_W\mathcal{F}(u_h^*)-\mathcal{F}(u_h^*)\|_T. \end{aligned} \tag{2.51}$$

Now we will need to consider the approximation error measured in a function space with additional regularity. For $T\in\mathcal{T}_h$ let $E_T : \boldsymbol{H}(\mathbf{div};T)\to\boldsymbol{H}(\mathbf{div};\mathbb{R}^d)$ be any local extension operator, and $Q_k(E_T(\boldsymbol{q}))\in\mathbb{P}_k(\mathbb{R}^d)$ the averaged averaged Taylor polynomial of degree k introduced in the proof of Lemma 2.7. Let also $E:\boldsymbol{H}(\mathbf{div};\mathcal{T}_h)\to\boldsymbol{H}(\mathbf{div};\mathbb{R}^d)$ be a global extension such that $E(\boldsymbol{v})|_T := E_T(\boldsymbol{v})$ for all $T\in\mathcal{T}_h$ and $\boldsymbol{v}\in\boldsymbol{H}(\mathbf{div};\mathcal{T}_h)$. Note that

$$\begin{aligned} \|\boldsymbol{q}-\boldsymbol{q}_h\|_{\Omega_h^c} &\leq \|Q_k(E(\boldsymbol{q}))-\boldsymbol{q}_h\|_{\Omega_h^c} + \|\boldsymbol{q}-Q_k(E(\boldsymbol{q}))\|_{\Omega_h^c} \\ &= \|\boldsymbol{q}-Q_k(E(\boldsymbol{q}))\|_{\Omega_h^c} + \left(\sum_{e\in\mathcal{E}_h^\partial}\|Q_k(E(\boldsymbol{q}))-\boldsymbol{q}_h\|_{T_{ext}^e}^2\right)^{1/2} \\ &\leq \|E(\boldsymbol{q})-Q_k(E(\boldsymbol{q}))\|_{\Omega_h^c} + \left(\sum_{e\in\mathcal{E}_h^\partial} r_e\,(C_{ext}^e)^2\|Q_k(E(\boldsymbol{q}))-\boldsymbol{q}_h\|_{T^e}^2\right)^{1/2}. \end{aligned}$$

Since $\|Q_k(E(\boldsymbol{q}))-\boldsymbol{q}_h\|_{T^e}^2 = \|Q_k(E(\boldsymbol{q})-\boldsymbol{q}_h)\|_{T^e}^2 \lesssim \|E(\boldsymbol{q})-\boldsymbol{q}_h\|_{T^e}^2 = \|\boldsymbol{q}-\boldsymbol{q}_h\|_{T^e}^2$, we obtain

$$\|\boldsymbol{q}-\boldsymbol{q}_h\|_{\Omega_h^c} \lesssim \|E(\boldsymbol{q})-Q_k(E(\boldsymbol{q}))\|_{\Omega_h^c} + \max_{e\in\mathcal{E}_h^\partial}\{r_e^{1/2}\,C_{ext}^e\}\|\boldsymbol{q}-\boldsymbol{q}_h\|_{\Omega_h}.$$

Moreover, to bound the first term on the right hand side, we observe that

$$\begin{aligned} \|E_{T^e}(\boldsymbol{q})-Q_k(E_{T^e}(\boldsymbol{q}))\|_{T_{ext}^e}^2 &= \|E_{T^e}(\boldsymbol{q}-\boldsymbol{q}_h)-Q_k(E_{T^e}(\boldsymbol{q})-E_{T^e}(\boldsymbol{q_h}))\|_{T_{ext}^e}^2 \\ &\lesssim \|E_{T^e}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{T_{ext}^e}^2 \lesssim r_e^2\,\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{T^e} + r_e^2\,h_T^2\,\|\nabla\cdot(\boldsymbol{q}-\boldsymbol{q}_h)\|_{T^e}^2, \end{aligned}$$

where we have used the estimate in (1.14a). Thus,

$$\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h^c}^2 \lesssim \max_{e\in\mathcal{E}_h^\partial}\{r_e^2, r_e\,(C_{ext}^e)^2\}\,\left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + h_T^2\|\nabla\cdot(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2\right).$$

The result follows combining the last inequality with (2.50). ■

With all the pieces in place, we can now show that the error in the flux can be successfully estimated if one considers the data oscillation.

Theorem 2.3. *Assume that the hypotheses of Lemmas 2.8–2.10 hold. In addition, if*

$$\overline{C}_1\,\overline{C}_3\,\max_{e\in\mathcal{E}_h^\partial}\{r_e^2, r_e\,(C_{ext}^e)^2\} < 1/2, \tag{2.52a}$$

$$\overline{C}_2\,L\,(L\,h^2 + 2\,\overline{C}_1) < 1/2, \tag{2.52b}$$

$$(2\,\overline{C}_2+1)\,\overline{C}_3\,\max_{e\in\mathcal{E}_h^\partial}\{r_e^2, r_e\,(C_{ext}^e)^2\} < 1/2, \tag{2.52c}$$

where $\overline{C}_i$, $i\in\{1,2,3\}$ *are defined in the Lemmas 2.8, 2.9 and 2.10. Then, there exists a positive constant* C_{rel}, *such that*

$$e_h^2 \le C_{rel}\left(\eta^2 + osc^2(\mathcal{F},\mathcal{T}_h)\right).$$

Proof. We first replace the estimation of the Lemma 2.10 into Lemma 2.8 and, together with assumption (2.52a), obtain

$$\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 \le L\,(L\,h^2+2\,\overline{C}_1)\,\|u-u_h^*\|_{\Omega_h}^2 + (2\,\overline{C}_1+1)\left(\mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2\right). \tag{2.53}$$

Combining the assumption (2.52b) with (2.53) into the Lemma 2.9, we obtain

$$\|u-u_h^*\|_{\Omega_h}^2 \le 2\,\overline{C}_2\,\|h_e^{-1/2}\,(\varphi-\varphi_h)\|_{\Gamma_h}^2 + 2\,(\overline{C}_1+1)\,\overline{C}_2\,\left(\mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2\right). \tag{2.54}$$

Note that, thanks to (2.54) and assumption (2.52b), the estimation (2.53) can be rewritten as

$$\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 \le \|h_e^{-1/2}\,(\varphi-\varphi_h)\|_{\Gamma_h}^2 + (3\,\overline{C}_1+2)\,\left(\mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2\right). \tag{2.55}$$

Combining (2.54) with (2.55) and using the Lemma 2.10 again, we arrives at

$$\begin{aligned}&\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + \|u-u_h^*\|_{\Omega_h}^2\\ &\quad\le (2\,\overline{C}_2+1)\,\overline{C}_3\,\max_{e\in\mathcal{E}_h^\partial}\{r_e^2, r_e\,(C_{ext}^e)^2\}\,\left(\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + \|u-u_h^*\|_{\Omega_h}^2\right) + \widehat{c}\,\left(\mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2\right).\end{aligned}$$

Then, by assumption (2.52c), we deduce

$$\|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_{\Omega_h}^2 + \|u-u_h^*\|_{\Omega_h}^2 \lesssim \mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2.$$

Finally, observe that the above estimation allows us rewritten the Lemma 2.10 as

$$\|h_e^{-1/2}\,(\varphi-\varphi_h)\|_{\Gamma_h}^2 \lesssim \mathrm{osc}^2(\mathcal{F},\mathcal{T}_h)+\eta^2,$$

which concludes the proof. ∎

Having established the reliability of the estimator we can now adapt arguments from the linear case to show that the estimator is locally efficient as well. This will follow readily from the following estiamtes.

Theorem 2.4. *Suppose that $L\,h\leq 1$. Then we can assert the following local estimates*

$$\begin{aligned}
\|\kappa^{-1/2}\,\boldsymbol{q}_h+\kappa^{1/2}\,\nabla u_h^*\|_T &\lesssim \|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_T,\\
h_e^{-1}\|[\![u_h^*]\!]\|_e^2 &\lesssim \sum_{T\in\mathcal{U}_h(e)}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2 \quad \forall\, e\in\mathcal{E}_h^\circ,\\
h_e^{-1}\,\|\varphi_h-u_h^*\|_e &\lesssim \sum_{T\in\mathcal{U}_h(e)}\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2+\|h_e^{-1/2}\,(\varphi-\varphi_h)\|_e^2 \quad \forall\, e\in\mathcal{E}_h^\partial,\\
h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2 &\lesssim \sum_{T\in\mathcal{U}_h(e)}\Big(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2+h_T^2\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|_T^2\\
&\qquad +\|u-u_h^*\|_T^2\Big)+osc^2(\mathcal{F},\mathcal{U}_h(e)) \qquad \forall\, e\in\mathcal{E}_h^\circ\\
h_T^2\,\|P_W\mathcal{F}(u_h^*)-\nabla\cdot\boldsymbol{q}_h\|_T &\lesssim \|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2+h_T^2\,\|P_W\mathcal{F}(u_h^*)-\mathcal{F}(u_h^*)\|_T^2+\|u-u_h^*\|_T^2.
\end{aligned}$$

Proof. Note that due to the presence of the non-linear source term, the post-processing defining u_h^* is also non linear, and a direct application of the results in [24, Lemmas 4.4–4.5] and [25, Lemmas 3.4-3.7] is not possible. We then proceed to adapt those arguments to the current semi-linear case and treat each of the estimates above separately in what follows. Local efficiency will follow by combining these estimates.

Bound for $\|\kappa^{-1/2}\,\boldsymbol{q}_h+\kappa^{1/2}\,\nabla u_h^*\|_T$. This term can be bounded by an application of [25, Lemma 3.7], that is

$$\|\kappa^{-1/2}\,\boldsymbol{q}_h+\kappa^{1/2}\,\nabla u_h^*\|_T \lesssim \|\kappa^{-1/2}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_T. \tag{2.56}$$

Bound for $h_e^{-1}\|[\![u_h^*]\!]\|_e^2$. We begin by splitting $[\![u_h^*]\!]$ into its component in the space defined as $M_0:=\{\mu\in L^2(\partial\mathcal{T}_h):\mu|_e\in\mathbb{P}_0(e),\ \forall\, e\in\mathcal{E}_h\}$ and its orthogonal complement. Considering P_{M_0}, the $L^2(\Omega)$-orthogonal projection into M_0, and Id_M the identity operator on the same space we have

$$h_e^{-1}\,\|[\![u_h^*]\!]\|_e^2 \lesssim h_e^{-1}\,\|P_{M_0}[\![u_h^*]\!]\|_e^2+h_e^{-1}\,\|(Id_M-P_{M_0})[\![u_h^*]\!]\|_e^2. \tag{2.57}$$

Each of these terms can be bounded by an application of [25, Lemma 3.4. and Lemma 3.5.] to all the interior faces of the triangulations. That is, for each $e \in \mathcal{E}_h^\circ$,

$$h_e^{-1} \|P_{M_0}[\![u_h^*]\!]\|_e^2 \lesssim \sum_{T \in \mathcal{U}_h(e)} \|\kappa^{-1/2} \boldsymbol{q}_h + \kappa^{1/2} \nabla u_h^*\|_T^2, \tag{2.58a}$$

$$h_e^{-1} \|(Id_M - P_{M_0})[\![u_h^*]\!]\|_e^2 \lesssim \sum_{T \in \mathcal{U}_h(e)} \|\nabla(u - u_h^*)\|_T^2. \tag{2.58b}$$

Now, adding and subtracting $\kappa^{-1/2}\boldsymbol{q}_h$ to $\nabla(u - u_h^*)$ and using the definition of the flux it follows that

$$\|\nabla(u - u_h^*)\|_T^2 \lesssim \|\kappa^{-1/2} (\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|\kappa^{-1/2} \boldsymbol{q}_h + \kappa^{1/2} \nabla u_h^*\|_T^2. \tag{2.59}$$

Moreover, using the fact that $\|\kappa^{-1/2} \boldsymbol{q}_h + \kappa^{1/2} \nabla u_h^*\|_T^2 \lesssim \|\kappa^{-1/2} (\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2$ (see (2.56)), we can bound the second term above. The same argument can be applied to (2.58a), and combining these two results we arrive at

$$h_e^{-1} \|[\![u_h^*]\!]\|_e^2 \lesssim \sum_{T \in \mathcal{U}_h(e)} \|\kappa^{-1/2} (\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 \qquad \forall\, e \in \mathcal{E}_h^\circ. \tag{2.60}$$

Bound for $h_e^{-1}\|\varphi_h - u_h^*\|_e^2$. First, we define for each $T \in \mathcal{T}_h$, the local Raviart-Thomas [33] space of order k as

$$\mathbb{RT}_k(T) := [\mathbb{P}_k(T)]^d \oplus \mathbb{P}_k(T)\, \boldsymbol{x},$$

where $\mathbb{P}_k(T)$ denotes the space of polynomials of degree at most k defined in $T \in \mathcal{T}_h$ (see Section 2.2).

Taking as test in (2.6a) $\boldsymbol{v} \in \mathbb{RT}_0(T)$, it is possible to use the second equation defining the post processing u_h^*, (2.29b), for $\nabla \cdot \boldsymbol{v}$ belongs to the space of piece-wise constant functions $\mathbb{P}_0(T)$, to obtain

$$(\kappa^{-1} \boldsymbol{q}_h, \boldsymbol{v})_T - (u_h^*, \nabla \cdot \boldsymbol{v})_T + \langle \widehat{u}_h, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial T} = (\kappa^{-1} \boldsymbol{q}_h + \nabla u_h^*, \boldsymbol{v})_T + \langle \widehat{u}_h - u_h^*, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial T} = 0.$$

On the other hand, if we consider $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, \mathcal{U}_h(e))$ for each $e \in \mathcal{E}_h^\partial$, then by summing over all $T \in \mathcal{U}_h(e)$, we arrive at

$$\sum_{T \in \mathcal{U}_h(e)} (\kappa^{-1} \boldsymbol{q}_h + \nabla u_h^*, \boldsymbol{v})_T + \sum_{T \in \mathcal{U}_h(e)} \sum_{F \in \partial T \setminus e} \langle \widehat{u}_h - u_h^*, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial T} - \langle \varphi_h - u_h^*, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_e.$$

Since $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, \mathcal{U}_h(e))$ is arbitrary, we can choose it such that, on each $T \in \mathcal{U}_h(e)$, belongs to $\mathbb{RT}_0(T)$ and satisfies

$$\begin{aligned} &\int_e \boldsymbol{v} \cdot \boldsymbol{n} = \int_e P_{M_0}(\varphi_h - u_h^*) \cdot \boldsymbol{n} && \text{for the face } e, \\ &\int_F \boldsymbol{v} \cdot \boldsymbol{n} = 0 && \forall\, F \in \partial T \setminus e, \end{aligned}$$

we obtain

$$\|P_{M_0}(\varphi_h - u_h^*)\|_e^2 = \sum_{T \in \mathcal{U}_h(e)} (\kappa^{-1} \boldsymbol{q}_h + \nabla u_h^*, \boldsymbol{v})_T.$$

Then, from the Cauchy–Schwarz inequality and a standard scaling argument $\|\boldsymbol{v}\|_T \lesssim h_e^{1/2}\,\|\boldsymbol{v}\cdot\boldsymbol{n}\|_e$, we get

$$h_e^{-1}\,\|P_{M_0}(\varphi_h - u_h^*)\|_e^2 \lesssim \sum_{T\in\mathcal{U}_h(e)} \|\kappa^{-1/2}\boldsymbol{q}_h + \kappa^{1/2}\,\nabla u_h^*\|_T^2. \tag{2.61}$$

Now, analogously to [25, Lema 3.5], it follows that for each boundary face $e \in \Gamma_h$,

$$\begin{aligned} h_e^{-1}\,\|(Id_M - P_{M_0})(\varphi_h - u_h^*)\|_e^2 &= h_e^{-1}\,\|(Id_M - P_{M_0})(\varphi_h - u_h^*)\|_e^2 \\ &\lesssim h_e^{-1}\,\|(Id_M - P_{M_0})(u - u_h^*)\|_e^2 + h_e^{-1}\,\|(Id_M - P_{M_0})(\varphi - \varphi_h)\|_e^2 \\ &\lesssim \sum_{T\in\mathcal{U}_h(e)} \|\nabla(u - u_h^*)\|_T^2 + h_e^{-1}\,\|\varphi - \varphi_h\|_e^2. \end{aligned}$$

Therefore, by (2.59) it follows that

$$\begin{aligned} h_e^{-1}\,\|(Id_M - P_{M_0})(\varphi_h - u_h^*)\|_e^2 \;\lesssim\; & \sum_{T\in\mathcal{U}_h(e)} (\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|\kappa^{-1/2}\,\boldsymbol{q}_h + \kappa^{1/2}\,\nabla u_h^*\|_T^2) \\ & + h_e^{-1}\,\|\varphi - \varphi_h\|_e^2. \end{aligned} \tag{2.62}$$

Finally, we decompose

$$\varphi_h - u_h^* = P_{M_0}(\varphi_h - u_h^*) + (Id_M - P_{M_0})(\varphi_h - u_h^*),$$

and apply (2.61) and (2.62), to arrive at

$$\begin{aligned} h_e^{-1}\,\|\varphi_h - u_h^*\|_e &\lesssim \sum_{T\in\mathcal{U}_h(e)} (\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|\kappa^{-1/2}\,\boldsymbol{q}_h + \kappa^{1/2}\,\nabla u_h^*\|_T^2) + h_e^{-1}\,\|\varphi - \varphi_h\|_e^2 \\ &\lesssim \sum_{T\in\mathcal{U}_h(e)} \|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T^2 + \|h_e^{-1/2}\,(\varphi - \varphi_h)\|_e^2, \end{aligned} \tag{2.63}$$

where we have applied (2.56) in the second line.

Bound for $h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2$. For the interior faces, we have that for any $w \in H_0^1(\mathcal{U}_h(e))$, then

$$\begin{aligned} \langle [\![\boldsymbol{q}_h]\!], w\rangle_e &= \sum_{T\in\mathcal{U}_h(e)} \langle(\boldsymbol{q} - \boldsymbol{q}_h)\cdot\boldsymbol{n}, w\rangle_{\partial T} = \sum_{T\in\mathcal{U}_h(e)} \Big((\boldsymbol{q} - \boldsymbol{q}_h), \nabla w)_T + (\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h, w)_T\Big) \\ &\le \sum_{T\in\mathcal{U}_h(e)} \Big(\overline{\kappa}^{1/2}\,\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T\,\|\nabla w\|_T + h_T\,\|\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h\|_T\,h_T^{-1}\,\|w\|_T\Big) \\ &\le \sum_{T\in\mathcal{U}_h(e)} \Big(\overline{\kappa}^{1/2}\,\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T + h_T\,\|\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h\|_T\Big)\Big(\|\nabla w\|_T + h_T^{-1}\,\|w\|_T\Big). \end{aligned}$$

By choosing a test function of the form $w = B_e[\![\boldsymbol{q}_h]\!] \in \mathbb{P}_{k+d}(T)$, whith being B_e is a face bubble function defined in (1.15), it follows that

$$\int_e B_e[\![\boldsymbol{q}_h]\!]^2 \lesssim \sum_{T\in\mathcal{U}_h(e)} \Big(\|\kappa^{-1/2}(\boldsymbol{q} - \boldsymbol{q}_h)\|_T + h_T\,\|\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h\|_T\Big)h_T^{-1}\,h_e^{1/2}\|B_e[\![\boldsymbol{q}_h]\!]\|_e,$$

then, due to $h_T^{-1}\, h_e^{1/2} \lesssim h_e^{-1/2}$ and the fact that

$$\int_e B_e^2 [\![\boldsymbol{q}_h]\!]^2 \lesssim \int_e [\![\boldsymbol{q}_h]\!]^2 \lesssim \int_e B_e [\![\boldsymbol{q}_h]\!]^2,$$

one arrives at

$$h_e \, \|[\![\boldsymbol{q}_h]\!]\|_e^2 \lesssim \sum_{T\in\mathcal{U}_h(e)} \left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2 + h_T^2 \|\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h\|_T^2 \right).$$

Now, using (2.51) and the Lipschitz continuity of $\mathcal{F}$, due to $L\,h<1$ we get

$$\begin{aligned} h_e\|[\![\boldsymbol{q}_h]\!]\|_e^2 \;\lesssim\; & \sum_{T\in\mathcal{U}_h(e)} \left(\|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2 + h_T^2 \|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|_T^2 + \|u-u_h^*\|_T^2 \right) \\ & + \mathrm{osc}^2(\mathcal{F},\mathcal{U}_h(e)). \end{aligned} \tag{2.64}$$

Bound for $h_T^2\|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|_T^2$. For each element $T\in\mathcal{T}_h$ and any function $w\in H_0^1(T)$, we have that

$$\begin{aligned} (P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h, w)_T &= (P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*), w)_T + (\mathcal{F}(u_h^*) - \mathcal{F}(u), w)_T + (\mathcal{F}(u) - \nabla\cdot\boldsymbol{q}_h, w)_T \\ &= (P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*), w)_T + (\mathcal{F}(u_h^*) - \mathcal{F}(u), w)_T - (\boldsymbol{q}-\boldsymbol{q}_h, \nabla w)_T. \end{aligned}$$

We now consider the element bubble function B_T defined in Lemma 1.4 and take $w := B_T v$, with $v := P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h$. Then, the equation above yields

$$\begin{aligned} &(v, B_T v)_T \\ &\quad\lesssim \big(h_T^{-1}\|\kappa^{-1}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T + \|P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*)\|_T + L\|u-u_h^*\|_T\big)\big(h_T\|\nabla(B_T v)\|_T + \|B_T v\|_T\big). \end{aligned}$$

Then, due to (1.15) and the inverse inequality $h_T\,\|\nabla w\|_T + \|w\|_T \lesssim \|w\|_T$, we obtain

$$\|v\|_T^2 \lesssim \left(h_T^{-1}\,\|\kappa^{-1}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_T + \|P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*)\|_T + L\,\|u-u_h^*\|_T \right)\|B_T v\|_T.$$

Since $\|B_T v\|_T \lesssim \|v\|_T$ also by (1.15), we have

$$\|v\|_T \lesssim h_T^{-1}\,\|\kappa^{-1}\,(\boldsymbol{q}-\boldsymbol{q}_h)\|_T + \|P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*)\|_T + L\,\|u-u_h^*\|_T.$$

Equivalently, since $Lh<1$, the estimate above can be rewritten as

$$h_T^2\,\|P_W\mathcal{F}(u_h^*) - \nabla\cdot\boldsymbol{q}_h\|_T \lesssim \|\kappa^{-1/2}(\boldsymbol{q}-\boldsymbol{q}_h)\|_T^2 + h_T^2\,\|P_W\mathcal{F}(u_h^*) - \mathcal{F}(u_h^*)\|_T^2 + \|u-u_h^*\|_T^2. \tag{2.65}$$

■

CHAPTER 3

Error analysis of an unfitted HDG method for a class of non-linear elliptic problems

In this chapter we study Hibridizable Discontinuous Galerkin (HDG) discretizations for a class of non-linear interior elliptic boundary value problems posed in curved domains where both the source term and the diffusion coefficient are non-linear. We consider the cases where the non-linear diffusion coefficient depends on the solution and on the gradient of the solution. To sidestep the need for curved elements, the discrete solution is computed on a polygonal subdomain that is not assumed to interpolate the true boundary, giving rise to an unfitted computational mesh. We show that, under mild assumptions on the source term and the computational domain, the discrete systems are well posed. Furthermore, we provide *a priori* error estimates showing that the discrete solution will have optimal order of convergence as long as the distance between the curved boundary and the computational boundary remains of the same order of magnitude as the mesh parameter.

3.1 Introduction

In this work we will study a discretization based on the hybridizable discontinous Galerkin (HDG) method [15] for a class of quasilinear elliptic boundary value problems of the form

$$-\nabla\cdot(\kappa\,\nabla u) = f(u) \qquad \text{in } \Omega, \tag{3.1a}$$

$$u = g \qquad \text{on } \Gamma := \partial\Omega, \tag{3.1b}$$

where the domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is not necessarily polygonal/polyhedral, the diffusion coefficient, κ, is a positive function that depends on the solution, u, in one of the following functional forms

$$\kappa = \begin{cases} \kappa(u) \\ \kappa(\nabla u) \end{cases}, \tag{3.1c}$$

and that will be assumed to be a bounded and Lipschitz—in a sense that will be made precise in due time. In addition, the source function f will be taken to be a Lipschitz-continuous mapping from

$L^2(\Omega)$ to $L^2(\Omega)$, so that there exists $L_f > 0$ such that

$$\|f(u_1) - f(u_2)\|_\Omega \leq L_f \, \|u_1 - u_2\|_\Omega \qquad \forall\, u_1, u_2 \in L^2(\Omega). \tag{3.2}$$

In this model, the unknown u is the stream function of the poloidal magnetic field, the source term f is a nonlinear function of u accounting for the effects of the hydrostatic pressure and total electric currents present in the device, and the coefficient κ encodes the magnetic properties of the system. The case where κ is a constant leads to a semi-linear equation for which an HDG discretization was proposed and implemented in [73, 74], and analyzed in detail in [71]. However, in the presence of ferroelectric materials, the permeability is affected by the total magnetic field $\mathbf{B}$—which is proportional to the gradient of u—and the coefficient then takes the form $\kappa = \kappa(\nabla u)$, leading to a quasi-linear equation that requires the more detailed treatment that will be the subject of this article. Some theoretical studies of the HDG method applied to quasilinear problems have been pursued recently [27, 37, 38], however these efforts are limited to polygonal domains. Moreover, the first reference does not consider non-linearities of the form $\kappa(\nabla u)$, while in [37, 38] the authors analyzed an augmented HDG discretization for a strictly quasi-linear problem arising from a non-linear Stokes flow using an approach based on a nonlinear version of the Babŭska–Brezzi theory. As we will show, our analysis will be valid for both quasi-linear and semi-linear problems, will not require an augmented formulation and the domain may be piecewise smooth.

Having established the basic setting in the Chapter 1, we then proceed to study separately the HDG discretizations for the case when the diffusion coefficient is a function of u only (Section 3.2), and the case where the diffusion coefficient depends on ∇u (Section 3.3). In these sections the well posedness of the corresponding discrete HDG formulations are established, and *a priori* error analyses on the discretizations are performed.

We will also need to make two technical assumptions relating the proximity constant to the and the diffusion coefficient κ and the degree of the polynomial approximation. For each $e \in \mathcal{E}_h^\partial$ we will require the following to hold:

$$H_e^\perp \leq \frac{1}{3}\, \underline{\kappa}\, \overline{\tau}^{\,-1}, \tag{3.3a}$$

$$\overline{\kappa}\, \underline{\kappa}^{-1}\, r_e^3\, (C_{ext}^e\, C_{inv}^e)^2 \leq 1, \tag{3.3b}$$

where $\underline{\kappa}$ and $\overline{\kappa}$ are the lower and upper bounds of κ, resp., specified in (3.6), whereas $\overline{\tau}$ is the maximum of the stabilization parameter τ of the HDG scheme .

The first of these two conditions, (3.3a), states the well known fact that for small values of the diffusivity, small scale behavior can be expected near the physical boundary, and therefore fine extension patches are required. However, it also provides the additional insight that the distance between the boundaries can be increased at the cost of accepting smaller values of the stabilization factor τ—and hence larger discontinuities in the discrete solution. In a similar vein, (3.3b) relates the range of values of the diffusion coefficient with the maximum separation between the computational and physical boundaries, thus making sure that the external patches are fine enough to resolve possible boundary behavior induced by large variations in diffusivity over the domain. Moreover, it sets a hard upper limit to the mechanism that allows for a larger separation by decreasing τ.

By combining (1.5) with (3.3) it is not hard to show that, for $k > 0$, $H_e^\perp$ must be bounded as

$$H_e^\perp \leq \min\left\{ \frac{h_e^\perp}{(C_1 C_2)^2} \left(\frac{\overline{\kappa}^{-1}\underline{\kappa}}{k^4(k+1)^4(3\beta+1)^{2k}} \right)^{1/3}, \frac{1}{3}\frac{\underline{\kappa}}{\overline{\tau}^{-1}} \right\}.$$

This expression provides insight into the way in which the physics of the problem—through the range of values for κ—interacts with the discretization—through the parameters $H_e^\perp$, $h_e^\perp$, k, β, and τ—and determines the maximum separation between the physical boundary and that of an admissible triangulation. Of particular note is the role played by the polynomial degree of the approximation: for larger values of k the distance between the mesh and the boundary must decrease. The reason for this will become apparent soon, as we will resort to extrapolation to approximate some quantities over the extension patches.

Recasting (3.1) in mixed form and restricting the resulting equivalent first order system to Ω_h leads to

$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega_h, \tag{3.4a}$$

$$\nabla \cdot \boldsymbol{q} = f(u) \qquad \text{in } \Omega_h, \tag{3.4b}$$

$$u = \varphi \qquad \text{on } \Gamma_h := \partial\Omega_h, \tag{3.4c}$$

where the specific relation between κ, u and ∇u has not been made explicit, and the—a priori unknown—function φ encodes the restriction of u to the computational boundary Γ_h. We can recover φ following the method proposed by [17] (in one dimension) and extended to higher dimensions by [21]. The idea consists of transferring the Dirichlet data g from Γ to Γ_h along segments called *transfer paths* by computing a line integral of the flux $\boldsymbol{q}$.

To be precise, given $\boldsymbol{x} \in \Gamma_h$ and $\overline{\boldsymbol{x}} \in \Gamma$, equation (3.4a) can be integrated along the segment connecting them. Let us denote by $\boldsymbol{t}(\boldsymbol{x})$ the unit vector anchored at $\boldsymbol{x}$ pointing towards $\overline{\boldsymbol{x}}$, and by $l(\boldsymbol{x})$ the length of the segment connecting them. We then have the following representation for φ:

$$\varphi := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} (\kappa^{-1}\,\boldsymbol{q})(\boldsymbol{x} + \boldsymbol{t}(\boldsymbol{x})s) \cdot \boldsymbol{t}(\boldsymbol{x}) ds. \tag{3.5}$$

Note that φ depends on the values of either u or ∇u (through κ^{-1}), and $\boldsymbol{q}$ over the extended domain Ω_h^c. As such, we should write $\varphi = \varphi(u, \nabla u, \boldsymbol{q}, \boldsymbol{x})$ however, to keep notation simple, we will abstain from this and will write simply φ. In a similar fashion, $g(\overline{\boldsymbol{x}})$ is in fact a function of $\boldsymbol{x}$, since the point $\overline{\boldsymbol{x}}$ varies smoothly with $\boldsymbol{x}$. To avoid the use of cumbersome notation we will write either $g(\overline{\boldsymbol{x}})$ or simply $\overline{\mathrm{g}} := g(\overline{\boldsymbol{x}}(\boldsymbol{x}))$.

3.2 Non-linearities of the form $\kappa(u)$

3.2.1 The HDG formulation

We will first consider the case when the coefficient κ depends on the solution in the form

$$\begin{aligned} \kappa : L^2(\Omega) &\longrightarrow L^\infty(\overline{\Omega}) \\ u \qquad &\longmapsto \quad \kappa(u). \end{aligned}$$

For the analysis, we will require the existence of positive constants $\underline{\kappa}$ and $\overline{\kappa}$ such that for all $u \in L^2(\Omega)$

$$\underline{\kappa} \leq \kappa(u) \leq \overline{\kappa} \qquad \text{almost everywhere in } \Omega. \tag{3.6}$$

Moreover, κ will be assumed to be Lipschitz-continuous on $L^2(\Omega)$, i.e, there exists $\tilde{L} > 0$ such that

$$\|\kappa(u_1) - \kappa(u_2)\|_{L^\infty(\Omega)} \leq \tilde{L}\|u_1 - u_2\|_{L^2(\Omega)} \qquad \forall\, u_1, u_2 \in L^2(\Omega). \tag{3.7}$$

The two conditions above, together, imply the existence of constants $\widehat{L}$ and L such that

$$\|\kappa^{1/2}(u_1) - \kappa^{1/2}(u_2)\|_{L^\infty(\Gamma)} \leq \widehat{L}\|u_1 - u_2\|_{L^2(\Omega)} \qquad \forall\, u_1, u_2 \in L^2(\Omega), \tag{3.8}$$

$$\|\kappa^{-1}(u_1) - \kappa^{-1}(u_2)\|_{L^\infty(\overline{\Omega})} \leq L\|u_1 - u_2\|_{L^2(\Omega)} \qquad \forall\, u_1, u_2 \in L^2(\Omega). \tag{3.9}$$

Note that all these assumptions imply the Lipschitz continuity of $\kappa, \kappa^{1/2}$, and κ^{-1} on the subdomain $\Omega_h \subset \Omega$ with corresponding Lipschitz constants equal to or smaller than those stated above.

Before introducing the discrete formulation we will recall here the mixed form (3.4), but now we make explicit the dependence $\kappa = \kappa(u)$

$$\boldsymbol{q} + \kappa(u)\,\nabla u = 0 \qquad \text{in } \Omega_h, \tag{3.10a}$$

$$\nabla \cdot \boldsymbol{q} = f(u) \qquad \text{in } \Omega_h, \tag{3.10b}$$

$$u = \varphi \qquad \text{on } \partial\Omega_h. \tag{3.10c}$$

The boundary data φ on the computational boundary Γ_h is transferred according to (3.5).

Taking an admissible triangulation $\mathcal{T}_h$ of the computational domain Ω_h, the HDG discretization of (3.10) reads: Find $(\boldsymbol{q}_h, u_h, \widehat{u}_h) \in \boldsymbol{V}_h \times W_h \times M_h$, such that

$$(\kappa^{-1}(u_h)\,\boldsymbol{q}_h, \boldsymbol{v})_{\mathcal{T}_h} - (u_h, \nabla \cdot \boldsymbol{v})_{\mathcal{T}_h} + \langle \widehat{u}_h, \boldsymbol{v} \cdot \boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.11a}$$

$$-(\boldsymbol{q}_h, \nabla w)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} = (f(u_h), w)_{\mathcal{T}_h}, \tag{3.11b}$$

$$\langle \widehat{u}_h, \mu\rangle_{\Gamma_h} = \langle \varphi_h(u_h), \mu\rangle_{\Gamma_h}, \tag{3.11c}$$

$$\langle \widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n}, \mu\rangle_{\partial\mathcal{T}_h \setminus \Gamma_h} = 0, \tag{3.11d}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$. Here

$$\widehat{\boldsymbol{q}}_h \cdot \boldsymbol{n} := \boldsymbol{q}_h \cdot \boldsymbol{n} + \tau(u_h - \widehat{u}_h) \quad \text{on} \quad \partial\mathcal{T}_h,$$

with τ being a positive stabilization function, whose maximum will be denoted by $\overline{\tau}$. The approximate boundary condition φ_h on the right hand side of (3.11c) is given by the discrete counterpart of (3.5)

$$\varphi_h(u_h)(\boldsymbol{x}) := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} (\kappa^{-1}(u_h)\, E_h \boldsymbol{q}_h)(\boldsymbol{x} + \boldsymbol{t}(\boldsymbol{x})s) \cdot \boldsymbol{t}(\boldsymbol{x}))ds, \quad \text{for} \quad \boldsymbol{x} \in \Gamma_h. \tag{3.11e}$$

In the definition above, we have used the extrapolation $E_h \boldsymbol{q}_h$ due to the fact that the approximation $\boldsymbol{q}_h$ is available only inside of the computational domain Ω_h, but the transfer paths along which the integral is computed are defined over the complementary extended region Ω_h^c.

3.2.2 Well-posedness

In this section we employ a Banach fixed-point argument to ensure the well-posedness of the discrete problem (3.11). To that end we will define an operator $\mathcal{J} : W_h \to W_h$ mapping ζ to the second component of the triplet $(\boldsymbol{q}, u, \widehat{u}) \in \boldsymbol{V}_h \times W_h \times M_h$ satisfying, for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$, the HDG system (3.11) where the source has been evaluated at ζ, namely

$$(\kappa^{-1}(\zeta)\, \boldsymbol{q}, \boldsymbol{v})_{\mathcal{T}_h} - (u, \nabla \cdot \boldsymbol{v})_{\mathcal{T}_h} + \langle \widehat{u}, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} = 0, \tag{3.12a}$$

$$-(\boldsymbol{q}, \nabla w)_{\mathcal{T}_h} + \langle \widehat{\boldsymbol{q}} \cdot \boldsymbol{n}, w \rangle_{\partial \mathcal{T}_h} = (f(\zeta), w)_{\mathcal{T}_h}, \tag{3.12b}$$

$$\langle \widehat{u}, \mu \rangle_{\Gamma_h} = \langle \varphi(\zeta), \mu \rangle_{\Gamma_h}, \tag{3.12c}$$

$$\langle \widehat{\boldsymbol{q}} \cdot \boldsymbol{n}, \mu \rangle_{\partial \mathcal{T}_h \setminus \Gamma_h} = 0. \tag{3.12d}$$

Above, the term $\varphi(\zeta)$ corresponds to the boundary condition transferred to the computational domain by means of (3.11e). The mapping $\mathcal{J}$ is well defined, as the linearized system (3.12) is uniquely solvable as proven in [18].

The main result of this section—that the mapping $\mathcal{J}$ defined above is a contraction—relies on the validity of a particular inequality—estimate (3.13) below—but is otherwise a simple argument. Since the proof of (3.13) requires a sequence of technical arguments, in the interest of clarity (we will first prove the main theorem assuming that the aforementioned inequality is valid. After having established the well posedness of the discrete problem, the reminder of the section will be devoted to verifying the validity of (3.13). This will be finally established in Lemma 3.3, after a series of auxiliary results.

Theorem 3.1 (Well-posedness of the discrete problem)**.** *Suppose that Assumptions* (3.3) *are satisfied and that additionally*

$$(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2}\, R_h^{1/2})\, \|l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}\, \widehat{L}\, h^{1/2} < 1/4,$$
$$\max\{\widehat{c}^2 h, 1\}\, L_f < 1/8,$$

where $\widehat{c}$ is the constant given in Lemma 3.3. Then $\mathcal{J}$ is a contraction operator.

Proof. Let $\zeta_1, \zeta_2 \in W_h$ and define $u_1 := \mathcal{J}(\zeta_1)$ and $u_2 := \mathcal{J}(\zeta_2)$. Then u_1 and u_2 are the second

components of solutions to (3.12) and Lemma 3.3 guarantees that

$$\begin{aligned}\|\mathcal{J}(\zeta_1)-\mathcal{J}(\zeta_2)\|_{\Omega_h} &= \|u_1-u_2\|_{\Omega_h}\\ &\leq 4\,\max\{\widehat{c}^2 h,1\}\,\|f(\zeta_1)-f(\zeta_2)\|_{\Omega_h}+2\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R_h {R_h}^{1/2})\,h^{1/2}\,\|(\kappa^{1/2}(\zeta_1)-\kappa^{1/2}(\zeta_2))\,l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}.\end{aligned} \tag{3.13}$$

Then, applying the Lipschitz-continuity of f and $\kappa^{1/2}$—given respectively in (3.2) and (3.7)—we get

$$\begin{aligned}&\leq 4\,\max\{\widehat{c}^2 h,1\}\,L_f\,\|\zeta_1-\zeta_2\|_{\Omega_h}+2\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R^{1/2})\,h^{1/2}\,\|\kappa^{1/2}(\zeta_1)-\kappa^{1/2}(\zeta_2)\|_{L^\infty(\Gamma_h)}\,\|l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}\\ &\leq 4\,\max\{\widehat{c}^2 h,1\}\,L_f\,\|\zeta_1-\zeta_2\|_{\Omega_h}+2\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R_h^{1/2})\,\widehat{L}\,h^{1/2}\,\|\zeta_1-\zeta_2\|_{\Omega_h}\,\|l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}.\end{aligned}$$

The result follows from the hypothesis for $\widehat{L}$ and L_f . ■

Combined with a standard fixed-point argument, the theorem above guarantees the existence and uniqueness of the solution to the discrete HDG system. We will now direct our efforts to showing that inequality (3.13) holds. To that end we will make use of the following auxiliary function and its properties listed in the lemma below—the proof of which can be found on [18, Lemma 5.2].

For the subsequent analysis, we will also make use of the following norm. Given $(\boldsymbol{v},w,\mu)\in \boldsymbol{V}_h\times W_h\times M_h$ and $\xi\in M_h$ we define

$$|\!|\!|(\boldsymbol{v},w,\mu)|\!|\!|_\xi := \Big(\|\kappa^{-1/2}(\xi)\boldsymbol{v}\|^2_{\Omega_h}+\|\tau^{1/2}\,w\|^2_{\partial\mathcal{T}_h}+\|\kappa^{1/2}(\xi)\,l^{-1/2}\,\mu(\xi)\|^2_{\Gamma_h}\Big)^{1/2}. \tag{3.14}$$

The proof of (3.13) requires considering the solution ϕ to an auxiliary problem (c.f. (1.5)) and using a duality argument that connects a stability estimate for ϕ with our variables of interest. This will be done in Lemma 3.3. Lemmas 3.1 and 3.2 below establish estimates relating the norm (3.14) of the solution $(\boldsymbol{q},u,\widehat{u})$ to problem data that will be used in the final step of Lemma 3.3.

Lemma 3.1. *Let φ be the transferred boundary condition appearing in (3.12) and suppose that assumptions (3.3) are satisfied. It holds*

$$\langle\varphi(\zeta),\delta_{\boldsymbol{q}}\rangle_{\Gamma_h}\leq\frac{1}{6}\,\|\kappa^{1/2}(\zeta)\,l^{-1/2}\,\varphi(\zeta)\|^2_{\Gamma_h}+\frac{1}{2}\,\|\kappa^{-1/2}(\zeta)\,\boldsymbol{q}\|^2_{\Omega_h},$$

$$\langle\varphi(\zeta),\tau(u-\widehat{u})\rangle_{\Gamma_h}\leq\frac{1}{6}\|\kappa^{1/2}(\zeta)\,l^{-1/2}\,\varphi(\zeta)\|^2_{\Gamma_h}+\frac{1}{2}\|\tau^{1/2}(u-\widehat{u})\|^2_{\partial\mathcal{T}_h},$$

$$\langle\varphi(\zeta),\kappa(\zeta)\,l^{-1}\,\overline{\mathrm{g}}\rangle_{\Gamma_h}\leq\frac{1}{6}\|\kappa^{1/2}(\zeta)\,l^{-1/2}\,\varphi(\zeta)\|^2_{\Gamma_h}+\frac{3}{2}\|\kappa^{1/2}(\zeta)\,l^{-1/2}\,\overline{\mathrm{g}}\|^2_{\Gamma_h},$$

where $\overline{\mathrm{g}}(\boldsymbol{x})=g(\overline{\boldsymbol{x}}(\boldsymbol{x}))\,\forall\,\boldsymbol{x}\in\Gamma_h$.

Proof. The first inequality is obtained after applying Young's inequality and estimate (1.12a), whereas the second inequality follows from assumption (1.2) and (3.3a). The third inequality follows from Young's inequality exclusively. ■

Lemma 3.2. *If Assumptions (3.3) hold, then*

$$|||(\boldsymbol{q}, u-\widehat{u}, \varphi)|||_{\zeta}^2 \leq 2\|f(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + 3\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}^2.$$

Proof. Let $\zeta \in W_h$ and $u = \mathcal{J}(\zeta) \in W_h$. Since u defined this way is the solution to the discrete system (3.12), then testing (3.12) with

$$\boldsymbol{v} = \boldsymbol{q}, \quad w = u, \quad \mu := \begin{cases} -\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} & \text{on } \Gamma_h \\ -\widehat{u} & \text{on } \partial\mathcal{T}_h \setminus \Gamma_h \end{cases},$$

we deduce that

$$\|\kappa^{-1/2}(\zeta)\,\boldsymbol{q}\|_{\Omega_h}^2 + \|\tau^{1/2}\,(u-\widehat{u})\|_{\partial\mathcal{T}_h}^2 = -\,\langle\varphi(\zeta), \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}\rangle_{\Gamma_h} + (f(\zeta), u)_{\mathcal{T}_h}. \tag{3.15}$$

On the other hand, we can use the definition of φ and $\delta_{\boldsymbol{q}}$ (cf. (3.11e) and (1.11)) to show that

$$\widehat{\boldsymbol{q}}\cdot\boldsymbol{n} = \kappa(\zeta)\, l^{-1}(\varphi(\zeta) - \overline{\mathrm{g}}) - \delta_{\boldsymbol{q}} + \tau(u-\widehat{u}).$$

Substituting the expression for $\widehat{\boldsymbol{q}}\cdot\boldsymbol{n}$ above in (3.15) , we obtain

$$\begin{aligned} \|\kappa^{-1/2}(\zeta)\,\boldsymbol{q}\|_{\Omega_h}^2 \quad &+ \|\tau^{1/2}\,(u-\widehat{u})\|_{\partial\mathcal{T}_h}^2 + \|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\varphi(\zeta)\|_{\Gamma_h}^2 \\ &= |||(\boldsymbol{q}, u-\widehat{u}, \varphi)|||_{\zeta}^2 \\ &\leq |\langle\varphi(\zeta), \kappa(\zeta)\, l^{-1}\,\overline{\mathrm{g}}\rangle_{\Gamma_h}| + |\langle\varphi(\zeta), \delta_{\boldsymbol{q}}\rangle_{\Gamma_h}| + |\langle\varphi(\zeta), \tau(u-\widehat{u})\rangle_{\Gamma_h}| + |(f(\zeta), u)_{\mathcal{T}_h}|. \end{aligned}$$

Now, using Lemma 3.1 to estimate the first three terms in the right hand of this expression we obtain

$$\frac{1}{2}\,|||(\boldsymbol{q}, u-\widehat{u}, \varphi)|||_{\zeta}^2 \leq \|f(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + \frac{3}{2}\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}^2,$$

whereupon the proof is concluded. ∎

For the following result, we will make use of the properties of the HDG projectors $\boldsymbol{\Pi}_{\boldsymbol{V}}$ and Π_W onto the discrete spaces $\boldsymbol{V}_h$ and W_h. This projection was first introduced in [16] and we include its definition and main properties in the Section 1.6. The L^2 projector onto the space M_h will be denoted by P_M, while Id_M will denote the identity on M_h.

Lemma 3.3. *Suppose that Assumptions (3.3) and the regularity (1.8) are satisfied. Then, there exists $\widehat{c} > 0$, independent of h such that*

$$\|u\|_{\Omega_h} \leq 4\,\max\{\widehat{c}^2 h, 1\}\,\|f(\zeta)\|_{\Omega_h} + 2\,\left(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2}\,R_h^{1/2}\right)\,h^{1/2}\,\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}. \tag{3.16}$$

Proof. Consider $\Theta \in L^2(\Omega)$ and let ϕ and ψ be the solutions to the dual problem (1.5) associated to Θ. If we define

$$\mathbb{T}_{\boldsymbol{q}} := (\kappa^{-1}(\zeta)\boldsymbol{q}, \boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{\phi} - \boldsymbol{\phi})_{\mathcal{T}_h}, \quad \mathbb{T}_u := \langle\widehat{u}, P_M(\boldsymbol{\phi}\cdot\boldsymbol{n})\rangle_{\Gamma_h} - \langle\widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, \Pi_W\psi\rangle_{\Gamma_h} \quad \text{and} \quad \mathbb{T}_f := (f(\zeta), \Pi_W\psi)_{\mathcal{T}_h},$$

it is possible to verify that

$$(u,\Theta)_{\mathcal{T}_h} = \mathbb{T}_{\boldsymbol{q}} + \mathbb{T}_f + \mathbb{T}_u. \tag{3.17}$$

The terms $\mathbb{T}_{\boldsymbol{q}}$ and $\mathbb{T}_f$ appearing on the expression above can be easily estimated by

$$|\mathbb{T}_{\boldsymbol{q}}| \lesssim \underline{\kappa}^{-1/2}\, h \|\kappa^{-1/2}(\zeta)\,\boldsymbol{q}\|_{\Omega_h}\,\|\Theta\|_{\Omega}, \quad \text{and} \quad |\mathbb{T}_f| \lesssim \|f(\zeta)\|_{\Omega_h}\,\|\Theta\|_{\Omega}. \tag{3.18}$$

In order to bound the final term of the decomposition of $(u,\Theta)_{\mathcal{T}_h}$, we rewrite $\mathbb{T}_u = \sum_{i=1}^{5}\mathbb{T}_u^i$ where

$$\begin{aligned}
\mathbb{T}_u^1 &:= -\langle \kappa(\zeta) l^{-1}\varphi(\zeta), \psi + l\partial_n\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^4 &:= -\langle \tau(u-\widehat{u}), P_M\psi\rangle_{\Gamma_h},\\
\mathbb{T}_u^2 &:= \langle \kappa(\zeta)\varphi(\zeta), (P_M - Id_M)\partial_n\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^5 &:= \langle \kappa(\zeta)\, l^{-1}\,\overline{\mathrm{g}}, \psi\rangle_{\Gamma_h},\\
\mathbb{T}_u^3 &:= \langle \delta_{\boldsymbol{q}}, \psi\rangle_{\Gamma_h}.
\end{aligned}$$

It is not hard, if cumbersome, to verify that for the terms above the following estimates hold

$$\begin{aligned}
|\mathbb{T}_u^1| &\lesssim \overline{\kappa}^{1/2}\, R_h\, h\, \|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\varphi(\zeta)\|_{\Gamma_h}\|\Theta\|_{\Omega}, & |\mathbb{T}_u^2| &\lesssim \overline{\kappa}^{1/2}\, R_h^{1/2}\, h\, \|\kappa^{1/2}(\zeta) l^{-1/2}\varphi(\zeta)\|_{\Gamma_h}\|\Theta\|_{\Omega},\\
|\mathbb{T}_u^3| &\lesssim \overline{\kappa}^{1/2}\, R_h^2\, h^{1/2}\|\kappa^{-1/2}(\zeta)\boldsymbol{q}\|_{\Omega_h}\|\Theta\|_{\Omega}, & |\mathbb{T}_u^4| &\lesssim \overline{\tau}^{1/2}\, R_h\, h\, \|\tau^{1/2}(u-\widehat{u})\|_{\partial\mathcal{T}_h}\|\Theta\|_{\Omega},\\
|\mathbb{T}_u^5| &\lesssim \overline{\kappa}^{1/2}\,(R_h h)^{1/2}\,\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}\|\Theta\|_{\Omega}.
\end{aligned}$$

Taking $\Theta = u$ in (3.17) and combining all of the above estimates with (3.18), we obtain

$$\|u\|_{\Omega_h} \le \widehat{c}\, h^{1/2}|\!|\!|(\boldsymbol{\sigma}, u-\widehat{u}, \varphi)|\!|\!|_{\zeta} + \overline{\kappa}^{1/2}\,(R_h h)^{1/2}\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h} + \|f(\zeta)\|_{\Omega_h},$$

where $\widehat{c} := C\,\max\{\underline{\kappa}^{-1/2}, \overline{\kappa}^{1/2}R_h, \overline{\kappa}^{1/2}R_h^2, \overline{\kappa}^{1/2}R_h^{1/2}, \overline{\tau}^{1/2}R_h\}$, and $C>0$ is the constant hidden in the symbol $\lesssim$. Then, applying Lemma 3.2, we get

$$\begin{aligned}
\|u\|_{\Omega_h} \le\ & \widehat{c}\, h^{1/2}\left(\sqrt{2}\|f(\zeta)\|_{\Omega_h}^{1/2}\|u\|_{\Omega_h}^{1/2} + \sqrt{3}\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h}\right)\\
&+ \overline{\kappa}^{1/2}\,(R_h h)^{1/2}\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h} + \|f(\zeta)\|_{\Omega_h}.
\end{aligned}$$

$$\|u\|_{\Omega_h} \le 4\,\max\{\widehat{c}^2 h, 1\}\,\|f(\zeta)\|_{\Omega_h} + 2\,(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2}\,R_h^{1/2})\,h^{1/2}\,\|\kappa^{1/2}(\zeta)\, l^{-1/2}\,\overline{\mathrm{g}}\|_{\Gamma_h},$$

with which the proof is concluded. ■

3.2.3 *A prior* error analysis

We now provide the *a priori* error bounds for the discretization error. The main results of the section are Theorem 3.19 and Corollary 3.1 immediately after it. As we will see, several of the results leading to the main presented in this section can be proven by using similar arguments to those of Section 3.2.2 and we will omit some of the arguments. The analysis will be performed by decomposing the approximation errors in two components using the properties of the HDG projection (see Appendix

1.6). The *projection of the errors* is defined as

$$\boldsymbol{\varepsilon}^{\boldsymbol{q}} := \boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q} - \boldsymbol{q}_h \quad \text{and} \quad \varepsilon^u := \Pi_W u - u_h,$$

and the *error of the projections* are given by

$$\boldsymbol{I}^{\boldsymbol{q}} := \boldsymbol{q} - \boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q} \quad \text{and} \quad I^u := u - \Pi_W u.$$

This allows to express the approximation errors as

$$\boldsymbol{q} - \boldsymbol{q}_h = \boldsymbol{\varepsilon}^{\boldsymbol{q}} + \boldsymbol{I}^{\boldsymbol{q}} \quad \text{and} \quad u - u_h = \varepsilon^u + I^u.$$

In addition, recalling that P_M is the L^2 projection into M_h, we define the projection error for the hybrid unknown $\widehat{u}_h$ as $\varepsilon^{\widehat{u}} := P_M u - \widehat{u}_h$. The L^2-projection of the error for the numerical flux on $\partial\mathcal{T}_h$ can be expressed as $\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} = \boldsymbol{\varepsilon}^{\boldsymbol{q}} \cdot \boldsymbol{n} + \tau(\varepsilon^u - \varepsilon^{\widehat{u}})$. It is not difficult show that $(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u, \varepsilon^{\widehat{u}})$ belongs to $\boldsymbol{V}_h \times W_h \times M_h$ and satisfies

$$\begin{aligned}
(\kappa^{-1}(u_h)\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \boldsymbol{v})_{\mathcal{T}_h} - (\varepsilon^u, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} + \langle \varepsilon^{\widehat{u}}, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} &= -(\kappa^{-1}(u)\boldsymbol{I}^{\boldsymbol{q}}, v)_{\mathcal{T}_h} \\
&\quad - \left((\kappa^{-1}(u) - \kappa^{-1}(u_h))\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q}, v\right)_{\mathcal{T}_h}, && \text{(3.19a)}\\
-(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \nabla w)_{\mathcal{T}_h} + \langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, w\rangle_{\partial\mathcal{T}_h} &= (f(u) - f(u_h), w)_{\mathcal{T}_h}, && \text{(3.19b)}\\
\langle \varepsilon^{\widehat{u}}, \mu\rangle_{\Gamma_h} &= \langle \varphi(u) - \varphi_h(u_h), \mu\rangle_{\Gamma_h}, && \text{(3.19c)}\\
\langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, \mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} &= 0, && \text{(3.19d)}
\end{aligned}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$.

To try and keep the notation compact, we will define the following two quantities involving only the errors in the projections $\boldsymbol{I}^{\boldsymbol{v}}$, $\boldsymbol{I}^{\boldsymbol{q}}$, and I^u measured in the three relevant domains Ω_h, Ω_h^c and Γ_h

$$\begin{aligned}
\Lambda_{\boldsymbol{q}} &:= \left(\|\boldsymbol{I}^{\boldsymbol{q}}\|^2_{\Omega_h} + \|h^{\perp}\partial_n(\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Omega_h^c} + \|(h^{\perp})^{1/2}\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Gamma_h}\right)^{1/2}, && \text{(3.20a)}\\
\Lambda_u &:= \left(\|(h^{\perp})^{1/2} I^u\|_{\Gamma_h} + \|I^u\|_{\Omega_h}\right)^{1/2}. && \text{(3.20b)}
\end{aligned}$$

We note that, as pointed out in [16,71] by using the properties of the projectors and scaling arguments, if $\boldsymbol{q} \in \boldsymbol{H}^{k+1}(\Omega)$, $u \in H^{k+1}(\Omega)$ and τ is of order one, then $\Lambda_{\boldsymbol{q}}$ and Λ_u are of order h^{k+1}. As stated in the theorem below, these quantities are in fact the key to estimating the approximation error of the method.

Theorem 3.2. *If L is small enough, the regularity (1.8) holds and the discrete spaces are of polynomial degree $k \geq 1$, then there exists $h_0 > 0$ such that, for all $h \leq h_0$, we have*

$$|||(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|||^2_{u_h} \lesssim \Lambda^2_{\boldsymbol{q}} + \Lambda^2_u. \tag{3.21}$$

The proof of this result will follow straightforwardly from lemmas 3.4 and 3.5 below.

Before setting out to prove these two lemmas (and therefore the theorem above) we first state the convergence order of the method—the main result of the section—which thanks to the remark made

just above Theorem 3.2 follows as a corollary.

Corollary 3.1 (Order of convergence)**.** *Suppose that the assumptions of Theorem 3.2 hold. If, in addition, $u \in H^{k+1}(\Omega)$ and $\boldsymbol{q} \in \boldsymbol{H}^{k+1}(\Omega)$, then*

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_\Omega + \|u - u_h\|_\Omega \leq C h^{k+1} \left(|u|_{k+1,\Omega} + |\boldsymbol{q}|_{k+1,\Omega}\right).$$

Having stated the main results of the section, we now set out to prove the two lemmas leading to Theorem 3.2. The first part of the analysis will require using an energy argument on the error equations (3.19) and a meticulous study of the error contribution due to the transferred boundary conditions $\varphi_h(u_h)$. This will be done in the following

Lemma 3.4. *There exist positive constants, independent of h, such that*

$$|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|\!|\!|_{u_h}^2 \leq 12 \max\{C_1 h, C_2\} L^2 \left(\|\varepsilon^u\|_{\Omega_h}^2 + \|I^u\|_{\Omega_h}^2\right) + C_3 \Lambda_{\boldsymbol{q}}^2. \tag{3.22}$$

Proof. Starting from (3.19) and letting

$$\boldsymbol{v} = \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \quad \text{and} \quad w = \varepsilon^u \ \text{ in } \mathcal{T}_h, \quad \text{and} \quad \mu := \begin{cases} -\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} & \text{on } \Gamma_h \\ -\varepsilon^{\widehat{u}} & \text{on } \partial\mathcal{T}_h \setminus \Gamma_h \end{cases},$$

it follows that

$$\begin{aligned} &\|\kappa^{-1/2}(u_h)\,\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|_{\Omega_h}^2 + \|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|_{\partial\mathcal{T}_h}^2 = -(\kappa^{-1}(u)\boldsymbol{I}^{\boldsymbol{q}}, \boldsymbol{\varepsilon}^{\boldsymbol{q}})_{\mathcal{T}_h} \\ &-((\kappa^{-1}(u) - \kappa^{-1}(u_h))\Pi_{\boldsymbol{V}}\boldsymbol{q}, \boldsymbol{\varepsilon}^{\boldsymbol{q}})_{\mathcal{T}_h} + (f(u) - f(u_h), \varepsilon^u)_{\mathcal{T}_h} - \langle \varphi(u) - \varphi_h(u_h), \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n}\rangle_{\Gamma_h}. \end{aligned} \tag{3.23}$$

We will now manipulate the final term in the expression above to include a term involving the norm of the difference $\varphi(u) - \varphi_h(u_h)$, thus allowing us to estimate the transfer error. Using definitions of φ_h and φ (cf. (3.5) and (3.11e) respectively), as well as the definition of $\delta_{\boldsymbol{q}}$ it follows that

$$\varphi(u) - \overline{g} = \kappa^{-1}(u)\ell\left(\delta_{\boldsymbol{q}} + \boldsymbol{q}\cdot\boldsymbol{n}\right) \quad \text{and} \quad \varphi_h(u_h) - \overline{g} = \kappa^{-1}(u_h)\ell\left(\delta_{\boldsymbol{q}_h} + \boldsymbol{q}_h\cdot\boldsymbol{n}\right).$$

Subtracting the second expression from the first one and adding zero in the form of $\pm\kappa^{-1}(u_h)\left(\delta_{\boldsymbol{q}} - \boldsymbol{q}\cdot\boldsymbol{n}\right)$ it is possible to express the difference as

$$\begin{aligned} \varphi(u) - \varphi_h(u_h) &= \ell\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\left(\delta_{\boldsymbol{q}} + \boldsymbol{q}\cdot\boldsymbol{n}\right) + \ell\,\kappa^{-1}(u_h)\left(\delta_{\boldsymbol{q}-\boldsymbol{q}_h} + (\boldsymbol{q} - \boldsymbol{q}_h)\cdot\boldsymbol{n}\right) \\ &= \ell\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\left(\delta_{\boldsymbol{q}} + \boldsymbol{q}\cdot\boldsymbol{n}\right) + \ell\,\kappa^{-1}(u_h)\left(\delta_{\boldsymbol{\varepsilon}^{\boldsymbol{q}}} + \delta_{\boldsymbol{I}^{\boldsymbol{q}}} + \left(\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} + \boldsymbol{I}^{\boldsymbol{q}}\right)\cdot\boldsymbol{n}\right) \\ &\quad - \ell\,\kappa^{-1}(u_h)\left(\tau\left(\varepsilon^u - \varepsilon^{\widehat{u}}\right)\right) \\ &= \ell\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\left(\delta_{\boldsymbol{I}^{\boldsymbol{q}}} + \delta_{\Pi_v\boldsymbol{q}} + \left(\boldsymbol{I}^{\boldsymbol{q}} + \Pi_{\boldsymbol{V}}\boldsymbol{q}\right)\cdot\boldsymbol{n}\right) \\ &\quad + \ell\kappa^{-1}(u_h)\left(\delta_{\boldsymbol{\varepsilon}^{\boldsymbol{q}}} + \delta_{\boldsymbol{I}^{\boldsymbol{q}}} + \left(\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} + \boldsymbol{I}^{\boldsymbol{q}}\right)\cdot\boldsymbol{n} - \tau\left(\varepsilon^u - \varepsilon^{\widehat{u}}\right)\right), \end{aligned} \tag{3.24}$$

where the first equality comes from the substitutions

$$(\boldsymbol{q} - \boldsymbol{q}_h)\cdot\boldsymbol{n} = \left(\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} + \boldsymbol{I}^{\boldsymbol{q}}\right)\cdot\boldsymbol{n} - \tau\left(\varepsilon^u - \varepsilon^{\widehat{u}}\right), \quad \text{and} \quad \delta_{\boldsymbol{q}+\boldsymbol{q}_h} = \delta_{\boldsymbol{q}} + \delta_{\boldsymbol{q}_h},$$

while the second one is obtained by replacing $\boldsymbol{q} = \boldsymbol{I^q} + \Pi_{\boldsymbol{V}}\boldsymbol{q}$. The expression (3.24) allows us to write the term $\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n}$ in terms of the transfer error

$$\begin{aligned}\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} = \kappa(u_h)\, l^{-1}\left(\varphi(u) - \varphi_h(u_h)\right) - \kappa(u_h)\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)(\delta_{\boldsymbol{I^q}} + \delta_{\Pi_{\boldsymbol{V}}\boldsymbol{q}} + \boldsymbol{I^q}\cdot\boldsymbol{n} + \Pi_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n}) \\ - \delta_{\boldsymbol{\varepsilon^q}} - \delta_{\boldsymbol{I}^q} - \boldsymbol{I^q}\cdot\boldsymbol{n} + \tau(\varepsilon^u - \varepsilon^{\widehat{u}}).\end{aligned}$$

Substituting this expression back into (3.23) and rearranging terms, it follows that

$$\begin{aligned}\|\kappa^{-1/2}(u_h)\,\boldsymbol{\varepsilon^q}\|^2_{\Omega_h} + \quad \|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial\mathcal{T}_h} + \|\kappa^{1/2}(u_h)\, l^{-1/2}\left(\varphi(u) - \varphi_h(u_h)\right)\|^2_{\Gamma_h} \\ \leq |(\kappa^{-1}(u)\boldsymbol{I^q}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h}| + |((\kappa^{-1}(u) - \kappa^{-1}(u_h))\Pi_{\boldsymbol{V}}\boldsymbol{q}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h}| + |\mathbb{T}^f| + |\mathbb{T}_\varphi|,\end{aligned} \tag{3.25}$$

with $\mathbb{T}^f := (f(u) - f(u_h), \varepsilon^u)_{\mathcal{T}_h}$ and $\mathbb{T}_\varphi := \sum_{i=1}^8 |\mathrm{T}^i_\varphi|$, where

$$\begin{aligned}
\mathbb{T}^1_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \delta_{\boldsymbol{\varepsilon^q}}\rangle_{\Gamma_h} & \mathbb{T}^5_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \kappa(u_h)\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\delta_{\Pi_{\boldsymbol{V}}\boldsymbol{q}}\rangle_{\Gamma_h} \\
\mathbb{T}^2_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \tau(\varepsilon^{\widehat{u}} - \varepsilon^u)\rangle_{\Gamma_h} & \mathbb{T}^6_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \kappa(u_h)\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\delta_{\boldsymbol{I^q}}\rangle_{\Gamma_h} \\
\mathbb{T}^3_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \boldsymbol{I^q}\cdot\boldsymbol{n}\rangle_{\Gamma_h} & \mathbb{T}^7_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \kappa(u_h)\left(\kappa^{-1}(u) - \kappa^{-1}(u_h)\right)\boldsymbol{I^q}\cdot\boldsymbol{n}\rangle_{\Gamma_h} \\
\mathbb{T}^4_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \delta_{\boldsymbol{I^q}}\rangle_{\Gamma_h} & \mathbb{T}^8_\varphi &:= \langle \varphi(u) - \varphi_h(u_h), \kappa(u_h)(\kappa^{-1}(u) - \kappa^{-1}(u_h))\Pi_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n}\rangle_{\Gamma_h}.
\end{aligned}$$

To determine upper bounds the terms in the right hand side of (3.25), we will make use of Young's inequality, the Lipschitz continuity of f and κ^{-1} and the fact that $\|v\|_{L^2(e)} \lesssim h_e^{1/2}\|v\|_e$ for all $e \in \mathcal{E}_h^\partial$ and for each $v \in \mathbb{P}_k(e)$. A combination of these with arguments similar as those in [71, Lemma 5] results in the following

$$\begin{aligned}
|\mathbb{T}^1_\varphi| &\leq \frac{1}{2\delta_1}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_1}{6}\|\kappa^{-1/2}(u_h)\boldsymbol{\varepsilon^q}\|^2_{\Omega_h}, \\
|\mathbb{T}^2_\varphi| &\leq \frac{1}{2\delta_1}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_1}{6}\|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial\mathcal{T}_h}, \\
|\mathbb{T}^3_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_2}{2}R_h\,\underline{\kappa}^{-1}\|(h^\perp)^{1/2}\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Gamma_h}, \\
|\mathbb{T}^4_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_2}{6}\,\underline{\kappa}^{-1}\max_{e\in\mathcal{E}_h^\partial}\{r_e^2\}\,\|h^\perp\partial_n(\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Omega_h^c}, \\
|\mathbb{T}^5_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_2}{6}\,\overline{\kappa}\,\|\Pi_{\boldsymbol{V}}\boldsymbol{q}\|^2_{L^\infty(\Omega_h)}\, h\, L^2\,\left(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h}\right)^2, \\
|\mathbb{T}^6_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{2\delta_2}{3}\,\overline{\kappa}\,\underline{\kappa}^{-2}\max_{e\in\mathcal{E}_h^\partial} r_e^2\,\|h^\perp\partial_n(\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Omega_h^c}, \\
|\mathbb{T}^7_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + 2\,\delta_2\, R_h\,\overline{\kappa}\,\underline{\kappa}^{-2}\,\|(h^\perp)^{1/2}\boldsymbol{I^q}\cdot\boldsymbol{n})\|^2_{\Gamma_h}, \\
|\mathbb{T}^8_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|^2_{\Gamma_h} + \frac{\delta_2}{2}\overline{\kappa}\|(h^\perp)^{1/2}\Pi_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n}\|_{L^\infty(\Gamma_h)}L^2 h\,\left(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h}\right)^2,
\end{aligned}$$

as well as

$$|(\kappa^{-1}(u)\boldsymbol{I}^{\boldsymbol{q}},\boldsymbol{\varepsilon}^{\boldsymbol{q}})_{\mathcal{T}_h}| \leq \frac{1}{2\,\delta_3}\|\kappa^{-1/2}(u_h)\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|^2_{\Omega_h} + \frac{\delta_3}{2}\,\underline{\kappa}^{-2}\,\overline{\kappa}\|\boldsymbol{I}^{\boldsymbol{q}}\|^2_{\Omega_h},$$
$$|(\kappa^{-1}(u)-\kappa^{-1}(u_h))\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q},\boldsymbol{\varepsilon}^{\boldsymbol{q}})_{\mathcal{T}_h}| \leq \frac{1}{2\,\delta_3}\,\|\kappa^{-1/2}(u_h)\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|^2_{\Omega_h} + \frac{\delta_3}{2}\|\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q}\|^2_{L^\infty(\Omega_h)}\,\overline{\kappa}\,L^2\,(\|\varepsilon^u\|_{\Omega_h}+\|I^u\|_{\Omega_h})^2,$$
$$|\mathbb{T}^f| \leq L_f\,(\|\varepsilon^u\|_{\Omega_h}+\|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h},$$

where $\delta_1,\delta_2,\delta_3$ are free positive parameters arising from applications of Young's inequality, $\overline{\kappa}$ and $\underline{\kappa}$ are the upper and lower bounds for the diffusivity, and L and L_f are the Lipschitz constants from κ^{-1} and f respectively.

If we let $\delta_1=4$, $\delta_2=12$, and $\delta_3=6$ in the above estimates and substitute back into (3.25) we obtain

$$\begin{aligned}&|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|^2_{u_h}\\ &\quad\leq 12\,\max\{C_1\,h,C_2\}L^2\,\left(\|\varepsilon^u\|^2_{\Omega_h}+\|I^u\|^2_{\Omega_h}\right)+C\,L_f(\|\varepsilon^u\|_{\Omega_h}+\|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h}+C_3\,\Lambda^2_{\boldsymbol{q}}.\end{aligned}$$

where C_1,C_2 and C_3 only depend on $\overline{\kappa},\underline{\kappa},R_h$, and the projections $\|(h^\perp)^{1/2}\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n}\|^2_{L^\infty(\Gamma_h)}$ and $\|\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{q}\|^2_{L^\infty(\Omega_h)}$. ■

We now proceed to show that the approximation error in u can be indeed controlled by the errors in the approximation of the flux, the hybrid variable and the transfer error, modulo the approximation properties of the discrete spaces. To show that, in the next lemma we will build upon the ideas as in [71] and use a duality argument.

Lemma 3.5. *Assume that the Lipschitz constant is such that L_f is small enough, and consider the discrete spaces to be of polynomial degree $k\geq 1$. Then,*

$$\|\varepsilon^u\|_{\Omega_h} \lesssim h^{1/2}\,|\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\varepsilon^u-\varepsilon^{\widehat{u}},\varphi-\varphi_h)|\!|\!|_{u_h} + (h^{1/2}+L\,h)\Lambda_{\boldsymbol{q}} + (L+h^{1/2})\Lambda_u. \tag{3.26}$$

Proof. The first part of the proof follows very closely the argument used in the proof of Lemma 3.3. Given $\Theta\in L^2(\Omega)$ we will denote by $\boldsymbol{\phi}$ the solution to the dual problem 1.5 associated to Θ. Considering then the equations (3.19), together with the dual system, it is possible to show that

$$(\varepsilon^u,\Theta)_{\mathcal{T}_h} = \mathbb{T}^1_{\boldsymbol{q}}+\mathbb{T}^2_{\boldsymbol{q}}+\mathbb{T}_u+\mathbb{T}_f, \tag{3.27}$$

where

$$\mathbb{T}^1_{\boldsymbol{q}} := (\kappa^{-1}(u_h)(\boldsymbol{q}-\boldsymbol{q}_h),\boldsymbol{\Pi}_{\boldsymbol{V}}\boldsymbol{\phi})_{\mathcal{T}_h}+(\boldsymbol{\varepsilon}^{\boldsymbol{q}},\nabla\psi)_{\mathcal{T}_h},\qquad \mathbb{T}^2_{\boldsymbol{q}} := (\kappa^{-1}(u)-\kappa^{-1}(u_h))(\boldsymbol{I}^{\boldsymbol{q}}+\Pi_V\boldsymbol{q}),\Pi_V\phi)_{\mathcal{T}_h},$$
$$\mathbb{T}_u := \langle\varepsilon^{\widehat{u}},P_M(\boldsymbol{\phi}\cdot\boldsymbol{n})\rangle_{\Gamma_h}-\langle\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n},\Pi_W\psi\rangle_{\Gamma_h}\qquad \mathbb{T}_f := (f(u)-f(u_h),\Pi_W\psi)_{\mathcal{T}_h}.$$

To prove the result (3.26), we will bound each of the terms $\mathbb{T}_\star$, with $\star\in\{\boldsymbol{q},u,f\}$ in the decomposition (3.27).

Bound for $\mathbb{T}_f^i$. The simplest term to bound is $\mathbb{T}_f$, for which an application of Cauchy-Schwartz, the properties of the HDG projector and the Lipschitz continuity of the source term f, together with the dual estimate (1.8) yield

$$|\mathbb{T}_f| \leq C_{\text{reg}}\, L_f\, (\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\|\Theta\|_\Omega. \tag{3.28}$$

Where the constant C_{reg} is the stability constant from the dual problem (1.5).

Bound for $\mathbb{T}_{\boldsymbol{q}}^i$. Using the Lipschitz-continuity of κ (c.f. (3.7)) and following the arguments leading to equation (4.8) in [71, Lemma 5], the terms $\mathbb{T}_{\boldsymbol{q}}^1$ and $\mathbb{T}_{\boldsymbol{q}}^2$ can be bounded like

$$\begin{aligned}|\mathbb{T}_{\boldsymbol{q}}^1| &\leq \underline{\kappa}^{-1/2}\, \|\kappa^{-1/2}(u_h)(\boldsymbol{\varepsilon^q} + \boldsymbol{I^q})\|_{\Omega_h}\, \|\boldsymbol{\Pi_V}\phi - \phi\|_{\Omega_h} + \|\boldsymbol{I^q}\|_{\Omega_h}\, \|\nabla(\psi - \psi_h)\|_{\Omega_h} \\ &\leq CC_{\text{reg}}\underline{\kappa}^{-1/2} h^{\min\{1,k\}} \|\kappa^{-1/2}(u_h)\boldsymbol{\varepsilon^q}\|_{\Omega_h}\|\Theta\|_\Omega + 2CC_{\text{reg}} \max\{\underline{\kappa}^{-1/2}\overline{\kappa}^{-1/2}, 1\} h^{\min\{1,k\}} \|\boldsymbol{I^q}\|_{\Omega_h}\|\Theta\|_\Omega\end{aligned} \tag{3.29a}$$

and

$$|\mathbb{T}_{\boldsymbol{q}}^2| \leq C_{\text{reg}}(\|\boldsymbol{I^q}\|_\infty + \|\boldsymbol{\Pi_V q}\|_\infty)\, L\, (\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\|\theta\|_\Omega. \tag{3.29b}$$

Bound for $\mathbb{T}_u$. To estimate this term we will have to decompose it and treat each of the parts separately. We will write then write $\mathbb{T}_u := \sum_{i=1}^{11} \mathbb{T}_u^i$, where:

$$\begin{aligned}
\mathbb{T}_u^1 &:= -\langle \kappa(u_h)\, l^{-1}\, (\varphi(u) - \varphi(u_h)), \psi + l\partial_{\boldsymbol{n}}\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^7 &:= -\langle \tau(\varepsilon^u - \varepsilon^{\widehat{u}}), P_M\psi\rangle_{\Gamma_h}, \\
\mathbb{T}_u^2 &:= \langle \kappa(u_h)(\varphi(u) - \varphi_h(u_h)), (P_M - Id_M)\partial_{\boldsymbol{n}}\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^8 &:= -\langle \kappa(u_h)\, (\varphi(u) - \varphi_h(u_h))\, \delta_{\boldsymbol{I^q}}, \psi\rangle_{\Gamma_h}, \\
\mathbb{T}_u^3 &:= \langle \delta_{\boldsymbol{I}^q}, \psi\rangle_{\Gamma_h}, & \mathbb{T}_u^9 &:= -\langle \kappa(u_h)\, (\varphi(u) - \varphi_h(u_h))\, \delta_{\boldsymbol{\Pi_V q}}, \psi\rangle_{\Gamma_h}, \\
\mathbb{T}_u^4 &:= \langle \boldsymbol{I^q}\cdot\boldsymbol{n}, (Id_M - P_M)\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^{10} &:= -\langle \kappa(u_h)\, (\varphi(u) - \varphi_h(u_h))\, \boldsymbol{I^q}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}, \\
\mathbb{T}_u^5 &:= -\langle \tau P_M I^u, \psi\rangle_{\Gamma_h}, & \mathbb{T}_u^{11} &:= -\langle \kappa(u_h)(\varphi(u) - \varphi_h(u_h))\boldsymbol{\Pi_V q}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}. \\
\mathbb{T}_u^6 &:= \langle \delta_{\boldsymbol{\varepsilon^q}}, \psi\rangle_{\Gamma_h}, &&
\end{aligned}$$

Bounds for $\mathbb{T}_u^1 - \mathbb{T}_u^7$: These terms can be estimated by we applying the same techniques of [71, Lemma 6]. We will omit most of the the details here. Recalling that the length of the transfer path $l(\boldsymbol{x}) \leq c\, R_h\, h \quad \forall \boldsymbol{x} \in \Gamma_h$ and considering the constant $\tilde{c}$ from Lemma 1.2 ([71, Lemma 6]), we have

$$\begin{aligned}
|\mathbb{T}_u^1| &\leq c\,\tilde{c}\,\overline{\kappa}^{1/2}\, R_h\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^2| &\leq c\,\tilde{c}\,\overline{\kappa}^{1/2}\, R_h^{1/2}\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^3| &\leq \frac{1}{\sqrt{3}}\, c^{1/2}\, \tilde{c}\, R_h^{3/2}\, h^{1/2}\|h^\perp \partial_n \boldsymbol{I^q}\cdot\boldsymbol{n}\|_{\Omega_h^c}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^4| &\leq \tilde{c}\, h\|(h^\perp)^{1/2}\boldsymbol{I^q}\cdot\boldsymbol{n}\|_{\Gamma_h}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^5| &\leq \tilde{c}\,\overline{\tau}\, R_h\, h^{1/2}\|(h^\perp)^{1/2} I^u\|_{\Gamma_h}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^6| &\leq \frac{1}{\sqrt{3}}\, c^{1/2}\, \tilde{c}\,\overline{\kappa}^{1/2} \max_{e\in\mathcal{E}_h^\partial}\{C_{\text{ext}}^e, C_{\text{inv}}^e\} R_h^2\, h^{1/2}\, \|\kappa^{-1/2}(u_h)\, \boldsymbol{\varepsilon^q}\|_{\Omega_h}\|\Theta\|_\Omega, \\
|\mathbb{T}_u^7| &\leq c\,\tilde{c}\,\overline{\tau}^{1/2}\, R_h\, h\|\tau^{1/2}\, (\varepsilon^u - \varepsilon^{\widehat{u}})\|_{\partial\mathcal{T}_h}\|\Theta\|_\Omega.
\end{aligned}$$

Bounds for $\mathbb{T}_u^8 - \mathbb{T}_u^9$: Let us first notice that by definition of $\mathbb{T}_u^8$, we can obtain

$$|\mathbb{T}_u^8| = \left| \left\langle \kappa^{1/2}(u_h)\, l\, \kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\, l^{1/2}\, \delta_{\boldsymbol{I}^{\boldsymbol{q}}}, l^{-1}\, \psi \right\rangle_{\Gamma_h} \right|.$$

Then, by Cauchy-Schwartz, the fact that $l(\boldsymbol{x}) \leq c\, R_h\, h \quad \forall \boldsymbol{x} \in \Gamma_h$ and the boundedness of κ, we can obtain

$$|\mathbb{T}_u^8| \leq c\, \overline{\kappa}^{1/2}\, R_h\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\, l^{1/2}\, \delta_{\boldsymbol{I}^{\boldsymbol{q}}}\|_{\Gamma_h} \|l^{-1}\, \psi\|_{\Gamma_h}.$$

Finally, a direct application of (1.12c) to the factor involving the function $\delta_{\boldsymbol{I}^q}$, and using the estimation (1.13d) for the factor $\|l^{-1}\, \psi\|_{\Gamma_h}$, results in

$$|\mathbb{T}_u^8| \quad \leq \tfrac{1}{\sqrt{3}}\, c\, \tilde{c}\, \overline{\kappa}^{1/2}\, R_h^2\, h \sup_{\boldsymbol{x} \in \Gamma_h} \|(h^{\perp}\, \partial_n \boldsymbol{I}^{\boldsymbol{q}} \cdot \boldsymbol{n}\|_{l(\boldsymbol{x})}\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\, \|\Theta\|_{\Omega}.$$

Analogously, we can show that

$$|\mathbb{T}_u^9| \quad \leq \tfrac{1}{\sqrt{3}}\, c\, \tilde{c}\, \overline{\kappa}^{1/2}\, R_h^2\, h \sup_{\boldsymbol{x} \in \Gamma_h} \|(h^{\perp}\, \partial_n \boldsymbol{\Pi}_{\boldsymbol{v}} \boldsymbol{q} \cdot \boldsymbol{n}\|_{l(\boldsymbol{x})}\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\, \|\Theta\|_{\Omega}.$$

Bounds for $\mathbb{T}_u^{10} - \mathbb{T}_u^{11}$: We start as in the case for $\mathbb{T}_u^{10} - \mathbb{T}_u^{11}$ by combining, Cauchy-Schwartz , the bounds for κ and $l \lesssim R_h h$, to obtain

$$\begin{aligned} |\mathbb{T}_u^{10}| \quad &= \left| \left\langle \kappa^{1/2}(u_h)\, l\, \kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\, l^{1/2}\, \boldsymbol{I}^{\boldsymbol{q}} \cdot \boldsymbol{n}, l^{-1}\, \psi \right\rangle_{\Gamma_h} \right| \\ &\leq c\, \overline{\kappa}^{1/2}\, R_h\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\, l^{1/2} \boldsymbol{I}^{\boldsymbol{q}} \cdot \boldsymbol{n}\|_{\Gamma_h} \|l^{-1}\, \psi\|_{\Gamma_h} \\ &\leq c\, \overline{\kappa}^{1/2}\, R_h\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}(\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h} \|l^{1/2} \boldsymbol{I}^{\boldsymbol{q}} \cdot \boldsymbol{n}\|_{L^{\infty}(\Gamma_h)} \|l^{-1}\, \psi\|_{\Gamma_h}. \end{aligned}$$

From here, we will use the inequality $l(\boldsymbol{x}) \leq r_e h_e^{\perp}$ together with the estimate (1.13d) for $\|l^{-1}\, \psi\|_{\Gamma_h}$, we get

$$|\mathbb{T}_u^{10}| \quad \leq c\, \tilde{c}\, \overline{\kappa}^{1/2}\, R_h^{3/2}\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\, \|(h^{\perp})^{1/2}\, \boldsymbol{I}^{\boldsymbol{q}} \cdot \boldsymbol{n}\|_{L^{\infty}(\Gamma_h)} \|\Theta\|_{\Omega}.$$

Similar arguments can be used to derive the following analogous bound

$$|\mathbb{T}_u^{11}| \quad \leq c\, \tilde{c}\, \overline{\kappa}^{1/2}\, R_h^{3/2}\, h\, \|\kappa^{1/2}(u_h)\, l^{-1/2}\, (\varphi(u) - \varphi_h(u_h))\|_{\Gamma_h}\, \|(h^{\perp})^{1/2}\, \boldsymbol{\Pi}_{\boldsymbol{V}} \boldsymbol{q} \cdot \boldsymbol{n}\|_{L^{\infty}(\Gamma_h)} \|\Theta\|_{\Omega}.$$

Finally, letting $\Theta = \varepsilon^u$ in Ω_h and $\Theta = 0$ in Ω_h^c and using the estimates derived above for all the terms $\mathbb{T}_\star$ in the decomposition (3.27), one arrives at the desired estimate:

$$\|\varepsilon^u\|_{\Omega_h} \lesssim h^{1/2}\, |\!|\!|(\boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|\!|\!|_{u_h} + (h^{1/2} + L\, h) \Lambda_{\boldsymbol{q}} + (L + h^{1/2}) \Lambda_u.$$

■

This concludes the analysis of the discretization for problems with nonlinear diffusivities of the form $\kappa = \kappa(u)$. The reminder of the article will be devoted to the analysis of cases where the nonlinearity appears as a dependence to the gradient of the unknown. This functional dependence will require a different reformulation of the problem.

3.3 Non-linearities of the form $\kappa(\nabla u)$

3.3.1 Problem statement

In some applications, the diffusivity coefficient depends on the gradient of the solution, rather than on the solution itself. This is indeed the case, for instance, in the plasma equilibrium problem, where the coefficient is the inverse of the magnetic permeability. In ferromagnetic materials, the magnetic permeability becomes a function of the total magnetic field and therefore the coefficient has the functional dependence $\kappa = \kappa(\nabla u)$. In cases like this, we will be interested in boundary value problems of the form

$$-\nabla \cdot (\kappa(\nabla u)\nabla u) = f(u) \qquad \text{in } \Omega, \tag{3.30a}$$
$$u = g \qquad \text{on } \Gamma := \partial\Omega. \tag{3.30b}$$

where, just like in the previous section, the source term f will be assumed to be Lipschitz-continuous in Ω, with Lipschitz constant $L_f > 0$. We will also maintain the assumption (3.6) on boundedness of the permeability. Note that, since we will be searching for solutions with $H^1(\Omega)$ regularity, the hypothesis (3.6) will guarantee that the permeability remains bounded as a function of ∇u. The Lipschitz-continuity assumptions (3.7), (3.8), and (3.9) will be replaced by their following vector counterparts

$$\begin{aligned}
\|\kappa^{-1}(\boldsymbol{\sigma}_1) - \kappa^{-1}(\boldsymbol{\sigma}_2)\|_{L^2(\Gamma)} &\leq L\|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)} \qquad \forall\, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \in \boldsymbol{L}^2(\Omega), \\
\|\kappa^{1/2}(\boldsymbol{\sigma}_1) - \kappa^{1/2}(\boldsymbol{\sigma}_2)\|_{L^\infty(\Gamma)} &\leq \widehat{L}\|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)} \qquad \forall\, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \in \boldsymbol{L}^2(\Omega), \\
\|\kappa(\boldsymbol{\sigma}_1) - \kappa(\boldsymbol{\sigma}_2)\|_{L^2(\Omega)} &\leq \widetilde{L}\|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)} \qquad \forall\, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \in \boldsymbol{L}^2(\Omega).
\end{aligned} \tag{3.31}$$

Following the spirit of reformulating the problem in a mixed form, the functional dependence $\kappa(\nabla u)$ will require us to introduce a new auxiliary variable. Therefore, we introduce $\boldsymbol{\sigma} := \nabla u$ and will express the the flux as $\boldsymbol{q} := -\kappa(\boldsymbol{\sigma})\boldsymbol{\sigma}$, thus introducing two additional unknowns to the problem. With these definitions, it is possible to write (3.30) as the equivalent system

$$\begin{aligned}
\boldsymbol{\sigma} - \nabla u = 0 \qquad & \text{in } \Omega, \\
\boldsymbol{q} + \kappa(\boldsymbol{\sigma})\,\boldsymbol{\sigma} = 0 \qquad & \text{in } \Omega, \\
\nabla \cdot \boldsymbol{q} = f(u) \qquad & \text{in } \Omega, \\
u = g \qquad & \text{on } \partial\Omega.
\end{aligned}$$

We shall analyze the discretization of this system when restricted to the subdomain in Ω_h. In view of this, our target formulation is

$$\boldsymbol{\sigma} - \nabla u = 0 \qquad \text{in } \Omega_h, \tag{3.32a}$$
$$\boldsymbol{q} + \kappa(\boldsymbol{\sigma})\,\boldsymbol{\sigma} = 0 \qquad \text{in } \Omega_h, \tag{3.32b}$$
$$\nabla \cdot \boldsymbol{q} = f(u) \qquad \text{in } \Omega_h, \tag{3.32c}$$
$$u = \varphi \qquad \text{on } \Gamma_h = \partial\Omega_h. \tag{3.32d}$$

where the boundary conditions have been transferred by means of

$$\varphi(\boldsymbol{\sigma},\boldsymbol{x}) := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} \big(\kappa^{-1}(\boldsymbol{\sigma})\boldsymbol{q}\big)(\boldsymbol{x}+\boldsymbol{t}(\boldsymbol{x})s)\cdot\boldsymbol{t}(\boldsymbol{x})ds.$$

The HDG scheme associated to (3.32) reads: Find $(\boldsymbol{q}_h,\boldsymbol{\sigma}_h,u_h,\widehat{u}_h)\in \boldsymbol{V}_h\times\boldsymbol{V}_h\times W_h\times M_h$, such that

$$(\boldsymbol{\sigma}_h,\boldsymbol{v})_{\mathcal{T}_h} + (u_h,\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} - \langle\widehat{u}_h,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.33a}$$
$$(\boldsymbol{q}_h,\boldsymbol{s})_{\mathcal{T}_h} + (\kappa(\boldsymbol{\sigma}_h)\,\boldsymbol{\sigma}_h,\boldsymbol{s})_{\mathcal{T}_h} = 0, \tag{3.33b}$$
$$-(\boldsymbol{q}_h,\nabla w)_{\mathcal{T}_h} + \langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n},w\rangle_{\partial\mathcal{T}_h} = (f(u_h),w)_{\mathcal{T}_h}, \tag{3.33c}$$
$$\langle\widehat{u}_h,\mu\rangle_{\Gamma_h} = \langle\varphi_h(\boldsymbol{\sigma}_h),\mu\rangle_{\Gamma_h}, \tag{3.33d}$$
$$\langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n},\mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} = 0, \tag{3.33e}$$

for all $(\boldsymbol{v},\boldsymbol{s},w,\mu)\in\boldsymbol{V}_h\times\boldsymbol{V}_h\times W_h\times M_h$. Here, the spaces $\boldsymbol{V}_h$, W_h, and M_h have been defined in (1.6), the restriction to the mesh skeleton of the numerical flux has been defined as

$$\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n} := \boldsymbol{q}_h\cdot\boldsymbol{n} + \tau(u_h-\widehat{u}_h) \quad \text{on} \quad \partial\mathcal{T}_h,$$

and the approximate boundary condition given by

$$\varphi_h(\boldsymbol{\sigma}_h,\boldsymbol{x}) := g(\overline{\boldsymbol{x}}) + \int_0^{l(\boldsymbol{x})} \big(\kappa^{-1}(\boldsymbol{\sigma}_h)\boldsymbol{q}_h\big)(\boldsymbol{x}+\boldsymbol{t}(\boldsymbol{x})s)\cdot\boldsymbol{t}(\boldsymbol{x})ds. \tag{3.33f}$$

As before, the maximum value of the positive stabilization function τ will be denoted by $\overline{\tau}$.

In this section we will analyze an HDG scheme for problems of this form. We will first have to reformulate the problem in terms of a mixed system with one additional unknown when compared to the case analyzed in the previous section. Many of the arguments required for the analysis will be similar to those developed in the previous section, and the analysis technique is similar as well. We will therefore omit many of the technical details and indicate the main steps in the analysis, focusing on those that are different from the previous section.

3.3.2 Well-posedness

The proof that the system (3.11) is well-posed will rely on a fixed point argument. As in the previous section, we define the operator $\mathcal{J}_2 : \boldsymbol{V}_h\times W_h\to\boldsymbol{V}_h\times W_h$ that maps a pair of functions $(\boldsymbol{\eta},\zeta)$ to the first and third component of the solution $(\boldsymbol{q},\boldsymbol{\sigma},u,\widehat{u})\in\boldsymbol{V}_h\times\boldsymbol{V}_h\times W_h\times M_h$ to the HDG system (3.33) where the source has been evaluated at $(\boldsymbol{\eta},\zeta)$. Namely

$$(\boldsymbol{\sigma},\boldsymbol{v})_{\mathcal{T}_h} + (u,\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} - \langle\widehat{u},\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{3.34a}$$
$$(\boldsymbol{q},\boldsymbol{s})_{\mathcal{T}_h} + (\kappa(\boldsymbol{\eta})\,\boldsymbol{\sigma},\boldsymbol{s})_{\mathcal{T}_h} = 0, \tag{3.34b}$$
$$-(\boldsymbol{q},\nabla w)_{\mathcal{T}_h} + \langle\widehat{\boldsymbol{q}}\cdot\boldsymbol{n},w\rangle_{\partial\mathcal{T}_h} = (f(\zeta),w)_{\mathcal{T}_h}, \tag{3.34c}$$
$$\langle\widehat{u},\mu\rangle_{\Gamma_h} = \langle\varphi(\boldsymbol{\eta}),\mu\rangle_{\Gamma_h}, \tag{3.34d}$$
$$\langle\widehat{\boldsymbol{q}}\cdot\boldsymbol{n},\mu\rangle_{\partial\mathcal{T}_h\setminus\Gamma_h} = 0, \tag{3.34e}$$

for all $(\boldsymbol{v}, \boldsymbol{s}, w, \mu) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times M_h$. Just as before, $\varphi(\zeta)$ accounts for the transferred boundary conditions and, since the discrete linearized system above is uniquely solvable [18], $\mathcal{J}_2$ is well defined.

Given a function $\boldsymbol{\eta} \in \boldsymbol{V}_h$, we define the following norm over the product space $\boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times M_h$

$$|\!|\!|(\boldsymbol{s}, \boldsymbol{v}, w, \lambda)|\!|\!|_{\boldsymbol{\eta}} := \Big(\|\boldsymbol{s}\|^2_{\Omega_h} + \|\kappa^{1/2}(\boldsymbol{\eta})\boldsymbol{v}\|^2_{\Omega_h} + \|\tau^{1/2}\, w\|^2_{\partial\mathcal{T}_h} + \|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \lambda(\boldsymbol{\eta})\|^2_{\Gamma_h} \Big)^{1/2}. \tag{3.35}$$

The general road map for the proof is as follows. Lemmas 3.6 and 3.7 below, will allow us to control $(\boldsymbol{q}, \boldsymbol{\sigma}, u - \widehat{u}, \varphi)$ by the linearized source term $f(\zeta)$ and the boundary condition *at the physical boundary*, $\overline{g}$. An application of these two results will then allow us—modulo some technical assumptions involving the bound of the diffusivity and the distance between the physical and computational domains—to use the Lipschitz continuity of f and κ to prove that the mapping is indeed a contraction. This will be done in Theorem 3.3, from which the well posedness of the HDG system (3.33) will follow as a simple corollary.

Lemma 3.6. *If Assumptions (3.3) hold, then*

$$|\!|\!|(\boldsymbol{q}, \boldsymbol{\sigma}, u - \widehat{u}, \varphi)|\!|\!|^2_{\boldsymbol{\eta}} \leq \max\{1, \overline{\kappa}\} \Big(4\|f(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + 6\|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|^2_{\Gamma_h} \Big). \tag{3.36}$$

Proof. Note that if we let $\boldsymbol{s} = \boldsymbol{q}$ in (3.34b) , we have

$$\|\boldsymbol{q}\|^2_{\Omega_h} \leq \overline{\kappa}\, \|\kappa^{1/2}(\boldsymbol{\eta})\, \boldsymbol{\sigma}\|^2_{\Omega_h}. \tag{3.37}$$

Then, following the process outlined in the proof of Lemma 3.2, we go back to (3.34) and choose the test functions as $\boldsymbol{v} = -\boldsymbol{q}, \boldsymbol{s} = \boldsymbol{\sigma}, w = u$, and

$$\mu = \begin{cases} -\widehat{u}, & \text{on } \partial\mathcal{T}_h \setminus \Gamma_h \\ -\widehat{\boldsymbol{q}} \cdot \boldsymbol{n}, & \text{on } \Gamma_h \end{cases},$$

This leads to the equality

$$\begin{aligned} \|\kappa^{1/2}(\boldsymbol{\eta})\boldsymbol{\sigma}\|^2_{\Omega_h} + \|\tau^{1/2}(u - \widehat{u})\|^2_{\partial\mathcal{T}_h} = \; & (f(\zeta), u)_{\mathcal{T}_h} - \langle \varphi(\boldsymbol{\eta}), \kappa(\boldsymbol{\eta}) l^{-1} \varphi(\boldsymbol{\eta}) \rangle_{\Gamma_h} + \langle \varphi(\boldsymbol{\eta}), \kappa(\boldsymbol{\sigma}) l^{-1} \overline{\mathrm{g}} \rangle_{\Gamma_h} \\ & + \langle \varphi(\boldsymbol{\eta}), \overline{\mathrm{g}} \rangle_{\Gamma_h} - \langle \varphi(\boldsymbol{\eta}), \tau(u - \widehat{u}) \rangle_{\Gamma_h}. \end{aligned}$$

The terms on the right hand side can be estimated by an application of Lemma 3.1, yielding

$$\|\kappa^{1/2}(\boldsymbol{\eta})\boldsymbol{\sigma}\|^2_{\Omega_h} + \|\tau^{1/2}\,(u - \widehat{u})\|^2_{\partial\mathcal{T}_h} + \|\kappa^{1/2}(\eta)\, l^{-1/2}\, \varphi(\boldsymbol{\eta})\|^2_{\Gamma_h} \leq 2\|f(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + 3\|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|^2_{\Gamma_h}.$$

Combining this estimate with (3.37), we obtain

$$|\!|\!|(\boldsymbol{q}, \boldsymbol{\sigma}, u - \widehat{u}, \varphi)|\!|\!|^2_{\boldsymbol{\eta}} \leq \max\{1, \overline{\kappa}\} \left(4\|f(\zeta)\|_{\Omega_h}\|u\|_{\Omega_h} + 6\|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|^2_{\Gamma_h} \right),$$

which concludes the proof. ∎

It only remains now to estimate the norm of u in terms of the sources and the boundary conditions. This will be done in the next lemma.

Lemma 3.7. *Suppose that Assumptions (3.3) and (1.8) are satisfied. Then, there exists $\widehat{c} > 0$, independent of h such that*

$$\|u\|_{\Omega_h} \leq 4 \max\{\widehat{c}^2 h, 1\} \|f(\zeta)\|_{\Omega_h} + 2\,(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2} R_h^{1/2})\, h^{1/2} \|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}. \tag{3.38}$$

Proof. The proof of this result is follows, with small variations, the same process as that of Lemma 3.3. By using $\boldsymbol{\eta}$ instead of ζ in the dual problem given in (**??**), and splitting the duality product as

$$(u, \Theta)_{\mathcal{T}_h} = \mathbb{T}_{\boldsymbol{\sigma}} + \mathbb{T}_u + \mathbb{T}_f,$$

where

$$\mathbb{T}_{\boldsymbol{\sigma}} := -(\boldsymbol{\sigma}, \boldsymbol{\Pi_V}\boldsymbol{\phi} - \boldsymbol{\phi})_{\mathcal{T}_h}, \quad \mathbb{T}_u := \langle \widehat{u}, P_M(\boldsymbol{\phi}\cdot\boldsymbol{n})\rangle_{\Gamma_h} - \langle \widehat{\boldsymbol{q}}\cdot\boldsymbol{n}, \Pi_W\psi\rangle_{\Gamma_h} \quad \text{and} \quad \mathbb{T}_f := (f(\zeta), \Pi_W\psi)_{\mathcal{T}_h}.$$

The terms $\mathbb{T}_{\boldsymbol{\sigma}}$ and $\mathbb{T}_f$ are bounded as

$$|\mathbb{T}_{\boldsymbol{\sigma}}| \lesssim \underline{\kappa}^{-1/2}\, h \|\kappa^{-1/2}(\boldsymbol{\eta})\,\boldsymbol{\sigma}\|_{\Omega_h} \|\Theta\|_{\Omega}, \quad \text{and} \quad |\mathbb{T}_f| \lesssim \|f(\zeta)\|_{\Omega_h} \|\Theta\|_{\Omega},$$

and, we rewrite $\mathbb{T}_u$ as $\mathbb{T}_u = \sum_{i=1}^5 \mathbb{T}_u^i$, with

$$\begin{aligned}
\mathbb{T}_u^1 &:= -\langle \kappa(\boldsymbol{\eta}) l^{-1}\varphi(\boldsymbol{\eta}), \psi + l\partial_n\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^4 &:= -\langle \tau(u-\widehat{u}), P_M\psi\rangle_{\Gamma_h},\\
\mathbb{T}_u^2 &:= \langle \kappa(\boldsymbol{\eta})\varphi(\boldsymbol{\eta}), (P_M - I_d)\partial_n\psi\rangle_{\Gamma_h}, & \mathbb{T}_u^5 &:= \langle \kappa(\boldsymbol{\eta})\, l^{-1}\,\overline{\mathrm{g}}, \psi\rangle_{\Gamma_h},\\
\mathbb{T}_u^3 &:= \langle \delta_{\boldsymbol{\sigma}}, \psi\rangle_{\Gamma_h}.
\end{aligned}$$

These terms can be bounded using arguments analogous to those in Lemma 3.3, yielding the desired estimate (3.38). ∎

The results in the two preceding lemmas can be combined to estimate $(\boldsymbol{q}, \boldsymbol{\sigma}, u - \widehat{u}, \varphi)$ in terms of the source $f(\zeta)$ and the boundary data $\overline{\mathrm{g}}$. This follows readily from an application of Lemma 3.7 to (3.36), yielding

$$\begin{aligned}
|\!|\!|(\boldsymbol{q}, \boldsymbol{\sigma}, u-\widehat{u}, \varphi)|\!|\!|_\eta^2 \leq\ & \left(16\,\max\{1,\overline{\kappa}\}^2 + 8\,\max\{\widehat{c}^2 h, 1\}^2\right) \|f(\zeta)\|_{\Omega_h}^2 \\
&+ \left(6\max\{1,\overline{\kappa}\} + 2\,(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2} R_h^{1/2})^2\, h\right) \|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}^2.
\end{aligned} \tag{3.39}$$

In turn, (3.38) implies that

$$\|u\|_{\Omega_h}^2 \leq 32\,\max\{\widehat{c}^2 h, 1\}^2 \|f(\zeta)\|_{\Omega_h}^2 + 8\,(\sqrt{3}\,\widehat{c} + \overline{\kappa}^{1/2} R_h^{1/2})^2\, h\, \|\kappa^{1/2}(\boldsymbol{\eta})\, l^{-1/2}\, \overline{\mathrm{g}}\|_{\Gamma_h}^2. \tag{3.40}$$

These two estimates will be used to prove the contractive properties of $\mathcal{J}_2$ as we will now show.

Theorem 3.3. *Suppose that the dual regularity (1.8) and the Assumptions (3.3) hold and suppose also that*

$$\left(16\,\max\{1,\overline{\kappa}\}^2 + 40\,\max\{\widehat{c}^2 h, 1\}^2\right) L_f^2 < \frac{1}{4}, \tag{3.41}$$

and

$$\left(6\max\{1,\overline{\kappa}\}+10\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R_h^{1/2})^2\,h\right)\widehat{L}^2\|l^{-1/2}\,\overline{\mathrm{g}}\|^2_{\Gamma_h}<\frac{1}{4}\tag{3.42}$$

are satisfied. Then $\mathcal{J}_2$ is a contraction operator.

Proof. Let $(\boldsymbol{\eta}_i,\zeta_i)\in \boldsymbol{V}_h\times W_h$ and define $(\boldsymbol{\sigma}_i,u_i):=\mathcal{J}_2\left((\boldsymbol{\eta}_i,\zeta_i)\right)\in \boldsymbol{V}_h\times W_h$ for $i\in\{1,2\}$. Then,

$$\|\mathcal{J}_2(\boldsymbol{\eta}_1,\zeta_1)-\mathcal{J}_2(\boldsymbol{\eta}_2,\zeta_2)\|_{\Omega_h}=\|(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2,u_1-u_2)\|_{\Omega_h}=\left(\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2_{\Omega_h}+\|u_1-u_2\|^2_{\Omega_h}\right)^{1/2}.$$

By applying the inequalities (3.39) and (3.40) respectively to $(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)$ and (u_1-u_2), we obtain

$$\begin{aligned}\|\mathcal{J}_2(\boldsymbol{\eta}_1,\zeta_1)-\mathcal{J}_2(\boldsymbol{\eta}_2,\zeta_2)\|_{\Omega_h}\leq\Bigg(&\left(16\,\max\{1,\overline{\kappa}\}^2+40\,\max\{\widehat{c}^2h,1\}^2\right)\|f(\zeta_1)-f(\zeta_2)\|^2_{\Omega_h}\\&+\left(6\max\{1,\overline{\kappa}\}+10\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R_h^{1/2})^2\,h\right)\|\kappa^{1/2}(\boldsymbol{\eta}_1)-\kappa^{1/2}(\boldsymbol{\eta}_2)\|^2_{L^\infty(\Gamma_h)}\|l^{-1/2}\,\overline{\mathrm{g}}\|^2_{\Gamma_h}\Bigg)^{1/2}.\end{aligned}$$

Then, using the Lipschitz continuity of f and $\kappa^{1/2}$, we get

$$\begin{aligned}\|\mathcal{J}_2(\boldsymbol{\eta}_1,\zeta_1)-\mathcal{J}_2(\boldsymbol{\eta}_2,\zeta_2)\|_{\Omega_h}\leq\Bigg(&\left(16\,\max\{1,\overline{\kappa}\}^2+40\,\max\{\widehat{c}^2h,1\}^2\right)L_f^2\|\zeta_1-\zeta_2\|^2_{\Omega_h}\\&+\left(6\max\{1,\overline{\kappa}\}+10\,(\sqrt{3}\,\widehat{c}+\overline{\kappa}^{1/2}\,R_h^{1/2})^2\,h\right)\widehat{L}^2\|\boldsymbol{\eta}_1-\boldsymbol{\eta}_2\|^2_{L^\infty(\Omega_h)}\|l^{-1/2}\,\overline{\mathrm{g}}\|^2_{\Gamma_h}\Bigg)^{1/2}.\end{aligned}$$

The proof is concluded, by applying the assumptions (3.41) and (3.42) to the right hand side of the preceding inequality. ■

As a result we can then conclude this section with with the following

Corollary 3.2. *If the hypotheses of Theorem 3.3 are satisfied, the HDG system (3.34) is well posed.*

Having established the well posedness of the discrete system (3.34), we will concentrate our efforts in the next section on establishing the convergence properties of the HDG scheme.

3.3.3 *A priori* error analysis

The study of the convergence properties and rates of our discretization follows similar steps as the ones laid out in Section 3.2.3, adapted for the extended system that arises from the introduction of the additional auxiliary variable $\boldsymbol{\sigma}=\nabla u$. To avoid unnecessary repetition of arguments, we will focus on the differences between these two cases and will omit most of the details that can be easily inferred from Section 3.2.3.

As before, we decompose the error with the aid of the HDG projection as :

$$\boldsymbol{q}-\boldsymbol{q}_h=\boldsymbol{\varepsilon}^{\boldsymbol{q}}+\boldsymbol{I}^{\boldsymbol{q}},\quad \boldsymbol{\sigma}-\boldsymbol{\sigma}_h=\boldsymbol{\varepsilon}^{\boldsymbol{\sigma}}+\boldsymbol{I}^{\boldsymbol{\sigma}},\quad\text{and}\quad u-u_h=\varepsilon^u+I^u,$$

where, similar to Section 3.2.3, we have defined

$$\begin{aligned}
&\boldsymbol{\varepsilon^q} := \boldsymbol{\Pi_V q} - \boldsymbol{q}_h, \qquad \boldsymbol{\varepsilon^\sigma} := \boldsymbol{\Pi_V \sigma} - \boldsymbol{\sigma}_h, \qquad \varepsilon^u := \Pi_W u - u_h, \qquad \textit{(Projection of the error)}\\
&\boldsymbol{I^q} := \ \boldsymbol{q} - \boldsymbol{\Pi_V q}, \qquad \boldsymbol{I^\sigma} := \ \boldsymbol{\sigma} - \boldsymbol{\Pi_V \sigma}, \qquad I^u := \ u - \Pi_W u. \qquad \textit{(Error of the projection)}
\end{aligned}$$

In addition, using the L^2 projection into M_h we have $\varepsilon^{\widehat{u}} := P_M u - \widehat{u}_h$. The vector of error projections $(\boldsymbol{\varepsilon^q}, \boldsymbol{\varepsilon^\sigma}, \varepsilon^u, \varepsilon^{\widehat{u}})$ belongs to $\boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times M_h$ and satisfies the error equations

$$(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{v})_{\mathcal{T}_h} + (\varepsilon^u, \nabla \cdot \boldsymbol{v})_{\mathcal{T}_h} - \langle \varepsilon^{\widehat{u}}, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} = -(\boldsymbol{I^\sigma}, \boldsymbol{v})_{\mathcal{T}_h} - (\boldsymbol{I}^u, \nabla \cdot \boldsymbol{v})_{\mathcal{T}_h} \tag{3.43a}$$

$$(\boldsymbol{\varepsilon^q}, \boldsymbol{s})_{\mathcal{T}_h} + (\kappa(\boldsymbol{\sigma}_h)\, \boldsymbol{\varepsilon}^\sigma, \boldsymbol{s})_{\mathcal{T}_h} = -(\boldsymbol{I^q}, \boldsymbol{s})_{\mathcal{T}_h} - (\kappa(\boldsymbol{\sigma})\, \boldsymbol{I^\sigma}, \boldsymbol{s})_{\mathcal{T}_h} \tag{3.43b}$$

$$- ((\kappa(\boldsymbol{\sigma}) - \kappa(\boldsymbol{\sigma}_h))\, \boldsymbol{\Pi_V \sigma}, s)_{\mathcal{T}_h}\,, \tag{3.43c}$$

$$-(\boldsymbol{\varepsilon^q}, \nabla w)_{\mathcal{T}_h} + \langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n}, w \rangle_{\partial \mathcal{T}_h} = (f(u) - f(u_h), w)_{\mathcal{T}_h}, \tag{3.43d}$$

$$\langle \varepsilon^{\widehat{u}}, \mu \rangle_{\Gamma_h} = \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \mu \rangle_{\Gamma_h}, \tag{3.43e}$$

$$\langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n}, \mu \rangle_{\partial \mathcal{T}_h \setminus \Gamma_h} = 0, \tag{3.43f}$$

for all $(\boldsymbol{v}, \boldsymbol{s}, w, \mu) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times W_h \times M_h$. Here, as before, on $\partial \mathcal{T}_h$ we have $\boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}} \cdot \boldsymbol{n} = \boldsymbol{\varepsilon^q} \cdot \boldsymbol{n} + \tau(\varepsilon^u - \varepsilon^{\widehat{u}})$.

Following the same arguments of Lemma 3.4 is possible to estimate the magnitude of $(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon^q}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)$, as measured by the norm $\|\!|\cdot|\!\|_{\boldsymbol{\sigma}}$ defined in (3.35). Choosing the vector of approximation errors both as test and trial in the error equations we obtain

$$\begin{aligned}
&\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, \boldsymbol{\varepsilon}^\sigma\|^2_{\Omega_h} + \|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial \mathcal{T}_h} + \|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}\, (\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h))\|^2_{\Gamma_h}\\
&\quad \leq |(\boldsymbol{I^q}, \boldsymbol{\varepsilon}^\sigma)_{\mathcal{T}_h}| + |(\boldsymbol{I^\sigma}, \boldsymbol{\varepsilon^q})_{\mathcal{T}_h}| + |(\kappa(\boldsymbol{\sigma}) \boldsymbol{I^\sigma}, \boldsymbol{\varepsilon}^\sigma)_{\mathcal{T}_h}| + |((\kappa(\boldsymbol{\sigma}) - \kappa(\boldsymbol{\sigma}_h)) \boldsymbol{\Pi_V \sigma}, \boldsymbol{\varepsilon}^\sigma)_{\mathcal{T}_h}| + |\mathbb{T}^f| + |\mathbb{T}_\varphi|.
\end{aligned} \tag{3.44}$$

The final two terms are defined as $\mathbb{T}^f := (f(u) - f(u_h), \varepsilon^u)_{\mathcal{T}_h}$ and $\mathbb{T}_\varphi := \sum_{i=1}^{8} |\mathrm{T}^i_\varphi|$, where

$$\begin{aligned}
\mathbb{T}^1_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \delta_{\boldsymbol{\varepsilon^q}} \rangle_{\Gamma_h}\\
\mathbb{T}^2_\varphi &:= -\langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \tau(\varepsilon^u - \varepsilon^{\widehat{u}}) \rangle_{\Gamma_h}\\
\mathbb{T}^3_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \boldsymbol{I^q} \cdot \boldsymbol{n} \rangle_{\Gamma_h}\\
\mathbb{T}^4_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \delta_{\boldsymbol{I^q}} \rangle_{\Gamma_h}\\
\mathbb{T}^5_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \kappa(\boldsymbol{\sigma}_h)\, (\kappa^{-1}(\boldsymbol{\sigma}) - \kappa^{-1}(\boldsymbol{\sigma}_h))\, \delta_{\boldsymbol{\Pi_V q}} \rangle_{\Gamma_h}\\
\mathbb{T}^6_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \kappa(\boldsymbol{\sigma}_h)\, (\kappa^{-1}(\boldsymbol{\sigma}) - \kappa^{-1}(\boldsymbol{\sigma}_h))\, \delta_{\boldsymbol{I^q}} \rangle_{\Gamma_h}\\
\mathbb{T}^7_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \kappa(\boldsymbol{\sigma}_h)\, (\kappa^{-1}(\boldsymbol{\sigma}) - \kappa^{-1}(\boldsymbol{\sigma}_h))\, \boldsymbol{I^q} \cdot \boldsymbol{n} \rangle_{\Gamma_h}\\
\mathbb{T}^8_\varphi &:= \langle \varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h), \kappa(\boldsymbol{\sigma}_h)\, (\kappa^{-1}(\boldsymbol{\sigma}) - \kappa^{-1}(\boldsymbol{\sigma}_h))\, \boldsymbol{\Pi_V q} \cdot \boldsymbol{n} \rangle_{\Gamma_h}.
\end{aligned}$$

By a combined use of arguments similar to those appearing in Lemma 3.4, it is possible to deduce

$$
\begin{aligned}
|\mathbb{T}^1_\varphi| &\leq \frac{1}{2\delta_1}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \frac{\delta_1}{6}\,\overline{\kappa}^{-1}\,\|\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|^2_{\Omega_h},\\
|\mathbb{T}^2_\varphi| &\leq \frac{1}{2\delta_2}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \frac{\delta_2}{6}\,\|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial\mathcal{T}_h},\\
|\mathbb{T}^3_\varphi| &\leq \frac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \frac{\delta_3}{2}\,R_h\,\underline{\kappa}^{-1}\,\|(h^\perp)^{1/2}\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Gamma_h},\\
|\mathbb{T}^4_\varphi| &\leq \frac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \frac{\delta_3}{6}\,\underline{\kappa}^{-1}\,R_h^2\,\|h^\perp\partial_n(\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Omega_h^c},\\
|\mathbb{T}^5_\varphi| &\leq \frac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h}\\
&\quad + \frac{\delta_3}{3}\,\overline{\kappa}\,R_h^2\,\sup_{\boldsymbol{x}\in e\subseteq\Gamma_h}\|h^\perp\partial_n(\Pi_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n})\|^2_{l(\boldsymbol{x})}\,L^2\,\left(\|\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h}\right),\\
|\mathbb{T}^6_\varphi| &\leq \frac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \frac{2\delta_3}{3}\,\overline{\kappa}\,\underline{\kappa}^{-2}\,R_h^2\,\|h^\perp\partial_n(\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Omega_h^c},\\
|\mathbb{T}^7_\varphi| &\leq \frac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\, l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + 2\,\delta_3\,R_h\,\overline{\kappa}\,\underline{\kappa}^{-2}\,\|(h^\perp)^{1/2}\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n})\|^2_{\Gamma_h},\\
|\mathbb{T}^8_\varphi| &\leq \tfrac{1}{2\delta_3}\|\kappa^{1/2}(\boldsymbol{\sigma}_h) l^{-1/2}(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)\|^2_{\Gamma_h} + \delta_3\overline{\kappa}R_h\|(h^\perp)^{1/2}\Pi_{\boldsymbol{V}}\boldsymbol{q}\cdot\boldsymbol{n}\|^2_\infty L^2\left(\|\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h}\right),
\end{aligned}
\tag{3.45}
$$

and

$$
\begin{aligned}
|(\boldsymbol{I}^{\boldsymbol{\sigma}}, \boldsymbol{\varepsilon}^{\boldsymbol{q}})_{\mathcal{T}_h}| &\leq \frac{1}{2\,\delta_4}\|\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|^2_{\Omega_h} + \frac{\delta_4}{2}\,\|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h},\\
|(\boldsymbol{I}^{\boldsymbol{q}}, \boldsymbol{\varepsilon}^{\sigma})_{\mathcal{T}_h}| &\leq \frac{1}{2\,\delta_5}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \frac{\delta_5}{2}\,\underline{\kappa}^{-1}\,\|\boldsymbol{I}^{\boldsymbol{q}}\|^2_{\Omega_h},\\
|(\kappa(\boldsymbol{\sigma})\boldsymbol{I}^{\boldsymbol{\sigma}}, \boldsymbol{\varepsilon}^{\sigma})_{\mathcal{T}_h}| &\leq \frac{1}{2\,\delta_5}\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \frac{\delta_5}{2}\,\underline{\kappa}^{-1}\,\overline{\kappa}^2\|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h},\\
|(\kappa(\boldsymbol{\sigma}) - \kappa(\boldsymbol{\sigma}_h))\,\Pi_{\boldsymbol{V}}\boldsymbol{\sigma}, \boldsymbol{\varepsilon}^{\sigma})_{\mathcal{T}_h}| &\leq \frac{1}{2\,\delta_5}\,\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \delta_5\,\underline{\kappa}^{-1}\,\|\Pi_{\boldsymbol{V}}\boldsymbol{q}\|^2_\infty\,L^2\,\left(\|\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h}\right).\\
|\mathbb{T}^f| &\leq L_f\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h}.
\end{aligned}
\tag{3.46}
$$

Then, taking the test function $s = \boldsymbol{\varepsilon}^{\boldsymbol{q}}$ in the second equation of (3.43), we get

$$
\|\boldsymbol{\varepsilon}^{\boldsymbol{q}}\|^2_{\Omega_h} \leq \left(4\,\overline{\kappa} + 8\,\underline{\kappa}^{-1}\,L^2\,\|\Pi_{\boldsymbol{V}}\boldsymbol{q}\|^2_\infty\right)\,\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\,\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + 4\,\|\boldsymbol{I}^{\boldsymbol{q}}\|^2_{\Omega_h} + 4\,\max\{\overline{\kappa}, 2\,L^2\,\|\Pi_{\boldsymbol{V}}\,\boldsymbol{\sigma}\|^2_\infty\}\,\|\boldsymbol{I}^{\boldsymbol{\sigma}}\|^2_{\Omega_h}.
\tag{3.47}
$$

If the Lipschitz constants L and L_f are sufficiently small, a direct application of (3.47) in the first equations of (3.45) and (3.46), together with the choices $\delta_1 = 1, \delta_2 = 3, \delta_3 = 12, \delta_4 = 24/\overline{\kappa}, \delta_5 = 18$ yield the following estimate for the right hand side of (3.44)

$$
\begin{aligned}
\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\,\boldsymbol{\varepsilon}^{\sigma}\|^2_{\Omega_h} + \|\tau^{1/2}(\varepsilon^u - \varepsilon^{\widehat{u}})\|^2_{\partial\mathcal{T}_h} + \|\kappa^{1/2}(\boldsymbol{\sigma}_h)\,l^{-1/2}\,(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h))\|^2_{\Gamma_h}&\\
\lesssim \Lambda^2_{\boldsymbol{q}} + \Lambda^2_{\boldsymbol{\sigma}} + L_f\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h}.&
\end{aligned}
\tag{3.48}
$$

In the expression above, the term $\Lambda_{\boldsymbol{\sigma}}$ has been defined according to (3.20a). By combining (3.47) and (3.48) we get

$$
|\!|\!|(\boldsymbol{\varepsilon}^{\sigma}, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|\!|\!|^2_{\boldsymbol{\sigma}_h} \lesssim \Lambda^2_{\boldsymbol{q}} + \Lambda^2_{\boldsymbol{\sigma}} + L_f\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\varepsilon^u\|_{\Omega_h}.
\tag{3.49}
$$

The following result allows us to estimate the term ε^u in (3.49) by means of a duality argument. The

proof technique is analogous to the one used for Lemma 3.5.

Lemma 3.8. *Given the regularity condition (1.8), assume that the Lipschitz constant is such that L_f is small enough, and consider the discrete spaces to be of polynomial degree $k \geq 1$. Then,*

$$\|\varepsilon^u\|^2_{\Omega_h} \lesssim 3\,h\|\!|(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)\|\!|^2_{\boldsymbol{\sigma}_h} + 6\Big((h + L^2\,h^2)\Lambda^2_{\boldsymbol{q}} + (L^2 + h)\,\Lambda^2_u\Big). \tag{3.50}$$

Proof. Consider the pair of functions function ϕ and ψ satisfying the dual problem (1.5). We will use them to define the following terms

$$\begin{aligned}
\mathbb{T}_{\boldsymbol{\sigma}} &:= -(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\Pi_V}\boldsymbol{\phi} - \boldsymbol{\phi})_{\mathcal{T}_h} + ((\kappa(\boldsymbol{\sigma}_h) - \kappa(\boldsymbol{\sigma}))(\boldsymbol{I}^{\boldsymbol{\sigma}} + \boldsymbol{\Pi_V}\boldsymbol{\sigma}), \nabla\psi)_{\mathcal{T}_h},\\
\mathbb{T}_{\boldsymbol{q}} &:= -(\boldsymbol{I}^{\boldsymbol{q}}, \nabla\psi)_{\mathcal{T}_h},\\
\mathbb{T}_f &:= (f(u) - f(u_h), \Pi_W\psi)_{\mathcal{T}_h},\\
\mathbb{T}_u &:= \langle \varepsilon^{\widehat{u}}, P_M(\boldsymbol{\phi}\cdot\boldsymbol{n})\rangle_{\Gamma_h} - \langle \boldsymbol{\varepsilon}^{\widehat{\boldsymbol{q}}}\cdot\boldsymbol{n}, \Pi_W\psi\rangle_{\Gamma_h}.
\end{aligned}$$

With all the definitions above and the equations in (1.5), it is possible to decompose the inner product between ε^u and the function Θ appearing as the source term of the dual problem in the form

$$(\varepsilon^u, \Theta)_{\mathcal{T}_h} = \mathbb{T}_{\boldsymbol{\sigma}} + \mathbb{T}_{\boldsymbol{q}} + \mathbb{T}_f + \mathbb{T}_u. \tag{3.51}$$

Following arguments similar to the ones applied in Lemma 3.7, it is possible to bound each of the terms in the decomposition as

$$\begin{aligned}
|\mathbb{T}_{\boldsymbol{\sigma}}| &\lesssim h^{\min\{1,k\}}\,\|\kappa^{1/2}(\boldsymbol{\sigma}_h)(\boldsymbol{\varepsilon}^\sigma + \boldsymbol{I}^{\boldsymbol{\sigma}})\|_{\Omega_h}\,\|\Theta\|_\Omega + L\,(\|\kappa^{1/2}(\boldsymbol{\sigma}_h)\boldsymbol{\varepsilon}^\sigma\| + \|\boldsymbol{I}^{\boldsymbol{\sigma}}\|)\,\|\Theta\|_\Omega,\\
|\mathbb{T}_{\boldsymbol{q}}| &\lesssim h^{\min\{1,k\}}\|\boldsymbol{I}\|_{\Omega_h}\,\|\Theta\|_\Omega,\\
|\mathbb{T}_f| &\lesssim L_f\,(\|\varepsilon^u\|_{\Omega_h} + \|I^u\|_{\Omega_h})\,\|\Theta\|_\Omega.
\end{aligned}$$

The bound for the final term in (3.51) requires decomposing it in the form $\mathbb{T}_u := \sum_{i=1}^{10}\mathbb{T}^i_u$, where:

$$\begin{aligned}
\mathbb{T}^1_u &:= -\langle \kappa(\boldsymbol{\sigma}_h)\,l^{-1}\,(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)), \psi + l\partial_{\boldsymbol{n}}\psi\rangle_{\Gamma_h},\\
\mathbb{T}^2_u &:= \langle \kappa(u_h)(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h)), (P_M - Id_M)\partial_{\boldsymbol{n}}\psi\rangle_{\Gamma_h},\\
\mathbb{T}^3_u &:= \langle \delta_{\boldsymbol{I}^q}, \psi\rangle_{\Gamma_h},\\
\mathbb{T}^4_u &:= \langle \boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n}, (Id_M - P_M)\psi\rangle_{\Gamma_h},\\
\mathbb{T}^5_u &:= -\langle \tau P_M I^u, \psi\rangle_{\Gamma_h},\\
\mathbb{T}^6_u &:= \langle \delta_{\boldsymbol{\varepsilon}^{\boldsymbol{q}}}, \psi\rangle_{\Gamma_h},\\
\mathbb{T}^7_u &:= -\langle \tau(\varepsilon^u - \varepsilon^{\widehat{u}}), P_M\psi\rangle_{\Gamma_h},\\
\mathbb{T}^8_u &:= -\langle \kappa(\boldsymbol{\sigma}_h)\,(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h))\,\delta_{\boldsymbol{I}^{\boldsymbol{q}}}, \psi\rangle_{\Gamma_h}\\
\mathbb{T}^9_u &:= -\langle \kappa(\boldsymbol{\sigma}_h)\,(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h))\,\delta_{\boldsymbol{\Pi_V}\boldsymbol{q}}, \psi\rangle_{\Gamma_h},\\
\mathbb{T}^{10}_u &:= -\langle \kappa(\boldsymbol{\sigma}_h)\,(\varphi(\boldsymbol{\sigma}) - \varphi_h(\boldsymbol{\sigma}_h))\,\boldsymbol{I}^{\boldsymbol{q}}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}.\\
\mathbb{T}^{11}_u &:= \langle \kappa(\boldsymbol{\sigma}_h)(\varphi_h(\boldsymbol{\sigma}_h) - \varphi(\boldsymbol{\sigma}))\boldsymbol{\Pi_V}\boldsymbol{q}\cdot\boldsymbol{n}, \psi\rangle_{\Gamma_h}.
\end{aligned}$$

These terms can be estimated by arguments like the ones detailed in Lemma 3.5. Finally, taking $\Theta = \varepsilon^u$ in (3.51) and considering the estimates for the components of $\mathbb{T}^i_u$ it is possible to deduce that

$$\|\varepsilon^u\|^2_{\Omega_h} \lesssim 3\,h\|\!|(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)\|\!|^2_{\boldsymbol{\sigma}_h} + 6\Big((h + L^2\,h^2)\Lambda^2_{\boldsymbol{q}} + (L^2 + h)\,\Lambda^2_u\Big).$$

■

The result of the previous Lemma allows us to estimate the error incurred by the HDG approximation by that of the HDG projection onto the discrete space, as we now show.

Theorem 3.4. *If L is small enough and the discrete spaces are of polynomial degree $k \geq 1$, then there exists $h_0 > 0$ such that, for all $h \leq h_0$, we have*

$$|||(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|||^2_{\boldsymbol{\sigma}_h} \lesssim \Lambda^2_{\boldsymbol{q}} + \Lambda^2_u + \Lambda^2_{\boldsymbol{\sigma}}. \tag{3.52}$$

Proof. Using simple algebraic arguments, note that the term (3.49) can be rewritten as

$$|||(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|||^2_{\boldsymbol{\sigma}} \lesssim \Lambda^2_{\boldsymbol{q}} + \Lambda^2_{\boldsymbol{\sigma}} + \frac{3}{2} L_f \|\varepsilon^u\|^2_{\Omega_h} + \frac{1}{2} L_f \, \Lambda_u \|\varepsilon^u\|_{\Omega_h}$$

Combined the above with the estimate given in the Lemma 3.8, we can deduce

$$\begin{aligned}\left(1 - \frac{9}{2} L_f \, h \, c\right) |||(\boldsymbol{\varepsilon}^\sigma, \boldsymbol{\varepsilon}^{\boldsymbol{q}}, \varepsilon^u - \varepsilon^{\widehat{u}}, \varphi - \varphi_h)|||^2_{\boldsymbol{\sigma}} \\ \lesssim 9 \, L_f \left((h + L^2 \, h^2) \Lambda^2_{\boldsymbol{q}} + (L^2 + h) \, \Lambda^2_u \right) + \Lambda^2_{\boldsymbol{q}} + \Lambda^2_{\boldsymbol{\sigma}} + \frac{1}{2} L_f \Lambda^2_u,\end{aligned}$$

where $c > 0$ is a constant independent of h arising from the symbol $\lesssim$. Assuming that L_f is small enough and considering that $h \leq h_0$, te proof is concluded. ■

As a consequence of this theorem, it follows that the HDG approximation of the linearized systems will indeed achieve optimal order of convergence with respect to the degree of the polynomial basis, provided that the true solutions are smooth enough.

Corollary 3.3. *Suppose that assumptions of Theorem 3.4 hold. If $u \in H^{k+1}(\Omega)$ and $\boldsymbol{q} \in \boldsymbol{H}^{k+1}(\Omega)$, then*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_\Omega + \|\boldsymbol{q} - \boldsymbol{q}_h\|_\Omega + \|u - u_h\|_\Omega \leq C h^{k+1} \left(|u|_{k+1,\Omega} + |\boldsymbol{q}|_{k+1,\Omega} + |\boldsymbol{\sigma}|_{k+1,\Omega} \right).$$

CHAPTER 4

HDG-BEM coupling for non-linear problems with curved boundaries

In this chapter we study an unfitted discretization scheme that couples HDG with the boundary element method (BEM) for the solution to a non-linear problem posed in an unbounded domain. The transfer of information between the non-touching grids is done via the method of transfer paths. The coupling is done using Costabel's symmetric approach [26]. However, the unfitted computational domain introduces a perturbation that breaks the symmetry of the scheme. We show that under a suitable local proximity condition on the grids, the influence of the perturbation vanishes as the mesh parameter tends to zero and, if the sources have small Lipschitz constant, the nonlinear discrete problem is well posed. This analysis constitutes a stepping stone towards the computational solution of a variant of the Grad–Shafranov problem known as the *free boundary problem*, where the location of the plasma is not known a priori and the equilibrium condition must be solved in free space in order to locate the plasma.

4.1 Introduction

In the reminder of this chapter we will follow the notations and definitions given in Sections 1.1 and 1.2. Consider a bounded domain $\Omega_0 \subset \mathbb{R}^d$ $(d = 2, 3)$ that is not necessarily polygonal/polyhedral, but has a Lipschitz boundary that will be denoted by $\Gamma_0 := \partial\Omega_0$. We will denote the unbounded complement of its closure by $\Omega_0^c =: \mathbb{R}^d \setminus \overline{\Omega_0}$. In this chapter, we will be concerned with the analysis of a discretization for following non-linear diffusion problem

$$\nabla \cdot \boldsymbol{q}^{\mathrm{tot}} = F(u^{\mathrm{tot}}) \qquad \text{in } \Omega_0^c, \tag{4.1a}$$

$$\boldsymbol{q}^{\mathrm{tot}} + \kappa \nabla u^{\mathrm{tot}} = 0 \qquad \text{in } \Omega_0^c, \tag{4.1b}$$

$$u^{\mathrm{tot}} = u_0 \qquad \text{on } \Gamma_0, \tag{4.1c}$$

$$\|u^{\mathrm{tot}}\| \to 0 \qquad \text{as } \boldsymbol{x} \to \infty. \tag{4.1d}$$

Above, the diffusion coefficient κ is a strictly positive constant. This problem is in fact motivated by an application in magnetic plasma confinement where a compactly supported plasma is surrounded by vacuum [5]. In that context, κ is related to the reciprocal of the magnetic permeability, which might

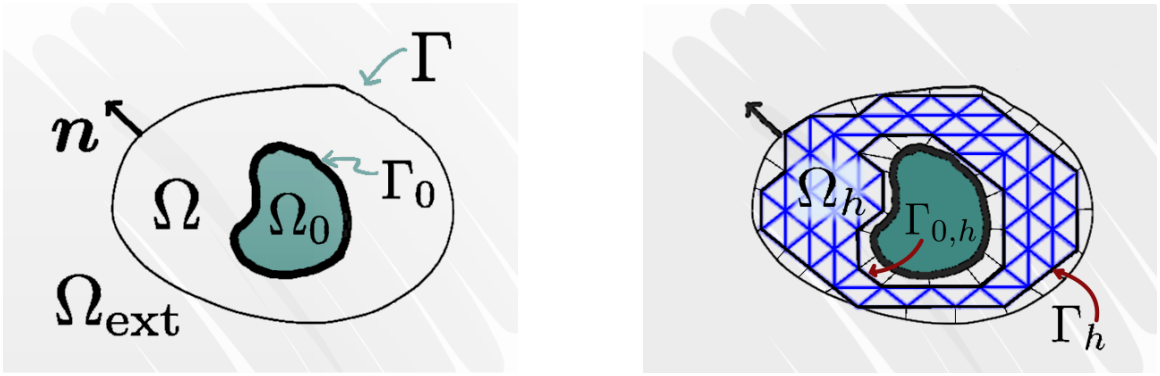

Figure 4.1: *Left: The artificial boundary Γ splits the domain of definition of Problem (4.1) into an unbounded region Ω_{ext} and a bounded annular domain Ω. Right: The computational domain Ω_h is discretized by an un-fitted triangulation (blue), with boundary $\Gamma_h \cup \Gamma_{0,h}$.*

vary with the position inside of the plasma and becomes a nonlinear function within ferromagnetic components of the reactor but is a constant in vacuum. The current work, with κ being a constant, is a first approximation to the aforementioned problem, which will be addressed in the near future.

The source term F depends on the solution as well; it will be assumed to be Lipschitz continuous as a function of u, and square integrable over Ω_0^c and compactly supported as a function of the space variables. More precisely, we will assume that there exists $L_F > 0$ such that

$$\|F(u_1) - F(u_2)\|_{\Omega_0^c} \leq L_F \|u_1 - u_2\|_{\Omega_0^c} \qquad \forall\, u_1, u_2 \in L^2(\Omega_0^c). \tag{4.2}$$

The Dirichlet boundary data u_0 will be considered to be an element of the trace space $H^{1/2}(\Gamma_0)$. The radiation condition at infinity (4.1d) is equivalent to assuming that there is a constant u_∞ such that $u = u_\infty + \mathcal{O}(|\boldsymbol{x}|^{-1})$ [57].

To deal with the unboundedness of the domain, we will make use of an integral representation that will reduce the computations to a bounded domain. To this avail, we introduce an artificial, smoothly parametrizable interface Γ enclosing Ω_0 and the support of F. We will denote the unit normal vector to Γ pointing in the direction of Ω_{ext} by $\boldsymbol{n}$ and will also require that $\Gamma \cap \Gamma_0 = \varnothing$. The bounded region interior to Γ and exterior to Γ_0 will be denoted by Ω, while the unbounded region exterior to Γ will be denoted by Ω_{ext}. This geometric decomposition, depicted in Figure 4.1, splits our region of interest into two disjoint domains and allows us to rewrite the problem (4.1) in terms of an interior and an exterior problem coupled by continuity conditions at the artificial boundary Γ. For reasons that will become clear later on, in the exterior domain we will prefer a second order formulation and will eliminate $\boldsymbol{q}$ from the system. This will lead to the representation of the solutions to (4.1) as the superposition

$$u^{\text{tot}} = u + u^{\text{ext}}, \quad \text{and} \quad \boldsymbol{q}^{\text{tot}} = \boldsymbol{q} + \kappa \nabla u^{\text{ext}}, \tag{4.3}$$

where the functions u and $\boldsymbol{q}$ are supported in Ω, while u^{ext} is supported in Ω_{ext}. The interior and exterior functions are coupled continuously at the artificial boundary Γ. The functions u and $\boldsymbol{q}$

appearing in the foregoing decomposition satisfy the interior problem

$$\nabla \cdot \boldsymbol{q} = F(u) \qquad \text{in } \Omega, \tag{4.4a}$$
$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega, \tag{4.4b}$$
$$u = g \qquad \text{on } \Gamma, \tag{4.4c}$$
$$\boldsymbol{q} \cdot \boldsymbol{n} = \lambda \qquad \text{on } \Gamma, \tag{4.4d}$$
$$u = u_0 \qquad \text{on } \Gamma_0. \tag{4.4e}$$

Above, the boundary value $g \in H^{1/2}(\Gamma)$ corresponds to the trace of u over the artificial boundary Γ, while $\lambda \in H^{-1/2}(\Gamma)$ is the value of the normal flux. These two functions are unknown at this point and will have to be retrieved as part the solution process.

On the other hand, the exterior function u^{ext} satisfies

$$-\nabla \cdot \left(\kappa \nabla u^{\text{ext}}\right) = 0 \qquad \text{in } \Omega_{\text{ext}}, \tag{4.5a}$$
$$u^{\text{ext}} = g \qquad \text{on } \Gamma, \tag{4.5b}$$
$$\kappa \nabla u^{\text{ext}} \cdot \boldsymbol{n} = -\lambda \qquad \text{on } \Gamma, \tag{4.5c}$$
$$u^{\text{ext}} \to 0 \qquad \text{as } |\boldsymbol{x}| \to \infty\ . \tag{4.5d}$$

Above, we made use of that, outside of Ω, the source F vanishes identically and of the fact that the normal vector $\boldsymbol{n}$ is *interior* to Ω_{ext}. Since the support of the nonlinear source term is contained entirely on Ω, the exterior boundary value problem above is in fact linear however, it has the additional challenge of being posed in an unbounded domain. In Section 4.3 we will see how to leverage the linearity to transform the problem into one defined uniquely on Γ; this will lead to a system of boundary integral equations.

The task is then to solve the interior boundary value problem (4.4) coupled to the exterior problem (4.5) through continuity conditions on the traces and normal fluxes on Γ. These conditions are reflected by the shared terms g and λ appearing in conditions (4.4c) and (4.5b), and (4.4d) and (4.5c) respectively. We will do so by, following the phrasing in [19], "coupling at a distance", meaning that the interior problem will indeed be posed over a polygonal subdomain $\Omega_h \subset \Omega$ and transferring data between the two problems following the method of *transfer paths* introduced in [18, 21]. The coupled system that we propose here will be different from the one used in [19] and will lead to an almost symmetric formulation, perturbed only by the transfer of data through the "gap" between Γ and Ω_h.

We will proceed as follows, Section 4.2 will detail how to deal with general *linear* problems of the form (4.4) posed on subdomains $\Omega_h \subset \Omega$ with polygonal boundaries. These problem will then be discretized through an augmented Hybridizable Discontinuous Galerkin (HDG) formulation. Next, in Section 4.3, we will introduce the basic elements of boundary integral equations (BIE) which will then be used to reformulate problems of the form (4.5) as systems of boundary integral equations. Having established the main results pertaining well posedness of said systems, the section will conclude by introducing a spectral discretization of the boundary integral equations, also known as a spectral boundary element method (BEM). Finally, in Section 4.4 we will return to the original non-linear problem (4.1) and will introduce an *almost symmetrical* coupled BEM-HDG discretization. We will then analyze the symmetry-braking perturbation introduced by the unfitted HDG scheme and show

that its norm is bounded by terms proportional to the mesh size. Combining this analysis with the results pretaining the HDG and BEM discretizations obtained in sections 4.2 and 4.3, we will then show that the *linearized* coupled BEM-HDG discretization is uniquely solvable if the discretization mesh is small enough. Finally, having shown that the linearized discrete system is well-posed, the non-linear problem will be tackled through a fixed-point argument under smallness conditions for the mesh size and the Lipschitz constant of the non-linear source term F.

4.2 An augmented HDG formulation for an interior problem

In this chapter we will step back and consider augmented HDG formulations for an interior boundary value model problem of the form

$$\nabla \cdot \boldsymbol{q} = f \qquad \text{in } \Omega, \tag{4.6a}$$
$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega, \tag{4.6b}$$
$$u = \xi_0 \qquad \text{on } \partial\Omega, \tag{4.6c}$$

posed over a bounded open domain Ω with Lipschitz continuous boundary $\partial\Omega$. The source term $f \in L^2(\Omega)$ and the Dirichlet boundary data $\xi_0 \in H^{1/2}(\partial\Omega)$.

In order to guarantee that the discrete HDG scheme arising from (4.6) is well-posed and derive the corresponding *a priori* error estimates, we propose to enrich the HDG formulation with two suitable equations. This augmented formulation will enable us to combine the powerful machinery of the Babuška–Brezzi theory. The use of these tools is certainly not new. In fact, they have been widely applied in the context of conforming mixed finite element method (see for instance [34] and [35], and the references therein). We also refer to [37,38] for an augmented HDG method for quasi-Newtonian Stokes flows. Even though the analysis of this interior problem has been done in [18] by considering the projection-based analysis in [16], we must resort to a different type of approach in order to analyze the discrete coupled problem by employing the tools for analyzing saddle-point schemes.

4.2.1 The Augmented HDG method

Given Dirichlet data $\xi_0 \in H^{1/2}(\partial\Omega)$, we will focus on a family of polygonal subdomains $\Omega_h \subset \Omega$, labeled by the parameter h, and for each Ω_h will consider an admissible triangulation $\mathcal{T}_h$ such that the pair $(\Omega_h, \mathcal{T}_h)$ satisfies the local proximity condition discussed in Section 1.1. In this setting, an HDG discretization of (4.6) seeks an approximation $(\boldsymbol{q}_h, u_h, \hat{u}_h) \in \boldsymbol{V}_h \times W_h \times M_h$ satisfying

$$(\kappa^{-1}\boldsymbol{q}_h, \boldsymbol{v})_{\mathcal{T}_h} - (u_h, \nabla_h \cdot \boldsymbol{v})_{\mathcal{T}_h} + \langle \hat{u}_h, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = 0, \tag{4.7a}$$
$$(\nabla \cdot \boldsymbol{q}_h, w)_{\mathcal{T}_h} + \langle \tau\, u_h, w\rangle_{\partial\mathcal{T}_h} - \langle \tau\, \hat{u}_h, w\rangle_{\partial\mathcal{T}_h} = (f, w)_{\mathcal{T}_h}, \tag{4.7b}$$
$$\langle \mu, \hat{\boldsymbol{q}}_h \cdot \boldsymbol{n}\rangle_{\partial\mathcal{T}_h \setminus \partial\Omega_h} = 0, \tag{4.7c}$$
$$\langle \hat{u}_h, \mu\rangle_{\partial\Omega_h} = \langle \varphi_0^{\boldsymbol{q}}, \mu\rangle_{\partial\Omega_h}, \tag{4.7d}$$

for any test $(\boldsymbol{v}, w, \lambda) \in \boldsymbol{V}_h \times W_h \times M_h$. Here, $\boldsymbol{V}_h, W_h$ and M_h are the finite dimensional spaces of piece-wise polynomial functions defined in (1.6) and we have abused the notation for the normal

vector $\boldsymbol{n}$ which, in this context, denotes the exterior normal to each mesh element. Following [21], the approximate boundary data on $\partial\Omega_h$ appearing on the right hand sides of (4.7d) is given by

$$\varphi_0^{\boldsymbol{q}}(\boldsymbol{x}) := \xi_0(\overline{\boldsymbol{x}}(\boldsymbol{x})) + \int_0^{l(\boldsymbol{x})} \kappa^{-1}\, \boldsymbol{q}_h(\boldsymbol{x}+\boldsymbol{t}s)\cdot\boldsymbol{t}\, ds \qquad \text{for } \boldsymbol{x}\in\partial\Omega_h \quad \text{and} \quad \overline{\boldsymbol{x}}\in\partial\Omega. \tag{4.8a}$$

The unit vector $\boldsymbol{t}$ appearing in the definition above is anchored on $\boldsymbol{x}$ and pointing in the direction of $\overline{\boldsymbol{x}}$. The transfer formula above requires a mapping

$$\begin{array}{rcl} \phi : \partial\Omega_h & \longrightarrow & \partial\Omega \\ \boldsymbol{x} & \longmapsto & \overline{\boldsymbol{x}} \end{array}, \tag{4.9}$$

associating a point $\overline{\boldsymbol{x}}\in\partial\Omega$ to every point $\boldsymbol{x}\in\partial\Omega_h$. As numerous numerical experiments have shown [21,22,73,74], the algorithm is robust with respect to the particular choice for this mapping, so long as distance between $\boldsymbol{x}$ and its corresponding $\overline{\boldsymbol{x}}$ remains comparable to the local mesh diameter. However, for a key argument used in Section 4.4.3, we will need to require from the mapping ϕ to be a diffeomorphism. With this condition, we will denote the image of an edge $e\in\partial\mathcal{T}_h$ under ϕ by $\Gamma_e := \phi(e)$.

The numerical flux in the normal direction $\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}$ is defined as

$$\hat{\boldsymbol{q}}_h\cdot\boldsymbol{n} = \boldsymbol{q}_h\cdot\boldsymbol{n} + \tau\,(u_h-\hat{u}_h) \qquad \text{on } \partial\mathcal{T}_h, \tag{4.10}$$

where τ is a non-negative stabilization operator, whose maximum will be denoted by $\overline{\tau}$. Throughout this analysis, we will assume that there exists a positive constant C_τ, such that $\tau\le C_\tau\, h$.

Note that, the terms $\langle\hat{u}_h,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h}$ and $\langle\tau\hat{u}_h,w\rangle_{\partial\mathcal{T}_h}$, given in (4.7a) and (4.7b), respectively, can be split into the contributions of the interior edges and of the boundary edges as

$$\langle\hat{u}_h,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h} = \langle\hat{u}_h,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} + \langle\varphi_0^{\boldsymbol{q}},\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\Omega_h}, \tag{4.11a}$$

$$\langle\tau\hat{u}_h,w\rangle_{\partial\mathcal{T}_h} = \langle\tau\,\hat{u}_h,w\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} + \langle\tau\varphi_0^{\boldsymbol{q}},w\rangle_{\partial\Omega_h}. \tag{4.11b}$$

Replacing now the numerical flux (4.10) in (4.7c), results in

$$\langle\mu,\boldsymbol{q}_h\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} + \langle\mu,\tau(u_h-\hat{u}_h)\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = 0. \tag{4.12}$$

As we mentioned above, in order to establish the unique solvability of the nonlinear problem (4.7), the HDG formulation will be augmented with the equilibrium equation

$$\rho(\nabla\cdot\boldsymbol{q}_h,\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} = \rho(f,\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} \qquad \forall\,\boldsymbol{v}\in\boldsymbol{V}_h, \tag{4.13}$$

where $\rho>0$ is a parameter whose value will be determined later on. Combining the estimates (4.11), (4.12) and (4.13), the HDG scheme (4.7) becomes:

Find $(\boldsymbol{q}_h, u_h, \hat{u}_h) \in \boldsymbol{V}_h \times W_h \times M_h$ such that

$$(\kappa^{-1}\boldsymbol{q}_h, \boldsymbol{v})_{\mathcal{T}_h} - (u_h, \nabla_h \cdot \boldsymbol{v})_{\mathcal{T}_h} + \langle \hat{u}_h, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = -\langle \varphi_0^{\boldsymbol{q}}, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\Omega_h}, \tag{4.14a}$$

$$(\nabla\cdot\boldsymbol{q}_h, w)_{\mathcal{T}_h} + \langle \tau\, u_h, w\rangle_{\partial\mathcal{T}_h} - \langle \tau\, \hat{u}_h, w\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = (f, w)_{\mathcal{T}_h} + \langle \tau\varphi_0^{\boldsymbol{q}}, w\rangle_{\partial\Omega_h}, \tag{4.14b}$$

$$\langle \mu, \boldsymbol{q}_h\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} + \langle \mu, \tau(u_h - \hat{u}_h)\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = 0, \tag{4.14c}$$

$$\rho(\nabla\cdot\boldsymbol{q}_h, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} = -\rho(f, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h}. \tag{4.14d}$$

for all $(\boldsymbol{v}, w, \mu) \in \boldsymbol{V}_h \times W_h \times M_h$. It is worth mentioning that the equivalent formulation (4.14) above serves only theoretical purposes. It will facilitate the forthcoming analysis but is not be used for the explicit numerical calculations for which (4.7) is better suited. The following section will be devoted to showing the well posedness of these systems.

4.2.2 Analysis of the augmented HDG method

In order to apply known results from functional analysis, we rewrite the numerical trace $\hat{u}_h$ in terms of averages and jumps. For this, we use the equation (4.14c) and separate the term featuring $\hat{u}_h$ as

$$\begin{aligned}
0 &= \langle \mu, \boldsymbol{q}_h\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} + \langle \mu, \tau u_h\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} - \langle \mu, \tau\hat{u}_h\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} \\
&= \sum_{T\in\mathcal{T}_h}\sum_{e\in\partial T\setminus\partial\Omega_h}\int_e (\mu\,\boldsymbol{q}_h\cdot\boldsymbol{n} + \tau\,\mu\,u_h - \tau\,\mu\,\hat{u}_h) \\
&= \sum_{e\in\mathcal{E}_h^\circ}\int_e (\llbracket \boldsymbol{q}_h\rrbracket\,\mu + 2\tau\,\{\!\{u_h\}\!\}\,\mu - 2\tau\,\hat{u}_h\,\mu) = \int_{\mathcal{E}_h^\circ} (\llbracket \boldsymbol{q}_h\rrbracket + 2\,\tau\,\{\!\{u_h\}\!\} - 2\,\tau\,\hat{u}_h)\,\mu \qquad \forall\,\mu\in M_h.
\end{aligned}$$

Above, the average $\{\!\{\cdot\}\!\}$ and jump $\llbracket\cdot\rrbracket$ operators above are defined as in Section 1.1, and we have used the fact that the hybrid variable $\hat{u}_h$ is single valued. Then, taking a test function $\mu = \llbracket \boldsymbol{q}_h\rrbracket + 2\,\tau\,\{\!\{u_h\}\!\} - 2\,\tau\,\hat{u}_h \in M_h$, we deduce that

$$\llbracket \boldsymbol{q}_h\rrbracket + 2\,\tau\,\{\!\{u_h\}\!\} - 2\,\tau\,\hat{u}_h = 0 \qquad \text{on } \mathcal{E}_h^\circ,$$

which yields

$$\hat{u}_h = \frac{1}{2}\tau^{-1}\llbracket \boldsymbol{q}_h\rrbracket + \{\!\{u_h\}\!\} \qquad \text{on } \mathcal{E}_h^\circ. \tag{4.15}$$

We now replace (4.15) in (4.14a) and (4.14b), to obtain

$$\langle \hat{u}_h, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = \sum_{T\in\mathcal{T}_h}\sum_{e\in\partial T\setminus\partial\Omega_h}\int_e \hat{u}_h\,\boldsymbol{v}\cdot\boldsymbol{n} = \int_{\mathcal{E}_h^\circ}\llbracket \boldsymbol{v}\rrbracket\,\hat{u}_h = \int_{\mathcal{E}_h^\circ}\left(\tfrac{1}{2}\tau^{-1}\llbracket \boldsymbol{q}_h\rrbracket\llbracket \boldsymbol{v}\rrbracket + \{\!\{u_h\}\!\}\llbracket \boldsymbol{v}\rrbracket\right), \tag{4.16}$$

$$\langle \tau\,\hat{u}_h, w\rangle_{\partial\mathcal{T}_h\setminus\partial\Omega_h} = \sum_{T\in\mathcal{T}_h}\sum_{e\in T\setminus\partial\Omega_h}\int_e \tau\,\hat{u}_h\,w = -2\int_{\mathcal{E}_h^\circ}\tau\{\!\{w\}\!\}\hat{u}_h = \int_{\mathcal{E}_h^\circ}(\llbracket \boldsymbol{q}_h\rrbracket\{\!\{w\}\!\} - 2\tau\{\!\{w\}\!\}\{\!\{u_h\}\!\}). \tag{4.17}$$

In this way, replacing the definition of $\varphi_0^{\boldsymbol{q}}$—see (4.8a)—in (4.14a) and (4.14b) together with the foregoing identities and the estimate (4.14d), we obtain that (4.14) is equivalent to finding $(\boldsymbol{q}_h, u_h) \in$

$\boldsymbol{V}_h \times W_h$ such that

$$\mathcal{A}_T(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{B}(\boldsymbol{v}, u_h) = \mathcal{F}_1(\boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{4.18a}$$
$$\mathcal{B}_T(\boldsymbol{q}_h, w) + \mathcal{B}(\boldsymbol{q}_h, w) - \mathcal{C}(u_h, w) = \mathcal{F}_2(w) \qquad \forall\, w \in W_h, \tag{4.18b}$$

where the bilinear forms $\mathcal{A}, \mathcal{A}_T : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{R}$, $\mathcal{B}, \mathcal{B}_T : \boldsymbol{V}_h \times W_h \to \mathbb{R}$, $\mathcal{C} : W_h \times W_h \to \mathbb{R}$, and the functionals $\mathcal{F}_1(\cdot) : \boldsymbol{V}_h \to \mathbb{R}$ and $\mathcal{F}_2(\cdot) : W_h \to \mathbb{R}$ are defined by

$$\mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v}) := (\kappa^{-1}\boldsymbol{q}_h, \boldsymbol{v})_{\mathcal{T}_h} + \frac{1}{2}\int_{\mathcal{E}_h^\circ} \tau^{-1}[\![\boldsymbol{q}_h]\!][\![\boldsymbol{v}]\!] + \rho(\nabla\cdot\boldsymbol{q}_h, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h}, \tag{4.19a}$$

$$\mathcal{A}_T(\boldsymbol{q}_h, \boldsymbol{v}) := \sum_{e\subset\Gamma_h}\int_e \left(\int_0^{l(\boldsymbol{x})} \kappa^{-1}\boldsymbol{q}_h(\boldsymbol{x}+\boldsymbol{t}s)\cdot\boldsymbol{t}\right)\boldsymbol{v}\cdot\boldsymbol{n}\, ds\, dS_{\boldsymbol{x}}, \tag{4.19b}$$

$$\mathcal{B}(\boldsymbol{q}_h, w) := -(w, \nabla\cdot\boldsymbol{q}_h)_{\mathcal{T}_h} + \int_{\mathcal{E}_h^\circ}[\![\boldsymbol{q}_h]\!]\{\!\{w\}\!\}, \tag{4.19c}$$

$$\mathcal{B}_T(\boldsymbol{q}_h, w) := \sum_{e\subset\Gamma_h}\int_e \tau\left(\int_0^{l(\boldsymbol{x})} \kappa^{-1}\boldsymbol{q}_h(\boldsymbol{x}+\boldsymbol{t}s)\cdot\boldsymbol{t}\right) w\, ds\, dS_{\boldsymbol{x}}, \tag{4.19d}$$

$$\mathcal{C}(u_h, w) := \langle \tau\, u_h, w\rangle_{\partial\mathcal{T}_h} + 2\int_{\mathcal{E}_h^\circ} \tau\,\{\!\{u_h\}\!\}\{\!\{w\}\!\}, \tag{4.19e}$$

$$\mathcal{F}_1(\boldsymbol{v}) := \rho(f, \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} - \langle \xi_0, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\Omega_h}, \tag{4.19f}$$
$$\mathcal{F}_2(w) := -(f, w)_{\mathcal{T}_h} - \langle \tau\,\xi_0, w\rangle_{\partial\Omega_h}. \tag{4.19g}$$

Note that if it were not for the the bilinear forms $\mathcal{A}_T$ and $\mathcal{B}_T$, the system (4.18) would have the standard form of an augmented formulation for a saddle point problem. These two additional terms, however, are due only to the transfer of boundary conditions from $\partial\Omega$ to $\partial\Omega_h$ (thus the subscript T in their definitions). It is natural then to interpret these forms as perturbations introduced by the transfer technique which—as we shall show in what follows—indeed vanish as $\partial\Omega_h \to \partial\Omega$. Indeed, if the boundaries $\partial\Omega$ and $\partial\Omega_h$ were equal, these two terms would vanish, due to the fact that, in that case, $l(\boldsymbol{x}) \equiv 0$. This will happen naturally when refinements are done along a sequence of admissible triangulations and computational domains—as defined in Section 1.1.

Due to the transfer of boundary conditions, (4.18) is a perturbation of the simpler problem of finding $(\boldsymbol{q}_h, u_h) \in \boldsymbol{V}_h \times W_h$ such that

$$\mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{B}(\boldsymbol{v}, u_h) = \mathcal{F}_1(\boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in \boldsymbol{V}_h, \tag{4.20a}$$
$$\mathcal{B}(\boldsymbol{q}_h, w) - \mathcal{C}(u_h, w) = \mathcal{F}_2(w) \qquad \forall\, w \in W_h. \tag{4.20b}$$

If we can first establish the well-posedness of this system and then control the size of the perturbation, the unique solvability of (4.18) will follow from the fact that the set of invertible operators over a Banach space is open. The first step can be deduced from the following abstract result derived from the Babŭska–Brezzi theorem whose proof can be found in [39, Lemma 3.2].

Theorem 4.1. *Let X and Y be Hilbert spaces and consider bilinear forms $A : X \times X \to \mathbb{R}$, B :*

$X \times Y \to \mathbb{R}$ *and* $C : Y \times Y \to \mathbb{R}$. *Suppose that* C *is positive semi-definite, that is*

$$C(y, y) \geq 0 \qquad \forall\, y \in Y,$$

and that there are positive constants α *and* β, *such that*

$$A(x, x) \geq \alpha \|x\|_X^2 \qquad \forall\, x \in X,$$

and

$$\sup_{\substack{x \in X \\ x \neq 0}} \frac{B(x, y)}{\|x\|_X} \geq \|y\|_Y \qquad \forall\, y \in Y.$$

Given the functionals $F : X \to \mathbb{R}$ *and* $G : Y \to \mathbb{R}$, *there is a unique* $(v, u) \in X \times Y$, *such that*

$$\begin{pmatrix} A & B \\ B^* & -C \end{pmatrix} \begin{pmatrix} v \\ u \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}.$$

In addition there is a constant $\hat{C} > 0$ *dependent of* $\|A\|, \|B\|, \alpha$ *and* β, *such that:*

$$\|v\|_X + \|u\|_Y \leq \hat{C}\, \left(\|F\|_{X'} + \|G\|_{Y'} \right).$$

The first hypothesis of the theorem described above requires that the bilinear form $\mathcal{C}$—defined in (4.19e)—to be a positive semi-definite operator, which is established in the following result.

Lemma 4.1. *Bilinear form* $\mathcal{C} : W_h \times W_h \to \mathbb{R}$ *defined by* (4.19e) *is positive semi-definite, that is,*

$$\mathcal{C}(w, w) \geq 0 \qquad \forall\, w \in W_h.$$

Proof. Thanks to the fact that τ is positive on $\mathcal{E}_h$, we have

$$\mathcal{C}(w, w) = \sum_{T \in \mathcal{T}_h} \sum_{e \subset \partial T} \int_e w^2 + 2 \int_{\mathcal{E}_h^\circ} \tau \, \{\!\{w\}\!\}^2 \geq 0 \qquad \forall\, w \in W_h.$$

■

In what follows it will be proved that the bilinear forms $\mathcal{A}$ and $\mathcal{B}$ satisfy the hypotheses of the Theorem 4.1, however some preliminary results are required to guarantee this. We begin by recalling the discrete trace inequality (cf. Lemma 1.46 in [29]).

Lemma 4.2. *Let* $T \in \mathcal{T}$ *and* $e \subset \partial T$. *There exists a positive constant* C, *independent of* h, *such that*

$$\|h_e^{1/2} \boldsymbol{v}\|_{0,e} \leq C \|\boldsymbol{v}\|_{0,T}.$$

The parameter τ introduced in (4.10) will play a key role in the solvability analysis of the fixed point problem associated to the fully coupled problem that will be shown later. This is reflected in the definition of the following norm onto $\boldsymbol{V}_h$ which will be used to guarantee the ellipticity of the bilinear

form $\mathcal{A}$—defined in (4.19).

$$\|\boldsymbol{v}\|_{1,h};=\left(\|\boldsymbol{v}\|_{0,\Omega_h}^2+\|\nabla\cdot\boldsymbol{v}\|_{0,\Omega_h}^2+\|\tau^{-1/2}[\![\boldsymbol{v}]\!]\|_{0,\mathcal{E}_h^\circ}^2\right)^{1/2}. \tag{4.21}$$

Having established a norm on $\boldsymbol{V}_h$, the ellipticity of the bilinear forms $\mathcal{A}$ and is proved in the following result, where we also show that the perturbed form $(\mathcal{A}+\mathcal{A}_T)$ is also elliptic under suitable conditions of proximity between the boundaries.

Lemma 4.3. *There exists $\alpha_{\mathcal{A}}>0$, independent of h, such that*

$$\mathcal{A}(\boldsymbol{v},\boldsymbol{v})\geq\alpha_{\mathcal{A}}\,\|\boldsymbol{v}\|_{1,h}^2\qquad\forall\,\boldsymbol{v}\in\boldsymbol{V}_h.$$

Moreover, if ρ and R satisfy

$$\widehat{C}\,\underline{\kappa}^{-1}R_h^{1/2}\leq\rho\leq\max\left\{\frac{1}{2},\overline{\kappa}^{-1}\right\}, \tag{4.22}$$

for a positive fixed constant $\widehat{C}$, independent of h, appearing in the proof, we have

$$\mathcal{A}(\boldsymbol{v},\boldsymbol{v})+\mathcal{A}_T(\boldsymbol{v},\boldsymbol{v})\geq(\alpha_{\mathcal{A}}-\alpha_{\mathcal{A}_T})\|\boldsymbol{v}\|_{1,h}^2,$$

where $\alpha_{\mathcal{A}_T}\to 0$ as $h\to 0$.

Proof. Given $\boldsymbol{\sigma},\boldsymbol{v}\in\boldsymbol{V}_h$, let first focus on the bilinear form $\mathcal{A}$ defined on (4.19a). Taking $\boldsymbol{\sigma}=\boldsymbol{v}$, we obtain

$$\begin{aligned}\mathcal{A}(\boldsymbol{v},\boldsymbol{v})&=(\kappa^{-1}\boldsymbol{v},\boldsymbol{v})_{\mathcal{T}_h}+\frac{1}{2}\int_{\mathcal{E}_h^\circ}\tau^{-1}[\![\boldsymbol{v}]\!]^2+\rho(\nabla\cdot\boldsymbol{v},\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h}\\&=\|\kappa^{-1/2}\boldsymbol{v}\|_{0,\Omega_h}^2+\frac{1}{2}\|\tau^{-1/2}[\![\boldsymbol{v}]\!]\|_{0,\mathcal{E}_h^\circ}^2+\rho\|\nabla\cdot\boldsymbol{v}\|_{0,\Omega_h}^2\\&\geq\min\left\{\overline{\kappa}^{-1},1/2,\rho\right\}\|\boldsymbol{v}\|_{1,h}^2.\end{aligned}$$

So, defining $\alpha_{\mathcal{A}}:=\min\left\{\overline{\kappa}^{-1},1/2,\rho\right\}$ it follows that

$$\mathcal{A}(\boldsymbol{v},\boldsymbol{v})\geq\alpha_{\mathcal{A}}\|\boldsymbol{v}\|_{1,h}^2. \tag{4.23}$$

Now, we verify that $\mathcal{A}_T$ is bounded. By the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}|\mathcal{A}_T(\boldsymbol{\sigma},\boldsymbol{v})|\leq&\underline{\kappa}^{-1}\sum_{e\subset\Gamma_h}\int_e l(\boldsymbol{x})^{1/2}\left(\int_0^{l(\boldsymbol{x})}|\boldsymbol{\sigma}(\boldsymbol{x}+\boldsymbol{t}s)|^2ds\right)^{1/2}\boldsymbol{v}\cdot\boldsymbol{n}\,dS_{\boldsymbol{x}}\\\leq&\underline{\kappa}^{-1}\sum_{e\subset\Gamma_h}|\!|\!|\boldsymbol{\sigma}|\!|\!|_e\,\|l^{1/2}\,\boldsymbol{v}\cdot\boldsymbol{n}\|_{0,e}\\\leq&\underline{\kappa}^{-1}\sum_{e\subset\Gamma_h}|\!|\!|\boldsymbol{\sigma}|\!|\!|_e\,\|l^{1/2}\,\boldsymbol{v}\cdot\boldsymbol{n}\|_{0,e},\end{aligned}$$

Since $l(\boldsymbol{x}) \leq H_e^{\perp} = r_e h_e^{\perp} \lesssim R_h h_e$, for all $\boldsymbol{x} \in e$ and $e \subset \Gamma_h$, we deduce that

$$|\mathcal{A}_T(\boldsymbol{\sigma}, \boldsymbol{v})| \leq \underline{\kappa}^{-1} R_h^{1/2} \sum_{e \subset \Gamma_h} |||\boldsymbol{\sigma}|||_e \, \|h_e^{1/2} \, \boldsymbol{v} \cdot \boldsymbol{n}\|_{0,e},$$

where

$$|||\boldsymbol{\sigma}|||_e := \left(\int_e \int_0^{l(\boldsymbol{x})} |\boldsymbol{\sigma}(\boldsymbol{x} + s\boldsymbol{t}(\boldsymbol{x}))|^2 \, ds \, dS_{\boldsymbol{x}} \right)^{1/2}. \tag{4.24}$$

By Lemma 3.4 in [64] (two dimensions) and Lemma A.1 in [65] (three dimensions), we have that the $|||\cdot|||_e$ norm is equivalent to the $L^2(T_{ext}^e)$-norm under certain conditions on the transferring segments associated to the vertices of the triangulation. Then, we deduce that there exists a constant $\hat{c} > 0$, independent of h, such that

$$|\mathcal{A}_T(\boldsymbol{\sigma}, \boldsymbol{v})| \leq \hat{c} \, \underline{\kappa}^{-1} \sum_{e \subset \Gamma_h} \|\boldsymbol{\sigma}\|_{T_{ext}^e} \, \|l^{1/2} \, \boldsymbol{v} \cdot \boldsymbol{n}\|_{0,\partial\Omega_h}.$$

Hence, by Lemma 4.2, (1.4) and (4.21), we deduce that

$$|\mathcal{A}_T(\boldsymbol{\sigma}, \boldsymbol{v})| \leq \hat{c} \, C \, \underline{\kappa}^{-1} \max_{e \subset \Gamma_h} (r_e^{1/2} C_{\text{ext}}^e) \|\boldsymbol{\sigma}\|_{1,h} \, \|\boldsymbol{v}_h\|_{1,h}.$$

In other words, there exists a positive constant $\widehat{C}$, independent of h, such that

$$|\mathcal{A}_T(\boldsymbol{\sigma}, \boldsymbol{v})| \leq \alpha_{\mathcal{A}_T} \|\boldsymbol{\sigma}\|_{1,h} \, \|\boldsymbol{v}_h\|_{1,h}, \qquad \text{with} \quad \alpha_{\mathcal{A}_T} := \widehat{C} \, \underline{\kappa}^{-1} R_h^{1/2}. \tag{4.25}$$

Note that $R_h \to 0$ as $h \to 0$ then $\alpha_{\mathcal{A}_T} \to 0$ as h vanishes.

Finally, using the ellipticity of $\mathcal{A}$ (see (4.23)) and the continuity of $\mathcal{A}_T$, it follows that

$$\mathcal{A}(\boldsymbol{v}, \boldsymbol{v}) + \mathcal{A}_T(\boldsymbol{v}, \boldsymbol{v}) \geq \mathcal{A}(\boldsymbol{v}, \boldsymbol{v}) - |\mathcal{A}_T(\boldsymbol{v}, \boldsymbol{v})| \geq (\alpha_{\mathcal{A}} - \alpha_{\mathcal{A}_T}) \|\boldsymbol{v}\|_{1,h}^2.$$

This expression, together with the lower bound of ρ given in the assumption (4.22), implies the result. ■

We now proceed similarly to [37] to derive the discrete inf-sup condition for the bilinear form $\mathcal{B}$. It will be useful to recall here the definition of the Raviart–Thomas space of order k and dimension d defined over a domain $\mathcal{O}$ (cf. [67]))

$$\boldsymbol{RT}_k(\mathcal{O}) := [\mathbb{P}_k(\mathcal{O})]^d \oplus \boldsymbol{x}\widetilde{\mathbb{P}}_k(\mathcal{O}), \tag{4.26}$$

where $\widetilde{P}_k(\mathcal{O})$ is the space of polynomials of total degree equal to k defined on $\mathcal{O}$, $\boldsymbol{x} \in \mathbb{R}^d$ and, as usual, for $k \geq 0$, $\mathbb{P}_k(\cdot)$ denotes the one dimensional space of polynomials of degree at most k.

Lemma 4.4. *There exists a constant $\beta > 0$ such that*

$$\sup_{\substack{\boldsymbol{v} \in \boldsymbol{V}_h \\ \boldsymbol{v} \neq 0}} \frac{|\mathcal{B}(\boldsymbol{v}, w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq \beta \, \|w\|_{\Omega_h} \qquad \forall \, w \in W_h.$$

Moreover, there exists a constant $\beta_T \to 0$ as $h \to 0$, such that

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{|\mathcal{B}(\boldsymbol{v},w)+\mathcal{B}_T(\boldsymbol{v},w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq (\beta-\beta_T)\,\|w\|_{\Omega_h} \qquad \forall\, w\in W_h.$$

So that the perturbed bilinear form $(\mathcal{B}+\mathcal{B}_T)$ satisfies an inf-sup condition if the mesh is fine enough.

Proof. Given $\boldsymbol{v}\in\boldsymbol{V}_h$ and $w\in W_h$, we start by noticing that the Raviart–Thomás space of degree $k-1$ (defined as in (4.26)) belongs to the discrete space $\boldsymbol{V}_h$, from wich it follows that

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{|\mathcal{B}(\boldsymbol{v},w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq \sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{\displaystyle\int_{\Omega_h} w\,\nabla\cdot\boldsymbol{v} - \int_{\mathcal{E}_h^\circ}[\![\boldsymbol{v}]\!]\,\{\!\{w\}\!\}}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq \sup_{\substack{\boldsymbol{v}\in\boldsymbol{RT}_{k-1}(\Omega_h)\setminus 0\\ \int_{\Omega_h}\mathrm{tr}(\boldsymbol{v})=0}} \frac{\displaystyle\int_{\Omega_h} w\,\nabla\cdot\boldsymbol{v}}{\|\boldsymbol{v}\|_{1,\Omega_h}}.$$

Following the classic result from mixed finite element methods (see e.g. [33, Section 4.2 and Lemma 2.6]), we have

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{|\mathcal{B}(\boldsymbol{v},w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq \beta\,\|w\|_{\Omega_h}. \tag{4.27}$$

For $\mathcal{B}_T$ we use the same arguments as for $\mathcal{A}_T$ to conclude that there exists a positive constant $\widetilde{C}$, independent of the mesh size such that

$$|\mathcal{B}_T(\boldsymbol{v},w)| \leq \widetilde{C}\,\underline{\kappa}^{-1}R_h^{1/2}\overline{\tau}\|\boldsymbol{v}\|_{1,h}\|w\|_{0,\Omega_h}. \tag{4.28}$$

We define then

$$\beta_T := \widetilde{C}\,\underline{\kappa}^{-1}R_h^{1/2}\overline{\tau},$$

which, since τ vanishes as $h\to 0$, the influence of β_T will disappear as the mesh is refined. Then,

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{|\mathcal{B}_T(\boldsymbol{v},w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \leq \beta_T\,\|w\|_{0,\Omega_h}. \tag{4.29}$$

Finally, from (4.27) and (4.29) we have

$$\sup_{\substack{\boldsymbol{v}\in\boldsymbol{V}_h\\ \boldsymbol{v}\neq 0}} \frac{|\mathcal{B}(\boldsymbol{v},w)+\mathcal{B}_T(\boldsymbol{v},w)|}{\|\boldsymbol{v}\|_{1,\Omega_h}} \geq (\beta-\beta_T)\,\|w\|_{0,\Omega_h}.$$

Note that for h small enough, the coefficient $\beta-\beta_T>0$. ∎

Let us now discuss the stability properties of the forms $\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{F}_1$ and $\mathcal{F}_2$ which are established in the following results.

Lemma 4.5. *Let $\boldsymbol{q}_h,\boldsymbol{v}\in\boldsymbol{V}_h$ and $u_h,w\in W_h$. The bilinear forms $\mathcal{A}$, and $\mathcal{B}$are continuous and there*

exist positive constants $C_{\mathcal{A}}$, $C_{\mathcal{B}}$, $C_{\mathcal{C}}$ *and* $C_{\mathcal{F}_2}$ *such that*

$$|\mathcal{A}(\boldsymbol{q}_h,\boldsymbol{v})+\mathcal{A}_T(\boldsymbol{q}_h,\boldsymbol{v})| \leq C_{\mathcal{A}}\,\|\boldsymbol{q}_h\|_{1,h}\|\boldsymbol{v}\|_{|,h}, \tag{4.30a}$$

$$|\mathcal{B}(\boldsymbol{q}_h,w)+\mathcal{B}_T(\boldsymbol{q}_h,w)| \leq C_{\mathcal{B}}\,\|w\|_{0,\Omega_h}\,\|\boldsymbol{q}_h\|_{1,h}, \tag{4.30b}$$

$$|\mathcal{C}(u_h,w)| \leq C_{\mathcal{C}}\,\|u_h\|_{0,\Omega_h}\,\|w\|_{0,\Omega_h}; \tag{4.30c}$$

and,

$$|\mathcal{F}_1(\boldsymbol{v})| \leq \max\{1,\rho\}\,\left(\|f\|_{0,\Omega_h}+\|\xi_0\|_{1/2,\partial\Omega_h}\right)\|\boldsymbol{v}\|_{1,h}, \tag{4.31}$$

$$|\mathcal{F}_2(w)| \leq C_{\mathcal{F}_2}\,\left(\|f\|_{0,\Omega_h}+\|\xi_0\|_{1/2,\partial\Omega_h}\right)\|w\|_{0,\Omega_h}. \tag{4.32}$$

Proof. For the first inequality we point out that the continuity bound for $\mathcal{A}_T$ was established in Lemma 4.3. Thus, it is enough to notice that

$$\begin{aligned}
|\mathcal{A}(\boldsymbol{q}_h,\boldsymbol{v})| &= (\kappa^{-1}\,\boldsymbol{q}_h,\boldsymbol{v})_{\mathcal{T}_h}+\frac{1}{2}\int_{\mathcal{E}_h^\circ}\tau^{-1/2}[\![\boldsymbol{q}_h]\!]\,\tau^{-1/2}[\![\boldsymbol{v}]\!]+\rho\,(\nabla\cdot\boldsymbol{q}_h,\nabla\cdot\boldsymbol{v})_{\mathcal{T}_h}\\
&\leq \underline{\kappa}^{-1}\,\|\boldsymbol{q}_h\|_{0,\Omega_h}\,\|\boldsymbol{v}\|_{0,\Omega_h}+\frac{1}{2}\,\|\tau^{-1/2}\,[\![\boldsymbol{q}_h]\!]\|_{\mathcal{E}_h^\circ}\,\|\tau^{-1/2}\,[\![\boldsymbol{v}]\!]\|_{\mathcal{E}_h^\circ}+\rho\,\|\nabla\cdot\boldsymbol{q}_h\|_{0,\Omega_h}\,\|\nabla\cdot\boldsymbol{v}\|_{0,\Omega_h}\\
&\leq \max\left\{\underline{\kappa}^{-1},\frac{1}{2},\rho\right\}\,\|\boldsymbol{q}_h\|_{1,h}\,\|\boldsymbol{v}\|_{1,h}.
\end{aligned}$$

Hence, combining this with the bound for $\mathcal{A}_T$ in (4.25), it follows that

$$|\mathcal{A}(\boldsymbol{q}_h,\boldsymbol{v})+\mathcal{A}_T(\boldsymbol{q}_h,\boldsymbol{v})| \leq C_{\mathcal{A}}\,\|\boldsymbol{q}_h\|_{1,h}\,\|\boldsymbol{v}\|_{1,h},$$

where, using the definition of $\alpha_{\mathcal{A}_T}$ given in (4.25), we defined $C_{\mathcal{A}} := \max\left\{\underline{\kappa}^{-1},\frac{1}{2},\rho\right\}+\alpha_{\mathcal{A}_T}>0$.

On the order hand, taking directly the definitions of $\mathcal{B}$, applying the Lemma 4.2 and considering the fact that τ is of order h, we have

$$\begin{aligned}
|\mathcal{B}(\boldsymbol{q}_h,w)| &\leq \|w\|_{0,\Omega_h}\,\|\nabla\cdot\boldsymbol{q}_h\|_{0,\Omega_h}+(\tau\,h^{-1})^{1/2}\,\|\tau^{-1/2}[\![\boldsymbol{q}_h]\!]\|_{\mathcal{E}_h^\circ}\|h^{1/2}\,\{\!\{w\}\!\}\|_{\mathcal{E}_h^\circ}\\
&\leq (1+C^2\,C_\tau)^{1/2}\,\|\boldsymbol{q}_h\|_{1,h}\,\|w\|_{0,\Omega_h},
\end{aligned}$$

which, combined with the estimation of $\mathcal{B}_T$ (see (4.29)) yields

$$|\mathcal{B}(\boldsymbol{q}_h,w)+\mathcal{B}_T(\boldsymbol{q}_h,w)| \leq C_{\mathcal{B}}\,\|\boldsymbol{q}_h\|_{1,h}\,\|w\|_{0,\Omega_h},$$

where $C_{\mathcal{B}} := 2\,c^2C_2\,C_3\,\overline{\tau}\,\underline{\kappa}^{-1}\,R_h^2+(1+C^2\,C_\tau)^{1/2}>0$.

Now, from Lemma 4.2 and the definition of $\mathcal{C}$ it follows

$$\begin{aligned}
|\mathcal{C}(u_h,w)| &= (\tau\,h^{-1}\,h^{1/2}\,u_h,h^{1/2}\,w)_{\Gamma_h}+2\int_{\mathcal{E}_h^\circ}\tau\,h^{-1}\,h^{1/2}\,\{\!\{u_h\}\!\}\,h^{1/2}\,\{\!\{w\}\!\}\\
&\leq \overline{\tau}\,h^{-1}\,\|h^{1/2}\,\{\!\{u_h\}\!\}\|_{0,\Gamma_h}\,\|h^{1/2}\,\{\!\{w\}\!\}\|_{0,\Gamma_h}+2\,\overline{\tau}\,h^{-1}\,\|h^{1/2}\,\{\!\{u_h\}\!\}\|_{0,\Gamma_h}\,\|h^{1/2}\,\{\!\{w\}\!\}\|_{0,\Gamma_h}\\
&\leq C^2\,\overline{\tau}\,h^{-1}\,\|u_h\|_{0,\Omega_h}\,\|w\|_{0,\Omega_h}+2C_1^2\,\overline{\tau}\,h^{-1}\,\|u_h\|_{0,\Omega_h}\,\|w\|_{0,\Omega_h}\\
&= 3\,C^2\,\overline{\tau}\,h^{-1}\,\|u_h\|_{0,\Omega_h}\,\|w\|_{0,\Omega_h}.
\end{aligned}$$

The estimation of $\mathcal{C}$ follows, remembering that τ is of order h.

Finally, the estimates of $\mathcal{F}_j$ with $j = 1, 2$, follow directly from Lemma 4.2 and the Cauchy-Schwarz inequality. ■

We are now ready to state the main result concerning the well-posedness of (4.18). For this, let us note that in Lemmas 4.1, 4.3, 4.4 and 4.5 it is shown that the bilinear forms $\mathcal{A}, \mathcal{B}, \mathcal{C}$ and the functionals $\mathcal{F}_j$ (for $j = 1, 2$) satisfy the hypotheses of the Theorem 4.1, which shows that the *unperturbed* problem (4.20) has a unique solution $(\boldsymbol{q}_h, u_h)$. However, since the space of invertible operators over a Banach space is open and, by the estimates (4.25) and (4.28), the norm of the perturbations $\mathcal{A}_T$ and $\mathcal{B}_T$ vanishes as $h \to 0$, we can conclude that if the mesh is fine enough, then the problem (4.18) is well posed. Moreover, there is a constant $C_{HDG} > 0$, independent of h, such that

$$\|\boldsymbol{q}_h\|_{1,h} + \|u_h\|_{0,\Omega_h} \leq C_{HDG} \left(\|f\|_{0,\Omega_h} + \|\xi_0\|_{1/2,\partial\Omega_h} \right). \tag{4.33}$$

4.3 A spectral BEM discretization for an exterior problem

The main motivation of considering this type of discretization is that the approximation provided by BEM makes use of trigonometric polynomials. Hence, the approximate solution converges to the exact solution spectrally fast. Moreover, a spectral BEM discretization is characterized by the simplicity of the corresponding equations. In order to exploit these features, the boundary Γ must be C^∞ and this is why we need to deal with a non-polyhedral domain.

4.3.1 Basic results from boundary integral equations

Before setting out to reformulate and analyze this problem, we need to introduce a few concepts from the boundary integral equation literature. The material presented in this subsection is standard machinery in the boundary integral equation literature. The reader is referred to classical sources like [48, 57] for further a comprehensive treatment or [47] for a concise overview.

Being linear, equation (4.5a), has an associated Green function $G(\boldsymbol{x}, \boldsymbol{y})$, given by

$$G(\boldsymbol{x}, \boldsymbol{y}) := \begin{cases} -\frac{\kappa}{2\pi} \log(|\boldsymbol{x} - \boldsymbol{y}|) & \text{if } d = 2 \\ \frac{\kappa}{4\pi} \frac{1}{|\boldsymbol{x}-\boldsymbol{y}|} & \text{if } d = 3 \end{cases},$$

that can be used to transfer the unbounded transmission problem into one posed solely over Γ. Green's representation theorem guarantees that u has a representation of the form

$$u(\boldsymbol{x}) = \mathcal{S}[\![\kappa u]\!]_{\text{Neu}} - \mathcal{D}[\![u]\!]_{\text{Dir}} \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^d \setminus \Gamma, \tag{4.34}$$

and the Dirichlet and Neumann *jump* operators are defined for almost every $\overline{\boldsymbol{x}} \in \Gamma$ as

$$[\![u]\!]_{\text{Dir}}(\overline{\boldsymbol{x}}) := \lim_{\epsilon \to 0} \left(u(\overline{\boldsymbol{x}} - \epsilon \boldsymbol{n}) - u(\overline{\boldsymbol{x}} + \epsilon \boldsymbol{n}) \right),$$
$$[\![u]\!]_{\text{Neu}}(\overline{\boldsymbol{x}}) := \lim_{\epsilon \to 0} \left(\nabla u(\overline{\boldsymbol{x}} - \epsilon \boldsymbol{n}) \cdot \boldsymbol{n} - \nabla u(\overline{\boldsymbol{x}} + \epsilon \boldsymbol{n}) \cdot \boldsymbol{n} \right).$$

The operators $\mathcal{D}$ and $\mathcal{S}$ are known respectively as *double layer potential* and *single layer potential*. They are defined respectively for any $\phi, \lambda \in L^2(\Gamma)$ by

$$\mathcal{D}\phi(\boldsymbol{x}) := \int_\Gamma \partial_{\boldsymbol{n}(\boldsymbol{y})} G(\boldsymbol{x},\boldsymbol{y})\, \phi(\boldsymbol{y})\, d\Gamma(\boldsymbol{y}) \quad \text{and} \quad \mathcal{S}\lambda(\boldsymbol{x}) := \int_\Gamma G(\boldsymbol{x},\boldsymbol{y})\, \lambda(\boldsymbol{y})\, d\Gamma(y).$$

The interior and exterior traces and normal derivatives of the functions defined through the single and double layer potential motivate the definition of the following integral operators for all ϕ in the trace space $\in H^{1/2}(\Gamma)$ and every λ in its dual space $H^{-1/2}(\Gamma)$:

$$\mathcal{V}\lambda := \tfrac{1}{2}\left(\gamma^-\mathcal{S}\lambda + \gamma^+\mathcal{S}\lambda\right) = \gamma^-\mathcal{S}\lambda = \gamma^+\mathcal{S}\lambda \qquad \text{(Single layer operator)}, \tag{4.35a}$$
$$\mathcal{K}\phi := \tfrac{1}{2}\left(\gamma^-\mathcal{D}\phi + \gamma^+\mathcal{D}\phi\right) \qquad \text{(Double layer operator)}, \tag{4.35b}$$
$$\mathcal{K}^t\lambda := \tfrac{1}{2}\left(\partial_{\boldsymbol{n}}^-\mathcal{S}\lambda + \partial_{\boldsymbol{n}}^+\mathcal{S}\lambda\right) \qquad \text{(Weakly singular operator)}, \tag{4.35c}$$
$$\mathcal{W}\phi := -\tfrac{1}{2}\left(\partial_{\boldsymbol{n}}^-\mathcal{D}\phi + \partial_{\boldsymbol{n}}^+\mathcal{D}\phi\right) = -\partial_{\boldsymbol{n}}^-\mathcal{D}\phi = -\partial_{\boldsymbol{n}}^+\mathcal{D}\phi \qquad \text{(Hypersingular operator)}. \tag{4.35d}$$

Moreover, the following *jump identities* hold

$$\partial_{\boldsymbol{n}}^\pm\mathcal{S}\lambda = \left(\mp\tfrac{1}{2} + \mathcal{K}^t\right)\lambda \qquad \forall \lambda \in H^{-1/2}(\Gamma), \tag{4.36a}$$
$$\gamma^\pm\mathcal{D}\phi = \left(\pm\tfrac{1}{2} + \mathcal{K}\right)\phi \qquad \forall \phi \in H^{1/2}(\Gamma). \tag{4.36b}$$

Above, the operators γ^+ (resp.γ^-) denote the restriction of a function defined over Ω^+ (resp. Ω^-) to the boundary Γ. The operators defined above and their one sided traces and normal derivatives define mappings between the spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$. The following well known result establishes their continuity.

Theorem 4.2. *The following mappings are continuous*

$$\mathcal{V} : H^{-1/2}(\Gamma) \longrightarrow H^{1/2}(\Gamma), \qquad \left(\pm\tfrac{1}{2} + \mathcal{K}^t\right) : H^{-1/2}(\Gamma) \longrightarrow H^{-1/2}(\Gamma),$$
$$\mathcal{W} : H^{1/2}(\Gamma) \longrightarrow H^{-1/2}(\Gamma), \qquad \left(\pm\tfrac{1}{2} + \mathcal{K}\right) : H^{1/2}(\Gamma) \longrightarrow H^{1/2}(\Gamma).$$

Finally, denoting by $\langle\cdot,\cdot\rangle$ the duality pairing between the space $H^{1/2}(\Gamma)$ and its dual $H^{-1/2}(\Gamma)$, we define the spaces

$$H_0^{1/2}(\Gamma) = \left\{\phi \in H^{1/2}(\Gamma) : \langle\phi, 1\rangle = 0\right\}, \quad \text{and} \quad H_0^{-1/2}(\Gamma) = \left\{\lambda \in H^{-1/2}(\Gamma) : \langle\mu, 1\rangle = 0\right\},$$

we can state the following coercivity estimates:

Theorem 4.3. *There exist positive constants $\alpha_{\mathcal{V}}$, $\alpha_{\mathcal{W}}$, $\alpha_{\mathcal{K}^t}$, and $\alpha_{\mathcal{K}}$ such that*

$$\alpha_{\mathcal{V}}\|\mu\|_{-1/2}^2 \le |\langle\mathcal{V}\mu,\mu\rangle| \qquad \forall \mu \in H_0^{-1/2}(\Gamma),$$
$$\alpha_W\|\phi\|_{1/2}^2 \le |\langle\mathcal{W}\phi,\phi\rangle| \qquad \forall \phi \in H_0^{1/2}(\Gamma),$$
$$c_{\mathcal{K}^t}\|\mu\|_{-1/2}^2 \le |\langle\left(\pm\tfrac{1}{2} + \mathcal{K}^t\right)\mu,\mu\rangle| \qquad \forall \lambda \in H_0^{-1/2}(\Gamma),$$
$$c_{\mathcal{K}}\|\phi\|_{1/2}^2 \le |\langle\left(\pm\tfrac{1}{2} + \mathcal{K}\right)\phi,\phi\rangle| \qquad \forall \phi \in H_0^{1/2}(\Gamma).$$

4.3.2 Boundary integral reformulation

Going back to (4.1), we note that, since the diffusivity coefficient becomes constant and F vanishes in Ω_{ext}, equation (4.5a) from the exterior problem is in fact linear. We will exploit this fact to recast the problem in terms of boundary integral equations posed over the bounded interface Γ rather than over the unbounded domain Ω_{ext}. We will illustrate the process with the following model problem

$$-\nabla \cdot (\kappa \nabla u) = 0 \qquad \text{in } \Omega^+, \tag{4.37a}$$

$$-\kappa \partial_{\boldsymbol{n}}^+ u = \chi_0 \qquad \text{on } \Gamma, \tag{4.37b}$$

$$u \to 0 \qquad \text{as } |\boldsymbol{x}| \to \infty, \tag{4.37c}$$

where Γ is a closed and bounded Lipschitz $d-1$ dimensional hypersurface, $d \in \{1, 2\}$, dividing the space into a bounded region Ω^- and an unbounded region Ω^+; the unit normal vector to Γ pointing in the direction of Ω^- is denoted by $\boldsymbol{n}$, $\kappa > 0$ is a constant, and $\partial_{\boldsymbol{n}}^+$ denotes the exterior normal derivative.

We will now proceed to use the tools introduced in the preceding section to reformulate (4.37a) as a boundary integral equation. We will start by representing u in Ω^+ as

$$u(\boldsymbol{x}) = \mathcal{D}\phi - \mathcal{S}\lambda.$$

Applying the boundary integral operators (4.35) and the jump properties (4.36) to the integral representation above, we can re-write the boundary condition as

$$\partial_{\boldsymbol{n}}^+ u_{\text{ext}} = -\mathcal{W}\phi - \left(-\tfrac{1}{2} + \mathcal{K}^+\right)\lambda = \chi_0, \tag{4.38}$$

where $\phi \in H^{1/2}(\Gamma)$ and $\lambda \in H^{-1/2}(\Gamma)$ are unknown functions that need to be determined. Since there are two unknowns, a second equation must be provided in order to close the system. This second relation arises naturally if we extend u by zero into Ω^-. In particular, this implies that the *interior* trace must vanish. Once again, using the integral operators and the jump conditions, the condition $\gamma^- u = 0$ leads to the following boundary integral equation

$$\gamma^- u = \left(-\tfrac{1}{2} + \mathcal{K}\right)\phi - \mathcal{V}\lambda = 0. \tag{4.39}$$

In the following section we describe a spectral discretization of the weak formulation that arises by testing (4.38) with $\eta \in H^{1/2}(\Gamma)$ and (4.39) with $\mu \in H^{-1/2}(\Gamma)$.

4.3.3 Spectral BEM discretization

Let us now explain the spectral BEM discretization that will be employed on the interface Γ. We will discretize the integral equations (4.38) and (4.39) using a spectral method. Towards this goal, we first translate boundary integrals and functions to a fixed interval through parametric representation of Γ. Let $\boldsymbol{x} : \mathbb{R} \to \Gamma$ be a a twice continuously differentiable regular 2π-periodic parametrization of Γ:

$$|\boldsymbol{x}'(s)| > 0 \quad \forall\, s \in \mathbb{R} \qquad \text{and} \qquad \boldsymbol{x}(s) \neq \boldsymbol{x}(t), \qquad 0 < |s - t| < 2\,\pi.$$

Now, we introduce the spaces of trigonometric polynomials

$$\mathbb{T}_n := \left\{ \sum_{j=0}^{n} a_j \cos(jt) + \sum_{j=1}^{n-1} b_j \sin(jt) : a_j, b_j \in \mathbb{R} \right\} \quad \text{and} \quad \mathbb{T}_n^0 := \left\{ \lambda_n \in \mathbb{T}_n : \int_0^{2\pi} \lambda_n(s) ds = 0 \right\}.$$

We will consider the parametric representations of the integral operators $\mathcal{V}$, $\mathcal{K}$, $\mathcal{K}^t$ and $\widetilde{\mathcal{W}}$. The latter of these is an appropriate regularization of the hypersinglular operator $\mathcal{W}$, which can be constructed through tangential differentiation. For instance, if the parametrization of Γ has C^2 regularity it can be shown [75] that

$$\left(\widetilde{\mathcal{W}}\phi\right) := \tfrac{d}{d\tau}\left(\mathcal{V}\tfrac{d}{ds}\phi(s)\right)(\tau) = \left(\mathcal{W}\phi\right)(\tau).$$

Note that if we take $g_n : \Gamma \to \mathbb{R}$ such that $\mathbb{T}_n \ni g_n \circ \boldsymbol{x} = \left(g_n^0 + c_n\right) \circ \boldsymbol{x}$, where $g_n^0 \in \mathbb{T}_n^0$ and c_n is a constant, it follows that

$$\left(-\tfrac{1}{2} g_n + \mathcal{K} g_n\right) = \left(-\tfrac{1}{2} + \mathcal{K}\right) g_n^0 + \left(-\tfrac{1}{2} + \mathcal{K}\right) c_n = \left(-\tfrac{1}{2} + \mathcal{K}\right) g_n^0 + c_n \left(-\tfrac{1}{2} + \tfrac{1}{2}\right) = \left(\tfrac{1}{2} + \mathcal{K}\right) g_n^0.$$

Analogously, for $\mathbb{T}_n \ni \lambda_n \circ \boldsymbol{x} = \left(\lambda_n^0 + c_n\right) \circ \boldsymbol{x}$ with $\lambda_n^0 \in \mathbb{T}_n^0$ we have

$$\left(-\tfrac{1}{2} + \mathcal{K}^t\right) \lambda_n = \left(-\tfrac{1}{2} + \mathcal{K}^t\right) \lambda_n^0.$$

We can now reformulate the problem defined by equations (4.38) and (4.39) as that of finding $\lambda_n^0 : \Gamma \to \mathbb{R}$ and $g_n^0 : \Gamma \to \mathbb{R}$ such that $\lambda_n^0 \circ \boldsymbol{x}|\boldsymbol{x}'(\cdot)| \in \mathbb{T}_n^0$, $g_n^0 \circ \boldsymbol{x}|\boldsymbol{x}'(\cdot)| \in \mathbb{T}_n^0$ and

$$\begin{aligned}
&\int_0^{2\pi} (\mathcal{V}\lambda_n^0)(\boldsymbol{x}(s))\psi(s)ds - \int_0^{2\pi} \left(\frac{1}{2} + \mathcal{K}\right) g_n^0(\boldsymbol{x}(s))\psi(s)ds = 0, && \forall\, \psi \in \mathbb{T}_n^0,\\
&\int_0^{2\pi} \left(\frac{1}{2} + \mathcal{K}^t\right) \lambda_n^0(\boldsymbol{x}(s))\mu(s)ds + \int_0^{2\pi} (\mathcal{W} g_n^0)(\boldsymbol{x}(s))\mu(s)ds = -\int_0^{2\pi} \chi_0\, \mu(s)\, ds && \forall\, \mu \in \mathbb{T}_n^0.
\end{aligned}$$

Defining the bilinear forms $a, b, c, d : \mathbb{T}_n^0 \times \mathbb{T}_n^0 \to \mathbb{R}$ as

$$a(\lambda_n^0, \psi) := \int_0^{2\pi} (\mathcal{V}\lambda_n)(\boldsymbol{x}(s))\psi(s)ds, \tag{4.40a}$$

$$b(g_n^0, \psi) := \int_0^{2\pi} \left(-\tfrac{1}{2} + \mathcal{K}\right) g_n^0(\boldsymbol{x}(s))\psi(s)ds, \tag{4.40b}$$

$$c(\lambda_n^0, \mu) := \int_0^{2\pi} \left(-\tfrac{1}{2} + \mathcal{K}^t\right) \lambda_n^0(\boldsymbol{x}(s))\mu(s)ds, \tag{4.40c}$$

$$d(g_n^0, \mu) := \int_0^{2\pi} (\mathcal{W} g_n^0)(\boldsymbol{x}(s))\mu(s)ds, \tag{4.40d}$$

the weak formulation above can be expressed compactly as that of finding $(\lambda_n^0, g_n^0) \in \mathbb{T}_n^0 \times \mathbb{T}_n^0$ such that

$$a(\lambda_n^0, \psi) + b(g_n^0, \psi) + c(\lambda_n^0, \mu) + d(g_n^0, \mu) = -\int_0^{2\,\pi} \chi_0 \mu(s) ds \qquad \forall\, \psi \times \mu \in \mathbb{T}_n^0 \times \mathbb{T}_n^0. \tag{4.41}$$

The proof of the result below follows from using $(\psi, \mu) = (\lambda_n^0, g_n^0)$ as tests in (4.41), the linearity of the right hand side, the continuity and coercivity results in theorems 4.2 and 4.3, and a Lax–Milgram argument.

Theorem 4.4. *The variational problem* (4.41) *is uniquely solvable. Moreover, there exists a constant* $\alpha_{BEM} > 0$ *such that*

$$\|(\lambda_n^0, g_n^0)\|_{-1/2,1/2} \leq \frac{\|\chi_0|\|_{-1/2}}{\alpha_{BEM}},$$

where α_{BEM} *is the smallest of the constants* $\alpha_{\mathcal{V}}$, $\alpha_{\mathcal{W}}$, $\alpha_{\mathcal{K}^t}$, *and* $\alpha_{\mathcal{K}}$ *appearing in Theorem 4.3 and*

$$\|(\lambda_n^0, g_n^0)\|_{-1/2,1/2}^2 = \|\lambda_n^0\|_{-1/2}^2 + \|g_n^0\|_{1/2}^2.$$

Spectrally accurate approximations of the bilinear forms in (4.40) can be built by discretizing the parameter space $[0.2\pi)$ with N equispaced points and approximating the integrals by the trapezoidal rule, which is exponentially convergent due to the periodicity and smoothness of the parametrization [81].

4.4 A perturbed symmetrically coupled formulation

We are now in the position of addressing the original problem (4.1) for the interior functions $(\boldsymbol{q}, u)$ and the exterior function u^{ext}. The idea of a symmetrically coupled boundary-field formulation was introduced by Costabel in [26], where it was used to analyze a linear problem posed in a domain with polygonal boundaries, and geared towards a BEM-FEM discretization. This formulation gained prominence due to the fact that, until recently [77], it was considered that other celebrated alternative requiring only one boundary unknown (e.g. Johnson-Nédélec [49]) was not stable unless the domain has smooth boundaries. This technique that has been recently used successfully even in time domain problems [8, 45, 46, 72]. While requiring two boundary unknowns, Costabel's formulation has the advantage of being symmetric when the computational domains share a common boundary, which greatly simplifies the analysis. In this chapter we will exploit this latter advantage while in fact choosing a smooth interface Γ at the continuous level.

Some of the techniques of the discretization that we will use here can be traced back to [59], where the authors analyzed a coupled BEM-FEM scheme for a quasilinear elliptic boundary value problem. In that work the authors formulate the interior equation in second order form and deal with the curvature of the interior domain via a curved triangulation that interpolates the boundary $\Gamma \cup \Gamma_0$. For the BEM part of the problem, they eliminate the trace g from the system and discretize the integral operators using a spectral approach like the one we describe below. One of the first combination of spectral and finite element methods for exterior problems can be found in [60]. There, trigonometric polynomials were used to approximate the unknown in the artificial boundary and curved triangles were employed to fit the boundary. In order to avoid the use of curved triangles and thus simplify the construction of the mesh, the authors in [19] coupled the HDG method with the spectral method in [60] through the data transferring technique developed in [21], however no analysis was provided. The novelty of our work in this chapter is to develop the analysis of the discrete method in [19] and include the additional difficulty of handling a nonlinear source.

For the discrete interior problem we will consider an unfitted region $\Omega_h \subset \Omega$, as our computational domain, which will be triangulated by a shape-regular and admissible (as defined in Section 1.1) triangulation $\mathcal{T}_h$. The computational boundary $\partial\Omega_h$ can be split into *interior* and *exterior* components as $\partial\Omega_h = \Gamma_h \cup \Gamma_{h,0}$, where

$$\Gamma_h := \left\{ e \in \mathcal{E}_h^\partial : d(e,\Gamma) \leq d(e,\Gamma_0) \right\} \quad \text{and} \quad \Gamma_{h,0} := \left\{ e \in \mathcal{E}_h^\partial : d(e,\Gamma_0) < d(e,\Gamma) \right\}.$$

The computational domain and the geometric setting are depicted schematically in Figure 4.1.

4.4.1 A strong integro-differential formulation

With the geometric discretization introduced above, the *strong* forms of the problems (4.4) and (4.5) become

$$\nabla \cdot \boldsymbol{q} = F(u) \qquad \text{in } \Omega_h, \tag{4.42a}$$

$$\boldsymbol{q} + \kappa \nabla u = 0 \qquad \text{in } \Omega_h, \tag{4.42b}$$

$$u = g \circ \phi - \int_0^{\ell(\boldsymbol{x})} \kappa^{-1} \boldsymbol{q}(\phi(\boldsymbol{x}) + s\boldsymbol{t}) \cdot \boldsymbol{t}\, ds \qquad \text{on } \Gamma_h, \tag{4.42c}$$

$$u = u_0 \circ \phi - \int_0^{\ell(\boldsymbol{x})} \kappa^{-1} \boldsymbol{q}(\phi(\boldsymbol{x}) + s\boldsymbol{t}) \cdot \boldsymbol{t}\, ds \qquad \text{on } \Gamma_{0,h}, \tag{4.42d}$$

$$\boldsymbol{q} \cdot \boldsymbol{n} = \lambda \qquad \text{on } \Gamma, \tag{4.42e}$$

$$-\nabla \cdot \left(\kappa \nabla u^{\text{ext}}\right) = 0 \qquad \text{in } \Omega_{\text{ext}}, \tag{4.42f}$$

$$\kappa \nabla u^{\text{ext}} \cdot \boldsymbol{n} = -\lambda \qquad \text{on } \Gamma, \tag{4.42g}$$

$$u^{\text{ext}} = g \qquad \text{on } \Gamma, \tag{4.42h}$$

$$\gamma^- u^{\text{ext}} = 0 \qquad \text{on } \Gamma, \tag{4.42i}$$

$$u^{\text{ext}} \to 0 \qquad \text{as } |\boldsymbol{x}| \to \infty\,. \tag{4.42j}$$

Some clarifications about this system are in order. First of all, it is important to recall that the Dirichlet and Neumann traces g and λ (defined on Γ) are both problem unknowns that must be determined. The second point of note is the addition of the condition (4.42i) pertaining the *interior* trace of u^{ext}. As discussed in the previous section, this condition has been added to enforce that the extension of u^{ext} into Ω^- vanishes identically. This will ensure that the decompositions

$$u^{\text{tot}} = u^{\text{ext}} + u \quad \text{and} \quad \boldsymbol{q}^{\text{tot}} = \nabla u^{\text{ext}} + \boldsymbol{q}$$

accurately represent the solutions to (4.1) by forcing u^{ext} to be supported exclusively on Ω^{ext}. Finally, the reader will note that the exterior problem is still posed in Ω_{ext} and conditions for this problem are still prescribed over Γ, while the interior problem has now been posed in the subdomain Ω_h with polygonal boundary Γ_h. Moreover, the boundary conditions on u have been transferred from the boundary $\Gamma \cup \Gamma_0$ to the computational boundary $\Gamma_h \cup \Gamma_{h,0}$, but the one for the normal flux $\boldsymbol{q} \cdot \boldsymbol{n}$ *was*

not. In fact, we will exploit the continuity condition

$$-\kappa\nabla u^{\text{ext}}\cdot\boldsymbol{n}=\lambda=\boldsymbol{q}\cdot\boldsymbol{n}\qquad\text{on }\Gamma$$

to merge these equalities into the single, *coupled* condition

$$\boldsymbol{q}\cdot\boldsymbol{n}+\kappa\nabla u^{\text{ext}}\cdot\boldsymbol{n}=0\qquad\text{on }\Gamma.$$

This important detail will introduce a perturbation in the system that will be discussed in detail later. If we now represent the exterior solution in the form

$$u_{\text{ext}}(\boldsymbol{x})=\int_\Gamma\partial_{\boldsymbol{n}(\boldsymbol{y})}G(\boldsymbol{x},\boldsymbol{y})\,g-\int_\Gamma G(\boldsymbol{x},\boldsymbol{y})\,\lambda\,d\Gamma(\boldsymbol{y}),\tag{4.43}$$

and use the boundary integral operators introduced in Section 4.3 the conditions (4.42e) and (4.42i) become, respectively

$$\left(-\tfrac{1}{2}+\mathcal{K}^+\right)\lambda+\mathcal{W}g-\boldsymbol{q}\cdot\boldsymbol{n}=0,\qquad\text{and}\qquad\mathcal{V}\lambda-\left(-\tfrac{1}{2}+\mathcal{K}\right)g=0.$$

This equations, together with the integral representation (4.43), effectively replace equations (4.42e), (4.42f), (4.42g), (4.42h), (4.42i), and (4.42j). In this way, we finally arrive at the following coupled formulation

$$\begin{aligned}
\mathcal{V}\lambda-\left(-\tfrac{1}{2}+\mathcal{K}\right)g&=0 && \text{on }\Gamma, && (4.44a)\\
\left(-\tfrac{1}{2}+\mathcal{K}^+\right)\lambda+\mathcal{W}g-\boldsymbol{q}\cdot\boldsymbol{n}&=0, && \text{on }\Gamma, && (4.44b)\\
\boldsymbol{q}+\kappa\,\nabla u&=0 && \text{in }\Omega_h, && (4.44c)\\
\nabla\cdot\boldsymbol{q}&=F(u) && \text{in }\Omega_h, && (4.44d)\\
u&=g\circ\phi-\int_0^{\ell(\boldsymbol{x})}\kappa^{-1}\boldsymbol{q}(\phi(\boldsymbol{x})+s\boldsymbol{t})\cdot\boldsymbol{t}\,ds && \text{on }\Gamma_h, && (4.44e)\\
u&=u_0\circ\phi-\int_0^{\ell(\boldsymbol{x})}\kappa^{-1}\boldsymbol{q}(\phi(\boldsymbol{x})+s\boldsymbol{t})\cdot\boldsymbol{t}\,ds && \text{on }\Gamma_{0,h}, && (4.44f)\\
\rho\nabla\cdot\boldsymbol{q}&=\rho F(u) && \text{in }\Omega_h, && (4.44g)
\end{aligned}$$

for the unknowns $(\lambda,g,\boldsymbol{q},u)$. The final (seemingly redundant) equation of the system has been added to aid in the stabilization of the HDG discretization. In the following section we will pose this system weakly and will describe a discretization algorithm.

4.4.2 Discretizing the coupled system

We will approximate the solutions to the coupled system introduced in the previous section by coupling the HDG discretization described in Section 4.2 for the equations involving $(\boldsymbol{q},u)$, and by discretizing the weak form of the boundary integral operators acting on (λ,g) by following the technique described in Section 4.3.

Choosing test functions $(\eta,\psi,\boldsymbol{v},w)\in\mathbb{T}_n^0\times\mathbb{T}_n^0\times\boldsymbol{V}_h\times W_h$ and using the identities (4.16) and (4.17)

along with the definitions of the bilinear forms and linear functionals given in (4.19) and (4.40), the integro-differential system (4.44) introduced in the previous section can be posed weakly as

$$a(\lambda_n^0, \psi) + b(g_n^0, \psi) = 0, \tag{4.45a}$$
$$c(\lambda_n^0, \eta) + d(g_n^0, \eta) - \langle \boldsymbol{q}_h \cdot \boldsymbol{n}, \eta \rangle_\Gamma = 0 \tag{4.45b}$$
$$\langle g_n^0, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\Gamma_h} + \mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{A}_T(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{B}(\boldsymbol{v}, u_h) = \mathcal{F}_1(u_h; \boldsymbol{v}), \tag{4.45c}$$
$$\langle \tau g_n^0, w \rangle_{\Gamma_h} + \mathcal{B}(\boldsymbol{q}_h, w) + \mathcal{B}_T(\boldsymbol{q}_h, w) - \mathcal{C}(u_h, w) = \mathcal{F}_2(u_h; w), \tag{4.45d}$$

where we recall that, according to the notation set in Section 1.3, $\boldsymbol{q}_h$ evaluated in Γ should be understood as the piecewise local extrapolation of $\boldsymbol{q}_h$ from Ω_h to Γ.

The attentive reader will notice a few details about this system. Firstly, the last term in (4.45b) and the first term in (4.45c) are *almost* the transpose of each other, which should be expected from a symmetrical coupling like the one carried over in this work. The only difference between these two terms is the fact that the one arising from the boundary integral equation is integrated over Γ, while the one stemming from the HDG discretization is integrated over the computational boundary Γ_h. The second point is the presence of the term $\langle \tau g_n^0, w \rangle_{\Gamma_h}$ which has no counterpart on the second equation of the system which, once again, would be normal in a symmetric coupling. Next, is the presence of the term involving the bilinear form $\mathcal{B}_T$ in the final equation. This bilinear form, involving an integral over Γ_h, also lacks a counterpart in the third equation of the system and therefore breaks the expected symmetry of the discretization. Finally, and perhaps less obvious, there is the presence of the term involving $\mathcal{A}_T$ in the third equation; while this term does not break the symmetry of the system, it does have an impact on the ellipticity constant of the system.

The common element in all these symmetry-breaking terms is the presence of an integral over the computational boundary Γ_h. Indeed, all these terms arise from the transfer of boundary conditions and, as we shall now see, it is possible to rewrite the system above in terms of two separate groups of operators: one corresponding to a symmetrically coupled system discretized over grids aligned perfectly along the interface Γ, and another one accounting for the perturbation in the symmetric system introduced by the non-matching grids and the transfer of boundary conditions.

We will devote the reminder of this section to showing how to split the system into the aforementioned components and will then show in the following section that indeed, the geometric perturbation vanishes as the mesh is refined (pushing the computational interfaces closer and closer).

Let us begin with the term involving g in equation (4.45c). By adding and subtracting $\langle g, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_\Gamma$, we have that

$$\begin{aligned}\langle g(\overline{\boldsymbol{x}}(\cdot)), \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\Gamma_h} &= \langle g, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_\Gamma + \sum_{e \in \Gamma_h} \int_e g(\overline{\boldsymbol{x}}(\boldsymbol{x}))(\boldsymbol{v} \cdot \boldsymbol{n}_{\Gamma_h})(\boldsymbol{x}) ds_{\boldsymbol{x}} - \langle g, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_\Gamma \\ &= \langle g, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_\Gamma + \sum_{e \in \Gamma_h} \frac{|e|}{|\Gamma_e|} \int_{\Gamma_e} g(\overline{\boldsymbol{x}})(\boldsymbol{v} \cdot \boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) ds_{\overline{\boldsymbol{x}}} - \langle g, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_\Gamma.\end{aligned}$$

Here, $\boldsymbol{n}_{\Gamma_h}$ is the exterior unit normal vector on Γ_h. Since in this section we will be dealing with terms involving the two boundaries Γ and Γ_h, to avoid confusion we will also denote by $\boldsymbol{n}_\Gamma$ the normal vector on Γ. Moreover, we are using the notation $|e|$ and $|\Gamma_e|$ respectively to denote the $d-1$ dimensional

Lebsegue measure of the face e and of its image under ϕ, $\Gamma_e = \phi(e)$.

Adding and subtracting appropriate terms to the previous identity, we obtain that

$$\begin{aligned}\langle g(\overline{\boldsymbol{x}}(\cdot)), \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\Gamma_h} = \langle g, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\Gamma} + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e} g(\overline{\boldsymbol{x}})(\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}} - \sum_{e\in\Gamma_h}\int_{\Gamma_e} g(\overline{\boldsymbol{x}})(\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}}\\ + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e} g(\overline{\boldsymbol{x}})[(\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) - (\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}},\end{aligned}$$

or equivalently,

$$\begin{aligned}\langle g(\overline{\boldsymbol{x}}(\cdot)), \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\Gamma_h} = \quad \langle g, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\Gamma} + \sum_{e\in\Gamma_h}\left(\frac{|e|}{|\Gamma_e|}-1\right)\int_{\Gamma_e} g(\overline{\boldsymbol{x}})(\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}}\\ + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e} g(\overline{\boldsymbol{x}})[(\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) - (\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}}.\end{aligned} \tag{4.46}$$

A re-writing for the term involving g in (4.45d) can be derived by similar arguments, resulting in

$$\langle \tau\, g(\overline{\boldsymbol{x}}(\cdot)), w\rangle_{\Gamma_h} = \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e}\tau\, g(\overline{\boldsymbol{x}})\, w(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}} + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e}\tau\, g(\overline{\boldsymbol{x}})[w(\phi^{-1}(\overline{\boldsymbol{x}})) - w(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}}. \tag{4.47}$$

If we then define the bilinear forms $\mathcal{G} : \boldsymbol{V}_h \times \mathbb{T}^0_n \to \mathbb{R}$, $T_1 : \mathbb{T}^0_n \times \boldsymbol{V}_h \to \mathbb{R}$ and $T_2 : \mathbb{T}^0_n \times W_h \to \mathbb{R}$ as

$$\begin{aligned}\mathcal{G}(\boldsymbol{q}_h, \eta) &:= \langle \boldsymbol{q}_h\cdot\boldsymbol{n}, \eta\rangle_\Gamma,\\ T_1(g^0_n, \boldsymbol{v}) &:= \sum_{e\in\Gamma_h}\left(\frac{|e|}{|\Gamma_e|}-1\right)\int_{\Gamma_e} g^0_n(\overline{\boldsymbol{x}})(\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}}\\ &\quad + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e} g^0_n(\overline{\boldsymbol{x}})[(\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) - (\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}}, \qquad (4.48)\\ T_2(g^0_n, w) &:= \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e}\tau\, g^0_n(\overline{\boldsymbol{x}})\, w(\overline{\boldsymbol{x}})\,ds_{\overline{\boldsymbol{x}}} + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e}\tau\, g^0_n(\overline{\boldsymbol{x}})[w(\phi^{-1}(\overline{\boldsymbol{x}})) - w(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}}, \qquad (4.49)\end{aligned}$$

the problem (4.45) can be rewritten in matrix form, equivalently as

$$\left(\begin{pmatrix} a & b & 0 & 0\\ c & d & -\mathcal{G} & 0\\ 0 & \mathcal{G}^* & \mathcal{A} & \mathcal{B}\\ 0 & 0 & \mathcal{B}^* & -\mathcal{C}\end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & T_1 & \mathcal{A}_T & 0\\ 0 & T_2 & \mathcal{B}_T & 0\end{pmatrix}\right)\begin{pmatrix}\lambda^0_n\\ g^0_n\\ \boldsymbol{q}_h\\ u_h\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \mathcal{F}_1(u_h)\\ \mathcal{F}_2(u_h)\end{pmatrix}. \tag{4.50}$$

Above, in abuse of notation, we have used the same symbol to denote the bilinear forms and their associated operators (i.e. the operators mapping the components of the column vector into the first argument of the bilinear form with the same symbol). The first matrix on the left hand side corresponds to the formulation that would arise from the analysis of this problem if the two grids were to line up perfectly at the artificial interface, rendering $\Gamma = \Gamma_h$. The second matrix encodes the action of the geometric perturbation induced in the system by the gap between the computational domains and the subsequent transferal of data between the two boundaries. In the following section, we will show

that the unperturbed matrix is indeed continuous and coercive, and that if the two boundaries are sufficiently close (or equivalently if the mesh is sufficiently fine) the perturbation does not preclude the invertibility of a linearization of (4.50). Moreover, as the mesh is refined, all the terms in the perturbation vanish.

4.4.3 Well-posedness of a linearized formulation

In this section we will consider the source terms in the problem to be independent of the solution, which effectively linearizes the system. We will show under what conditions this linearization is in fact a well-posed problem in the sense of Haddamard. This results will be useful once we return to consider the non-linearity in the sources.

The matricial form of the problem given in (4.50) gives a clear roadmap of the steps required to show that the operator on the left hand side is invertible. To simplify the discussion, let us define the following formal matrices of operators (by identifying bilinear forms with their respective operators)

$$\mathcal{M}_D := \begin{pmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & \mathcal{A} & \mathcal{B} \\ 0 & 0 & \mathcal{B}^* & -\mathcal{C} \end{pmatrix}, \quad \mathcal{M}_G := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\mathcal{G} & 0 \\ 0 & \mathcal{G}^* & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{M}_T := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & T_1 & \mathcal{A}_T & 0 \\ 0 & T_2 & \mathcal{B}_T & 0 \end{pmatrix}.$$

The first step is to prove that the matrix of operators is continuous. Then, formally speaking, the second step would be to show that there exists a positive constant $\widetilde{\alpha}$ such that

$$\left(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h\right) \left(\mathcal{M}_D + \mathcal{M}_G + \mathcal{M}_T\right) \left(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h\right)^t \geq \widetilde{\alpha} \|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|_H^2.$$

As we shall show in the rest of this section, 1) the continuity and coercivity of the term $\mathcal{M}_D$ will follow readily from the analysis carried out in sections 4.2 and 4.3 for the interior and exterior problems, 2) due to the symmetric nature of the coupling, the term $\mathcal{M}_G$ will drop about of the coercivity analysis and will easily be shown to be continuous, and 3) for the perturbation component $\mathcal{M}_T$, the terms in the third column have already been shown in Section 4.2 to be continuous and with norm proportional to the mesh size, therefore their effect on the coercivity of the system will decrease as the mesh is refined. A similar statement holds true for the terms appearing in the second column of $\mathcal{M}_T$, as we will prove below. After establishing these facts, the invertibility the system will be proven at the end of the section. The following lemma establishes the continuity of the bilinear forms $\mathcal{G}$, T_1, and T_2.

Lemma 4.6. *The bilinear form $\mathcal{G}(\cdot,\cdot)$ is bounded, that is,*

$$|\mathcal{G}(g_n^0, \boldsymbol{v})| \leq \|g_n^0\|_{1/2,\Gamma} \|\boldsymbol{v}\|_{1,h}.$$

Moreover, if we assume that there exist non-negative constants C_1 and C_2, and positive parameters s_1 and s_2, such that

$$\max_{e \in \Gamma_h} \left| \frac{|e|}{|\Gamma_e|} - 1 \right| \leq C_1 R_h \qquad \text{and} \qquad \max_{\overline{\boldsymbol{x}} \in \Gamma_h} |\boldsymbol{n}_{\Gamma_h} - \boldsymbol{n}_\Gamma(\overline{\boldsymbol{x}})| \leq C_2 R_h, \tag{4.51}$$

then,

$$|T_1(g_n^0, \boldsymbol{v})| \lesssim R_h^{1/2}(R_h^{1/2} + h^{1/2})\|g_n^0\|_{1/2,\Gamma}\|\boldsymbol{v}\|_{1,h},$$
$$|T_2(g_n^0, w)| \lesssim \overline{\tau}(h^{-1/2} + (R_h h)^{1/2})\,\|g_n^0\|_{1/2,\Gamma}\|w\|_{0,\Omega_h}.$$

Proof. We started the proof, by applying trace inequality to term $\mathcal{G}_1$,

$$|\mathcal{G}(g_n^0, \boldsymbol{v})| = |\langle g_n^0, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_\Gamma| \leq \|g_n^0\|_{1/2,\Gamma}\,\|\boldsymbol{v}\|_{1,h},$$

Adding suitable terms to T_1, defined in (4.48), and the assumption (4.51), we have

$$\begin{aligned}
T_1(g_n^0, \boldsymbol{v}) &= \sum_{e\in\Gamma_h}\left(\frac{|e|}{|\Gamma_e|} - 1\right)\langle g_n^0, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_\Gamma + \sum_{e\in\Gamma_h}\frac{|e|}{|\Gamma_e|}\int_{\Gamma_e} g_n^0(\overline{\boldsymbol{x}})[(\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) - (\boldsymbol{v}\cdot\boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}} \\
&= \sum_{e\in\Gamma_h}\left(\frac{|e|}{|\Gamma_e|} - 1\right)\langle g_n^0, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_\Gamma + \sum_{e\in\Gamma_h}\int_{\Gamma_e} g_n^0(\overline{\boldsymbol{x}})[\boldsymbol{v}\cdot(\boldsymbol{n}_{\Gamma_h} - \boldsymbol{n}_\Gamma)(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}} \\
&\quad + \sum_{e\in\Gamma_h}\int_{\Gamma_e} g_n^0(\overline{\boldsymbol{x}})[(\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h})(\phi^{-1}(\overline{\boldsymbol{x}})) - (\boldsymbol{v}\cdot\boldsymbol{n}_{\Gamma_h}).(\overline{\boldsymbol{x}})]ds_{\overline{\boldsymbol{x}}} \\
&\leq C_1\,R_h\,\|g_n^0\|_{1/2,\Gamma}\,\|\boldsymbol{v}\|_{1,h} + C_2\,R_h\,\|g_n^0\|_{1/2,\Gamma}\|w\|_{0,\Omega_h} \\
&\quad + \|l^{1/2}g_n^0\|_{1/2,\Gamma}\|l^{-1/2}((\boldsymbol{v}\cdot\boldsymbol{n})\circ\phi^{-1} - \boldsymbol{v}\cdot)\boldsymbol{n}\|_\Gamma.
\end{aligned}$$

The term $\|l^{-1/2}(\boldsymbol{v}\cdot\boldsymbol{n}\circ\phi^{-1} - \boldsymbol{v}\cdot\boldsymbol{n}\|_\Gamma$ is estimated thanks to [62, Lemma 1]. Since $l(\boldsymbol{x})$ is bounded by $R_h h$, for all $\boldsymbol{x}\in\Gamma_h$, then

$$|T_1(g_n^0, \boldsymbol{v})| \leq (C_1\,R_h + C_2\,R_h + (R_h h)^{1/2})\|g_n^0\|_{1/2,\Gamma}\|\boldsymbol{v}\|_{1,h}.$$

We adapt the same technique used for the bound of T_1, and estimate the term T_2. ■

Before making use of this result, a few words about the assumptions (4.51) in the lemma are pertinent. The segment $\Gamma_e = \phi(e)$ is in fact a subset of the "true" boundary Γ. As such, the ratio $|e|/|\Gamma_e|$ is a measure of the quality of the geometric approximation of Ω given by Ω_h. For a sequence of admissible geometric discretizations $(\Omega_h, \mathcal{T}_h)$ this ratio and should approach the value 1 asymptotically as $h\to 0$, therefore the first assumption is natural. It may in fact be a consequence of the admissibility criterion of the meshes.

In a similar vein, the second assumption on the difference of normal vectors also pertains the quality of the geometric approximation of the computational domain. The hypothesis amounts to demanding that the normal vector of the computational domain lines up asymptotically with $\boldsymbol{n}_\Gamma$. This natural requirement does in fact exclude certain families of computational domains that would otherwise be admissible. However many families of geometric approximations remain valid. It seems plausible, although it remains to be proven, that by requiring for a family of computational domains Ω_h to be admissible that they satisfy the second hypothesis, the first one will also be satisfied due to the local proximity condition.

We will now proceed to prove the unique solvability of a linearization of (4.50). We will first define rigorously the bilinear form associated to the problem. To this end, we start by simplifying notation, and define the space $H := \mathbb{T}_n^0 \times \mathbb{T}_n^0 \times \boldsymbol{V}_h \times W_h$, and the bilinear form $\mathcal{M} : H\times H \to \mathbb{R}$ associated to

the weak formulation (4.45) as

$$\begin{aligned}\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w),(\psi, \eta, \boldsymbol{v}, w)) := \\ & a(\lambda_n^0, \psi) + b(g_n^0, \psi) + c(\lambda_n^0, \eta) + d(g_n^0, \eta) \\ & - \mathcal{G}(\eta, \boldsymbol{q}_h) + \mathcal{G}(g_n^0, \boldsymbol{v}) \\ & + \mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{A}_T(\boldsymbol{q}_h, \boldsymbol{v}) + \mathcal{B}(\boldsymbol{v}, u_h) + \mathcal{B}(\boldsymbol{q}_h, w) + \mathcal{B}_T(\boldsymbol{q}_h, w) - \mathcal{C}(u_h, w) \\ & + T_1(g_n^0, \boldsymbol{v}) + T_2(g_n^0, w). \end{aligned} \tag{4.52}$$

Then, for a continuous linear functional $\mathcal{F} \in H'$ we have the following.

Theorem 4.5. *If assumptions (4.22) holds true and the mesh is sufficiently fine, then the problem of finding $(\lambda_n^0, g_n^0, \boldsymbol{q}_h, w_u) \in H$ such that*

$$\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w), (\psi, \eta, \boldsymbol{v}, w)) = \mathcal{F}((\psi, \eta, \boldsymbol{v}, w)), \qquad \forall\, (\psi, \eta, \boldsymbol{v}, w) \in H. \tag{4.53}$$

is uniquely solvable, moreover, for the solution vector $(\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)$ it holds that

$$\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)\|_H \leq \frac{\|\mathcal{F}\|_{H'}}{(\alpha - \alpha_T)}. \tag{4.54}$$

Proof. The proof of this statements will follow a Lax–Milgram argument. To that end, we will first show that operator $\mathcal{M}$, defined in (4.52), is continuous. The boundedness of the bilinear forms $\mathcal{A}, \mathcal{A}_T, \mathcal{B}, \mathcal{B}_T$ and $\mathcal{C}$ had alread been proven in Lemma 4.5. As for the bilinear forms a, b, c, d arising from the BEM discretization, their continuity is given by Theorem 4.2. Finally $\mathcal{G}$, T_1, and T_2 are bounded using Lemma 4.6. Hence,

$$\begin{aligned} |\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w), (\psi, \eta, \boldsymbol{v}, w)| \quad \leq \quad & |a(\lambda_n^0, \psi))| + |b(g_n^0, \psi)| + |c(\lambda_n^0, \eta)| + |d(g_n^0, \eta)| \\ & + |\mathcal{G}(\eta, \boldsymbol{q}_h)| + |\mathcal{G}(g_n^0, \boldsymbol{v})| + |\mathcal{A}(\boldsymbol{q}_h, \boldsymbol{v})| + |\mathcal{A}_T(\boldsymbol{q}_h, \boldsymbol{v})| \\ & + |\mathcal{B}(\boldsymbol{v}, u_h)| + |\mathcal{B}(\boldsymbol{q}_h, w)| + |\mathcal{B}_T(\boldsymbol{q}_h, w)| + |\mathcal{C}(u_h, w)| \\ & + |T_1(g_n^0, \boldsymbol{v})| + |T_2(g_n^0, w)| \\ \leq \quad & C_{\mathcal{M}} \, \|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)\|_H \, \|(\psi, \eta, \boldsymbol{v}, w)\|_H, \end{aligned} \tag{4.55}$$

where $C_{\mathcal{M}} > 0$ depends only on $C_{\mathcal{A}}, C_{\mathcal{B}}, C_{BEM}$, and the continuity coefficients for T_1, T_2, $\mathcal{A}_T$, and $\mathcal{B}_T$. The latter four all vanish as $h \to 0$.

Let us now discuss the ellipticity of $\mathcal{M}$. We take $(\psi, \eta, \boldsymbol{v}, w) = (\lambda_n^0, g_n^0, \boldsymbol{q}_h, w) \in H$ as the second argument of $\mathcal{M}$ and, as a consequence of Lemma 4.3 and the analysis of the interior problem studied in Section 4.2.2, we establish the following result

$$\begin{aligned}\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w),(\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)) \geq & \alpha_{BEM} \|(\lambda_n^0, g_n^0)\|^2 + (\alpha_{HDG} - \max\{\alpha_{\mathcal{A}_T}, \beta_T\}) \|(\boldsymbol{q}_h, u_h)\|^2 \\ & - |T_1(g_n^0, \boldsymbol{q}_h)| - |T_2(g_n^0, u_h)|. \end{aligned} \tag{4.56}$$

Where $\alpha_{HDG} := \min\{\alpha_{\mathcal{A}}, \alpha_{\mathcal{B}}\}$ and α_{BEM} are the ellipticity constants for the *uncoupled* HDG and BEM discretizations.

On the other hand, the estimations for T_1 and T_2 given in the Lemma 4.6 yield to the existence of

a constant $C > 0$, independent of the meshsize, such that

$$|T_1(g_n^0, \boldsymbol{q}_h)| + |T_2(g_n^0, u_h)| \leq C\,\overline{\tau}(h^{-1/2} + R_h + (R_h h)^{1/2})\,\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|^2. \tag{4.57}$$

So, from (4.56) and (4.57), and by denoting $\alpha_\tau = C\overline{\tau}(h^{-1/2} + R_h + (R_h h)^{1/2})$, we deduce that

$$\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w), (\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)) \geq (\min\{\alpha_{HDG}, \alpha_{BEM}\} - \alpha_\tau)\,\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|_H^2. \tag{4.58}$$

Finally, applying the Lax–Milgram theorem, we deduce that, for a sufficiently fine mesh, (4.53) has a unique solution and there holds

$$\begin{aligned}(\min\{\alpha_{HDG}, \alpha_{BEM}\} - \alpha_\tau)\,\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|_H^2 &\leq \mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w), (\lambda_n^0, g_n^0, \boldsymbol{q}_h, w)) \\ &= \mathcal{F}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w))) \\ &\leq \|\mathcal{F}\|_{H'}\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|_H.\end{aligned}$$

From which we conclude that

$$\|(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h)\|_H \leq \frac{\|\mathcal{F}\|_{H'}}{(\alpha - \alpha_T)},$$

where we had defined

$$\alpha := \min\{\alpha_{HDG}, \alpha_{BEM}\} \quad \text{and} \quad \alpha_T := \max\{\alpha_\tau, \alpha_{\mathcal{A}_T}, \beta_T\}.$$

We point out that $\alpha_T \to 0$ as $h \to 0$. ■

4.4.4 The non-linear problem: A fixed-point approach

In the preceding sections we have slowly built all the machinery necessary to tackle the non linear coupled problem. We will do so by leveraging the Banach fixed-point theorem. We start by defining the functional $\mathcal{F}(z;\cdot) : H \to \mathbb{R}$, with $z \in W_h$ given by

$$\mathcal{F}(z;(\psi, \eta, \boldsymbol{v}, w)) := \mathcal{F}_1(z;\boldsymbol{v}) + \mathcal{F}_2(z;w). \tag{4.59}$$

Where $\mathcal{F}_1$ and $\mathcal{F}_2$ are defined in terms of the source function F appearing on the right hand side of (4.1a) as

$$\mathcal{F}_1(\boldsymbol{v}) := -\rho(F(z), \nabla\cdot\boldsymbol{v})_{\mathcal{T}_h} - \langle\xi_0, \boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial\Omega_h}, \quad \text{and} \quad \mathcal{F}_2(w) := -(F(z), w)_{\mathcal{T}_h} - \langle\tau\,\xi_0, w\rangle_{\partial\Omega_h}.$$

The continuity of $\mathcal{F}$ follows from Lemma 4.5, that is,

$$|\mathcal{F}(z;(\psi, \eta, \boldsymbol{v}, w))| \leq C_{\mathcal{F}}\,\left(\|F(z)\|_{0,\Omega_h} + \|u_0\|_{1/2,\partial\Omega_h}\right)\,\|(\psi, \eta, \boldsymbol{v}, w)\|_H, \tag{4.60}$$

where $C_{\mathcal{F}} := \max\{1, \rho, C_2, C_{\mathcal{F}_2}\} > 0$.

Now, we can define the operator $\mathcal{J}_h : W_h \to W_h$ as

$$\mathcal{J}_h(z) := u_h \qquad \forall\, z \in W_h, \tag{4.61}$$

where u_h is the fourth component of the solution $(\lambda_n^0, g_n^0, \boldsymbol{q}_h, u_h) \in H$ to the problem:

$$\mathcal{M}((\lambda_n^0, g_n^0, \boldsymbol{q}_h, w), (\psi, \eta, \boldsymbol{v}, w)) = \mathcal{F}(z; (\psi, \eta, \boldsymbol{v}, w)), \qquad \forall\, (\psi, \eta, \boldsymbol{v}, w) \in H.$$

It follows from Theorem 4.5 that the mapping $\mathcal{J}_h$ is well-defined. Thus, we realize that solving (4.45) is equivalent to finding $u_h \in W_h$ such that

$$\mathcal{J}_h(u_h) = u_h.$$

We can now prove the main result of the section: under suitable conditions, $\mathcal{J}_h$ has indeed a fixed point and thus, the discrete nonlinear coupled problem (4.45) has a unique solution.

Theorem 4.6. *Assume the same hypothesis of Theorem 4.5. In addition, if the Lipschitz constant of the source term, L_F, satisfies*

$$L_F < \frac{(\alpha - \alpha_T)}{(\rho + 1)},$$

then the mapping $\mathcal{J}_h$ has a unique fixed point and, equivalently, there exists a unique solution of (4.45).

Proof. Choose $z_1, z_2 \in W_h$ and define $u_j := \mathcal{J}_h(z_j)$ $(j = 1, 2)$ using the mapping $\mathcal{J}_h$ given in (4.61).

By construction, for all $(\psi, \eta, \boldsymbol{v}, w) \in H$, the function $u_1 - u_2$ is the fourth component of the solution to the problem

$$\begin{aligned}\mathcal{M}((\lambda_1 - \lambda_2, g_1 - g_2, \boldsymbol{q}_1 - \boldsymbol{q}_1, u_1 - u_2),(\psi, \eta, \boldsymbol{v}, w)) &= \mathcal{F}(z_1 - z_2; (\psi, \eta, \boldsymbol{v}, w)) \\ &= -\rho(F(z_1) - F(z_2), \nabla \cdot \boldsymbol{v})_{\mathcal{T}_h} - (F(z_1) - F(z_2), w)_{\mathcal{T}_h},\end{aligned}$$

and therefore, from (4.54) it follows that

$$\begin{aligned}\|\mathcal{J}_h(z_1) - \mathcal{J}_h(z_2)\|_{W_h} = \|u_1 - u_2\|_{W_h} &\leq \|(\lambda_1 - \lambda_2, g_1 - g_2, \boldsymbol{q}_1 - \boldsymbol{q}_1, u_1 - u_2)\|_H \\ &\leq \frac{C}{(\alpha - \alpha_T)} \|F(z_1) - F(z_2)\|_{0,\Omega_h} \\ &\leq \frac{(\rho + 1) L_F}{(\alpha - \alpha_T)} \|z_1 - z_2\|_{W_h}.\end{aligned}$$

Therefore, if

$$\frac{(\rho + 1) L_F}{(\alpha - \alpha_T)} < 1,$$

then the mapping is a contraction and the result follows from Banach's fixed-point theorem. ■

Conclusions, and future work

In this thesis we analyzed nonlinear elliptic problems of physical interest. More precisely, those from plasma physics, in the form

$$-\nabla\cdot(\kappa(u,\nabla u))\nabla u) = \begin{cases} F(u) & \text{in } \Omega_P(u) \\ I_i & \text{in } \Omega_{C_i} \\ 0 & \text{elsewhere} \end{cases}, \qquad (**)$$

subject to domains with curved boundaries. Due to the complexity of the domain, we used a high-order transfer technique for the boundary data. This technique is known as *transferring paths* and was proposed in [17], in which the definition of flux is used. Using this approach, we developed new optimally-convergent Hybridizable Discontinuous Galerkin (HDG) methods for all the problems analyzed. Below we present details of the study carried out for the different situations that derive from the problem (**).

In Chapter 2, we considered the case in which the parameter κ is a positive function independent of the solution, and we provide a rigorous justification for the numerical results obtained in [73, 74], where the use of transfer techniques was applied to semi-linear problems in curved geometries. To prove stability of the discretizations proposed for these kinds of problems, we made certain assumptions about the *transfer paths* that appear in (2.8). In particular, (2.8b) imposes the geometric constraint that the family of triangulations should be such that the distance between the computational boundary and the true boundary remains locally of the same order of magnitude as the face mesh parameter. The assumption (2.8d) establishes that the minimum size of κ determines an upper bound for the admissible mesh size and the distance between the real and artificial boundary, that is, the smaller $\underline{\kappa}$ is, the finer the mesh must be and therefore the distance between Γ and Γ_h is reduced.

Under the assumptions required in (2.8) and a duality argument given in Section 1.5 we present an error analysis *a priori* that guarantees the order of convergence. Additionally, we established a non-linear local post-processing of the scalar unknown that guarantees an additional order of convergence. This result is corroborated with the *a posteriori* error estimator presented in Section 2.5.2. This estimator is shown to be reliable and locally efficient and includes the approximation error between the real and transferred boundary data.

In Chapter 3, we extended the analysis carried out in [71], considering the source term and the diffusion coefficient κ as non-linear. In practice, this arises due to the presence of ferroelectric materials.

In Sections 3.2, and 3.3, we independently analyzed the cases when $\kappa = \kappa(u)$ and $\kappa = \kappa(\nabla u)$, respectively. The study for both cases is not trivial, since new terms appear both in the stability analysis and in the error analysis, however, we corroborate that, under certain assumptions about the source term and the computational domain—that appear by the techniques of transfer—both schemes are well-posed. We also provide *a priori* error estimates that guarantee the optimal order of convergence for the discrete solution, provided that the distance between the curved boundary and the computational boundary are of the same order of magnitude as the mesh parameter. Motivated by what was done in [71], we are interested in performing an error analysis *a posteriori* for κ variable.

In Chapters 2 and 3 we analyzed the interior problem of (**), that is, we only consider the first row of said equation, known as the Grad–Shafranov equation. However, the results obtained are not limited to plasma applications and remain valid for general semi-linear elliptic equations.

In Chapter 4, the discrete method proposed in [19] was developed and we added the difficulty of including a non-linear source term. It is worth mentioning that the authors, in [19] coupled the HDG method with the spectral method in [60] by means of the transfer technique used in both Chapter 2 and Chapter 3 however they did not present any analysis at a theoretical level. In this chapter, we studied the coupled problem associated to (**) that involves an interior problem analyzed by HDG discretization and an external problem in which we use BEM. The coupling occurs through continuity conditions on the traces and normal fluxes of Γ. Therefore, we proposed an almost symmetric coupled BEM-HDG discretization. The presence of perturbation terms that appear due to the transfer of data between the non-matching grids was analyzed and we showed that their norm is limited by terms proportional to the size of the mesh. Finally, we proved that the coupled BEM-HDG discretization has a unique solution as long as the Lipschitz constant of the sources and the discretization mesh are small enough . It is important to mention that throughout Chapter 4, the stabilization parameter τ from the HDG scheme was considered to be of order h. However, the authors in [19] reported optimal experimental rates of convergence when τ is of order one. This is why we intend to extend our analysis to this case.

We are currently working on the analysis of the coupled problem at a continuous level and we intend to obtain error estimates showing optimal order of convergence. Another alternative is to analyze a different integro-differential formulation of the coupled problem as a non-symmetric system of three unknowns.

Conclusiones y trabajos futuros

En esta tesis analizamos problemas elípticos no lineales de interés físico. Más precismante, aquellos provenientes de la física de plasmas, de la forma

$$-\nabla \cdot (\kappa(u, \nabla u))\nabla u) = \begin{cases} F(u) & \text{en } \Omega_P(u) \\ I_i & \text{en } \Omega_{C_i} \\ 0 & \text{en otras partes} \end{cases}, \qquad (**)$$

sujetos a dominios con fronteras curvas. Debido a la complejidad del dominio, utilizamos una técnica de transferencia de alto orden para los datos de frontera. Esta técnica es conocida como *caminos de transferencia* y fue propuesta en [17], en la cual se usa la definición del flujo. Usando este enfoque, desarrollamos nuevos métodos de Galerkin Discontinuo Hibridizable (HDG) óptimamente convergente para todos los problemas analizados. A continuación presentamos detalles del estudio realizado para las diferentes situaciones que se derivan del problema (**) .

En el Capítulo 2, consideramos el caso en que el parámtro κ es una función positiva independiente de la solución, y proporcionamos una justificación rigurosa de los resultados numéricos obtenidos en [73, 74], donde el uso de las técnicas de transferencia es aplicada a problemas semi-lineales en geometrias curvas. Para probar la estabilidad de la discretización prupuesta para este tipo de problemas, adoptamos ciertas suposiciones sobre los *caminos de transferencia* que aparecen en (2.8). En particular, (2.8b) impone la restricción geométrica de que la familia de triangulaciones debe ser tal que la distancia entre la frontera computacional y la frontera original, mantengan localmente, el mismo orden de magnitud que el parámetro de malla en las caras, h_e. La condición (2.8d) establece que el tamaño mínimo de κ determina una cota superior para el tamaño de malla admisible y la distancia entre la frontera real y artificial, es decir, mientras más pequeño es $\underline{\kappa}$, más fina debe de ser la malla y por lo tanto la distancia ente Γ y Γ_h se reduce.

Bajo las suposiciones requeridas en (2.8) y un argumento de dualidad dado en la Sección 1.5 presentamos un análisis de error *a priori* que garantiza el orden de convergencia. Adicionalmente, establecemos un posprocesamiento local no lineal de la incógnita escalar que garantiza un orden adicional de convergencia. Este resultado se corrobora con un estimador de error *a posteriori* presentado en la Sección 2.5.2. Se muestra que este estimador es confiable y localmente eficiente e incluye el error de aproximacón entre los datos de la frontera original y transferida.

En el Capítulo 3, extendemos el análisis realizado en [71], considerando al término fuente y al coeficiente de difusión κ como no lineales. En la práctica, ésto surge debido a la presencia de material ferroeléctrico. En la Sección 3.2, y 3.3, analizamos de manera independiente, los casos cuando $\kappa = \kappa(u)$

y $\kappa = \kappa(\nabla u)$, respectivamente. El estudio para ambos casos no es trivial, pues aparecen nuevos términos tanto en el análisis de estabilidad como en el análisis de error, sin embargo, corroboramos que, bajo ciertos supuestos sobre el término fuente y el dominio computacional—que aparecen por las técnicas de transferencia—ambos esquemas están bien puestos. Proporcionamos también estimaciones de error *a priori* que garantizan el orden de convergencia óptimo para la solución discreta, siempre que la distancia entre la frontera curva y la frontera computacional sean del mismo orden de magnitud que el parámetro de la malla. Motivados por lo realizado en [71], estamos interesados en realizar un análisis de error *a posteriori* para κ variable.

En los Capítulos 2 y 3 analizamos el problema interior de (**), es decir sólo consideramos la primera fila de dicha ecuación, conocida como ecuación de Grad Shafranov. Sin embargo, los resultados obtenidos, no se limitan a las aplicaciones de plasma y siguen siendo válidos para ecuaciones elípticas semilineales generales.

En el Capítulo 4, se desarrolló el método discreto propuesto en [19] y agregamos la dificultad de incluir un término fuente no lineal. Cabe mencionar que los autores en [19] acoplaron el método HDG con el método espectral en [60] mediante la técnica de transferencia utilizada tanto en el Capítulo 2 como el Capítulo 3, sin embargo no presentaron ningún análisis a nivel teórico. En este capítulo, estudiamos el problema acoplado asociado a (**) que involucra un problema interior analizado mediante una discretización HDG y un problema exterior en el que utilizamos BEM. El acoplamiento se da mediante condiciones de continuidad sobre las trazas y flujos normales de Γ. Por consiguiente, propusimos una discretización BEM-HDG acoplada casi simétrica. La presencia de términos de perturbación que aparecen debido a la transferencia entre las mallas no ajustadas fue analizada y mostramos que su norma está limitada por términos proporcionales al tamaño de la malla. Finalmente, probamos que la discretización BEM-HDG acoplada tiene solución única siempre que la constante de Lipschitz del término fuente y la malla de discretización sean lo suficientemente pequeñas. Es importante mencionar que a los largo del Capítulo 4, se consideró al parámetro de estabilización τ, proveniente del esquema HDG, de orden h. Sin embargo, los autores en [19] reportaron tasas de convergencias óptimas cuando τ es de orden 1. Es por ello que pretendemos extender nuestro analisis a este caso.

Actualmente estamos trabajando en el análisis del problema acoplado a nivel continuo y pretendemos obtener las estimaciones de error *a priori* mostrando orden de convergencia óptimo. Otra de las alternativas, es analizar una formulación integral diferente del problema acoplado como un sistema no simétrico de tres incógnitas.

References

[1] D. Adak, S. Natarajan, and E. Natarajan. Virtual element method for semilinear elliptic problems on polygonal meshes. Applied Numerical Mathematics, 145:175–187, 2019.

[2] M. Amrein. Adaptive fixed point iterations for semilinear elliptic partial differential equations. Calcolo, 56(30), 2019.

[3] M. Amrein and T. P. Wihler. Fully adaptive Newton-Galerkin methods for semilinear elliptic partial differential equations. SIAM Journal on Scientific Computing, 37(4):A1637–A1657, 2015.

[4] L. Artsimovich, G. Bobrovskii, and E. Gorbunov. Experiments in Tokamak devices. Plasma Physics and Controlled Nuclear Fusion Research, 1:157–173, 1969.

[5] J. Blum. Numerical simulation and optimal control in plasma physics: with applications to Tokamaks. Wiley/Gauthier-Villars series in modern applied mathematics. Gauthier-Villars ; J. Wiley, Paris : Chichester ; New York, 1989.

[6] J. Bramble, T. Dupont, and V. Thomé. Projection methods for Dirichlet's problem in approximating polygonal domains with boundary-value corrections. Math. Comp., 26:869–879, 1972.

[7] S. Brenner and R. Scott. The mathematical theory of finite element methods, volume 15 of Texts in Applied Mathematics. Springer, New York, third edition, 2008.

[8] Brown, Thomas S., Sánchez-Vizuet, Tonatiuh, and Sayas, Francisco-Javier. Evolution of a semi-discrete system modeling the scattering of acoustic waves by a piezoelectric solid. ESAIM: Mathematical Modeling and Numerical Analysis (M2AN), 52(2):423–455, 2018.

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

[10] E. Burman and P. Hansbo. Fictitious domain finite element methods using cut elements: I. a stabilized lagrange multiplier method. Comput. Methods Appl. Mech. Engrg., 199(41):2680–2686, 2010.

[11] E. Burman and P. Hansbo. Fictitious domain finite element methods using cut elements: Ii. a stabilized Nitsche method. Appl. Numer. Math., 62(4):328–341, 2012. third Chilean Workshop on Numerical Analysis of Partial Differential Equations (WONAPDE 2010).

[12] E. Burman and P. Hansbo. Fictitious domain methods using cut elements: Iii. a stabilized Nitsche method for Stoke's problem. ESAIM: Math. Model. Numer. Anal., 48(3):859–874, 2014.

[13] P. Clément. Approximation by finite element functions using local regularization. ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, 9(R-2):77–84, 1975.

[14] B. Cockburn. The hybridizable discontinuous Galerkin methods. In Proceedings of the International Congress of Mathematicians., volume 4, pages 2749–2775, India, 2010. Hyderabad.

[15] B. Cockburn, J. Gopalakrishnan, and R. D. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal., 47:1319–1365, 2009.

[16] B. Cockburn, J. Gopalakrishnan, and F. Sayas. A projection-based error analysis of HDG methods. Math. Comp., 79(271):1351–1367, 2010.

[17] B. Cockburn, D. Gupta, and F. Reitich. Boundary-conforming discontinuous Galerkin methods via extensions from subdomains. Journal of Scientific Computing, 42(1):144–184, Aug 2009.

[18] B. Cockburn, W. Qiu, and M. Solano. A priori error analysis for HDG methods using extensions from subdomains to achieve boundary conformity. Mathematics of computation, 83(286):665–699, 2014.

[19] B. Cockburn, F.-J. Sayas, and M. Solano. Coupling at a distance HDG and BEM. SIAM Journal on Scientific Computing, 34(1):A28–A47, 2012.

[20] B. Cockburn, J. Singler, and Y. Zhang. Interpolatory HDG method for parabolic semilinear PDEs. Journal of Scientific Computing, (79):1777–1800, 2019.

[21] B. Cockburn and M. Solano. Solving Dirichlet boundary-value problems on curved domains by extensions from subdomains. SIAM Journal on Scientific Computing, 34(1):A497–A519, 2012.

[22] B. Cockburn and M. Solano. Solving convection-diffusion problems on curved domains by extensions from subdomain. Journal of Scientific Computing, 59:512–543, 2014.

[23] B. Cockburn and W. Zhang. A posteriori error estimates for HDG methods. J. Sci. Comput., 51(3):582–607, 2012.

[24] B. Cockburn and W. Zhang. A posteriori error analysis for hybridizable discontinuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal., 51(1):676–693, 2013.

[25] B. Cockburn and W. Zhang. An a posteriori error estimate for the variable-degree Raviart-Thomas method. Math. Comp., 83(287):1063–1082, 2014.

[26] M. Costabel. Symmetric methods for the coupling of finite elements and boundary elements (invited contribution). In Mathematical and Computational Aspects, pages 411–420. Springer Berlin Heidelberg, 1987.

[27] R. Z. Dautov and E. M. Fedotov. Abstract theory of hybridizable discontinuous Galerkin methods for second-order quasilinear elliptic problems. Computational Mathematics and Mathematical Physics, 54(3):474–490, mar 2014.

[28] M. David, C. Ocampo Martínez, and R. Sánchez Peña. Advances in alkaline water electrolyzers: A review. Journal of Energy Storage, 23:392–403, 2019.

[29] D. A. Di Pietro and A. Ern. Mathematical aspects of discontinuous Galerkin methods, volume 69. Springer Science & Business Media, 2011.

[30] H. Elman, J. Liang, and T. Sánchez-Vizuet. Surrogate approximation of the Grad–Shafranov free boundary problem via stochastic collocation on sparse grids. Journal of Computational Physics, 2021. In press.

[31] F. M. Fowkes. Attractive forces at interfaces. Industrial and Engineering Chemistry, 56(12):40–52, 1964.

[32] J. Freidberg. Plasma Physics and Fusion Energy. Cambridge University Press, 2008.

[33] G. N. Gatica. A simple introduction to the mixed finite element method: theory and applications. Springer Briefs in Mathematics. Springer, Heidelberg, 2014.

[34] G. N. Gatica, J. A. Almonacid, and R. Oyarzúa. A mixed-primal finite element method for the boussinesq problem with temperature-dependent viscosity. Computer Methods in Applied Mechanics and Engineering, 55(3):411–438, 2018.

[35] G. N. Gatica, B. Gomez-Vargas, and R. Ruiz-Baier. Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems. Computer Methods in Applied Mechanics and Engineering, 13:411–438, 2018.

[36] G. N. Gatica, H. Norbert, and M. Salim. On the numerical analysis of nonlinear twofold saddle point problems. Journal of Numerical Analysis, 23:301–330, 2013.

[37] G. N. Gatica and F. A. Sequeira. Analysis of an augmented HDG method for a class of Quasi-Newtonian Stokes flows. Journal of Scientific Computing, 65(3):1270–1308, Mar. 2015.

[38] G. N. Gatica and F. A. Sequeira. A priori and a posteriori error analyses of an augmented HDG method for a class of Quasi-Newtonian Stokes flows. J. Sci. Comput., 69:1192–1250, 2016.

[39] G. N. Gatica and W. L. Wendland. Coupling of mixed finite elements and boundary elements for linear and nonlinear elliptic problems. Applicable Analysis, 63(1-2):39–75, 1996.

[40] R. J. Goldston and P. H. Rutherford. Introduction to plasma physics. Institute of Physics Publishing, Bristol, U.K., 1995.

[41] H. Grad and H. Rubin. Hydromagnetic equilibria and force-free fields. In Proc. Second international conference on the peaceful uses of atomic energy, Geneva, volume 31,190, New York, Oct 1958. United Nations.

[42] E. M. Harrell and W. J. Layton. L2 estimates for Galerkin methods for semilinear elliptic equations. SIAM Journal on Numerical Analysis, 24(1):52–58, 1987.

[43] P. Heid and T. P. Wihler. Adaptive iterative linearization Galerkin methods for nonlinear problems. Mathematics of Computation, 89(326):2707–2734, July 2020.

[44] P. Houston and T. P. Wihler. An *hp*-adaptive Newton-discontinuous-Galerkin finite element approach for semilinear elliptic boundary value problems. Mathematics of Computation, (87), 2018.

[45] G. C. Hsiao, T. Sánchez-Vizuet, and F.-J. Sayas. Boundary and coupled boundary-finite element methods for transient wave-structure interaction. IMA Journal of Numerical Analysis, 37(1):237–265, 2016.

[46] G. C. Hsiao, T. Sánchez-Vizuet, F.-J. Sayas, and R. J. Weinacht. A time-dependent wave-thermoelastic solid interaction. IMA Journal of Numerical Analysis, 39(2):924–956, 04 2018.

[47] G. C. Hsiao, O. Steinbach, and W. L. Wendland. Boundary Element Methods: Foundation and Error Analysis, pages 1–62. American Cancer Society, 2017.

[48] G. C. Hsiao and W. L. Wendland. Boundary Element Methods: Foundation and Error Analysis, chapter 12. John Wiley & Sons, 2004.

[49] C. Johnson and J. C. Nédélec. On the Coupling of Boundary Integral and Finite Element Methods. Mathematics of Computation, 35(152):1063–1079, 1980.

[50] O. A. Karakashian and F. Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. SIAM J. Numer. Anal., 41(6):2374–2399, 2003.

[51] I. Langmuir. Oscillations in ionized gases. Proceedings of the National Academy of Sciences, 14(8):627–637, 1928.

[52] C. Lehrenfeld and A. Reusken. High order unfitted finite element methods for interface problems and PDE's on surfaces. Springer International Publishing, Cham., pages 33–63, 2017.

[53] M. Lenoir. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. SIAM J. Numer. Anal., 23(3):562–580, 1986.

[54] R. J. LeVeque and Z. Li. Immersed interface methods for stokes flow with elastic boundaries or surface tension. SIAM J. Sci. Comput., 18:709–735, 1997.

[55] B. S. Liley, S. Potter, and M. C. Kelley. Plasma. Technical report, 2021. https://www.britannica.com/science/plasma-state-of-matter.

[56] R. Lüst and A. Schlüter. Axialsymmetrische magnetohydrodynamische Gleichgewichts konfigurationen. Z. Naturf, 12a:850–854, 1957.

[57] W. MCLean. Strongly elliptic systems and boundary integral equations. Cambridge University Press, Cambridge, UK, 2002.

[58] S. Meddahi. An optimal iterative process for the Johnson-Nédélec method of coupling boundary and finite elements. Journal on Numerical Analysis, 35(4):1393–1415, 1998.

[59] S. Meddahi, M. González, and P. Pérez. On a fem–bem formulation for an exterior quasilinear problem in the plane. SIAM Journal on Numerical Analysis, 37(6):1820–1837, 2000.

[60] S. Meddahi and A. Márquez. A combination of spectral and finite elements for an exterior problem in the plane. Applied Numerical Mathematics, 43(3):275–295, 2002.

[61] B. Müller, S. Kräme-Eis, F. Kummer, and M. Oberlack. A high-order discontinuous Galerkin method for compressible flows with immersed boundaries. Internat. J. Numer. Methods Engrg., 110(1):3–30, 2017.

[62] N. Nguyen, J. Peraire, M. Solano, and S. Terrana. An HDG method for non-matching meshes. Preprint ,Centro de Investigación de Ongeniería MAtemática (CI^{2}MA), 2020-09.

[63] J. Nitsche. Über ein variationsprinzip zur lösung von Dirichlet-problemen bei verwendung von teilräumen, die keinen randbedingungen unterworfen sind. Abh. Math. Semin. Univ. Hambg., 36(1):9–15, 1971.

[64] R. Oyarzúa, M. Solano, and P. Zúñiga. A high order mixed-FEM for diffusion problems on curved domains. Journal of Scientific Computing, 79(1):49–78, 2019.

[65] R. Oyarzúa, M. Solano, and P. Zúñiga. A priori and a posteriori error analyses of a high order unfitted mixed-fem for stokes flow. Computer Methods in Applied Mechanics and Engineering, 360:112780, 2020.

[66] C. S. Peskin. Flow patterns around heart valves: A numerical method. Journal of Computational Physics, 10:252–271, 1972.

[67] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In Lecture Notes in Mathematics, pages 292–315. Springer Berlin Heidelberg, 1977.

[68] N. Sánchez, T. Sánchez-Vizuet, and M. E. Solano. Afternote to "*Coupling at a distance*": convergence analysis and *a priori* error estimates. (In preparation), 2021.

[69] N. Sánchez, T. Sánchez-Vizuet, and M. E. Solano. Analysis of a coupled HDG-BEM formulation for non-linear elliptic problems with curved interfaces. (In preparation), 2021.

[70] N. Sánchez, T. Sánchez-Vizuet, and M. E. Solano. Error analysis of an unfitted HDG method for a class of non-linear elliptic problems. (Submitted), 2021. https://arxiv.org/abs/2105.03560.

[71] N. Sánchez, T. Sánchez-Vizuet, and M. E. Solano. A priori and a posteriori error analysis of an unfitted HDG method for semi-linear elliptic problems in curved domains. Numerische Mathematik, 148:919–958, 2021.

[72] T. Sánchez-Vizuet and F.-J. Sayas. Symmetric boundary-finite element discretization of time dependent acoustic scattering by elastic obstacles with piezoelectric behavior. Journal of Scientific Computing, 70(3):1290–1315, 2017.

[73] T. Sánchez-Vizuet and M. E. Solano. A hybridizable discontinuous Galerkin solver for the Grad-Shafranov equation. Computer Physics Communications, 235:120–132, Feb 2019.

[74] T. Sánchez-Vizuet, M. E. Solano, and A. J. Cerfon. Adaptive hybridizable discontinuous Galerkin discretization of the Grad–Shafranov equation by extension from polygonal subdomains. Computer Physics Communications, 255:107239, 2020.

[75] J. Saranen and G. Vainikko. Periodic Integral and Pseudodifferential Equations with Numerical Approximation. Springer Berlin Heidelberg, 2002.

[76] S. A. Sauter and C. Schwab. Boundary Element Methods. Springer Berlin Heidelberg, 2011.

[77] F.-J. Sayas. The validity of Johnson-Nédélec's BEM-FEM coupling on polygonal interfaces. SIAM J. Numer. Anal., 47(5):3451–3463, 2009.

[78] V. D. Shafranov. On magnetohydrodynamical equilibrium configurations. Soviet Physics JETP, 6:545–554, 1958.

[79] F. E. Tabarés. Plasma Applications for Material Modification:. Jenny Stanford Publishing, New York, 2021.

[80] V. Thomée. Polygonal domain approximation in Dirichlet's problem. J. Inst. Math. Appl., 44:33–44, 1973.

[81] L. N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. SIAM Review, 56(3):385–458, 2014.

[82] R. Verfürth. A posteriori error estimators for convection-diffusion equations. Numer. Math., 80:641–663, 1998.

[83] M. R. Wertheimer, A. C. Fozza, and A. Holländer. Industrial processing of polymers by low-pressure plasmas: the role of vuv radiation. Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, 151(1-4):65–75, 1999.

[84] J. Wesson. Tokamaks. Oxford University Press, fourth edition, 2011.

[85] Z. Xie and C. Chen. The interpolated coefficient FEM and its application in computing the multiple solutions of semilinear elliptic problems. International Journal of Numerical Analysis and Modeling, 2(1):97–106, 2005.

[86] J. Xu. A novel two-Grid method for semilinear elliptic equations. SIAM Journal on Scientific Computing, 15(1):231–237, 1994.

[87] J. Xu. Two-Grid discretization techniques for linear and nonlinear PDEs. SIAM Journal on Numerical Analysis, 33(5):1759–1777, 1996.

[88] J. Zhan, L. Zhong, and J. Peng. Discontinuous Galerkin methods for semilinear elliptic boundary value problem, 2021. arXiv: 2101.10664.